University of Essex

MSc in Artificial Intelligence

September 2024

Mitigating machine learning bias in criminal justice: An ontological approach to predicting recidivism in England and Wales

Author: Leigh Feaviour

First supervisor: Godfried Williams

Second supervisor: Samuel Danso

Word count: 13,918

Abstract

Recidivism is when someone who has previously been convicted of a crime reoffends. Reducing recidivism has benefits for society by reducing crime and its socioeconomic costs. Machine learning has been used to predict recidivism, but the impact of getting it wrong has significant implications. However, machine learning models in judicial use and academia have been found wanting, with racial biases caused by the underlying data and a paradoxical conflict between accuracy and fairness.

Ontology is proposed as a mitigation to reduce bias by classifying features according to their "safe use" based upon metaknowledge of the degree of correlation between the feature and a protected characteristic such as race. The ontology also allows the prediction results to be compared with the actual profiles for ethnicity and gender to ensure parity.

The ontology concept was initially tested using an anonymised recidivism dataset from the USA. Statistical analysis identified relationships between age, race, gender, and recidivism and calculated the correlations between all features and race and recidivism by gender. The results were used to classify the features in the ontology to guide data scientists creating machine learning algorithms to minimise the risk of bias. An ontology was then designed and built for England and Wales using publicly available metadata plus features from other studies to demonstrate the concept working with the complex data model of a real-world application.

There were challenges storing the feature correlations in the ontology because ontologies store knowledge, not metaknowledge. This was addressed by storing the metaknowledge as data properties, but it created separation between the knowledge and metaknowledge which would be prone to human error, so an enhancement to the semantic web is recommended.

The study shows that ontology is a practicable way of mitigating the risk of bias in machine learning models by identifying high-risk features to avoid (transparency of input) and then ensuring that results conform to expected distributions by comparing the ratios of positive and negative predictions with the training data across the protected classes (transparency of output). However, limitations with storing metaknowledge were discovered which meant the ontology was less robust than planned.

Future work is recommended to populate the ontology with actual data from England and Wales, to then train and test machine learning solutions and feed the results back into the ontology to fully prove the concept end-to-end. Modifications to the semantic web are also recommended to provide a means to store metaknowledge as knowledge within the ontology.

Table of Contents

Table of Figures

Files are available at https://github.com/feaviolp/msc-project/

1 Introduction

A standard definition of recidivism is "whether a person returns to prison within three years of release" King & Elderbroom (2014: 2). This is an overly simplistic measure, however, and a more rounded definition would include "rearrest, reconviction, reincarceration for a new crime, or return to prison [for the same crime]" King & Elderbroom (2014: 2), who also note that observing for longer than three years would provide more data that could be used to analyse links between supervision and reoffending. For this study, the simple definition of "a prisoner returning to prison within three years of release" will be used unless otherwise stated.

Lowering recidivism is intuitively beneficial because it logically correlates to reducing crime, which is desirable in society. There are political and financial reasons too, including the justice reinvestment initiative in the United States of America (USA), whose goals include reducing the prison population, in part by lowering recidivism, to use the savings to invest in public safety (La Vigne et al., 2014). So, there is a socioeconomic case to reduce recidivism, which means that predicting who will recidivate is vital to providing interventions for targeted rehabilitation or for deciding who to release on parole.

There are machine learning models to predict recidivism. However, there are important questions about ethics and fairness, which are critical in this domain because getting it wrong can have serious implications; unnecessarily depriving someone of their liberty, or releasing someone who is a danger to the public. There are challenges with providing fair and ethical models due to the available features (Dressel & Farid, 2018), biases in the baseline data (Wadsworth et al., 2018; Biddle, 2022), and even the metrics chosen to assess the results (Caton & Haas, 2020). These challenges are analysed in a literature review with mitigation strategies suggested.

This project proposes the use of ontology to mitigate biases, focusing mainly on racial bias. A detailed statistical analysis is provided using publicly available anonymised American data to determine how the features correlate with recidivism and race. A proof of concept ontology is then built in Protégé using the same dataset. Having proven the concept, a more robust ontology is built in Protégé using metadata from England and Wales criminal justice, with additional features from the literature review.

Research question: “Can ontology mitigate bias when using machine learning to predict recidivism?”

Aim: Machine learning is used to predict recidivism, but previous studies have indicated ethical issues such as racial bias. This study will show if biases can be identified and mitigated with the use of ontology by creating an ontology of criminal justice in England and Wales. Features will be identified for safely predicting recidivism and features to be used with caution. Furthermore, the protected characteristic profiles can be compared with predicted profiles to check for parity.

Objectives:

- Identify features that predict recidivism.
- Identify features that potentially introduce biases when predicting recidivism.
- Assess if recidivism varies between characteristics such as ethnicity, gender and age.
- Create an ontology of criminal justice using available metadata and illustrate how the ontology can manage the features to reduce biases, as well as highlight potential biases in the output to further mitigate bias risks.

2 Ethical and Professional Considerations

This project did not include primary research, so considerations regarding participant safety including data storage and access were not relevant.

Ethical and professional considerations were assessed using the BCS code of conduct (BCS, 2022) (figure1).

	Code of conduct. You shall:	Compliance statement for this project
Public Interest	have due regard for public health, privacy, security and wellbeing of others and the environment.	N/A. No opportunity
	have due regard for the legitimate rights of Third Parties.	Compliant. No direct third-party interaction, but rights always considered
	conduct your professional activities without discrimination on the grounds of sex, sexual orientation, marital status, nationality, colour, race, ethnic origin, religion, age or disability, or of any other condition or requirement.	Compliant. This project aims to reduce discrimination
	promote equal access to the benefits of IT and seek to promote the inclusion of all sectors in society wherever opportunities arise.	N/A. The project does not discuss who would use the solution

Professional Competence and Integrity	only undertake to do work or provide a service that is within your professional competence.	Compliant. The project is an extension of learning undertaken on the MSc
	NOT claim any level of competence that you do not possess.	Compliant
	develop your professional knowledge, skills and competence on a continuing basis, maintaining awareness of technological developments, procedures, and standards that are relevant to your field.	Compliant. The project, including this report, are examples
	ensure that you have the knowledge and understanding of Legislation and that you comply with such Legislation, in carrying out your professional responsibilities.	Compliant. Legislation has been considered and discussed
	respect and value alternative viewpoints and, seek, accept and offer honest criticisms of work.	Compliant. Input from supervisors has shaped the scope
	avoid injuring others, their property, reputation, or employment by false or malicious or negligent action or inaction.	Compliant
	reject and will not make any offer of bribery or unethical inducement.	Compliant
Duty to Relevant Authority	carry out your professional responsibilities with due care and diligence in accordance with the Relevant Authority's requirements whilst exercising your professional judgement at all times.	Compliant
	seek to avoid any situation that may give rise to a conflict of interest between you and your Relevant Authority.	Compliant
	accept professional responsibility for your work and for the work of colleagues who are defined in a given context as working under your supervision.	Compliant
	NOT disclose or authorise to be disclosed, or use for personal gain, or to benefit a third party, confidential information except with the permission of your Relevant Authority, or as required by Legislation.	Compliant. No personal data used in the project. All data was public domain.
	NOT misrepresent or withhold information on the performance of products, systems or services (unless lawfully bound by a duty of confidentiality not to disclose such information), or take advantage of the lack of relevant knowledge or inexperience of others.	Compliant
Duty to the Profession	accept your personal duty to uphold the reputation of the profession and not take any action which could bring the profession into disrepute.	Compliant. All project activities were professional
	seek to improve professional standards through participation in their development, use and enforcement.	N/A. No opportunity
	uphold the reputation and good standing of BCS, the Chartered Institute for IT.	Compliant. The project addresses bias which is reputationally positive
	act with integrity and respect in your professional relationships with all members of BCS and with members of other professions with whom you work in a professional capacity.	Compliant. All research and supervisor discussions were respectful and professional
	encourage and support fellow members in their professional development.	N/A. No opportunity

Figure 1. Compliance with BCS code of conduct (BCS, 2022)

3 Literature Review

Reducing recidivism is the topic of many studies, as are the potentially biased results. This literature review examines how recidivism has been predicted and how machine learning has helped. It then examines the ethics and fairness of using machine learning to predict recidivism. It looks at how racial and other biases are inextricably intertwined in the results, with no meaningful way to completely untangle them. The importance of transparency is then explored as a mitigation to bias. Finally, the use of ontologies to improve ethics and fairness is briefly examined, although this is an area with little progress to date, hence the topic of this dissertation.

3.1 Why and How to Predict Recidivism?

Predicting recidivism for parole decisions has been used in America since the 1920s, using factors such as age, intelligence, nationality and criminal history (Borden, 1928), leading to the birth of actuarial measures (Burgess, 1928). Interestingly, Borden (1928) found higher intelligence had a positive correlation with higher recidivism, which is the opposite of more recent studies (Richter et al., 1996; Lotar Rihtarić et al., 2017), which could be due to demographic changes in the prison population over the past one hundred years, or other societal factors. This demonstrates that factors used to assess recidivism and how they are used must be adapted over time to remain relevant. This is an important point that will be discussed later.

Correctional Service of Canada (CSC) uses Statistical Information on Recidivism – Revised (SIR-R1) to assess the likelihood of reoffending within three years by scoring 15 items (Figure 2). The tool is considered effective because it gives results that are “17% better than chance” (Nafekh & Motiuk, 2002: ii). However, the results did not differentiate between false positives and false negatives, so it is unclear if the tool is skewed in either direction. Examining the outcome measures (Figure 3), the very good group (predicted to succeed) are significantly more accurate than the poor group (predicted to fail), which suggests a higher proportion of false positives overall. Additionally, comparing to chance is not as compelling as comparing to expert judgment without the tool, which is the natural alternative.

1. Current Offence
2. Age at Admission
3. Previous Incarceration
4. Revocation or Forfeiture
5. Act of Escape
6. Security Classification
7. Age at First Adult Conviction
8. Previous Convictions for Assault
9. Marital Status at Most Recent Admission
10. Interval at Risk Since Last Offence
11. Number of Dependants at Most Recent Admission
12. Current Total Aggregate Sentence
13. Previous Convictions for Sex Offences
14. Previous Convictions for Break and Enter
15. Employment Status at Arrest

Figure 2. SIR-R1 scoring items (Nafekh & Motiuk, 2002: 26)

Outcome measures by SIR-RI groupings
Male Non-Aboriginal Offenders
(N=6,881)

SIR-RI Group	General recidivism***		Violent***		Sexual	
	Successes %	Failures %	Successes %	Failures %	Successes %	Failures %
Very Good	94.4	5.6	99.2	0.8	99.6	0.4
Good	84.3	15.7	98.6	1.4	99.7	0.3
Fair	75.7	24.3	95.4	4.6	99.4	0.6
Fair / Poor	68.7	31.3	94.5	5.5	99.4	0.6
Poor	56.3	43.7	93.3	6.7	99.4	0.6
All Cases	78.5	21.5	96.6	3.4	99.5	0.5

Notes: ***p<.001

Figure 3. SIR-R1 outcome measures (Nafekh & Motiuk, 2002: 41)

Brown et al. (2009) compared the static SIR-R1 predictive score with a dynamic three-wave assessment, assessing pre-release, one month post-release and three months post-release. The dynamic variables significantly increased the accuracy of the predictions, with difficulties with employment, finances and substance abuse all correlating with a higher likelihood of reoffending. A word of caution: for ethical reasons, participants had to consent to be part of

the study, with a consent rate of 56.4% (Brown et al., 2009: 27), which potentially means an unrepresentative sample, although the findings were consistent with Lloyd et al. (2020) and Davies et al. (2022).

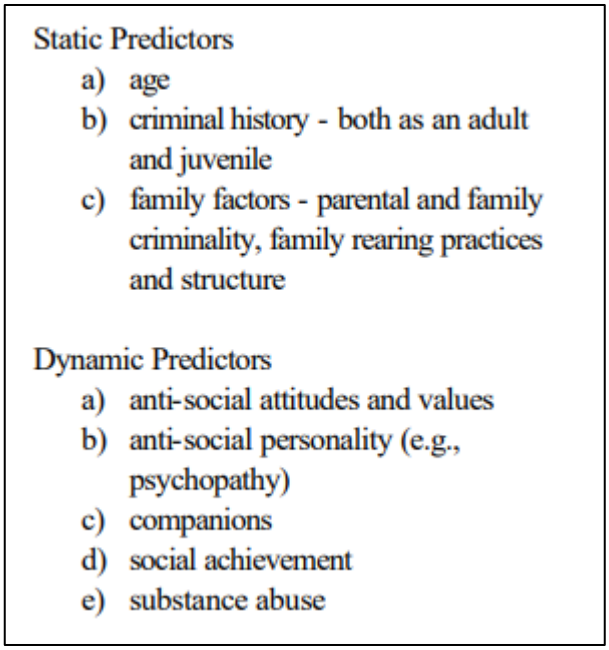
Dynamic factors have been proven to link to desistance (the opposite of recidivism). Getting married and having children (Farrington & West, 1995), obtaining employment (Farrington et al., 2017), and moving house to a lower crime area (Osborn, 1980) have all been shown to reduce patterns of offending. Andrews & Bonda (2024) posit that many biological and environmental factors cause a person to commit a crime, so it should follow that the factors to predict whether or not someone will recidivate are also complex. The unavoidable fact is that a person's propensity to reoffend is based upon dynamic and static factors, so monitoring and reassessing post-release provides a more accurate view. It could also facilitate interventions around the key indicators to allow probation officers to address issues and actively reduce recidivism rather than just predicting it.

Lloyd et al. (2020) found that regular reassessment significantly improved the accuracy of predicting recidivism. Using the Dynamic Risk Assessment for Offender Re-entry (DRAYOR; Serin, 2007), supervision officers working with ex-offenders on parole in New Zealand reassessed their scores every few weeks with improved accuracy compared with the original baseline scores. Davies et al. (2022) performed a similar study looking only at high-risk parolees with similar results. This is hardly surprising, demonstrating again that the factors that impact someone's propensity to recidivate are dynamic, changing over time according to the individual's circumstances. Furthermore, the simple passage of time itself changes the propensity to reoffend. The likelihood of reoffending reduces the longer a person remains in the community without offending, to the point where they eventually pose no greater threat than someone without a criminal conviction (Baumann et al., 2022).

Factors that increase the risk of recidivism from a Canadian study in 2010 were "young age, prior criminal history, negative peer associations, substance abuse, and antisocial personality disorder" (Hanson, 2010: 1); who also found that "Offenders of minority race were at increased risk for general and violent recidivism, but not sexual recidivism" (Hanson, 2010: 2). Tollenaar & van der Heijden (2013: 578) concurred that age is a significant factor stating "recidivism tends to decline with age; the odds lower by 3% for each extra year".

Bonta (1997) analysed 131 separate studies, producing 1,141 correlations. Characteristics that were the best recidivism predictors were criminal attitudes, criminal lifestyle, criminal

history, social achievement, age/gender/race, and family factors, concurring with Hanson (2010). However, actuarial measures using a variety of variables were found most effective, with Level of Service Inventory-Revised (LSI-R) performing best (Bonta, 1997), concluding that composite actuarial measures should be used, including the constituents in Figure 4.



Static Predictors

a) age
b) criminal history - both as an adult and juvenile
c) family factors - parental and family criminality, family rearing practices and structure

Dynamic Predictors

a) anti-social attitudes and values
b) anti-social personality (e.g., psychopathy)
c) companions
d) social achievement
e) substance abuse

Figure 4. Recommended constituents for recidivism assessment (Bonta, 1997: 2)

3.2 How is Recidivism Predicted in England and Wales?

The prison and probation service in England and Wales use various actuarial risk assessment tools, including "Risk of Serious Recidivism (RSR), OASys Sexual reoffending Predictor (OSP) and the Offender Group Reconviction Scale (OGRS)" (HM Prison & Probation Service, 2023: 6), who note that the tools need to be evidence-based, fair and without bias. The OGRS is the generic recidivism prediction tool that will be examined further.

OGRS was launched in 1996 as a manual "pen and paper" tool using six demographic and criminal history factors and updated with a computerised OGRS2 in 2000 with ten factors (Moore, 2015). In 2008, OGRS3 was introduced as an ordinal logistic regression model with fewer risk factors that can be scored more easily to be as simple as possible for probation and prison staff to use, being less prone to error while improving the predictive reliability, with Area Under Curve (AUC) of 80% compared with 78% for OGRS2 (Howard et al., 2009).

All versions used the same fundamental factors of age, gender, and offence history (Figure 5), but how they were used was adapted with each iteration. For example, OGRS3 used age when the offence was committed and age at release whereas previous versions excluded age at release (Moore, 2015), potentially skewing results for offenders with longer sentences.

An issue with OGRS is that it uses static characteristics (Stephens & Brown, 2001). The Home Office found evidence that social variables such as accommodation, employment, financial status and relationship with alcohol and drugs are "significantly related to reconviction" (May, 1999), albeit referring to reconviction following a community sentence rather than prison, but the factors should be similar. This supports the earlier view that dynamic variables impact recidivism (Davies et al., 2022; Farrington et al., 2017; Farrington & West, 1995; Hanson, 2010; Osborn, 1980). However, they state that criminal history factors such as those used in OGRS have a much stronger relationship (May, 1999), and Ministry of Justice (2024) describes how recidivism correlates with static factors such as age, gender, offence type and number of previous offences. Whilst criminal history factors might be more robust, they are static; the offender cannot influence them. Excluding dynamic factors that people can influence seems unfair, even if the impact is relatively small.

> "Like most other actuarial measures, OGRS[…] is insensitive to changes in mental health-related risk and the effects on risk of therapeutic engagement which are important considerations in practice" (Hill et al., 2024).

The effort of including dynamic factors possibly outweighed the benefit when OGRS1 was a manual calculation in 1999, and Farrington & Davies (2007) note "the need for practitioners to have a simple method to use in practice". However, even OGCR3 excludes dynamic factors, and with computerised solutions today, it would be much easier to include them.

	Version		
	1	2	3
Factors included in the model:			
Age and gender			
Gender	√	√	(AG)
Age at time of sentence	√, (C)	√, (C)	(C)
Age at release or start of order			(AG)
Combination of age and gender			√
Age at first conviction	(C)	√, (C)	(C)
Offence/offending history			
Type of offence (number of categories)	√ (9)	√ (27)	√ (20)
Number of previous convictions	(C)	(C)	
Current or previous breach		√	
Current or previous burglary		√	
Number of previous youth custodial sentences	√	√	
Number of previous sanctions (convictions and cautions/reprimands/final warnings (CRFW))			(C)
Offending history status (first conviction; other conviction; first CRFW; second CRFW, or other CRFW)			√
Is current sanction a conviction or another sanction?			(O)
'Copas rate'[108]	√	√	√

Key: √ Included in its own right; (AG) Part of age/gender: (C) Part of 'Copas rate'; (O) Part of offending history status.

Figure 5. OGRS Factors (Moore, 2015: 154).

3.3 Machine Learning to Predict Recidivism

Machine learning solutions from industry and academia were reviewed next to discover the models and approaches tested, the features used, and how they compared with the key variables already discussed.

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) (Equivant, 2019) is widely used in the USA's judicial system. It is a proprietary tool with many features, but the ones of interest here are the recidivism risk scales. COMPAS generates two scales: the General Recidivism Risk Scale (GRRS) and the Violent Recidivism Risk Scale (VRRS), with details about the differences between the scores in the practitioner's guide (Equivant, 2019). COMPAS is discussed at length in the ethics and fairness section. It was introduced at the beginning of this section because many machine learning solutions that have come since compare themselves to COMPAS.

When comparing Random Forest (RF) and XGBoost to predict recidivism, Curtis (2018) found that RF outperformed XGBoost, with AUC scores of 0.85 and 0.831 respectively, and similarly impressive accuracy and count of true positives and true negatives. However, there is a problem with measuring machine learning output: what do we wish to achieve? With 25.6% belonging to the target class of "reincarcerated" (Curtis, 2018), the dataset is imbalanced, and accuracy works best when the datasets are symmetric (Ghoneim, 2019). Area Under the Curve (AUC) is less problematic with imbalanced datasets. However, it treats the positive and negative classes equally, which might not always be desirable.

Considering the implications of false positives; offenders being wrongly predicted to reoffend and potentially losing their liberty as a result, or false negatives; offenders being wrongly predicted not to reoffend, care should be taken over which metrics to use. Indeed, "the choice of performance measure(s) itself may even harbor, disguise, or create new underlying ethical concerns" (Caton & Haas, 2020: 21). Some popular machine learning performance metrics are illustrated in Figure 6 alongside their respective implications for measuring models that predict recidivism. Curtis (2018) does not provide a confusion matrix with actual values. However, the sample size, percentage of offenders who reoffended, percentage of offenders who did not reoffend, and percentage of true positives and true negatives were provided, so the confusion matrix was reverse engineered from which precision, recall, F1-Scores and specificity were calculated (Figure 7). Here, it can be seen that both algorithms scored significantly higher on specificity than recall, meaning that true negatives were more accurately predicted than true positives. This is good news for offenders, but not for the public.

Performance measure	Summary	Implications	Favours
$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$	Overall accuracy. Treats false positive and false negatives equally.	Cannot determine if results are skewed towards over or under predicting recidivism.	Indeterminate
$Precision = \frac{TP}{TP + FP}$	How accurate the positive predictions were. Does not penalise false negatives.	Could score highly whilst still incorrectly predicting people not to recidivate who do.	Offenders
$Recall = \frac{TP}{TP + FN}$	How many of the total positive cases were predicted, without penalising for false positives.	Could score highly whilst incorrectly predicting people to recidivate who did not.	Public
$F1\ Score = \frac{2(Precision * Recall)}{Precision + Recall}$	Combination of precision and recall with a balanced score of positive predictions. However it excludes true negatives.	More balanced than accuracy, cannot determine if results are skewed towards over or under predicting recidivism.	Indeterminate
$Specificity = \frac{TN}{TN + FP}$	How many negative cases were predicted, without penalising for false negatives.	Could score highly whilst still incorrectly predicting people not to recidivate who do.	Offenders

Figure 6. Machine learning performance measures

	Random forest	XGBoost
AUC	0.85	0.831
Accuracy	76.53	76.87
True Positives %	79.52	72.99
True Negatives %	75.5	78.22
True Positives (TP)	52,571	48,254
False Positives (FP)	13,540	17,857
True Negatives (TN)	145,063	150,290
False Negatives (FN)	47,074	41,847
Precision	0.795	0.730
Recall (Sensitivity)	0.528	0.536
F1-Score	0.634	0.618
Specificity	0.915	0.894

Figure 7. Recidivism model comparison adapted from Curtis (2018)

Kovalchuk et al. (2023) compared various models using Ukrainian offender static data. They found that Decision Tree (DT) was the best for precision, recall and F-Measure (Figure 8), with the greatest computational efficiency and explainability due to its relative simplicity. The importance of explainability will be covered in more detail later when discussing ethics and fairness.

Model	Accuracy	Precision	Recall	F Measure	Sensitivity	Specificity	AUC
Naive Bayes	86.7%	78.9%	99.5%	88.0%	99.5%	74.4%	0.96
Generalized Linear Model	95.8%	92.8%	99.1%	95.8%	99.1%	92.6%	0.99
Logistic Regression	91.1%	85.0%	99.5%	91.7%	99.5%	83.1%	0.99
Fast Large Margin	80.5%	98.7%	61.3%	75.6%	61.3%	99.2%	0.99
Deep Learning	84.4%	76.1%	99.5%	86.3%	99.5%	69.6%	0.99
Decision Tree	98.3%	97.7%	98.8%	98.3%	98.8%	97.8%	0.99
Random Forest	98.3%	97.7%	98.8%	98.3%	98.8%	97.8%	0.99
Gradient Boosted Trees	98.3%	97.7%	98.8%	98.3%	98.8%	97.8%	0.99

Figure 8. Model Comparison (Kovalchuk et al., 2023: 8)

Wang et al. (2010) compared logistic regression, Artificial Neural Networks (ANN) and Support Vector Machine (SVM) using static variables. The performance varied between the models, and looking more closely at the results, specific instances of false positives and false negatives also varied more than expected between the models, so a hybrid solution combining all models was tested with improved results. The performance measures focused on sensitivity (the same as recall) and specificity, meaning they were concerned with true match rates. This is concerning because, in a real-world scenario, there would need to be consideration for the implications of false positives and false negatives, as already discussed. The hybrid solution was reported as good, with recall circa 0.4, specificity circa 0.85 and accuracy circa 0.7. However, a recall of 0.4 leaves many true positives undetected, which would not be in the public interest.

In another comparative study, Zeng et al. (2017) compared machine learning models, including various DT, RF, SVM and linear regression (LR) models using static variables commonly accessible to judicial officers. The prediction results, using AUC, were similar between the best RF and SVM models, LR and Supersparse Linear Integer Model (SLIM) (Ustun & Rudin, 2015), while DT fared worst. They recommend using SLIM because it is the simplest to use (Farrington & Davies, 2007; Tollenaar & van der Heijden, 2013) and the easiest to interpret. Zhang (2022) separately compared K Nearest Neighbour (KNN), RF, SVM and LR using static variables from the COMPAS dataset. The performance metrics focused on accuracy and AUC but with limited success.

Tollenaar & van der Heijden (2013) compared eleven models using static data and found no meaningful increase in accuracy with ML models over LR or Linear Discriminant Analysis (LDA). They found women recidivated significantly less than men, and conviction density (the number of convictions over time) was a significant predictor where “the number of previous offences […] show by far the largest effect on the probability of recidivism” (Tollenaar & van

der Heijden, 2013: 579). They also considered real-world implementation challenges, preferring a small number of variables to make prediction models quick and easy, and "regardless of the complexity of the final model, the formulation of the model can always be translated into a set of equations in an Excel spreadsheet so it can readily be used by a probation worker" (Tollenaar & van der Heijden, 2013: 582).

Tollenaar & van der Heijden (2013) also make good use of the performance metrics to assess their models for real-world use. They preferred not to use accuracy alone for reasons similar to those already discussed, and they focused on sensitivity and specificity to predict positive and negative results. This way, the cost of a false positive or false negative can be altered according to the situation. However, they note that where the cost of false positives is the same as false negatives, the highest accuracy that they could achieve was 70%, which

> "is too low to rely solely on […] for decision making about individuals. We can use the predicted score for group-based predictions […] but, when individual predictions are concerned, additional information concerning dynamic factors is required" (Tollenaar & van der Heijden, 2013: 583).

So even at 70% accuracy with static data, the results alone should not be used to make decisions about individuals, and yet that is a similar accuracy to COMPAS (Dieterich et al., 2016)!

Lin et al. (2020) tested logistic regression models using enhanced data with dynamic factors from LSI-R added to the static factors used in COMPAS. The model outperformed humans with almost 90% accuracy compared with less than 60% accuracy for humans (Miller, 2020), although that comparison is misleading because it is based upon human assessment without feedback. When the human participants were "trained" with feedback, their accuracy was over 80%, which was still lower than the LR model but significantly higher than using static data alone, which was around 70% accuracy. Adding dynamic data improves prediction accuracy (Brown et al., 2009; Davies et al., 2022; Farrington et al., 2017; Farrington & West, 1995; Hanson, 2010; Osborn, 1980); however, it also complicates the assessment.

Lin et al. (2020) tested with and without feedback for the human participants, but feedback is crucial for machine learning. Recidivism datasets show who reoffended once released, so by their nature; they are skewed because decisions have already been made about whom to release, which "makes it hard to evaluate counterfactual decision rules based on algorithmic

predictions" (Kleinberg et al., 2018). Dealing with this issue is outside the scope of this research.

3.4 Ethics and Fairness

This section discusses the ethical challenges of using machine learning to predict recidivism, starting with a detailed examination of challenges and rebuttals of COMPAS before moving on to more general considerations of ethics and fairness.

There are nine protected characteristics against which it is illegal to discriminate in England, Scotland and Wales: age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex, and sexual orientation (Equality and Human Rights Commission, 2021). Similar legislation exists elsewhere. AI-based recidivism predictors would be considered high-risk AI systems under the proposed European Commission AI Act (European Commission, 2021). As such, potential biases need to be identified and addressed (Van Dijck, 2022).

3.5 Recidivism and Race

Angwin et al. (2016) challenged COMPAS's fairness, claiming it was racially biased. They found that while the accuracy was similar for black and white people, the error types varied significantly. Black people were more likely to be predicted to reoffend who did not (false positives), and white people were more likely to be predicted not to reoffend who did (false negatives). This again illustrates the problem with using accuracy as the sole performance measure.

Dieterich et al. (2016) refuted the findings of Angwin et al. (2016), unsurprisingly considering the commercial nature of COMPAS. They claimed that COMPAS has predictive parity and that Angwin et al. (2016) were selective in what they measured and reported and "did not present any valid evidence that the risk scales are biased against blacks" (Dieterich et al., 2016: 22).

The updated practitioner's guide shows a statistically significant correlation between COMPAS predictions and the widely used LSI-R (Farabee et al., 2010). However, a closer examination of the results reveals that while Pearson's coefficient did indeed have highly

significant correlations, the actual correlations were moderate, with the highest (criminal involvement) being 0.64 (Figure 9). That does not prove that COMPAS is unreliable. However, it does illustrate that all papers need to be reviewed critically, especially when there is commercial interest in the outcome, as in this case. The conclusion is misleading even though the figures are accurate.

COMPAS	LSI-R	Correlation
Criminal Involvement	Criminal History	0.64 ($p < .0001$)
Criminal Associates/Peers	Companions	0.48 ($p < .0001$)
Substance Abuse	Alcohol/Drug Problem	0.53 ($p < .0001$)
Financial	Financial	0.49 ($p < .0001$)
Vocation/Education	Education/Employment	0.51 ($p < .0001$)
Family Criminality	Family/Marital	0.16 ($p > .10$)
Leisure	Leisure/Recreation	0.05 ($p > .10$)
Residential Instability	Accommodation	0.57 ($p < .0001$)
Criminal Attitudes	Attitudes/Orientation	0.20 ($p = .08$)

Figure 9. Correlations between COMPAS and LSI-R (Equivant, 2019: 25)

In a report commissioned by Broward Sheriff's Office, Blomberg et al. (2010) found that the predictive accuracy of COMPAS was slightly higher for black offenders, and "COMPAS is highly predictive of future recidivism" (Blomberg et al., 2010: 91). This is unsurprising when accuracy has already been found to be similar, but proven not to be a reliable measure on its own. This seems like a 98-page report to tell the Broward Sheriff's Office what they wanted to hear and what they already knew.

An LR model using seven features: age, sex, number of juvenile misdemeanours, number of juvenile felonies, number of prior (nonjuvenile) crimes, crime degree, and crime charge outperformed COMPAS, with 137 features, on accuracy and false positives (Dressel & Farid, 2018), although it still demonstrated a false positive bias towards black people. The use of 137 features has since been clarified by the company that makes COMPAS, stating "the vast number of these 137 are needs factors and are NOT used as predictors in the COMPAS risk assessment. The COMPAS risk assessment has six inputs only" (Equivant, 2018).

A simpler LR model using two features: age and total number of previous convictions, was comparable with COMPAS on accuracy and false positives, which supports the assertions that age and criminal history are the most important features for predicting recidivism (Rudin

et al., 2020). However, analysis of the data revealed that black offenders were disproportionately younger than white offenders (Rudin et al., 2018), which explains why modelling primarily on age still produced racially biased results.

Examining the data more closely, the rate of incarceration of black people was significantly higher than that of white people across the USA, within Florida (where COMPAS was validated), and within the dataset used by COMPAS (Dressel & Farid, 2021). That inescapable fact appears to drive the racial imbalance of the results. The same racial imbalance was even replicated when untrained humans were given a cut-down COMPAS dataset with seven features. Black people had more prior convictions because they were more likely to be arrested and imprisoned than white people (Fenton, 2016; Home Office, 2023; Mohdin & Garcia, 2023), so relying on prior convictions could contribute to racial bias (Wadsworth et al., 2018). Even though race was not a variable, it was encoded in the criminal history features because of the racial disparity in prisons, as eloquently described by Biddle (2022):

> "In the case of penal systems that are populated disproportionately by historically marginalized groups, if researchers collect information about individuals who are already in these systems (as all of the developers of these tools do), then the baseline data will reflect the discriminatory practices that have led to these injustices." (Biddle, 2022: 332)

However, predicting recidivism without knowing criminal history, as suggested by Wadsworth et al. (2018), would be significantly less accurate (Berk, 2012), so there is a dilemma.

Looking beyond race, socioeconomic features are predictors of recidivism (Eaglin, 2017); however, including them could imply that poorer people are at a higher risk of recidivism than wealthy people, which creates a feedback loop in sentencing that becomes self-fulfilling (Biddle, 2022), just like the apparent feedback loop for black offenders. "In the design of recidivism-prediction algorithms, the trade-offs that have been made have tended to disproportionately harm groups that are already unjustly disadvantaged" (Biddle, 2022: 399).

Delving into the abstract to illustrate the issue of an already biased dataset, Ensign et al. (2017) describe an apple-tasting example of partial monitoring, whereby tasting apples before selling ensures only apples that are good to sell are sold, but the already-tasted apple cannot be sold, which results in a loss. Not tasting them could result in bad apples being sold, resulting in a loss. Randomly sampling apples to determine the characteristics of good and

bad apples without tasting them all overcomes this problem. The relevance to recidivism prediction is that, to create a model that accurately predicts recidivism without bias, a random sample must be taken, which equates to releasing parolees or not at random to see who will recidivate. In reality, a judgement is made about whom to release and when, so there is already an inherent bias in the data. This is exacerbated by historic racial inequalities whereby a disproportionate number of black people have been convicted of crimes, which makes the criminal dataset racially imbalanced (Thomas, 2023).

Soares & Angelov (2019) developed a prototype-based algorithm that modelled black and white defendants separately using the COMPAS dataset. The accuracy of predictions for black defendants, white defendants, and overall was improved. Also, the number of false positives for black defendants and false negatives for white defendants was reduced. However, the number of false positives for white defendants and false negatives for black defendants increased, which was not widely acknowledged in the paper. The percentage of false positives for white defendants and false negatives for black defendants were both significantly higher than their corresponding black or white equivalent, suggesting that the model, in pursuit of fairness, has overcompensated. This suggests that assessing the ethnic classes separately might not be the optimal approach. Instead, ethnicity and all factors from which ethnicity can be derived could be removed from the model. Then, the results could be assessed separately for the ethnic groups to ensure a fair distribution. However, Rudin (2015) found that removing race did not materially impact accuracy, and removing anything that correlates with race would be troublesome because, as already discussed, predictive features such as age correlate with race (Rudin et al., 2018).

Skeem & Lowenkamp (2020) examined ways of debiasing algorithms and concluded that any debiasing method creates a trade-off between predictive accuracy, where the focus is crime prevention, and error rates, where the focus is individual fairness. When algorithms were debiased to remove any correlation with race, the error rates were higher, but disfavouring white defendants rather than black defendants, and so "providing algorithms with access to race (rather than omitting race or "blinding" its effects) can maximize calibration and minimize imbalanced error rates" (Skeem & Lowenkamp, 2020: 259). Furthermore, they found that COMPAS was not biased when assessed using fairness as a balance of accuracy rates for predictions; a measure of accuracy across classes, but it was racially biased when assessed using fairness as a balance in error rates for outcomes; a measure of the proportion of false positives across classes, in which black defendants were more likely to be classified as false

positive (predicted to reoffend but did not) than white defendants (Angwin et al., 2016; Biddle, 2022; Skeem & Lowenkamp, 2020). It seems that, concerning including race in the features for fairness to everyone, you are damned if you do and damned if you do not.

Squadrone et al. (2022) prototyped an "ethics by design" framework that essentially created rules for the impacts of different types of false positives and false negatives to address predicted (or known) biases. For example, they ruled that classifying a black defendant as not likely to reoffend was low risk, and classifying a black defendant as likely to reoffend was high risk. The results were reported as successfully reducing bias. However, the results show that accuracy decreased as "ethical compliance" increased. As false positives decreased (in both black and white defendants), false negatives increased. Once again, it is a trade-off.

Kleinberg et al. (2016: 4) defined three fairness measures: 1. Calibration within groups equates to the expected proportion of individuals in each group. 2. Balance for the negative class equates to individuals in the negative class having the same average predicted score between classes. 3. Balance for the positive class equates to individuals in the positive class having the same average predicted score between classes. Intuitively, they all seem appropriate; however, "except in highly constrained special cases, it is not possible to satisfy these three constraints simultaneously" (Kleinberg et al., 2016: 17). Yet again – trade-offs are needed.

3.6 Recidivism and Age

Age is a good predictor for recidivism (Bushway & Piehl, 2007; Kleiman et al., 2007; Stevenson & Slobogin, 2018), where "an early onset of offending typically predicts a long criminal career and the commission of many offences" Farrington & Davies (2007: 12). However, age is "really a proxy indicator of many time-related changes in life such as reduced physical and mental abilities, impulsivity, and improvement in maturity, judgment, education, and life skills" (Olver & Wong, 2015: 104). Therefore, age alone should not be used to assess recidivism likelihood because many of the features for which age is a proxy, such as "judgment, education, life skills, and even maturity together with antisocial tendencies and problematic lifestyles", can be addressed without the need to wait for ageing (Olver & Wong, 2015: 104).

Age is a primary feature in every model assessed in this literature review. This is justified because of its strong correlation to recidivism. However, age correlates with race and cannot be influenced by the defendant, so over-indexing on age could be unfair.

3.7 Recidivism and Gender

A Supreme Court ruling about a capital punishment case ruled "It would be patently unconstitutional for a state to argue that a defendant is liable to be a future danger because of his race" Buck v Davis (2016). However, "in Wisconsin v. Loomis (2016), the court allowed for COMPAS risk estimates to differ by gender [because] failure to consider gender would make risk scores less accurate and overestimate the risk that women pose" (Skeem & Lowenkamp, 2020: 274).

Excluding gender as a factor results in overestimating the likelihood of women to recidivate (Skeem et al., 2015), and removing race, age and gender results in a substantial loss of performance (Berk, 2012), so while gender is a protected characteristic (Equality and Human Rights Commission, 2021), including it improves the accuracy and fairness of the model. However, is it fair to once again use a protected characteristic that cannot be influenced by the defendant to predict their future behaviour? Based upon the evidence, not doing so seems unfair (Skeem & Lowenkamp, 2020), however doing so also seems unfair.

Returning to the earlier discussion about performance metrics, overall accuracy equality is where the accuracy is the same between protected classes. However, this also does not differentiate between false positives and false negatives. Conditional use accuracy equality, or predictive parity (Chouldechova, 2017: 3-4) is where the precision and specificity of predictions are the same between different protected classes. This better balances the differing implications of false positives and false negatives between the protected classes.

> "When base rates for protected group classes differ, one cannot have simultaneously conditional use accuracy equality and, across protected group classes, equal false positive and false rates." (Berk et al., 2021: 20).

This is a necessary implication of this method of measuring fairness. It is known that females are less likely to recidivate than males (Skeem et al., 2015), so it is better to ensure the prediction profile matches reality than to ensure there is parity between the classes. There is, however, a risk that the actual results to which the predictions are being compared are

already unjustly skewed, and by fitting the results to the same profile those pre-existing biases are perpetuated, just like the earlier discussion about race profiles being skewed by historic racial injustices. This is a genuine concern, but not one that is addressed in this dissertation. In the words of Berk et al. (2021: 35) “One cannot expect any risk assessment tool to reverse centuries of racial injustice or gender inequality. That bar is far too high. But, one can hope to do better”.

3.8 The Importance of Explainability

Black box solutions are justified based on outcome-reasoning, which is where the performance of the model is assessed based on its output. This is much easier to discuss with non-technical stakeholders, however, it omits the detailed understanding of relationships between the parameters and the model outputs that are present with model-reasoning (Rodu & Baiocchi, 2023).

It is better to use an explainable ML model such as DT than to try to explain a black box model such as SVM or ANN, especially when the results are similar and there is marginal incremental performance from the black box model (Rudin, 2019). The explainable output is, by its nature transparent, and transparency is important for ethical machine learning models (Walmsley, 2021). Three forms of opacity, (the opposite of transparency), are “(1) opacity as intentional corporate or state secrecy, (2) opacity as technical illiteracy, and (3) an opacity that arises from the characteristics of machine learning algorithms and the scale required to apply them usefully” Burrell (2016: 1). So anything that increases opacity, intentionally or otherwise, reduces transparency and can therefore be considered to be less ethical, even if the results are balanced and accurate.

Curtis (2018) compared RF and XGBoost, both of which are black box and so difficult to explain, while Wang et al. (2023) found that interpretable machine learning models performed as well as black box models, and better than COMPAS. This is another trade-off, but one that is easier to reconcile. Firstly, when comparing machine learning models for recidivism, the candidate models should always include interpretable options such as DT and LR, and the performance measures should be considered alongside explainability. A slightly worse-performing explainable model is likely to be preferred to a slightly better-performing black box model, notwithstanding the challenges with reconciling the different performance measures already discussed.

Babad & Chun (2023) take explainability for fairness to an extreme level by reducing the number of parameters to a minimal number to make recidivism prediction models explainable. This appears to comply with the spirit of explainability and fairness, however, the final features are all static and therefore give the individuals no ability to influence their prediction. This appears to be unfair, so simply being transparent is not enough. In cases with societal impacts, a fair model must be simple enough to be explainable, but sophisticated enough to consider dynamic as well as static factors.

Butler et al. (2022) posit that even logistical regression models can be considered black box because there will still be unknown factors due to missing data, so the true nature of the input-output relationship will be unknown. This is missing the point. Even if somehow every possible data point were known the output is still only a prediction and therefore not 100% accurate. It is not the lack of accuracy but rather the lack of transparency that makes a machine learning model a black box. If input variables can be used to explain outputs, which is the case with logistical regression and DT, for example, then they are not black boxes.

Finally, Shaikh et al. (2017) proposed a system to interpret policies of fairness and bias from policy documents using natural language processing (NLP), to monitor the development of the ML systems and then test them to ensure compliance with those policies. This is a potential area for further study.

3.9 Ontology to Increase Fairness Through Transparency

This section examines how ontologies have been used to address ethical challenges with machine learning. It is relatively short because there is not much literature on the subject, hence this dissertation is new and innovative.

Kasirzadeh & Smart (2021) describe how counterfactuals can be used in ontology to test the fairness of machine learning algorithms by imagining a world where characteristics such as race are different. Examples given in the paper are trivial, where causal relationships are obvious. In the case of recidivism, the relationships between the parameters are complex, and there would be no way of saying with reasonable certainty if an individual either would or would not recidivate if their race were different. However, it would be relatively easy to check if the model would predict a different outcome if their race were different but everything else remained the same. This could be worth exploring further.

Franklin et al. (2023) created an ontology of fairness metrics, many of which have been discussed throughout this literature review. This helps to highlight and reason about different metrics and how they contribute, or not, to fairness. However, it does not address the fairness itself.

Lehnert (2021) discussed the need to be careful about the terms and relationships used in ontologies in order the reduce the risk of introducing bias. Again, this does not address the issue of bias already existing in the baseline dataset.

3.10 Summary

The factors that can cause recidivism predictions to be biased are complex and interdependent, including the very data from which the models are built and the metrics chosen to measure their performance. It appears to be impossible to balance everything and arrive at a completely fair and unbiased solution, so perhaps the best that can be done is to make the solution as fair as possible, which means including dynamic features alongside static features, and using explainable models so the reasons for predictions are clear and can potentially be challenged. The choice of performance metrics should also be justified, highlighting the risks; once again transparency is the key to fairness. Finally, how the models are used is important, and with scope for potential bias remaining, there should always be a human decision-maker at the end of the process to justify and explain the outcome. With decisions as important as this, the model should *guide* the answer, not *provide* the answer.

> "An actuarial device may be able to tell you quite accurately that two-thirds of all cases in a particular risk category will fail, but it cannot tell which ones will fail. When a particular inmate comes up for parole, the decision-maker still will not know whether he will succeed or fail on parole" (Hoffman & Beck 1994: 203).

4 Methodology

Ontology is proposed to address the risks of bias in predicting recidivism in England and Wales by making correlations with race transparent, so features that present a higher risk of racial bias can be avoided, or at least used with caution with robust justification, and by allowing prediction results to be compared to training data to highlight racial biases.

Metadata from Crown Courts, Magistrate Courts, Prison Service and Probation Service in England and Wales was obtained from Ministry of Justice (2020), henceforth referred to as MoJ metadata. An ontology could be built from the metadata, but without data to populate the ontology the concept of identifying and monitoring the correlations with race before selecting features could not be fully tested. A recidivism dataset containing 25,835 anonymised records from Georgia, USA was obtained from National Institute of Justice (N.D.), henceforth referred to as the NIJ dataset, however, the features were different and the data model was significantly simpler because it had already been pre-processed and consolidated from two sources; the Georgia Department of Community Supervision and the Georgia Bureau of Investigation (Office of Justice Programs, N.D.). A multi-step approach was therefore adopted to prove the concept of maintaining the correlations within a complex ontology using the MoJ metadata after first proving the concept using the simpler NIJ dataset.

The methodology was divided into four steps:

1. Exploratory Data Analysis (EDA) using Python on the NIJ dataset to identify correlations on which to form hypotheses for detailed statistical analysis.
2. Detailed statistical analysis on the NIJ dataset using R to test the hypotheses.
3. Build an ontology in Protégé using the NIJ dataset as the domain scope and the results of the statistical analysis to highlight features with a higher risk of racial bias to prove and test the concept.
4. Build an ontology in Protégé using the MoJ metadata as the initial domain scope to further prove the concept using the MoJ taxonomy, with additional features informed by the literature review and the NIJ ontology, to extend the scope and illustrate how it could work in England and Wales. This did not include real data, however, some dummy records were created for illustration and testing.

5 Implementation

NIJ analysis files are at https://github.com/feaviolp/msc-project/tree/main/NIJ%20Analysis

NIJ ontology files are at https://github.com/feaviolp/msc-project/tree/main/NIJ%20Ontology

MoJ ontology files are at https://github.com/feaviolp/msc-project/tree/main/MoJ%20Ontology

5.1 NIJ Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed on the NIJ dataset using Python in a Jupyter Notebook (Appendix A).

First, the dataset was examined and pre-processed to deal with null values and some features were transformed into integers.

Next, the data were checked for correlations to identify features on which to form hypotheses to test with statistical analysis. It was clear, however, that while some features correlated more than others with recidivism, gender and race, there were no features with strong correlation so the statistical analysis should examine all features. This was still a useful outcome of the EDA to set the scope of the statistical analysis.

5.2 NIJ Statistical Analysis

Statistical analysis was performed on the NIJ dataset using R in a Jupyter Notebook (Appendix B).

Once again the data were pre-processed to deal with null values, and this time all ordinal and categorical features were converted to integers to enable significance testing.

The following hypotheses formed from the literature review were tested:

- There is a difference in recidivism by race, $\alpha=0.01$
- There is a difference in recidivism by gender, $\alpha=0.01$
- There is a difference in recidivism by age, $\alpha=0.01$
- There is a difference in offender age by race, $\alpha=0.01$

Chi-square was used to test if there was a difference in recidivism by race because both features are categorical. The results (Figure 10) show that, with p 0.00017, the null hypothesis

is rejected. There is a difference in recidivism between black and white offenders, with black offenders being more likely to reoffend (Figure 11).

```
       BLACK WHITE
  false  6134  4797
  true   8713  6191
       Pearson's Chi-squared test with Yates' continuity correction

data:  table(NIJ_orig$Recidivism_Within_3years, NIJ_orig$Race)
X-squared = 14.094, df = 1, p-value = 0.0001739
```

Figure 10. Chi-square test: difference in recidivism by race

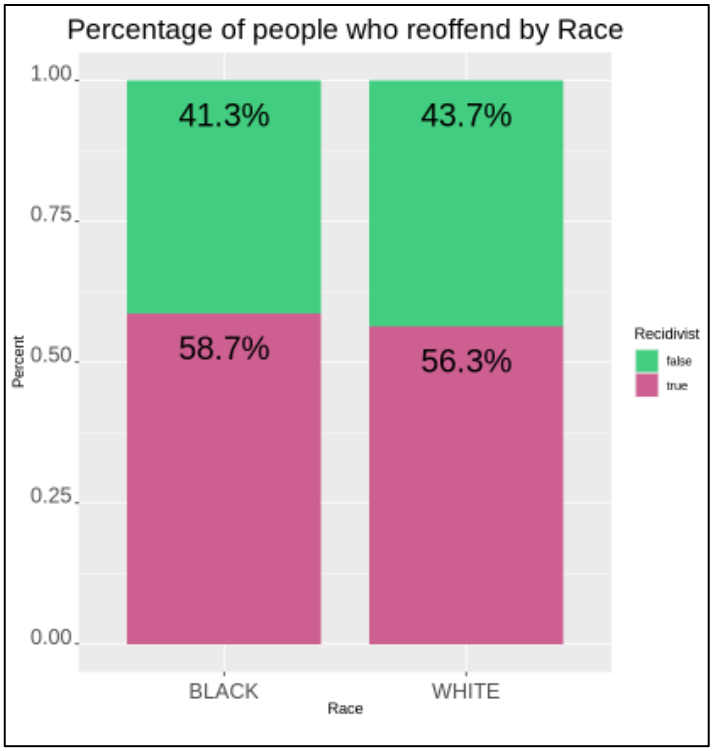


Figure 11. Percentage of people who reoffend by race

Recidivism and gender are both categorical so Chi-square was again used to test if there was a difference in recidivism by gender. The results (Figure 12) show, with $p < 2.2e^{-16}$, the null hypothesis is rejected. There is a difference in recidivism between male and female offenders, with male offenders being more likely to reoffend (Figure 13).

```
        Male Female
  false  9206   1725
  true  13462   1442
        Pearson's Chi-squared test with Yates' continuity correction

data:  table(NIJ_orig$Recidivism_Within_3years, NIJ_orig$Gender)
X-squared = 217.99, df = 1, p-value < 2.2e-16
```

Figure 12. Chi-square test: difference in recidivism by gender

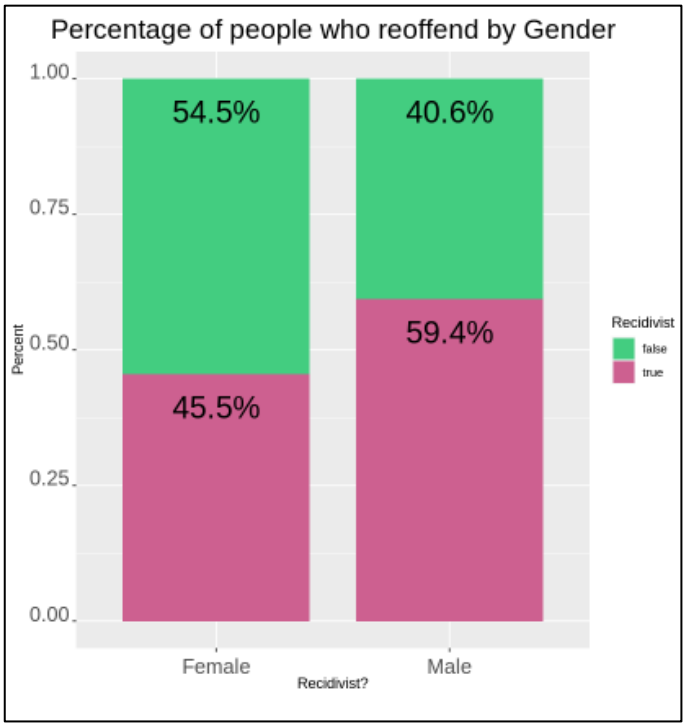


Figure 13. Percentage of people who reoffend by gender

The distribution of age was checked with a histogram and Q-Q plot and found to be not normal, which would typically mean using the non-parametric Mann-Whitney U test. However with a large sample size the central limit theorem allows the use of a t-test, so both were used to test if there was a difference in recidivism by age. The results show that with $p < 2.2e^{-16}$, the null hypothesis is rejected. There is a difference in age between reoffenders and non-reoffenders, with reoffenders being younger; mean age of 31.2 than non-reoffenders; mean age of 34.5 (Figure 14). This is further illustrated in the boxplot and bar chart in Figure 15.

```
        Wilcoxon rank sum test with continuity correction

data:  Age_at_Release by Recidivism_Within_3years
W = 98018098, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Welch Two Sample t-test

data:  Age_at_Release by as_factor(Recidivism_Within_3years)
t = 28.822, df = 22988, p-value < 2.2e-16
alternative hypothesis: true difference in means between group false and group true is not equal to 0
95 percent confidence interval:
 3.145913 3.605023
sample estimates:
mean in group false  mean in group true
          34.53737            31.16190
```

Figure 14. Mann-Whitney U test and t-test: difference in recidivism by age

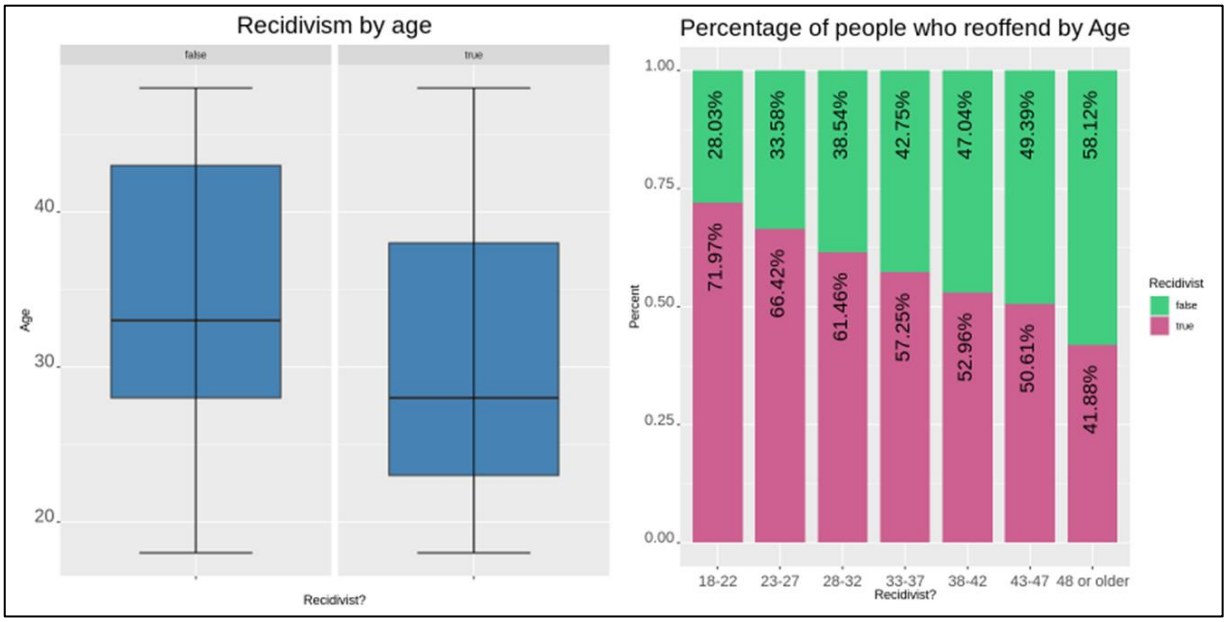


Figure 15. Recidivism by age

Mann-Whitney U test and t-test were used again to test if there was a difference in age by race, with the results showing that with $p < 2.2e^{-16}$, the null hypothesis is rejected. There is a difference in age between black offenders and white offenders, with black offenders younger; mean age of 31.6 than white offenders; mean age of 33.9 (Figure 16). This is further illustrated in the boxplot in Figure 17.

```
        Wilcoxon rank sum test with continuity correction

data:  Age_at_Release by Race
W = 69826119, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Welch Two Sample t-test

data:  Age_at_Release by as_factor(Race)
t = -19.514, df = 24010, p-value < 2.2e-16
alternative hypothesis: true difference in means between group BLACK and group WHITE is not equal to 0
95 percent confidence interval:
 -2.509302 -2.051234
sample estimates:
mean in group BLACK mean in group WHITE
          31.62026            33.90053
```

Figure 16. Mann-Whitney U test and t-test: difference in age by race

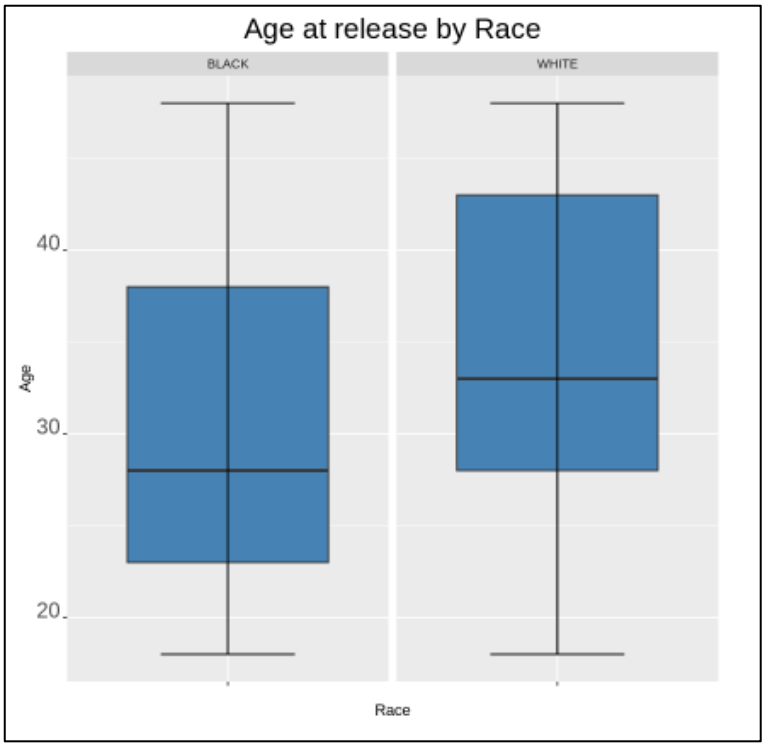


Figure 17. Age at release by race

The first three null hypotheses were all rejected, with statistically higher reoffending in black over white, male over female, and younger over older, all of which is consistent with the literature review. The fourth null hypothesis was also rejected, with black offenders being statistically younger than white offenders. This highlights one of the challenges identified in the literature review; age is a reliable predictor of recidivism, but age also correlates with race, so focusing too heavily on age risks racially biased results.

With the rejection of all four null hypotheses and the best practice from the literature review of modelling male and female recidivism separately, it was decided that the ontology will differentiate between genders when modelling correlations with recidivism and race.

Every feature was tested for correlation with recidivism and race for male and female offenders separately using Spearman's Rho, because the data are a mix of nominal and ordinal, with α=0.01 due to the large sample size (Kim & Choi, 2021). Age had a correlation for male offenders of r_s -0.177 with recidivism with p $2.2e^{-16}$ and r_s 0.121 with race with p $2.2e^{-16}$. Those are not typically strong correlations, however, considering the previously discussed correlations between age and race with recidivism, it was decided to use r_s 0.1 (weak correlation; Figure 18; Xiao et al., 2016) as the cut-off for feature risk and importance.

Value of *r*	Strength of relationship
−1.0 to −0.5 or 1.0 to 0.5	Strong
−0.5 to −0.3 or 0.3 to 0.5	Moderate
−0.3 to −0.1 or 0.1 to 0.3	Weak
−0.1 to 0.1	None or very weak

Figure 18. Strength of relationship (Xiao et al., 2016)

Figure 19 shows all r_s values, highlighted in green where $r_s \geq 0.1$ for recidivism, to represent features likely to be important to a recidivism prediction model, and highlighted in red where $r_s \geq 0.1$ for race, to highlight features that carry a higher risk of introducing racial bias. All negative values were converted to positive because the direction of correlation was unimportant for this analysis. Most values were statistically significant at α=0.01, but some were not, which are highlighted in amber and set to $r_s = 0$.

The correlations between features in Figure 19 are metadata about the NIJ dataset. They were used in the design and build of the NIJ and MoJ ontologies to create ontologies with knowledge and metaknowledge stored alongside each other, in a novel use of ontology to address the increasing need for transparency in feature engineering.

Feature	Correlation with:				
	Everyone	Male		Female	
	Recidivism	Recidivism	Race	Recidivism	Race
Age_at_Release	0.176	0.177	0.121	0.133	0.072
Residence_PUMA	0.025	0.026	0.139	0	0.187
Gang_Affiliated	0.185	0.185	0.086	N/A	N/A
Supervision_Risk_Score_First	0.180	0.185	0.053	0.146	0.046
Supervision_Level_First	0.061	0.053	0	0.069	0
Education_Level	0.088	0.088	0.057	0	0
Dependents	0.031	0.031	0.096	0	0.064
Prison_Offense	0.018	0.024	0.033	0	0.260
Prison_Years	0.130	0.134	0.066	0.186	0.109
Prior_Arrest_Episodes_Felony	0.199	0.187	0.025	0.262	0
Prior_Arrest_Episodes_Misd	0.178	0.161	0.094	0.279	0
Prior_Arrest_Episodes_Violent	0.065	0.055	0.111	0	0.213
Prior_Arrest_Episodes_Property	0.182	0.181	0.103	0.233	0
Prior_Arrest_Episodes_Drug	0.081	0.071	0	0.107	0.279
Prior_Arrest_Episodes_PPViolationCharges	0.229	0.218	0.063	0.303	0.067
Prior_Arrest_Episodes_DVCharges	0.066	0.062	0.052	0.052	0
Prior_Arrest_Episodes_GunCharges	0.044	0.036	0.104	0	0
Prior_Conviction_Episodes_Felony	0.105	0.094	0.032	0.169	0.047
Prior_Conviction_Episodes_Misd	0.175	0.160	0.070	0.247	0
Prior_Conviction_Episodes_Viol	0.047	0.043	0.088	0	0.161
Prior_Conviction_Episodes_Prop	0.161	0.157	0.104	0.232	0.073
Prior_Conviction_Episodes_Drug	0.065	0.059	0	0.077	0.235
Prior_Conviction_Episodes_PPViolationCharges	0.096	0.088	0.050	0.137	0
Prior_Conviction_Episodes_DomesticViolenceCharges	0.059	0.057	0.017	0	0
Prior_Conviction_Episodes_GunCharges	0.031	0.024	0.058	0	0
Prior_Revocations_Parole	0.058	0.051	0.037	0.060	0
Prior_Revocations_Probation	0.039	0.036	0.065	0.076	0.059
Condition_MH_SA	0.114	0.121	0.131	0.149	0.259
Condition_Cog_Ed	0.038	0.050	0.039	0	0
Condition_Other	0	0	0	0	0.065
Violations_ElectronicMonitoring	0.004	0	0.069	0	0.075
Violations_Instruction	0.064	0.058	0.046	0.087	0
Violations_FailToReport	0.030	0.024	0	0.069	0
Violations_MoveWithoutPermission	0.032	0.029	0	0.057	0
Delinquency_Reports	0.041	0.028	0	0.102	0.068
Program_Attendances	0.060	0.065	0.072	0	0.190
Program_UnexcusedAbsences	0.060	0.050	0.043	0.108	0
Residence_Changes	0.054	0.052	0.047	0.079	0
Avg_Days_per_DrugTest	0.011	0	0.078	0	0.135
DrugTests_THC_Positive	0.082	0.078	0.161	0	0.089
DrugTests_Cocaine_Positive	0.011	0	0.128	0	0.121
DrugTests_Meth_Positive	0.055	0.055	0.279	0.091	0.227
DrugTests_Other_Positive	0.004	0	0.121	0.053	0.126
Percent_Days_Employed	0.217	0.217	0.126	0.227	0.059
Jobs_Per_Year	0.074	0.074	0.120	0.088	0.060
Employment_Exempt	0.050	0.048	0.021	0	0

Legend	
	Recidivism correlation ≥ 0.1
	Race correlation ≥ 0.1
	No statitically significant correlation at α=0.01

Figure 19. Feature correlations with recidivism and race by gender

5.3 NIJ Ontology

The dataset and associated metadata from National Institute of Justice (N.D.) were used to derive knowledge to build an ontology. The Noy & McGuiness (2001) seven-step approach was superficially used with emphasis on the class hierarchy and properties of classes. It was superficial because the dataset was already presented in a simple flat structure with a single row per offender, so there was limited opportunity to address the overall scope. DeBellis (2021) was also used as a technical reference.

