PROJECT REPORT: ISAS Iteration of Analysis of gender bias in policing, to address the problem of gender inequality.

By Ahmed Suhail
asuh702
for INFOSYS722

Contents

1. Situation Understanding 2
2. Data Understanding 2
3. Data Preparation: 4
4. Data Transformation 6
5. Data Mining Method Selection: 6
6. Data Mining Algorithm Selection 7
7. Data Mining: 8
8. Interpretation: 10
9. Future Work: 12
10. Acknowledgement 12
11. References 12

1. Situation Understanding

1.1. Objectives of the situation: Analysis of gender bias in policing, to address the problem of gender inequality. How often does gender come into play when traffic cops stop motorists on the road? For several years now, especially since the Black Lives Matter movement started, people have advocated for better policing. There is an immediate need for analysis of how policing is done, spot discrimination (if any!). There is also an immediate need to empower women and address the problem of Gender Equality, as is stated in the UN's Sustainable Development Goals (https://bit.ly/1Qk5cqI).

1.2. An interdisciplinary team of statisticians, computer scientists, and journalists at Stanford University came together to study possible bias in policing. They "filed public record requests with all 50 states to obtain details of each stop carried out by state patrol officers over the last 10 years" (Pierson et al., 2017). They "found that black and Hispanic drivers were often searched based on lesser evidence than whites. This double standard was widespread, not confined to any one state or geographic region." Data was then made to the available to the public, to facilitate further discussion of policing practices. The team hoped, "this resource helps researchers, policymakers, and journalists assess and improve police contact with the public."(Sharad, Goel & Philips, 2017) A problem with crime analysis is data is often private, and not all parts are safe for release to the public. Similar was the case with this project, wherein to protect the privacy of individuals, number plate details of the drivers and the location of the stop was removed. Also, police officers might have been biased in their stops and might have falsely added some attributes, such as listing suspicious behaviour as reason for search when in actuality the person was just fine.Data Mining Objectives

1.3. Data Mining Objectives

i. Analysis of gender discrimination in police stops in parts of USA (Rhode Island for this report)
ii. Evaluate the effects of time, race, and other features available on traffic stops by the police.

1.4. Data Mining Success Criteria

i. Successful examination and study of the rate at which male and female drivers are stopped and arrested, of the hypotheses that arrests are discriminatory and being able to state with confidence if they are/are not.
ii. Perform different types of modelling and gather successful insights

2. Data Understanding

2.1. Data Collection: Data Collection: An issue when collecting data for crime analysis is that law enforcement agencies do not make data publicly available because of vested interests and privacy concerns. Police data is often inaccessible, making rigorous analysis difficult. A database of over 100 million records from 31 states was collected by a research team at Stanford University. All the data collected was cleaned to some extent and made available for public use, in hope that data analysts over the world can dive deeper into the data to more insights and build upon their work. Link to the project: https://openpolicing.stanford.edu/

2.2. Data Description:

i. The data is in files of the format .csv (Comma Separated Values) and .rds for RStudio. The data was converted to .xls for loading into SPSS Modeler, which was because SPSS has problems loading .csv using the Var. File from Source nodes.
ii. Data was publicly available for download around June 2017, for analysis and help in getting better insights. By studying location-specific data and evaluating effects of different features on stops, better analyses could be made.

iii. For my SPSS Iteration, I've chosen to focus on data from Rhode Island. The State Patrol from Rhode Island had provided data of all vehicular stops from January 2005 to December 2015.

iv. The available data consisted of the following features:

- raw_row_number: A record ID for reference to the raw data
- Date: Stop date in the format of YYYY-MM-DD
- Stop Time: Time of the Stop, in a 24 hour, HHMM format.
- Stop Location: Textual, non-standardised location.
- Driver Race: Race of the stopped subject.
- Driver Age: Age of the stopped subject.
- Driver Sex: Gender of the stopped subject.
- Search Conducted: Boolean value that indicates whether a search was conducted.
- Contraband Found: Boolean value that indicates if contraband was found
- Citation Issued: Boolean value indicating outcome, if a citation was issued.
- Warning Issued: If a stop resulted in the issue of warning, this Boolean was selected.
- Frisk Performed: If a quick Frisk was performed, this Boolean value was set True
- Arrest Made: If the person were arrested, this Boolean value was set True.
- reason_for_stop: The visible reason for stopping the driver.