Initially, the class hierarchy was designed with every feature having its own class because a key requirement of the ontology was to highlight if there was a correlation between race and any other feature, so the features themselves were important, and it is preferred to have a separate class where the attribute is important to the domain (Noy & McGuiness, 2001). However, it became apparent that this approach was impracticable because the classes needed to contain the data properties about each offender as an instance, and also be members of inferred correlation classes to highlight if a feature was high or low risk or importance, which introduced circular references. The design was therefore changed so each feature was both an instance against which the race and recidivism correlations metaknowledge can be recorded via data properties (Figure 20) and an object property in the class hierarchy (Figure 21).

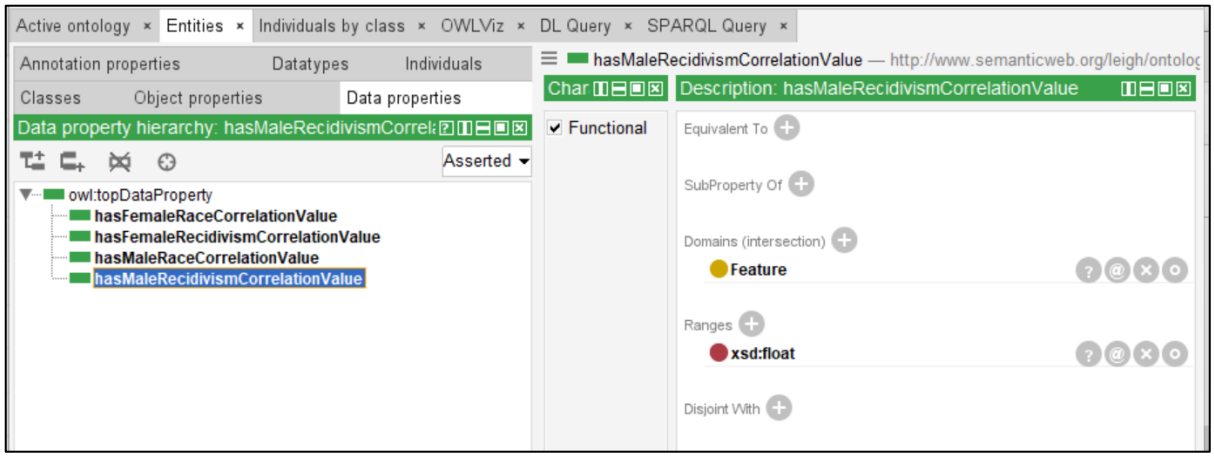


Figure 20. Data property hierarchy

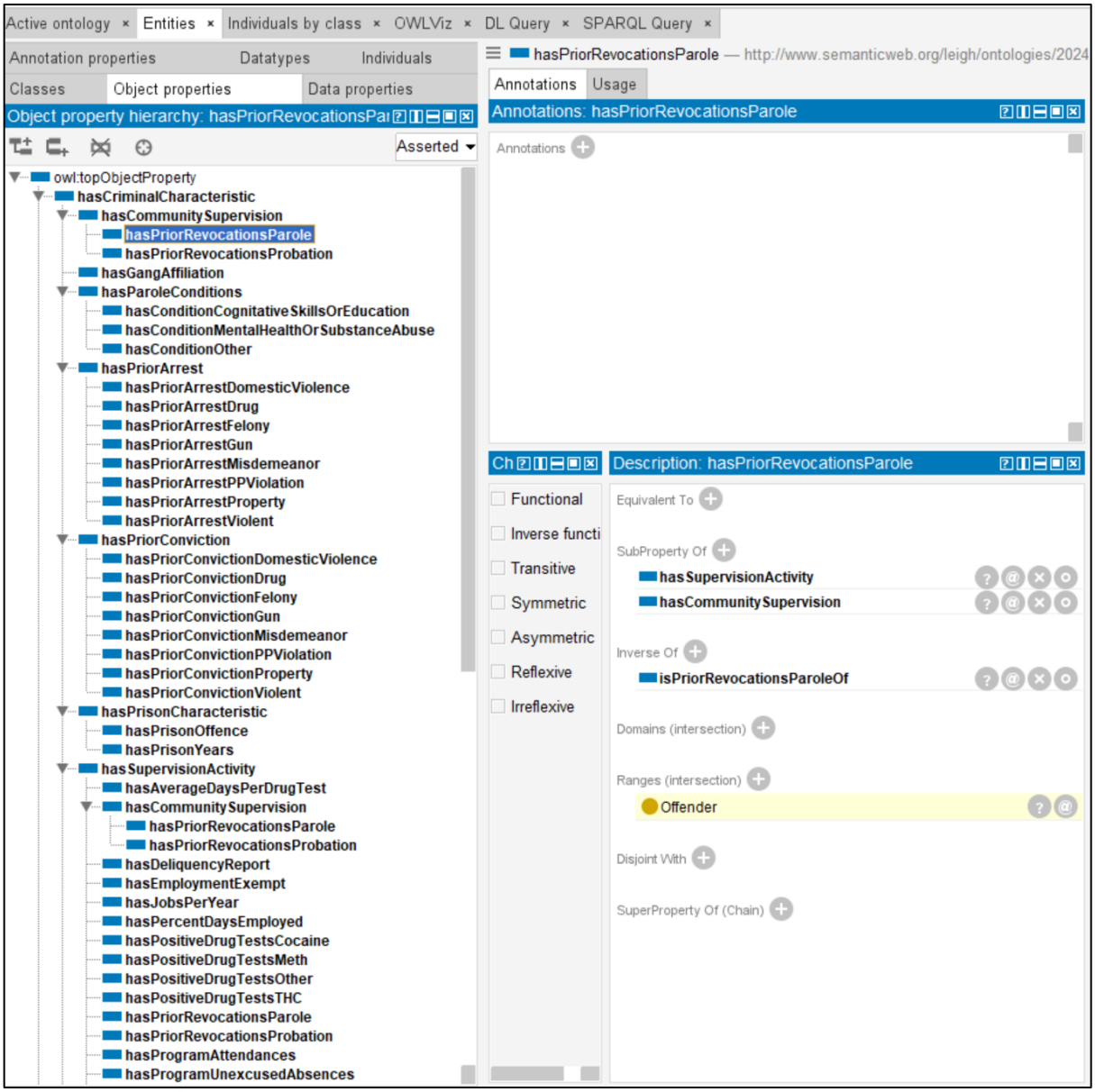


Figure 21. Object property hierarchy (truncated)

Defined classes were added to infer which features were high risk (race correlation ≥ 0.1 and < 1) and high importance (recidivism correlation ≥ 0.1) for male and female classes. Defined classes for low risk and low importance (with correlations < 0.1) and prohibited (with race correlation = 1, which is the race feature itself) were also added.

UML notation for ontologies (Bārzdiņš et al., 2010), adapted to include object properties, shows the final ontology design (Figure 22), while Figure 23 shows the inferred hierarchy from Protégé. The structure is relatively flat due to the simplicity of the dataset, with everything captured in one row per offender.

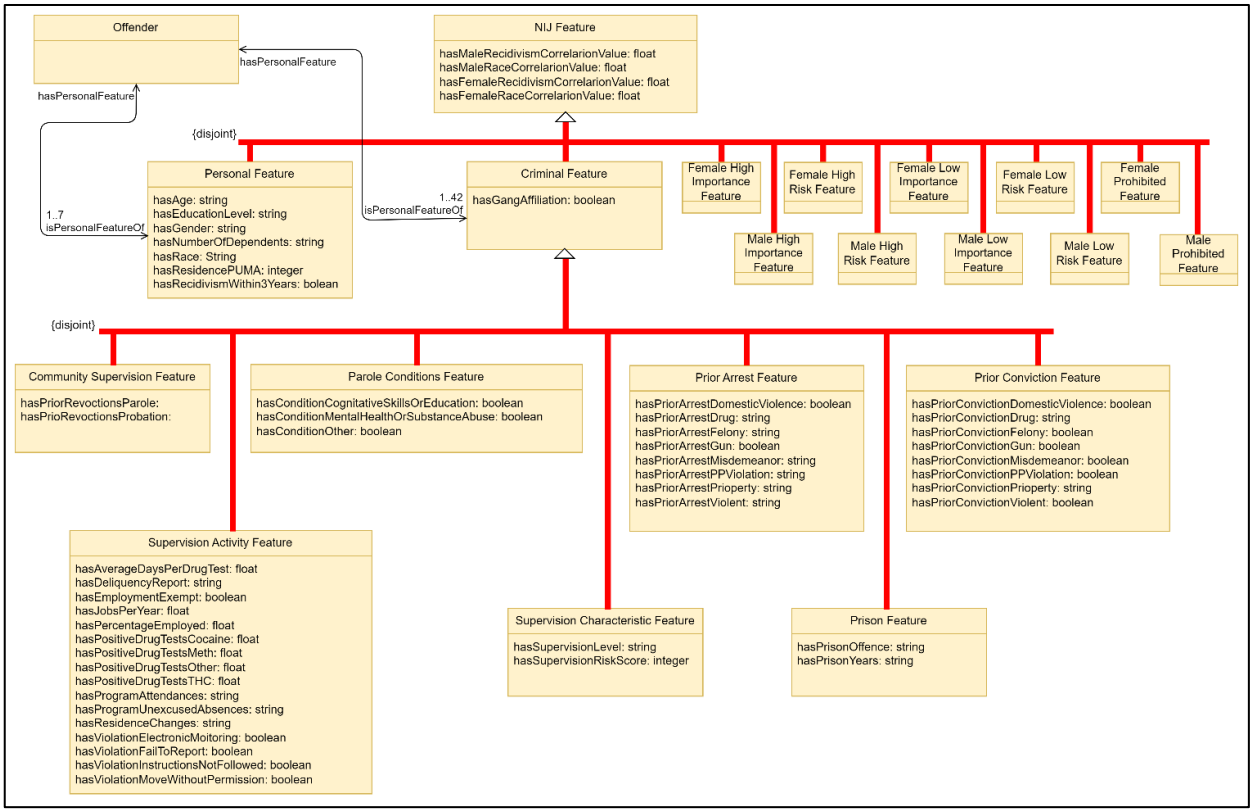


Figure 22. NIJ ontology design

Figure 23. Inferred hierarchy

The NIJ dataset was imported as offender instances using a Cellfie import script (Appendix C). Importing the instances in this way without first defining ranges of and dependencies between features is using the ontology more like a database, however, it was the most efficient way to prove the concept given the available data. The recidivism and race correlation values from Figure 19 were entered manually.

The ontology successfully stored all of the data from the NIJ dataset alongside the feature correlation metadata, assigning every feature to high/low risk/importance (Appendix D). Furthermore, simple DL Queries demonstrated, for example, how to find features that were both high importance and high risk for males (Figure 24) which are the features to be most cautious about.

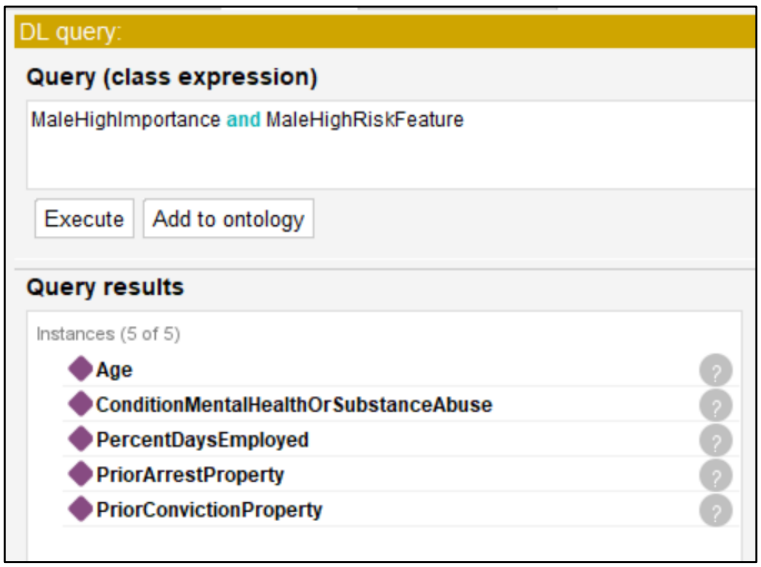


Figure 24. DL Query for male high-risk and high-importance

SPARQL was used for additional filtering and showing the actual correlation values, for example, features where male correlation with race is higher than male correlation with recidivism (Figure 25).

Using that data, a data scientist would select features that correlate more highly with recidivism than race to build prediction models that are less racially biased.

Snap SPARQL Query:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX nij: <http://www.semanticweb.org/leigh/ontologies/2024/4/NIJ#>

#Show features where maleRaceCorrelationValue is greater than maleRecidivismCorrelationValue
SELECT ?feature ?maleRaceCorrelationValue ?maleRecidivismCorrelationValue
WHERE {
  ?feature  nij:hasMaleRaceCorrelationValue ?maleRaceCorrelationValue ;
            nij:hasMaleRecidivismCorrelationValue ?maleRecidivismCorrelationValue .

  FILTER(?maleRaceCorrelationValue > ?maleRecidivismCorrelationValue)
}
ORDER BY DESC(?maleRaceCorrelationValue)
```

Execute

?feature	?maleRaceCorrelationValue	?maleRecidivismCorrelationValue
nij:PositiveDrugTestsMeth	0.279	0.055
nij:PostiveDrugTestsTHC	0.161	0.078
nij:ResidencePUMA	0.139	0.026
nij:ConditionMentalHealthOrSubstanceAbuse	0.131	0.121
nij:PositiveDrugTestsCocaine	0.128	0.0
nij:PositiveDrugTestsOther	0.121	0.0
nij:JobsPerYear	0.12	0.074
nij:PriorArrestViolent	0.111	0.055
nij:PriorArrestGun	0.104	0.036
nij:NumberOfDependents	0.096	0.031
nij:PriorConvictionViolent	0.088	0.043
nij:AverageDaysPerDrugTest	0.078	0.0
nij:ProgramAttendances	0.072	0.065
nij:ViolationElectronicMonitoring	0.069	0.0
nij:PriorRevocationsProbation	0.065	0.036
nij:PriorConvictionGun	0.058	0.024
nij:PrisonOffence	0.033	0.024

17 results

Figure 25. Features with male race correlation > male recidivism correlation

Having trained the model, the results would be loaded back into the ontology to check that the profiles are consistent. SPARQL would again be used to find the total number of people and the number who reoffended, by gender and race (Figure 26), which would then be re-run to ensure that the ratios remain consistent. The first order logic to build the query is:

Let Offender(x) denote that x is an offender.

Let Gender(x,g) denote that offender x has gender g.

Let Race(x,r) denote that offender x has race r.

Let Reoffended(x) denote that offender x has reoffended within three years.

Total(g,r,n)≡n=#{x|Offender(x)∧Gender(x,g)∧Race(x,r)}

ReoffendedCount(g,r,m)≡m=#{x|Offender(x)∧Gender(x,g)∧Race(x,r)∧Reoffended(x)}

$\forall g,r\ \exists n,m$ (Total(g,r,n)∧ReoffendedCount(g,r,m))

Where:

- n is the total count of offenders with gender g and race r.
- m is the count of reoffended offenders with gender g and race r.

Snap SPARQL Query:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX nij: <http://www.semanticweb.org/leigh/ontologies/2024/4/NIJ#>

#Total people and number who reoffended number gender and race
SELECT ?gender ?race (COUNT(?person) AS ?total) (COUNT(?reoffender) AS
?reoffended)
WHERE {
 ?person a nij:Offender ;
          nij:hasGender ?gender ;
          nij:hasRace ?race .

 OPTIONAL {
  ?person nij:hasRecidivismWithin3Years nij:true .
  BIND(IF(BOUND(?person), ?person, "") AS ?reoffender)
 }
}
GROUP BY ?gender ?race
ORDER BY ?gender ?race
```

Execute

?gender	?race	?total	?reoffended
nij:F	nij:BLACK	1082	474
nij:F	nij:WHITE	2085	968
nij:M	nij:BLACK	13765	8239
nij:M	nij:WHITE	8903	5223

Figure 26. Count of total people and people who reoffended by gender and race

The ontology now has everything needed to inform a data scientist about the features that are more likely to create an accurate model to predict recidivism and the features more likely

to introduce racial bias, and the data themselves, so they can create a model with transparency and report on the risk levels of the features selected.

Taking a worked example, a data scientist building a model to predict male recidivism might initially look for features with high importance and low risk (Figure 27). They know that age is a highly important feature so even though it is high-risk they decide to add that too. That leaves them with 8 features on which to train their model. Before proceeding further they would check the recidivism and race correlations of the 8 selected features (Figure 28). As expected, all have a correlation value ≥ 0.1 for recidivism, and only age has a correlation ≥ 0.1 with race, which is still lower than recidivism.

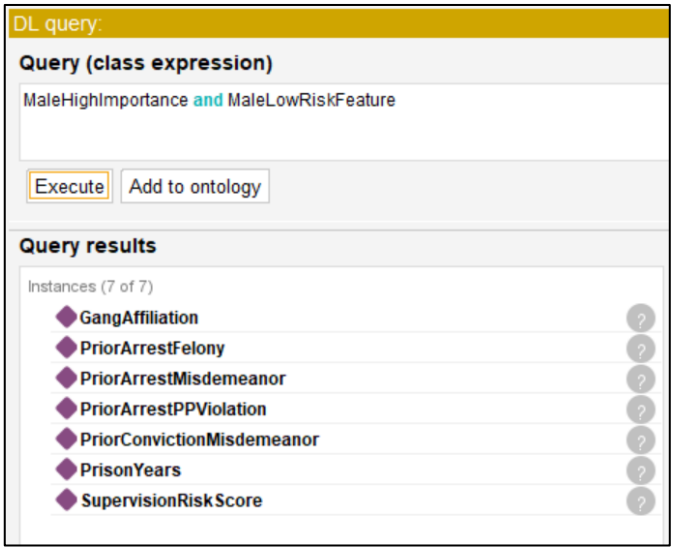


Figure 27. Features with high importance and low risk for males

Snap SPARQL Query:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX nij: <http://www.semanticweb.org/leigh/ontologies/2024/4/NIJ#>

#Show MaleImportantFeatures where maleRaceCorrelationValue is greater than maleRecidivismCorrelationValue
SELECT ?feature ?maleRecidivismCorrelationValue ?maleRaceCorrelationValue
WHERE {
  ?feature a nij:Feature ;
           nij:hasMaleRaceCorrelationValue ?maleRaceCorrelationValue ;
           nij:hasMaleRecidivismCorrelationValue ?maleRecidivismCorrelationValue .

  FILTER((?feature = nij:Age) || (?feature = nij:GangAffiliation) || (?feature = nij:PriorArrestFelony) || (?feature = nij:PriorArrestMisdemeanor) ||
(?feature = nij:PriorArrestPPViolation) || (?feature = nij:PriorConvictionMisdemeanor) || (?feature = nij:PrisonYears) || (?feature =
nij:SupervisionRiskScore))
}
ORDER BY DESC(?maleRecidivismCorrelationValue)
```

Execute

?feature	?maleRecidivismCorrelationValue	?maleRaceCorrelationValue
nij:PriorArrestPPViolation	0.218	0.063
nij:PriorArrestFelony	0.187	0.025
nij:SupervisionRiskScore	0.185	0.053
nij:GangAffiliation	0.185	0.086
nij:Age	0.177	0.121
nij:PriorArrestMisdemeanor	0.161	0.094
nij:PriorConvictionMisdemeanor	0.16	0.07
nij:PrisonYears	0.134	0.066

8 results

Figure 28. Recidivism and race correlation for selected features

They consider the overall risk level acceptable so they export the male data for only those eight features from the ontology (Figure 29). Once the model has been trained and the data scientist is happy with the results, they will be imported back into the ontology to be checked for deviations from the existing racial ratios. The concept of creating an ontology that provides transparency about the risks of using specific features and the ability to check the prediction results against actual profiles has been proven using the NIJ dataset.

Snap SPARQL Query:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX nij: <http://www.semanticweb.org/leigh/ontologies/2024/4/NIJ#>

SELECT ?offender
(REPLACE(STR(?ageValue), STR(nij:), "") AS ?age)
(REPLACE(STR(?gangAffiliationValue), STR(nij:), "") AS ?gangAffiliation)
(REPLACE(STR(?priorArrestFelonyValue), STR(nij:), "") AS ?priorArrestFelony)
(REPLACE(STR(?priorArrestMisdemeanorValue), STR(nij:), "") AS ?priorArrestMisdemeanor)
(REPLACE(STR(?priorArrestPPViolationValue), STR(nij:), "") AS ?priorArrestPPViolation)
(REPLACE(STR(?priorConvictionMisdemeanorValue), STR(nij:), "") AS ?priorConvictionMisdemeanor)
(REPLACE(STR(?prisonYearsValue), STR(nij:), "") AS ?prisonYears)
(REPLACE(STR(?supervisionRiskScoreValue), STR(nij:), "") AS ?supervisionRiskScore)
WHERE {
  ?offender a nij:Offender .
  ?offender nij:hasGender nij:M .
  ?offender nij:hasAge ?ageValue .
  OPTIONAL {?offender nij:hasGangAffiliation ?gangAffiliationValue . }
  OPTIONAL {?offender nij:hasPriorArrestFelony ?priorArrestFelonyValue . }
  OPTIONAL {?offender nij:hasPriorArrestMisdemeanor ?priorArrestMisdemeanorValue . }
  OPTIONAL {?offender nij:hasPriorArrestPPViolation ?priorArrestPPViolationValue . }
  OPTIONAL {?offender nij:hasPriorConvictionMisdemeanor ?priorConvictionMisdemeanorValue . }
  OPTIONAL {?offender nij:hasPrisonYears ?prisonYearsValue . }
  OPTIONAL {?offender nij:hasSupervisionRiskScore ?supervisionRiskScoreValue . }
}
ORDER BY ?offender
```

Execute

?offender	?age	?gangAffiliation	?priorArrestFel...	?priorArrestMis...	?priorArrestPP...	?priorConvictio...	?prisonYears	?supervisionRi...
nij:1	43-47	false	6	6OrMore	4	3	MoreThan3Years	3
nij:10	43-47	false	10OrMore	6OrMore	5OrMore	4OrMore	MoreThan3Years	5
nij:100	48OrOlder	false	10OrMore	6OrMore	5OrMore	4OrMore	MoreThan3Years	6
nij:1000	28-32	false	5	6OrMore	3	4OrMore	GreaterThan2To...	10
nij:10000	43-47	false	5	1	2	0	LessThan1Year	3
nij:10002	38-42	false	10OrMore	6OrMore	2	4OrMore	MoreThan3Years	8
nij:10003	28-32	false	7	3	1	0	1-2Years	10
nij:10004	28-32	false	8	4	5OrMore	2	1-2Years	8
nij:10005	28-32	false	5	6OrMore	5OrMore	2	LessThan1Year	7
nij:10006	48OrOlder	false	5	3	1	2	GreaterThan2To...	4
nij:10007	23-27	false	7	6OrMore	5OrMore	3	LessThan1Year	8
nij:10008	28-32	true	9	5	2	2	MoreThan3Years	9
nij:10009	23-27	true	3	0	1	1	GreaterThan2To...	5
nij:1001	38-42	false	9	6OrMore	5OrMore	4OrMore	1-2Years	8

22668 results

Figure 29. Data export of selected features

5.4 MoJ Ontology

Having proven the concept with the NIJ dataset, the ontology for recidivism prediction in England and Wales was designed and built using metadata from Data First, a programme combining data from Crown Courts, Magistrate Courts, Prison Service and Probation Service (Ministry of Justice, 2020) using the Noy & McGuiness (2001) seven-step approach. The amalgamation of four datasets, each representing a sub-domain of criminal justice, created a significantly more complex design than the NIJ ontology.

First, the metadata for each dataset was examined to determine the scope and constructs of the data in the absence of a domain expert to interview as would be normal practice (Kendal & Green, 2019). The features covered a wide scope of criminal justice which could all have been built into a single ontology. Indeed, having everything in a single "Criminal Justice" ontology would enable additional use cases in the future. However, to prove the concept of using an ontology to manage the knowledge and the metaknowledge to enable ethical transparency for the specific use case of recidivism prediction, only the features that were considered useful were selected, excluding, for example, features about proceedings such as pleas and mode of trial. Verdict features were also omitted on the assumption that for this use case, all of the data pertains to guilty verdicts, with proceedings resulting in not-guilty outcomes omitted. As such, the domain "Recidivism in England and Wales" is a sub-domain of "Criminal Justice in England and Wales".

There was a risk that, with no domain expert to interview, feature selection was not optimal. Since the exercise was to prove a concept and not to train machine learning models, the risk was acceptable. Domain expert input would be needed before taking the concept into production.

The features were grouped into classes according to their sub-domain and usage. Offender has one instance per offender with local data properties and data properties about the object properties of other classes such as gender and ethnicity. Each offender has 1-n cases. Each case has one offender (cases with multiple offenders would have separate case IDs in the ontology), one court, one principal offence (the offence which attracted the most severe sentence), one most serious offence (the offence that has the most severe potential sentence, which can be the same as the principal offence), and multiple sentence outcomes and probation requirements.

Properties were included for features that would be calculated externally from the ontology. For example, the number of previous offences would be calculated from the number of cases for each offender during data pre-processing outside the ontology before using the results in a machine learning model, so the results can then be stored in the ontology as another feature with its correlation values.

Additional properties were then added to capture information not in the Data First datasets that could potentially improve recidivism prediction according to the literature review and the statistical analysis of the NIJ dataset. The additional properties are denoted with a blue background in the ontology design, including a new class called “Prison Conduct” with features held in the prison system but not included in the Data First datasets (HM Prison and Probation Service, 2024; Ministry of Justice, 2023; Ministry of Justice, 2016; Prison Reform Trust, 2022). Once again UML notation for ontologies (Bārzdiņš et al., 2010) was adapted to include object properties (Figure 30). The inferred hierarchy, object property hierarchy and data property hierarchy are in Appendix E.

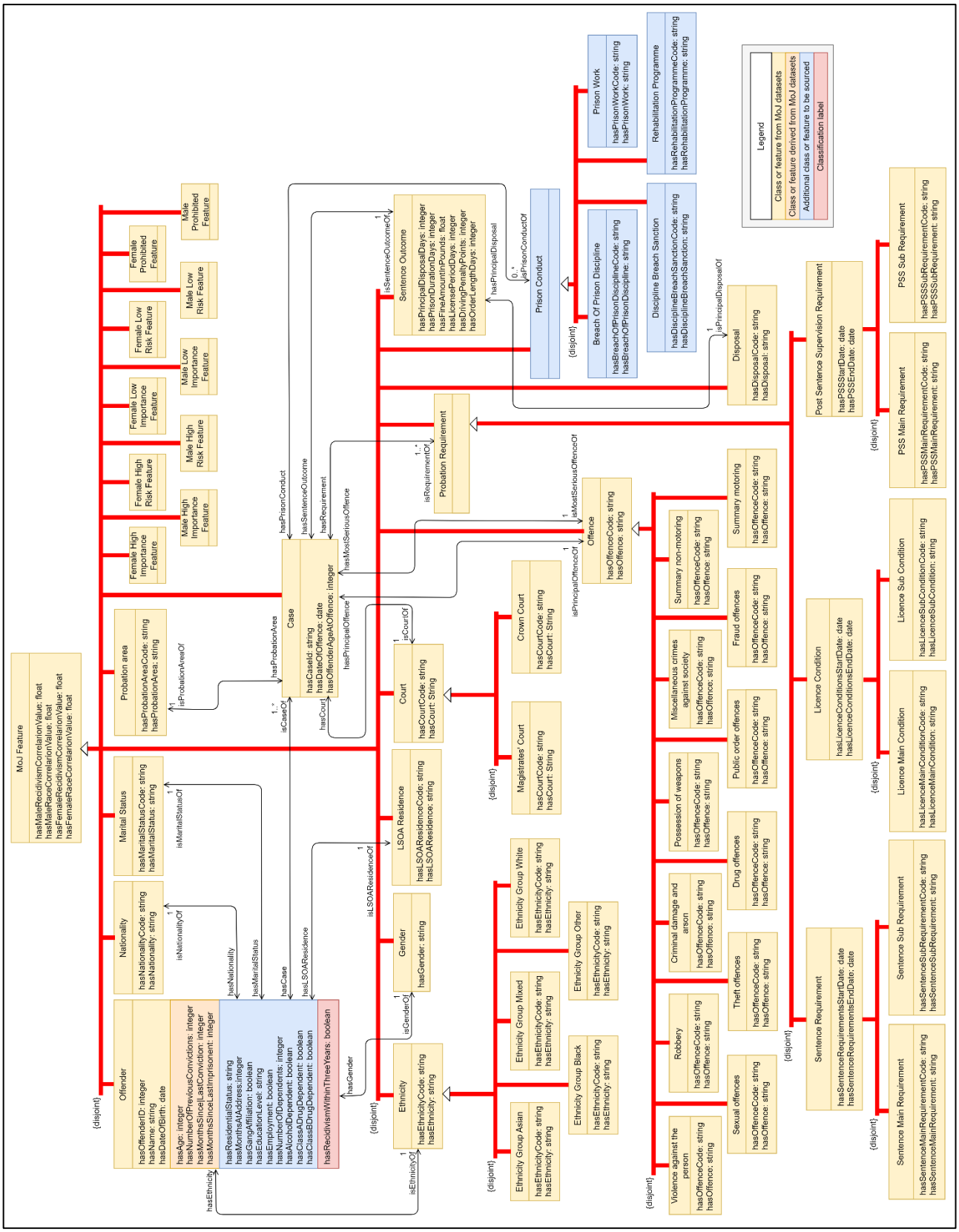


Figure 30. MoJ ontology design

Object and data instances were imported using Cellfie scripts (Appendix F) and an input spreadsheet (Appendix G) was created from lookup tables (HM Prison and Probation Service, 2024; Ministry of Justice, 2023; Ministry of Justice, 2016; Prison Reform Trust, 2022) plus Lower Super Output Areas (LSOAs) from Office for National Statistics (2023). Prefixes were added to instances of code objects to ensure uniqueness across classes. There were 35,672 LSOAs so only the first 100 were imported to prove the ontology without unnecessarily impacting performance during testing.

Horn clauses were added via Semantic Web Rule Language (SWRL) to infer descriptions from codes for code/description pairs so only code instances needed to be added to ensure data consistency (Figure 31).

Another spreadsheet (Appendix H) and Cellfie script (Appendix I) were created to import the recidivism and race correlation values to demonstrate how the metadata would be loaded (and re-loaded following regular re-reviews of the statistical analysis) in a real-world scenario. This time a statistical analysis was not possible because there was no real data, so dummy data were used to demonstrate the process. To make it look credible the dummy data were partially based upon the NIJ correlation values, but no statistical inference should be drawn. The process was to demonstrate the concept, not to provide actual results.

The race correlations in the MoJ ontology used Black and White as per the NIJ ontology, however, the MoJ ontology has a richer categorisation of ethnicity, with 19 instances across 6 classes. Without a domain expert to discuss how ethnicity could and should be classified, it was kept simple with White and Non-white as substitutes for White and Black, where White was considered to be the White class with its 4 instances of “British”, “Irish”, “Gypsy or Irish Traveler” and “Any other White background”. All other ethnicities were aligned with Non-white as the nearest equivalent to Black from the NIJ ontology. This was not intended to cause anyone any offence, it was simply a means of simplifying the MoJ ontology by aligning it to the NIJ ontology in the absence of a domain expert.

Horn clause (SWRL rule)
Offender(?o) ^ hasEthnicityCode(?o, ?code) ^ hasEthnicityCode(?e, ?code) ^ hasEthnicity(?e, ?ethnicity) -> hasEthnicity(?o, ?ethnicity)
Offender(?o) ^ hasLSOAResidenceCode(?o, ?code) ^ hasLSOAResidenceCode(?r, ?code) ^ hasLSOAResidence(?r, ?residence) -> hasLSOAResidence(?o, ?residence)
Offender(?o) ^ hasMaritalStatusCode(?o, ?code) ^ hasMaritalStatusCode(?m, ?code) ^ hasMaritalStatus (?m, ?status) -> hasMaritalStatus (?o, ?status)
Offender(?o) ^ hasNationalityCode(?o, ?code) ^ hasNationalityCode (?n, ?code) ^ hasNationality(?n, ?nationality) -> hasNationality(?o, ?nationality)
Case(?c) ^ hasCourtCode(?c, ?code) ^ hasCourtCode(?ct, ?code) ^ hasCourt(?ct, ?court) -> hasCourt(?c, ?court)
Case(?c) ^ hasPrincipleOffenceCode(?c, ?code) ^ hasOffenceCode(?o, ?code) ^ hasOffence(?o, ?offence) -> hasPrincipleOffence(?c, ?offence)
Case(?c) ^ hasMostSeriousOffenceCode(?c, ?code) ^ hasOffenceCode(?o, ?code) ^ hasOffence(?o, ?offence) -> hasMostSeriousOffenceCode(?c, ?offence)
Case(?c) ^ hasPrincipleDisposalCode(?c, ?code) ^ hasDisposalCode(?d, ?code) ^ hasDisposal(?d, ?disposal) -> hasPrincipleDisposal(?c, ?disposal)
Case(?c) ^ hasProbationAreaCode(?c, ?code) ^ hasProbationAreaCode(?p, ?code) ^ hasProbationArea(?p, ?area) -> hasProbationArea(?c, ?area)
Case(?c) ^ hasSentenceMainRequirementCode(?c, ?code) ^ hasSentenceMainRequirementCode(?r, ?code) ^ hasSentenceMainRequirement(?r, ?requirement) -> hasSentenceMainRequirement(?c, ?requirement)
Case(?c) ^ hasSentenceSubRequirementCode(?c, ?code) ^ hasSentenceSubRequirementCode(?r, ?code) ^ hasSentenceSubRequirement(?r, ?requirement) -> hasSentenceSubRequirement(?c, ?requirement)
Case(?c) ^ hasLicenceMainConditionCode(?c, ?code) ^ hasLicenceMainConditionCode(?cond, ?code) ^ hasLicenceMainCondition(?cond, ?condition) -> hasLicenceMainCondition(?c, ?condition)
Case(?c) ^ hasLicenceSubConditionCode(?c, ?code) ^ hasLicenceSubConditionCode(?cond, ?code) ^ hasLicenceSubCondition(?cond, ?condition) -> hasLicenceSubCondition(?c, ?condition)
Case(?c) ^ hasPSSMainRequirementCode(?c, ?code) ^ hasPSSMainRequirementCode(?r, ?code) ^ hasPSSMainRequirement(?r, ?requirement) -> hasPSSMainRequirement(?c, ?requirement)
Case(?c) ^ hasPSSSubRequirementCode(?c, ?code) ^ hasPSSSubRequirementCode(?r, ?code) ^ hasPSSSubRequirement(?r, ?requirement) -> hasPSSSubRequirement(?c, ?requirement)
Case(?c) ^ hasBreachOfPrisonDisciplineCode(?c, ?code) ^ hasBreachOfPrisonDisciplineCode(?b, ?code) ^ hasBreachOfPrisonDiscipline(?b, ?breach) -> hasBreachOfPrisonDiscipline(?c, ?breach)
Case(?c) ^ hasDisciplineBreachSanctionCode(?c, ?code) ^ hasDisciplineBreachSanctionCode(?s, ?code) ^ hasDisciplineBreachSanction(?s, ?sanction) -> hasDisciplineBreachSanction(?c, ?sanction)
Case(?c) ^ hasPrisonWorkCode(?c, ?code) ^ hasPrisonWorkCode(?w, ?code) ^ hasPrisonWork(?w, ?work) -> hasPrisonWork(?c, ?work)
Case(?c) ^ hasRehabilitationProgrammeCode(?c, ?code) ^ hasRehabilitationProgrammeCode(?p, ?code) ^ hasRehabilitationProgramme(?p, ?programme) -> hasRehabilitationProgramme(?c, ?programme)

Figure 31. Horn clauses to infer descriptions from codes

The initial ontology suffered from a performance issue, taking excessive time to reason running on a Microsoft Surface Laptop with Intel Core i7-1255U processor and 32GB DDR DRAM. By testing with and without different data loaded (Figure 32) and using different reasoners, which did not help and so not shown in the table, the issue was found to be caused by reasoning the high and low risk and importance defined classes with the correlation values. Having the correlation data properties set at the root class level meant that the reasoner was trying to reason all properties of all classes into one of the defined classes, even without correlation values. Ironically this is closer to the initial desired design which was not possible because data properties cannot be set against object properties or other data properties, hence why separate data properties for each feature were created against which to store the correlation values.

Design approach	Ontology memory pre-reasoning	Time to run reasoner	Ontology memory post-reasoning
Importance and risk defined classes alongside other classes and disjoint	235MB	2 hours 36 minutes	11GB
Importance and risk defined classes alongside other classes and not disjoint	235MB	29 minutes	10.2GB
Importance and risk defined classes deleted	235MB	0.5 seconds	380MB
MoJ Feature class containing importance and risk defined classes separated from other classes at ontology root	235MB	3 seconds	520MB

Figure 32. MoJ ontology performance tests

The design was modified to separate the “MoJ Feature” class holding the correlation data properties from the other classes (Figure 33) which had a significant positive impact, reducing the reasoner time from 29 minutes (without disjoint classes) to 3 seconds with only correlation instances loaded. The drawback of the updated design is further separation between the classes with the knowledge and the class with the metaknowledge, but there is separation in practical terms in both designs due to the construct of ontologies so the drawback was negligible.

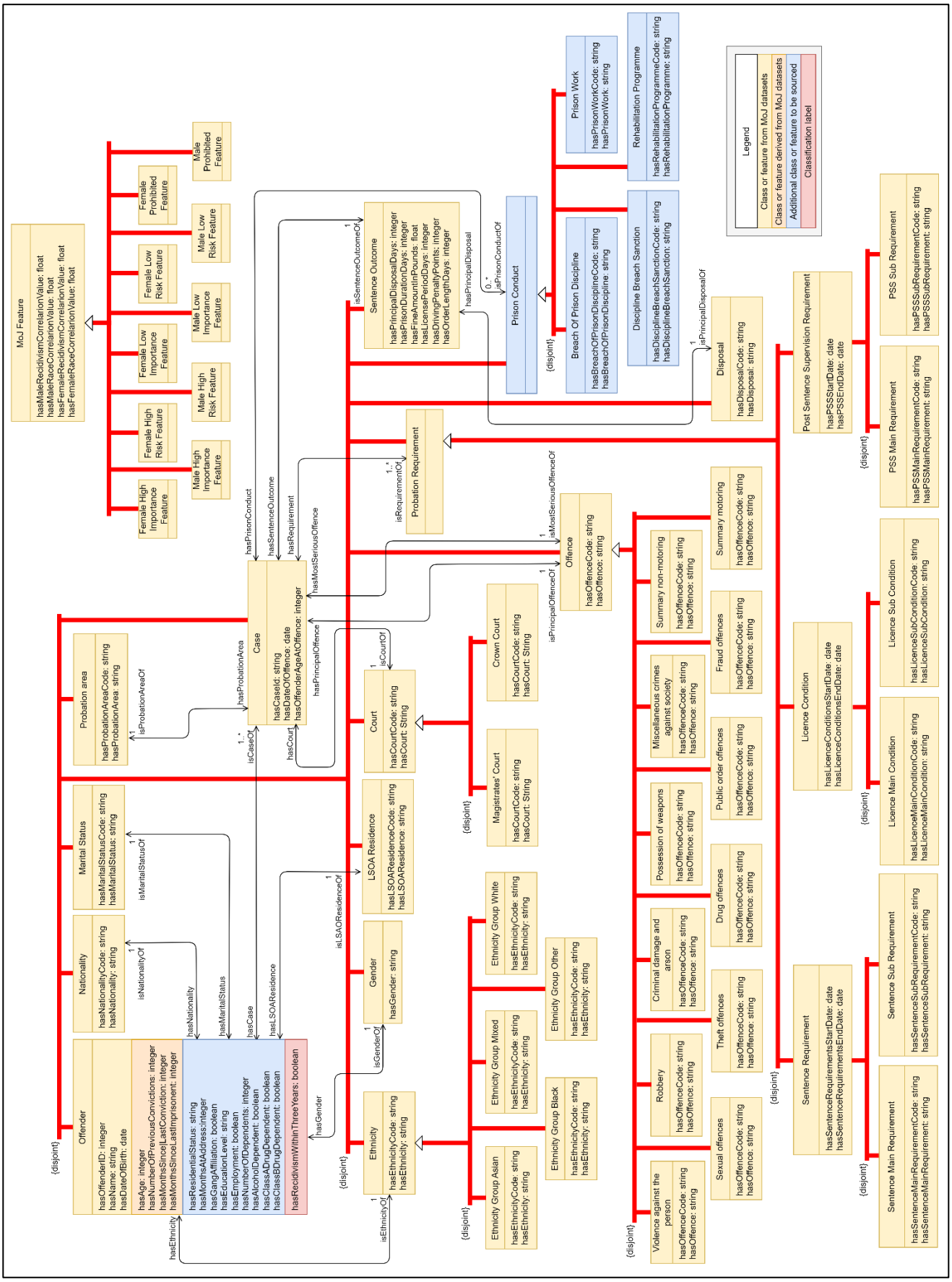


Figure 33. Updated MoJ ontology design

With the performance of the ontology improved; dummy data were created to represent sample offenders along with their cases and outcomes. Twelve offender instances and thirty-five case instances were created with credible data, including sentences aligned with Sentencing Council (N.D.), credible durations between dates of offence, sentence, licence etc., and court location aligned with probation area. The LSOA residence code was not aligned with the court or probation area due to only 100 instances being loaded, so in this case random instances were selected.

While it would have been easier to create a few offender and case instances manually, a spreadsheet was created (Appendix J) along with a Cellfie import script (Appendix K) to demonstrate how loading data from the external Data First dataset could work in practice.

The ontology was tested using manual data entry and import scripts to ensure data validation was working according to the object properties. Figure 34 shows, using manual entry, data that are not valid instances cannot be entered, and data that is valid for a different object property is rejected by the reasoner, so validation was successful.

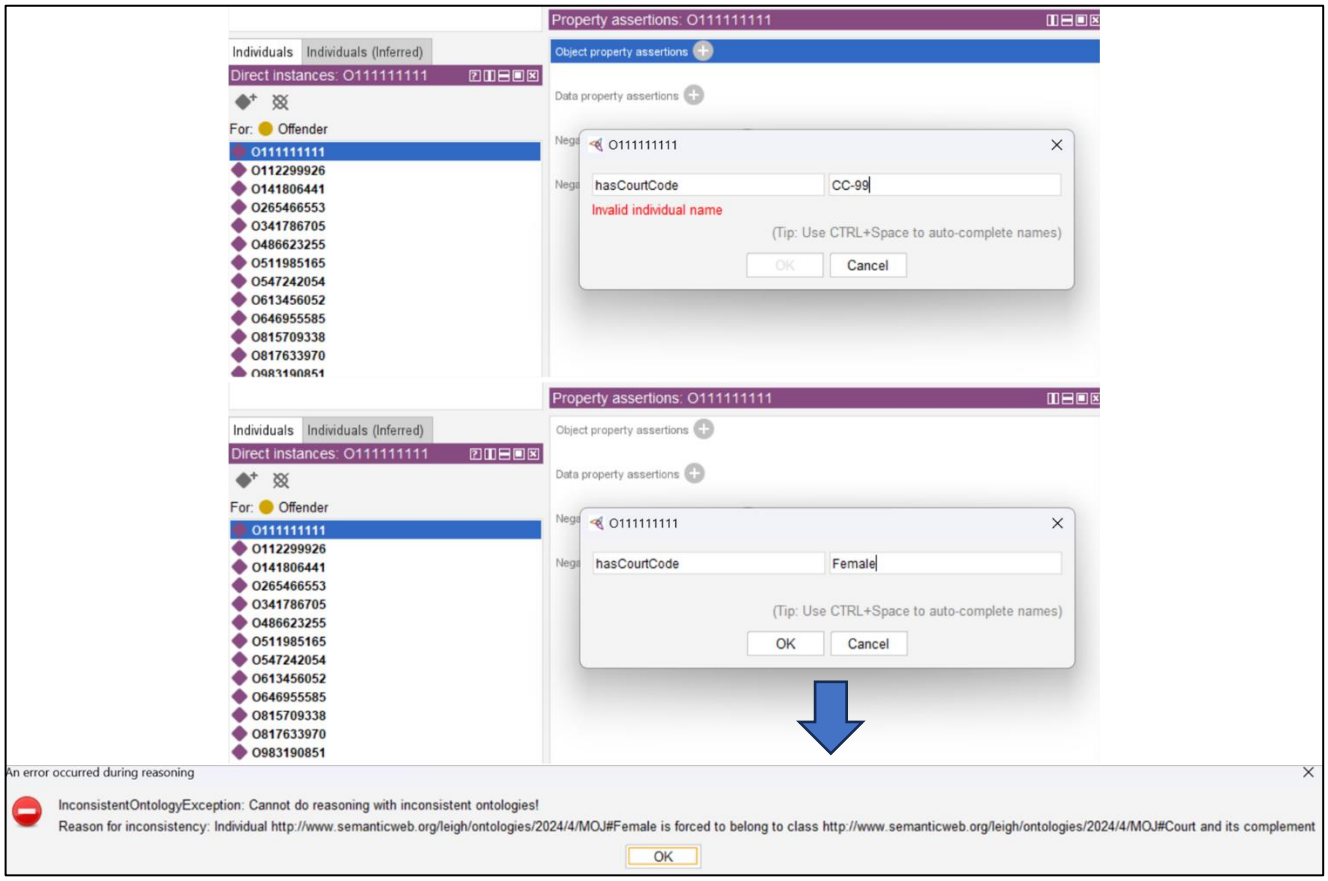


Figure 34. MoJ ontology data entry validation

Similar to the NIJ ontology, the MoJ ontology was tested using the defined classes to check which features were high or low risk or importance, and DL Queries and SPARQL queries were used for more detailed interrogation (examples in Appendix E). The SPARQL queries were more complicated than the NIJ ontology, as were the results, due to the more complex ontology with many-to-one mappings of numerous properties. Notwithstanding the additional complexity, the results were clear and accurate.

A similar worked example to the NIJ ontology follows, which first checked male high-risk and low-importance features (figure 35) to be excluded from all models.

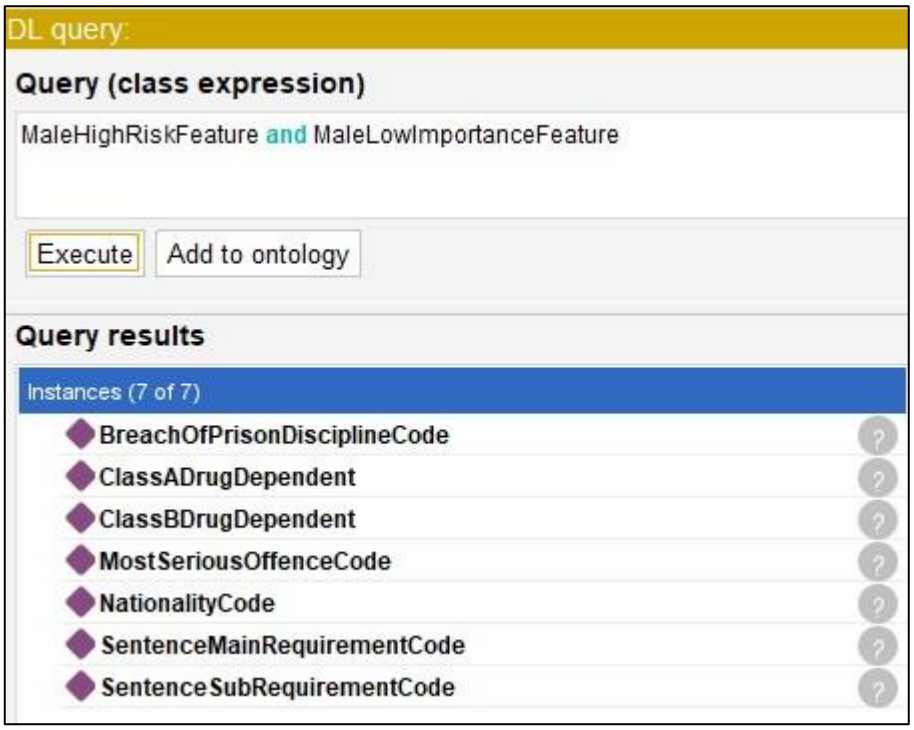


Figure 35. DL Query for male high-risk and low-importance

Next the features with higher correlation to recidivism than race were examined (figure 36). Fifteen features were selected with recidivism correlation greater than 0.12. However, some of those features also had race correlations greater than 0.1. If the final machine learning model was shown to have racial biases when tested in the ontology, the model would be re-trained by systematically removing the features with the highest correlation with race.

Snap SPARQL Query:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX moj: <http://www.semanticweb.org/leigh/ontologies/2024/4/MOJ#>

#Show features where maleRecidivismCorrelationValue is greater than maleRaceCorrelationValue
SELECT ?feature ?maleRecidivismCorrelationValue ?maleRaceCorrelationValue
WHERE {
 ?feature  moj:hasMaleRecidivismCorrelationValue ?maleRecidivismCorrelationValue ;
           moj:hasMaleRaceCorrelationValue ?maleRaceCorrelationValue .

 FILTER(?maleRecidivismCorrelationValue > ?maleRaceCorrelationValue)
}
ORDER BY DESC(?maleRecidivismCorrelationValue)
```

Execute

?feature	?maleRecidivismCorrelationValue	?maleRaceCorrelationValue
moj:MonthsSinceLastImprisonment	0.281	0.142
moj:MonthsSinceLastConviction	0.242	0.132
moj:Employment	0.217	0.126
moj:MonthsAtAddress	0.213	0.095
moj:GangAffiliation	0.185	0.086
moj:Age	0.177	0.121
moj:NumberOfPreviousConvictions	0.164	0.079
moj:OffenderAgeAtOffence	0.162	0.102
moj:PrisonDurationDays	0.162	0.092
moj:RehabilitationProgrammeCode	0.162	0.053
moj:PrincipalOffenceCode	0.161	0.092
moj:MaritalStatusCode	0.153	0.032
moj:PSSSubRequirementCode	0.152	0.056
moj:AlcoholDependent	0.132	0.098
moj:PrisonWorkCode	0.123	0.078
moj:OrderLengthDays	0.112	0.101
moj:LicenceSubConditionCode	0.099	0.045
moj:PSSMainRequirementCode	0.096	0.023
moj:LicenceMainConditionCode	0.092	0.076
moj:PrincipalDisposalDays	0.092	0.021
moj:EducationLevel	0.088	0.057
moj:PrincipalDisposalCode	0.075	0.021
moj:DrivingPenaltyPoints	0.052	0.024
moj:FineAmountPounds	0.043	0.012
moj:ProbationAreaCode	0.02	0.01

25 results

Figure 36. Features with male recidivism correlation > male race correlation

The descriptive instances of the fifteen features were examined using SPARQL (figure 37). They were exported using the code instances for training the machine learning model (Appendix E), which is beyond the scope of this project. The predictions would then be checked to ensure they matched the actual racial profiles (Appendix E) in the same way as the NIJ ontology.

Snap SPARQL Query:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX moj: <http://www.semanticweb.org/leigh/ontologies/2024/4/MOJ#>

SELECT ?offender
    (REPLACE(STR(?monthsSinceImprisonmentValue), STR(moj:), "") AS ?monthsSinceImprisonment)
    (REPLACE(STR(?monthsSinceConvictionValue), STR(moj:), "") AS ?monthsSinceConviction)
    (REPLACE(STR(?employmentValue), STR(moj:), "") AS ?employment)
    (REPLACE(STR(?monthsAtAddressValue), STR(moj:), "") AS ?monthsAtAddress)
    (REPLACE(STR(?gangAffiliationValue), STR(moj:), "") AS ?gangAffiliation)
    (REPLACE(STR(?ageValue), STR(moj:), "") AS ?age)
    (REPLACE(STR(?previousConvictionsValue), STR(moj:), "") AS ?previousConvictions)
    (REPLACE(STR(?maritalStatusValue), STR(moj:), "") AS ?maritalStatus)
    (REPLACE(STR(?alcoholDependentValue), STR(moj:), "") AS ?alcoholDependent)
    (REPLACE(STR(?caseValue), STR(moj:), "") AS ?case)
    (REPLACE(STR(?rehabilitationProgrammeValue), STR(moj:), "") AS ?rehabilitationProgramme)
    (REPLACE(STR(?principalOffenceValue), STR(moj:), "") AS ?principalOffence)
    (REPLACE(STR(?PSSSubRequirementValue), STR(moj:), "") AS ?PSSSubRequirement)
    (REPLACE(STR(?prisonWorkValue), STR(moj:), "") AS ?prisonWork)
    (REPLACE(STR(?prisonDurationValue), STR(moj:), "") AS ?prisonDuration)
WHERE {
 ?offender a moj:Offender
            FILTER(REGEX(STR(?offender), CONCAT("^", STR(moj:), "O")))
            OPTIONAL { ?offender moj:hasMonthsSinceLastImprisonment ?monthsSinceImprisonmentValue. }
            OPTIONAL { ?offender moj:hasMonthsSinceLastConviction ?monthsSinceConvictionValue. }
            OPTIONAL { ?offender moj:hasEmployment ?employmentValue. }
            OPTIONAL { ?offender moj:hasMonthsAtAddress ?monthsAtAddressValue. }
            OPTIONAL { ?offender moj:hasGangAffiliation ?gangAffiliationValue. }
            OPTIONAL { ?offender moj:hasAge ?ageValue. }
            OPTIONAL { ?offender moj:hasNumberOfPreviousConvictions ?previousConvictionsValue. }
            OPTIONAL { ?offender moj:hasMaritalStatus ?maritalStatusValue. }
            OPTIONAL { ?offender moj:hasAlcoholDependent ?alcoholDependentValue. }
 ?offender moj:hasCase ?case
            OPTIONAL { ?case moj:hasPrisonDurationDays ?prisonDurationValue. }
            OPTIONAL { ?case moj:hasPrisonDurationDays ?prisonDurationValue. }
            OPTIONAL { ?case moj:hasRehabilitationProgramme ?rehabilitationProgrammeValue. }
            OPTIONAL { ?case moj:hasPrincipalOffence ?principalOffenceValue. }
            OPTIONAL { ?case moj:hasPSSSubRequirement ?PSSSubRequirementValue. }
            OPTIONAL { ?case moj:hasPrisonWork ?prisonWorkValue. }
}
ORDER BY ?offender
```

Execute

?offender	?m...	?m...	?emp...	?m...	?gan...	?age	?p...	?maritalStatus	?alco...	?case	?rehabilitationPr...	?principalOffence	?PSSSubRequ...	?prisonWork	?priso...
moj:O112299926		11	false	13	false	63	4	Divorced or dissolv...	true	moj:C647325366		46 Theft from Shops			
moj:O112299926		11	false	13	false	63	4	Divorced or dissolv...	true	moj:C214350956		46 Theft from Shops			
moj:O112299926		11	false	13	false	63	4	Divorced or dissolv...	true	moj:C691205173		46 Theft from Shops			
moj:O112299926		11	false	13	false	63	4	Divorced or dissolv...	true	moj:C986467303		46 Theft from Shops			
moj:O112299926		11	false	13	false	63	4	Divorced or dissolv...	true	moj:C745987478		46 Theft from Shops			
moj:O141806441		2	false	52	false	59	0	Married or in civil p...	false	moj:C637473147		8.01 Assault occasioning actu...			
moj:O265466553		16	false	15	false	24	0	Married or in civil p...	true	moj:C196164791		46 Theft from Shops			
moj:O341786705	0	8	false	51	false	46	3	Married or in civil p...	true	moj:C527235880		92E.01 Possession of a contr...			
moj:O341786705	0	8	false	51	false	46	3	Married or in civil p...	true	moj:C905424044	The Bridge Progr...	4.6 Causing Death by Careles...			730
moj:O341786705	0	8	false	51	false	46	3	Married or in civil p...	true	moj:C794276181		92D.01 Possession of a contr...			
moj:O341786705	0	8	false	51	false	46	3	Married or in civil p...	true	moj:C830828808		92D.01 Possession of a contr...			
moj:O486623255	0	10	false	13	false	73	3	Widowed	false	moj:C478663808		46 Theft from Shops			
moj:O486623255	0	10	false	13	false	73	3	Widowed	false	moj:C338060576	Building Better R...	8.10 Breach of a restraining or...			730
moj:O486623255	0	10	false	13	false	73	3	Widowed	false	moj:C338060576	Becoming New M...	8.10 Breach of a restraining or...			730
moj:O486623255	0	10	false	13	false	73	3	Widowed	false	moj:C348844983		8.01 Assault occasioning actu...			
moj:O511985165	299	10	false	36	false	61	5	Single-not married/...	true	moj:C587659755		46 Theft from Shops			
moj:O511985165	299	10	false	36	false	61	5	Single-not married/...	true	moj:C618323430	Living as New Me	34 Robbery	Restorative Ju...	Workshop	365
moj:O511985165	299	10	false	36	false	61	5	Single-not married/...	true	moj:C535754077		45 Theft from Vehicle			
moj:O511985165	299	10	false	36	false	61	5	Single-not married/...	true	moj:C499266153		46 Theft from Shops			
moj:O511985165	299	10	false	36	false	61	5	Single-not married/...	true	moj:C113345537		46 Theft from Shops			
moj:O511985165	299	10	false	36	false	61	5	Single-not married/...	true	moj:C904791425		34 Robbery			
moj:O547242054	0	1	false	31	false	23	4	Single-not married/...	false	moj:C375098237		92D.01 Possession of a contr...			
moj:O547242054	0	1	false	31	false	23	4	Single-not married/...	false	moj:C851372657		46 Theft from Shops			
moj:O547242054	0	1	false	31	false	23	4	Single-not married/...	false	moj:C607333684		92D.01 Possession of a contr...			
moj:O547242054	0	1	false	31	false	23	4	Single-not married/...	false	moj:C569872265	Living as New Me	92A.09 Production supply and ...		Servery	1277
moj:O547242054	0	1	false	31	false	23	4	Single-not married/...	false	moj:C569872265	Living as New Me	92A.09 Production supply and ...		Cleaner	1277
moj:O547242054	0	1	false	31	false	23	4	Single-not married/...	false	moj:C254596091		92D.01 Possession of a contr...			
moj:O613456052	3	27	false	62	false	69	1	Married or in civil p...	true	moj:C808774974	Identity Matters	53D Fraud by false representa...		Servery	760
moj:O613456052	3	27	false	62	false	69	1	Married or in civil p...	true	moj:C808774974	Identity Matters	53D Fraud by false representa...		Cleaner	760
moj:O613456052	3	27	false	62	false	69	1	Married or in civil p...	true	moj:C723372549		46 Theft from Shops			
moj:O646955585		7	false	9	false	42	0	Single-not married/...	false	moj:C595710701		46 Theft from Shops			
moj:O815709338		4	false	58	false	20	0	Single-not married/...	false	moj:C755833865		46 Theft from Shops			
moj:O815709338		4	false	58	false	20	0	Single-not married/...	false	moj:C178906509		46 Theft from Shops			
moj:O815709338		4	false	58	false	20	0	Single-not married/...	false	moj:C861603678		46 Theft from Shops			
moj:O815709338		4	false	58	false	20	0	Single-not married/...	false	moj:C202280393		46 Theft from Shops			
moj:O815709338		4	false	58	false	20	0	Single-not married/...	false	moj:C112529789		46 Theft from Shops			
moj:O817633970		1	false	37	false	29	4	Single-not married/...	true	moj:C782821873		8.07 Racially or religiously ag...			
moj:O983190851		15	false	3	false	26	0	Married or in civil p...	false	moj:C281645560		8.07 Racially or religiously ag...			

38 results

Figure 37. Examination of features

6 Evaluation

The four artefacts were evaluated in two groups; data analysis evaluation covering the EDA and statistical analysis, and ontology evaluation covering both the NIJ and MoJ ontologies.

6.1 Data Analysis Evaluation

The methods used in both the EDA and the statistical analysis reflected best practice techniques that have not fundamentally changed since their inception (Tukey, 1977), other than the availability of computational tools to aid the analysis, and the need to transform the data.

The initial EDA (Appendix A) replaced null values with appropriate values and transformed features that were almost discrete into discrete features, for example, a feature with a range “0, 1, 2, 3 or more” was changed to “0, 1, 2, 3”. This, alongside additional transformations detailed in the EDA, allowed numerical tests to be performed. Correlations were assessed between all of the numerical features and there were no strong correlations.

The dython library was used to further assess correlations between all features, including categorical features. Again, this produced no strong results.

Finally, the categorical features were one-hot encoded and correlation heatmaps were produced to determine which features correlated most closely with recidivism, race=black, race=white, gender=male, gender=female for everyone, and race=black and race=white for males, and race=black and race=white for females, and recidivism for black males, white males, black females and white females. The intention was to identify correlations on which to form hypotheses for a more detailed statistical analysis. However while some features correlated more highly and more consistently than others, the differences were too small to restrict the detailed analysis, so instead, the hypotheses were based upon the findings from the literature review. This was not a failing in the initial EDA because the lack of strong correlations was itself a meaningful finding.

The statistical analysis continued the EDA in R (Appendix B), testing four hypotheses:

- There is a difference in recidivism by race, $\alpha=0.01$
- There is a difference in recidivism by gender, $\alpha=0.01$
- There is a difference in recidivism by age, $\alpha=0.01$
- There is a difference in offender age by race, $\alpha=0.01$

Similar data pre-processing was performed to prepare the data, with the addition of converting all ordinal features to integers to allow statistical tests to be performed. Chi-square was used to test the hypotheses for categorical features. For numerical features, the distribution was first checked, as is standard practice, using histograms and q-q plots, which showed non-normal distributions in all cases. For smaller sample sizes this would require the use of the non-parametric Mann-Whitney U test, however, because the sample size was large, with 25,835 records, the central limit theorem meant that the t-test could still be used. R both were used, producing similar results. Following the rejection of all four null hypotheses, box plots and bar charts were produced to identify the direction and scale of the differences.

The results aligned with the literature review, with recidivism higher in black over white, male over female, and young over old, and black offenders being younger than white offenders. This created similar concerns to those identified in the literature review about how to deal with the different offending patterns. The same “least bad” method from the literature review was adopted; to differentiate between genders based on fairness, even though gender is a protected characteristic, and to protect against differentiating on race. On that basis, Spearman’s rank with $\alpha=0.01$ was used to calculate the Spearman’s Rho correlation between recidivism and race for males and females for every other feature to build a table of correlations for males and females separately since they will be treated separately.

The analysis stopped there because that gave enough information regarding the risks of the different features and potential racial biases to be used in the ontology. There was much more analysis that could have been performed. Indeed, if the objective was to create machine learning models to predict recidivism it would have gone further, but that was not the purpose of this study.

Overall, the evaluation of the data analysis is that it provided all of the information necessary to move to the ontology design and so fulfilled its purpose and was considered complete and successful.

6.2 Ontology Evaluation

The initial evaluation concerned how the ontologies stored the semantic knowledge. Methods to evaluate the ontologies were examined from Raad & Cruz (2015):

- Gold standard-based was not suitable because there was no similar ontology with which to compare.
- Corpus-based was discounted because without a corpus to use as a reference, or a domain expert to interrogate, it is impossible to determine if full coverage has been achieved. They could have been evaluated against the datasets used to design them, but that would be self-fulfilling and inappropriate.
- Task-based was considered suitable because likely tasks for the ontology can be suggested without a domain expert, although having one would still be preferable.
- Criteria-based was possible, but not desirable, because without domain expertise the appropriate criteria would be subjective and potentially just fitted to the solution to ensure a positive evaluation.

Task-based evaluation was selected, using tasks such as the ease of formulating queries to interrogate the ontology, the accuracy of the responses, and how explainable the system is (Obrst et al., 2007: 12). Performance was also considered in general terms, however, cognisant of the fact that the ontologies were developed on a laptop whereas production deployment would likely be on a cloud with significantly higher computational power, so specific performance criteria were not defined.

Both ontologies enabled offender and case data to be extracted using DL Queries and SPARQL queries as described in the implementation section, with results checked and confirmed to be accurate. Only the MoJ ontology provided data validation for offender and case instances, which was tested and proven to be accurate, because the NIJ data were loaded directly without defining the object properties first. The data structure and queries were simpler in the NIJ ontology, however, the MoJ ontology was more representative of a usable ontology taking data from various sources including courts, prisons and probation and linking offender and case data. This means that data taken from the MoJ ontology would require more pre-processing, but that is because the NIJ ontology was built using data that had already been pre-processed.

The class structure of the MoJ ontology is unambiguous with relationships clearly shown in the design, and inferences are logical and explainable. The clarity of the NIJ ontology was not assessed due to its simplicity.

Next, the ontologies were evaluated for their ability to increase the transparency of potential biases; the aim of this project.

The SPARQL queries to show the ratios of actual and predicted recidivists by gender and ethnicity demonstrated that models trained to predict recidivism can be objectively proven to predict, or not, according to actual profiles. This is a big step towards transparency of output.

The data properties capturing the correlations with race and recidivism by gender for every feature provided demonstrable transparency to expose potential risks of selecting certain features for training recidivism prediction models. This is an example of improving the transparency of input, which should improve the ethical quality of output, to be measured by the transparency of output.

However, there remains a challenge. The correlation features are separate from the object and data properties used to store the criminal justice data in the ontologies. The naming convention ensured a direct mapping, but it relies upon a robust process to ensure a) the correlation values are kept up-to-date, and b) that the correlations are checked and robustly justified before training a prediction model. There is nothing in the ontologies that forces compliance. Indeed, there is nothing in the ontologies that mandates that the correlation values exist at all because they exist as separate axioms. A robust and auditable process should address this, but it was still considered a weakness in the final designs.

Notwithstanding that challenge, the MoJ ontology was successful at providing a semantic knowledge base of criminal justice for England and Wales to store and retrieve knowledge about offenders and their cases as well as metaknowledge about the correlations of features with race and recidivism, thereby increasing transparency at both the input and output stages of creating machine learning models to predict recidivism. The objective of the ontology for this project has been met.

7 Learning

The project has generated significant learning and new skills.

The literature review uncovered ethical challenges with using machine learning within the recidivism prediction domain that apply across all domains. The need to be mindful of hidden biases within the data as well as the more obvious overt biases in feature selection were critical takeaways. Additionally, the need to be careful about which performance metrics to use when measuring the performance of machine learning algorithms was well noted because the choice can have a profound impact on hiding or exposing potential biases.

First-hand knowledge and skills in ontology, specifically Protégé, increased markedly as a result of this project. Limitations of Protégé, and the semantic web in general, that were previously unknown to this author were uncovered which required changes in design and approach to fulfil the aims of the project, all of which contributed to a rich learning experience.

First Order Logic was practiced and knowledge and skills improved during this project. Horn clauses were used in the production of rules to map object descriptions to code values using SWRL. First order logic was also used to describe a SPARQL query. First order logic was a challenging area when encountered in other modules of the MSc, so self-directed learning was necessary to improve the skills and knowledge to complete this project.

SPARQL was an area of significant self-directed learning too. Having not used it previously, a lot of research, along with trial and error, was required to produce working queries, first in the simple NIJ ontology, and then in the more complex MoJ ontology. Additional code was also included to remove prefixes for a more useful output. This took time and effort to complete, demonstrating continuous learning and application even after the core requirements had been achieved.

8 Conclusions and Recommendations

Features that predict recidivism in machine learning models and less sophisticated solutions have been identified in the literature review and by analysing a publicly available dataset. They were a mix of static and dynamic features. Static features were the easiest to measure, but dynamic features improved predictive accuracy and were fairer because they can be influenced by the offender.

Features that can introduce racial bias were also identified, and an overlap was found with recidivism prediction features, which created an ethical conundrum. Using features with the highest predictive accuracy would create an inherent racial bias. For example, age correlated with recidivism, and so was used in practically every recidivism prediction model. However, it also correlated with race, so predicting on age alone would create a racially biased solution.

In some cases, the data themselves were found to be biased because they were the results of historically biased decisions and actions.

Recidivism was found to vary by gender, but unlike race, gender was found to be a characteristic that should be differentiated on in the interest of fairness.

A key conclusion was that transparency is key to an ethical solution; being able to explain why a machine learning model made a prediction.

The choice of performance metrics to assess machine learning models was found to have a profound effect on exposing potential biases. For example, accuracy can score highly while producing biased results. The most appropriate metric should be chosen depending on the context, and ideally, multiple measures should be used to mitigate the risk of hidden bias.

Ontology was found to be a credible solution to mitigate, but not prevent, the risk of racial bias. This was achieved in two ways, termed transparency of input, and transparency of output:

1. **Transparency of input:** Clear exposure of feature correlations with recidivism and race to give data scientists information about the risk of bias they could be introducing for each feature.
2. **Transparency of output:** Comparing prediction results to racial profiles to ensure parity.

Transparency of output does not automatically equate to an unbiased solution, even if the prediction profile matches the training profile due to the issue with the data themselves often being biased. It does, however, allow for the right conversations and decisions to be made to improve the data over time.

It is recommended that the prototype MoJ ontology be industrialised in three ways:

1. The design should be reviewed with a subject matter expert to validate assumptions made from the metadata to ensure it is a true reflection of the knowledge.
2. The domain should be extended beyond recidivism prediction, with the help of the subject matter expert, to cover all of criminal justice in England and Wales.
3. The ontology should be populated with real data to validate if biases exist within existing tools such as OGRS3, and to provide guardrails to protect against bias with future machine learning models.

It is further recommended to include dynamic features in the next iteration of OGRS, or whatever tool comes next, to provide a fairer prediction that allows offenders to improve their scores through rehabilitative actions.

Additionally, the semantic web should be reviewed with the World Wide Web Consortium (W3C) to produce a means for storing metaknowledge as knowledge within the ontology. The requirement is to store metaknowledge as properties of classes and other object and data properties. If this were possible in Protégé it would have been possible to store the correlation values directly against the classes and properties rather than having to create separate data property instances to store the correlation values, which was effective but prone to human error. Thought would need to be given to reasoning capabilities and optimisation for this approach because it could result in very complicated relationships.

9 References

Andrews, D.A. & Bonta, J. (2024) *The psychology of criminal conduct.* 7th ed. New York: Routledge.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016) Machine Bias. Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. [Accessed 29 April 2024].

Babad, T. & Chun, S.A. (2023) 'Enhancing Accuracy and Explainability of Recidivism Prediction Models', *36th International Florida Artificial Intelligence Research Society Conference*. Clearwater Beach, Florida, 14-17 May. 1-3.

Bārzdiņš, J., Bārzdiņš, G., Čerāns, K., Liepiņš, R. & Sproģis, A. (2010) 'UML style graphical notation and editor for OWL 2', *Perspectives in Business Informatics Research: 9th International Conference*, Rostock Germany, 29 September – 1 October 1. Berlin Heidelberg: Springer.102-114.

Baumann, G., Vegetabile, B.G., Remi, L., Kalra, N. & Bushway, S.D. (2022) *Providing Another Chance: Resetting Recidivism Risk in Criminal Background Checks*. Santa Monica, CA: RAND Corporation. Available from: https://www.rand.org/pubs/research_reports/RRA1360-1.html [Accessed 29 April 2024].

BCS (2022) Code of Conduct for BCS Members. Available from: https://www.bcs.org/media/2211/bcs-code-of-conduct.pdf [Accessed 15 July 2024].

Berk, R. (2012) *Criminal justice forecasts of risk: A machine learning approach*. New York: Springer Science & Business Media.

Berk, R., Heidari, H., Jabbari, S., Kearns, M. & Roth, A. (2021) Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1): 3-44.

Biddle, J.B. (2022) On predicting recidivism: epistemic risk, tradeoffs, and values in machine learning. *Canadian Journal of Philosophy*, 52(3): 321-341. https://doi.org/10.1017/can.2020.27

Blomberg, T., Bales, W., Mann, K., Meldrum, R. & Nedelec, J. (2010) *Validation of the COMPAS risk assessment classification instrument*. Available from: https://criminology.fsu.edu/sites/g/files/upcbnu3076/files/2021-03/Validation-of-the-COMPAS-Risk-Assessment-Classification-Instrument.pdf [Accessed 30 April 2024].

Bonta, J. (1997) Predicting adult offender recidivism. *Solicitor General Canada Research Summary*, 2(2): 1-2.

Borden, H.G. (1928) Factors for predicting parole success. *Journal of the American Institute of Criminal Law and Criminology*, 19(3): 328-336.

Brown, S.L., St. Amand, M.D. & Zamble, E. (2009) The dynamic prediction of criminal recidivism: A three-wave prospective study. *Law and human behavior*, 33: 25-45.

Buck v Davis (2017) 137 S.Ct. 759

Burgess, E.W. (1928) 'Factors determining success or failure on parole', in: Bruce, A., Burgess, E. & Harno, A. (eds) *The workings of the indeterminate sentence law and the parole system in Illinois*. Springfield, IL.: Illinois State Board of Parole. 221–234

Burrell, J. (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, 3(1), p.2053951715622512. https://doi.org/10.1177/2053951715622512

Bushway, S.D. & Piehl, A.M. (2007) The inextricable link between age and criminal history in sentencing. *Crime & Delinquency*, 53(1): 156-183. https://doi.org/10.1177/0011128706294444

Butler, L., Gunturkun, F., Karabayir, I. & Akbilgic, O. (2022) 'Logistic Regression is also a Black Box. Machine Learning Can Help', in: Shaban-Nejad, A., Michalowski, M., & Bianco. S. (eds) *AI for Disease Surveillance and Pandemic Intelligence: Intelligent Disease Detection in Action*. Springer Cham. 323-331.

Caton, S. & Haas, C. (2020) Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7): 1-38. https://doi.org/10.1145/3616865

Chouldechova, A. (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2): 153-163.

Curtis, J. (2018) On using machine learning to predict recidivism. Ph.D. thesis, Texas Tech University. Available from: https://ttu-ir.tdl.org/server/api/core/bitstreams/8e745777-200b-45d2-94a1-84355598d2ba/content [Accessed 22 April 2024].

Davies, S.T., Lloyd, C.D. & Polaschek, D.L. (2022) Does reassessment enhance the prediction of imminent criminal recidivism? Replicating Lloyd et al.(2020) with high-risk parolees. *Assessment*, 29(5): 962-980. https://doi.org/10.1177/1073191121993216

DeBellis, M. (2021) *A Practical Guide To Building OWL Ontologies, Using Protégé 5.5 and Plugins*. Available from: https://www.researchgate.net/publication/351037551_A_Practical_Guide_to_Building_OWL_Ontologies_Using_Protege_55_and_Plugins [Accessed 05 May 2024].

Dieterich, W., Mendoza, C. & Brennan, T. (2016) *COMPAS risk scales: Demonstrating accuracy equity and predictive parity*. Northpointe Inc. Research Department.

Dressel, J. & Farid, H. (2018) The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1): 1-5. https://doi.org/10.1126/sciadv.aao5580

Dressel, J. & Farid, H. (2021) The dangers of risk prediction in the criminal justice system. *MIT Case Studies in Social and Ethical Responsibilities of Computing*. https://doi.org/10.21428/2c646de5.f5896f9f

Eaglin, J.M. (2017) Constructing recidivism risk. Emory Law Journal, 67(1): 59-122.

Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C. & Venkatasubramanian, S. (2017) 'Decision making with limited feedback: Error bounds for recidivism prediction and predictive policing', *FAT/ML 2017*. Halifax, Nova Scotia, Canada, August. 1-5.

Equality and Human Rights Commission (2021) Protected characteristics. Available from: https://www.equalityhumanrights.com/equality/equality-act-2010/protected-characteristics [Accessed 27 April 2024].

Equivant (2018) Official Response to Science Advances. Available from: https://www.equivant.com/official-response-to-science-advances/ [Accessed 30 April 2024].

Equivant (2019) Practitioner's Guide to COMPAS Core. Available from: https://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf [Accessed 30 April 2024].

European Commission (2021) Proposal for a Regulation of the European Parliament and of the Council. Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain union Legislative Acts. Brussels: European Commission. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 [Accessed 24 April 2024].

Farabee, D., Zhang, S., Roberts, R.E. & Yang, J. (2010) *COMPAS validation study: Final report*. Los Angeles: Semel Institute for Neuroscience and Human Behavior.

Farrington, D.P. & Davies, D.T. (2007) Repeated contacts with the criminal justice system and offender outcomes. Statistics Canada. Available from: https://www.crim.cam.ac.uk/sites/www.crim.cam.ac.uk/files/statcanf.pdf [Accessed 25 April 2024].

Farrington, D.P. & West, D. J. (1995) 'Effects of marriage, separation and children on offending by adult males'. in Hagan, J, (ed) *Current Perspectives on Aging and the Life Cycle. vol. 4: Delinquency and Disrepute in the Life Course*. Greenwich, CT: JAI Press. 249-281.

Farrington, D.P., Gallagher, B., Morley, L., Ledger, R.J.S. & West, D.J. (2017) 'Unemployment, school leaving, and crime'. in Farrall, S. (ed) *The Termination of Criminal Careers*. London: Routledge. 101-122.

Fenton, S. (2016) Black people much more likely to be jailed over criminal offences than white people, research suggests. Available from: https://www.independent.co.uk/news/uk/home-news/courts-prison-justice-racism-black-asian-white-conviction-a7419426.html. [Accessed 29 April 2024].

Franklin, J.S., Powers, H., Erickson, J.S., McCusker, J., McGuinness, D.L. & Bennett, K.P. (2023) 'An Ontology for Reasoning About Fairness in Regression and Machine Learning', *Fifth Iberoamerican and the Fourth Indo-American Knowledge Graphs and Semantic Web Conference*. University of Zaragoza, Zaragoza, Spain, 13-15 November. Cham, Switzerland: Springer Nature. 243-261.

Ghoneim, S. (2019) Accuracy, Recall, Precision, F-Score & Specificity, which to optimize on? Available from: https://towardsdatascience.com/accuracy-recall- Page 15 precision-f-score-specificity-which-to-optimize-on-867d3f11124 [Accessed 22 April 2022].

Hanson, R. (2010) The same risk factors predict most types of recidivism. *Public Safety Canada Research Summary*, 15(4): 1-2.

Hill, C., Bagshaw, R., Hewlett, P., Perham, N., Davies, J., Maden, A. & Watt, A. (2024) Estimating the effects of secure services on reconviction. Part 1–Predictive validity of the Offending Groups Reconviction Scale (OGRS-2) and redundancy of patient social and clinical features. *International Journal of Forensic Mental Health*, 23(1): 85-91.

HM Prison & Probation Service (2023) Risk of Serious Harm Guidance 2020 v3. London: HM Prison & Probation Service. Available from:

https://assets.publishing.service.gov.uk/media/652cf8c9697260000dccf834/Risk_of_Serious_Harm_Guidance_v3.pdf [Accessed 17 April 2024].

HM Prison and Probation Service (2024) Prisoner Discipline Procedures (Adjudications). Available from: https://assets.publishing.service.gov.uk/media/664f4730ae748c43d3794155/adjudications-pf.pdf [Accessed 17 June 2024].

Hoffman, P.B. & Beck, J.L. (1974) Parole decision-making: A salient factor score. *Journal of criminal justice*, 2(3): 195-206.

Home Office (2023) Ethnicity facts and figures. Available from: https://www.ethnicity-facts-figures.service.gov.uk/crime-justice-and-the-law/policing/number-of-arrests/latest/ [Accessed: 29 April 2024].

Howard, P., Francis, B., Soothill, K. & Humphreys, L. (2009) OGRS 3: The revised offender group reconviction scale. London: Ministry of Justice.

Kasirzadeh, A. & Smart, A. (2021) March. 'The use and misuse of counterfactuals in ethical machine learning', *ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event, Canada, 3-10 March. New York: Association for Computing Machinery. 228-236. https://doi.org/10.1145/3442188.3445886

Kim, J.H. and Choi, I. (2021) Choosing the level of significance: A decision-theoretic approach. *Abacus*, 57(1): 27-71.

King, R.S. & Elderbroom, B. (2014) *Improving recidivism as a performance measure*. Washington, DC: Urban Institute. Available from: http://hint-magazine.com/wp-content/uploads/2014/10/413247-improving-recidivism.pdf [Accessed 01 April 2024].

Kleiman, M., Ostrom, B.J. & Cheesman, F.L. (2007) Using risk assessment to inform sentencing decisions for nonviolent offenders in Virginia. *Crime & Delinquency*, 53(1): 106-132. https://doi.org/10.1177/0011128706294442

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. & Mullainathan, S. (2018) Human decisions and machine predictions. *The quarterly journal of economics*, 133(1): 237-293. https://doi.org/10.1093/qje/qjx032

Kleinberg, J., Mullainathan, S. and Raghavan, M. (2016) Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*. https://doi.org/10.48550/arXiv.1609.05807

Kovalchuk, O., Karpinski, M., Banakh, S., Kasianchuk, M., Shevchuk, R. & Zagorodna, N. (2023) Prediction machine learning models on propensity convicts to criminal recidivism. *Information*, 14(3): 1-15. https://doi.org/10.3390/info14030161

La Vigne, N.G. et al. (2014) *Justice reinvestment initiative state assessment report*. Washington, DC: Urban Institute. Available from: https://www.ajc.state.ak.us/acjc/docs/resources/reform/reinvest.pdf [Accessed 01 April 2024].

Lehnert, A. (2021) Ontologies and Ethical AI. Available from: https://www.synaptica.com/ontologies-and-ethical-ai/ [Accessed 02 May 2024].

Lin, Z.J., Jung, J., Goel, S. & Skeem, J. (2020) The limits of human predictions of recidivism. *Science advances*, 6(7): 1-8.

Lloyd, C.D., Hanson, R.K., Richards, D.K. & Serin, R.C. (2020) Reassessment improves prediction of criminal recidivism: A prospective study of 3,421 individuals in New Zealand. *Psychological Assessment*, 32(6): 568. https://doi.org/10.1037/pas0000813

Lotar Rihtarić, M., Vrselja, I. & Badurina Sertić, Đ. (2017) 'The relationship between intelligence and criminal recidivism: The mediation effect of empathy', *9th International Conference of the Faculty of Education and Rehabilitation Sciences*. University of Zagreb, Zagreb, 17-19 May, Zagreb: Faculty of Education and Rehabilitation Sciences, University of Zagreb. 41-42.

May, C. (1999) *Explaining reconviction following a community sentence: the role of social factors*. Croydon: Home Office Research.

Miller, S. (2020) AI pulls ahead in recidivism prediction. Available from: https://www.route-fifty.com/emerging-tech/2020/02/ai-pulls-ahead-in-recidivism-prediction/291184/ [Accessed 27 April 2024].

Ministry of Justice (2016) Prison National Offender Management Information System (p-NOMIS) and Inmate Information System (IIS). Available from: https://www.data.gov.uk/dataset/7237e18e-c1fe-443f-881a-1113b90b3351/prison-national-offender-management-information-system-p-nomis-and-inmate-information-system-iis [Accessed 10 May 2024].

Ministry of Justice (2020) Ministry of Justice: Data First. Available from: https://www.gov.uk/guidance/ministry-of-justice-data-first [Accessed 18 March 2024].

Ministry of Justice (2023) HMPPS Accredited Programmes. Available from: https://assets.publishing.service.gov.uk/media/64085767e90e0740d3cd6fa3/HMPPS_Accredited_Programmes.docx [Accessed 17 June 2024].

Ministry of Justice (2024) Proven reoffending statistics quarterly bulletin, January to March 2022. Available from: https://assets.publishing.service.gov.uk/media/65b0f5bff2718c0014fb1c49/PRSQ_Bulletin_January_to_March_2022.pdf [Accessed 18 April 2024].

Mohdin, A. & Garcia, C.M. (2023) Defendants of colour more likely to be charged than white people, finds CPS study. Available from: https://www.theguardian.com/world/2023/feb/07/defendants-of-colour-more-likely-to-be-charged-than-white-people-finds-cps-study. [Accessed 29 April 2024].

Moore, R. (2015) A compendium of research and analysis on the Offender Assessment System (OASys). London: National Offender Management Service. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/449357/research-analysis-offender-assessment-system.pdf [Accessed 18 April 2024].

Nafekh, M. & Motiuk, L.L. (2002) *The statistical information on recidivism, revised 1 (SIR-R1) scale: a psychometric examination*. Ottawa, Ontario: Correctional Service of Canada, Research Branch.

National Institute of Justice (N.D.) Recidivism Forecasting Challenge. Available from: https://nij.ojp.gov/funding/recidivism-forecasting-challenge [Accessed 18 March 2024].

Noy, N.F. & McGuinness, D.L. (2001) *Ontology Development 101: A Guide to Creating Your First Ontology*. Knowledge Systems Laboratory.

Obrst, L., Ceusters, W., Mani, I., Ray, S. & Smith, B. (2007) The evaluation of ontologies: Toward improved semantic interoperability. *Semantic web: Revolutionizing knowledge discovery in the life sciences*: 139-158.

Office for National Statistics (2023) Lower Super Output Areas (December 2021) Names and Codes in EW (V3). Available from: https://geoportal.statistics.gov.uk/maps/0f80c523f3cd4d0fab5111572f84a2fb [Accessed 17 June 2024].

Office of Justice Programs (N.D.) NIJ Recidivism Challenge. Available from: https://data.ojp.usdoj.gov/stories/s/NIJ-s-Recidivism-Challenge-Data/daxx-hznc/ [Accessed 04 June 2024].

Olver, M.E. & Wong, S.C. (2015) Short-and long-term recidivism prediction of the PCL-R and the effects of age: a 24-year follow-up. *Personality Disorders: Theory, Research, and Treatment*, 6(1): 97-105.

Osborn, S.G. (1980) Moving home, leaving London and delinquent trends. *British Journal of Criminology*, 20(1): 54-61.

Prison Reform Trust (2022) Prison Rules and Adjudications. Available from: https://prisonreformtrust.org.uk/wp-content/uploads/2019/07/23-Prison-Rules-and-Adjudications.pdf [Accessed 14 June 2024].

Protegeproject (N.D.) cellfie-plugin. Available from: https://github.com/protegeproject/cellfie-plugin. [Accessed 12 May 2024].

Raad, J. & Cruz, C. (2015) 'A survey on ontology evaluation methods', *International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Lisbon, Portugal, 12-14 November 12-14. Cham, Springer. 179-186. https://doi.org/10.5220/0005591001790186

Richter, P., Scheurer, H., Barnett, W. & Kröber, H.L. (1996) Forecasting recidivism in delinquency by intelligence and related constructs. *Medicine, Science and the Law*, 36(4): 337-342.

Rodu, J. & Baiocchi, M. (2023) When black box algorithms are (not) appropriate. *Observational Studies*, 9(2): 79-101.

Rudin, C. (2015) New models to predict recidivism could provide better way to deter repeat crime. Available from: https://theconversation.com/new-models-to-predict-recidivism-could-provide-better-way-to-deter-repeat-crime-44165 [Accessed 27 April 2024].

Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5): 206-215.

Rudin, C., Wang, C. & Coker, B. (2020) The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1): 1-53. https://doi.org/10.1162/99608f92.6ed64b30

Sentencing Council (N.D.) Sentencing Guidelines: Magistrates. Crown Court. Available from: https://www.sentencingcouncil.org.uk/offences/ [Accessed 03 July 2024].

Serin, R.C., Mailloux, D.L. & Wilson, N.J. (2007) The dynamic risk assessment for offender re-entry (DRAOR). *Unpublished user manual*.

Shaikh, S., Vishwakarma, H., Mehta, S., Varshney, K.R., Ramamurthy, K.N. & Wei, D. (2017) 'An end-to-end machine learning pipeline that ensures fairness policies', Bloomberg Data for Good Exchange Conference. Chicago, Illinoi, 24 September. 1-5. https://doi.org/10.48550/arXiv.1710.06876

Skeem, J. & Lowenkamp, C. (2020) Using algorithms to address trade-offs inherent in predicting recidivism. *Behavioral Sciences & the Law*, 38(3): 259-278. https://doi.org/10.1002/bsl.2465

Skeem, J., Monahan, J. & Lowenkamp, C. (2016) Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and human behavior*, 40(5): 580-593. https://doi.org/10.1037/lhb0000206

Soares, E. & Angelov, P. (2019) Fair-by-design explainable models for prediction of recidivism. arXiv 2019. *arXiv preprint arXiv:1910.02043*, 1-5. https://doi.org/10.48550/arXiv.1910.02043

Squadrone, L., Croce, D. & Basili, R. (2022) 'Ethics by design for intelligent and sustainable adaptive systems', *International Conference of the Italian Association for Artificial Intelligence*. Udine, Italy, 28 November – 2 December. Cham: Springer International Publishing. 154-167.

Stephens, K. & Brown, I. (2001) OGRS2 in practice: an elastic ruler?. *Probation Journal*, 48(3): 179-187. https://doi.org/10.1177/026455050104800303

Stevenson, M.T. & Slobogin, C. (2018) Algorithmic risk assessments and the double-edged sword of youth. *Behavioral sciences & the law*, 36(5): 638-656. https://doi.org/10.1002/bsl.2384

Thomas, S. (2023) *The Fairness Fallacy: Northpointe and the COMPAS Recidivism Prediction Algorithm*. Ph.D. thesis, Columbia University. Available from: https://academiccommons.columbia.edu/doi/10.7916/ab13-jf83 [Accessed 29 April 2024].

Tollenaar, N. & van der Heijden, P.G. (2013) Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *Journal of*

the Royal Statistical Society Series A: Statistics in Society, 176(2): 565-584. https://doi.org/10.1111/j.1467-985X.2012.01056.x

Tukey, J.W. (1977) *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Ustun, B. & Rudin, C. (2015) Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102: 349-391. https://doi.org/10.1007/s10994-015-5528-6

Van Dijck, G. (2022) Predicting recidivism risk meets AI Act. *European Journal on Criminal Policy and Research*, 28(3): 407-423.

Wadsworth, C., Vera, F. & Piech, C. (2018) Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*. https://doi.org/10.48550/arXiv.1807.00199

Walmsley, J. (2021) Artificial intelligence and the value of transparency. *AI & Society*, 36(2): 585-595.

Wang, C., Han, B., Patel, B. & Rudin, C. (2023) In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *Journal of Quantitative Criminology*, 39(2): 519-581.

Wang, P., Mathieu, R., Ke, J. & Cai, H.J. (2010) 'Predicting criminal recidivism with support vector machine'. International Conference on Management and Service Science. Wuhan, China, 24-26 August. IEEE. 1-9. https://doi.org/10.1109/ICMSS.2010.5575352

Wisconsin v Loomis (2016) 881 N.W.2d 749.

Xiao, C., Ye, J., Esteves, R.M. & Rong, C. (2016) Using Spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation: Practice and Experience*, 28(14): 3866-3878. https://doi.org/10.1002/cpe.3745

Zeng, J., Ustun, B. & Rudin, C. (2017) Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 180(3): 689-722. https://doi.org/10.1111/rssa.12227

Zhang, J. (2022) 'Research on the Criminal Recidivism Prediction Based on Machine Learning Algorithm', *2nd International Conference on Business Administration and Data Science (BADS 2022)*. Hangzhou, China, 28-30 October. Atlantis Press. 1297-1306. https://doi.org/10.2991/978-94-6463-102-9_134

Appendix A. NIJ Exploratory Data Analysis

Jupyter Notebook available at: https://github.com/feaviolp/msc-project/blob/0a726abe16a82ed34d3f891361d0cca09d408edc/NIJ%20Analysis/NIJ%20EDA.ipynb

Exploratory Data Analysis of the NIJ recidivism dataset

Data pre-processing

Import libraries

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.image as mpimg
import seaborn as sns
import scipy.stats as st
from sklearn import linear_model
from sklearn.preprocessing import LabelEncoder, StandardScaler
import warnings
# ignore future deprecation
warnings.filterwarnings('ignore')
```

Load the CSV file into a pandas dataframe

In [2]:
```
url = "https://raw.githubusercontent.com/feaviolp/msc-project/main/NIJ%20Ontology/NIJ_s_Recidivism_Challenge_Full_Datase
NIJ = pd.read_csv(url)
```

Look at the shape and features of the dataset

In [3]:
```
NIJ.shape
```

Out[3]: (25835, 54)

In [4]:
```
NIJ.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25835 entries, 0 to 25834
Data columns (total 54 columns):
 #   Column                                      Non-Null Count  Dtype
---  ------                                      --------------  -----
 0   ID                                          25835 non-null  int64
 1   Gender                                      25835 non-null  object
 2   Race                                        25835 non-null  object
 3   Age_at_Release                              25835 non-null  object
 4   Residence_PUMA                              25835 non-null  int64
 5   Gang_Affiliated                             22668 non-null  object
 6   Supervision_Risk_Score_First                25360 non-null  float64
 7   Supervision_Level_First                     24115 non-null  object
 8   Education_Level                             25835 non-null  object
 9   Dependents                                  25835 non-null  object
 10  Prison_Offense                              22558 non-null  object
 11  Prison_Years                                25835 non-null  object
 12  Prior_Arrest_Episodes_Felony                25835 non-null  object
 13  Prior_Arrest_Episodes_Misd                  25835 non-null  object
 14  Prior_Arrest_Episodes_Violent               25835 non-null  object
 15  Prior_Arrest_Episodes_Property              25835 non-null  object
 16  Prior_Arrest_Episodes_Drug                  25835 non-null  object
 17  Prior_Arrest_Episodes_PPViolationCharges    25835 non-null  object
 18  Prior_Arrest_Episodes_DVCharges             25835 non-null  bool
 19  Prior_Arrest_Episodes_GunCharges            25835 non-null  bool
 20  Prior_Conviction_Episodes_Felony            25835 non-null  object
 21  Prior_Conviction_Episodes_Misd              25835 non-null  object
 22  Prior_Conviction_Episodes_Viol              25835 non-null  bool
 23  Prior_Conviction_Episodes_Prop              25835 non-null  object
 24  Prior_Conviction_Episodes_Drug              25835 non-null  object
 25  Prior_Conviction_Episodes_PPViolationCharges  25835 non-null  bool
 26  Prior_Conviction_Episodes_DomesticViolenceCharges  25835 non-null  bool
 27  Prior_Conviction_Episodes_GunCharges        25835 non-null  bool
 28  Prior_Revocations_Parole                    25835 non-null  bool
 29  Prior_Revocations_Probation                 25835 non-null  bool
 30  Condition_MH_SA                             25835 non-null  bool
 31  Condition_Cog_Ed                            25835 non-null  bool
 32  Condition_Other                             25835 non-null  bool
 33  Violations_ElectronicMonitoring             25835 non-null  bool
```

```
 34  Violations_Instruction                  25835 non-null  bool
 35  Violations_FailToReport                 25835 non-null  bool
 36  Violations_MoveWithoutPermission        25835 non-null  bool
 37  Delinquency_Reports                     25835 non-null  object
 38  Program_Attendances                     25835 non-null  object
 39  Program_UnexcusedAbsences               25835 non-null  object
 40  Residence_Changes                       25835 non-null  object
 41  Avg_Days_per_DrugTest                   19732 non-null  float64
 42  DrugTests_THC_Positive                  20663 non-null  float64
 43  DrugTests_Cocaine_Positive              20663 non-null  float64
 44  DrugTests_Meth_Positive                 20663 non-null  float64
 45  DrugTests_Other_Positive                20663 non-null  float64
 46  Percent_Days_Employed                   25373 non-null  float64
 47  Jobs_Per_Year                           25027 non-null  float64
 48  Employment_Exempt                       25835 non-null  bool
 49  Recidivism_Within_3years                25835 non-null  bool
 50  Recidivism_Arrest_Year1                 25835 non-null  bool
 51  Recidivism_Arrest_Year2                 25835 non-null  bool
 52  Recidivism_Arrest_Year3                 25835 non-null  bool
 53  Training_Sample                         25835 non-null  int64
dtypes: bool(20), float64(8), int64(3), object(23)
memory usage: 7.2+ MB
```

The dataset has 25,836 rows and 54 columns.

Some of the columns have missing values (less than 25,836 values).

Next describe the dataset, then look at the top and bottom 4 rows.

In [5]:
```
NIJ.describe()
```

Out[5]:

	ID	Residence_PUMA	Supervision_Risk_Score_First	Avg_Days_per_DrugTest	DrugTests_THC_Positive	DrugTests_Cocain
count	25835.000000	25835.000000	25360.000000	19732.000000	20663.000000	206
mean	13314.004838	12.361796	6.082216	93.890044	0.063350	
std	7722.206327	7.133742	2.381442	117.169847	0.138453	
min	1.000000	1.000000	1.000000	0.500000	0.000000	
25%	6626.500000	6.000000	4.000000	28.837366	0.000000	
50%	13270.000000	12.000000	6.000000	55.424812	0.000000	
75%	20021.500000	18.000000	8.000000	110.333333	0.071429	
max	26761.000000	25.000000	10.000000	1088.500000	1.000000	

In [6]:
```
NIJ.head()
```

Out[6]:

	ID	Gender	Race	Age_at_Release	Residence_PUMA	Gang_Affiliated	Supervision_Risk_Score_First	Supervision_Level_First	Educat
0	1	M	BLACK	43-47	16	False	3.0	Standard	At I
1	2	M	BLACK	33-37	16	False	6.0	Specialized	Le
2	3	M	BLACK	48 or older	24	False	7.0	High	At I
3	4	M	WHITE	38-42	16	False	7.0	High	Le
4	5	M	WHITE	33-37	16	False	4.0	Specialized	Le

5 rows × 54 columns

In [7]:
```
NIJ.tail()
```

Out[7]:

	ID	Gender	Race	Age_at_Release	Residence_PUMA	Gang_Affiliated	Supervision_Risk_Score_First	Supervision_Level_First

25830	26756	M	BLACK	23-27	9	False	5.0	Standard
25831	26758	M	WHITE	38-42	25	False	5.0	Standard
25832	26759	M	BLACK	33-37	15	False	5.0	Standard
25833	26760	F	WHITE	33-37	15	NaN	5.0	Standard
25834	26761	M	WHITE	28-32	12	False	5.0	Standard

5 rows × 54 columns

Drop ID and Training_Sample because they will not be requried for the EDA.

In [8]:
```
NIJ.drop(columns=['ID', 'Training_Sample'], inplace=True)
```

Also drop Recidivism_Arrest_Year1, Recidivism_Arrest_Year2 and Recidivism_Arrest_Year3 because we're only interested in Recidivism_Within_3years

In [9]:
```
NIJ.drop(columns=['Recidivism_Arrest_Year1', 'Recidivism_Arrest_Year2', 'Recidivism_Arrest_Year3'], inplace=True)
```

Now take a look at the rows with missing values.

In [10]:
```
NIJ.isna().sum()
```

Out[10]:
```
Gender                                          0
Race                                            0
Age_at_Release                                  0
Residence_PUMA                                  0
Gang_Affiliated                              3167
Supervision_Risk_Score_First                  475
Supervision_Level_First                      1720
Education_Level                                 0
Dependents                                      0
Prison_Offense                               3277
Prison_Years                                    0
Prior_Arrest_Episodes_Felony                    0
Prior_Arrest_Episodes_Misd                      0
Prior_Arrest_Episodes_Violent                   0
Prior_Arrest_Episodes_Property                  0
Prior_Arrest_Episodes_Drug                      0
Prior_Arrest_Episodes_PPViolationCharges        0
Prior_Arrest_Episodes_DVCharges                 0
Prior_Arrest_Episodes_GunCharges                0
Prior_Conviction_Episodes_Felony                0
Prior_Conviction_Episodes_Misd                  0
Prior_Conviction_Episodes_Viol                  0
Prior_Conviction_Episodes_Prop                  0
Prior_Conviction_Episodes_Drug                  0
Prior_Conviction_Episodes_PPViolationCharges    0
Prior_Conviction_Episodes_DomesticViolenceCharges    0
Prior_Conviction_Episodes_GunCharges            0
Prior_Revocations_Parole                        0
Prior_Revocations_Probation                     0
Condition_MH_SA                                 0
Condition_Cog_Ed                                0
Condition_Other                                 0
Violations_ElectronicMonitoring                 0
Violations_Instruction                          0
Violations_FailToReport                         0
Violations_MoveWithoutPermission                0
Delinquency_Reports                             0
Program_Attendances                             0
Program_UnexcusedAbsences                       0
Residence_Changes                               0
Avg_Days_per_DrugTest                        6103
DrugTests_THC_Positive                       5172
DrugTests_Cocaine_Positive                   5172
```

```
DrugTests_Meth_Positive                5172
DrugTests_Other_Positive               5172
Percent_Days_Employed                   462
Jobs_Per_Year                           808
Employment_Exempt                         0
Recidivism_Within_3years                  0
dtype: int64
```

Replace null values with appropriate values:

- Gang_Affiliated is missing only for Female offenders so replace with NA
- Supervision_Risk_Score_First is INTEGER so replace with the most frequently occuring value
- Supervision_Level_First is CATEGORICAL so replace with the most frequently occuring value
- Prison_Offense is categorical and includes "Other" so replace with "Other"
- Avg_Days_per_DrugTest is FLOAT so replace with average value
- DrugTests_THC_Positive is FLOAT but replace with 0
- DrugTests_Cocaine_Positive is FLOAT but replace with 0
- DrugTests_Meth_Positive is FLOAT but replace with 0
- DrugTests_Other_Positive is FLOAT but replace with 0
- Percent_Days_Employed is FLOAT so replace with average value
- Jobs_Per_Year is FLOAT so replace with average value

In [11]:
```
NIJ['Gang_Affiliated'].fillna('NA', inplace=True)
NIJ['Supervision_Risk_Score_First'].fillna(NIJ['Supervision_Risk_Score_First'].mode().iloc[0], inplace=True)
NIJ['Supervision_Level_First'].fillna(NIJ['Supervision_Level_First'].mode().iloc[0], inplace=True)
NIJ['Prison_Offense'].fillna('Other', inplace=True)
NIJ['Avg_Days_per_DrugTest'].fillna(NIJ['Avg_Days_per_DrugTest'].mean(), inplace=True)
NIJ['DrugTests_THC_Positive'].fillna('0', inplace=True)
NIJ['DrugTests_Cocaine_Positive'].fillna('0', inplace=True)
NIJ['DrugTests_Meth_Positive'].fillna('0', inplace=True)
NIJ['DrugTests_Other_Positive'].fillna('0', inplace=True)
NIJ['Percent_Days_Employed'].fillna(NIJ['Percent_Days_Employed'].mean(), inplace=True)
NIJ['Jobs_Per_Year'].fillna(NIJ['Jobs_Per_Year'].mean(), inplace=True)
```

Re-check to make sure there are no missing values

In [12]:
```
NIJ.isna().sum()
```

Out[12]:
```
Gender                                          0
Race                                            0
Age_at_Release                                  0
Residence_PUMA                                  0
Gang_Affiliated                                 0
Supervision_Risk_Score_First                    0
Supervision_Level_First                         0
Education_Level                                 0
Dependents                                      0
Prison_Offense                                  0
Prison_Years                                    0
Prior_Arrest_Episodes_Felony                    0
Prior_Arrest_Episodes_Misd                      0
Prior_Arrest_Episodes_Violent                   0
Prior_Arrest_Episodes_Property                  0
Prior_Arrest_Episodes_Drug                      0
Prior_Arrest_Episodes_PPViolationCharges        0
Prior_Arrest_Episodes_DVCharges                 0
Prior_Arrest_Episodes_GunCharges                0
Prior_Conviction_Episodes_Felony                0
Prior_Conviction_Episodes_Misd                  0
Prior_Conviction_Episodes_Viol                  0
Prior_Conviction_Episodes_Prop                  0
Prior_Conviction_Episodes_Drug                  0
Prior_Conviction_Episodes_PPViolationCharges    0
Prior_Conviction_Episodes_DomesticViolenceCharges 0
Prior_Conviction_Episodes_GunCharges            0
Prior_Revocations_Parole                        0
Prior_Revocations_Probation                     0
Condition_MH_SA                                 0
Condition_Cog_Ed                                0
Condition_Other                                 0
Violations_ElectronicMonitoring                 0
Violations_Instruction                          0
Violations_FailToReport                         0
Violations_MoveWithoutPermission                0
Delinquency_Reports                             0
Program_Attendances                             0
```

```
Program_UnexcusedAbsences                0
Residence_Changes                        0
Avg_Days_per_DrugTest                    0
DrugTests_THC_Positive                   0
DrugTests_Cocaine_Positive               0
DrugTests_Meth_Positive                  0
DrugTests_Other_Positive                 0
Percent_Days_Employed                    0
Jobs_Per_Year                            0
Employment_Exempt                        0
Recidivism_Within_3years                 0
dtype: int64
```

Some features contain integers but max out with, "x or more". To enable numerical analysis the "or more" will be removed:

- Dependents: "3 or more" changed to "3"
- Prior_Arrest_Episodes_Felony: "10 more more" changed to "10"
- Prior_Arrest_Episodes_Misd: "6 or more" changed to "6"
- Prior_Arrest_Episodes_Violent: "3 or more" changed to "3"
- Prior_Arrest_Episodes_Property: "5 or more" changed to "5"
- Prior_Arrest_Episodes_Drug: "5 or more" changed to "5"
- Prior_Arrest_Episodes_PPViolationCharges: "5 or more" changed to "5"
- Prior_Conviction_Episodes_Felony.replace: "3 or more" changed to "3"
- Prior_Conviction_Episodes_Misd.replace: "4 or more" changed to "4"
- Prior_Conviction_Episodes_Prop.replace: "3 or more" changed to "3"
- Prior_Conviction_Episodes_Drug.replace: "2 or more" changed to "2"
- Delinquency_Reports: "4 or more" changed to "4"
- Program_Attendances: "10 more more" changed to "10"
- Program_UnexcusedAbsences: "3 or more" changed to "3"
- Residence_Changes: "3 or more" changed to "3"

In [13]:
```
NIJ.Dependents.replace("3 or more","3", inplace=True)
NIJ.Prior_Arrest_Episodes_Felony.replace("10 or more","10", inplace=True)
NIJ.Prior_Arrest_Episodes_Misd.replace("6 or more","6", inplace=True)
NIJ.Prior_Arrest_Episodes_Violent.replace("3 or more","3", inplace=True)
NIJ.Prior_Arrest_Episodes_Property.replace("5 or more","5", inplace=True)
NIJ.Prior_Arrest_Episodes_Drug.replace("5 or more","5", inplace=True)
NIJ.Prior_Arrest_Episodes_PPViolationCharges.replace("5 or more","5", inplace=True)
NIJ.Prior_Conviction_Episodes_Felony.replace("3 or more","3", inplace=True)
NIJ.Prior_Conviction_Episodes_Misd.replace("4 or more","4", inplace=True)
NIJ.Prior_Conviction_Episodes_Prop.replace("3 or more","3", inplace=True)
NIJ.Prior_Conviction_Episodes_Drug.replace("2 or more","2", inplace=True)
NIJ.Delinquency_Reports.replace("4 or more","4", inplace=True)
NIJ.Program_Attendances.replace("10 or more","10", inplace=True)
NIJ.Program_UnexcusedAbsences.replace("3 or more","3", inplace=True)
NIJ.Residence_Changes.replace("3 or more","3", inplace=True)
```

Age_at_Release is populated with ranges which won't be assessed as numeric, so change each value to the first number in each range

In [14]:
```
sorted(NIJ['Age_at_Release'].unique())
```

Out[14]:
```
['18-22', '23-27', '28-32', '33-37', '38-42', '43-47', '48 or older']
```

In [15]:
```
NIJ.Age_at_Release.replace("18-22","18", inplace=True)
NIJ.Age_at_Release.replace("23-27","23", inplace=True)
NIJ.Age_at_Release.replace("28-32","28", inplace=True)
NIJ.Age_at_Release.replace("33-37","33", inplace=True)
NIJ.Age_at_Release.replace("38-42","38", inplace=True)
NIJ.Age_at_Release.replace("43-47","43", inplace=True)
NIJ.Age_at_Release.replace("48 or older","48", inplace=True)
```

In [16]:
```
sorted(NIJ['Age_at_Release'].unique())
```

Out[16]:
```
['18', '23', '28', '33', '38', '43', '48']
```

Now replace M with Male and F with Female in Gender to make it easier to read later after the are one-hot encoded

In [17]:
```
NIJ.Gender.replace("M","Male", inplace=True)
NIJ.Gender.replace("F","Female", inplace=True)
```

In [18]:
```
NIJ
```

Out[18]:

	Gender	Race	Age_at_Release	Residence_PUMA	Gang_Affiliated	Supervision_Risk_Score_First	Supervision_Level_First	Educa
0	Male	BLACK	43	16	False	3.0	Standard	At
1	Male	BLACK	33	16	False	6.0	Specialized	Le
2	Male	BLACK	48	24	False	7.0	High	At
3	Male	WHITE	38	16	False	7.0	High	Le
4	Male	WHITE	33	16	False	4.0	Specialized	Le
...	...	...	...	...	...	...	...	
25830	Male	BLACK	23	9	False	5.0	Standard	At
25831	Male	WHITE	38	25	False	5.0	Standard	At
25832	Male	BLACK	33	15	False	5.0	Standard	At
25833	Female	WHITE	33	15	NA	5.0	Standard	At
25834	Male	WHITE	28	12	False	5.0	Standard	H

25835 rows × 49 columns

Now convert all of the modified columns to numeric.

In [19]:
```
NIJ[["Dependents"]] = NIJ[["Dependents"]].apply(pd.to_numeric)
NIJ[["Prior_Arrest_Episodes_Felony", "Prior_Arrest_Episodes_Misd"]] = NIJ[["Prior_Arrest_Episodes_Felony", "Prior_Arrest
NIJ[["Prior_Arrest_Episodes_Violent", "Prior_Arrest_Episodes_Property"]] = NIJ[["Prior_Arrest_Episodes_Violent", "Prior_
NIJ[["Prior_Arrest_Episodes_Drug", "Prior_Arrest_Episodes_PPViolationCharges"]] = NIJ[["Prior_Arrest_Episodes_Drug", "Pr
NIJ[["Prior_Conviction_Episodes_Felony", "Prior_Conviction_Episodes_Misd"]] = NIJ[["Prior_Conviction_Episodes_Felony", "
NIJ[["Prior_Conviction_Episodes_Prop", "Prior_Conviction_Episodes_Drug"]] = NIJ[["Prior_Conviction_Episodes_Prop", "Prio
NIJ[["Delinquency_Reports", "Program_Attendances"]] = NIJ[["Delinquency_Reports", "Program_Attendances"]].apply(pd.to_nu
NIJ[["Program_UnexcusedAbsences", "Residence_Changes"]] = NIJ[["Program_UnexcusedAbsences", "Residence_Changes"]].apply(

NIJ[["DrugTests_THC_Positive", "DrugTests_Cocaine_Positive"]] = NIJ[["DrugTests_THC_Positive", "DrugTests_Cocaine_Positi
NIJ[["DrugTests_Meth_Positive", "DrugTests_Other_Positive"]] = NIJ[["DrugTests_Meth_Positive", "DrugTests_Other_Positive

NIJ[["Age_at_Release"]] = NIJ[["Age_at_Release"]].apply(pd.to_numeric)
```

In [20]:
```
NIJ['Prior_Arrest_Episodes_Misd'].unique()
```

Out[20]:
```
array([6, 4, 0, 1, 3, 5, 2])
```

In [21]:
```
NIJ['Gang_Affiliated'] = NIJ['Gang_Affiliated'].astype('bool')
```

Exploratory Data Analysis

Examine numerical and categorical features.

In [22]:
```
numeric_features = NIJ.select_dtypes(include=[np.number])
numeric_features.columns
```

Out[22]:
```
Index(['Age_at_Release', 'Residence_PUMA', 'Supervision_Risk_Score_First',
       'Dependents', 'Prior_Arrest_Episodes_Felony',
       'Prior_Arrest_Episodes_Misd', 'Prior_Arrest_Episodes_Violent',
```

```
       'Prior_Arrest_Episodes_Property', 'Prior_Arrest_Episodes_Drug',
       'Prior_Arrest_Episodes_PPViolationCharges',
       'Prior_Conviction_Episodes_Felony', 'Prior_Conviction_Episodes_Misd',
       'Prior_Conviction_Episodes_Prop', 'Prior_Conviction_Episodes_Drug',
       'Delinquency_Reports', 'Program_Attendances',
       'Program_UnexcusedAbsences', 'Residence_Changes',
       'Avg_Days_per_DrugTest', 'DrugTests_THC_Positive',
       'DrugTests_Cocaine_Positive', 'DrugTests_Meth_Positive',
       'DrugTests_Other_Positive', 'Percent_Days_Employed', 'Jobs_Per_Year'],
      dtype='object')
```

In [23]:
```
categorical_features = NIJ.select_dtypes(include=[object])
categorical_features.columns
```

Out[23]:
```
Index(['Gender', 'Race', 'Supervision_Level_First', 'Education_Level',
       'Prison_Offense', 'Prison_Years'],
      dtype='object')
```

Estimate Skewness and Kurtosis of numerical features

In [24]:
```
numeric_features.skew()
```

Out[24]:
```
Age_at_Release                        0.262007
Residence_PUMA                        0.098660
Supervision_Risk_Score_First         -0.136975
Dependents                            0.050562
Prior_Arrest_Episodes_Felony          0.040836
Prior_Arrest_Episodes_Misd           -0.115781
Prior_Arrest_Episodes_Violent         0.688265
Prior_Arrest_Episodes_Property        0.321845
Prior_Arrest_Episodes_Drug            0.616521
Prior_Arrest_Episodes_PPViolationCharges    0.196785
Prior_Conviction_Episodes_Felony      0.160138
Prior_Conviction_Episodes_Misd        0.295852
Prior_Conviction_Episodes_Prop        0.546661
Prior_Conviction_Episodes_Drug        0.442740
Delinquency_Reports                   1.162612
Program_Attendances                   0.819652
Program_UnexcusedAbsences             2.093484
Residence_Changes                     0.941644
Avg_Days_per_DrugTest                 3.998832
DrugTests_THC_Positive                3.948409
DrugTests_Cocaine_Positive            8.049606
DrugTests_Meth_Positive              10.485032
DrugTests_Other_Positive             12.752432
Percent_Days_Employed                 0.035112
Jobs_Per_Year                         1.485389
dtype: float64
```

In [25]:
```
numeric_features.kurt()
```

Out[25]:
```
Age_at_Release                        -1.092770
Residence_PUMA                        -1.239440
Supervision_Risk_Score_First          -0.734744
Dependents                            -1.560358
Prior_Arrest_Episodes_Felony          -1.362785
Prior_Arrest_Episodes_Misd            -1.515902
Prior_Arrest_Episodes_Violent         -0.857591
Prior_Arrest_Episodes_Property        -1.373212
Prior_Arrest_Episodes_Drug            -0.880327
Prior_Arrest_Episodes_PPViolationCharges    -1.488162
Prior_Conviction_Episodes_Felony      -1.446276
Prior_Conviction_Episodes_Misd        -1.415104
Prior_Conviction_Episodes_Prop        -1.237449
Prior_Conviction_Episodes_Drug        -1.427408
Delinquency_Reports                   -0.425493
Program_Attendances                   -0.877871
Program_UnexcusedAbsences              2.852693
Residence_Changes                     -0.395310
Avg_Days_per_DrugTest                 22.621673
DrugTests_THC_Positive                20.278025
DrugTests_Cocaine_Positive            88.639150
DrugTests_Meth_Positive              154.769982
DrugTests_Other_Positive             247.210002
Percent_Days_Employed                 -1.720806
Jobs_Per_Year                          3.325226
dtype: float64
```

Correlations

In [26]:
```
# NIJ.corr(method ='pearson')
numeric_features.corr()
```

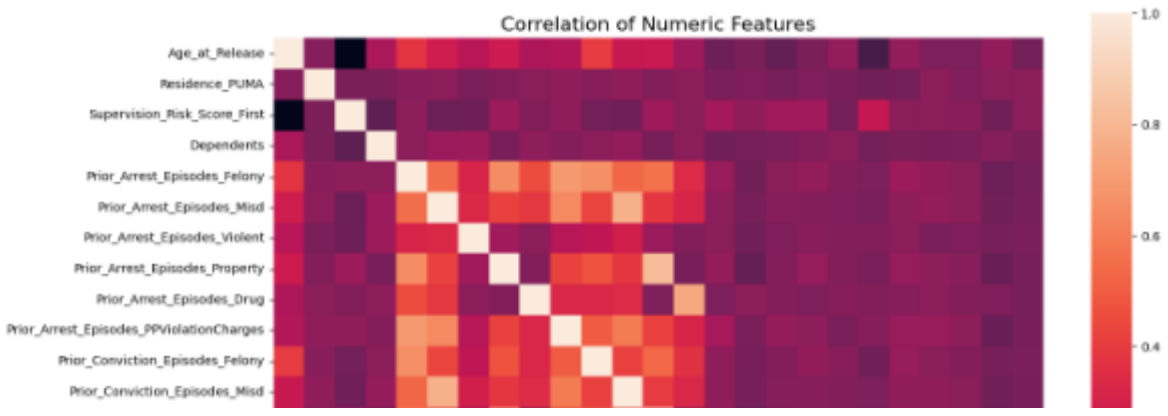
Out[26]:

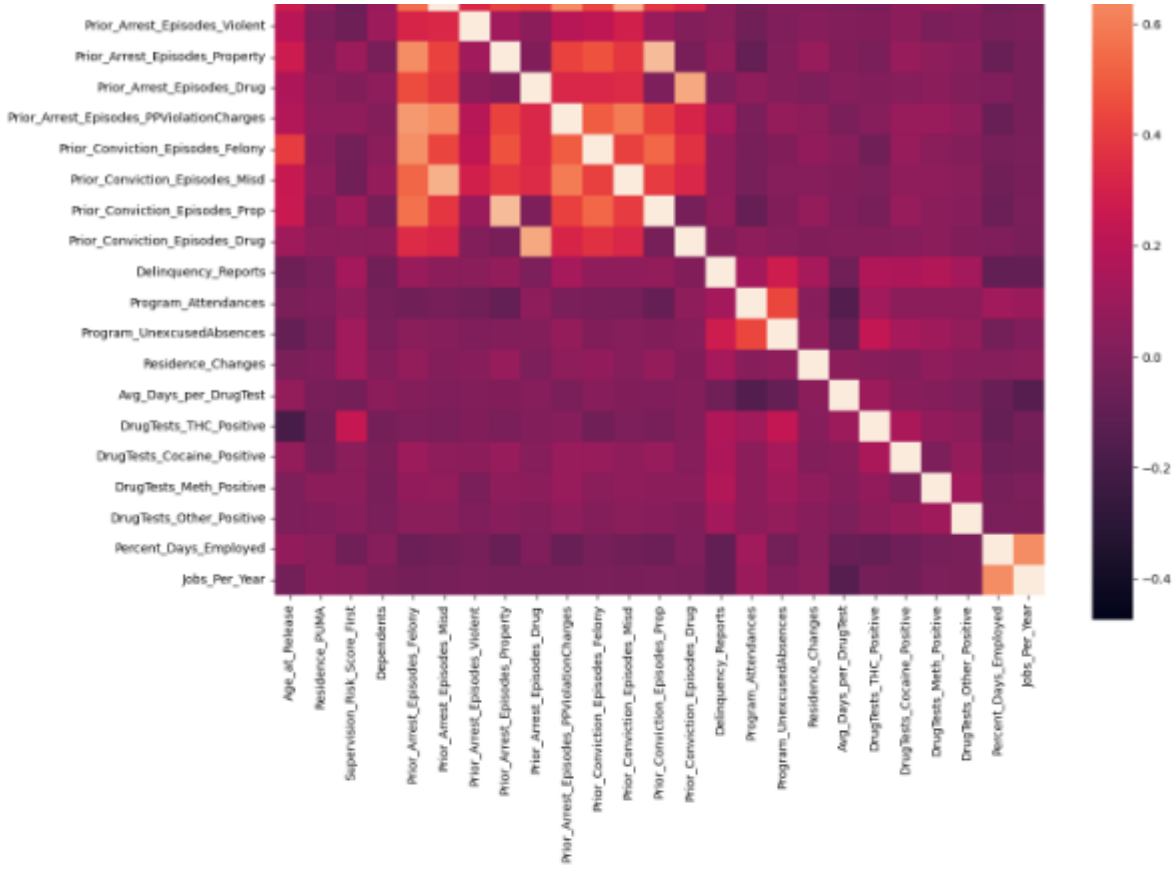
	Age_at_Release	Residence_PUMA	Supervision_Risk_Score_First	Dependents	Prior_Arrest_Epi
Age_at_Release	1.000000	0.022132	-0.474483	0.149536	
Residence_PUMA	0.022132	1.000000	-0.009568	-0.010828	
Supervision_Risk_Score_First	-0.474483	-0.009568	1.000000	-0.113169	
Dependents	0.149536	-0.010828	-0.113169	1.000000	
Prior_Arrest_Episodes_Felony	0.370477	0.028485	0.049855	0.046847	
Prior_Arrest_Episodes_Misd	0.272936	0.047212	-0.068378	0.092013	
Prior_Arrest_Episodes_Violent	0.193949	-0.013974	-0.062500	0.101109	
Prior_Arrest_Episodes_Property	0.263956	0.006619	0.097777	-0.020090	
Prior_Arrest_Episodes_Drug	0.155872	0.036650	0.006608	0.051614	
Prior_Arrest_Episodes_PPViolationCharges	0.173532	0.060351	0.060859	0.018629	
Prior_Conviction_Episodes_Felony	0.398924	0.027633	-0.041738	0.040165	
Prior_Conviction_Episodes_Misd	0.246866	0.058480	-0.051761	0.072735	
Prior_Conviction_Episodes_Prop	0.254130	0.009848	0.101852	-0.026460	
Prior_Conviction_Episodes_Drug	0.106459	0.046625	0.033248	0.040581	
Delinquency_Reports	-0.059886	-0.017291	0.126960	-0.049473	
Program_Attendances	-0.009654	-0.000652	0.061068	-0.025537	
Program_UnexcusedAbsences	-0.092027	-0.028561	0.117546	-0.017636	
Residence_Changes	-0.009472	-0.001424	0.118164	0.010702	
Avg_Days_per_DrugTest	0.067365	-0.024877	-0.036588	0.041916	
DrugTests_THC_Positive	-0.195092	-0.044197	0.240734	-0.034796	
DrugTests_Cocaine_Positive	0.067101	-0.031516	0.032673	-0.011982	
DrugTests_Meth_Positive	-0.002315	0.040276	0.047783	-0.024194	
DrugTests_Other_Positive	-0.002634	0.003403	0.029040	-0.015315	
Percent_Days_Employed	0.062651	0.034806	-0.045773	0.024515	
Jobs_Per_Year	-0.039791	0.044123	0.032629	-0.012196	

25 rows × 25 columns

In [27]:
```
correlation = numeric_features.corr()
f , ax = plt.subplots(figsize = (14,12))
plt.title('Correlation of Numeric Features',y=1,size=16)
sns.heatmap(correlation,square = True)
```

Out[27]: <Axes: title={'center': 'Correlation of Numeric Features'}>





In [28]:
```
pip install dython
```

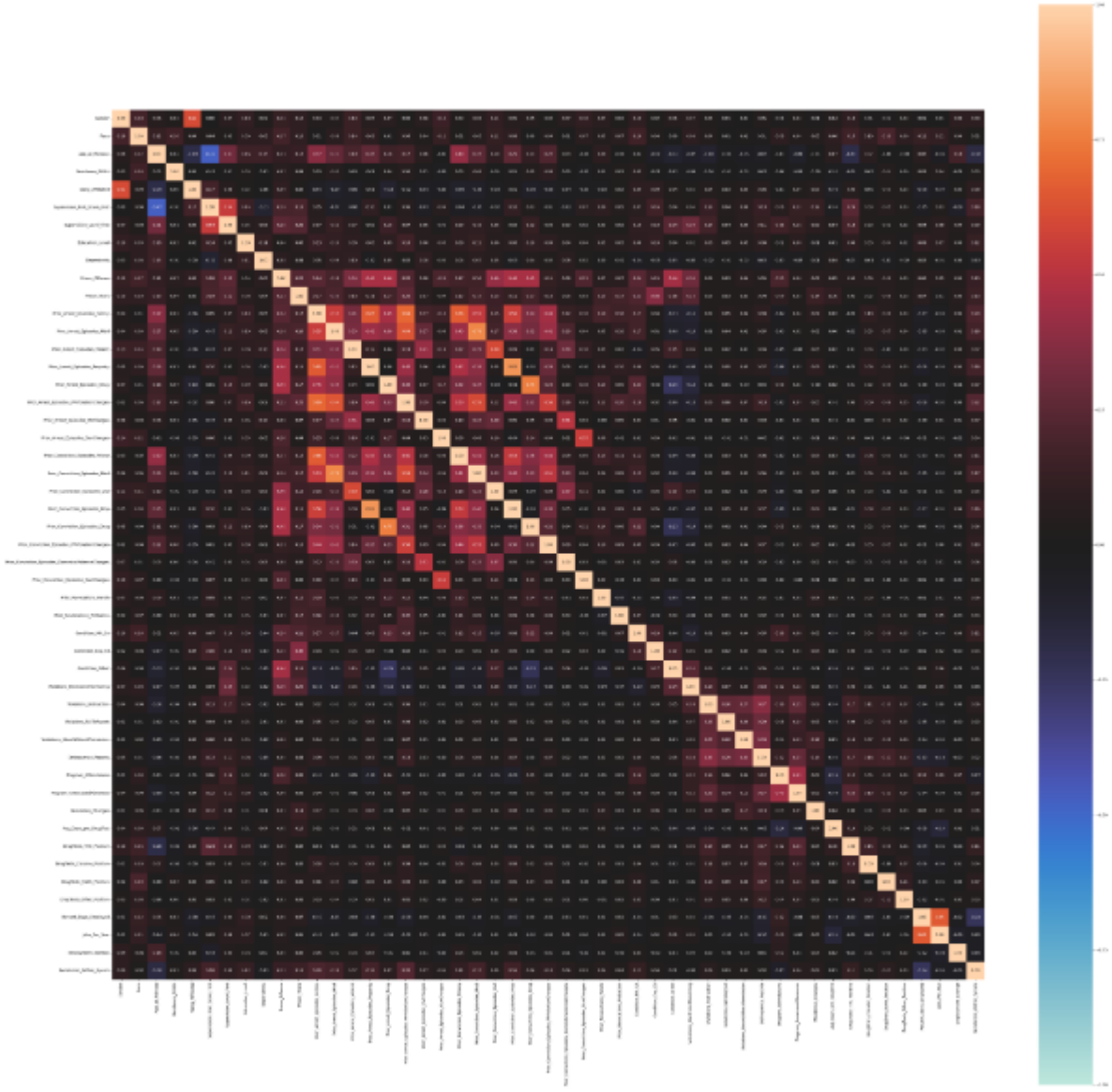
```
Requirement already satisfied: dython in /usr/local/lib/python3.10/dist-packages (0.7.6)
Requirement already satisfied: numpy>=1.23.0 in /usr/local/lib/python3.10/dist-packages (from dython) (1.25.2)
Requirement already satisfied: pandas>=1.4.2 in /usr/local/lib/python3.10/dist-packages (from dython) (2.0.3)
Requirement already satisfied: seaborn>=0.12.0 in /usr/local/lib/python3.10/dist-packages (from dython) (0.13.1)
Requirement already satisfied: scipy>=1.7.1 in /usr/local/lib/python3.10/dist-packages (from dython) (1.11.4)
Requirement already satisfied: matplotlib>=3.6.0 in /usr/local/lib/python3.10/dist-packages (from dython) (3.7.1)
Requirement already satisfied: scikit-learn>=0.24.2 in /usr/local/lib/python3.10/dist-packages (from dython) (1.2.2)
Requirement already satisfied: psutil>=5.9.1 in /usr/local/lib/python3.10/dist-packages (from dython) (5.9.5)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.6.0->dytho
n) (1.2.1)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.6.0->dython)
(0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.6.0->dytho
n) (4.53.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.6.0->dytho
n) (1.4.5)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.6.0->dython)
(24.1)
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.6.0->dython)
(9.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.6.0->dytho
n) (3.1.2)
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.6.0->dy
thon) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.4.2->dython) (2023.
4)
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.4.2->dython) (202
4.1)
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.10/dist-packages (from scikit-learn>=0.24.2->dytho
n) (1.4.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn>=0.24.2-
>dython) (3.5.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.7->matplotlib>
=3.6.0->dython) (1.16.0)
```

In [29]:
```
from dython.nominal import identify_nominal_columns
categorical_features=identify_nominal_columns(NIJ)
categorical_features
```

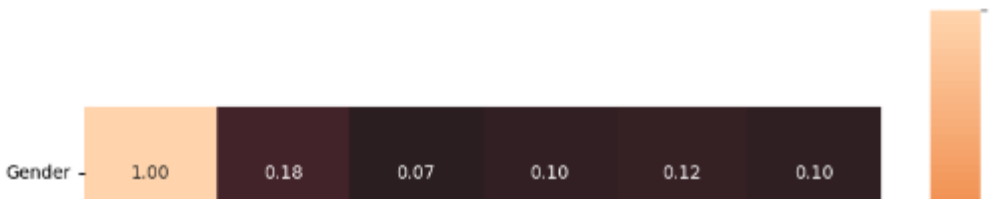
Out[29]:
```
['Gender',
 'Race',
 'Supervision_Level_First',
 'Education_Level',
 'Prison_Offense',
 'Prison_Years']
```

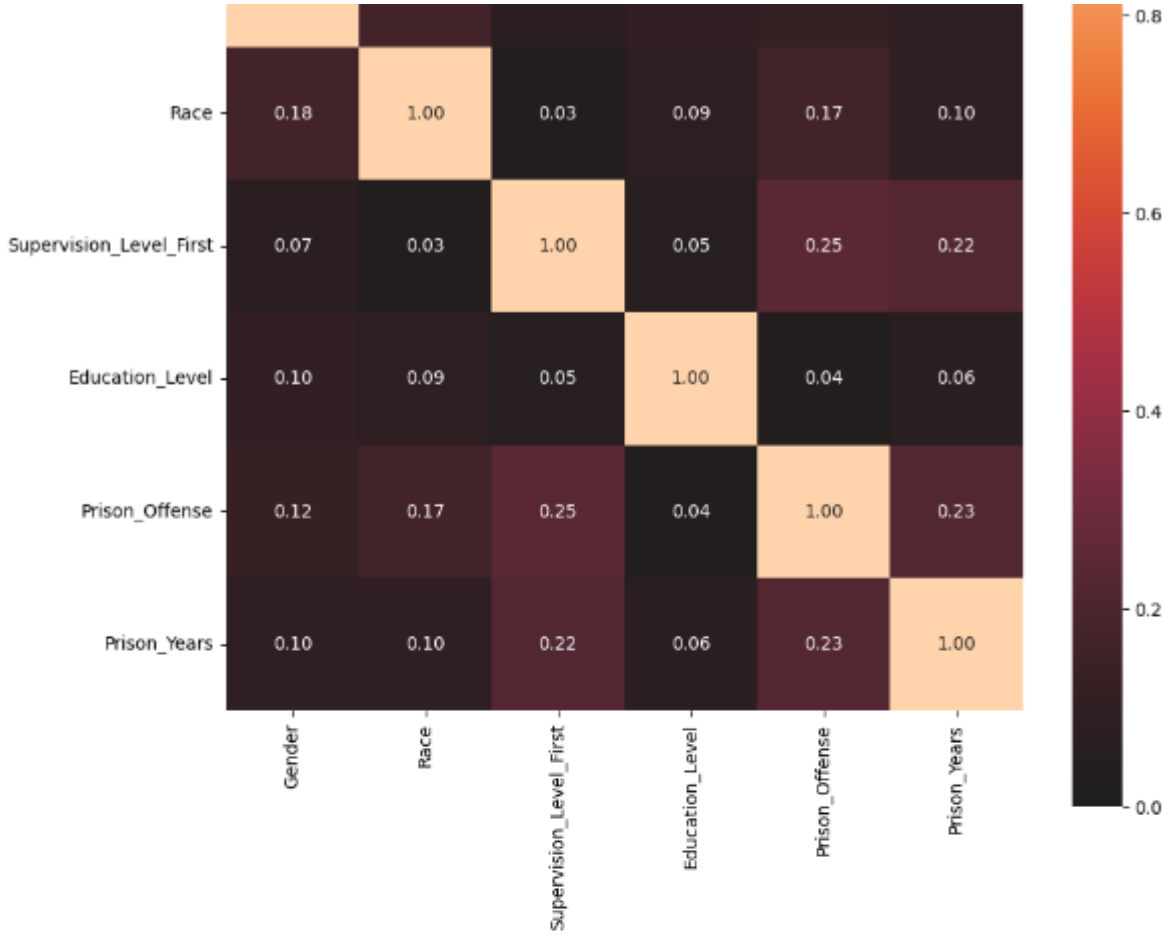
In [30]:
```
from dython.nominal import associations
```

In [31]:
```
complete_correlation=associations(NIJ, figsize=(50,50))
```



In [32]:
```
selected_column= NIJ[categorical_features]
categorical_df = selected_column.copy()

categorical_correlation= associations(categorical_df, filename= 'categorical_correlation.png', figsize=(10,10))
```



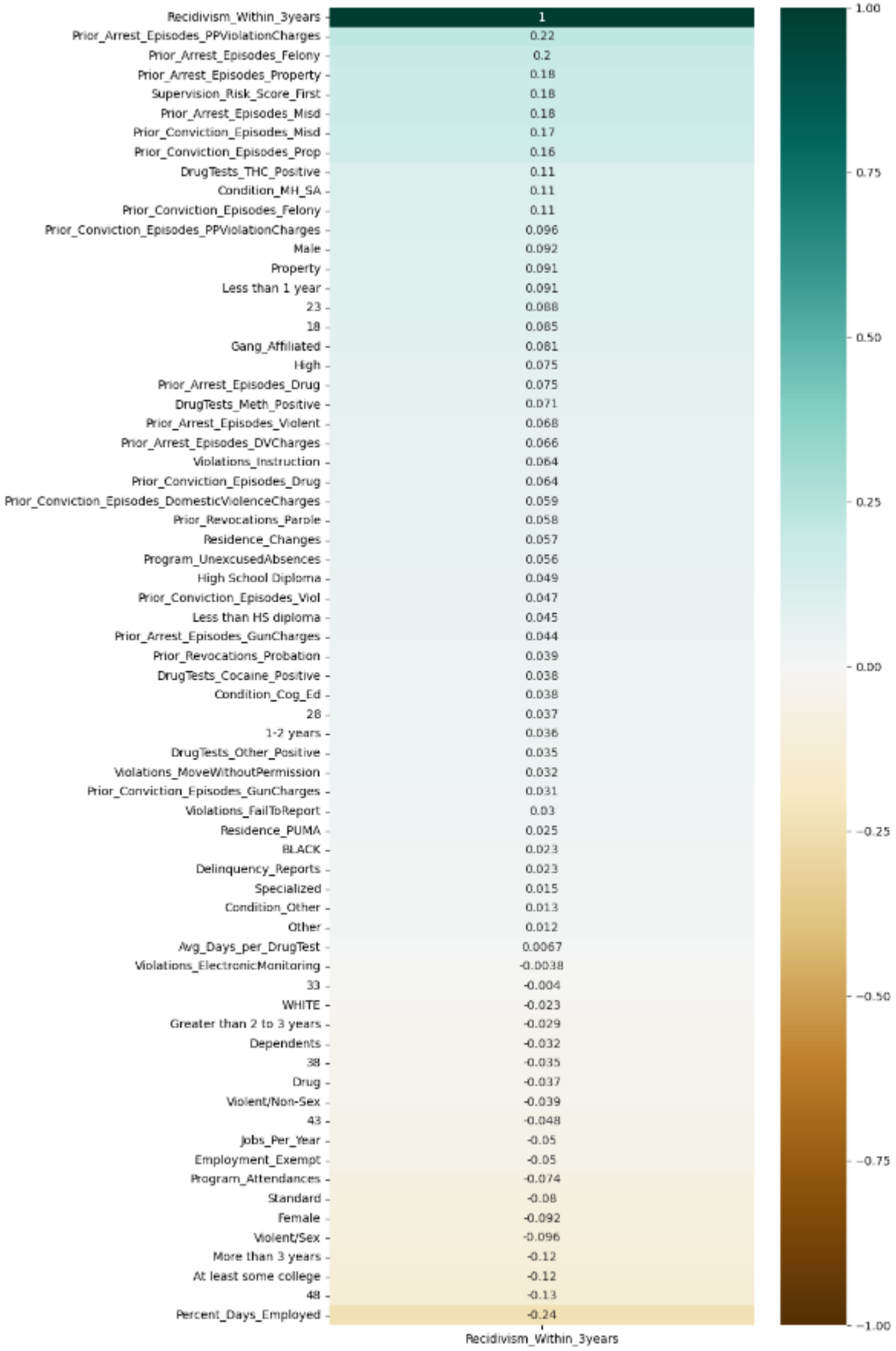


Create a new dataframe for key features with that feature one hot encoded test correlations with that feature.

In [33]:
```
one_hot = pd.get_dummies(NIJ.Gender)
NIJ = pd.concat([NIJ, one_hot], axis=1)

one_hot = pd.get_dummies(NIJ.Race)
NIJ = pd.concat([NIJ, one_hot], axis=1)

one_hot = pd.get_dummies(NIJ.Age_at_Release)
NIJ = pd.concat([NIJ, one_hot], axis=1)

one_hot = pd.get_dummies(NIJ.Supervision_Level_First)
NIJ = pd.concat([NIJ, one_hot], axis=1)

one_hot = pd.get_dummies(NIJ.Education_Level)
NIJ = pd.concat([NIJ, one_hot], axis=1)

one_hot = pd.get_dummies(NIJ.Prison_Offense)
NIJ = pd.concat([NIJ, one_hot], axis=1)

one_hot = pd.get_dummies(NIJ.Prison_Years)
NIJ = pd.concat([NIJ, one_hot], axis=1)
NIJ.drop(columns=['Gender', 'Race', 'Age_at_Release', 'Supervision_Level_First', 'Education_Level', 'Prison_Offense', 'P
```

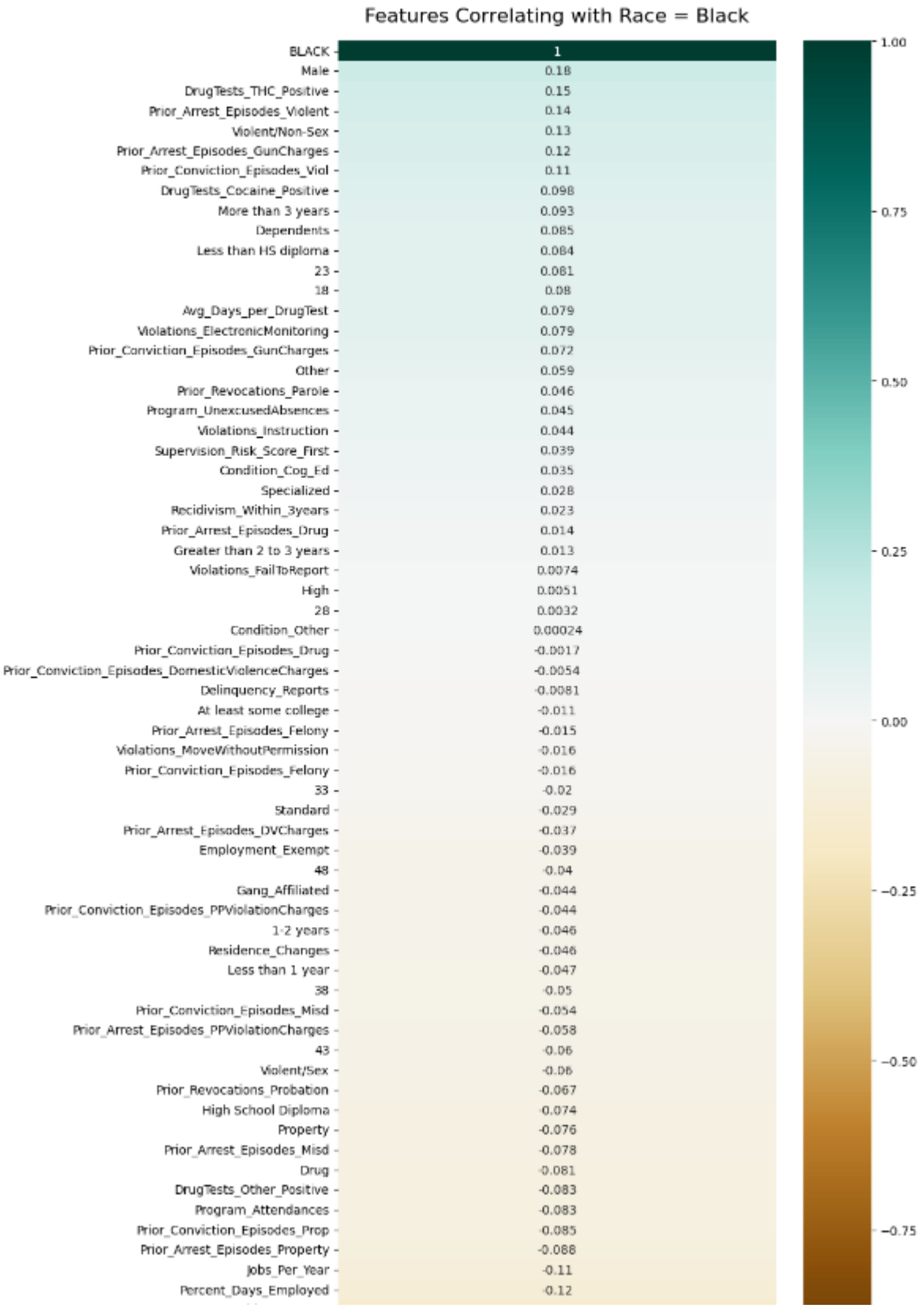
In [34]:
```
plt.figure(figsize=(8, 20))
heatmap = sns.heatmap(NIJ.corr(method ='pearson')[['Recidivism_Within_3years']].sort_values(by='Recidivism_Within_3years
heatmap.set_title('Features Correlating with Recidivism', fontdict={'fontsize':16}, pad=16);
```

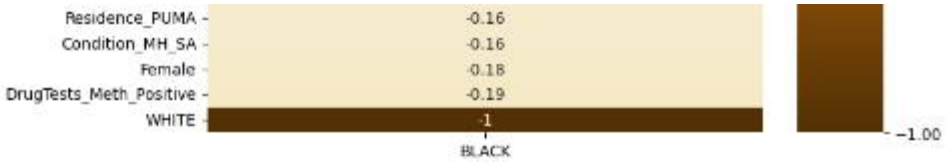
Features Correlating with Recidivism


Recidivism_Within_3years 1
Prior_Arrest_Episodes_PPViolationCharges 0.22
Prior_Arrest_Episodes_Felony 0.2
Prior_Arrest_Episodes_Property 0.18
Supervision_Risk_Score_First 0.18
Prior_Arrest_Episodes_Misd 0.18
Prior_Conviction_Episodes_Misd 0.17
Prior_Conviction_Episodes_Prop 0.16
DrugTests_THC_Positive 0.11
Condition_MH_SA 0.11
Prior_Conviction_Episodes_Felony 0.11
Prior_Conviction_Episodes_PPViolationCharges 0.096
Male 0.092
Property 0.091
Less than 1 year 0.091
23 0.088
18 0.085
Gang_Affiliated 0.081
High 0.075
Prior_Arrest_Episodes_Drug 0.075
DrugTests_Meth_Positive 0.071
Prior_Arrest_Episodes_Violent 0.068
Prior_Arrest_Episodes_DVCharges 0.066
Violations_Instruction 0.064
Prior_Conviction_Episodes_Drug 0.064
Prior_Conviction_Episodes_DomesticViolenceCharges 0.059
Prior_Revocations_Parole 0.058
Residence_Changes 0.057
Program_UnexcusedAbsences 0.056
High School Diploma 0.049
Prior_Conviction_Episodes_Viol 0.047
Less than HS diploma 0.045
Prior_Arrest_Episodes_GunCharges 0.044
Prior_Revocations_Probation 0.039
DrugTests_Cocaine_Positive 0.038
Condition_Cog_Ed 0.038
28 0.037
1-2 years 0.036
DrugTests_Other_Positive 0.035
Violations_MoveWithoutPermission 0.032
Prior_Conviction_Episodes_GunCharges 0.031
Violations_FailToReport 0.03
Residence_PUMA 0.025
BLACK 0.023
Delinquency_Reports 0.023
Specialized 0.015
Condition_Other 0.013
Other 0.012
Avg_Days_per_DrugTest 0.0067
Violations_ElectronicMonitoring -0.0038
33 -0.004
WHITE -0.023
Greater than 2 to 3 years -0.029
Dependents -0.032
38 -0.035
Drug -0.037
Violent/Non-Sex -0.039
43 -0.048
Jobs_Per_Year -0.05
Employment_Exempt -0.05
Program_Attendances -0.074
Standard -0.08
Female -0.092
Violent/Sex -0.096
More than 3 years -0.12
At least some college -0.12
48 -0.13
Percent_Days_Employed -0.24
Recidivism_Within_3years
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

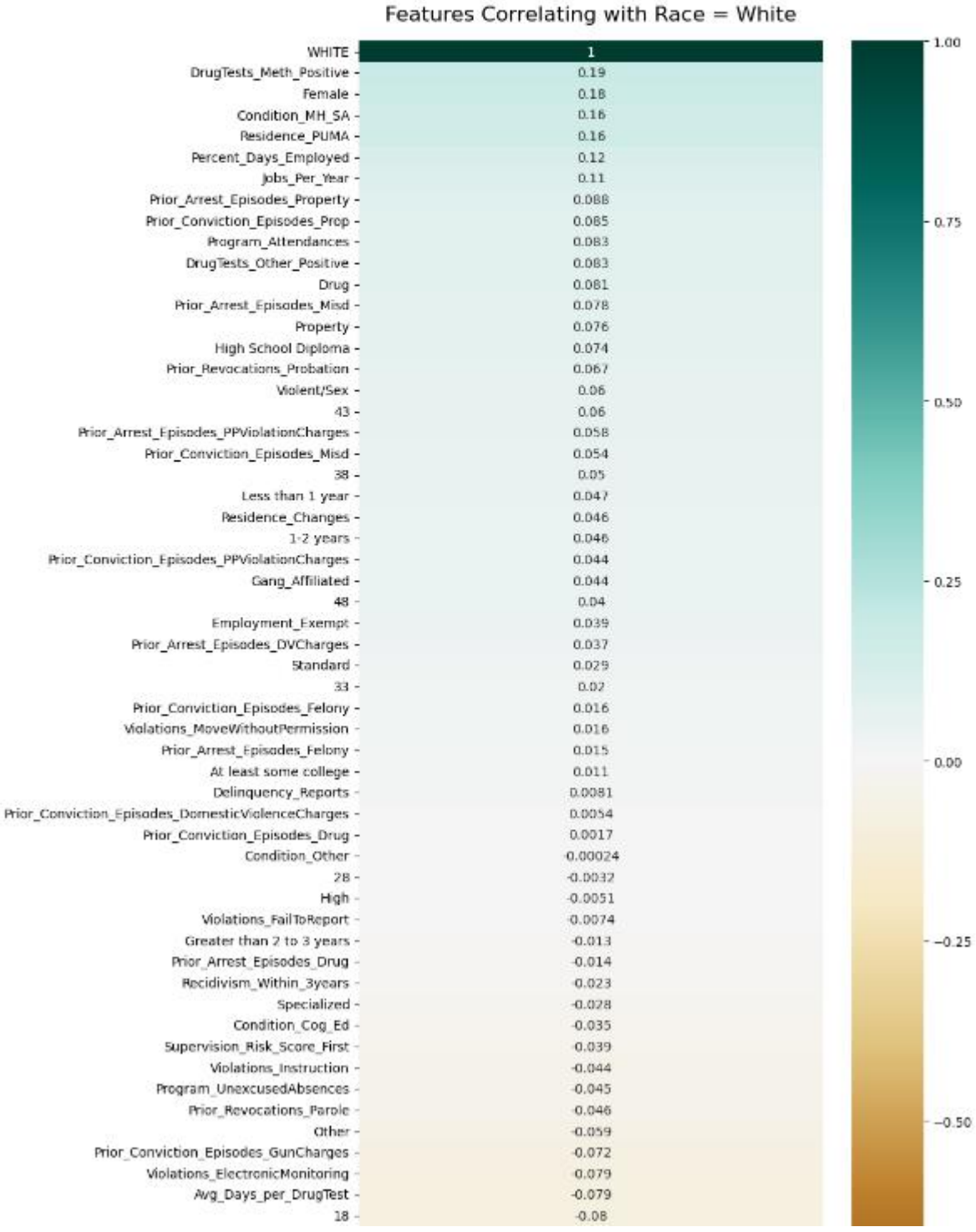
In [35]:

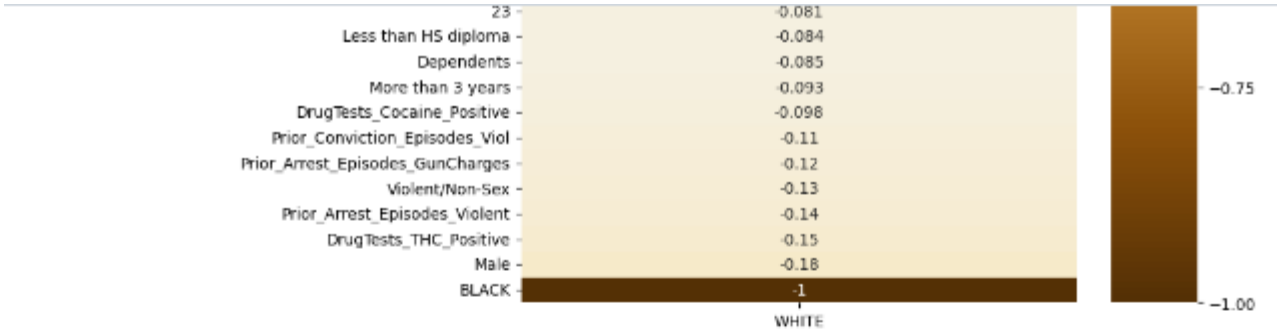
```
plt.figure(figsize=(8, 20))
heatmap = sns.heatmap(NIJ.corr(method ='pearson')[['BLACK']].sort_values(by='BLACK', ascending=False), vmin=-1, vmax=1,
heatmap.set_title('Features Correlating with Race = Black', fontdict={'fontsize':16}, pad=16);
```





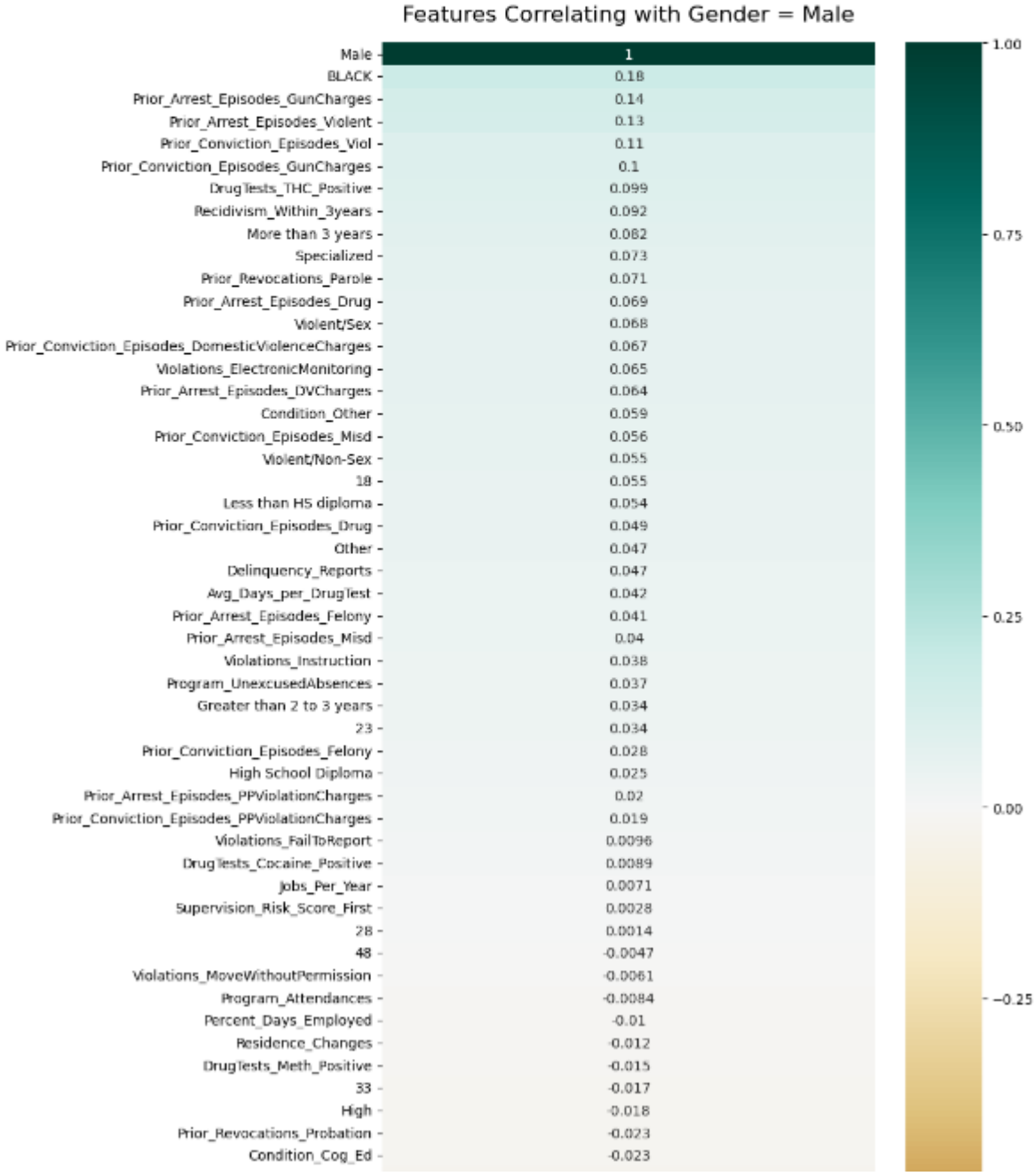
In [36]:
```
plt.figure(figsize=(8, 20))
heatmap = sns.heatmap(NIJ.corr(method ='pearson')[['WHITE']].sort_values(by='WHITE', ascending=False), vmin=-1, vmax=1,
heatmap.set_title('Features Correlating with Race = White', fontdict={'fontsize':16}, pad=16);
```

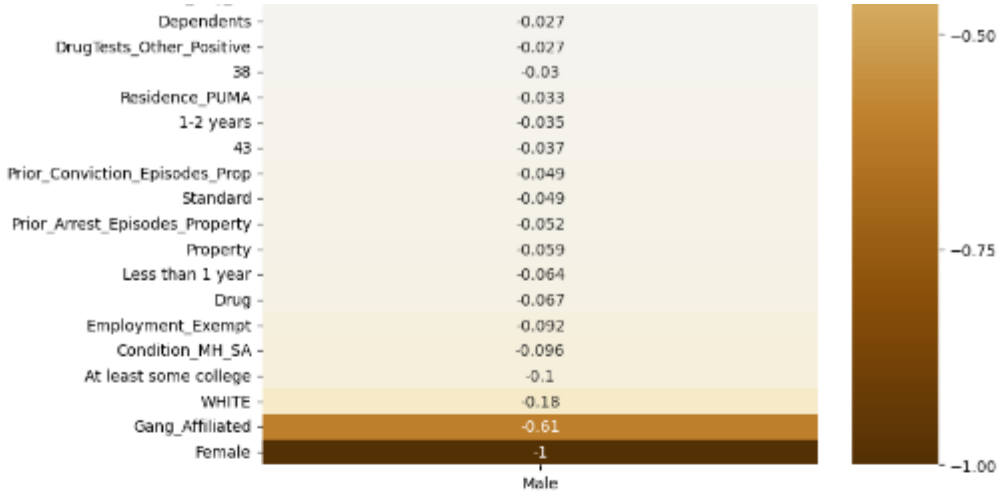




In [37]:

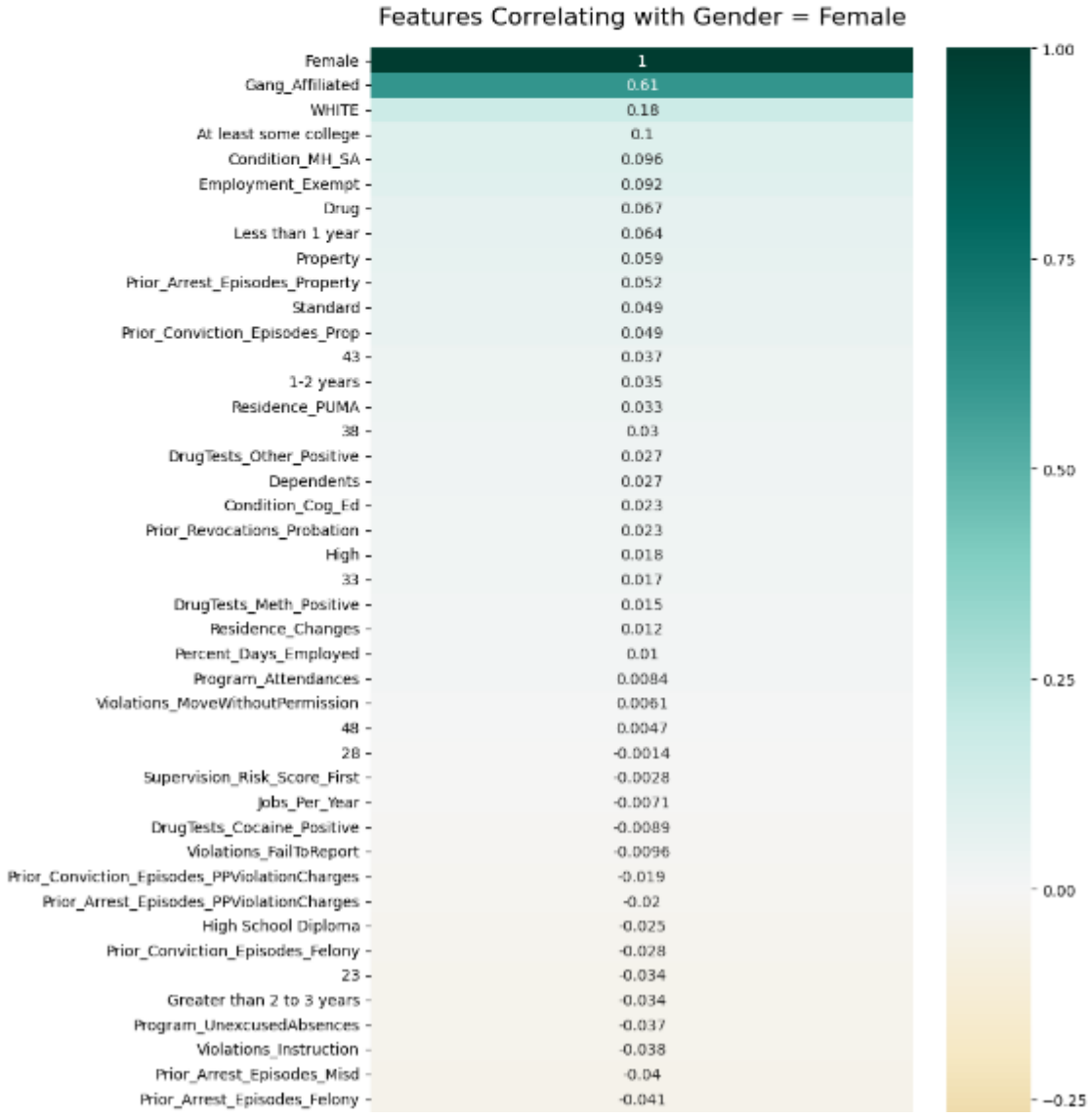
```
plt.figure(figsize=(8, 20))
heatmap = sns.heatmap(NIJ.corr(method ='pearson')[['Male']].sort_values(by='Male', ascending=False), vmin=-1, vmax=1, an
heatmap.set_title('Features Correlating with Gender = Male', fontdict={'fontsize':16}, pad=16);
```

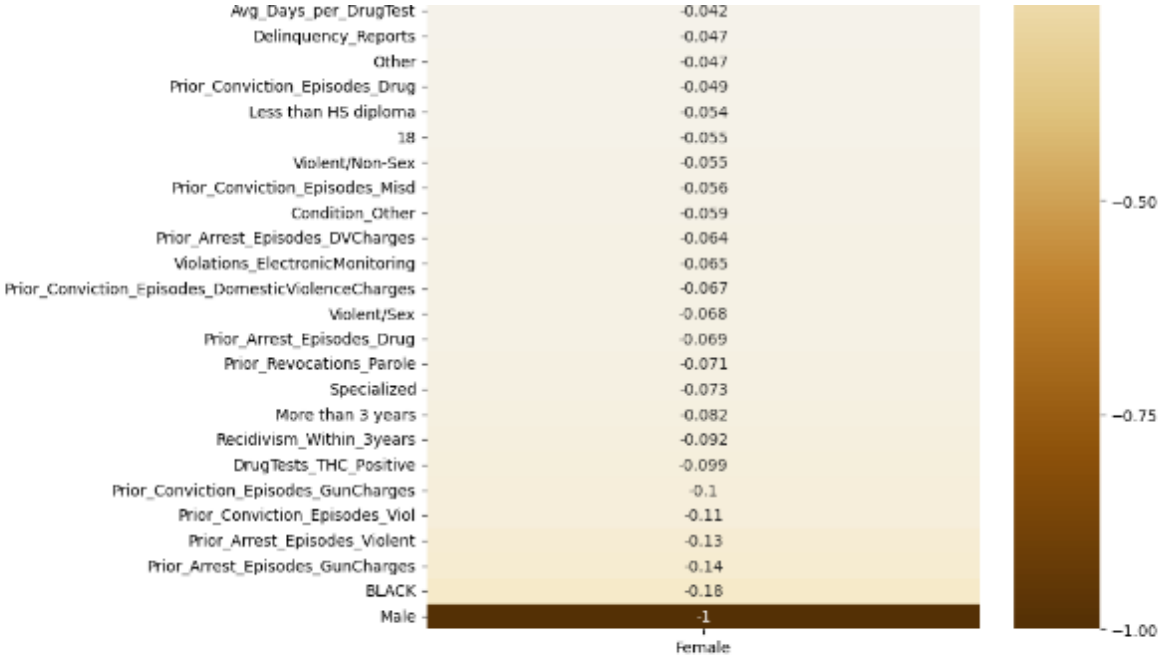




In [38]:

```
plt.figure(figsize=(8, 20))
heatmap = sns.heatmap(NIJ.corr(method ='pearson')[['Female']].sort_values(by='Female', ascending=False), vmin=-1, vmax=1
heatmap.set_title('Features Correlating with Gender = Female', fontdict={'fontsize':16}, pad=16);
```





Check again with dataset with only male offenders and female offenders

In [39]:
```
Male = NIJ.copy()
Male.drop(Male[Male.Male == False].index, inplace=True)
Male.drop(columns=['Male', 'Female'], inplace=True)
Male
```

Out[39]:

	Residence_PUMA	Gang_Affiliated	Supervision_Risk_Score_First	Dependents	Prior_Arrest_Episodes_Felony	Prior_Arrest_Episode
0	16	False	3.0	3	6	
1	16	False	6.0	1	7	
2	24	False	7.0	3	6	
3	16	False	7.0	1	8	
4	16	False	4.0	3	4	
...	...	...	...	...	...	
25829	5	False	5.0	3	1	
25830	9	False	5.0	1	2	
25831	25	False	5.0	3	0	
25832	15	False	5.0	3	0	
25834	12	False	5.0	3	6	

22668 rows × 66 columns

In [40]:
```
Female = NIJ.copy()
Female.drop(Female[Female.Female == False].index, inplace=True)
Female.drop(columns=['Male', 'Female'], inplace=True)
Female
```

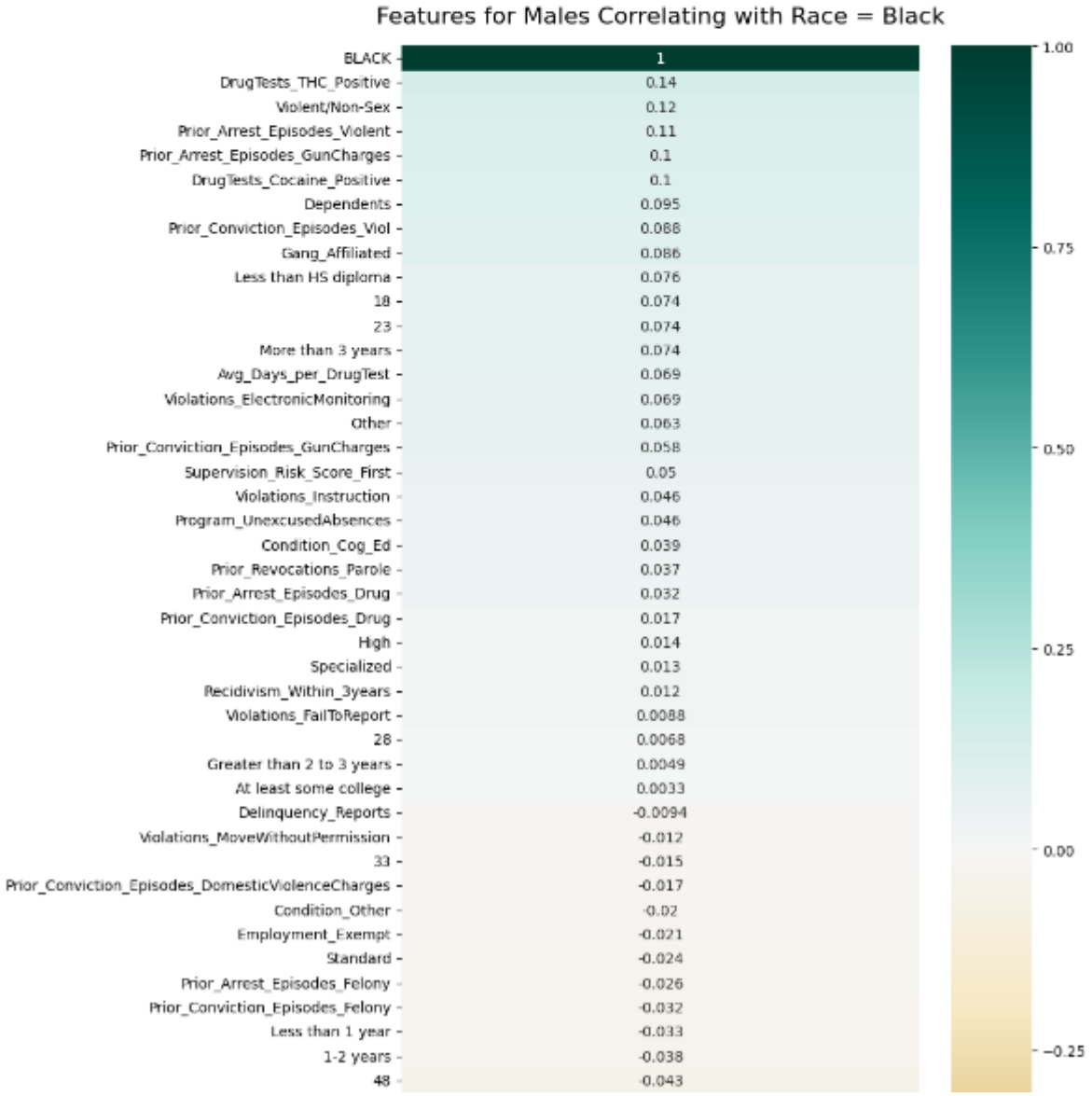
Out[40]:

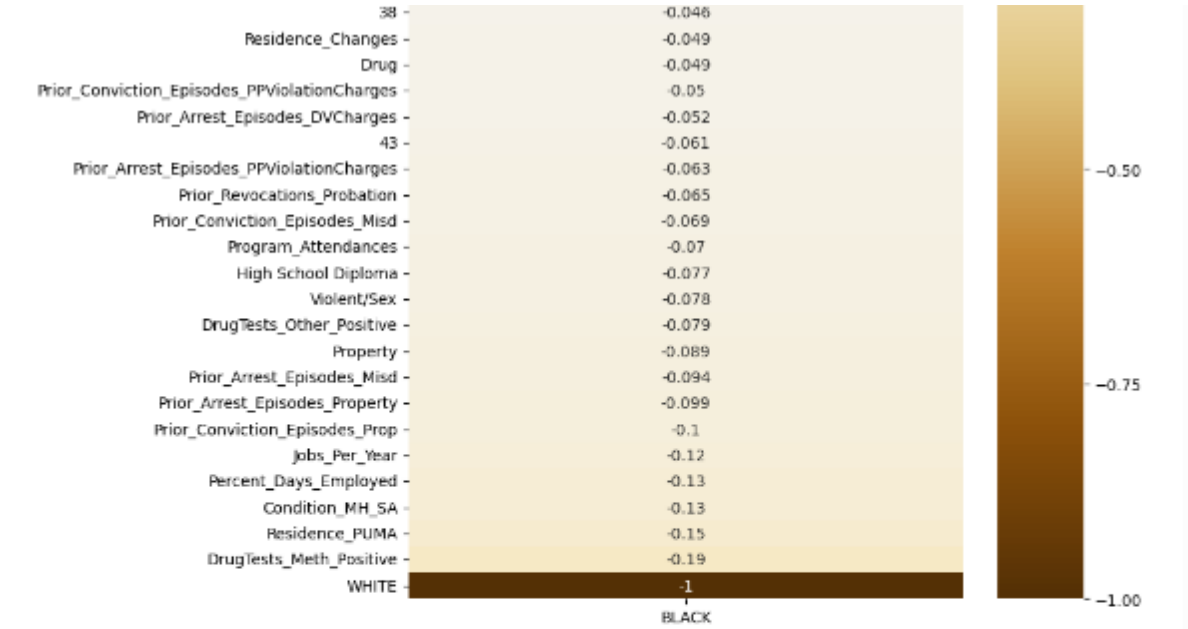
	Residence_PUMA	Gang_Affiliated	Supervision_Risk_Score_First	Dependents	Prior_Arrest_Episodes_Felony	Prior_Arrest_Episode
8	5	True	7.0	0	10	

22	5	True	4.0	0	2
23	16	True	5.0	0	10
31	16	True	9.0	0	10
33	12	True	10.0	3	10
...	...	...	...	...	...
25779	22	True	8.0	3	1
25783	22	True	9.0	1	3
25790	20	True	7.0	3	1
25803	15	True	5.0	0	5
25833	15	True	5.0	3	0

3167 rows × 66 columns

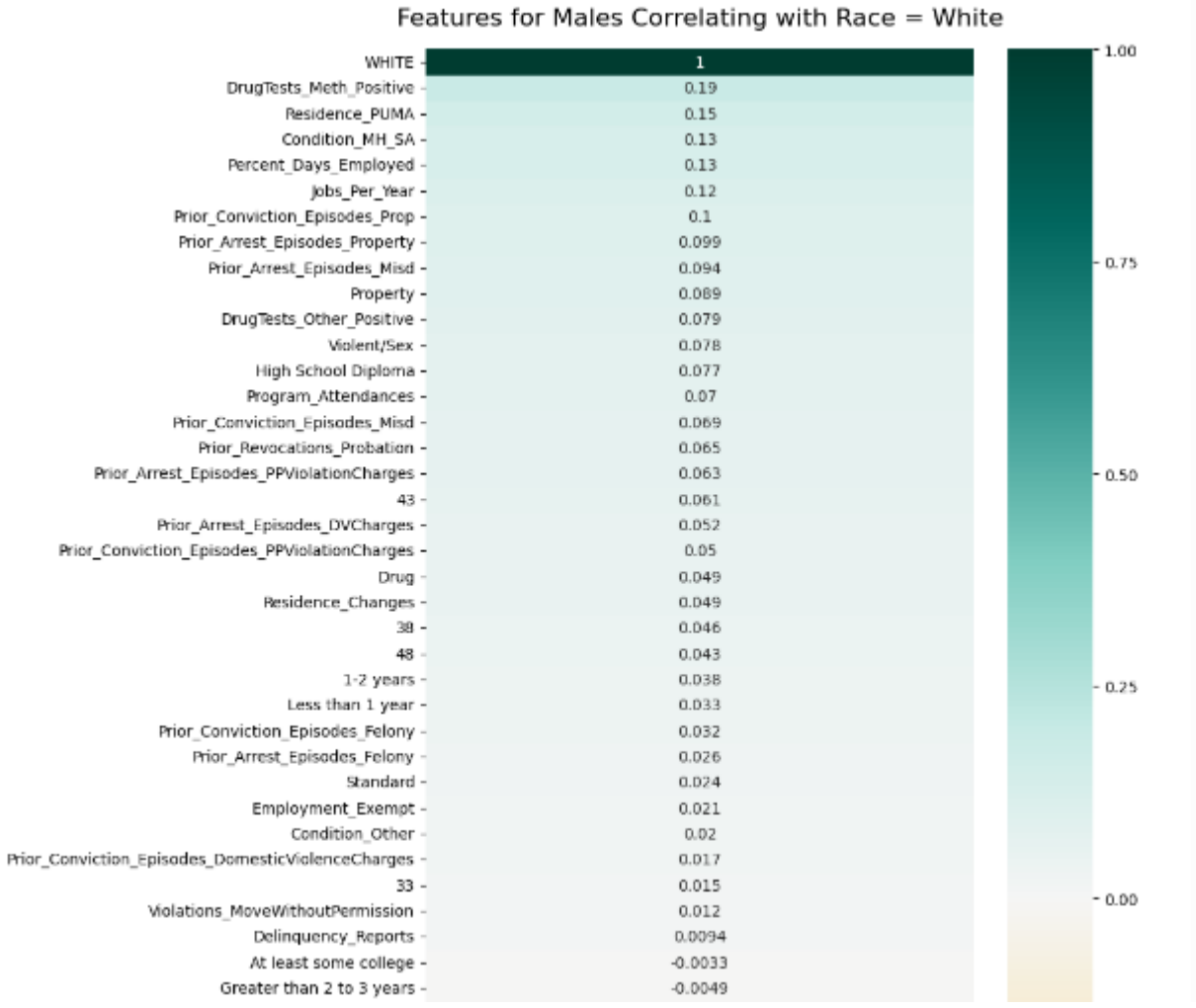
In [41]:
```
plt.figure(figsize=(8, 20))
heatmap = sns.heatmap(Male.corr(method ='pearson')[['BLACK']].sort_values(by='BLACK', ascending=False), vmin=-1, vmax=1,
heatmap.set_title('Features for Males Correlating with Race = Black', fontdict={'fontsize':16}, pad=16);
```

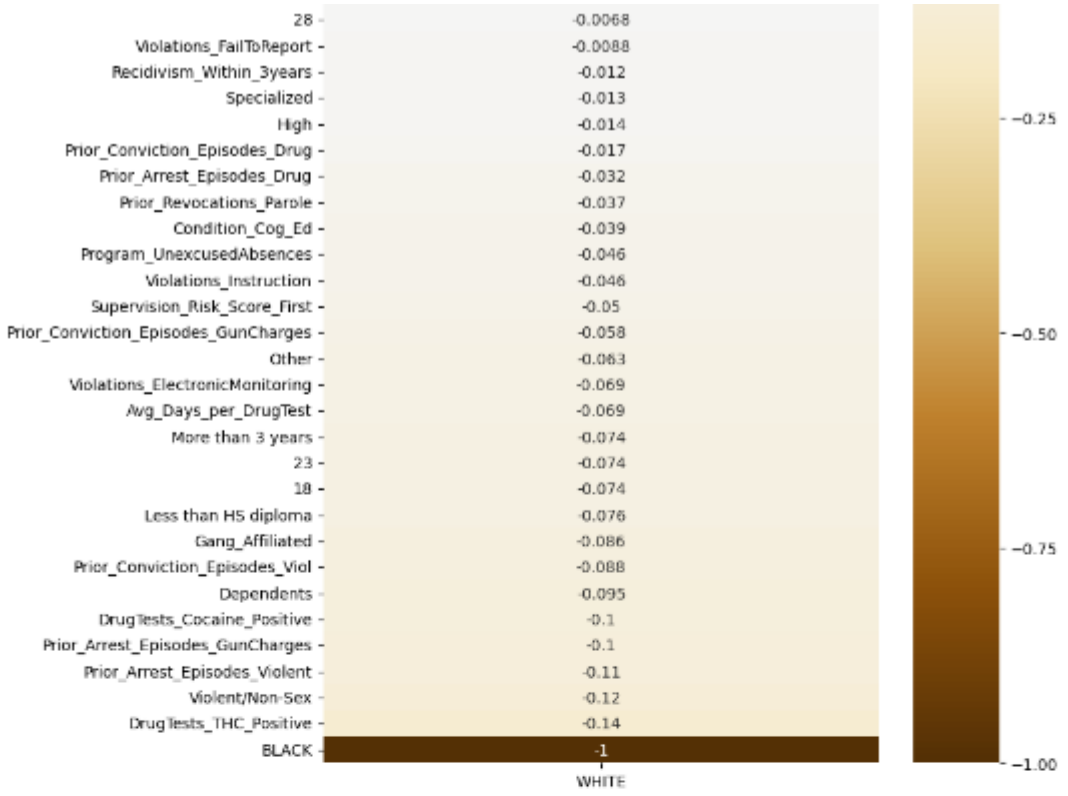




In [42]:

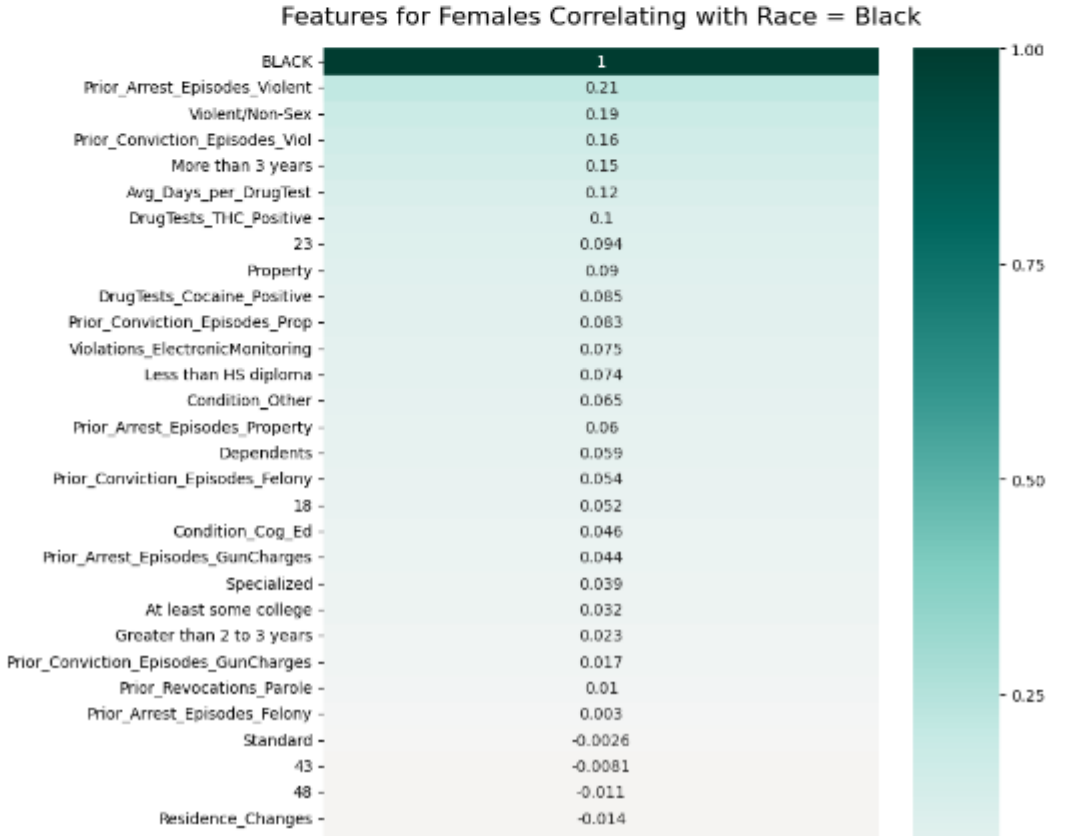
```
plt.figure(figsize=(8, 20))
heatmap = sns.heatmap(Male.corr(method ='pearson')[['WHITE']].sort_values(by='WHITE', ascending=False), vmin=-1, vmax=1,
heatmap.set_title('Features for Males Correlating with Race = White', fontdict={'fontsize':16}, pad=16);
```





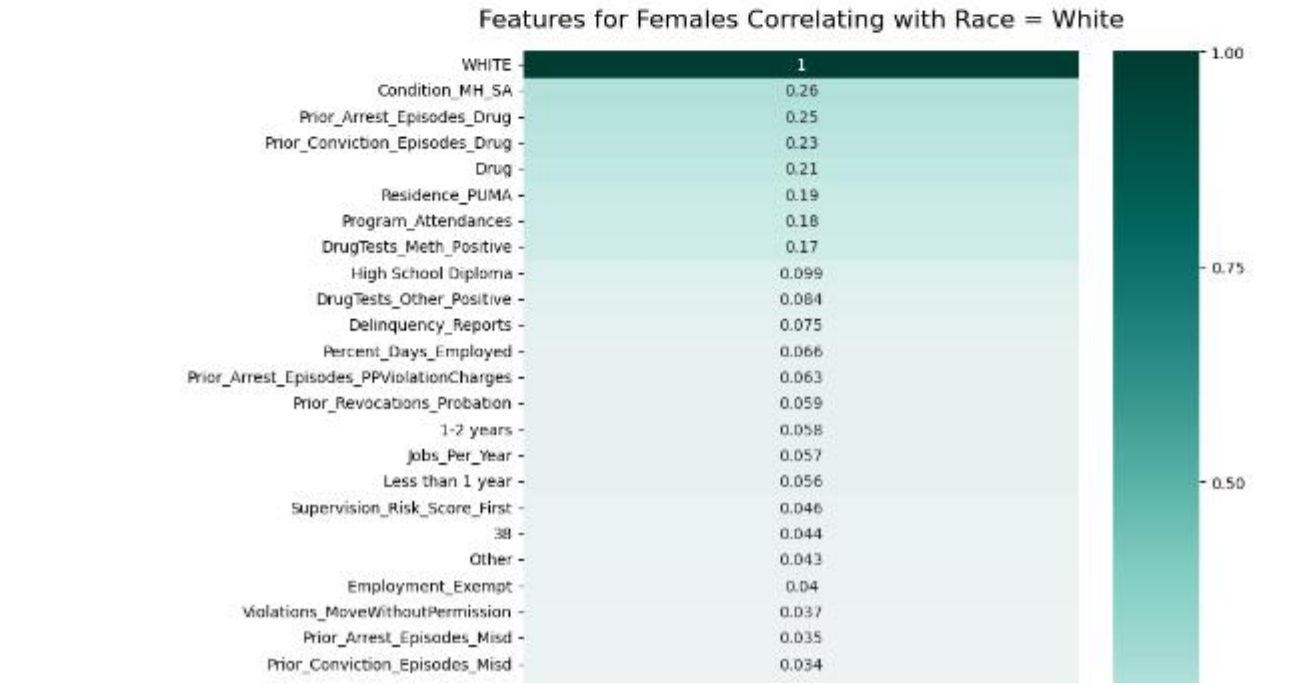
In [43]:

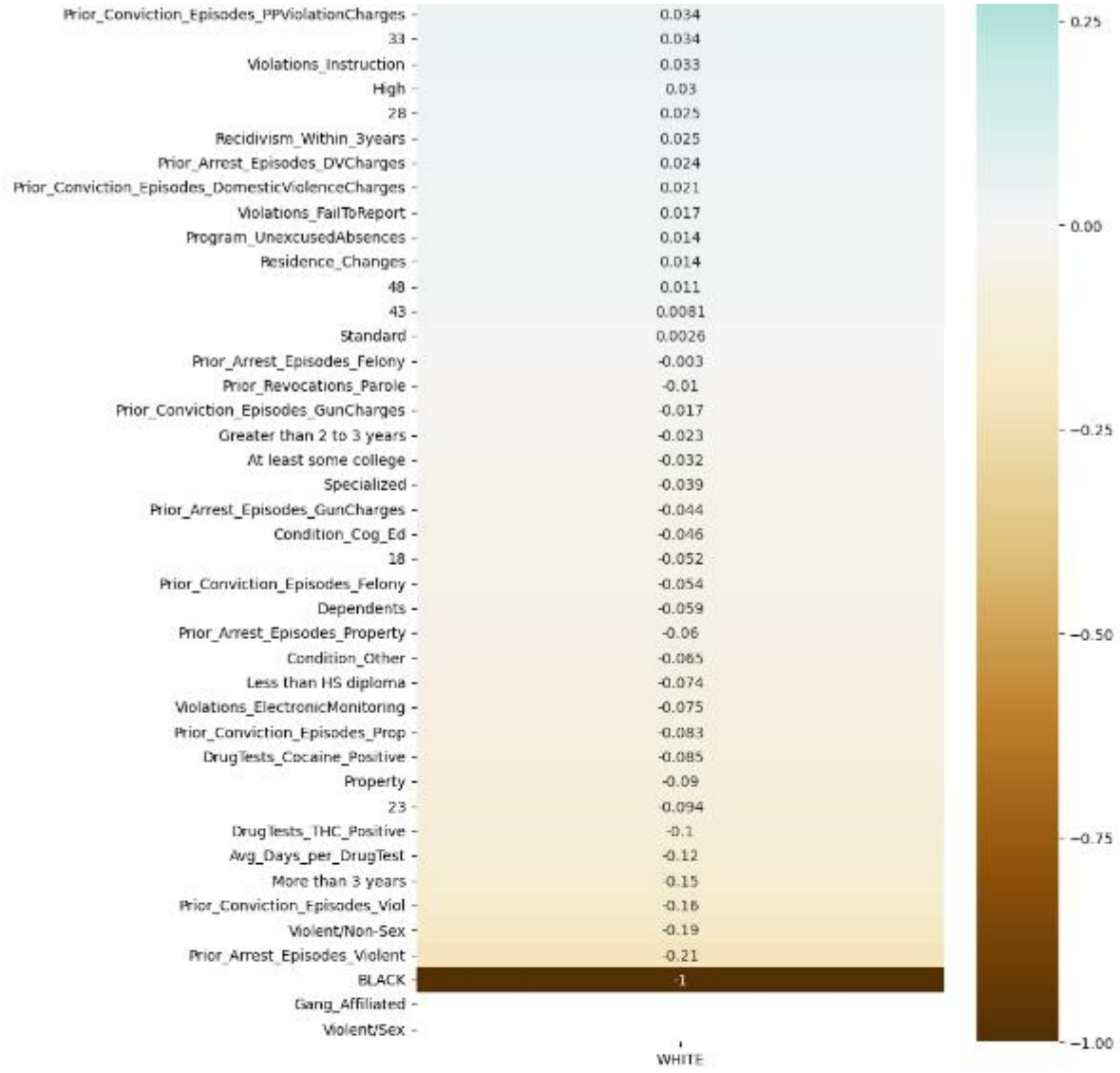
```
plt.figure(figsize=(8, 20))
heatmap = sns.heatmap(Female.corr(method ='pearson')[['BLACK']].sort_values(by='BLACK', ascending=False), vmin=-1, vmax=
heatmap.set_title('Features for Females Correlating with Race = Black', fontdict={'fontsize':16}, pad=16);
```



In [44]:

```
plt.figure(figsize=(8, 20))
heatmap = sns.heatmap(Female.corr(method ='pearson')[['WHITE']].sort_values(by='WHITE', ascending=False), vmin=-1, vmax=
heatmap.set_title('Features for Females Correlating with Race = White', fontdict={'fontsize':16}, pad=16);
```





In [45]:
```
BlackMale = Male.copy()
BlackMale.drop(BlackMale[BlackMale.WHITE == True].index, inplace=True)
BlackMale.drop(columns=['BLACK', 'WHITE'], inplace=True)
BlackMale
```

Out[45]:

	Residence_PUMA	Gang_Affiliated	Supervision_Risk_Score_First	Dependents	Prior_Arrest_Episodes_Felony	Prior_Arrest_Episode
0	16	False	3.0	3	6	
1	16	False	6.0	1	7	
2	24	False	7.0	3	6	
6	18	False	2.0	2	10	
7	16	False	5.0	3	6	
...	...	...	...	...	...	
25825	6	False	5.0	3	0	
25826	7	False	5.0	3	0	
25827	6	False	5.0	3	0	
25830	9	False	5.0	1	2	

25832	15	False	5.0	3	0

13765 rows × 64 columns

In [46]:
```
WhiteMale = Male.copy()
WhiteMale.drop(WhiteMale[WhiteMale.BLACK == True].index, inplace=True)
WhiteMale.drop(columns=['BLACK', 'WHITE'], inplace=True)
WhiteMale
```

Out[46]:

	Residence_PUMA	Gang_Affiliated	Supervision_Risk_Score_First	Dependents	Prior_Arrest_Episodes_Felony	Prior_Arrest_Episode
3	16	False	7.0	1	8	
4	16	False	4.0	3	4	
5	17	False	5.0	0	4	
10	5	False	3.0	1	3	
12	18	False	3.0	1	8	
...	...	...	...	...	...	
25824	18	False	5.0	3	0	
25828	14	False	5.0	3	0	
25829	5	False	5.0	3	1	
25831	25	False	5.0	3	0	
25834	12	False	5.0	3	6	

8903 rows × 64 columns

In [47]:
```
BlackFemale = Female.copy()
BlackFemale.drop(BlackFemale[BlackFemale.WHITE == True].index, inplace=True)
BlackFemale.drop(columns=['BLACK', 'WHITE'], inplace=True)
BlackFemale
```

Out[47]:

	Residence_PUMA	Gang_Affiliated	Supervision_Risk_Score_First	Dependents	Prior_Arrest_Episodes_Felony	Prior_Arrest_Episode
8	5	True	7.0	0	10	
39	2	True	2.0	1	2	
84	3	True	1.0	3	10	
87	3	True	7.0	2	10	
100	12	True	5.0	0	4	
...	...	...	...	...	...	
25684	3	True	4.0	3	1	
25698	1	True	1.0	3	1	
25710	4	True	3.0	3	10	
25734	10	True	7.0	0	1	
25790	20	True	7.0	3	1	

1082 rows × 64 columns

In [48]:
```
WhiteFemale = Female.copy()
WhiteFemale.drop(WhiteFemale[WhiteFemale.BLACK == True].index, inplace=True)
WhiteFemale.drop(columns=['BLACK', 'WHITE'], inplace=True)
```

```
WhiteFemale
```

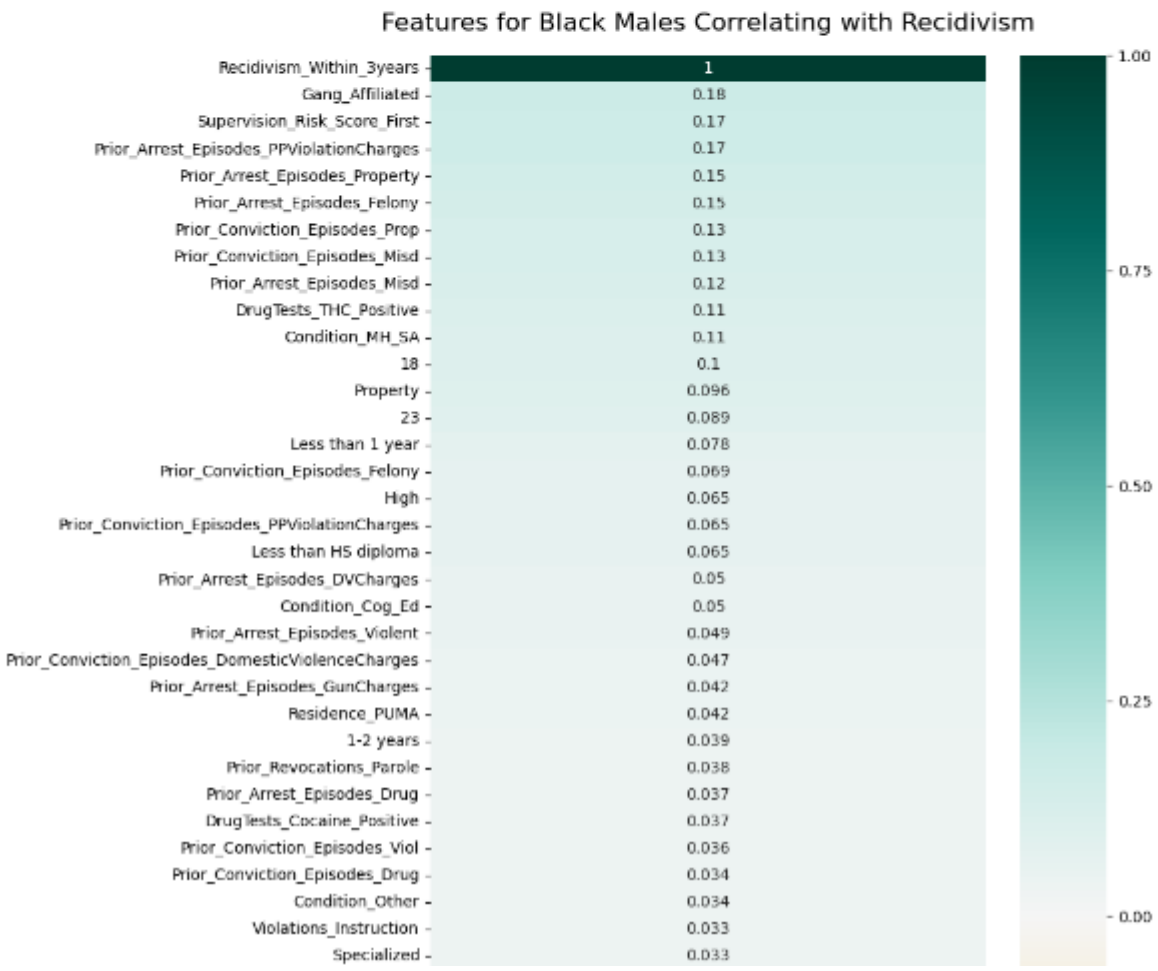
Out[48]:

	Residence_PUMA	Gang_Affiliated	Supervision_Risk_Score_First	Dependents	Prior_Arrest_Episodes_Felony	Prior_Arrest_Episode
22	5	True	4.0	0	2	
23	16	True	5.0	0	10	
31	16	True	9.0	0	10	
33	12	True	10.0	3	10	
42	16	True	2.0	0	10	
...	...	...	...	...	...	
25775	22	True	7.0	0	2	
25779	22	True	8.0	3	1	
25783	22	True	9.0	1	3	
25803	15	True	5.0	0	5	
25833	15	True	5.0	3	0	

2085 rows × 64 columns

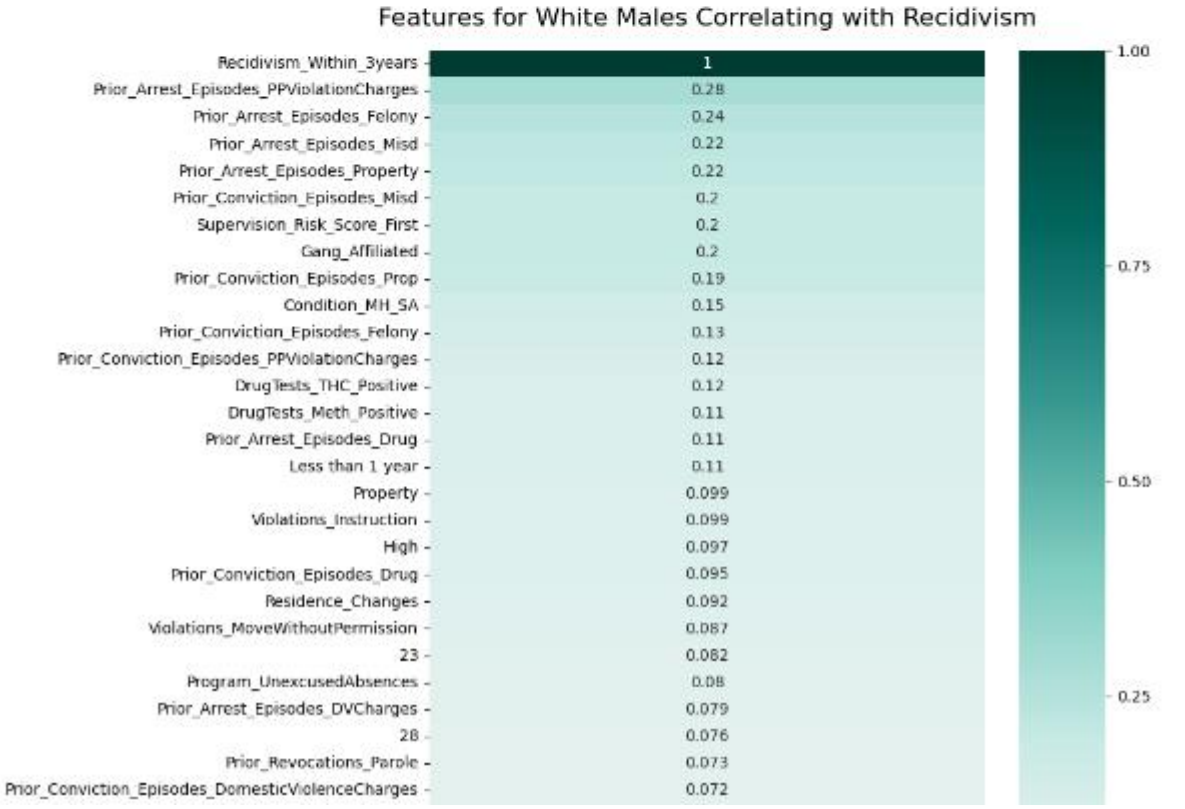
In [49]:

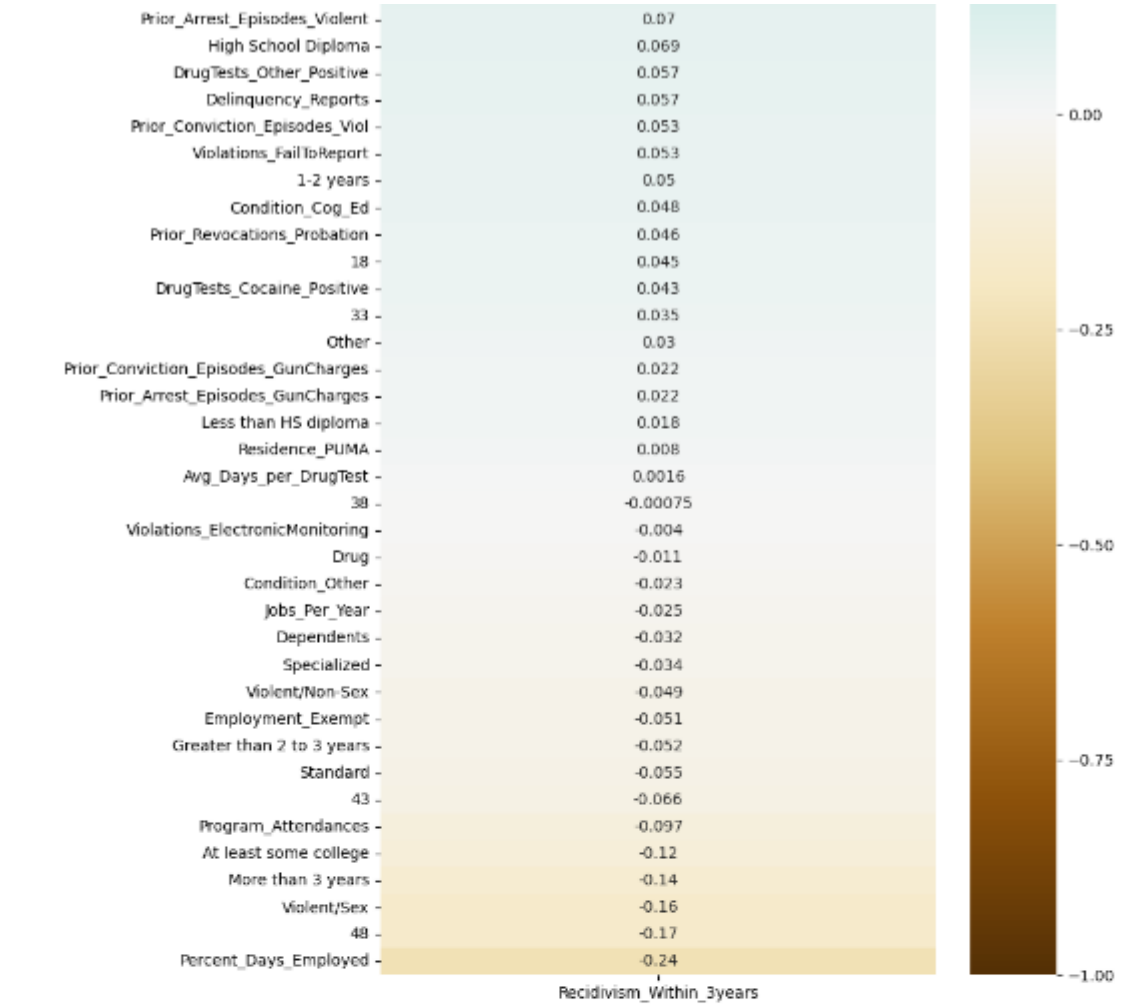
```
plt.figure(figsize=(8, 20))
heatmap = sns.heatmap(BlackMale.corr(method ='pearson')[['Recidivism_Within_3years']].sort_values(by='Recidivism_Within_
heatmap.set_title('Features for Black Males Correlating with Recidivism', fontdict={'fontsize':16}, pad=16);
```



In [50]:

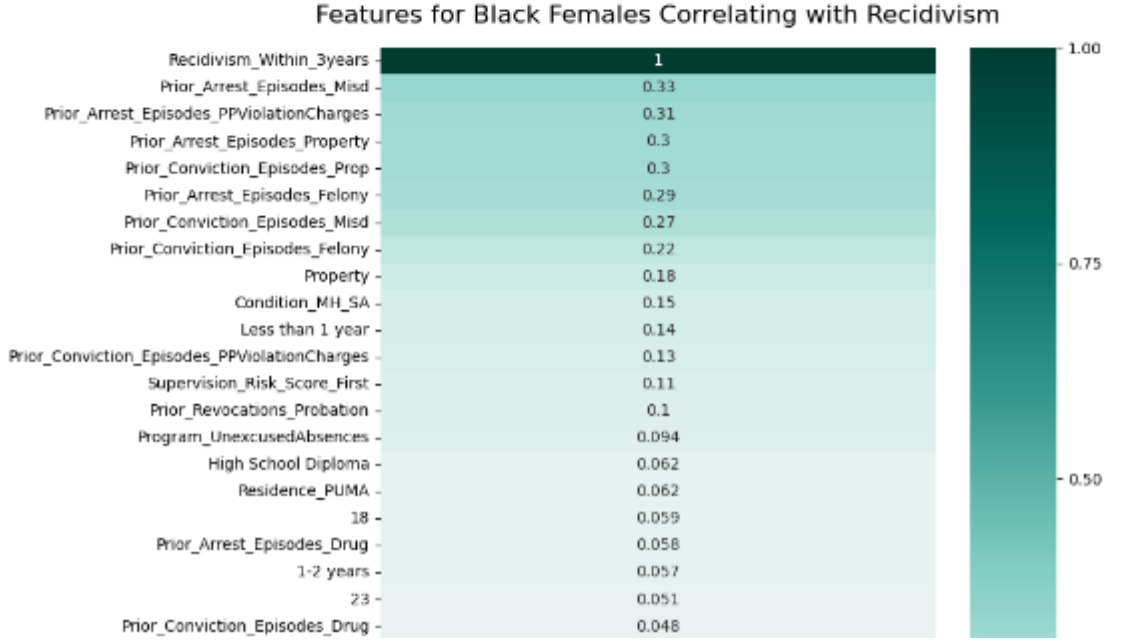
```
plt.figure(figsize=(8, 20))
heatmap = sns.heatmap(WhiteMale.corr(method ='pearson')[['Recidivism_Within_3years']].sort_values(by='Recidivism_Within_
heatmap.set_title('Features for White Males Correlating with Recidivism', fontdict={'fontsize':16}, pad=16);
```

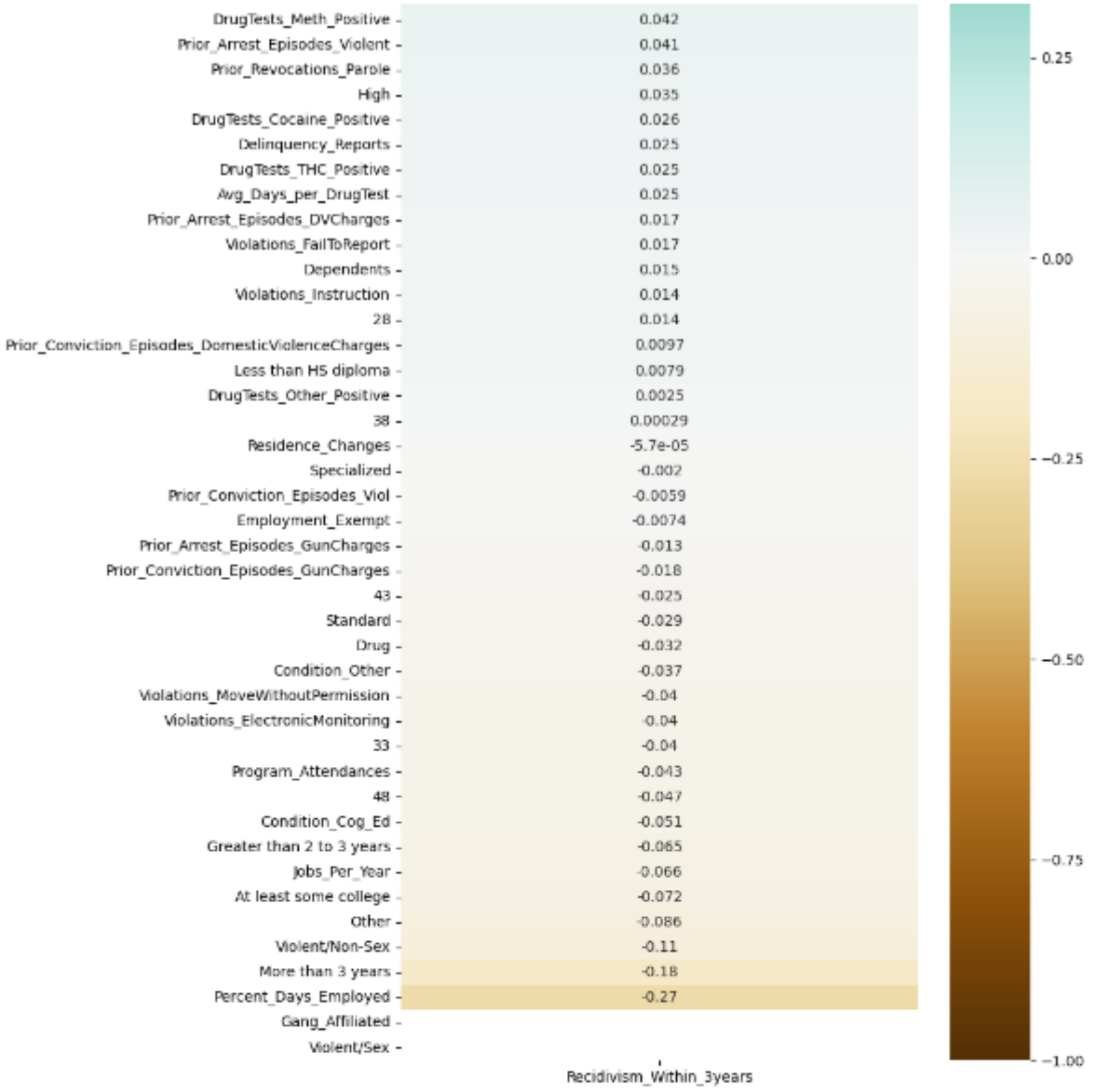




In [51]:

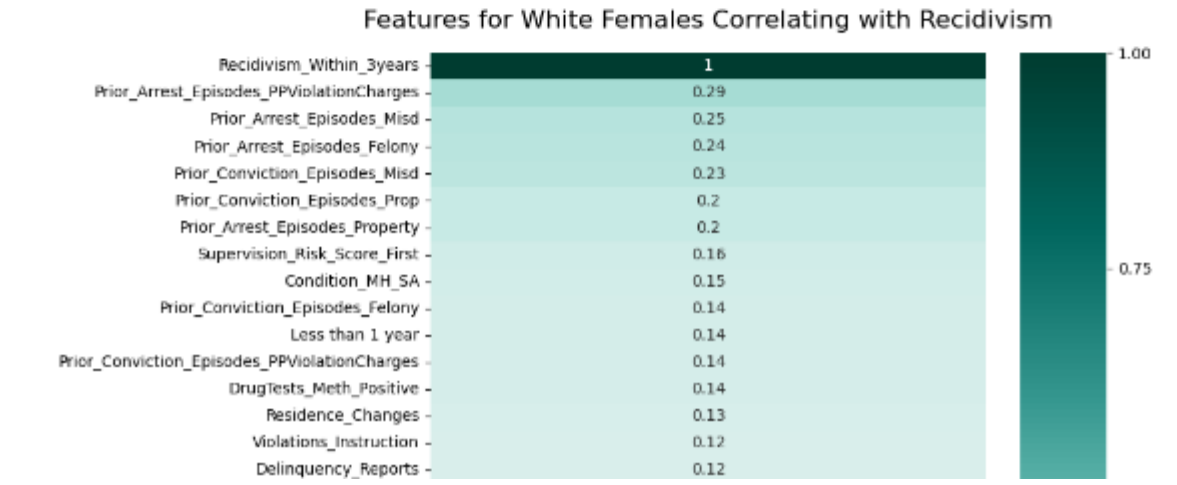
```
plt.figure(figsize=(8, 20))
heatmap = sns.heatmap(BlackFemale.corr(method ='pearson')[['Recidivism_Within_3years']].sort_values(by='Recidivism_Withi
heatmap.set_title('Features for Black Females Correlating with Recidivism', fontdict={'fontsize':16}, pad=16);
```

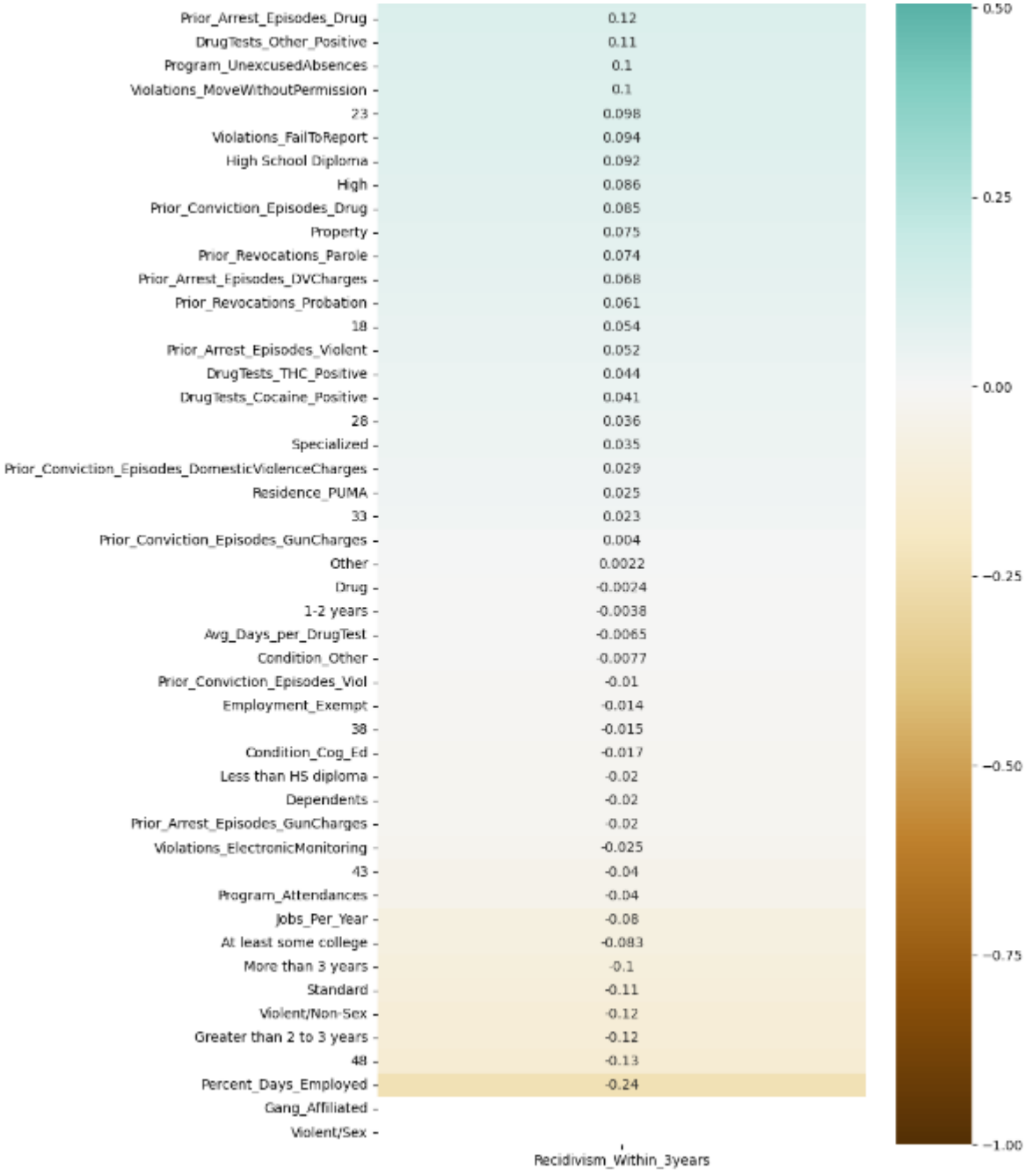




In [52]:

```
plt.figure(figsize=(8, 20))
heatmap = sns.heatmap(WhiteFemale.corr(method ='pearson')[['Recidivism_Within_3years']].sort_values(by='Recidivism_Withi
heatmap.set_title('Features for White Females Correlating with Recidivism', fontdict={'fontsize':16}, pad=16);
```




Prior_Arrest_Episodes_Drug 0.12
DrugTests_Other_Positive 0.11
Program_UnexcusedAbsences 0.1
Violations_MoveWithoutPermission 0.1
23 0.098
Violations_FailToReport 0.094
High School Diploma 0.092
High 0.086
Prior_Conviction_Episodes_Drug 0.085
Property 0.075
Prior_Revocations_Parole 0.074
Prior_Arrest_Episodes_DVCharges 0.068
Prior_Revocations_Probation 0.061
18 0.054
Prior_Arrest_Episodes_Violent 0.052
DrugTests_THC_Positive 0.044
DrugTests_Cocaine_Positive 0.041
28 0.036
Specialized 0.035
Prior_Conviction_Episodes_DomesticViolenceCharges 0.029
Residence_PUMA 0.025
33 0.023
Prior_Conviction_Episodes_GunCharges 0.004
Other 0.0022
Drug -0.0024
1-2 years -0.0038
Avg_Days_per_DrugTest -0.0065
Condition_Other -0.0077
Prior_Conviction_Episodes_Viol -0.01
Employment_Exempt -0.014
38 -0.015
Condition_Cog_Ed -0.017
Less than HS diploma -0.02
Dependents -0.02
Prior_Arrest_Episodes_GunCharges -0.02
Violations_ElectronicMonitoring -0.025
43 -0.04
Program_Attendances -0.04
Jobs_Per_Year -0.08
At least some college -0.083
More than 3 years -0.1
Standard -0.11
Violent/Non-Sex -0.12
Greater than 2 to 3 years -0.12
48 -0.13
Percent_Days_Employed -0.24
Gang_Affiliated
Violent/Sex
Recidivism_Within_3years
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Appendix B. NIJ Statistical Analysis

Jupyter Notebook available at: https://github.com/feaviolp/msc-project/blob/0a726abe16a82ed34d3f891361d0cca09d408edc/NIJ%20Analysis/NIJ%20statistics.ipynb

Statistical analysis of the NIJ recidivism dataset

Data pre-processing

Import libraries

In [1]:
```
library(haven)
library(dplyr)
library(ggplot2)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

In [2]:
```
NIJ <- read.csv("https://raw.githubusercontent.com/feaviolp/msc-project/main/NIJ%20Ontology/NIJ_s_Recidivism_Challenge_Full_
View(NIJ)
```

ID <int>	Gender <chr>	Race <chr>	Age_at_Release <chr>	Residence_PUMA <int>	Gang_Affiliated <chr>	Supervision_Risk_Score_First <int>	Supervision_Level_First <chr>	Educatior
A data.frame: 25835 × 54								
1	M	BLACK	43-47	16	false	3	Standard	At leas
2	M	BLACK	33-37	16	false	6	Specialized	Less t d
3	M	BLACK	48 or older	24	false	7	High	At leas
4	M	WHITE	38-42	16	false	7	High	Less t d
5	M	WHITE	33-37	16	false	4	Specialized	Less t d
6	M	WHITE	38-42	17	false	5	Standard	High D
7	M	BLACK	48 or older	18	false	2	Standard	Less t d
8	M	BLACK	38-42	16	false	5	High	High D
9	F	BLACK	43-47	5		7	High	High D
10	M	BLACK	43-47	16	false	5	Standard	High D
11	M	WHITE	43-47	5	false	3	Specialized	Less t d
12	M	BLACK	33-37	16	false	5	Specialized	High D
13	M	WHITE	48 or older	18	false	3	Standard	Less t d

14	M	BLACK	33-37	3	true	7		D
15	M	WHITE	33-37	5	false	7	Standard	Less t d
16	M	BLACK	33-37	3	false	4	Standard	Less t d
17	M	BLACK	38-42	16	false	6	High	Less t d
18	M	WHITE	43-47	24	false	1	Standard	High D
19	M	BLACK	48 or older	12	false	7	High	At leas
20	M	WHITE	48 or older	16	false	3	Standard	Less t d
21	M	BLACK	38-42	5	false	3	Standard	High D
22	M	BLACK	33-37	16	false	5	Standard	Less t d
23	F	WHITE	48 or older	5		4		High D
24	F	WHITE	43-47	16		5	Standard	Less t d
25	M	BLACK	38-42	18	false	5	Standard	Less t d
26	M	WHITE	48 or older	17	false	2		Less t d
27	M	WHITE	48 or older	14	false	2	Specialized	Less t d
28	M	WHITE	48 or older	23	false	3	Standard	Less t d
29	M	BLACK	43-47	6	false	6	High	High D
30	M	BLACK	18-22	2	false	10	Specialized	Less t d
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
26731	M	WHITE	43-47	7	false	3	Standard	High D
26732	M	BLACK	38-42	5	false	5	Standard	High D
26733	M	WHITE	23-27	17	false	5	Standard	High D
26734	M	BLACK	28-32	12	false	4	Standard	High D
26735	M	BLACK	23-27	19	false	7	High	At leas
26736	M	WHITE	38-42	12	true	5	Standard	At leas
26737	M	BLACK	28-32	8	false	5	Standard	At leas
26738	M	BLACK	33-37	11	false	5	Specialized	High D
26739	M	WHITE	43-47	2	false	5	Standard	At leas
26740	M	WHITE	48 or older	12	false	5	Standard	At leas
26741	M	WHITE	28-32	3	false	5	Standard	At leas
26742	M	WHITE	23-27	9	false	5	Standard	At leas

26743	M	WHITE	43-47	8	false	3	Standard	At leas
26744	M	BLACK	48 or older	21	false	7	Standard	Less t d
26745	M	WHITE	48 or older	24	false	5	Standard	At leas
26746	M	BLACK	43-47	6	false	5	Standard	At leas
26747	M	WHITE	33-37	8	false	5	Standard	At leas
26748	M	WHITE	48 or older	7	false	5	Standard	At leas
26749	M	BLACK	28-32	2	false	5	Standard	High D
26750	M	WHITE	23-27	18	false	5	Standard	High D
26751	M	BLACK	38-42	6	false	5	Standard	At leas
26752	M	BLACK	28-32	7	false	5	Standard	At leas
26753	M	BLACK	23-27	6	false	5	Standard	At leas
26754	M	WHITE	43-47	14	false	5	Standard	At leas
26755	M	WHITE	33-37	5	false	5	Standard	At leas
26756	M	BLACK	23-27	9	false	5	Standard	At leas
26758	M	WHITE	38-42	25	false	5	Standard	At leas
26759	M	BLACK	33-37	15	false	5	Standard	At leas
26760	F	WHITE	33-37	15		5	Standard	At leas
26761	M	WHITE	28-32	12	false	5	Standard	High D

Remove redundant columns from the dataframe

In [3]:
```
NIJ = subset(NIJ, select = -c(Recidivism_Arrest_Year1,Recidivism_Arrest_Year2,Recidivism_Arrest_Year3,Training_Sample))
```

Replace null values with appropriate values:

- Gang_Affiliated is missing only for Female offenders so replace with FALSE
- Supervision_Risk_Score_First is INTEGER so replace with the most frequently occuring value
- Supervision_Level_First is CATEGORICAL so replace with the most frequently occuring value
- Prison_Offense is categorical and includes "Other" so replace with "Other"
- Avg_Days_per_DrugTest is FLOAT so replace with average value
- DrugTests_THC_Positive is FLOAT but replace with 0
- DrugTests_Cocaine_Positive is FLOAT but replace with 0
- DrugTests_Meth_Positive is FLOAT but replace with 0
- DrugTests_Other_Positive is FLOAT but replace with 0
- Percent_Days_Employed is FLOAT so replace with average value
- Jobs_Per_Year is FLOAT so replace with average value

In [4]:
```
NIJ$Gang_Affiliated <- replace(NIJ$Gang_Affiliated, NIJ$Gang_Affiliated == '', 'false')
NIJ$Supervision_Risk_Score_First[is.na(NIJ$Supervision_Risk_Score_First)]<- names(which.max(table(NIJ$Supervision_Risk_Scor
NIJ$Supervision_Level_First <- replace(NIJ$Supervision_Level_First, NIJ$Supervision_Level_First == '', names(which.max(tabl
NIJ$Prison_Offense <- replace(NIJ$Prison_Offense, NIJ$Prison_Offense == '', 'Other')
NIJ$Avg_Days_per_DrugTest[is.na(NIJ$Avg_Days_per_DrugTest)]<-mean(NIJ$Avg_Days_per_DrugTest ,na.rm=TRUE)
```

```
NIJ$DrugTests_THC_Positive[is.na(NIJ$DrugTests_THC_Positive)]<-0
NIJ$DrugTests_Cocaine_Positive[is.na(NIJ$DrugTests_Cocaine_Positive)]<-0
NIJ$DrugTests_Meth_Positive[is.na(NIJ$DrugTests_Meth_Positive)]<-0
NIJ$DrugTests_Other_Positive[is.na(NIJ$DrugTests_Other_Positive)]<-0
NIJ$Percent_Days_Employed[is.na(NIJ$Percent_Days_Employed)]<-mean(NIJ$Percent_Days_Employed ,na.rm=TRUE)
NIJ$Jobs_Per_Year[is.na(NIJ$Jobs_Per_Year)]<-mean(NIJ$Jobs_Per_Year ,na.rm=TRUE)
```

Copy the dataframe to use for descriptive analysis later and replace M and F with Male and Female.

In [5]:
```
NIJ_orig <- NIJ
NIJ_orig$Gender <- replace(NIJ_orig$Gender, NIJ_orig$Gender == 'M', 'Male')
NIJ_orig$Gender <- replace(NIJ_orig$Gender, NIJ_orig$Gender == 'F', 'Female')
```

Some features contain integers but max out with, "x or more". To enable numerical analysis the "or more" will be removed and the column changed to integer:

- Dependents: "3 or more" changed to "3"
- Prior_Arrest_Episodes_Felony: "10 more more" changed to "10"
- Prior_Arrest_Episodes_Misd: "6 or more" changed to "6"
- Prior_Arrest_Episodes_Violent: "3 or more" changed to "3"
- Prior_Arrest_Episodes_Property: "5 or more" changed to "5"
- Prior_Arrest_Episodes_Drug: "5 or more" changed to "5"
- Prior_Arrest_Episodes_PPViolationCharges: "5 or more" changed to "5"
- Prior_Conviction_Episodes_Felony.replace: "3 or more" changed to "3"
- Prior_Conviction_Episodes_Misd.replace: "4 or more" changed to "4"
- Prior_Conviction_Episodes_Prop.replace: "3 or more" changed to "3"
- Prior_Conviction_Episodes_Drug.replace: "2 or more" changed to "2"
- Delinquency_Reports: "4 or more" changed to "4"
- Program_Attendances: "10 more more" changed to "10"
- Program_UnexcusedAbsences: "3 or more" changed to "3"
- Residence_Changes: "3 or more" changed to "3"

In [6]:
```
NIJ$Dependents <- replace(NIJ$Dependents, NIJ$Dependents == '3 or more', '3')
NIJ$Dependents <-  as.integer(NIJ$Dependents)

NIJ$Prior_Arrest_Episodes_Felony <- replace(NIJ$Prior_Arrest_Episodes_Felony, NIJ$Prior_Arrest_Episodes_Felony == '10 or mo
NIJ$Prior_Arrest_Episodes_Felony <-  as.integer(NIJ$Prior_Arrest_Episodes_Felony)

NIJ$Prior_Arrest_Episodes_Misd <- replace(NIJ$Prior_Arrest_Episodes_Misd, NIJ$Prior_Arrest_Episodes_Misd == '6 or more', '6
NIJ$Prior_Arrest_Episodes_Misd <-  as.integer(NIJ$Prior_Arrest_Episodes_Misd)

NIJ$Prior_Arrest_Episodes_Violent <- replace(NIJ$Prior_Arrest_Episodes_Violent, NIJ$Prior_Arrest_Episodes_Violent == '3 or
NIJ$Prior_Arrest_Episodes_Violent <-  as.integer(NIJ$Prior_Arrest_Episodes_Violent)

NIJ$Prior_Arrest_Episodes_Property <- replace(NIJ$Prior_Arrest_Episodes_Property, NIJ$Prior_Arrest_Episodes_Property == '5
NIJ$Prior_Arrest_Episodes_Property <-  as.integer(NIJ$Prior_Arrest_Episodes_Property)

NIJ$Prior_Arrest_Episodes_Drug <- replace(NIJ$Prior_Arrest_Episodes_Drug, NIJ$Prior_Arrest_Episodes_Drug == '5 or more', '5
NIJ$Prior_Arrest_Episodes_Drug <-  as.integer(NIJ$Prior_Arrest_Episodes_Drug)

NIJ$Prior_Arrest_Episodes_PPViolationCharges <- replace(NIJ$Prior_Arrest_Episodes_PPViolationCharges, NIJ$Prior_Arrest_Epis
NIJ$Prior_Arrest_Episodes_PPViolationCharges <-  as.integer(NIJ$Prior_Arrest_Episodes_PPViolationCharges)

NIJ$Prior_Conviction_Episodes_Felony <- replace(NIJ$Prior_Conviction_Episodes_Felony, NIJ$Prior_Conviction_Episodes_Felony
NIJ$Prior_Conviction_Episodes_Felony <-  as.integer(NIJ$Prior_Conviction_Episodes_Felony)

NIJ$Prior_Conviction_Episodes_Misd <- replace(NIJ$Prior_Conviction_Episodes_Misd, NIJ$Prior_Conviction_Episodes_Misd == '4
NIJ$Prior_Conviction_Episodes_Misd <-  as.integer(NIJ$Prior_Conviction_Episodes_Misd)

NIJ$Prior_Conviction_Episodes_Prop <- replace(NIJ$Prior_Conviction_Episodes_Prop, NIJ$Prior_Conviction_Episodes_Prop == '3
NIJ$Prior_Conviction_Episodes_Prop <-  as.integer(NIJ$Prior_Conviction_Episodes_Prop)

NIJ$Prior_Conviction_Episodes_Drug <- replace(NIJ$Prior_Conviction_Episodes_Drug, NIJ$Prior_Conviction_Episodes_Drug == '2
NIJ$Prior_Conviction_Episodes_Drug <-  as.integer(NIJ$Prior_Conviction_Episodes_Drug)

NIJ$Delinquency_Reports <- replace(NIJ$Delinquency_Reports, NIJ$Delinquency_Reports == '4 or more', '4')
NIJ$Delinquency_Reports <-  as.integer(NIJ$Delinquency_Reports)

NIJ$Program_Attendances <- replace(NIJ$Program_Attendances, NIJ$Program_Attendances == '10 or more', '10')
NIJ$Program_Attendances <-  as.integer(NIJ$Program_Attendances)

NIJ$Program_UnexcusedAbsences <- replace(NIJ$Program_UnexcusedAbsences, NIJ$Program_UnexcusedAbsences == '3 or more', '3')
```

```
NIJ$Program_UnexcusedAbsences <-  as.integer(NIJ$Program_UnexcusedAbsences)

NIJ$Residence_Changes <- replace(NIJ$Residence_Changes, NIJ$Residence_Changes == '3 or more', '3')
NIJ$Residence_Changes <-  as.integer(NIJ$Residence_Changes)
```

Age_at_Release is populated with ranges which won't be assessed as numeric, so change each value to the first number in each range and convert to integer:

In [7]:
```
NIJ$Age_at_Release <- replace(NIJ$Age_at_Release, NIJ$Age_at_Release == '18-22', '18')
NIJ$Age_at_Release <- replace(NIJ$Age_at_Release, NIJ$Age_at_Release == '23-27', '23')
NIJ$Age_at_Release <- replace(NIJ$Age_at_Release, NIJ$Age_at_Release == '28-32', '28')
NIJ$Age_at_Release <- replace(NIJ$Age_at_Release, NIJ$Age_at_Release == '33-37', '33')
NIJ$Age_at_Release <- replace(NIJ$Age_at_Release, NIJ$Age_at_Release == '38-42', '38')
NIJ$Age_at_Release <- replace(NIJ$Age_at_Release, NIJ$Age_at_Release == '43-47', '43')
NIJ$Age_at_Release <- replace(NIJ$Age_at_Release, NIJ$Age_at_Release == '48 or older', '48')

NIJ$Age_at_Release <-  as.integer(NIJ$Age_at_Release)
```

Prison_Years is populated with range descriptions so change to integer:

In [8]:
```
NIJ$Prison_Years <- replace(NIJ$Prison_Years, NIJ$Prison_Years == 'Less than 1 year', '0')
NIJ$Prison_Years <- replace(NIJ$Prison_Years, NIJ$Prison_Years == '1-2 years', '1')
NIJ$Prison_Years <- replace(NIJ$Prison_Years, NIJ$Prison_Years == 'Greater than 2 to 3 years', '2')
NIJ$Prison_Years <- replace(NIJ$Prison_Years, NIJ$Prison_Years == 'More than 3 years', '3')

NIJ$Prison_Years <-  as.integer(NIJ$Prison_Years)
```

Convert ordinal values to integers:

Supervision_Level_First

- Standard = 1
- High = 2
- Specialized = 3

Education_Level

- Less than HS diploma = 1
- High School Diploma = 2
- At least some college = 3

In [9]:
```
NIJ$Supervision_Level_First <- replace(NIJ$Supervision_Level_First, NIJ$Supervision_Level_First == 'Standard', '1')
NIJ$Supervision_Level_First <- replace(NIJ$Supervision_Level_First, NIJ$Supervision_Level_First == 'High', '2')
NIJ$Supervision_Level_First <- replace(NIJ$Supervision_Level_First, NIJ$Supervision_Level_First == 'Specialized', '3')

NIJ$Supervision_Level_First <-  as.integer(NIJ$Supervision_Level_First)

NIJ$Education_Level <- replace(NIJ$Education_Level, NIJ$Education_Level == 'Less than HS diploma', '1')
NIJ$Education_Level <- replace(NIJ$Education_Level, NIJ$Education_Level == 'High School Diploma', '2')
NIJ$Education_Level <- replace(NIJ$Education_Level, NIJ$Education_Level == 'At least some college', '3')

NIJ$Education_Level <-  as.integer(NIJ$Education_Level)
```

Convert boolean and character features to integer

In [10]:
```
glimpse(NIJ)
```

```
Rows: 25,835
Columns: 50
$ ID                                    <int> 1, 2, 3, 4, 5, 6, 7,…
$ Gender                                <chr> "M", "M", "M", "M", …
$ Race                                  <chr> "BLACK", "BLACK", "B…
$ Age_at_Release                        <int> 43, 33, 48, 38, 33, …
$ Residence_PUMA                        <int> 16, 16, 24, 16, 16, …
$ Gang_Affiliated                       <chr> "false", "false", "f…
$ Supervision_Risk_Score_First          <chr> "3", "6", "7", "7", …
$ Supervision_Level_First               <int> 1, 3, 2, 2, 3, 1, 1,…
$ Education_Level                       <int> 3, 1, 3, 1, 1, 2, 1,…
$ Dependents                            <int> 3, 1, 3, 1, 3, 0, 2,…
$ Prison_Offense                        <chr> "Drug", "Violent/Non…
$ Prison_Years                          <int> 3, 3, 1, 1, 1, 3, 0,…
$ Prior_Arrest_Episodes_Felony          <int> 6, 7, 6, 8, 4, 4, 10…
```

```
$ Prior_Arrest_Episodes_Misd                        <int> 6, 6, 6, 6, 4, 0, 6,…
$ Prior_Arrest_Episodes_Violent                     <int> 1, 3, 3, 0, 3, 1, 1,…
$ Prior_Arrest_Episodes_Property                    <int> 3, 0, 2, 3, 2, 3, 5,…
$ Prior_Arrest_Episodes_Drug                        <int> 3, 3, 2, 3, 1, 0, 1,…
$ Prior_Arrest_Episodes_PPViolationCharges          <int> 4, 5, 5, 3, 3, 0, 5,…
$ Prior_Arrest_Episodes_DVCharges                   <chr> "false", "true", "tr…
$ Prior_Arrest_Episodes_GunCharges                  <chr> "false", "false", "f…
$ Prior_Conviction_Episodes_Felony                  <int> 3, 3, 3, 3, 1, 1, 3,…
$ Prior_Conviction_Episodes_Misd                    <int> 3, 4, 2, 4, 0, 0, 1,…
$ Prior_Conviction_Episodes_Viol                    <chr> "false", "true", "tr…
$ Prior_Conviction_Episodes_Prop                    <int> 2, 0, 1, 3, 0, 2, 3,…
$ Prior_Conviction_Episodes_Drug                    <int> 2, 2, 2, 2, 1, 0, 0,…
$ Prior_Conviction_Episodes_PPViolationCharges      <chr> "false", "true", "fa…
$ Prior_Conviction_Episodes_DomesticViolenceCharges <chr> "false", "true", "tr…
$ Prior_Conviction_Episodes_GunCharges              <chr> "false", "true", "fa…
$ Prior_Revocations_Parole                          <chr> "false", "false", "f…
$ Prior_Revocations_Probation                       <chr> "false", "false", "f…
$ Condition_MH_SA                                   <chr> "true", "false", "tr…
$ Condition_Cog_Ed                                  <chr> "true", "false", "tr…
$ Condition_Other                                   <chr> "false", "false", "f…
$ Violations_ElectronicMonitoring                   <chr> "false", "false", "f…
$ Violations_Instruction                            <chr> "false", "true", "tr…
$ Violations_FailToReport                           <chr> "false", "false", "f…
$ Violations_MoveWithoutPermission                  <chr> "false", "false", "t…
$ Delinquency_Reports                               <int> 0, 4, 4, 0, 0, 0, 0,…
$ Program_Attendances                               <int> 6, 0, 6, 6, 7, 0, 0,…
$ Program_UnexcusedAbsences                         <int> 0, 0, 0, 0, 0, 0, 0,…
$ Residence_Changes                                 <int> 2, 2, 0, 3, 0, 3, 1,…
$ Avg_Days_per_DrugTest                             <dbl> 612.00000, 35.66667,…
$ DrugTests_THC_Positive                            <dbl> 0.0000000, 0.0000000…
$ DrugTests_Cocaine_Positive                        <dbl> 0, 0, 0, 0, 0, 0, 0,…
$ DrugTests_Meth_Positive                           <dbl> 0.00000000, 0.000000…
$ DrugTests_Other_Positive                          <dbl> 0.000000000, 0.00000…
$ Percent_Days_Employed                             <dbl> 0.4885621, 0.4252336…
$ Jobs_Per_Year                                     <dbl> 0.4476103, 2.0000000…
$ Employment_Exempt                                 <chr> "false", "false", "f…
$ Recidivism_Within_3years                          <chr> "false", "true", "tr…
```

In [11]:

```
NIJ$Gender <- as.integer(factor(NIJ$Gender))
NIJ$Race <- as.integer(factor(NIJ$Race))
NIJ$Gang_Affiliated <- as.integer(factor(NIJ$Gang_Affiliated))
NIJ$Supervision_Risk_Score_First <- as.integer(NIJ$Supervision_Risk_Score_First)
NIJ$Prison_Offense <- as.integer(factor(NIJ$Prison_Offense))
NIJ$Prior_Arrest_Episodes_DVCharges <- as.integer(factor(NIJ$Prior_Arrest_Episodes_DVCharges))
NIJ$Prior_Arrest_Episodes_GunCharges <- as.integer(factor(NIJ$Prior_Arrest_Episodes_GunCharges))
NIJ$Prior_Conviction_Episodes_Viol <- as.integer(factor(NIJ$Prior_Conviction_Episodes_Viol))
NIJ$Prior_Conviction_Episodes_PPViolationCharges <- as.integer(factor(NIJ$Prior_Conviction_Episodes_PPViolationCharges))
NIJ$Prior_Conviction_Episodes_DomesticViolenceCharges <- as.integer(factor(NIJ$Prior_Conviction_Episodes_DomesticViolenceCh
NIJ$Prior_Conviction_Episodes_GunCharges <- as.integer(factor(NIJ$Prior_Conviction_Episodes_GunCharges))
NIJ$Prior_Revocations_Parole  <- as.integer(factor(NIJ$Prior_Revocations_Parole ))
NIJ$Prior_Revocations_Probation <- as.integer(factor(NIJ$Prior_Revocations_Probation))
NIJ$Condition_MH_SA <- as.integer(factor(NIJ$Condition_MH_SA))
NIJ$Condition_Cog_Ed <- as.integer(factor(NIJ$Condition_Cog_Ed))
NIJ$Condition_Other <- as.integer(factor(NIJ$Condition_Other))
NIJ$Violations_ElectronicMonitoring <- as.integer(factor(NIJ$Violations_ElectronicMonitoring))
NIJ$Violations_Instruction <- as.integer(factor(NIJ$Violations_Instruction))
NIJ$Violations_FailToReport <- as.integer(factor(NIJ$Violations_FailToReport))
NIJ$Violations_MoveWithoutPermission <- as.integer(factor(NIJ$Violations_MoveWithoutPermission))
NIJ$Employment_Exempt <- as.integer(factor(NIJ$Employment_Exempt))
NIJ$Recidivism_Within_3years <- as.integer(factor(NIJ$Recidivism_Within_3years))
```

In [12]:

```
glimpse(NIJ)
```

```
Rows: 25,835
Columns: 50
$ ID                                <int> 1, 2, 3, 4, 5, 6, 7,…
$ Gender                            <int> 2, 2, 2, 2, 2, 2, 2,…
$ Race                              <int> 1, 1, 1, 2, 2, 2, 1,…
$ Age_at_Release                    <int> 43, 33, 48, 38, 33, …
$ Residence_PUMA                    <int> 16, 16, 24, 16, 16, …
$ Gang_Affiliated                   <int> 1, 1, 1, 1, 1, 1, 1,…
$ Supervision_Risk_Score_First      <int> 3, 6, 7, 7, 4, 5, 2,…
$ Supervision_Level_First           <int> 1, 3, 2, 2, 3, 1, 1,…
$ Education_Level                   <int> 3, 1, 3, 1, 1, 2, 1,…
$ Dependents                        <int> 3, 1, 3, 1, 3, 0, 2,…
$ Prison_Offense                    <int> 1, 4, 1, 3, 4, 3, 2,…
$ Prison_Years                      <int> 3, 3, 1, 1, 1, 3, 0,…
$ Prior_Arrest_Episodes_Felony      <int> 6, 7, 6, 8, 4, 4, 10…
$ Prior_Arrest_Episodes_Misd        <int> 6, 6, 6, 6, 4, 0, 6,…
```

```
$ Prior_Arrest_Episodes_Violent                    <int> 1, 3, 3, 0, 3, 1, 1,…
$ Prior_Arrest_Episodes_Property                   <int> 3, 0, 2, 3, 2, 3, 5,…
$ Prior_Arrest_Episodes_Drug                       <int> 3, 3, 2, 3, 1, 0, 1,…
$ Prior_Arrest_Episodes_PPViolationCharges         <int> 4, 5, 5, 3, 3, 0, 5,…
$ Prior_Arrest_Episodes_DVCharges                  <int> 1, 2, 2, 1, 2, 1, 1,…
$ Prior_Arrest_Episodes_GunCharges                 <int> 1, 1, 1, 1, 1, 1, 2,…
$ Prior_Conviction_Episodes_Felony                 <int> 3, 3, 3, 3, 1, 1, 3,…
$ Prior_Conviction_Episodes_Misd                   <int> 3, 4, 2, 4, 0, 0, 1,…
$ Prior_Conviction_Episodes_Viol                   <int> 1, 2, 2, 1, 2, 1, 1,…
$ Prior_Conviction_Episodes_Prop                   <int> 2, 0, 1, 3, 0, 2, 3,…
$ Prior_Conviction_Episodes_Drug                   <int> 2, 2, 2, 2, 1, 0, 0,…
$ Prior_Conviction_Episodes_PPViolationCharges     <int> 1, 2, 1, 1, 1, 1, 2,…
$ Prior_Conviction_Episodes_DomesticViolenceCharges <int> 1, 2, 2, 1, 1, 1, 1,…
$ Prior_Conviction_Episodes_GunCharges             <int> 1, 2, 1, 1, 1, 1, 2,…
$ Prior_Revocations_Parole                         <int> 1, 1, 1, 1, 1, 1, 1,…
$ Prior_Revocations_Probation                      <int> 1, 1, 1, 2, 1, 1, 1,…
$ Condition_MH_SA                                  <int> 2, 1, 2, 2, 2, 1, 1,…
$ Condition_Cog_Ed                                 <int> 2, 1, 2, 2, 2, 1, 1,…
$ Condition_Other                                  <int> 1, 1, 1, 1, 2, 2, 1,…
$ Violations_ElectronicMonitoring                  <int> 1, 1, 1, 1, 1, 1, 1,…
$ Violations_Instruction                           <int> 1, 2, 2, 1, 1, 1, 1,…
$ Violations_FailToReport                          <int> 1, 1, 1, 1, 1, 1, 1,…
$ Violations_MoveWithoutPermission                 <int> 1, 1, 2, 1, 1, 2, 1,…
$ Delinquency_Reports                              <int> 0, 4, 4, 0, 0, 0, 0,…
$ Program_Attendances                              <int> 6, 0, 6, 6, 7, 0, 0,…
$ Program_UnexcusedAbsences                        <int> 0, 0, 0, 0, 0, 0, 0,…
$ Residence_Changes                                <int> 2, 2, 0, 3, 0, 3, 1,…
$ Avg_Days_per_DrugTest                            <dbl> 612.00000, 35.66667,…
$ DrugTests_THC_Positive                           <dbl> 0.0000000, 0.0000000…
$ DrugTests_Cocaine_Positive                       <dbl> 0, 0, 0, 0, 0, 0, 0,…
$ DrugTests_Meth_Positive                          <dbl> 0.00000000, 0.000000…
$ DrugTests_Other_Positive                         <dbl> 0.000000000, 0.00000…
$ Percent_Days_Employed                            <dbl> 0.4885621, 0.4252336…
$ Jobs_Per_Year                                    <dbl> 0.4476103, 2.0000000…
$ Employment_Exempt                                <int> 1, 1, 1, 1, 1, 1, 1,…
$ Recidivism_Within_3years                         <int> 1, 2, 2, 1, 2, 1, 2,…
```

In [13]:
```
View(NIJ)
```

A data.frame: 25835 × 50

	ID	Gender	Race	Age_at_Release	Residence_PUMA	Gang_Affiliated	Supervision_Risk_Score_First	Supervision_Level_First	Ec
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	
1	1	2	1	43	16	1	3	1	
2	2	2	1	33	16	1	6	3	
3	3	2	1	48	24	1	7	2	
4	4	2	2	38	16	1	7	2	
5	5	2	2	33	16	1	4	3	
6	6	2	2	38	17	1	5	1	
7	7	2	1	48	18	1	2	1	
8	8	2	1	38	16	1	5	2	
9	9	1	1	43	5	1	7	2	
10	10	2	1	43	16	1	5	1	
11	11	2	2	43	5	1	3	3	
12	12	2	1	33	16	1	5	3	
13	13	2	2	48	18	1	3	1	
14	14	2	1	33	3	2	7	1	
15	15	2	2	33	5	1	7	1	
16	16	2	1	33	3	1	4	1	
17	17	2	1	38	16	1	6	2	
18	18	2	2	43	24	1	1	1	
19	19	2	1	48	12	1	7	2	

20	20	2	2	48	16	1	3	1
21	21	2	1	38	5	1	3	1
22	22	2	1	33	16	1	5	1
23	23	1	2	48	5	1	4	1
24	24	1	2	43	16	1	5	1
25	25	2	1	38	18	1	5	1
26	26	2	2	48	17	1	2	1
27	27	2	2	48	14	1	2	3
28	28	2	2	48	23	1	3	1
29	29	2	1	43	6	1	6	2
30	30	2	1	18	2	1	10	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
25806	26731	2	2	43	7	1	3	1
25807	26732	2	1	38	5	1	5	1
25808	26733	2	2	23	17	1	5	1
25809	26734	2	1	28	12	1	4	1
25810	26735	2	1	23	19	1	7	2
25811	26736	2	2	38	12	2	5	1
25812	26737	2	1	28	8	1	5	1
25813	26738	2	1	33	11	1	5	3
25814	26739	2	2	43	2	1	5	1
25815	26740	2	2	48	12	1	5	1
25816	26741	2	2	28	3	1	5	1
25817	26742	2	2	23	9	1	5	1
25818	26743	2	2	43	8	1	3	1
25819	26744	2	1	48	21	1	7	1
25820	26745	2	2	48	24	1	5	1
25821	26746	2	1	43	6	1	5	1
25822	26747	2	2	33	8	1	5	1
25823	26748	2	2	48	7	1	5	1
25824	26749	2	1	28	2	1	5	1
25825	26750	2	2	23	18	1	5	1
25826	26751	2	1	38	6	1	5	1
25827	26752	2	1	28	7	1	5	1
25828	26753	2	1	23	6	1	5	1
25829	26754	2	2	43	14	1	5	1
25830	26755	2	2	33	5	1	5	1
25831	26756	2	1	23	9	1	5	1
25832	26758	2	2	38	25	1	5	1
25833	26759	2	1	33	15	1	5	1
25834	26760	1	2	33	15	1	5	1
25835	26761	2	2	28	12	1	5	1

In order to present age plots based on Race and Gender, a new datafram is created based upon the original NIJ datafame with categorical values and the Age ranges converted to integers from the transformed dataframe.

In [14]:
```
RaceGenderAge <- NIJ_orig
RaceGenderAge$Age_at_Release <- NIJ$Age_at_Release
```

Separate the dataset into Male and Female for more detailed comparrisons later.

In [15]:
```
Male <- subset(NIJ, Gender == 2)
Female <- subset(NIJ, Gender == 1)
```

Statistical analysis

Examine statistical differences in the data between the protected characteristics

Is there a difference in recidivism by race?

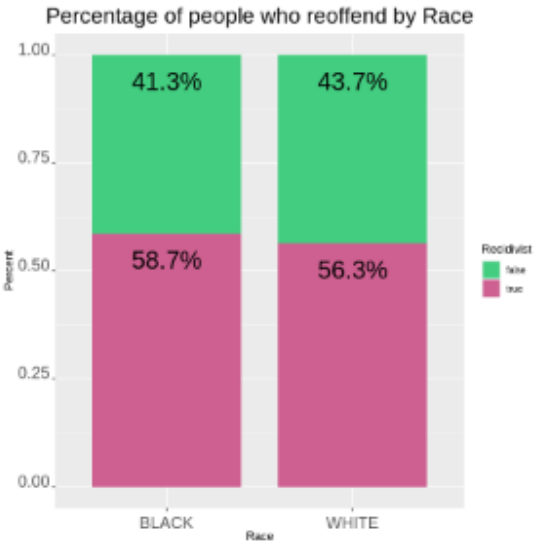
Both Race and Recidivism_Within_3years are categorical so Chi-square test used to see if there is a difference in the level of recidivism between Black and White offenders.

In [16]:
```
table(as_factor(NIJ_orig$Recidivism_Within_3years), as_factor(NIJ_orig$Race))
chisq.test(table(NIJ_orig$Recidivism_Within_3years, NIJ_orig$Race))
```

```
       BLACK WHITE
 false  6134  4797
 true   8713  6191
```

```
	Pearson's Chi-squared test with Yates' continuity correction

data:  table(NIJ_orig$Recidivism_Within_3years, NIJ_orig$Race)
X-squared = 14.094, df = 1, p-value = 0.0001739
```

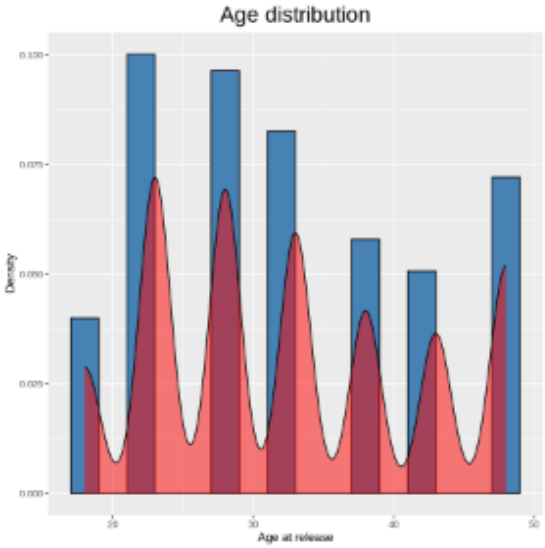
In [17]:
```
ggplot(NIJ_orig, aes(fill=Recidivism_Within_3years, x=Race)) +
  geom_bar(position = "fill", width=0.8) +
  labs(title='Percentage of people who reoffend by Race', x='Race', y='Percent') +
  geom_text(
    aes(label = after_stat(
      scales::percent(
        ave(count, x, FUN = function(x) x / sum(x))
      )
    )),
    stat = "count", position = "fill"
  , vjust=2, hjust=0.5, size = 8) +
  guides(fill = guide_legend(title = "Recidivist")) +
  scale_fill_manual(values = c('seagreen3', 'hotpink3')) +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 20)) +
  theme(axis.text.x=element_text(angle=0,hjust=0.5,vjust=0, size=15)) +
  theme(axis.text.y=element_text(angle=0,hjust=0.5,vjust=0, size=15))
```



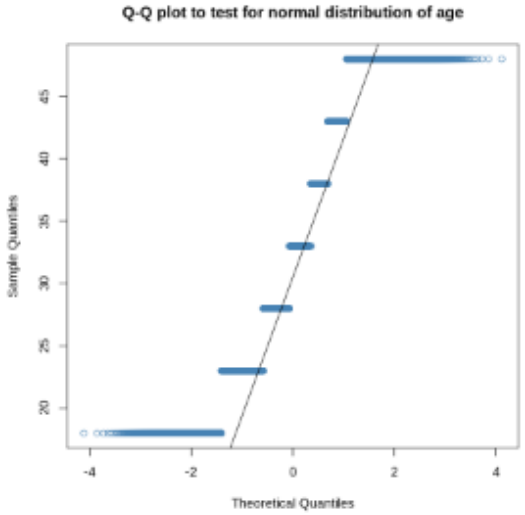
Is there a difference in recidivism by Age?

First check if Age is normally distributed in total, by race, and by gender.

In [18]:
```
ggplot(RaceGenderAge, aes(x=Age_at_Release)) +
  geom_histogram(aes(y=after_stat(density)), binwidth=2,
                 colour='black', fill='steel blue') +
  geom_density(alpha=0.5, fill='red') +
  labs(title = "Age distribution", y = 'Density', x='Age at release') +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 20))
```

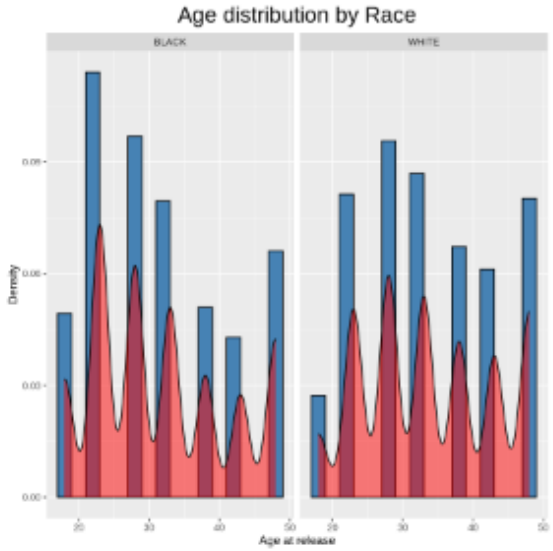


In [19]:
```
qqnorm(RaceGenderAge$Age_at_Release,
       col='steelblue',
       main='Q-Q plot to test for normal distribution of age')
qqline(RaceGenderAge$Age_at_Release)
```



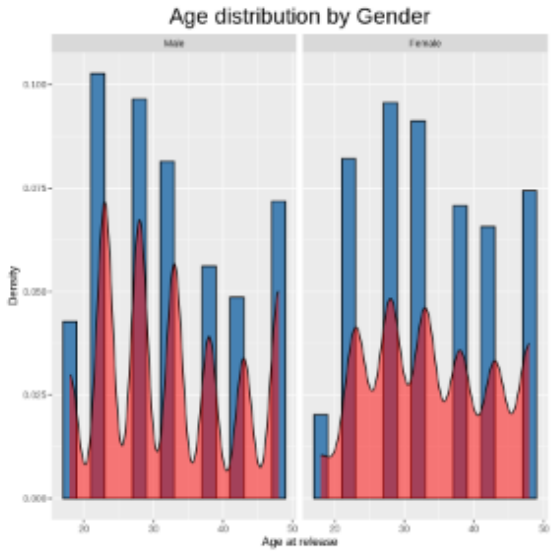
In [20]:
```
ggplot(RaceGenderAge, aes(x=Age_at_Release)) +
  geom_histogram(aes(y=after_stat(density)), binwidth=2,
                 colour='black', fill='steel blue') +
  geom_density(alpha=0.5, fill='red') +
  facet_grid(~as_factor(Race)) +
  labs(title = "Age distribution by Race", y = 'Density', x='Age at release') +
  theme_grey() +
```

```
  theme(plot.title = element_text(hjust = 0.5, size = 20))
```



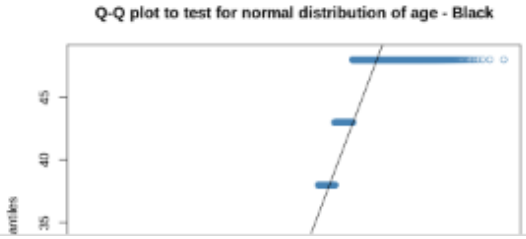
In [21]:

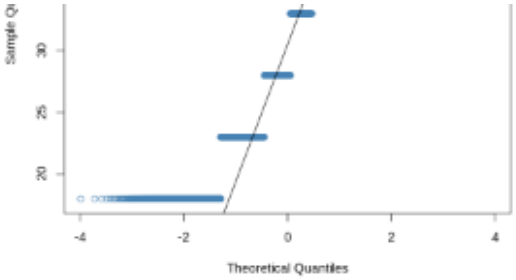
```
ggplot(RaceGenderAge, aes(x=Age_at_Release)) +
  geom_histogram(aes(y=after_stat(density)), binwidth=2,
                 colour='black', fill='steel blue') +
  geom_density(alpha=0.5, fill='red') +
  facet_grid(~as_factor(Gender)) +
  labs(title = "Age distribution by Gender", y = 'Density', x='Age at release') +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 20))
```



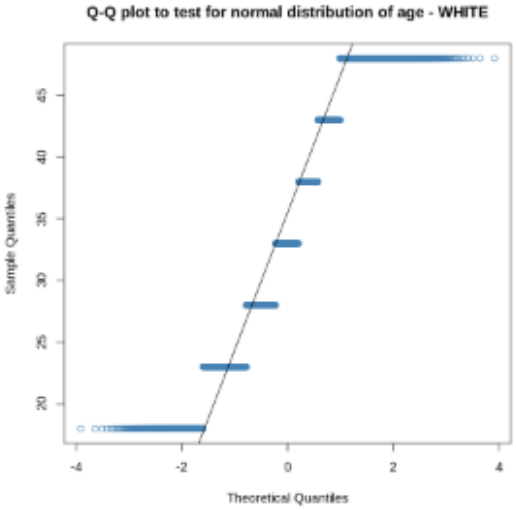
In [22]:

```
qqnorm(RaceGenderAge$Age_at_Release[as_factor(RaceGenderAge$Race)=='BLACK'],
       col='steelblue',
       main='Q-Q plot to test for normal distribution of age - Black')
qqline(RaceGenderAge$Age_at_Release[as_factor(RaceGenderAge$Race)=='BLACK'])
```





In [23]:
```
qqnorm(RaceGenderAge$Age_at_Release[as_factor(RaceGenderAge$Race)=='WHITE'],
       col='steelblue',
       main='Q-Q plot to test for normal distribution of age - WHITE')
qqline(RaceGenderAge$Age_at_Release[as_factor(RaceGenderAge$Race)=='WHITE'])
```

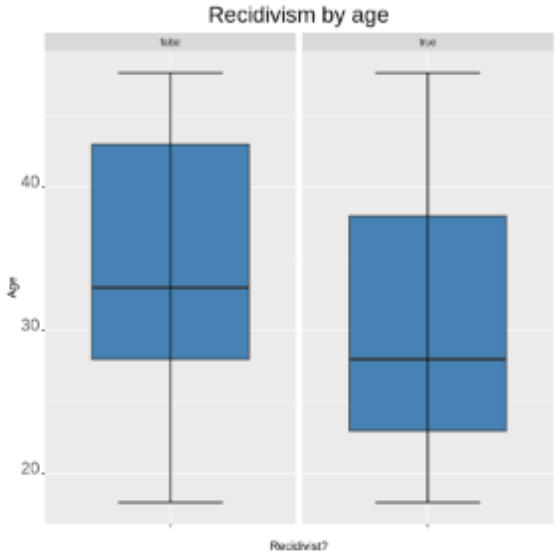


The distribution of Age is not normal so the non-parametric Mann-Whitney U test can be used. However, with a large sample size the central limit theorem allows the use of a t-test, so both will be checked.

In [24]:
```
wilcox.test(Age_at_Release ~ Recidivism_Within_3years, data=RaceGenderAge)
t.test(Age_at_Release ~ as_factor(Recidivism_Within_3years), data=RaceGenderAge)
```

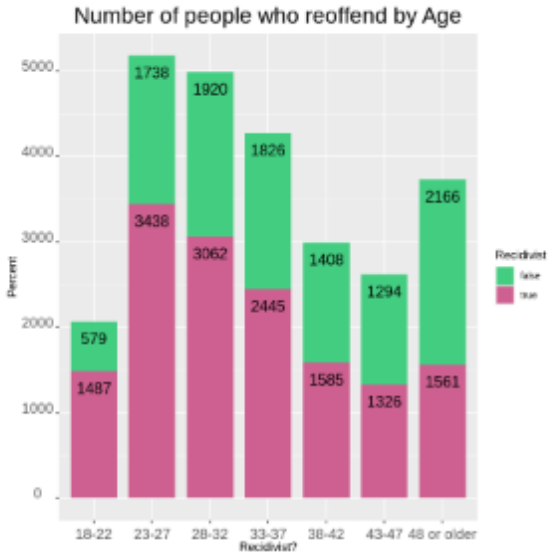
```
        Wilcoxon rank sum test with continuity correction

data:  Age_at_Release by Recidivism_Within_3years
W = 98018098, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Welch Two Sample t-test

data:  Age_at_Release by as_factor(Recidivism_Within_3years)
t = 28.822, df = 22988, p-value < 2.2e-16
alternative hypothesis: true difference in means between group false and group true is not equal to 0
95 percent confidence interval:
 3.145913 3.605023
sample estimates:
mean in group false  mean in group true
           34.53737            31.16190
```

In [25]:
```
ggplot(RaceGenderAge, aes(x='', y=Age_at_Release)) +
  geom_boxplot(fill='steel blue') +
  stat_boxplot(geom ='errorbar', width=0.5) +
  labs(title = "Recidivism by age", y = 'Age', x='Recidivist?') +
  facet_grid(~as_factor(Recidivism_Within_3years)) +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 20)) +
  theme(axis.text.y=element_text(angle=0,hjust=0.5,vjust=0, size=15))
```



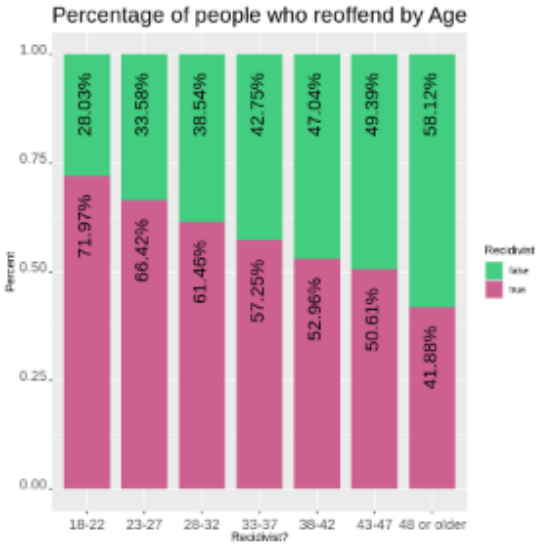
In [26]:

```
ggplot(NIJ_orig, aes(fill=Recidivism_Within_3years, x=Age_at_Release)) +
  geom_bar(position='stack', stat='count', width=0.8) +
  labs(title='Number of people who reoffend by Age', x='Recidivist?', y='Percent') +
  geom_text(aes(label = after_stat(count)), stat = "count", position = "stack", vjust=2, size = 5) +
  guides(fill = guide_legend(title = "Recidivist")) +
  scale_fill_manual(values = c('seagreen3', 'hotpink3')) +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 20)) +
  theme(axis.text.x=element_text(angle=0,hjust=0.5,vjust=0, size=13)) +
  theme(axis.text.y=element_text(angle=0,hjust=0.5,vjust=0, size=13))
```



In [27]:

```
ggplot(NIJ_orig, aes(fill=Recidivism_Within_3years, x=Age_at_Release)) +
  geom_bar(position = "fill", width=0.8) +
  labs(title='Percentage of people who reoffend by Age', x='Recidivist?', y='Percent') +
  geom_text(
    aes(label = after_stat(
      scales::percent(
        ave(count, x, FUN = function(x) x / sum(x))
      )
    )),
    stat = "count", position = "fill"
   , vjust=0.3, hjust=1.3, angle=90, size = 6) +
   guides(fill = guide_legend(title = "Recidivist")) +
   scale_fill_manual(values = c('seagreen3', 'hotpink3')) +
   theme_grey() +
   theme(plot.title = element_text(hjust = 0.5, size = 20)) +
   theme(axis.text.x=element_text(angle=0,hjust=0.5,vjust=0, size=13)) +
   theme(axis.text.y=element_text(angle=0,hjust=0.5,vjust=0, size=13))
```

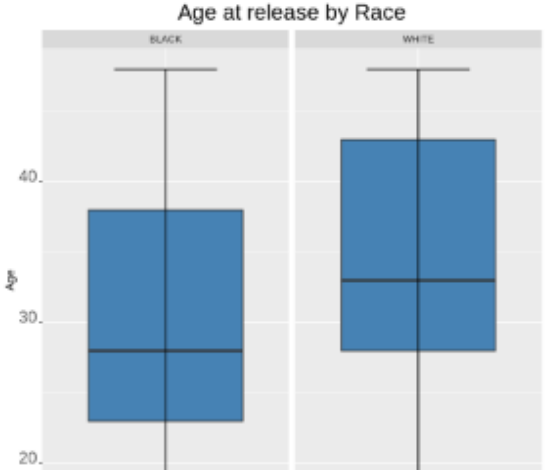


Is there a difference in Age by Race?

In [28]:
```
wilcox.test(Age_at_Release ~ Race, data=RaceGenderAge)
t.test(Age_at_Release ~ as_factor(Race), data=RaceGenderAge)
```

```
	Wilcoxon rank sum test with continuity correction

data:  Age_at_Release by Race
W = 69826119, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
	Welch Two Sample t-test

data:  Age_at_Release by as_factor(Race)
t = -19.514, df = 24010, p-value < 2.2e-16
alternative hypothesis: true difference in means between group BLACK and group WHITE is not equal to 0
95 percent confidence interval:
 -2.509302 -2.051234
sample estimates:
mean in group BLACK mean in group WHITE
          31.62026            33.90053
```

In [29]:
```
ggplot(RaceGenderAge, aes(x='', y=Age_at_Release)) +
  geom_boxplot(fill='steel blue') +
  stat_boxplot(geom ='errorbar', width=0.5) +
  labs(title = "Age at release by Race", y = 'Age', x='Race') +
  facet_grid(~as_factor(Race)) +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 20)) +
  theme(axis.text.y=element_text(angle=0,hjust=0.5,vjust=0, size=15))
```



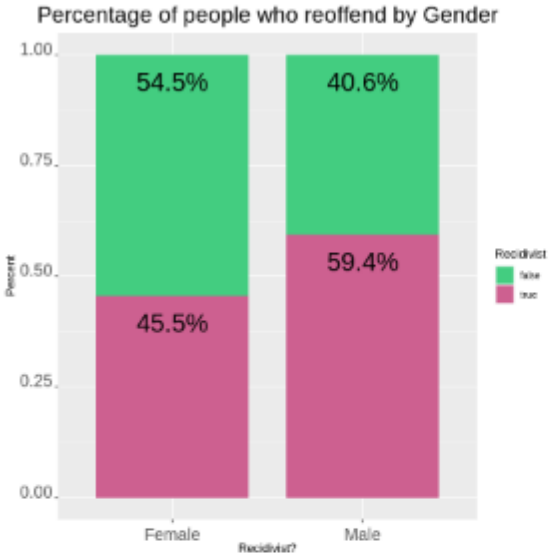
Is there a difference in recidivism by gender?

Both Gender and Recidivism_Within_3years are categorical so Chi-Square is used.

In [30]:
```
table(as_factor(NIJ_orig$Recidivism_Within_3years), as_factor(NIJ_orig$Gender))
chisq.test(table(NIJ_orig$Recidivism_Within_3years, NIJ_orig$Gender))
```

```
       Male Female
false  9206   1725
true  13462   1442
	Pearson's Chi-squared test with Yates' continuity correction

data:  table(NIJ_orig$Recidivism_Within_3years, NIJ_orig$Gender)
X-squared = 217.99, df = 1, p-value < 2.2e-16
```

In [31]:
```
ggplot(NIJ_orig, aes(fill=Recidivism_Within_3years, x=Gender)) +
  geom_bar(position = "fill", width=0.8) +
  labs(title='Percentage of people who reoffend by Gender', x='Recidivist?', y='Percent') +
  geom_text(
    aes(label = after_stat(
      scales::percent(
        ave(count, x, FUN = function(x) x / sum(x))
      )
    )),
    stat = "count", position = "fill"
  , vjust=2, hjust=0.5, size = 8) +
   guides(fill = guide_legend(title = "Recidivist")) +
   scale_fill_manual(values = c('seagreen3', 'hotpink3')) +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 20)) +
  theme(axis.text.x=element_text(angle=0,hjust=0.5,vjust=0, size=15)) +
  theme(axis.text.y=element_text(angle=0,hjust=0.5,vjust=0, size=15))
```



Correlations with recidivism and race

Is there a correlation between Age_at_Release and recidivism and/or race for everyone, or just males, or just females?

In [32]:
```
wilcox.test(Age_at_Release ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Age_at_Release, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
	Wilcoxon rank sum test with continuity correction

data:  Age_at_Release by Recidivism_Within_3years
W = 98018098, p-value < 2.2e-16
```

```
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Age_at_Release and NIJ$Recidivism_Within_3years
S = 3.3805e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1762601
```

In [33]:
```
wilcox.test(Age_at_Release ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Age_at_Release, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Age_at_Release by Recidivism_Within_3years
W = 74667771, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Age_at_Release and Male$Recidivism_Within_3years
S = 2.2843e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1766892
```

In [34]:
```
wilcox.test(Age_at_Release ~ Race, data=Male)
cor.test(Male$Age_at_Release, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Age_at_Release by Race
W = 52608014, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Age_at_Release and Male$Race
S = 1.7059e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1212351
```

In [35]:
```
wilcox.test(Age_at_Release ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Age_at_Release, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Age_at_Release by Recidivism_Within_3years
W = 1432521, p-value = 8.335e-14
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Age_at_Release and Female$Recidivism_Within_3years
S = 5996475425, p-value = 6.563e-14
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1326694
```

In [36]:
```
wilcox.test(Age_at_Release ~ Race, data=Female)
cor.test(Female$Age_at_Release, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Age_at_Release by Race
W = 1029906, p-value = 4.66e-05
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Age_at_Release and Female$Race
S = 4.911e+09, p-value = 4.572e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.07237061
```

Is there a correlation between Residence_PUMA and recidivism and/or race for everyone, or just males, or just females?

In [37]:
```
wilcox.test(Residence_PUMA ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Residence_PUMA, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Residence_PUMA by Recidivism_Within_3years
W = 79078201, p-value = 5.764e-05
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Residence_PUMA and NIJ$Recidivism_Within_3years
S = 2.802e+12, p-value = 5.751e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.02502519
```

In [38]:
```
wilcox.test(Residence_PUMA ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Residence_PUMA, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Residence_PUMA by Recidivism_Within_3years
W = 60068431, p-value = 8.666e-05
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Residence_PUMA and Male$Recidivism_Within_3years
S = 1.8907e+12, p-value = 8.647e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.02607135
```

In [39]:
```
wilcox.test(Residence_PUMA ~ Race, data=Male)
cor.test(Male$Residence_PUMA, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Residence_PUMA by Race
W = 51225560, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Residence_PUMA and Male$Race
S = 1.6717e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1388776
```

In [40]:
```
wilcox.test(Residence_PUMA ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Residence_PUMA, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Residence_PUMA by Recidivism_Within_3years
W = 1183842, p-value = 0.01928
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Residence_PUMA and Female$Recidivism_Within_3years
S = 5073939404, p-value = 0.01926
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.0415877
```

In [41]:
```
wilcox.test(Residence_PUMA ~ Race, data=Female)
cor.test(Female$Residence_PUMA, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Residence_PUMA by Race
W = 870884, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Residence_PUMA and Female$Race
S = 4301514099, p-value < 2.2e-16
```

```
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1874905
```

Is there a correlation between Gang_Affiliated and recidivism and/or race for everyone, or just males?

(Females do not have Gang_Affiliated recorded)

In [42]:
```
wilcox.test(Gang_Affiliated ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Gang_Affiliated, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Gang_Affiliated by Recidivism_Within_3years
W = 70514106, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Gang_Affiliated and NIJ$Recidivism_Within_3years
S = 2.3432e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1846643
```

In [43]:
```
wilcox.test(Gang_Affiliated ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Gang_Affiliated, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Gang_Affiliated by Recidivism_Within_3years
W = 53121518, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Gang_Affiliated and Male$Recidivism_Within_3years
S = 1.5821e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1850057
```

In [44]:
```
wilcox.test(Gang_Affiliated ~ Race, data=Male)
cor.test(Male$Gang_Affiliated, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Gang_Affiliated by Race
W = 65362780, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Gang_Affiliated and Male$Race
S = 2.1082e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.08599346
```

Is there a correlation between Supervision_Risk_Score_First and recidivism and/or race for everyone, or just males, or just females?

In [45]:
```
wilcox.test(Supervision_Risk_Score_First ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Supervision_Risk_Score_First, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE
```

```
        Wilcoxon rank sum test with continuity correction

data:  Supervision_Risk_Score_First by Recidivism_Within_3years
W = 64434499, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Supervision_Risk_Score_First and NIJ$Recidivism_Within_3years
S = 2.3561e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1801646
```

In [46]:
```
wilcox.test(Supervision_Risk_Score_First ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Supervision_Risk_Score_First, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALS
```

```
        Wilcoxon rank sum test with continuity correction

data:  Supervision_Risk_Score_First by Recidivism_Within_3years
W = 48568409, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Supervision_Risk_Score_First and Male$Recidivism_Within_3years
S = 1.5816e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1852754
```

In [47]:
```
wilcox.test(Supervision_Risk_Score_First ~ Race, data=Male)
cor.test(Male$Supervision_Risk_Score_First, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Supervision_Risk_Score_First by Race
W = 65061744, p-value = 2.211e-15
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Supervision_Risk_Score_First and Male$Race
S = 2.0435e+12, p-value = 2.119e-15
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.05266429
```

In [48]:
```
wilcox.test(Supervision_Risk_Score_First ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Supervision_Risk_Score_First, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact
```

```
        Wilcoxon rank sum test with continuity correction

data:  Supervision_Risk_Score_First by Recidivism_Within_3years
W = 1034138, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Supervision_Risk_Score_First and Female$Recidivism_Within_3years
S = 4518704309, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1464656
```

In [49]:
```
wilcox.test(Supervision_Risk_Score_First ~ Race, data=Female)
cor.test(Female$Supervision_Risk_Score_First, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Supervision_Risk_Score_First by Race
W = 1064797, p-value = 0.009081
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Supervision_Risk_Score_First and Female$Race
S = 5048632463, p-value = 0.00906
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.0463679
```

Is there a correlation between Supervision_Level_First and recidivism and/or race for everyone, or just males, or just females?

In [50]:
```
wilcox.test(Supervision_Level_First ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Supervision_Level_First, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction
```

```
data:  Supervision_Level_First by Recidivism_Within_3years
W = 76067222, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Supervision_Level_First and NIJ$Recidivism_Within_3years
S = 2.699e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.06084827
```

In [51]:
```
wilcox.test(Supervision_Level_First ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Supervision_Level_First, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Supervision_Level_First by Recidivism_Within_3years
W = 58355268, p-value = 1.22e-15
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Supervision_Level_First and Male$Recidivism_Within_3years
S = 1.8381e+12, p-value = 1.168e-15
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.05315242
```

In [52]:
```
wilcox.test(Supervision_Level_First ~ Race, data=Male)
cor.test(Male$Supervision_Level_First, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Supervision_Level_First by Race
W = 62723506, p-value = 0.001243
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Supervision_Level_First and Male$Race
S = 1.9829e+12, p-value = 0.001241
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.02144681
```

In [53]:
```
wilcox.test(Supervision_Level_First ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Supervision_Level_First, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALS
```

```
        Wilcoxon rank sum test with continuity correction

data:  Supervision_Level_First by Recidivism_Within_3years
W = 1153571, p-value = 0.0001114
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Supervision_Level_First and Female$Recidivism_Within_3years
S = 4930517683, p-value = 0.0001097
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.06867851
```

In [54]:
```
wilcox.test(Supervision_Level_First ~ Race, data=Female)
cor.test(Female$Supervision_Level_First, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Supervision_Level_First by Race
W = 1147841, p-value = 0.3715
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Supervision_Level_First and Female$Race
S = 5378196639, p-value = 0.3716
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.01588321
```

Is there a correlation between Education_Level and recidivism and/or race for everyone, or just males, or just females?

In [55]:
```
wilcox.test(Education_Level ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Education_Level, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Education_Level by Recidivism_Within_3years
W = 89183922, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Education_Level and NIJ$Recidivism_Within_3years
S = 3.1264e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.08785932
```

In [56]:
```
wilcox.test(Education_Level ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Education_Level, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Education_Level by Recidivism_Within_3years
W = 67810954, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Education_Level and Male$Recidivism_Within_3years
S = 2.1105e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.08715922
```

In [57]:
```
wilcox.test(Education_Level ~ Race, data=Male)
cor.test(Male$Education_Level, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Education_Level by Race
W = 57504234, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Education_Level and Male$Race
S = 1.8315e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.05653968
```

In [58]:
```
wilcox.test(Education_Level ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Education_Level, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Education_Level by Recidivism_Within_3years
W = 1295912, p-value = 0.02993
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Education_Level and Female$Recidivism_Within_3years
S = 5498378072, p-value = 0.02991
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.03858418
```

In [59]:
```
wilcox.test(Education_Level ~ Race, data=Female)
cor.test(Female$Education_Level, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Education_Level by Race
W = 1092961, p-value = 0.126
alternative hypothesis: true location shift is not equal to 0
```

```
        Spearman's rank correlation rho

data:  Female$Education_Level and Female$Race
S = 5150157751, p-value = 0.126
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.02719088
```

Is there a correlation between Dependents and recidivism and/or race for everyone, or just males, or just females?

In [60]:
```
wilcox.test(Dependents ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Dependents, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Dependents by Recidivism_Within_3years
W = 84298694, p-value = 6.358e-07
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Dependents and NIJ$Recidivism_Within_3years
S = 2.963e+12, p-value = 6.324e-07
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.03098377
```

In [61]:
```
wilcox.test(Dependents ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Dependents, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Dependents by Recidivism_Within_3years
W = 64158533, p-value = 2.519e-06
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Dependents and Male$Recidivism_Within_3years
S = 2.002e+12, p-value = 2.506e-06
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.03126164
```

In [62]:
```
wilcox.test(Dependents ~ Race, data=Male)
cor.test(Male$Dependents, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Dependents by Race
W = 67958992, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Dependents and Male$Race
S = 2.1273e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.09582089
```

In [63]:
```
wilcox.test(Dependents ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Dependents, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Dependents by Recidivism_Within_3years
W = 1253358, p-value = 0.6966
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Dependents and Female$Recidivism_Within_3years
S = 5330799694, p-value = 0.6966
alternative hypothesis: true rho is not equal to 0
sample estimates:
         rho
-0.006930435
```

In [64]:
```
wilcox.test(Dependents ~ Race, data=Female)
```

```
cor.test(Female$Dependents, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Dependents by Race
W = 1212521, p-value = 0.0003262
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Dependents and Female$Race
S = 5632225875, p-value = 0.0003226
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.06386658
```

Is there a correlation between Prison_Offense and recidivism and/or race for everyone, or just males, or just females?

In [65]:
```
wilcox.test(Prison_Offense ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prison_Offense, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prison_Offense by Recidivism_Within_3years
W = 83076946, p-value = 0.004727
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prison_Offense and NIJ$Recidivism_Within_3years
S = 2.9244e+12, p-value = 0.004725
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.01757645
```

In [66]:
```
wilcox.test(Prison_Offense ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prison_Offense, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prison_Offense by Recidivism_Within_3years
W = 63649651, p-value = 0.0003267
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prison_Offense and Male$Recidivism_Within_3years
S = 1.9876e+12, p-value = 0.0003262
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.02386608
```

In [67]:
```
wilcox.test(Prison_Offense ~ Race, data=Male)
cor.test(Male$Prison_Offense, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prison_Offense by Race
W = 63609402, p-value = 5.472e-07
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prison_Offense and Male$Race
S = 2.0059e+12, p-value = 5.438e-07
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.03326985
```

In [68]:
```
wilcox.test(Prison_Offense ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prison_Offense, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prison_Offense by Recidivism_Within_3years
W = 1280305, p-value = 0.1343
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho
```

```
data:  Female$Prison_Offense and Female$Recidivism_Within_3years
S = 5.435e+09, p-value = 0.1343
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.0266105
```

In [69]:
```
wilcox.test(Prison_Offense ~ Race, data=Female)
cor.test(Female$Prison_Offense, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prison_Offense by Race
W = 1468656, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prison_Offense and Female$Race
S = 6671785940, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.2602282
```

Is there a correlation between Prison_Years episodes and recidivism and/or race for everyone, or just males, or just females?

In [70]:
```
wilcox.test(Prison_Years ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prison_Years, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prison_Years by Recidivism_Within_3years
W = 93350168, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prison_Years and NIJ$Recidivism_Within_3years
S = 3.2472e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1298835
```

In [71]:
```
wilcox.test(Prison_Years ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prison_Years, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prison_Years by Recidivism_Within_3years
W = 71377101, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prison_Years and Male$Recidivism_Within_3years
S = 2.2016e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1341215
```

In [72]:
```
wilcox.test(Prison_Years ~ Race, data=Male)
cor.test(Male$Prison_Years, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prison_Years by Race
W = 65906564, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prison_Years and Male$Race
S = 2.0701e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.06637585
```

In [73]:
```
wilcox.test(Prison_Years ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prison_Years, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prison_Years by Recidivism_Within_3years
W = 1496864, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prison_Years and Female$Recidivism_Within_3years
S = 6278922029, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1860205
```

In [74]:
```
wilcox.test(Prison_Years ~ Race, data=Female)
cor.test(Female$Prison_Years, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prison_Years by Race
W = 1269646, p-value = 7.72e-10
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prison_Years and Female$Race
S = 5872811843, p-value = 6.932e-10
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1093107
```

Is there a correlation between Prior_Arrest_Episodes_Felony and recidivism and/or race for everyone, or just males, or just females?

In [75]:
```
wilcox.test(Prior_Arrest_Episodes_Felony ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prior_Arrest_Episodes_Felony, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Felony by Recidivism_Within_3years
W = 62707353, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Arrest_Episodes_Felony and NIJ$Recidivism_Within_3years
S = 2.3022e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1989344
```

In [76]:
```
wilcox.test(Prior_Arrest_Episodes_Felony ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Arrest_Episodes_Felony, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FAL
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Felony by Recidivism_Within_3years
W = 48507358, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Arrest_Episodes_Felony and Male$Recidivism_Within_3years
S = 1.579e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1866431
```

In [77]:
```
wilcox.test(Prior_Arrest_Episodes_Felony ~ Race, data=Male)
cor.test(Male$Prior_Arrest_Episodes_Felony, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Felony by Race
W = 59468355, p-value = 0.0001487
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho
```

```
data:  Male$Prior_Arrest_Episodes_Felony and Male$Race
S = 1.8924e+12, p-value = 0.0001484
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.02519452
```

In [78]:
```
wilcox.test(Prior_Arrest_Episodes_Felony ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Arrest_Episodes_Felony, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Felony by Recidivism_Within_3years
W = 869366, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Arrest_Episodes_Felony and Female$Recidivism_Within_3years
S = 3908319286, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2617607
```

In [79]:
```
wilcox.test(Prior_Arrest_Episodes_Felony ~ Race, data=Female)
cor.test(Female$Prior_Arrest_Episodes_Felony, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Felony by Race
W = 1117905, p-value = 0.6771
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Arrest_Episodes_Felony and Female$Race
S = 5254927796, p-value = 0.6772
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
0.007400947
```

Is there a correlation between Prior_Arrest_Episodes_Misd and recidivism and/or race for everyone, or just males, or just females?

In [80]:
```
wilcox.test(Prior_Arrest_Episodes_Misd ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prior_Arrest_Episodes_Misd, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Misd by Recidivism_Within_3years
W = 64852604, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Arrest_Episodes_Misd and NIJ$Recidivism_Within_3years
S = 2.3616e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1782683
```

In [81]:
```
wilcox.test(Prior_Arrest_Episodes_Misd ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Arrest_Episodes_Misd, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Misd by Recidivism_Within_3years
W = 50501961, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Arrest_Episodes_Misd and Male$Recidivism_Within_3years
S = 1.6288e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1609577
```

In [82]:
```
wilcox.test(Prior_Arrest_Episodes_Misd ~ Race, data=Male)
cor.test(Male$Prior_Arrest_Episodes_Misd, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Misd by Race
W = 54645267, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Arrest_Episodes_Misd and Male$Race
S = 1.7596e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.09360806
```

In [83]:
```
wilcox.test(Prior_Arrest_Episodes_Misd ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Arrest_Episodes_Misd, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FA
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Misd by Recidivism_Within_3years
W = 848096, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Arrest_Episodes_Misd and Female$Recidivism_Within_3years
S = 3816603155, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2790849
```

In [84]:
```
wilcox.test(Prior_Arrest_Episodes_Misd ~ Race, data=Female)
cor.test(Female$Prior_Arrest_Episodes_Misd, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Misd by Race
W = 1079116, p-value = 0.04167
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Arrest_Episodes_Misd and Female$Race
S = 5102467471, p-value = 0.04165
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.03619905
```

Is there a correlation between Prior_Arrest_Episodes_Violent and recidivism and/or race for everyone, or just males, or just females?

In [85]:
```
wilcox.test(Prior_Arrest_Episodes_Violent ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prior_Arrest_Episodes_Violent, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Violent by Recidivism_Within_3years
W = 75588806, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Arrest_Episodes_Violent and NIJ$Recidivism_Within_3years
S = 2.6864e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.06523665
```

In [86]:
```
wilcox.test(Prior_Arrest_Episodes_Violent ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Arrest_Episodes_Violent, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FAL
```

```
        Wilcoxon rank sum test with continuity correction
```

```
data:  Prior_Arrest_Episodes_Violent by Recidivism_Within_3years
W = 58131862, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Arrest_Episodes_Violent and Male$Recidivism_Within_3years
S = 1.8338e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.05537433
```

In [87]:
```
wilcox.test(Prior_Arrest_Episodes_Violent ~ Race, data=Male)
cor.test(Male$Prior_Arrest_Episodes_Violent, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Violent by Race
W = 68920160, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Arrest_Episodes_Violent and Male$Race
S = 2.1569e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1110488
```

In [88]:
```
wilcox.test(Prior_Arrest_Episodes_Violent ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Arrest_Episodes_Violent, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exac
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Violent by Recidivism_Within_3years
W = 1198582, p-value = 0.04395
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Arrest_Episodes_Violent and Female$Recidivism_Within_3years
S = 5104560413, p-value = 0.04393
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.03580372
```

In [89]:
```
wilcox.test(Prior_Arrest_Episodes_Violent ~ Race, data=Female)
cor.test(Female$Prior_Arrest_Episodes_Violent, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Violent by Race
W = 1383962, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Arrest_Episodes_Violent and Female$Race
S = 6422727039, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.2131837
```

Is there a correlation between Prior_Arrest_Episodes_Property and recidivism and/or race for everyone, or just males, or just females?

In [90]:
```
wilcox.test(Prior_Arrest_Episodes_Property ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prior_Arrest_Episodes_Property, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FAL
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Property by Recidivism_Within_3years
W = 64485580, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Arrest_Episodes_Property and NIJ$Recidivism_Within_3years
S = 2.351e+12, p-value < 2.2e-16
```

```
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1819637
```

In [91]:
```
wilcox.test(Prior_Arrest_Episodes_Property ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Arrest_Episodes_Property, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=F
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Property by Recidivism_Within_3years
W = 49017653, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Arrest_Episodes_Property and Male$Recidivism_Within_3years
S = 1.5891e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1814225
```

In [92]:
```
wilcox.test(Prior_Arrest_Episodes_Property ~ Race, data=Male)
cor.test(Male$Prior_Arrest_Episodes_Property, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Property by Race
W = 53947021, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Arrest_Episodes_Property and Male$Race
S = 1.7408e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.103253
```

In [93]:
```
wilcox.test(Prior_Arrest_Episodes_Property ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Arrest_Episodes_Property, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exa
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Property by Recidivism_Within_3years
W = 915387, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Arrest_Episodes_Property and Female$Recidivism_Within_3years
S = 4062578159, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2326229
```

In [94]:
```
wilcox.test(Prior_Arrest_Episodes_Property ~ Race, data=Female)
cor.test(Female$Prior_Arrest_Episodes_Property, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Property by Race
W = 1206106, p-value = 0.001075
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Arrest_Episodes_Property and Female$Race
S = 5601793500, p-value = 0.001067
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.05811824
```

Is there a correlation between Prior_Arrest_Episodes_Drug and recidivism and/or race for everyone, or just males, or just females?

In [95]:
```
wilcox.test(Prior_Arrest_Episodes_Drug ~ Recidivism_Within_3years, data=NIJ)
```

```
cor.test(NIJ$Prior_Arrest_Episodes_Drug, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Drug by Recidivism_Within_3years
W = 73955160, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Arrest_Episodes_Drug and NIJ$Recidivism_Within_3years
S = 2.6418e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.08076391
```

In [96]:
```
wilcox.test(Prior_Arrest_Episodes_Drug ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Arrest_Episodes_Drug, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Drug by Recidivism_Within_3years
W = 56888106, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Arrest_Episodes_Drug and Male$Recidivism_Within_3years
S = 1.8028e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.07135533
```

In [97]:
```
wilcox.test(Prior_Arrest_Episodes_Drug ~ Race, data=Male)
cor.test(Male$Prior_Arrest_Episodes_Drug, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Drug by Race
W = 62377202, p-value = 0.01901
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Arrest_Episodes_Drug and Male$Race
S = 1.9715e+12, p-value = 0.01901
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.01557807
```

In [98]:
```
wilcox.test(Prior_Arrest_Episodes_Drug ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Arrest_Episodes_Drug, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=Fa
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Drug by Recidivism_Within_3years
W = 1094805, p-value = 1.877e-09
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Arrest_Episodes_Drug and Female$Recidivism_Within_3years
S = 4728813189, p-value = 1.702e-09
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1067783
```

In [99]:
```
wilcox.test(Prior_Arrest_Episodes_Drug ~ Race, data=Female)
cor.test(Female$Prior_Arrest_Episodes_Drug, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_Drug by Race
W = 756876, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Arrest_Episodes_Drug and Female$Race
```

```
S = 3814879926, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2794104
```

Is there a correlation between Prior_Arrest_Episodes_PPViolationCharges and recidivism and/or race for everyone, or just males, or just females?

In [100…]
```
wilcox.test(Prior_Arrest_Episodes_PPViolationCharges ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prior_Arrest_Episodes_PPViolationCharges, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs',
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_PPViolationCharges by Recidivism_Within_3years
W = 60091696, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Arrest_Episodes_PPViolationCharges and NIJ$Recidivism_Within_3years
S = 2.2152e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2292122
```

In [101…]
```
wilcox.test(Prior_Arrest_Episodes_PPViolationCharges ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Arrest_Episodes_PPViolationCharges, Male$Recidivism_Within_3years, method='spearman', use='complete.obs
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_PPViolationCharges by Recidivism_Within_3years
W = 46411974, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Arrest_Episodes_PPViolationCharges and Male$Recidivism_Within_3years
S = 1.5179e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2180697
```

In [102…]
```
wilcox.test(Prior_Arrest_Episodes_PPViolationCharges ~ Race, data=Male)
cor.test(Male$Prior_Arrest_Episodes_PPViolationCharges, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_PPViolationCharges by Race
W = 56798607, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Arrest_Episodes_PPViolationCharges and Male$Race
S = 1.8188e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.06311263
```

In [103…]
```
wilcox.test(Prior_Arrest_Episodes_PPViolationCharges ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Arrest_Episodes_PPViolationCharges, Female$Recidivism_Within_3years, method='spearman', use='complete
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_PPViolationCharges by Recidivism_Within_3years
W = 815493, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Arrest_Episodes_PPViolationCharges and Female$Recidivism_Within_3years
S = 3689494796, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
```

```
0.3030943
```

In [104…
```
wilcox.test(Prior_Arrest_Episodes_PPViolationCharges ~ Race, data=Female)
cor.test(Female$Prior_Arrest_Episodes_PPViolationCharges, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_PPViolationCharges by Race
W = 1038206, p-value = 0.0001738
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Arrest_Episodes_PPViolationCharges and Female$Race
S = 4940861570, p-value = 0.0001715
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.06672466
```

Is there a correlation between Prior_Arrest_Episodes_DVCharges and recidivism and/or race for everyone, or just males, or just females?

In [105…
```
wilcox.test(Prior_Arrest_Episodes_DVCharges ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prior_Arrest_Episodes_DVCharges, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FA
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_DVCharges by Recidivism_Within_3years
W = 77416886, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Arrest_Episodes_DVCharges and NIJ$Recidivism_Within_3years
S = 2.6842e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.06599674
```

In [106…
```
wilcox.test(Prior_Arrest_Episodes_DVCharges ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Arrest_Episodes_DVCharges, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_DVCharges by Recidivism_Within_3years
W = 59019107, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Arrest_Episodes_DVCharges and Male$Recidivism_Within_3years
S = 1.8217e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.06158075
```

In [107…
```
wilcox.test(Prior_Arrest_Episodes_DVCharges ~ Race, data=Male)
cor.test(Male$Prior_Arrest_Episodes_DVCharges, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_DVCharges by Race
W = 58806456, p-value = 5.683e-15
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Arrest_Episodes_DVCharges and Male$Race
S = 1.8406e+12, p-value = 5.461e-15
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.05187982
```

In [108…
```
wilcox.test(Prior_Arrest_Episodes_DVCharges ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Arrest_Episodes_DVCharges, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', ex
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_DVCharges by Recidivism_Within_3years
W = 1204388, p-value = 0.003319
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Arrest_Episodes_DVCharges and Female$Recidivism_Within_3years
S = 5017808186, p-value = 0.003304
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.05219028
```

In [109…

```
wilcox.test(Prior_Arrest_Episodes_DVCharges ~ Race, data=Female)
cor.test(Female$Prior_Arrest_Episodes_DVCharges, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_DVCharges by Race
W = 1111090, p-value = 0.1854
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Arrest_Episodes_DVCharges and Female$Race
S = 5169505323, p-value = 0.1854
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.02353633
```

Is there a correlation between Prior_Arrest_Episodes_GunCharges and recidivism and/or race for everyone, or just males, or just females?

In [110…

```
wilcox.test(Prior_Arrest_Episodes_GunCharges ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prior_Arrest_Episodes_GunCharges, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=F
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_GunCharges by Recidivism_Within_3years
W = 78283810, p-value = 2.252e-12
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Arrest_Episodes_GunCharges and NIJ$Recidivism_Within_3years
S = 2.7484e+12, p-value = 2.202e-12
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.04366299
```

In [111…

```
wilcox.test(Prior_Arrest_Episodes_GunCharges ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Arrest_Episodes_GunCharges, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_GunCharges by Recidivism_Within_3years
W = 59908620, p-value = 5.836e-08
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Arrest_Episodes_GunCharges and Male$Recidivism_Within_3years
S = 1.8713e+12, p-value = 5.785e-08
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.03602491
```

In [112…

```
wilcox.test(Prior_Arrest_Episodes_GunCharges ~ Race, data=Male)
cor.test(Male$Prior_Arrest_Episodes_GunCharges, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_GunCharges by Race
W = 67184610, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho
```

```
data:  Male$Prior_Arrest_Episodes_GunCharges and Male$Race
S = 2.1433e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1040821
```

In [113…

```
wilcox.test(Prior_Arrest_Episodes_GunCharges ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Arrest_Episodes_GunCharges, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', e
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_GunCharges by Recidivism_Within_3years
W = 1257350, p-value = 0.3023
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Arrest_Episodes_GunCharges and Female$Recidivism_Within_3years
S = 5391168898, p-value = 0.3023
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.01833352
```

In [114…

```
wilcox.test(Prior_Arrest_Episodes_GunCharges ~ Race, data=Female)
cor.test(Female$Prior_Arrest_Episodes_GunCharges, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Arrest_Episodes_GunCharges by Race
W = 1159255, p-value = 0.01292
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Arrest_Episodes_GunCharges and Female$Race
S = 5.528e+09, p-value = 0.0129
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.04418063
```

Is there a correlation between Prior_Conviction_Episodes_Felony and recidivism and/or race for everyone, or just males, or just females?

In [115…

```
wilcox.test(Prior_Conviction_Episodes_Felony ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prior_Conviction_Episodes_Felony, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=F
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Felony by Recidivism_Within_3years
W = 71820576, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Conviction_Episodes_Felony and NIJ$Recidivism_Within_3years
S = 2.5724e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.104915
```

In [116…

```
wilcox.test(Prior_Conviction_Episodes_Felony ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Conviction_Episodes_Felony, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Felony by Recidivism_Within_3years
W = 55387404, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Conviction_Episodes_Felony and Male$Recidivism_Within_3years
S = 1.7596e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.09356382
```

In [117...]
```
wilcox.test(Prior_Conviction_Episodes_Felony ~ Race, data=Male)
cor.test(Male$Prior_Conviction_Episodes_Felony, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Felony by Race
W = 59036107, p-value = 1.428e-06
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Conviction_Episodes_Felony and Male$Race
S = 1.8791e+12, p-value = 1.42e-06
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.03202208
```

In [118...]
```
wilcox.test(Prior_Conviction_Episodes_Felony ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Conviction_Episodes_Felony, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', e
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Felony by Recidivism_Within_3years
W = 1009364, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Conviction_Episodes_Felony and Female$Recidivism_Within_3years
S = 4400670768, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1687609
```

In [119...]
```
wilcox.test(Prior_Conviction_Episodes_Felony ~ Race, data=Female)
cor.test(Female$Prior_Conviction_Episodes_Felony, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Felony by Race
W = 1189712, p-value = 0.008634
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Conviction_Episodes_Felony and Female$Race
S = 5541207330, p-value = 0.008613
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.04667416
```

Is there a correlation between Prior_Conviction_Episodes_Misd and recidivism and/or race for everyone, or just males, or just females?

In [120...]
```
wilcox.test(Prior_Conviction_Episodes_Misd ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prior_Conviction_Episodes_Misd, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FAL
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Misd by Recidivism_Within_3years
W = 65279298, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Conviction_Episodes_Misd and NIJ$Recidivism_Within_3years
S = 2.3718e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1747248
```

In [121...]
```
wilcox.test(Prior_Conviction_Episodes_Misd ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Conviction_Episodes_Misd, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=F
```

```
        Wilcoxon rank sum test with continuity correction
```

```
data:  Prior_Conviction_Episodes_Misd by Recidivism_Within_3years
W = 50631872, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Conviction_Episodes_Misd and Male$Recidivism_Within_3years
S = 1.6309e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.159869
```

In [122…

```
wilcox.test(Prior_Conviction_Episodes_Misd ~ Race, data=Male)
cor.test(Male$Prior_Conviction_Episodes_Misd, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Misd by Race
W = 56344087, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Conviction_Episodes_Misd and Male$Race
S = 1.8055e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.06994302
```

In [123…

```
wilcox.test(Prior_Conviction_Episodes_Misd ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Conviction_Episodes_Misd, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exa
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Misd by Recidivism_Within_3years
W = 898976, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Conviction_Episodes_Misd and Female$Recidivism_Within_3years
S = 3984628403, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2473468
```

In [124…

```
wilcox.test(Prior_Conviction_Episodes_Misd ~ Race, data=Female)
cor.test(Female$Prior_Conviction_Episodes_Misd, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Misd by Race
W = 1074738, p-value = 0.024
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Conviction_Episodes_Misd and Female$Race
S = 5081735000, p-value = 0.02397
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.04011519
```

Is there a correlation between Prior_Conviction_Episodes_Viol and recidivism and/or race for everyone, or just males, or just females?

In [125…

```
wilcox.test(Prior_Conviction_Episodes_Viol ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prior_Conviction_Episodes_Viol, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FAL
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Viol by Recidivism_Within_3years
W = 77853423, p-value = 6.195e-14
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Conviction_Episodes_Viol and NIJ$Recidivism_Within_3years
```

```
S = 2.7397e+12, p-value = 6.014e-14
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.04668643
```

In [126…

```
wilcox.test(Prior_Conviction_Episodes_Viol ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Conviction_Episodes_Viol, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=F
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Viol by Recidivism_Within_3years
W = 59387315, p-value = 9.273e-11
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Conviction_Episodes_Viol and Male$Recidivism_Within_3years
S = 1.8577e+12, p-value = 9.104e-11
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.04302958
```

In [127…

```
wilcox.test(Prior_Conviction_Episodes_Viol ~ Race, data=Male)
cor.test(Male$Prior_Conviction_Episodes_Viol, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Viol by Race
W = 66492996, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Conviction_Episodes_Viol and Male$Race
S = 2.1113e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.08757594
```

In [128…

```
wilcox.test(Prior_Conviction_Episodes_Viol ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Conviction_Episodes_Viol, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exa
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Viol by Recidivism_Within_3years
W = 1255835, p-value = 0.4866
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Conviction_Episodes_Viol and Female$Recidivism_Within_3years
S = 5359576495, p-value = 0.4866
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.01236606
```

In [129…

```
wilcox.test(Prior_Conviction_Episodes_Viol ~ Race, data=Female)
cor.test(Female$Prior_Conviction_Episodes_Viol, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Viol by Race
W = 1277894, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Conviction_Episodes_Viol and Female$Race
S = 6145087303, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.1607406
```

Is there a correlation between Prior_Conviction_Episodes_Prop and recidivism and/or race for everyone, or just males, or just females?

In [130…

```
wilcox.test(Prior_Conviction_Episodes_Prop ~ Recidivism_Within_3years, data=NIJ)
```

```
cor.test(NIJ$Prior_Conviction_Episodes_Prop, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALS
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Prop by Recidivism_Within_3years
W = 66961644, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Conviction_Episodes_Prop and NIJ$Recidivism_Within_3years
S = 2.4114e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1609443
```

In [131…

```
wilcox.test(Prior_Conviction_Episodes_Prop ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Conviction_Episodes_Prop, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=F
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Prop by Recidivism_Within_3years
W = 51172761, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Conviction_Episodes_Prop and Male$Recidivism_Within_3years
S = 1.6367e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1568742
```

In [132…

```
wilcox.test(Prior_Conviction_Episodes_Prop ~ Race, data=Male)
cor.test(Male$Prior_Conviction_Episodes_Prop, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Prop by Race
W = 54186755, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Conviction_Episodes_Prop and Male$Race
S = 1.7402e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1036055
```

In [133…

```
wilcox.test(Prior_Conviction_Episodes_Prop ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Conviction_Episodes_Prop, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exa
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Prop by Recidivism_Within_3years
W = 923958, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Conviction_Episodes_Prop and Female$Recidivism_Within_3years
S = 4065717970, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2320298
```

In [134…

```
wilcox.test(Prior_Conviction_Episodes_Prop ~ Race, data=Female)
cor.test(Female$Prior_Conviction_Episodes_Prop, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Prop by Race
W = 1224242, p-value = 3.679e-05
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho
```

```
data:  Female$Prior_Conviction_Episodes_Prop and Female$Race
S = 5682390620, p-value = 3.606e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.07334216
```

Is there a correlation between Prior_Conviction_Episodes_Drug and recidivism and/or race for everyone, or just males, or just females?

In [135…
```
wilcox.test(Prior_Conviction_Episodes_Drug ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prior_Conviction_Episodes_Drug, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FAL
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Drug by Recidivism_Within_3years
W = 75714134, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Conviction_Episodes_Drug and NIJ$Recidivism_Within_3years
S = 2.6859e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.06542195
```

In [136…
```
wilcox.test(Prior_Conviction_Episodes_Drug ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Conviction_Episodes_Drug, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=Fa
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Drug by Recidivism_Within_3years
W = 57978966, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Conviction_Episodes_Drug and Male$Recidivism_Within_3years
S = 1.8263e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.05922146
```

In [137…
```
wilcox.test(Prior_Conviction_Episodes_Drug ~ Race, data=Male)
cor.test(Male$Prior_Conviction_Episodes_Drug, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Drug by Race
W = 62162406, p-value = 0.04593
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Conviction_Episodes_Drug and Male$Race
S = 1.967e+12, p-value = 0.04592
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.01325808
```

In [138…
```
wilcox.test(Prior_Conviction_Episodes_Drug ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Conviction_Episodes_Drug, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exa
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Drug by Recidivism_Within_3years
W = 1142982, p-value = 1.411e-05
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Conviction_Episodes_Drug and Female$Recidivism_Within_3years
S = 4885557823, p-value = 1.376e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
```

```
      rho
0.07717094
```

In [139…

```
wilcox.test(Prior_Conviction_Episodes_Drug ~ Race, data=Female)
cor.test(Female$Prior_Conviction_Episodes_Drug, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_Drug by Race
W = 835706, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Conviction_Episodes_Drug and Female$Race
S = 4049473944, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2350981
```

Is there a correlation between Prior_Conviction_Episodes_PPViolationCharges and recidivism and/or race for everyone, or just males, or just females?

In [140…

```
wilcox.test(Prior_Conviction_Episodes_PPViolationCharges ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prior_Conviction_Episodes_PPViolationCharges, NIJ$Recidivism_Within_3years, method='spearman', use='complete.ol
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_PPViolationCharges by Recidivism_Within_3years
W = 74057633, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Conviction_Episodes_PPViolationCharges and NIJ$Recidivism_Within_3years
S = 2.5991e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.09561764
```

In [141…

```
wilcox.test(Prior_Conviction_Episodes_PPViolationCharges ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Conviction_Episodes_PPViolationCharges, Male$Recidivism_Within_3years, method='spearman', use='complete
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_PPViolationCharges by Recidivism_Within_3years
W = 56720589, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Conviction_Episodes_PPViolationCharges and Male$Recidivism_Within_3years
S = 1.7698e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.08832585
```

In [142…

```
wilcox.test(Prior_Conviction_Episodes_PPViolationCharges ~ Race, data=Male)
cor.test(Male$Prior_Conviction_Episodes_PPViolationCharges, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_PPViolationCharges by Race
W = 58298834, p-value = 3.254e-14
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Conviction_Episodes_PPViolationCharges and Male$Race
S = 1.8434e+12, p-value = 3.141e-14
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.05039863
```

In [143…

```
wilcox.test(Prior_Conviction_Episodes_PPViolationCharges ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Conviction_Episodes_PPViolationCharges, Female$Recidivism_Within_3years, method='spearman', use='comp
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_PPViolationCharges by Recidivism_Within_3years
W = 1086266, p-value = 1.198e-14
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Conviction_Episodes_PPViolationCharges and Female$Recidivism_Within_3years
S = 4568091239, p-value = 9.108e-15
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1371369
```

In [144…

```
wilcox.test(Prior_Conviction_Episodes_PPViolationCharges ~ Race, data=Female)
cor.test(Female$Prior_Conviction_Episodes_PPViolationCharges, Female$Race, method='spearman', use='complete.obs', exact=FAL
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_PPViolationCharges by Race
W = 1091052, p-value = 0.05737
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Conviction_Episodes_PPViolationCharges and Female$Race
S = 5115290797, p-value = 0.05735
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.03377687
```

Is there a correlation between Prior_Conviction_Episodes_DomesticViolenceCharges and recidivism and/or race for everyone, or just males, or just females?

In [145…

```
wilcox.test(Prior_Conviction_Episodes_DomesticViolenceCharges ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prior_Conviction_Episodes_DomesticViolenceCharges, NIJ$Recidivism_Within_3years, method='spearman', use='compl
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_DomesticViolenceCharges by Recidivism_Within_3years
W = 78779560, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Conviction_Episodes_DomesticViolenceCharges and NIJ$Recidivism_Within_3years
S = 2.7034e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.05932661
```

In [146…

```
wilcox.test(Prior_Conviction_Episodes_DomesticViolenceCharges ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Conviction_Episodes_DomesticViolenceCharges, Male$Recidivism_Within_3years, method='spearman', use='com
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_DomesticViolenceCharges by Recidivism_Within_3years
W = 59933661, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Conviction_Episodes_DomesticViolenceCharges and Male$Recidivism_Within_3years
S = 1.8312e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.05671442
```

In [147…

```
wilcox.test(Prior_Conviction_Episodes_DomesticViolenceCharges ~ Race, data=Male)
cor.test(Male$Prior_Conviction_Episodes_DomesticViolenceCharges, Male$Race, method='spearman', use='complete.obs', exact=FA
```

```
        Wilcoxon rank sum test with continuity correction
```

```
data:  Prior_Conviction_Episodes_DomesticViolenceCharges by Race
W = 60656927, p-value = 0.009016
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Conviction_Episodes_DomesticViolenceCharges and Male$Race
S = 1.9076e+12, p-value = 0.009013
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.01734553
```

In [148…

```
wilcox.test(Prior_Conviction_Episodes_DomesticViolenceCharges ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Conviction_Episodes_DomesticViolenceCharges, Female$Recidivism_Within_3years, method='spearman', use=
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_DomesticViolenceCharges by Recidivism_Within_3years
W = 1233200, p-value = 0.1833
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Conviction_Episodes_DomesticViolenceCharges and Female$Recidivism_Within_3years
S = 5168916907, p-value = 0.1834
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.02364747
```

In [149…

```
wilcox.test(Prior_Conviction_Episodes_DomesticViolenceCharges ~ Race, data=Female)
cor.test(Female$Prior_Conviction_Episodes_DomesticViolenceCharges, Female$Race, method='spearman', use='complete.obs', exac
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_DomesticViolenceCharges by Race
W = 1119226, p-value = 0.245
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Conviction_Episodes_DomesticViolenceCharges and Female$Race
S = 5184708393, p-value = 0.245
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.02066463
```

Is there a correlation between Prior_Conviction_Episodes_GunCharges and recidivism and/or race for everyone, or just males, or just females?

In [150…

```
wilcox.test(Prior_Conviction_Episodes_GunCharges ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prior_Conviction_Episodes_GunCharges, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exa
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_GunCharges by Recidivism_Within_3years
W = 79728950, p-value = 7.358e-07
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Conviction_Episodes_GunCharges and NIJ$Recidivism_Within_3years
S = 2.7854e+12, p-value = 7.319e-07
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.0308074
```

In [151…

```
wilcox.test(Prior_Conviction_Episodes_GunCharges ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Conviction_Episodes_GunCharges, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', e
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_GunCharges by Recidivism_Within_3years
W = 60900371, p-value = 0.0003296
alternative hypothesis: true location shift is not equal to 0
```

```
        Spearman's rank correlation rho

data:  Male$Prior_Conviction_Episodes_GunCharges and Male$Recidivism_Within_3years
S = 1.895e+12, p-value = 0.0003291
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.02385081
```

In [152…

```
wilcox.test(Prior_Conviction_Episodes_GunCharges ~ Race, data=Male)
cor.test(Male$Prior_Conviction_Episodes_GunCharges, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_GunCharges by Race
W = 63864555, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Conviction_Episodes_GunCharges and Male$Race
S = 2.0545e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.05830988
```

In [153…

```
wilcox.test(Prior_Conviction_Episodes_GunCharges ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Conviction_Episodes_GunCharges, Female$Recidivism_Within_3years, method='spearman', use='complete.obs
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_GunCharges by Recidivism_Within_3years
W = 1245895, p-value = 0.8025
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Conviction_Episodes_GunCharges and Female$Recidivism_Within_3years
S = 5317644095, p-value = 0.8025
alternative hypothesis: true rho is not equal to 0
sample estimates:
         rho
-0.004445484
```

In [154…

```
wilcox.test(Prior_Conviction_Episodes_GunCharges ~ Race, data=Female)
cor.test(Female$Prior_Conviction_Episodes_GunCharges, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Conviction_Episodes_GunCharges by Race
W = 1135827, p-value = 0.3426
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Conviction_Episodes_GunCharges and Female$Race
S = 5383417056, p-value = 0.3426
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.01686929
```

Is there a correlation between Prior_Revocations_Parole and recidivism and/or race for everyone, or just males, or just females?

In [155…

```
wilcox.test(Prior_Revocations_Parole ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prior_Revocations_Parole, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Revocations_Parole by Recidivism_Within_3years
W = 78663156, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Revocations_Parole and NIJ$Recidivism_Within_3years
S = 2.7083e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.05761161
```

In [156…
```
wilcox.test(Prior_Revocations_Parole ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Revocations_Parole, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Revocations_Parole by Recidivism_Within_3years
W = 60002865, p-value = 1.504e-14
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Revocations_Parole and Male$Recidivism_Within_3years
S = 1.8422e+12, p-value = 1.449e-14
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.05105864
```

In [157…
```
wilcox.test(Prior_Revocations_Parole ~ Race, data=Male)
cor.test(Male$Prior_Revocations_Parole, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Revocations_Parole by Race
W = 62678036, p-value = 3.269e-08
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Revocations_Parole and Male$Race
S = 2.0125e+12, p-value = 3.238e-08
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.03670668
```

In [158…
```
wilcox.test(Prior_Revocations_Parole ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Revocations_Parole, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALS
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Revocations_Parole by Recidivism_Within_3years
W = 1214225, p-value = 0.0006729
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Revocations_Parole and Female$Recidivism_Within_3years
S = 4974165013, p-value = 0.0006671
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.060434
```

In [159…
```
wilcox.test(Prior_Revocations_Parole ~ Race, data=Female)
cor.test(Female$Prior_Revocations_Parole, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Revocations_Parole by Race
W = 1132660, p-value = 0.5715
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Revocations_Parole and Female$Race
S = 5347349991, p-value = 0.5716
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.01005661
```

Is there a correlation between Prior_Revocations_Probation and recidivism and/or race for everyone, or just males, or just females?

In [160…
```
wilcox.test(Prior_Revocations_Probation ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Prior_Revocations_Probation, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Revocations_Probation by Recidivism_Within_3years
W = 79186288, p-value = 3.507e-10
```

```
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Prior_Revocations_Probation and NIJ$Recidivism_Within_3years
S = 2.7617e+12, p-value = 3.457e-10
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.03903791
```

In [161…

```
wilcox.test(Prior_Revocations_Probation ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Prior_Revocations_Probation, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Revocations_Probation by Recidivism_Within_3years
W = 60371791, p-value = 5.46e-08
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Revocations_Probation and Male$Recidivism_Within_3years
S = 1.8712e+12, p-value = 5.412e-08
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.03610383
```

In [162…

```
wilcox.test(Prior_Revocations_Probation ~ Race, data=Male)
cor.test(Male$Prior_Revocations_Probation, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Revocations_Probation by Race
W = 58411387, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Prior_Revocations_Probation and Male$Race
S = 1.8146e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.06523093
```

In [163…

```
wilcox.test(Prior_Revocations_Probation ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Prior_Revocations_Probation, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=F
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Revocations_Probation by Recidivism_Within_3years
W = 1172974, p-value = 1.953e-05
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Revocations_Probation and Female$Recidivism_Within_3years
S = 4892332009, p-value = 1.908e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.07589137
```

In [164…

```
wilcox.test(Prior_Revocations_Probation ~ Race, data=Female)
cor.test(Female$Prior_Revocations_Probation, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Prior_Revocations_Probation by Race
W = 1075613, p-value = 0.0009031
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Prior_Revocations_Probation and Female$Race
S = 4981816114, p-value = 0.0008962
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.05898879
```

Is there a correlation between Condition_MH_SA and recidivism and/or race for everyone, or just males, or just females?

In [165…
```
wilcox.test(Condition_MH_SA ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Condition_MH_SA, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Condition_MH_SA by Recidivism_Within_3years
W = 72554760, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Condition_MH_SA and NIJ$Recidivism_Within_3years
S = 2.5467e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1138556
```

In [166…
```
wilcox.test(Condition_MH_SA ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Condition_MH_SA, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Condition_MH_SA by Recidivism_Within_3years
W = 54645947, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Condition_MH_SA and Male$Recidivism_Within_3years
S = 1.7065e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1209539
```

In [167…
```
wilcox.test(Condition_MH_SA ~ Race, data=Male)
cor.test(Male$Condition_MH_SA, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Condition_MH_SA by Race
W = 53392083, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Condition_MH_SA and Male$Race
S = 1.687e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1309923
```

In [168…
```
wilcox.test(Condition_MH_SA ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Condition_MH_SA, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Condition_MH_SA by Recidivism_Within_3years
W = 1090279, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Condition_MH_SA and Female$Recidivism_Within_3years
S = 4508583709, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1483773
```

In [169…
```
wilcox.test(Condition_MH_SA ~ Race, data=Female)
cor.test(Female$Condition_MH_SA, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Condition_MH_SA by Race
W = 872704, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho
```

```
data:  Female$Condition_MH_SA and Female$Race
S = 3921860347, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.259203
```

Is there a correlation between Condition_Cog_Ed and recidivism and/or race for everyone, or just males, or just females?

In [170…
```
wilcox.test(Condition_Cog_Ed ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Condition_Cog_Ed, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Condition_Cog_Ed by Recidivism_Within_3years
W = 78386279, p-value = 1.659e-09
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Condition_Cog_Ed and NIJ$Recidivism_Within_3years
S = 2.7661e+12, p-value = 1.64e-09
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.0375043
```

In [171…
```
wilcox.test(Condition_Cog_Ed ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Condition_Cog_Ed, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Condition_Cog_Ed by Recidivism_Within_3years
W = 58850274, p-value = 6.813e-14
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Condition_Cog_Ed and Male$Recidivism_Within_3years
S = 1.8447e+12, p-value = 6.588e-14
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.04975858
```

In [172…
```
wilcox.test(Condition_Cog_Ed ~ Race, data=Male)
cor.test(Male$Condition_Cog_Ed, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Condition_Cog_Ed by Race
W = 63723016, p-value = 3.217e-09
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Condition_Cog_Ed and Male$Race
S = 2.0176e+12, p-value = 3.177e-09
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.03932174
```

In [173…
```
wilcox.test(Condition_Cog_Ed ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Condition_Cog_Ed, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Condition_Cog_Ed by Recidivism_Within_3years
W = 1280836, p-value = 0.09404
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Condition_Cog_Ed and Female$Recidivism_Within_3years
S = 5451659984, p-value = 0.09404
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.02975964
```

In [174…
```
wilcox.test(Condition_Cog_Ed ~ Race, data=Female)
cor.test(Female$Condition_Cog_Ed, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
	Wilcoxon rank sum test with continuity correction

data:  Condition_Cog_Ed by Race
W = 1182118, p-value = 0.01032
alternative hypothesis: true location shift is not equal to 0
	Spearman's rank correlation rho

data:  Female$Condition_Cog_Ed and Female$Race
S = 5535429871, p-value = 0.0103
alternative hypothesis: true rho is not equal to 0
sample estimates:
	rho
-0.04558286
```

Is there a correlation between Condition_Other and recidivism and/or race for everyone, or just males, or just females?

In [175…
```
wilcox.test(Condition_Other ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Condition_Other, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
	Wilcoxon rank sum test with continuity correction

data:  Condition_Other by Recidivism_Within_3years
W = 80463756, p-value = 0.03843
alternative hypothesis: true location shift is not equal to 0
	Spearman's rank correlation rho

data:  NIJ$Condition_Other and NIJ$Recidivism_Within_3years
S = 2.8369e+12, p-value = 0.03842
alternative hypothesis: true rho is not equal to 0
sample estimates:
	rho
0.01288046
```

In [176…
```
wilcox.test(Condition_Other ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Condition_Other, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
	Wilcoxon rank sum test with continuity correction

data:  Condition_Other by Recidivism_Within_3years
W = 61307782, p-value = 0.09621
alternative hypothesis: true location shift is not equal to 0
	Spearman's rank correlation rho

data:  Male$Condition_Other and Male$Recidivism_Within_3years
S = 1.9198e+12, p-value = 0.09621
alternative hypothesis: true rho is not equal to 0
sample estimates:
	rho
0.01104915
```

In [177…
```
wilcox.test(Condition_Other ~ Race, data=Male)
cor.test(Male$Condition_Other, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
	Wilcoxon rank sum test with continuity correction

data:  Condition_Other by Race
W = 60103686, p-value = 0.002897
alternative hypothesis: true location shift is not equal to 0
	Spearman's rank correlation rho

data:  Male$Condition_Other and Male$Race
S = 1.9029e+12, p-value = 0.002895
alternative hypothesis: true rho is not equal to 0
sample estimates:
	rho
0.01978344
```

In [178…
```
wilcox.test(Condition_Other ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Condition_Other, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
	Wilcoxon rank sum test with continuity correction

data:  Condition_Other by Recidivism_Within_3years
W = 1265282, p-value = 0.2621
alternative hypothesis: true location shift is not equal to 0
	Spearman's rank correlation rho

data:  Female$Condition_Other and Female$Recidivism_Within_3years
```

```
S = 5399632843, p-value = 0.2621
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.01993227
```

In [179…]
```
wilcox.test(Condition_Other ~ Race, data=Female)
cor.test(Female$Condition_Other, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Condition_Other by Race
W = 1195148, p-value = 0.0002434
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Condition_Other and Female$Race
S = 5639328585, p-value = 0.0002405
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.06520821
```

Is there a correlation between Violations_ElectronicMonitoring and recidivism and/or race for everyone, or just males, or just females?

In [180…]
```
wilcox.test(Violations_ElectronicMonitoring ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Violations_ElectronicMonitoring, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FA
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_ElectronicMonitoring by Recidivism_Within_3years
W = 81621988, p-value = 0.5406
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Violations_ElectronicMonitoring and NIJ$Recidivism_Within_3years
S = 2.8849e+12, p-value = 0.5406
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.003807369
```

In [181…]
```
wilcox.test(Violations_ElectronicMonitoring ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Violations_ElectronicMonitoring, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_ElectronicMonitoring by Recidivism_Within_3years
W = 62240571, p-value = 0.2271
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Violations_ElectronicMonitoring and Male$Recidivism_Within_3years
S = 1.9569e+12, p-value = 0.2271
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.008022131
```

In [182…]
```
wilcox.test(Violations_ElectronicMonitoring ~ Race, data=Male)
cor.test(Male$Violations_ElectronicMonitoring, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_ElectronicMonitoring by Race
W = 63614504, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Violations_ElectronicMonitoring and Male$Race
S = 2.0745e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.06863688
```

In [183…]
```
wilcox.test(Violations_ElectronicMonitoring ~ Recidivism_Within_3years, data=Female)
```

```
cor.test(Female$Violations_ElectronicMonitoring, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', ex
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_ElectronicMonitoring by Recidivism_Within_3years
W = 1257222, p-value = 0.06578
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Violations_ElectronicMonitoring and Female$Recidivism_Within_3years
S = 5467231769, p-value = 0.06576
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.03270098
```

In [184…

```
wilcox.test(Violations_ElectronicMonitoring ~ Race, data=Female)
cor.test(Female$Violations_ElectronicMonitoring, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_ElectronicMonitoring by Race
W = 1157428, p-value = 2.502e-05
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Violations_ElectronicMonitoring and Female$Race
S = 5690662549, p-value = 2.447e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.07490464
```

Is there a correlation between Violations_Instruction and recidivism and/or race for everyone, or just males, or just females?

In [185…

```
wilcox.test(Violations_Instruction ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Violations_Instruction, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_Instruction by Recidivism_Within_3years
W = 77444129, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Violations_Instruction and NIJ$Recidivism_Within_3years
S = 2.689e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.06434993
```

In [186…

```
wilcox.test(Violations_Instruction ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Violations_Instruction, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_Instruction by Recidivism_Within_3years
W = 59161984, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Violations_Instruction and Male$Recidivism_Within_3years
S = 1.8286e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.05803821
```

In [187…

```
wilcox.test(Violations_Instruction ~ Race, data=Male)
cor.test(Male$Violations_Instruction, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_Instruction by Race
W = 63499836, p-value = 3.093e-12
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho
```

```
data:  Male$Violations_Instruction and Male$Race
S = 2.0312e+12, p-value = 3.017e-12
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.04631797
```

In [188…

```
wilcox.test(Violations_Instruction ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Violations_Instruction, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_Instruction by Recidivism_Within_3years
W = 1169968, p-value = 1.142e-06
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Violations_Instruction and Female$Recidivism_Within_3years
S = 4836329239, p-value = 1.097e-06
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.08646969
```

In [189…

```
wilcox.test(Violations_Instruction ~ Race, data=Female)
cor.test(Female$Violations_Instruction, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_Instruction by Race
W = 1101250, p-value = 0.06404
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Violations_Instruction and Female$Race
S = 5119866222, p-value = 0.06403
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.03291262
```

Is there a correlation between Violations_FailToReport and recidivism and/or race for everyone, or just males, or just females?

In [190…

```
wilcox.test(Violations_FailToReport ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Violations_FailToReport, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_FailToReport by Recidivism_Within_3years
W = 80130495, p-value = 9.9e-07
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Violations_FailToReport and NIJ$Recidivism_Within_3years
S = 2.7864e+12, p-value = 9.85e-07
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.03044632
```

In [191…

```
wilcox.test(Violations_FailToReport ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Violations_FailToReport, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_FailToReport by Recidivism_Within_3years
W = 61149200, p-value = 0.0002492
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Violations_FailToReport and Male$Recidivism_Within_3years
S = 1.894e+12, p-value = 0.0002487
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.02433072
```

In [192...]
```
wilcox.test(Violations_FailToReport ~ Race, data=Male)
cor.test(Male$Violations_FailToReport, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_FailToReport by Race
W = 61567598, p-value = 0.1866
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Violations_FailToReport and Male$Race
S = 1.9583e+12, p-value = 0.1866
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.008772355
```

In [193...]
```
wilcox.test(Violations_FailToReport ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Violations_FailToReport, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_FailToReport by Recidivism_Within_3years
W = 1199798, p-value = 9.268e-05
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Violations_FailToReport and Female$Recidivism_Within_3years
S = 4926313489, p-value = 9.122e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.06947264
```

In [194...]
```
wilcox.test(Violations_FailToReport ~ Race, data=Female)
cor.test(Female$Violations_FailToReport, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_FailToReport by Race
W = 1117725, p-value = 0.3377
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Violations_FailToReport and Female$Race
S = 5203902690, p-value = 0.3378
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.01703904
```

Is there a correlation between Violations_MoveWithoutPermission and recidivism and/or race for everyone, or just males, or just females?

In [195...]
```
wilcox.test(Violations_MoveWithoutPermission ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Violations_MoveWithoutPermission, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=F
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_MoveWithoutPermission by Recidivism_Within_3years
W = 79785276, p-value = 3.566e-07
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Violations_MoveWithoutPermission and NIJ$Recidivism_Within_3years
S = 2.7829e+12, p-value = 3.545e-07
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.03167298
```

In [196...]
```
wilcox.test(Violations_MoveWithoutPermission ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Violations_MoveWithoutPermission, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_MoveWithoutPermission by Recidivism_Within_3years
```

```
W = 60802650, p-value = 1.396e-05
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Violations_MoveWithoutPermission and Male$Recidivism_Within_3years
S = 1.8853e+12, p-value = 1.391e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.02885641
```

In [197…

```
wilcox.test(Violations_MoveWithoutPermission ~ Race, data=Male)
cor.test(Male$Violations_MoveWithoutPermission, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_MoveWithoutPermission by Race
W = 60795072, p-value = 0.07145
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Violations_MoveWithoutPermission and Male$Race
S = 1.918e+12, p-value = 0.07145
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.01197301
```

In [198…

```
wilcox.test(Violations_MoveWithoutPermission ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Violations_MoveWithoutPermission, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', e
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_MoveWithoutPermission by Recidivism_Within_3years
W = 1197555, p-value = 0.001439
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Violations_MoveWithoutPermission and Female$Recidivism_Within_3years
S = 4994273495, p-value = 0.00143
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.05663573
```

In [199…

```
wilcox.test(Violations_MoveWithoutPermission ~ Race, data=Female)
cor.test(Female$Violations_MoveWithoutPermission, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Violations_MoveWithoutPermission by Race
W = 1099176, p-value = 0.03681
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Violations_MoveWithoutPermission and Female$Race
S = 5097658027, p-value = 0.03678
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.03710751
```

Is there a correlation between Delinquency_Reports and recidivism and/or race for everyone, or just males, or just females?

In [200…

```
wilcox.test(Delinquency_Reports ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Delinquency_Reports, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Delinquency_Reports by Recidivism_Within_3years
W = 78229373, p-value = 3.869e-11
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Delinquency_Reports and NIJ$Recidivism_Within_3years
S = 2.7557e+12, p-value = 3.802e-11
alternative hypothesis: true rho is not equal to 0
sample estimates:
```

```
      rho
0.04111878
```

In [201…
```
wilcox.test(Delinquency_Reports ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Delinquency_Reports, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Delinquency_Reports by Recidivism_Within_3years
W = 60286503, p-value = 3.061e-05
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Delinquency_Reports and Male$Recidivism_Within_3years
S = 1.8875e+12, p-value = 3.052e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.02769022
```

In [202…
```
wilcox.test(Delinquency_Reports ~ Race, data=Male)
cor.test(Male$Delinquency_Reports, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Delinquency_Reports by Race
W = 60951532, p-value = 0.4194
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Delinquency_Reports and Male$Race
S = 1.9309e+12, p-value = 0.4195
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
0.005362688
```

In [203…
```
wilcox.test(Delinquency_Reports ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Delinquency_Reports, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Delinquency_Reports by Recidivism_Within_3years
W = 1131918, p-value = 9.61e-09
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Delinquency_Reports and Female$Recidivism_Within_3years
S = 4754276078, p-value = 8.865e-09
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1019686
```

In [204…
```
wilcox.test(Delinquency_Reports ~ Race, data=Female)
cor.test(Female$Delinquency_Reports, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Delinquency_Reports by Race
W = 1056483, p-value = 0.0001168
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Delinquency_Reports and Female$Race
S = 4931598209, p-value = 0.000115
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.06847441
```

Is there a correlation between Program_Attendances and recidivism and/or race for everyone, or just males, or just females?

In [205…
```
wilcox.test(Program_Attendances ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Program_Attendances, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Program_Attendances by Recidivism_Within_3years
```

```
W = 86704988, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Program_Attendances and NIJ$Recidivism_Within_3years
S = 3.046e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.05986484
```

In [206…

```
wilcox.test(Program_Attendances ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Program_Attendances, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Program_Attendances by Recidivism_Within_3years
W = 66285010, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Program_Attendances and Male$Recidivism_Within_3years
S = 2.0665e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.06450459
```

In [207…

```
wilcox.test(Program_Attendances ~ Race, data=Male)
cor.test(Male$Program_Attendances, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Program_Attendances by Race
W = 56476294, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Program_Attendances and Male$Race
S = 1.8014e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.0720632
```

In [208…

```
wilcox.test(Program_Attendances ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Program_Attendances, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Program_Attendances by Recidivism_Within_3years
W = 1270157, p-value = 0.268
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Program_Attendances and Female$Recidivism_Within_3years
S = 5398324963, p-value = 0.2681
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.01968523
```

In [209…

```
wilcox.test(Program_Attendances ~ Race, data=Female)
cor.test(Female$Program_Attendances, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Program_Attendances by Race
W = 884718, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Program_Attendances and Female$Race
S = 4.287e+09, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.190241
```

Is there a correlation between Program_UnexcusedAbsences and recidivism and/or race for everyone, or just males, or just females?

In [210…

```
wilcox.test(Program_UnexcusedAbsences ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Program_UnexcusedAbsences, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Program_UnexcusedAbsences by Recidivism_Within_3years
W = 77510396, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Program_UnexcusedAbsences and NIJ$Recidivism_Within_3years
S = 2.7014e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.06004243
```

In [211…

```
wilcox.test(Program_UnexcusedAbsences ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Program_UnexcusedAbsences, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Program_UnexcusedAbsences by Recidivism_Within_3years
W = 59407433, p-value = 3.851e-14
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Program_UnexcusedAbsences and Male$Recidivism_Within_3years
S = 1.8437e+12, p-value = 3.719e-14
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.05025331
```

In [212…

```
wilcox.test(Program_UnexcusedAbsences ~ Race, data=Male)
cor.test(Male$Program_UnexcusedAbsences, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Program_UnexcusedAbsences by Race
W = 63432572, p-value = 1.387e-10
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Program_UnexcusedAbsences and Male$Race
S = 2.024e+12, p-value = 1.362e-10
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.04262437
```

In [213…

```
wilcox.test(Program_UnexcusedAbsences ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Program_UnexcusedAbsences, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FA
```

```
        Wilcoxon rank sum test with continuity correction

data:  Program_UnexcusedAbsences by Recidivism_Within_3years
W = 1146698, p-value = 1.391e-09
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Program_UnexcusedAbsences and Female$Recidivism_Within_3years
S = 4724258046, p-value = 1.257e-09
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1076387
```

In [214…

```
wilcox.test(Program_UnexcusedAbsences ~ Race, data=Female)
cor.test(Female$Program_UnexcusedAbsences, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Program_UnexcusedAbsences by Race
W = 1113867, p-value = 0.3548
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho
```

```
data:  Female$Program_UnexcusedAbsences and Female$Race
S = 5.207e+09, p-value = 0.3549
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.01644606
```

Is there a correlation between Residence_Changes and recidivism and/or race for everyone, or just males, or just females?

In [215…
```
wilcox.test(Residence_Changes ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Residence_Changes, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Residence_Changes by Recidivism_Within_3years
W = 76768688, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Residence_Changes and NIJ$Recidivism_Within_3years
S = 2.72e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.05355539
```

In [216…
```
wilcox.test(Residence_Changes ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Residence_Changes, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Residence_Changes by Recidivism_Within_3years
W = 58515399, p-value = 8.532e-15
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Residence_Changes and Male$Recidivism_Within_3years
S = 1.8412e+12, p-value = 8.208e-15
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.05153858
```

In [217…
```
wilcox.test(Residence_Changes ~ Race, data=Male)
cor.test(Male$Residence_Changes, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Residence_Changes by Race
W = 58114046, p-value = 8.763e-13
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Residence_Changes and Male$Race
S = 1.8491e+12, p-value = 8.523e-13
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.04748188
```

In [218…
```
wilcox.test(Residence_Changes ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Residence_Changes, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Residence_Changes by Recidivism_Within_3years
W = 1138082, p-value = 8.412e-06
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Residence_Changes and Female$Recidivism_Within_3years
S = 4.875e+09, p-value = 8.181e-06
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.07916598
```

In [219…
```
wilcox.test(Residence_Changes ~ Race, data=Female)
cor.test(Female$Residence_Changes, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
	Wilcoxon rank sum test with continuity correction

data:  Residence_Changes by Race
W = 1110681, p-value = 0.4436
alternative hypothesis: true location shift is not equal to 0
	Spearman's rank correlation rho

data:  Female$Residence_Changes and Female$Race
S = 5.222e+09, p-value = 0.4437
alternative hypothesis: true rho is not equal to 0
sample estimates:
	rho
0.01361617
```

Is there a correlation between Avg_Days_per_DrugTest and recidivism and/or race for everyone, or just males, or just females?

In [220...
```
wilcox.test(Avg_Days_per_DrugTest ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Avg_Days_per_DrugTest, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
	Wilcoxon rank sum test with continuity correction

data:  Avg_Days_per_DrugTest by Recidivism_Within_3years
W = 82501010, p-value = 0.0762
alternative hypothesis: true location shift is not equal to 0
	Spearman's rank correlation rho

data:  NIJ$Avg_Days_per_DrugTest and NIJ$Recidivism_Within_3years
S = 2.9056e+12, p-value = 0.0762
alternative hypothesis: true rho is not equal to 0
sample estimates:
	rho
-0.01103187
```

In [221...
```
wilcox.test(Avg_Days_per_DrugTest ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Avg_Days_per_DrugTest, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
	Wilcoxon rank sum test with continuity correction

data:  Avg_Days_per_DrugTest by Recidivism_Within_3years
W = 63152875, p-value = 0.0135
alternative hypothesis: true location shift is not equal to 0
	Spearman's rank correlation rho

data:  Male$Avg_Days_per_DrugTest and Male$Recidivism_Within_3years
S = 1.9731e+12, p-value = 0.0135
alternative hypothesis: true rho is not equal to 0
sample estimates:
	rho
-0.0164076
```

In [222...
```
wilcox.test(Avg_Days_per_DrugTest ~ Race, data=Male)
cor.test(Male$Avg_Days_per_DrugTest, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
	Wilcoxon rank sum test with continuity correction

data:  Avg_Days_per_DrugTest by Race
W = 66885995, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
	Spearman's rank correlation rho

data:  Male$Avg_Days_per_DrugTest and Male$Race
S = 2.0927e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
	rho
-0.07797769
```

In [223...
```
wilcox.test(Avg_Days_per_DrugTest ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Avg_Days_per_DrugTest, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
	Wilcoxon rank sum test with continuity correction

data:  Avg_Days_per_DrugTest by Recidivism_Within_3years
W = 1239768, p-value = 0.8765
alternative hypothesis: true location shift is not equal to 0
	Spearman's rank correlation rho

data:  Female$Avg_Days_per_DrugTest and Female$Recidivism_Within_3years
```

```
S = 5279486704, p-value = 0.8765
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.002762034
```

In [224…]
```
wilcox.test(Avg_Days_per_DrugTest ~ Race, data=Female)
cor.test(Female$Avg_Days_per_DrugTest, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Avg_Days_per_DrugTest by Race
W = 1311614, p-value = 3.673e-14
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Avg_Days_per_DrugTest and Female$Race
S = 6006556687, p-value = 2.851e-14
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1345736
```

Is there a correlation between DrugTests_THC_Positive and recidivism and/or race for everyone, or just males, or just females?

In [225…]
```
wilcox.test(DrugTests_THC_Positive ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$DrugTests_THC_Positive, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_THC_Positive by Recidivism_Within_3years
W = 75487226, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$DrugTests_THC_Positive and NIJ$Recidivism_Within_3years
S = 2.6378e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.08214599
```

In [226…]
```
wilcox.test(DrugTests_THC_Positive ~ Recidivism_Within_3years, data=Male)
cor.test(Male$DrugTests_THC_Positive, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_THC_Positive by Recidivism_Within_3years
W = 57489348, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$DrugTests_THC_Positive and Male$Recidivism_Within_3years
S = 1.7889e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.07847997
```

In [227…]
```
wilcox.test(DrugTests_THC_Positive ~ Race, data=Male)
cor.test(Male$DrugTests_THC_Positive, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_THC_Positive by Race
W = 70385384, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$DrugTests_THC_Positive and Male$Race
S = 2.2531e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
-0.160628
```

In [228…]
```
wilcox.test(DrugTests_THC_Positive ~ Recidivism_Within_3years, data=Female)
cor.test(Female$DrugTests_THC_Positive, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_THC_Positive by Recidivism_Within_3years
W = 1226076, p-value = 0.2244
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$DrugTests_THC_Positive and Female$Recidivism_Within_3years
S = 5179808393, p-value = 0.2245
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.02159019
```

In [229…

```
wilcox.test(DrugTests_THC_Positive ~ Race, data=Female)
cor.test(Female$DrugTests_THC_Positive, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_THC_Positive by Race
W = 1197473, p-value = 5.107e-07
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$DrugTests_THC_Positive and Female$Race
S = 5766647806, p-value = 4.876e-07
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.08925743
```

Is there a correlation between DrugTests_Cocaine_Positive and recidivism and/or race for everyone, or just males, or just females?

In [230…

```
wilcox.test(DrugTests_Cocaine_Positive ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$DrugTests_Cocaine_Positive, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_Cocaine_Positive by Recidivism_Within_3years
W = 80963668, p-value = 0.06556
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$DrugTests_Cocaine_Positive and NIJ$Recidivism_Within_3years
S = 2.841e+12, p-value = 0.06556
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.01145671
```

In [231…

```
wilcox.test(DrugTests_Cocaine_Positive ~ Recidivism_Within_3years, data=Male)
cor.test(Male$DrugTests_Cocaine_Positive, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_Cocaine_Positive by Recidivism_Within_3years
W = 61613674, p-value = 0.1122
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$DrugTests_Cocaine_Positive and Male$Recidivism_Within_3years
S = 1.9208e+12, p-value = 0.1122
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.01055009
```

In [232…

```
wilcox.test(DrugTests_Cocaine_Positive ~ Race, data=Male)
cor.test(Male$DrugTests_Cocaine_Positive, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_Cocaine_Positive by Race
W = 65530882, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$DrugTests_Cocaine_Positive and Male$Race
```

```
S = 2.1904e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1283088
```

In [233…

```
wilcox.test(DrugTests_Cocaine_Positive ~ Recidivism_Within_3years, data=Female)
cor.test(Female$DrugTests_Cocaine_Positive, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=F
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_Cocaine_Positive by Recidivism_Within_3years
W = 1240552, p-value = 0.7663
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$DrugTests_Cocaine_Positive and Female$Recidivism_Within_3years
S = 5266141893, p-value = 0.7663
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
0.005282725
```

In [234…

```
wilcox.test(DrugTests_Cocaine_Positive ~ Race, data=Female)
cor.test(Female$DrugTests_Cocaine_Positive, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_Cocaine_Positive by Race
W = 1197177, p-value = 1.009e-11
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$DrugTests_Cocaine_Positive and Female$Race
S = 5934402294, p-value = 8.57e-12
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1209444
```

Is there a correlation between DrugTests_Meth_Positive and recidivism and/or race for everyone, or just males, or just females?

In [235…

```
wilcox.test(DrugTests_Meth_Positive ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$DrugTests_Meth_Positive, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_Meth_Positive by Recidivism_Within_3years
W = 79020430, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$DrugTests_Meth_Positive and NIJ$Recidivism_Within_3years
S = 2.7164e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.05479223
```

In [236…

```
wilcox.test(DrugTests_Meth_Positive ~ Recidivism_Within_3years, data=Male)
cor.test(Male$DrugTests_Meth_Positive, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_Meth_Positive by Recidivism_Within_3years
W = 60155968, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$DrugTests_Meth_Positive and Male$Recidivism_Within_3years
S = 1.8346e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.05496933
```

In [237…

```
wilcox.test(DrugTests_Meth_Positive ~ Race, data=Male)
cor.test(Male$DrugTests_Meth_Positive, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_Meth_Positive by Race
W = 52129654, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$DrugTests_Meth_Positive and Male$Race
S = 1.399e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.279359
```

In [238…]
```
wilcox.test(DrugTests_Meth_Positive ~ Recidivism_Within_3years, data=Female)
cor.test(Female$DrugTests_Meth_Positive, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_Meth_Positive by Recidivism_Within_3years
W = 1170252, p-value = 3.223e-07
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$DrugTests_Meth_Positive and Female$Recidivism_Within_3years
S = 4813322239, p-value = 3.067e-07
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.09081546
```

In [239…]
```
wilcox.test(DrugTests_Meth_Positive ~ Race, data=Female)
cor.test(Female$DrugTests_Meth_Positive, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_Meth_Positive by Race
W = 952752, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$DrugTests_Meth_Positive and Female$Race
S = 4.09e+09, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2274356
```

Is there a correlation between DrugTests_Other_Positive and recidivism and/or race for everyone, or just males, or just females?

In [240…]
```
wilcox.test(DrugTests_Other_Positive ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$DrugTests_Other_Positive, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_Other_Positive by Recidivism_Within_3years
W = 81321518, p-value = 0.5721
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$DrugTests_Other_Positive and NIJ$Recidivism_Within_3years
S = 2.8638e+12, p-value = 0.5721
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
0.003515093
```

In [241…]
```
wilcox.test(DrugTests_Other_Positive ~ Recidivism_Within_3years, data=Male)
cor.test(Male$DrugTests_Other_Positive, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_Other_Positive by Recidivism_Within_3years
W = 61980588, p-value = 0.9374
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho
```

```
data:  Male$DrugTests_Other_Positive and Male$Recidivism_Within_3years
S = 1.9423e+12, p-value = 0.9374
alternative hypothesis: true rho is not equal to 0
sample estimates:
         rho
-0.0005213972
```

In [242…]
```
wilcox.test(DrugTests_Other_Positive ~ Race, data=Male)
cor.test(Male$DrugTests_Other_Positive, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_Other_Positive by Race
W = 57820564, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$DrugTests_Other_Positive and Male$Race
S = 1.7069e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1207307
```

In [243…]
```
wilcox.test(DrugTests_Other_Positive ~ Recidivism_Within_3years, data=Female)
cor.test(Female$DrugTests_Other_Positive, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FAL
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_Other_Positive by Recidivism_Within_3years
W = 1206134, p-value = 0.002456
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$DrugTests_Other_Positive and Female$Recidivism_Within_3years
S = 5009135222, p-value = 0.002443
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.05382851
```

In [244…]
```
wilcox.test(DrugTests_Other_Positive ~ Race, data=Female)
cor.test(Female$DrugTests_Other_Positive, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  DrugTests_Other_Positive by Race
W = 1043939, p-value = 1.156e-12
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$DrugTests_Other_Positive and Female$Race
S = 4625082991, p-value = 9.502e-13
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.1263718
```

Is there a correlation between Percent_Days_Employed and recidivism and/or race for everyone, or just males, or just females?

In [245…]
```
wilcox.test(Percent_Days_Employed ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Percent_Days_Employed, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Percent_Days_Employed by Recidivism_Within_3years
W = 101742703, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Percent_Days_Employed and NIJ$Recidivism_Within_3years
S = 3.4989e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.2174536
```

In [246…]
```
wilcox.test(Percent_Days_Employed ~ Recidivism_Within_3years, data=Male)
```

```
cor.test(Male$Percent_Days_Employed, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Percent_Days_Employed by Recidivism_Within_3years
W = 77420973, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Percent_Days_Employed and Male$Recidivism_Within_3years
S = 2.3616e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.2165066
```

In [247…

```
wilcox.test(Percent_Days_Employed ~ Race, data=Male)
cor.test(Male$Percent_Days_Employed, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Percent_Days_Employed by Race
W = 52345212, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Percent_Days_Employed and Male$Race
S = 1.6971e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1257944
```

In [248…

```
wilcox.test(Percent_Days_Employed ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Percent_Days_Employed, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Percent_Days_Employed by Recidivism_Within_3years
W = 1564024, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Percent_Days_Employed and Female$Recidivism_Within_3years
S = 6.494e+09, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.2266397
```

In [249…

```
wilcox.test(Percent_Days_Employed ~ Race, data=Female)
cor.test(Female$Percent_Days_Employed, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Percent_Days_Employed by Race
W = 1048062, p-value = 0.0008339
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Percent_Days_Employed and Female$Race
S = 4979729207, p-value = 0.0008274
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.05938299
```

Is there a correlation between Jobs_Per_Year and recidivism and/or race for everyone, or just males, or just females?

In [250…

```
wilcox.test(Jobs_Per_Year ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Jobs_Per_Year, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Jobs_Per_Year by Recidivism_Within_3years
W = 88435116, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Jobs_Per_Year and NIJ$Recidivism_Within_3years
```

```
S = 3.0877e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.07439718
```

In [251…]
```
wilcox.test(Jobs_Per_Year ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Jobs_Per_Year, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Jobs_Per_Year by Recidivism_Within_3years
W = 67242287, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Jobs_Per_Year and Male$Recidivism_Within_3years
S = 2.084e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.07351884
```

In [252…]
```
wilcox.test(Jobs_Per_Year ~ Race, data=Male)
cor.test(Male$Jobs_Per_Year, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Jobs_Per_Year by Race
W = 52689576, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Jobs_Per_Year and Male$Race
S = 1.7078e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1202892
```

In [253…]
```
wilcox.test(Jobs_Per_Year ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Jobs_Per_Year, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Jobs_Per_Year by Recidivism_Within_3years
W = 1368512, p-value = 7.639e-07
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Jobs_Per_Year and Female$Recidivism_Within_3years
S = 5759319738, p-value = 7.316e-07
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.08787324
```

In [254…]
```
wilcox.test(Jobs_Per_Year ~ Race, data=Female)
cor.test(Female$Jobs_Per_Year, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Jobs_Per_Year by Race
W = 1047517, p-value = 0.0008142
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Jobs_Per_Year and Female$Race
S = 4979106780, p-value = 0.0008078
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.05950055
```

Is there a correlation between Employment_Exempt and recidivism and/or race for everyone, or just males, or just females?

In [255…]
```
wilcox.test(Employment_Exempt ~ Recidivism_Within_3years, data=NIJ)
cor.test(NIJ$Employment_Exempt, NIJ$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction
```

```
data:  Employment_Exempt by Recidivism_Within_3years
W = 84276414, p-value = 5.886e-16
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  NIJ$Employment_Exempt and NIJ$Recidivism_Within_3years
S = 3.0186e+12, p-value = 5.654e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.05034328
```

In [256…

```
wilcox.test(Employment_Exempt ~ Recidivism_Within_3years, data=Male)
cor.test(Male$Employment_Exempt, Male$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Employment_Exempt by Recidivism_Within_3years
W = 63946261, p-value = 4.568e-13
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Employment_Exempt and Male$Recidivism_Within_3years
S = 2.0346e+12, p-value = 4.437e-13
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.04807239
```

In [257…

```
wilcox.test(Employment_Exempt ~ Race, data=Male)
cor.test(Male$Employment_Exempt, Male$Race, method='spearman', use='complete.obs', exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Employment_Exempt by Race
W = 60430204, p-value = 0.00191
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Male$Employment_Exempt and Male$Race
S = 1.9013e+12, p-value = 0.001908
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.02061653
```

In [258…

```
wilcox.test(Employment_Exempt ~ Recidivism_Within_3years, data=Female)
cor.test(Female$Employment_Exempt, Female$Recidivism_Within_3years, method='spearman', use='complete.obs', exact=FALSE)
```

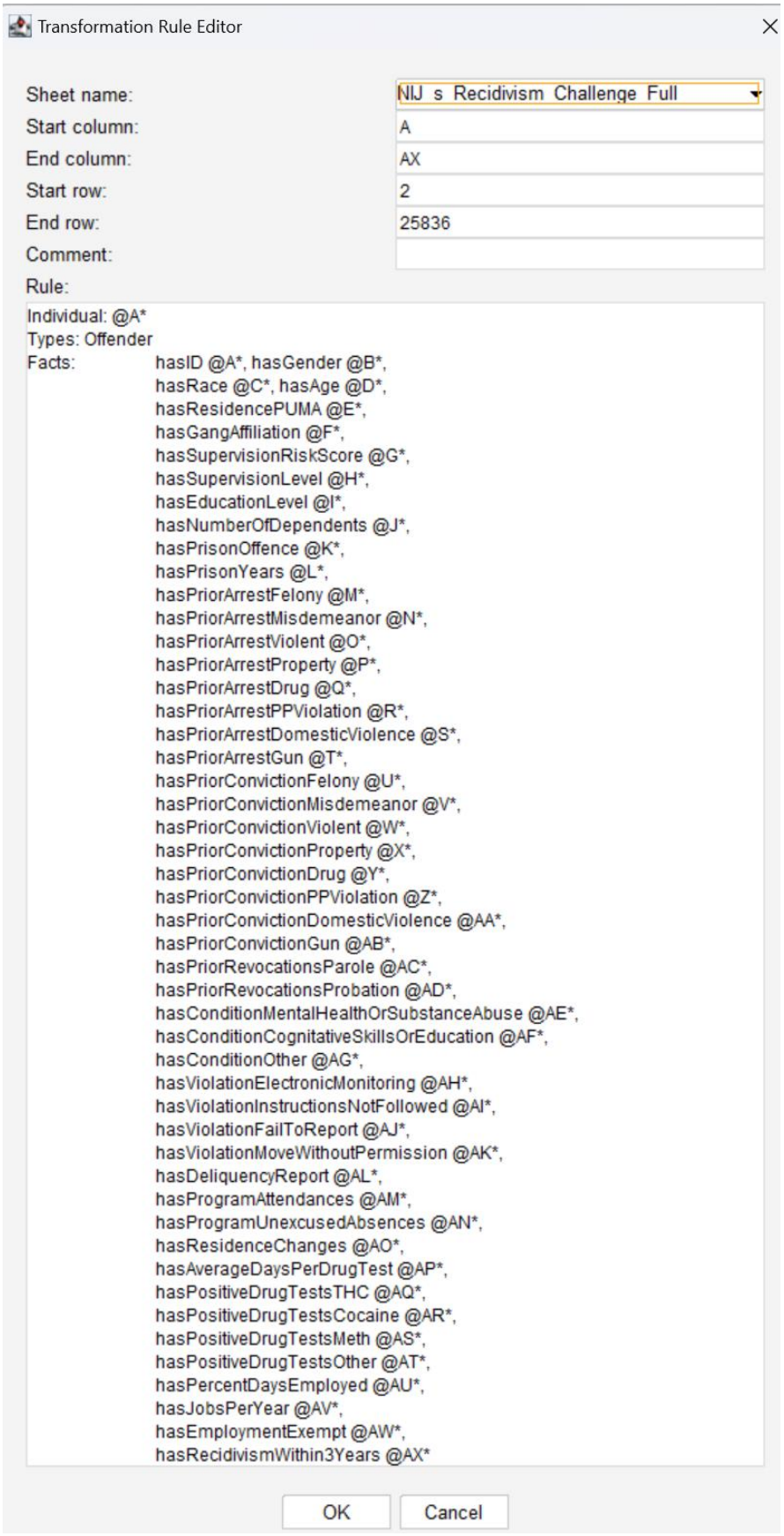
```
        Wilcoxon rank sum test with continuity correction

data:  Employment_Exempt by Recidivism_Within_3years
W = 1254642, p-value = 0.5502
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Employment_Exempt and Female$Recidivism_Within_3years
S = 5350320401, p-value = 0.5503
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.01061769
```

In [259…

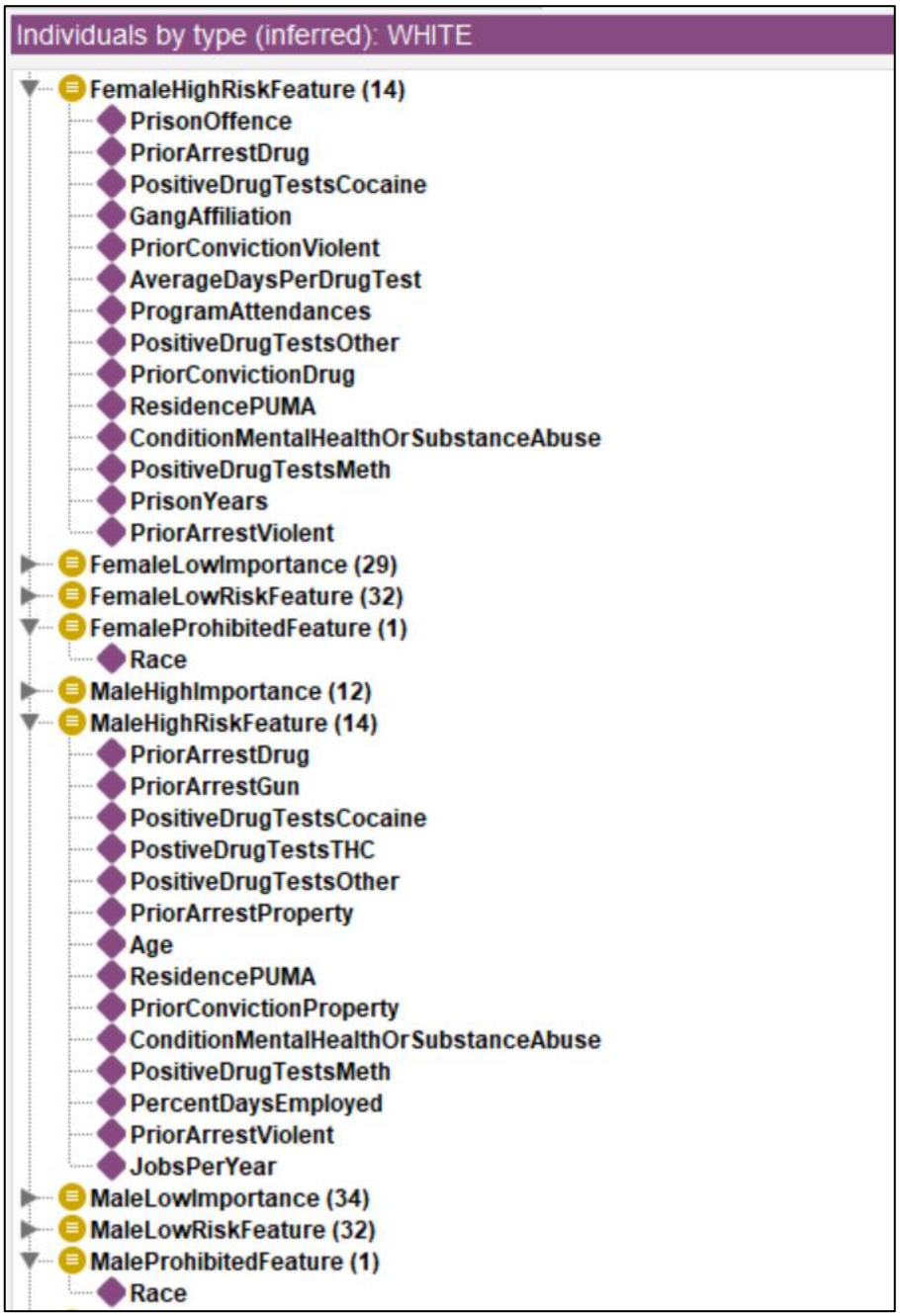
```
wilcox.test(Employment_Exempt ~ Race, data=Female)
cor.test(Female$Employment_Exempt, Female$Race, method='spearman', use='complete.obs', exact=FALSE)
```

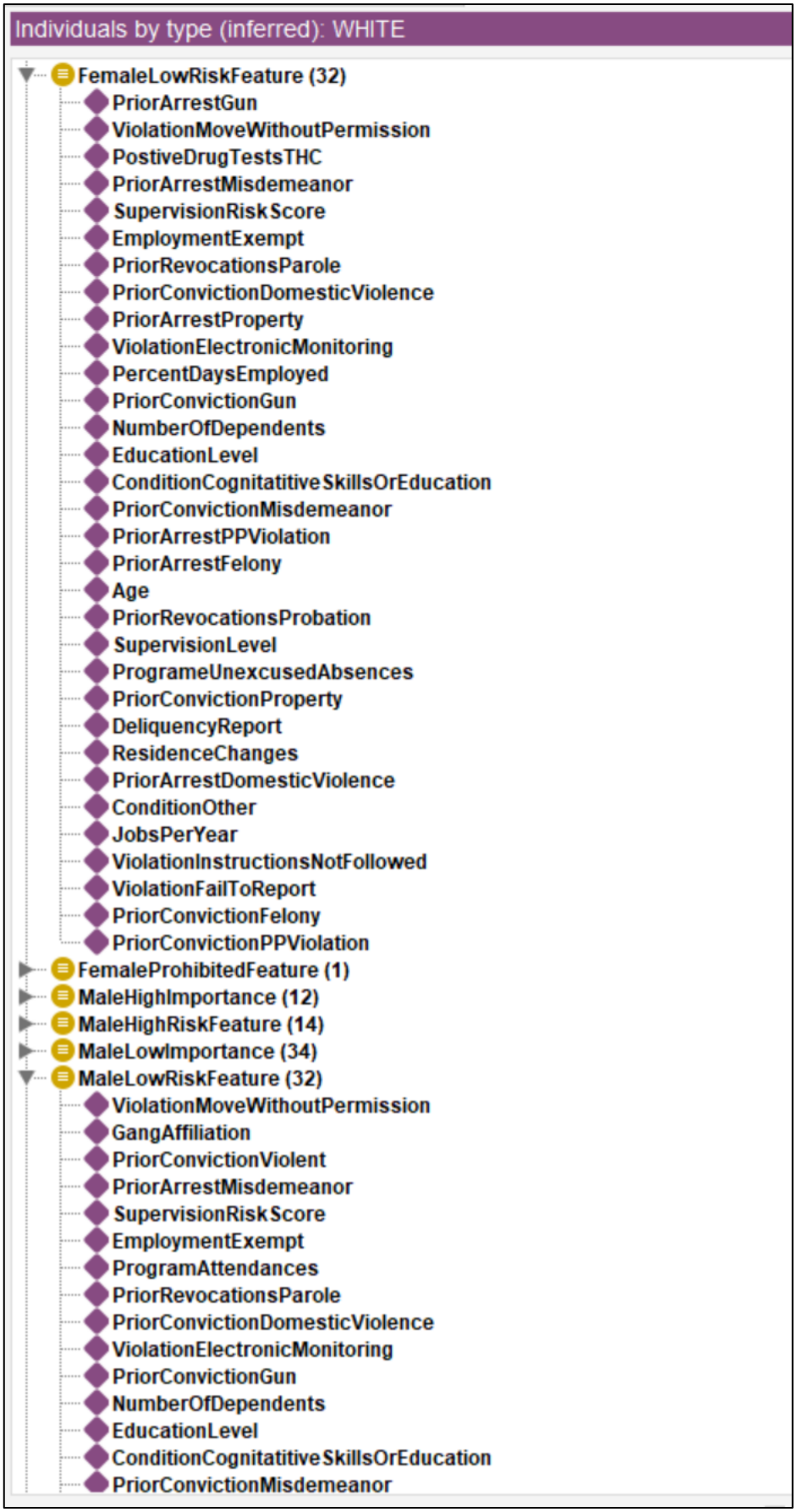
```
        Wilcoxon rank sum test with continuity correction

data:  Employment_Exempt by Race
W = 1088352, p-value = 0.02277
alternative hypothesis: true location shift is not equal to 0
        Spearman's rank correlation rho

data:  Female$Employment_Exempt and Female$Race
S = 5079833732, p-value = 0.02274
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.04047432
```

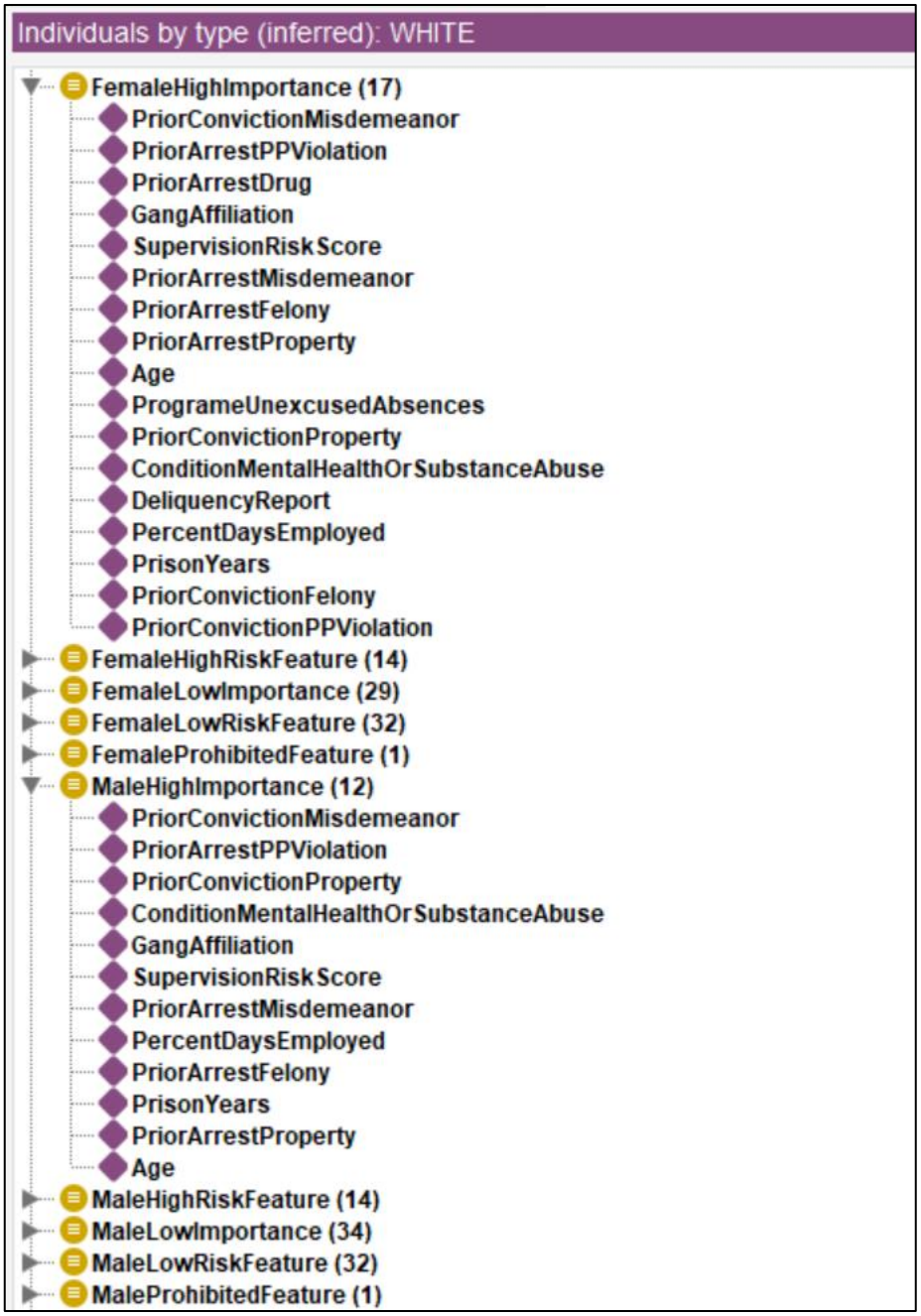
Appendix C. NIJ Ontology Cellfie Import Script

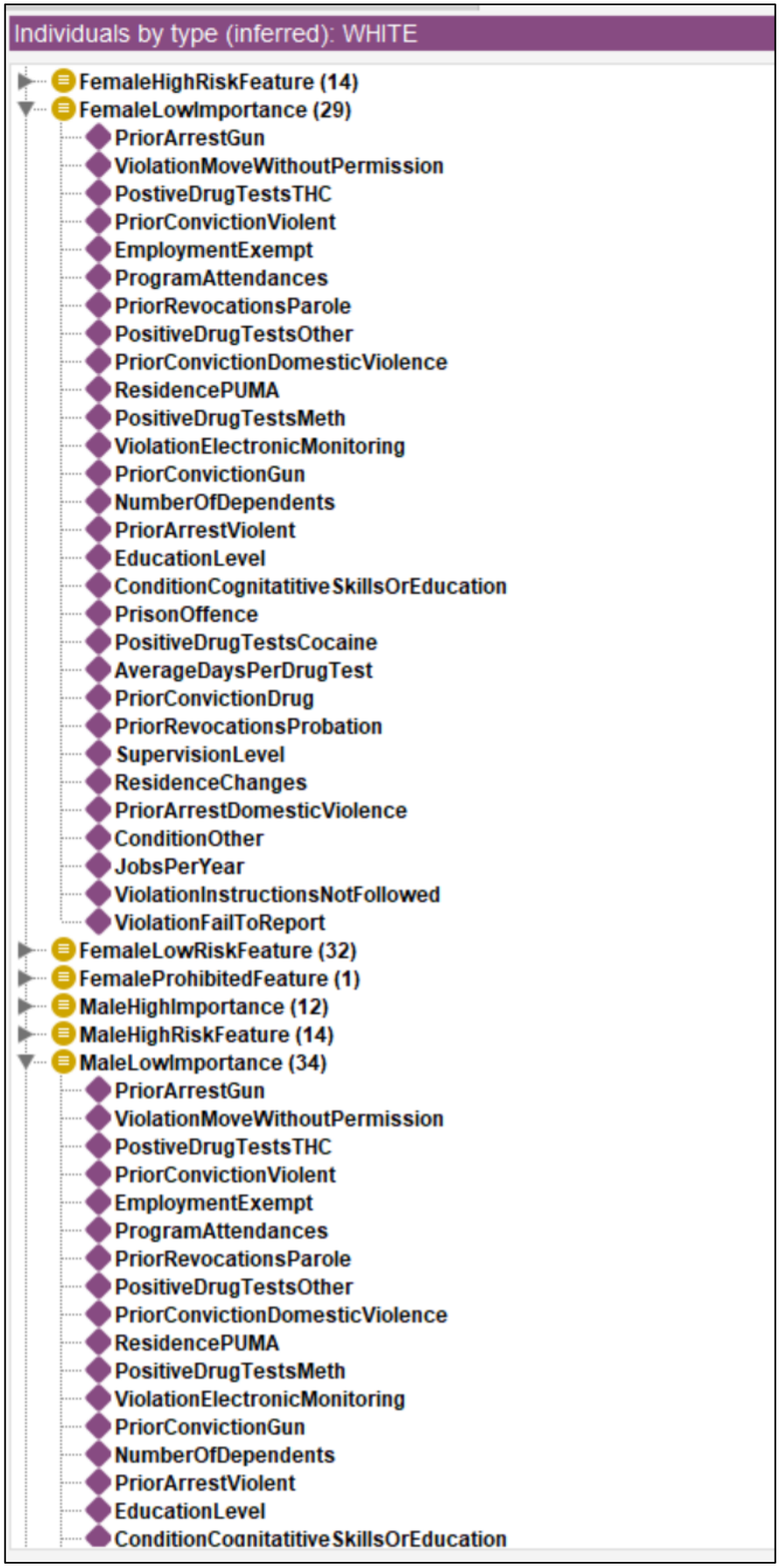


Appendix D. NIJ Ontology High/low Risk/importance Features

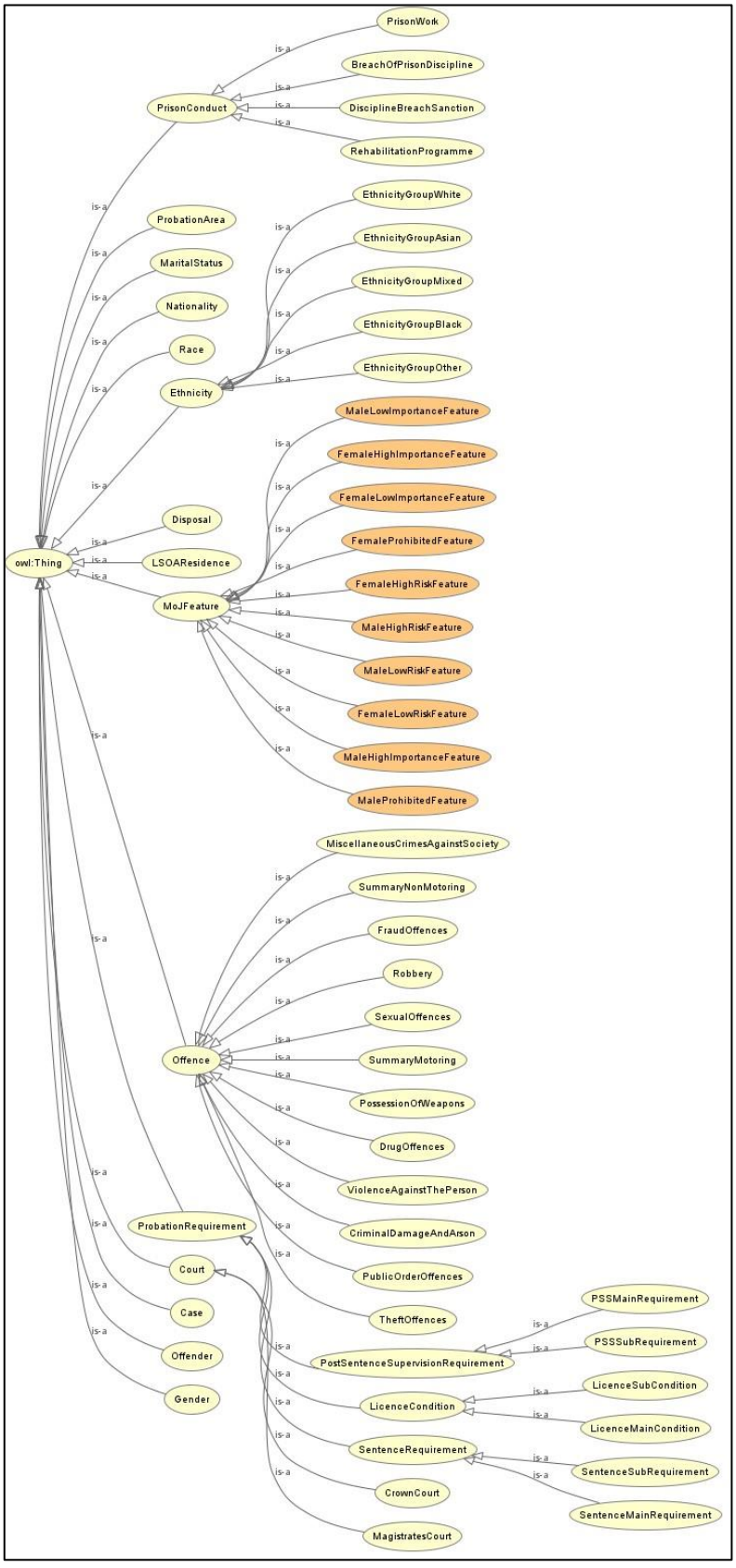



Individuals by type (inferred): WHITE
FemaleLowRiskFeature (32)
PriorArrestGun
ViolationMoveWithoutPermission
PostiveDrugTestsTHC
PriorArrestMisdemeanor
SupervisionRiskScore
EmploymentExempt
PriorRevocationsParole
PriorConvictionDomesticViolence
PriorArrestProperty
ViolationElectronicMonitoring
PercentDaysEmployed
PriorConvictionGun
NumberOfDependents
EducationLevel
ConditionCognitatitiveSkillsOrEducation
PriorConvictionMisdemeanor
PriorArrestPPViolation
PriorArrestFelony
Age
PriorRevocationsProbation
SupervisionLevel
ProgrameUnexcusedAbsences
PriorConvictionProperty
DeliquencyReport
ResidenceChanges
PriorArrestDomesticViolence
ConditionOther
JobsPerYear
ViolationInstructionsNotFollowed
ViolationFailToReport
PriorConvictionFelony
PriorConvictionPPViolation
FemaleProhibitedFeature (1)
MaleHighImportance (12)
MaleHighRiskFeature (14)
MaleLowImportance (34)
MaleLowRiskFeature (32)
ViolationMoveWithoutPermission
GangAffiliation
PriorConvictionViolent
PriorArrestMisdemeanor
SupervisionRiskScore
EmploymentExempt
ProgramAttendances
PriorRevocationsParole
PriorConvictionDomesticViolence
ViolationElectronicMonitoring
PriorConvictionGun
NumberOfDependents
EducationLevel
ConditionCognitatitiveSkillsOrEducation
PriorConvictionMisdemeanor


Individuals by type (inferred): WHITE
FemaleHighImportance (17)
PriorConvictionMisdemeanor
PriorArrestPPViolation
PriorArrestDrug
GangAffiliation
SupervisionRiskScore
PriorArrestMisdemeanor
PriorArrestFelony
PriorArrestProperty
Age
ProgrameUnexcusedAbsences
PriorConvictionProperty
ConditionMentalHealthOrSubstanceAbuse
DeliquencyReport
PercentDaysEmployed
PrisonYears
PriorConvictionFelony
PriorConvictionPPViolation
FemaleHighRiskFeature (14)
FemaleLowImportance (29)
FemaleLowRiskFeature (32)
FemaleProhibitedFeature (1)
MaleHighImportance (12)
PriorConvictionMisdemeanor
PriorArrestPPViolation
PriorConvictionProperty
ConditionMentalHealthOrSubstanceAbuse
GangAffiliation
SupervisionRiskScore
PriorArrestMisdemeanor
PercentDaysEmployed
PriorArrestFelony
PrisonYears
PriorArrestProperty
Age
MaleHighRiskFeature (14)
MaleLowImportance (34)
MaleLowRiskFeature (32)
MaleProhibitedFeature (1)


Individuals by type (inferred): WHITE
FemaleHighRiskFeature (14)
FemaleLowImportance (29)
PriorArrestGun
ViolationMoveWithoutPermission
PostiveDrugTestsTHC
PriorConvictionViolent
EmploymentExempt
ProgramAttendances
PriorRevocationsParole
PositiveDrugTestsOther
PriorConvictionDomesticViolence
ResidencePUMA
PositiveDrugTestsMeth
ViolationElectronicMonitoring
PriorConvictionGun
NumberOfDependents
PriorArrestViolent
EducationLevel
ConditionCognitatitiveSkillsOrEducation
PrisonOffence
PositiveDrugTestsCocaine
AverageDaysPerDrugTest
PriorConvictionDrug
PriorRevocationsProbation
SupervisionLevel
ResidenceChanges
PriorArrestDomesticViolence
ConditionOther
JobsPerYear
ViolationInstructionsNotFollowed
ViolationFailToReport
FemaleLowRiskFeature (32)
FemaleProhibitedFeature (1)
MaleHighImportance (12)
MaleHighRiskFeature (14)
MaleLowImportance (34)
PriorArrestGun
ViolationMoveWithoutPermission
PostiveDrugTestsTHC
PriorConvictionViolent
EmploymentExempt
ProgramAttendances
PriorRevocationsParole
PositiveDrugTestsOther
PriorConvictionDomesticViolence
ResidencePUMA
PositiveDrugTestsMeth
ViolationElectronicMonitoring
PriorConvictionGun
NumberOfDependents
PriorArrestViolent
EducationLevel
ConditionCognitatitiveSkillsOrEducation

Appendix E. MoJ Ontology Protégé Screenshots

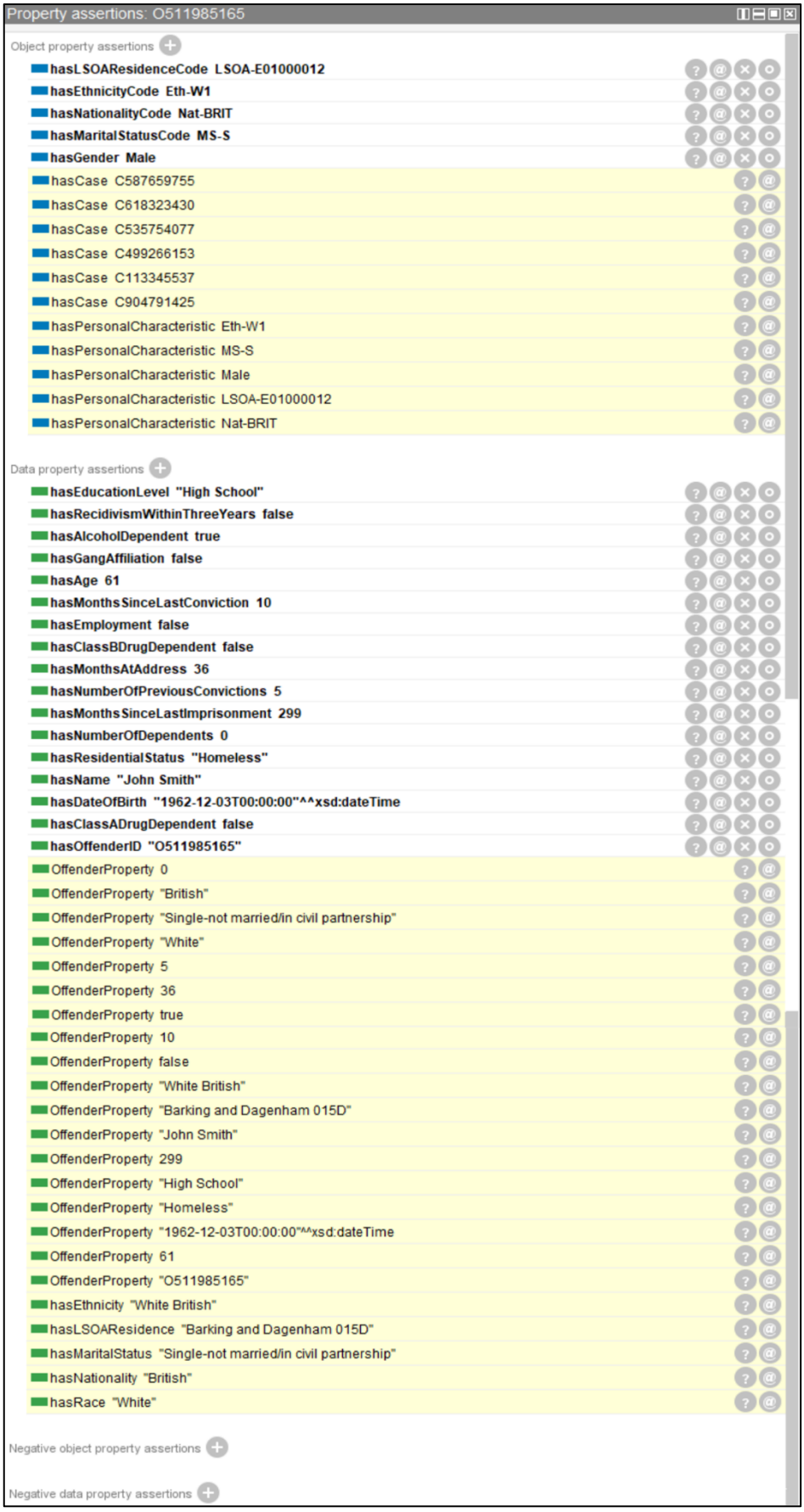


Inferred hierarchy

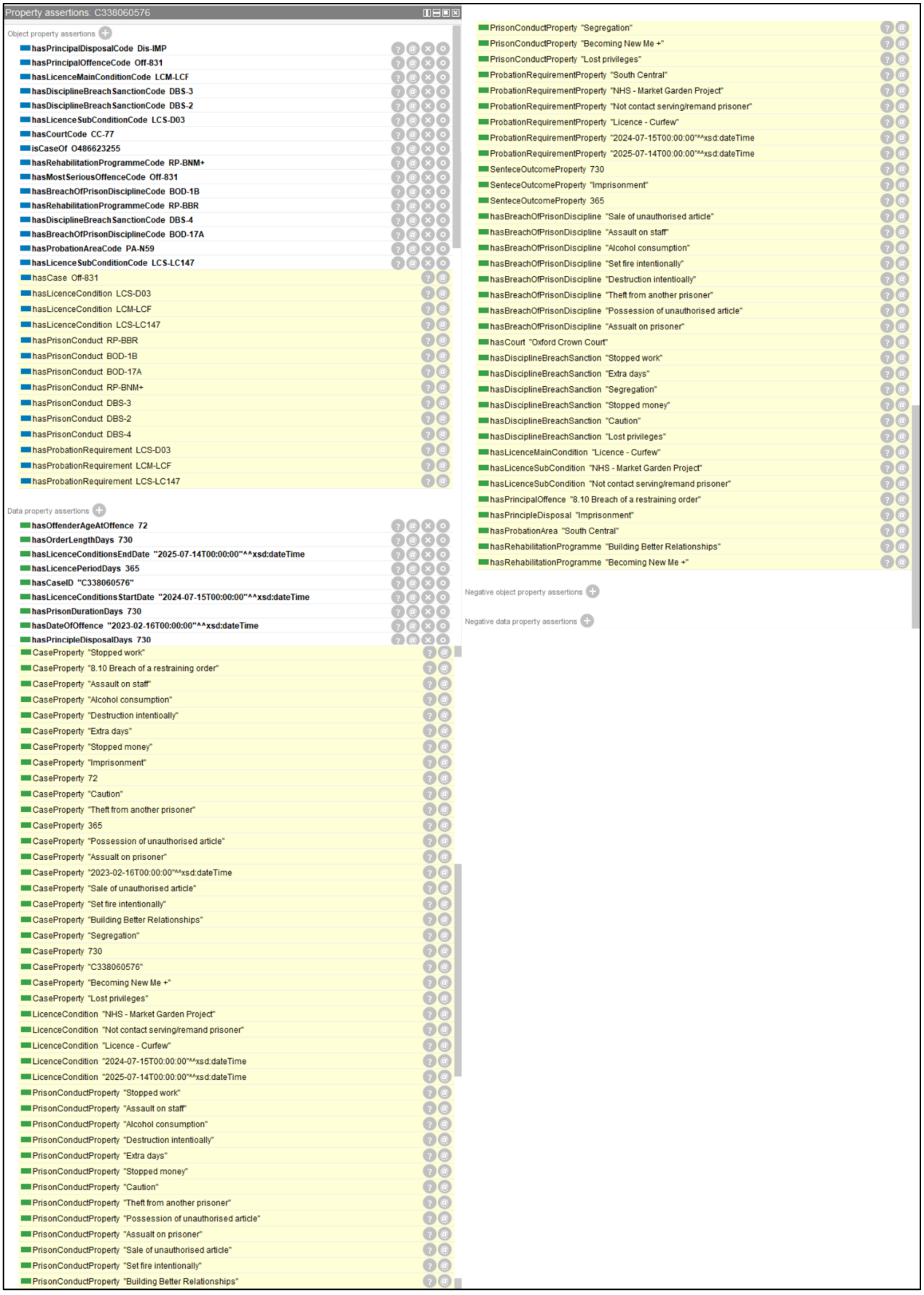
Object properties and data properties

Active ontology
Entities
Individuals by class
OWLViz
DL Query
SWRLTab
SPARQL Query
Name
Rule
Breach of prison discipline mappi...
MOJ:Case(?c) ^ MOJ:hasBreachOfPrisonDisciplineCode(?c, ?code) ^ MOJ:hasBreachOfPrisonDisciplineCode(?b, ?code) ^ MOJ:hasBreachOfPrisonDiscipline(?b, ?breach) -> MOJ:hasBreachOfPrisonDiscipline(?c, ?breach)
Court mapping
MOJ:Case(?c) ^ MOJ:hasCourtCode(?c, ?code) ^ MOJ:hasCourtCode(?ct, ?code) ^ MOJ:hasCourt(?ct, ?court) -> MOJ:hasCourt(?c, ?court)
Discipline breach sanction mappi...
MOJ:Case(?c) ^ MOJ:hasDisciplineBreachSanctionCode(?c, ?code) ^ MOJ:hasDisciplineBreachSanctionCode(?s, ?code) ^ MOJ:hasDisciplineBreachSanction(?s, ?sanction) -> MOJ:hasDisciplineBreachSanction(?c, ?sanction)
Ethnicity mapping
MOJ:Offender(?o) ^ MOJ:hasEthnicityCode(?o, ?code) ^ MOJ:hasEthnicityCode(?e, ?code) ^ MOJ:hasEthnicity(?e, ?ethnicity) -> MOJ:hasEthnicity(?o, ?ethnicity)
LSOA Residence mapping
MOJ:Offender(?o) ^ MOJ:hasLSOAResidenceCode(?o, ?code) ^ MOJ:hasLSOAResidence(?r, ?code) ^ MOJ:hasLSOAResidence(?r, ?residence) -> MOJ:hasLSOAResidence(?o, ?residence)
Licence main condition mapping
MOJ:Case(?c) ^ MOJ:hasLicenceMainConditionCode(?c, ?code) ^ MOJ:hasLicenceMainCondition(?cond, ?code) ^ MOJ:hasLicenceMainCondition(?cond, ?condition) -> MOJ:hasLicenceMainCondition(?c, ?condition)
Licence sub condition mapping
MOJ:Case(?c) ^ MOJ:hasLicenceSubConditionCode(?c, ?code) ^ MOJ:hasLicenceSubCondition(?cond, ?code) ^ MOJ:hasLicenceSubCondition(?cond, ?condition) -> MOJ:hasLicenceSubCondition(?c, ?condition)
Marital status mapping
MOJ:Offender(?o) ^ MOJ:hasMaritalStatusCode(?o, ?code) ^ MOJ:hasMaritalStatusCode(?m, ?code) ^ MOJ:hasMaritalStatus(?m, ?status) -> MOJ:hasMaritalStatus(?o, ?status)
Most serious offence mapping
MOJ:Case(?c) ^ MOJ:hasMostSeriousOffenceCode(?c, ?code) ^ MOJ:hasOffenceCode(?o, ?code) ^ MOJ:hasOffence(?o, ?offence) -> MOJ:hasMostSeriousOffenceCode(?c, ?offence)
Nationality mapping
MOJ:Offender(?o) ^ MOJ:hasNationalityCode(?o, ?code) ^ MOJ:hasNationalityCode(?n, ?code) ^ MOJ:hasNationality(?n, ?nationality) -> MOJ:hasNationality(?o, ?nationality)
PSS main requirement mapping
MOJ:Case(?c) ^ MOJ:hasPSSMainRequirementCode(?c, ?code) ^ MOJ:hasPSSMainRequirementCode(?r, ?code) ^ MOJ:hasPSSMainRequirement(?r, ?requirement) -> MOJ:hasPSSMainRequirement(?c, ?requirement)
PSS sub requirement mapping
MOJ:Case(?c) ^ MOJ:hasPSSSubRequirementCode(?c, ?code) ^ MOJ:hasPSSSubRequirementCode(?r, ?code) ^ MOJ:hasPSSSubRequirement(?r, ?requirement) -> MOJ:hasPSSSubRequirement(?c, ?requirement)
Principle disposal mapping
MOJ:Case(?c) ^ MOJ:hasDisposalCode(?d, ?code) ^ MOJ:hasDisposal(?d, ?disposal) ^ MOJ:hasPrincipalDisposalCode(?c, ?code) -> MOJ:hasPrincipleDisposal(?c, ?disposal)
Principle offence mapping
MOJ:Case(?c) ^ MOJ:hasOffence(?o, ?offence) ^ MOJ:hasOffenceCode(?o, ?code) ^ MOJ:hasPrincipalOffenceCode(?c, ?code) -> MOJ:hasPrincipalOffence(?c, ?offence)
Prison work mapping
MOJ:Case(?c) ^ MOJ:hasPrisonWorkCode(?c, ?code) ^ MOJ:hasPrisonWorkCode(?w, ?code) ^ MOJ:hasPrisonWork(?w, ?work) -> MOJ:hasPrisonWork(?c, ?work)
Probation area mapping
MOJ:Case(?c) ^ MOJ:hasProbationAreaCode(?c, ?code) ^ MOJ:hasProbationAreaCode(?p, ?code) ^ MOJ:hasProbationArea(?p, ?area) -> MOJ:hasProbationArea(?c, ?area)
Race mapping
MOJ:Offender(?o) ^ MOJ:hasEthnicityCode(?o, ?code) ^ MOJ:hasEthnicityCode(?e, ?code) ^ MOJ:hasRace(?e, ?race) -> MOJ:hasRace(?o, ?race)
Rehabilitation programme mapping
MOJ:Case(?c) ^ MOJ:hasRehabilitationProgrammeCode(?c, ?code) ^ MOJ:hasRehabilitationProgrammeCode(?p, ?code) ^ MOJ:hasRehabilitationProgramme(?p, ?programme) -> MOJ:hasRehabilitationProgramme(?c, ?programme)
Sentence main requirement mapp...
MOJ:Case(?c) ^ MOJ:hasSentenceMainRequirementCode(?c, ?code) ^ MOJ:hasSentenceMainRequirementCode(?r, ?code) ^ MOJ:hasSentenceMainRequirement(?r, ?requirement) -> MOJ:hasSentenceMainRequirement(?c, ?requirement)
Sentence sub requirement mappi...
MOJ:Case(?c) ^ MOJ:hasSentenceSubRequirementCode(?c, ?code) ^ MOJ:hasSentenceSubRequirementCode(?r, ?code) ^ MOJ:hasSentenceSubRequirement(?r, ?requirement) -> MOJ:hasSentenceSubRequirement(?c, ?requirement)
New
Edit
Clone
Delete

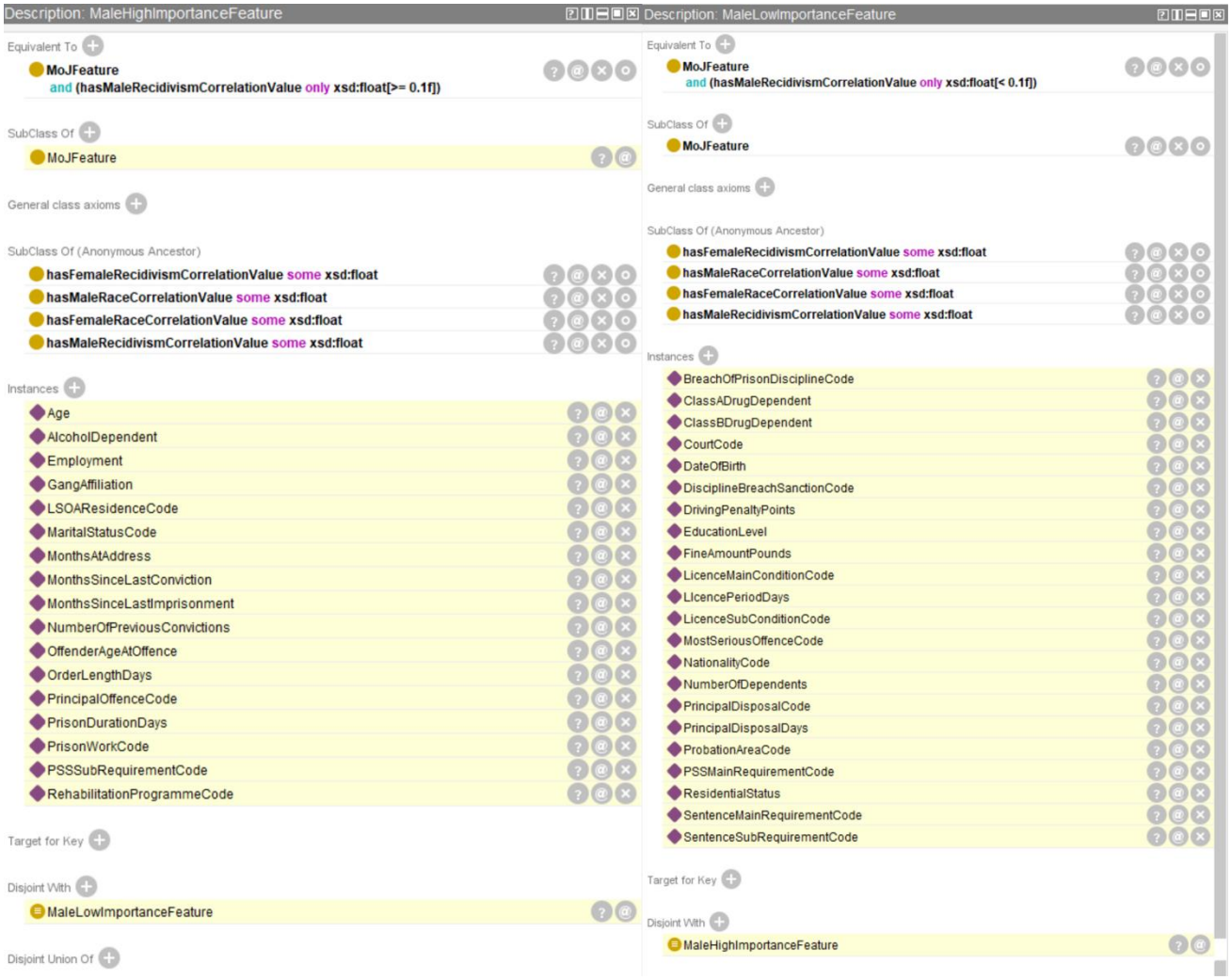
SWRL rules



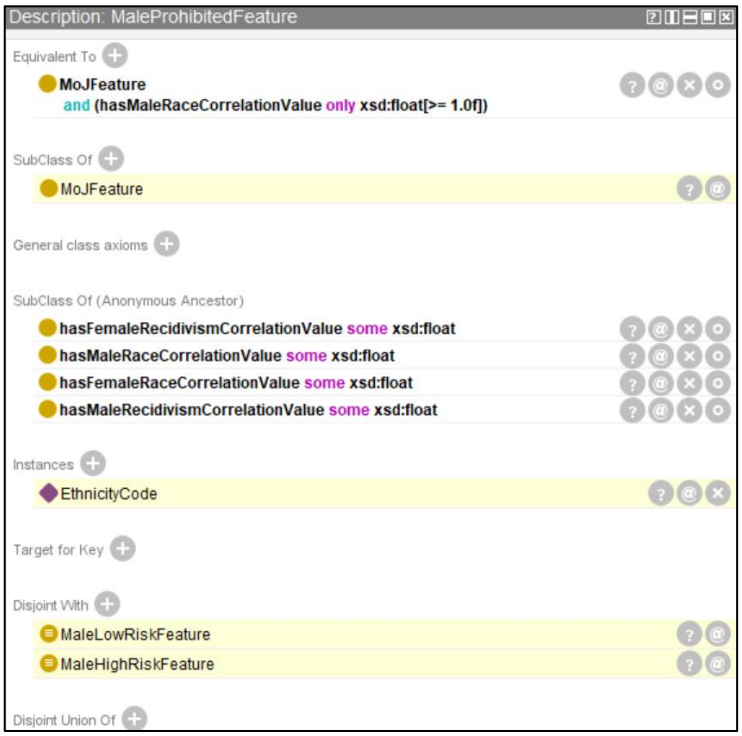
Sample offender property assertions



Sample case property assertions



High and low importance features



High risk, low risk, and prohibited features

Snap SPARQL Query:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX moj: <http://www.semanticweb.org/leigh/ontologies/2024/4/MOJ#>

SELECT ?offender
    (REPLACE(STR(?monthsSinceImprisonmentValue), STR(moj:), "") AS ?monthsSinceImprisonment)
    (REPLACE(STR(?monthsSinceConvictionValue), STR(moj:), "") AS ?monthsSinceConviction)
    (REPLACE(STR(?employmentValue), STR(moj:), "") AS ?employment)
    (REPLACE(STR(?monthsAtAddressValue), STR(moj:), "") AS ?monthsAtAddress)
    (REPLACE(STR(?gangAffiliationValue), STR(moj:), "") AS ?gangAffiliation)
    (REPLACE(STR(?ageValue), STR(moj:), "") AS ?age)
    (REPLACE(STR(?previousConvictionsValue), STR(moj:), "") AS ?previousConvictions)
    (REPLACE(STR(?maritalStatusValue), STR(moj:), "") AS ?maritalStatus)
    (REPLACE(STR(?alcoholDependentValue), STR(moj:), "") AS ?alcoholDependent)
    (REPLACE(STR(?caseValue), STR(moj:), "") AS ?case)
    (REPLACE(STR(?rehabilitationProgrammeValue), STR(moj:), "") AS ?rehabilitationProgramme)
    (REPLACE(STR(?principalOffenceValue), STR(moj:), "") AS ?principalOffence)
    (REPLACE(STR(?PSSSubRequirementValue), STR(moj:), "") AS ?PSSSubRequirement)
    (REPLACE(STR(?prisonWorkValue), STR(moj:), "") AS ?prisonWork)
    (REPLACE(STR(?prisonDurationValue), STR(moj:), "") AS ?prisonDuration)
WHERE {
 ?offender a moj:Offender
          FILTER(REGEX(STR(?offender), CONCAT("^", STR(moj:), "O")))
          OPTIONAL { ?offender moj:hasMonthsSinceLastImprisonment ?monthsSinceImprisonmentValue. }
          OPTIONAL { ?offender moj:hasMonthsSinceLastConviction ?monthsSinceConvictionValue. }
          OPTIONAL { ?offender moj:hasEmployment ?employmentValue. }
          OPTIONAL { ?offender moj:hasMonthsAtAddress ?monthsAtAddressValue. }
          OPTIONAL { ?offender moj:hasGangAffiliation ?gangAffiliationValue. }
          OPTIONAL { ?offender moj:hasAge ?ageValue. }
          OPTIONAL { ?offender moj:hasNumberOfPreviousConvictions ?previousConvictionsValue. }
          OPTIONAL { ?offender moj:hasMaritalStatusCode ?maritalStatusValue. }
          OPTIONAL { ?offender moj:hasAlcoholDependent ?alcoholDependentValue. }
 ?offender moj:hasCase ?case
          OPTIONAL { ?case moj:hasPrisonDurationDays ?prisonDurationValue. }
          OPTIONAL { ?case moj:hasPrisonDurationDays ?prisonDurationValue. }
          OPTIONAL { ?case moj:hasRehabilitationProgrammeCode ?rehabilitationProgrammeValue. }
          OPTIONAL { ?case moj:hasPrincipalOffenceCode ?principalOffenceValue. }
          OPTIONAL { ?case moj:hasPSSSubRequirementCode ?PSSSubRequirementValue. }
          OPTIONAL { ?case moj:hasPrisonWorkCode ?prisonWorkValue. }
}
ORDER BY ?offender
```

Execute

?offender	?mo...	?mo...	?emp...	?m...	?gan...	?a...	?p...	?marit...	?alco...	?case	?rehabili...	?princip...	?PSSSubR...	?prison...	?pris...
moj:O112299926		11	false	13	false	63	4	MS-D	true	moj:C647325366		Off-4600			
moj:O112299926		11	false	13	false	63	4	MS-D	true	moj:C214350956		Off-4600			
moj:O112299926		11	false	13	false	63	4	MS-D	true	moj:C691205173		Off-4600			
moj:O112299926		11	false	13	false	63	4	MS-D	true	moj:C986467303		Off-4600			
moj:O112299926		11	false	13	false	63	4	MS-D	true	moj:C745987478		Off-4600			
moj:O141806441		2	false	52	false	59	0	MS-M	false	moj:C637473147		Off-806			
moj:O265466553		16	false	15	false	24	0	MS-M	true	moj:C196164791		Off-4600			
moj:O341786705	0	8	false	51	false	46	3	MS-M	true	moj:C527235880		Off-9261			
moj:O341786705	0	8	false	51	false	46	3	MS-M	true	moj:C905424044	RP-TBP	Off-406			730
moj:O341786705	0	8	false	51	false	46	3	MS-M	true	moj:C794276181		Off-9250			
moj:O341786705	0	8	false	51	false	46	3	MS-M	true	moj:C830828808		Off-9250			
moj:O486623255	0	10	false	13	false	73	3	MS-W	false	moj:C478663808		Off-4600			
moj:O486623255	0	10	false	13	false	73	3	MS-W	false	moj:C338060576	RP-BBR	Off-831			730
moj:O486623255	0	10	false	13	false	73	3	MS-W	false	moj:C338060576	RP-BNM+	Off-831			730
moj:O486623255	0	10	false	13	false	73	3	MS-W	false	moj:C348844983		Off-806			
moj:O511985165	299	10	false	36	false	61	5	MS-S	true	moj:C587659755		Off-4600			
moj:O511985165	299	10	false	36	false	61	5	MS-S	true	moj:C618323430	RP-LNM	Off-3400	PSSS-SA20	PW-6	365
moj:O511985165	299	10	false	36	false	61	5	MS-S	true	moj:C535754077		Off-4510			
moj:O511985165	299	10	false	36	false	61	5	MS-S	true	moj:C499266153		Off-4600			
moj:O511985165	299	10	false	36	false	61	5	MS-S	true	moj:C113345537		Off-4600			
moj:O511985165	299	10	false	36	false	61	5	MS-S	true	moj:C904791425		Off-3400			
moj:O547242054	0	1	false	31	false	23	4	MS-S	false	moj:C375098237		Off-9250			
moj:O547242054	0	1	false	31	false	23	4	MS-S	false	moj:C851372657		Off-4600			
moj:O547242054	0	1	false	31	false	23	4	MS-S	false	moj:C607333684		Off-9250			
moj:O547242054	0	1	false	31	false	23	4	MS-S	false	moj:C569872265	RP-LNM	Off-9210		PW-1	1277
moj:O547242054	0	1	false	31	false	23	4	MS-S	false	moj:C254596091		Off-9250			
moj:O613456052	3	27	false	62	false	69	1	MS-M	true	moj:C808774974	RP-IM	Off-5301		PW-1	760
moj:O613456052	3	27	false	62	false	69	1	MS-M	true	moj:C808774974	RP-IM	Off-5301		PW-2	760
moj:O613456052	3	27	false	62	false	69	1	MS-M	true	moj:C723372549		Off-4600			
moj:O646955585		7	false	9	false	42	0	MS-S	false	moj:C595710701		Off-4600			
moj:O815709338		4	false	58	false	20	0	MS-S	false	moj:C755833865		Off-4600			
moj:O815709338		4	false	58	false	20	0	MS-S	false	moj:C178906509		Off-4600			
moj:O815709338		4	false	58	false	20	0	MS-S	false	moj:C861603678		Off-4600			
moj:O815709338		4	false	58	false	20	0	MS-S	false	moj:C202280393		Off-4600			
moj:O815709338		4	false	58	false	20	0	MS-S	false	moj:C112529789		Off-4600			
moj:O817633970		1	false	37	false	29	4	MS-S	true	moj:C782821873		Off-835			
moj:O983190851		15	false	3	false	26	0	MS-M	false	moj:C281645560		Off-835			

SPARQL query to export code instances

Snap SPARQL Query:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX moj: <http://www.semanticweb.org/leigh/ontologies/2024/4/MOJ#>

#Total people and number who reoffended by gender and race
SELECT ?gender ?race (COUNT(?person) AS ?total) (COUNT(?reoffender) AS ?reoffended)
WHERE {
  ?person a moj:Offender ;
            moj:hasGender ?gender ;
            moj:hasRace ?race .

  OPTIONAL {
    ?person moj:hasRecidivismWithinThreeYears true .
    BIND(IF(BOUND(?person), ?person, "") AS ?reoffender)
  }
}
GROUP BY ?gender ?race
ORDER BY ?gender ?race
```

Execute

?gender	?race	?total	?reoffended
moj:Female	Non-White^^xsd:string	3	1
moj:Female	White^^xsd:string	2	1
moj:Male	Non-White^^xsd:string	4	2
moj:Male	White^^xsd:string	3	2

4 results

SPARQL query to examine racial profile of recidivism

Appendix F. MoJ Object Instances (Truncated)

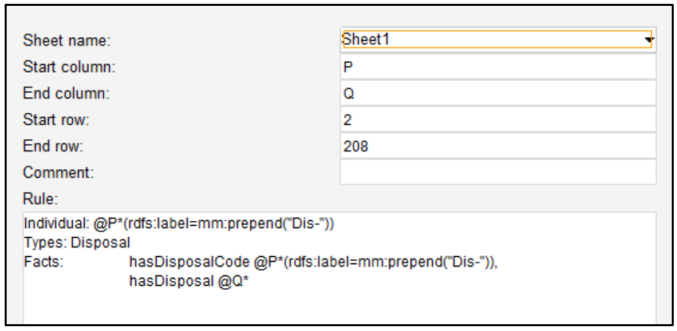
Ethnicity Code	Ethnicity Descrption	Marital status code	Marital status	Nationality code	Nationality	Magistrates' Court Code	Magistrates' Court Name	Crown Court Code	Crown Court Name	Offence code	Offence description
A1	Indian	C	Co-habiting (living with partner)	AB	Aruban	1013	Bristol Magistrates' Court	78	Aylesbury Crown Court	101	1 Murder
A2	Pakistani	D	Divorced or dissolved	AFGA	Afghan	1021	North Avon Magistrates' Court	30	Bradford Crown Court	102	1 Murder
A3	Bangladeshi	M	Married or in civil partnership	AG	Anguillan	1022	Bath and Wansdyke Magistrates' Court	66	Kingston-upon-Hull Crown Court	200	2 Attempted Murder
A9	Any other Asian background	N	Prefer not to say	ALBA	Albanian	1023	North Somerset Magistrates' Court	58	Birmingham Crown Court	301	3B Threats to kill
B1	Caribbean	P	Separated-not living with legal	ALGE	Algerian	1030	Somerset Magistrates' Court	63	Bournemouth Crown Court	302	3A Conspiracy to murder
B2	African	S	Single-not married/in civil partnership	AMER	American	1051	Bedford and Mid Bedfordshire Magistrates' Court	68	Dorchester Crown Court	303	3A Conspiracy to murder
B9	Any other Black background	W	Widowed	AN	Dutch Antillean	1055	Luton and South Bedfordshire Magistrates' Court	12	Bristol Crown Court	304	3A Conspiracy to murder
M1	White and Black Caribbean			ANGOL	Angolan	1072	East Berkshire Magistrates' Court	16	Burnley Crown Court	305	3A Conspiracy to murder
M2	White and Black African			ANTIG	Antiguan, Barbudan	1075	West Berkshire Magistrates' Court	17	Cambridge Crown Court	306	3A Conspiracy to murder
M3	White and Asian			ARGEN	Argentine	1076	Reading Magistrates' Court	55	Cardiff Crown Court	401	4.1 Manslaughter
M9	Any other mixed			ARME	Armenian	1080	Bedfordshire Magistrates' Court	62	Carlisle Crown Court	402	4.2 Infanticide
O1	Chinese			ASM	American Samoan	1124	Milton Keynes Magistrates' Court	41	Central Criminal Court	403	4.3 Child Destruction
O2	Arab			AUSI	Australian	1129	Central Buckinghamshire Magistrates' Court	11	Chelmsford Crown Court	404	4.4 Causing Death by Dangerous Driving (MOT)
O9	Any other ethnic group			AUST	Austrian	1130	Wycombe and Beaconsfield Magistrates' Court	69	Chester Crown Court	405	4.5 Manslaughter Due to Diminished Responsibility
W1	British			AZERB	Azerbaijani	1162	Peterborough Magistrates' Court	65	Chichester Crown Court	406	4.6 Causing Death by Careless Driving when under the influence of Drink or Drugs (MOT)
W2	Irish			BAHA	Bahamian	1165	Cambridge Magistrates' Court	60	Coventry Crown Court	407	4.7 Causing Death of a child or vulnerable person
W3	Gypsy or Irish Traveller			BAHR	Bahraini	1166	East Cambridgeshire Magistrates' Court	47	Croydon Crown Court	408	4.8 Causing Death by careless or inconsiderate driving (MOT)
W9	Any other White background			BANGL	Bangladeshi	1167	Fenland Magistrates' Court	26	Derby Crown Court	409	4.9 Causing death by driving unlicensed disqualified or uninsured drivers (MOT)
NS	Not stated			BARB	Barbadian or Bajuns	1168	Huntingdonshire Magistrates' Court	23	Doncaster Crown Court	410	4.10 Applicable organisation causing death by gross breach of duty of care
				BELA	Belarusian	1177	Halton Magistrates' Court	59	Wolverhampton Crown Court	411	4.11 Causing or allowing child or vulnerable adult to suffer serious physical harm
				BELG	Belgian	1178	Macclesfield Magistrates' Court	5	Durham Crown Court	412	4.12 Causing serious injury by dangerous driving (MOT)

Offence group code	Offence group description	Disposal Code	Disposal name	Probation area code	Probation area name	Requirement type main category code	Requirement type main category	Requirement type sub category code
1	Violence against the person	ABDIS	Absolute Discharge	ACI	Altcourse (HMP)	3	Residential	701
1	Violence against the person	ADISQU	Disqualification Order (from working with children) - ADULTS	AGI	Askham Grange (HMP & YOI)	7	Court - Accredited Programme	702
1	Violence against the person	ALTCT	Alternative Count: Jury Discharged from Verdict	ALL	No Trust or Trust Unknown	7A	Accredited Programme - Not specified at Court	703
1	Violence against the person	ALTRTCS	Alcohol Treatment	ASI	Ashfield (HMYOI)	A1	Local - CRAMS Activity	704
1	Violence against the person	ALTRTSS	Alcohol Treatment	ASP	Avon & Somerset	A2	Local - Guided Skills Learning	705
1	Violence against the person	ASBOSEN	Order re anti-social behaviour additional to sentence	AYI	Aylesbury (HMYOI)	A3	Local - Basic Skills	706
1	Violence against the person	ATCRQCS	Attendance Centre Requirements	BAI	Belmarsh (HMP)	A4	Local - ETE	707
1	Violence against the person	ATCRQSS	Attendance Centre Requirements	BCI	Buckley Hall (HMP)	A5	Local - Skills for Life - Literacy	708
1	Violence against the person	ATCRQYR	Attendance Centre Requirement	BED	Bedfordshire	A6	Local - Skills for Life - Numeracy	709
1	Violence against the person	ATTCEN	Attendance Centre Order	BFI	Bedford (HMP)	A7	Local - Skills for Life - ESOL	710
1	Violence against the person	BARR	May be placed on Barring List	BLI	Bristol (HMP)	A8	Local - Skills for Life - Other	711
1	Violence against the person	BOCOND	Bind Over with Conditions	BMI	Birmingham (HMP)	B	Education (YRO)	712
1	Violence against the person	BOGB	Bind Over	BNI	Bullingdon (HMP)	B4	Local - Non Accr Prog	713
1	Violence against the person	BOJDG	Bind Over to Appear for Judgment	BRI	Bure (HMP)	C	Intensive Supervision and Surveillance (YRO)	714
1	Violence against the person	BOJDGD	Bind Over to Appear for Judgment on Date given	BSI	Brinsford (HMYOI)	CSPE	Court Stated Punitive Element	715
1	Violence against the person	BWBAILB	Backed for Bail: Failed to answer bail	BT1	BVT CRC	E	Drug Testing (YRO)	716
1	Violence against the person	BWBAILN	Not Backed for Bail: Failed to answer bail	BVT	BVT NPS Division	ESXRM	Local - ESX Interventions	718
1	Violence against the person	BWSUMB	Backed for Bail: Failed to answer summons	BWI	Berwyn (HMP)	F	Rehabilitation Activity Requirement (RAR)	719
1	Violence against the person	BWSUMN	Not Backed for Bail: Failed to answer summons	BXI	Brixton (HMP)	G	Drug Rehabilitation	720
1	Violence against the person	CD	Conditional Discharge Order	BZI	Bronzefield (HMP)	GMPRM1	Local - GMPT Activities	721
1	Violence against the person	CJCO	Confiscation Order under S71 CJA 1988	C01	CPA Northumbria	H	Alcohol Treatment	722

Requirement type sub category	License condition main category code	License condition main category	License condition sub category code	License condition sub category	Post sentence supervision main category code	Post sentence supervision main category	Post sentence supervision sub category code
Think First	EM01	Licence - Location Monitoring (GPS Tagging)	701	Comply with any requirements for the purpose of addressing alcohol/ drug/ sexual/ gambling/ solvent abuse/ anger/ debt/ prolific offending behaviour problems	S08	Specified Activity	SA01
Enhanced Thinking Skills	ESXLC	Local - ESX Interventions	702	Participate in a prolific or other priority offender (PPO) project	S09	Drug Testing	SA02
One to One	GMPLC1	Local - GMPT Activities	799	Other (see notes) (799)	S10	Drug Appointments	SA03
Cognitive Skills Booster	GPT002	Local - GPT ETE With Award	A01	Attend medical/psychiatric appointments and co-operate with	SS7	Standard 7 Conditions	SA04
CALM	LA4	Local - ETE	A02	Receive home visits from mental health worker			SA05
Aggression Replacement Training	LA5	Local - Skills for Life - Literacy	A202	Entry Level 1			SA06
Integrated Domestic Abuse	LA6	Local - Skills for Life - Numeracy	A203	Entry Level 2			SA07
CDVP - Domestic Violence	LA8	Local - Skills for Life - Other	A204	Entry Level 3			SA08
TV-SOGP	LAP	Licence - Accredited Programme	A205	Level Unknown			SA09
NBR Sex Offender Rolling Prog	LC100	WTS ETE External Signposting	A206	Level 1			SA10
C-SOGP	LC101	WTS ETE Internal Management	A207	Level 2			SA11
ASRO	LC102	WTS Role	A99	Other (see notes)			SA12
Offender Substance Abuse Prog	LC103	WTS SARC	C01	Not to seek to approach or communicate with victim/family member			SA13
PRISM	LC16	Local - Health Trainer	C02	Not to have unsupervised contact with children			SA14
Drink Impaired Drivers	LC17	Local - Victim Liaison	C99	Other (see notes)			SA15
Womens Acquisitive Crime	LC18	Local - Restorative Justice	CE1	Not to contact person with terrorist related offence			SA16
Cognitive Self Change Block 6	LC27	Local - LCS ETE	D01	Not to contact/associate with named individual			SA17
Internet Sex Offender Treatment	LC28	Local - HBS Activities	D02	Not to contact known sex offender			SA18
COVAID (Control - Angry Impulsive Drinkers)	LC29	Local - Notts Activities	D03	Not contact serving/remand prisoner			SA19
LIAP (Low Intensity Alcohol Programme)	LC30	Licence - Programme	D04	Not to associate with any person linked with specific group/organisation			SA20
TSP (Thinking Skills)	LC32	Local - WY ETE	D99	Other (see notes)			SA21

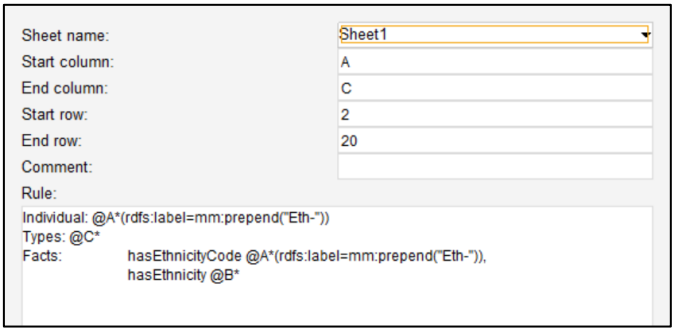
Post sentence supervision sub category	LSOA Residence Code	LSOA Residence Name	Prison discipline breach code	Prison discipline breach	Prison sanction code	Prison sanction	Prison work code	Prison work	Rehabilitation programme code	Rehabilitation programme name
Anger Management - High Intensity	E01000001	City of London 001A	1A	Assault on staff	1	Caution	1	Cleaner	ADTP	Alcohol Dependence Treatment Programme
Anger Management - Medium Intensity	E01000002	City of London 001B	1B	Assualt on prisoner	2	Lost privileges	2	Servery	BNM+	Becoming New Me +
Anger Management - Low Intensity	E01000003	City of London 001C	2	Detain person against their will	3	Stopped money	3	Prisoner representitive	BF	Breaking Free: Health and Justice Package
Compliance Breach - High Intensity	E01000005	City of London 001E	3	Deny access	4	Segregation	4	Laundry	BBR	Building Better Relationships
Compliance Breach - Medium Intensity	E01000006	Barking and Dagenham 016A	4A	Fight, multiple participants	5	Stopped work	5	Education	C2C	Challenge to Change
Compliance Breach - Low Intensity	E01000007	Barking and Dagenham 015A	4B	Fight, sustained attack	6	Extra days	6	Workshop	COVAID-GS	Control of Violence for Angry Impulsive Drinkers – Group Secure
Domestic Violence - High Intensity	E01000008	Barking and Dagenham 015B	5A	Intentional endangerment			7	Orderly, librarian	COVAID-GSW	Control of Violence for Angry Impulsive Drinkers – Group Secure Women
Domestic Violence - Medium Intensity	E01000009	Barking and Dagenham 016B	5B	Reckless endangerment			8	Orderly, reception	DTC	Democratic Therapeutic Community Mode
Domestic Violence - Low Intensity	E01000011	Barking and Dagenham 016C	6	Intentional obstruction			9	Orderly, healthcare	TC+	Therapeutic Communities Plus
ETE - High Intensity	E01000012	Barking and Dagenham 015D	7A	Escape					HII	Healthy Identity Intervention
ETE - Medium Intensity	E01000013	Barking and Dagenham 013A	7B	Abscond					HSP	Healthy Sex Programme
ETE - Low Intensity	E01000014	Barking and Dagenham 013B	8	Fail to comply with conditions					H	Horizon
Hate Crime - High Intensity	E01000015	Barking and Dagenham 009A	9A	Substance abuse, class A					IM	Identity Matters
Hate Crime - Medium Intensity	E01000016	Barking and Dagenham 009B	9B	Substance abuse, class B or C					K	Kaizen
Hate Crime - Low Intensity	E01000017	Barking and Dagenham 009C	9C	Substance abuce, non-prescribed medication					LNM	Living as New Me
Mental Health - High Intensity	E01000018	Barking and Dagenham 009D	10	Intoxicated					NMS	New Me Strengths
Mental Health - Medium Intensity	E01000019	Barking and Dagenham 023A	11	Alcohol consumption					TBP	The Bridge Programme
Mental Health - Low Intensity	E01000020	Barking and Dagenham 023B	12	Possession of unauthorised article					TSP	Thinking Skills Programme
Restorative Justice - High Intensity	E01000021	Barking and Dagenham 008A	13	Sale of unauthorised article						
Restorative Justice - Medium Intensity	E01000022	Barking and Dagenham 008B	14	Sale of article for personal use						
Restorative Justice - Low Intensity	E01000024	Barking and Dagenham 008D	15	Theft from another prisoner						

Appendix G. MoJ Ontology Cellfie Import Scripts



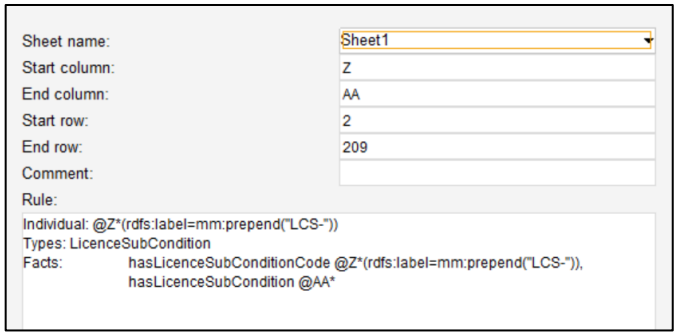
Sheet name: Sheet1
Start column: P
End column: Q
Start row: 2
End row: 208
Comment:
Rule:

```
Individual: @P*(rdfs:label=mm:prepend("Dis-"))
Types: Disposal
Facts:      hasDisposalCode @P*(rdfs:label=mm:prepend("Dis-")),
            hasDisposal @Q*
```



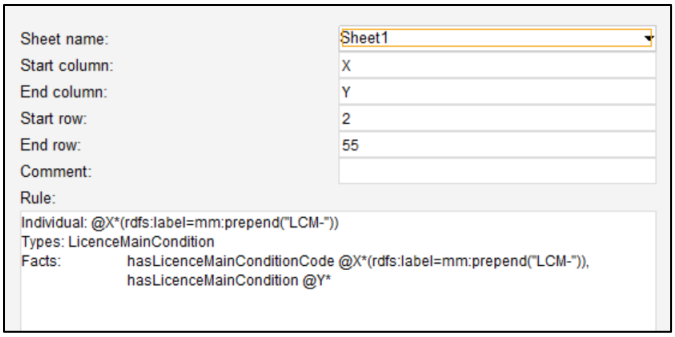
Sheet name: Sheet1
Start column: A
End column: C
Start row: 2
End row: 20
Comment:
Rule:

```
Individual: @A*(rdfs:label=mm:prepend("Eth-"))
Types: @C*
Facts:      hasEthnicityCode @A*(rdfs:label=mm:prepend("Eth-")),
            hasEthnicity @B*
```



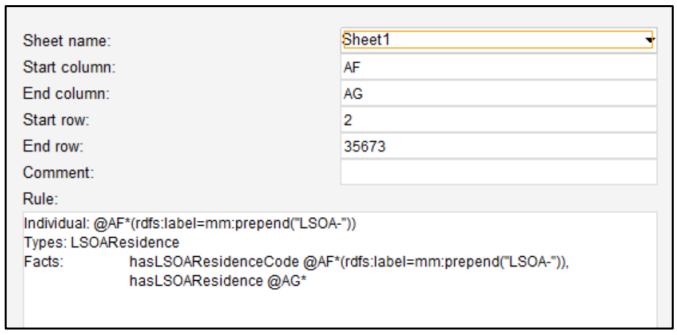
Sheet name: Sheet1
Start column: Z
End column: AA
Start row: 2
End row: 209
Comment:
Rule:

```
Individual: @Z*(rdfs:label=mm:prepend("LCS-"))
Types: LicenceSubCondition
Facts:      hasLicenceSubConditionCode @Z*(rdfs:label=mm:prepend("LCS-")),
            hasLicenceSubCondition @AA*
```



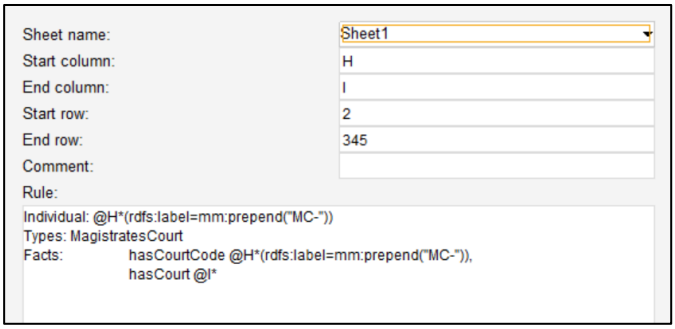
Sheet name: Sheet1
Start column: X
End column: Y
Start row: 2
End row: 55
Comment:
Rule:

```
Individual: @X*(rdfs:label=mm:prepend("LCM-"))
Types: LicenceMainCondition
Facts:      hasLicenceMainConditionCode @X*(rdfs:label=mm:prepend("LCM-")),
            hasLicenceMainCondition @Y*
```



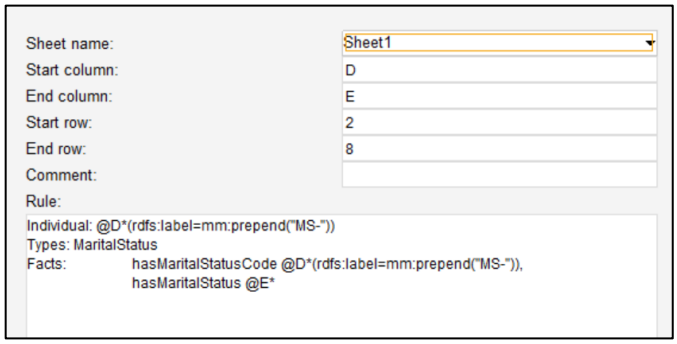
Sheet name: Sheet1
Start column: AF
End column: AG
Start row: 2
End row: 35673
Comment:
Rule:

```
Individual: @AF*(rdfs:label=mm:prepend("LSOA-"))
Types: LSOAResidence
Facts:      hasLSOAResidenceCode @AF*(rdfs:label=mm:prepend("LSOA-")),
            hasLSOAResidence @AG*
```



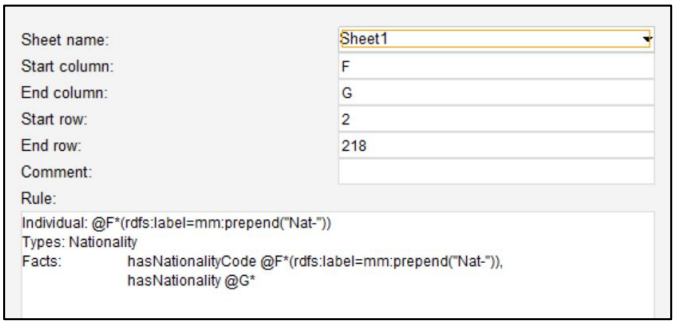
Sheet name: Sheet1
Start column: H
End column: I
Start row: 2
End row: 345
Comment:
Rule:

```
Individual: @H*(rdfs:label=mm:prepend("MC-"))
Types: MagistratesCourt
Facts:      hasCourtCode @H*(rdfs:label=mm:prepend("MC-")),
            hasCourt @I*
```



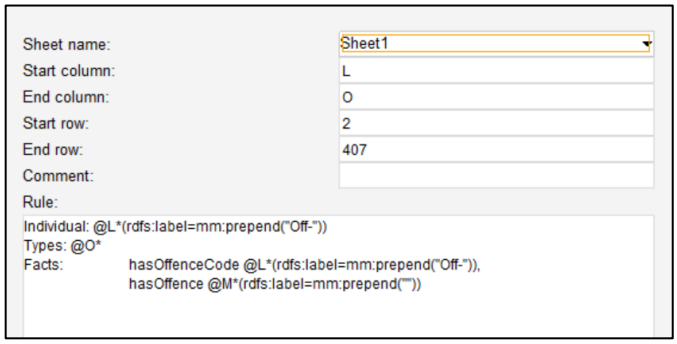
Sheet name: Sheet1
Start column: D
End column: E
Start row: 2
End row: 8
Comment:
Rule:

```
Individual: @D*(rdfs:label=mm:prepend("MS-"))
Types: MaritalStatus
Facts:      hasMaritalStatusCode @D*(rdfs:label=mm:prepend("MS-")),
            hasMaritalStatus @E*
```



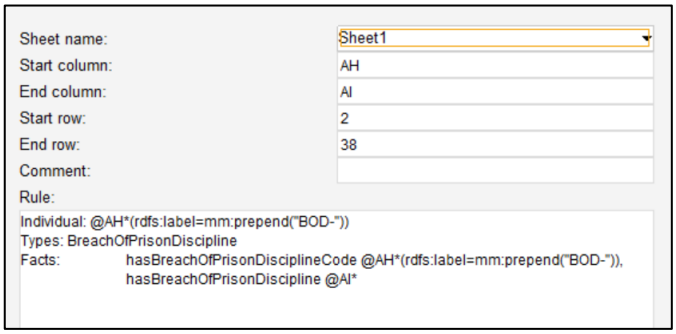
Sheet name: Sheet1
Start column: F
End column: G
Start row: 2
End row: 218
Comment:
Rule:

```
Individual: @F*(rdfs:label=mm:prepend("Nat-"))
Types: Nationality
Facts:      hasNationalityCode @F*(rdfs:label=mm:prepend("Nat-")),
            hasNationality @G*
```



Sheet name: Sheet1
Start column: L
End column: O
Start row: 2
End row: 407
Comment:
Rule:

```
Individual: @L*(rdfs:label=mm:prepend("Off-"))
Types: @O*
Facts:      hasOffenceCode @L*(rdfs:label=mm:prepend("Off-")),
            hasOffence @M*(rdfs:label=mm:prepend(""))
```



Sheet name: Sheet1
Start column: AH
End column: AI
Start row: 2
End row: 38
Comment:
Rule:

```
Individual: @AH*(rdfs:label=mm:prepend("BOD-"))
Types: BreachOfPrisonDiscipline
Facts:      hasBreachOfPrisonDisciplineCode @AH*(rdfs:label=mm:prepend("BOD-")),
            hasBreachOfPrisonDiscipline @AI*
```

Sheet name:	Sheet1
Start column:	AL
End column:	AM
Start row:	2
End row:	10
Comment:	

Rule:

```
Individual: @AL*(rdfs:label=mm:prepend("PW-"))
Types: PrisonWork
Facts:      hasPrisonWorkCode @AL*(rdfs:label=mm:prepend("PW-")),
            hasPrisonWork @AM*
```

Sheet name:	Sheet1
Start column:	R
End column:	S
Start row:	2
End row:	209
Comment:	

Rule:

```
Individual: @R*(rdfs:label=mm:prepend("PA-"))
Types: ProbationArea
Facts:      hasProbationAreaCode @R*(rdfs:label=mm:prepend("PA-")),
            hasProbationArea @S*
```

Sheet name:	Sheet1
Start column:	AB
End column:	AC
Start row:	2
End row:	5
Comment:	

Rule:

```
Individual: @AB*(rdfs:label=mm:prepend("PSSM-"))
Types: PSSMainRequirement
Facts:      hasPSSMainRequirementCode @AB*(rdfs:label=mm:prepend("PSSM-")),
            hasPSSMainRequirement @AC*
```

Sheet name:	Sheet1
Start column:	AD
End column:	AE
Start row:	2
End row:	47
Comment:	

Rule:

```
Individual: @AD*(rdfs:label=mm:prepend("PSSS-"))
Types: PSSSubRequirement
Facts:      hasPSSSubRequirementCode @AD*(rdfs:label=mm:prepend("PSSS-")),
            hasPSSSubRequirement @AE*
```

Sheet name:	Sheet1
Start column:	AN
End column:	AO
Start row:	2
End row:	18
Comment:	

Rule:

```
Individual: @AN*(rdfs:label=mm:prepend("RP-"))
Types: RehabilitationProgramme
Facts:      hasRehabilitationProgrammeCode @AN*(rdfs:label=mm:prepend("RP-")),
            hasRehabilitationProgramme @AO*
```

Sheet name:	Sheet1
Start column:	T
End column:	U
Start row:	2
End row:	71
Comment:	

Rule:

```
Individual: @T*(rdfs:label=mm:prepend("SRM-"))
Types: SentenceMainRequirement
Facts:      hasSentenceMainRequirementCode @T*(rdfs:label=mm:prepend("SRM-")),
            hasSentenceMainRequirement @U*
```

Sheet name:	Sheet1
Start column:	V
End column:	W
Start row:	2
End row:	328
Comment:	

Rule:

```
Individual: @V*(rdfs:label=mm:prepend("SRS-"))
Types: SentenceSubRequirement
Facts:      hasSentenceSubRequirementCode @V*(rdfs:label=mm:prepend("SRS-")),
            hasSentenceSubRequirement @W*
```

Sheet name:	Sheet1
Start column:	AJ
End column:	AK
Start row:	2
End row:	7
Comment:	

Rule:

```
Individual: @AJ*(rdfs:label=mm:prepend("DBS-"))
Types: DisciplineBreachSanction
Facts:      hasDisciplineBreachSanctionCode @AJ*(rdfs:label=mm:prepend("DBS-")),
            hasDisciplineBreachSanction @AK*
```

Sheet name:	Sheet1
Start column:	A
End column:	E
Start row:	2
End row:	41
Comment:	

Rule:

```
Individual: @A*(rdfs:label=mm:prepend("Feature:"))
Types: MoJFeature
Facts:      hasMaleRecidivismCorrelationValue @B*(xsd:float),
            hasMaleRaceCorrelationValue @C*(xsd:float),
            hasFemaleRecidivismCorrelationValue @D*(xsd:float),
            hasFemaleRaceCorrelationValue @E*(xsd:float)
```

Sheet name:	Sheet1
Start column:	J
End column:	K
Start row:	2
End row:	78
Comment:	

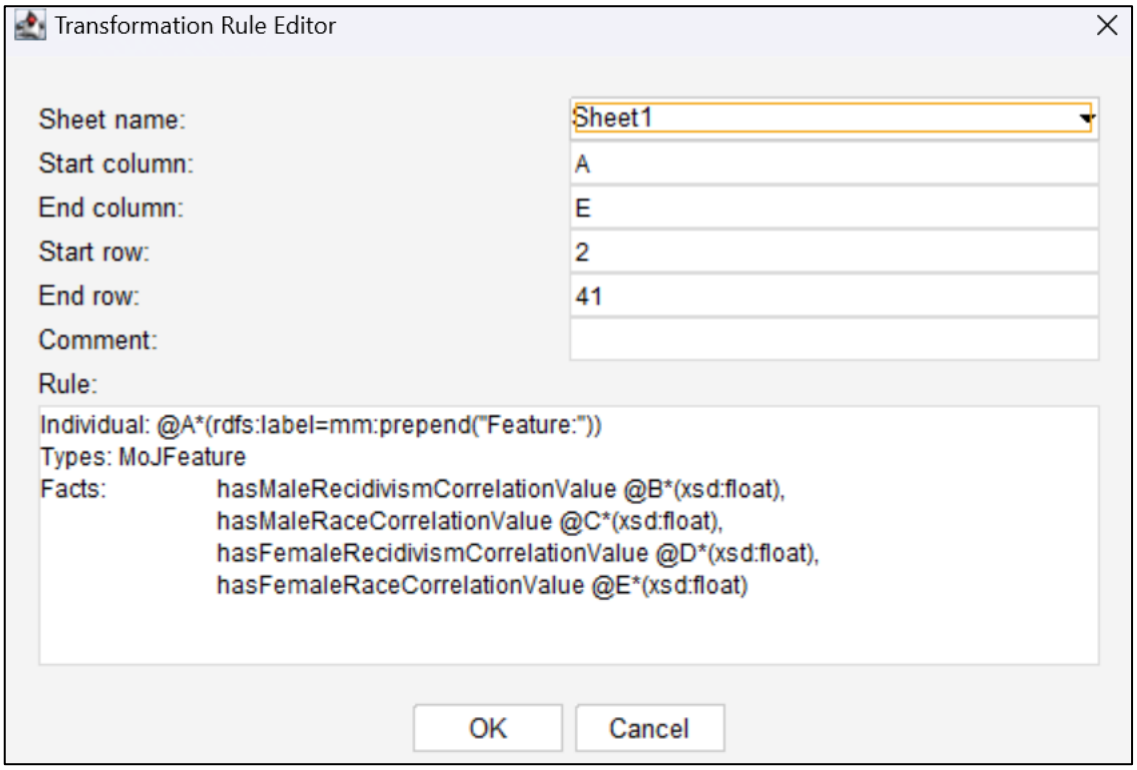
Rule:

```
Individual: @J*(rdfs:label=mm:prepend("CC-"))
Types: CrownCourt
Facts:      hasCourtCode @J*(rdfs:label=mm:prepend("CC-")),
            hasCourt @K*
```

Appendix H. MoJ Feature Correlation Instances

Feature	MaleRecidivismCorrelationValue	MaleRaceCorrelationValue	FemaleRecidivismCorrelationValue	FemaleRaceCorrelationValue
DateOfBirth	0.01	0.03	0.04	0.05
Age	0.177	0.121	0.133	0.072
NumberOfPreviousConvictions	0.164	0.079	0.247	0.04
MonthsSinceLastConviction	0.242	0.132	0.298	0.163
MonthsSinceLastImprisonment	0.281	0.142	0.312	0.172
ResidentialStatus	0.01	0.03	0.04	0.05
MonthsAtAddress	0.213	0.095	0.194	0.071
GangAffiliation	0.185	0.086	0.093	0.043
EducationLevel	0.088	0.057	0.039	0
Employment	0.217	0.126	0.227	0.059
NumberOfDependents	0.031	0.096	0.007	0.064
AlcoholDependent	0.132	0.098	0.103	0.084
ClassADrugDependent	0	0.121	0.053	0.126
ClassBDrugDependent	0	0.128	0	0.121
EthnicityCode		1		1
NationalityCode	0.042	0.132	0.016	0.187
MaritalStatusCode	0.153	0.032	0.198	0.021
LSOAResidenceCode	0.123	0.143	0.123	0.143
OffenderAgeAtOffence	0.162	0.102	0.12	0.043
PrincipleOffenceCode	0.161	0.092	0.141	0.089
MostSeriousOffenceCode	0.082	0.162	0.052	0.172
CourtCode	0.02	0.02	0.02	0.02
ProbationAreaCode	0.02	0.01	0.02	0.01
PrincipleDisposalCode	0.075	0.021	0.012	0.043
PrincipleDisposalDays	0.092	0.021	0.083	0.032
PrisonDurationDays	0.162	0.092	0.159	0.089
FineAmountPounds	0.043	0.012	0.032	0.011
LIcencePeriodDays	0.01	0.03	0.04	0.05
DrivingPenaltyPoints	0.052	0.024	0.063	0.012
OrderLengthDays	0.112	0.101	0.111	0.092
BreachOfPrisonDisciplineCode	0.096	0.123	0.076	0.072
DisciplineBreachSanctionCode	0.072	0.087	0.043	0.023
PrisonWorkCode	0.123	0.078	0.111	0.062
RehabilitationProgrammeCode	0.162	0.053	0.151	0.042
SentenceMainRequirementCode	0.092	0.121	0.078	0.101
SentenceSubRequirementCode	0.098	0.143	0.087	0.098
LicenceMainConditionCode	0.092	0.076	0.087	0.054
LicenceSubConditionCode	0.099	0.045	0.084	0.034
PSSMainRequirementCode	0.096	0.023	0.04	0.05
PSSSubRequirementCode	0.152	0.056	0.123	0.023

Appendix I. MoJ Feature Correlation Cellfie Import Script



Appendix J. MoJ Offender and Case Dummy Instances

Offender ID	Name	Date of Birth	Age	Number of previous convictions	Months since last conviction	Months since last imprisonme	Residential status	Months at address	Gang affiliation
511985165	John Smith	1962-12-03	61	5	10	299	Homeless	36	false
141806441	Lillian Harpole	1965-03-05	59	0	2		Homeowner	52	false
486623255	Zachary Nixon	1951-02-06	73	3	10	0	Rented	13	false
613456052	George Chadwick	1954-12-16	69	1	27	3	Homeowner	62	false
112299926	Scarlett Morgan	1961-04-05	63	4	11		Homeless	13	false
646955585	Muhammad King	1982-05-17	42	0	7		Rented	9	false
341786705	Grace Fitzgerald	1978-03-22	46	3	8	0	Homeowner	51	false
983190851	Jack Harrison	1998-02-21	26	0	15		Rented	3	false
815709338	Natasha Murphy	2003-10-03	20	0	4		Living with parents	58	false
817633970	Christopher Robson	1994-10-28	29	4	1		Homeowner	37	false
265466553	Aaliyah Martin	2000-04-16	24	0	16		Rented	15	false
547242054	Stephen Johnson	2001-06-12	23	4	1	0	Living with parents	31	false

Education level	Employment	Number of dependents	Alcohol dependent	Clas A drug dependent	Class B drug dependent	Gender	Ethnicity code	Nationality code	Marital status code	LSOA residence code
High School	false	0	true	false	false	Male	W1	BRIT	S	E01000012
None	false	2	false	false	false	Female	W1	BRIT	M	E01000031
High School	false	1	false	true	false	Male	O2	BRIT	W	E01000249
High School	false	3	true	true	false	Male	B2	BRIT	M	E01000250
High School	false	4	true	true	true	Female	W1	BRIT	D	E01000251
Graduate	false	0	false	false	false	Male	A9	IRAN	S	E01000252
High School	false	2	true	true	false	Female	M3	BRIT	M	E01000253
High School	false	1	false	false	false	Male	B2	BRIT	M	E01000254
High School	false	0	false	false	false	Female	B2	BRIT	S	E01000255
Post Graduate	false	2	true	false	false	Male	W1	BRIT	S	E01000256
High School	false	0	true	false	false	Female	O2	BRIT	M	E01000257
High School	false	0	false	true	true	Male	W1	BRIT	S	E01000258

Case ID	Offender ID	Date of offence	Offender age at offence	Principle offence code	Most serious offence code	Court code	Probation area code	Principle disposal code	Principle disposal days	Prison duration days	Fine amount pounds	Licence period days	Driving penalty points	Order length days	Breach of prison discipline code	Breach of prison discipline code	Breach of prison discipline code	Breach of prison discipline code	Breach of prison discipline code
904791425	511985165	1993-05-12	30	3400	3400	CC-36	WTS	SS	730		525			730					
618323430	511985165	1998-01-21	35	3400	3400	CC-36	WTS	IMP	365	365		183		365	11	16A	12	1A	13
587659755	511985165	2002-06-02	39	4600	4600	MC-1894	WTS	FINE			200								
113345537	511985165	2016-03-13	53	4600	4600	MC-1894	WTS	FINE			125								
535754077	511985165	2018-08-05	55	4510	4510	MC-1894	WTS	FINE			290								
499266153	511985165	2023-02-16	60	4600	4600	CC-36	WTS	SS	365					365					
637473147	141806441	2023-10-13	58	806	806	MC-2870	C14	CPO	365					365					
478663808	486623255	1992-12-03	41	4600	4600	MC-1922	N59	FINE			150								
348844983	486623255	2022-09-22	71	806	806	CC-77	N59	CPO	730					730					
338060576	486623255	2023-02-16	72	831	831	CC-77	N59	IMP	730	730		365		730	17A	1B			
723372549	613456052	2003-02-02	48	4600	4600	MC-1922	N59	FINE			255								
808774974	613456052	2021-09-04	66	5301	5301	CC-77	N59	IMP	760	760		365		760	16B				
691205173	112299926	2020-06-12	59	4600	4600	MC-2330	NTS	FINE			150								
214350956	112299926	2021-05-04	60	4600	4600	MC-2330	NTS	FINE			240								
647325366	112299926	2021-12-03	60	4600	4600	MC-2330	NTS	FINE			110								
745987478	112299926	2022-09-03	61	4600	4600	MC-2330	NTS	FINE			255								
986467303	112299926	2023-01-12	61	4600	4600	MC-2330	NTS	FINE			120								
595710701	646955585	2023-05-12	40	4600	4600	MC-1943	HLI	FINE			215								
527235880	341786705	2020-01-01	41	9261	4600	MC-2978	N55	FINE			100								
794276181	341786705	2021-03-12	42	9250	9250	MC-2978	N55	CPO	365					365					
830828808	341786705	2022-09-20	44	9250	9250	MC-2978	N55	SS	365					365					
905424044	341786705	2023-04-23	45	406	406	CC-30	N55	IMP	730	730		365	6	730	22				
281645560	983190851	2022-12-05	24	835	835	CC-30	N55	CPO	730					730					
782821873	817633970	2023-08-29	19	835	835	MC-1613	ESX	FINE			210								
112529789	815709338	2022-06-09	27	4600	4600	MC-1051	BED	FINE			175								
861603678	815709338	2022-12-21	28	4600	4600	MC-1051	BED	FINE			205								
202280393	815709338	2023-02-01	28	4600	4600	MC-1051	BED	FINE			180								
755833865	815709338	2023-10-10	28	4600	4600	MC-1051	BED	CPO	365		50			365					
178906509	815709338	2024-03-23	29	4600	4600	CC-79	BED	SS	548					548					
196164791	265466553	2022-11-24	22	4600	4600	MC-1051	BED	FINE			130								
851372657	547242054	2022-08-10	21	4600	9250	MC-1051	BED	FINE			210								
254596091	547242054	2022-11-27	21	9250	9250	CC-79	BED	CPO	365					365					
375098237	547242054	2023-03-01	21	9250	9250	CC-79	BED	CPO	365					365					
607333684	547242054	2023-09-26	22	9250	9250	CC-79	BED	SS	548					548					
569872265	547242054	2023-12-12	22	9210	9210	CC-79	BED	IMP	1277	1277		365		1277	13	15	1B	1A	

Discipline breach sanction code	Discipline breach sanction code	Discipline breach sanction code	Discipline breach sanction code	Discipline breach sanction code	Prison work code	Prison work code	Prison work code	Prison work code	Prison work code	Rehabilitation programme code	Rehabilitation programme code	Rehabilitation programme code	Rehabilitation programme code	Rehabilitation programme code
1	3	2	4	5	6					LNM				
4	2	3								BBR	BNM+			
1					1	2				IM				
2										TBP				
2	5	3	4	6	1					LNM				

Sentence requirements start date	Sentence requirements end date	Sentence main requirement code	Sentence sub requirement code	Sentence sub requirement code	Sentence sub requirement code	Sentence sub requirement code	Sentence sub requirement code	Licence conditions start date	Licence conditions end date	Licence main condition code	Licence sub condition code	Licence sub condition code	Licence sub condition code	Licence sub condition code	Licence sub condition code	PSS start date	PSS end date	PSS main requirement code	PSS sub requirement code	PSS sub requirement code	PSS sub requirement code	PSS sub requirement code	PSS sub requirement code
1993-12-02	1995-12-01	3	703	RS111																			
								1999-05-26	1999-11-25	EM01	701					1999-11-26	2000-05-25	3	SA20				
2023-05-11	2024-05-10	T	702	ALCMON																			
2023-10-17	2024-10-16	W	703	706																			
2023-01-18	2024-01-17	W	ALCMON	GMP007	706																		
								2024-07-15	2025-07-14	LCF	D03	LC147											
								2023-02-05	2024-02-04	EM01	701	LC139											
2021-08-02	2022-08-01	W	724																				
2023-01-19	2024-01-18	X	ESX011	RS156																			
								1999-05-26	1999-11-26	LA5	LC180	LC238	LC260	LC76									
2023-04-10	2025-04-09	3																					
2024-02-02	2025-02-01	W	731																				
2024-05-05	2025-05-04																						
2023-03-03	2024-03-02	W																					
2023-07-17	2024-07-16	W																					
2023-12-12	2025-06-11	X	731																				
								2026-02-16	2027-10-12	EM01	Y01	LC262											

Appendix K. MoJ Offender and Case Cellfie Import Scripts