2.3. Data Exploration: Initial exploration of data reveals the need to prepare data and fix errors, and missing values before plots for better understanding can be made. A Distribution of subject_sex reveals that more number of males have been stopped than females (which in no way can relate to discrimination unless we analyze base population statistics and adjust for other factors)

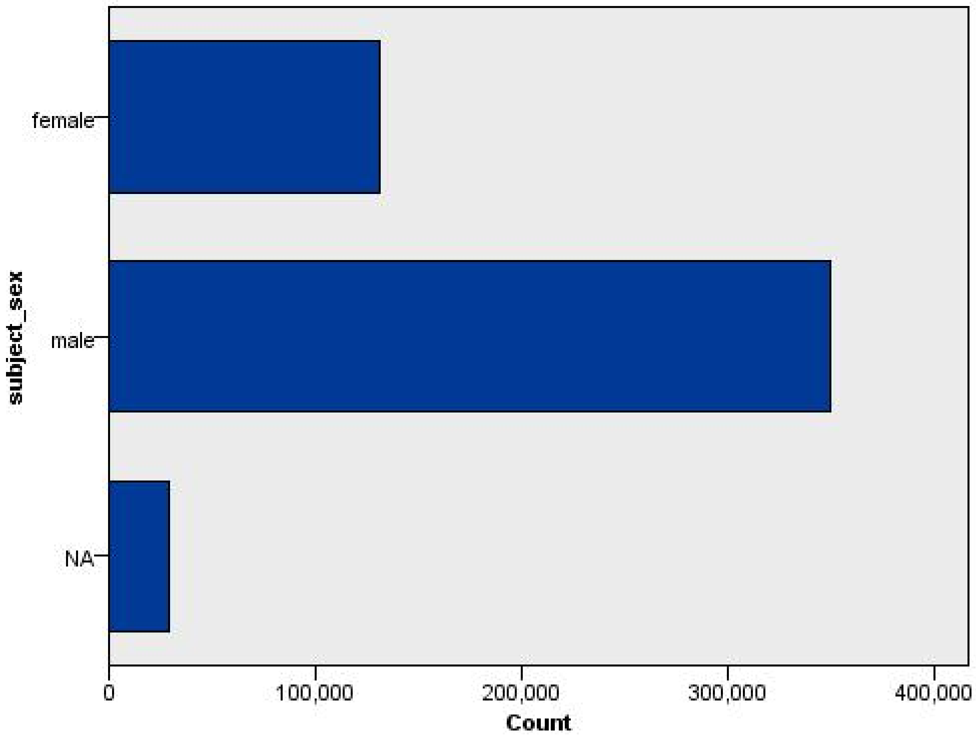
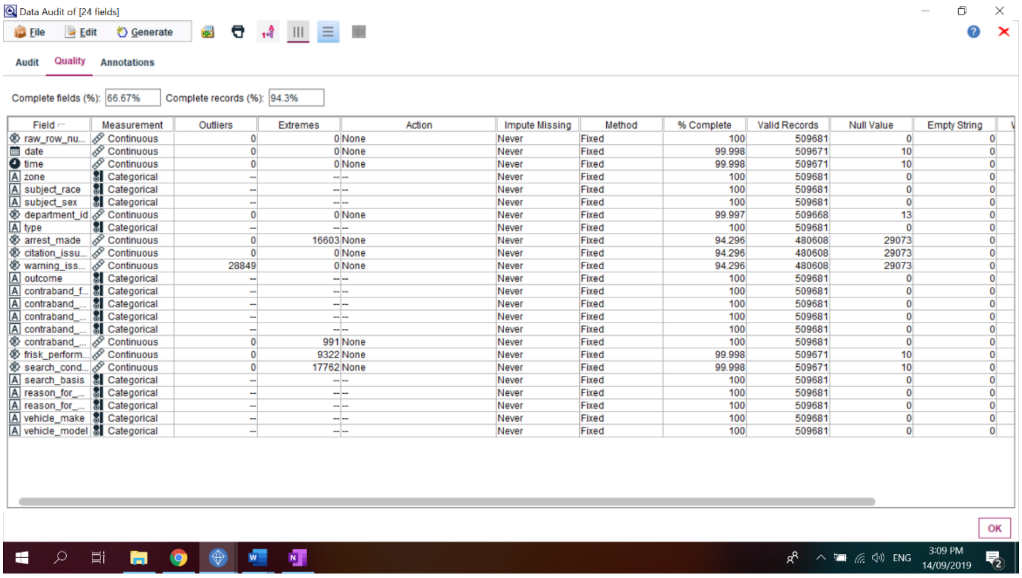


Figure 1: Distribution of subject_sex

2.4. Data Quality: The Data Audit node run from SPSS Modeler indicated good quality data, with thoughts that very less time would be spent on Data Preparation, but since I was also working on my OSAS Iteration, I knew that was not the case. Data being saved in .xls format resulted in NA values in the dataset from CSV to be rendered as valid values, which needed to be corrected (as seen from table node).


Data Audit of [24 fields]

File Edit Generate

Audit Quality Annotations

Complete fields (%): 66.67% Complete records (%): 94.3%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String
raw_row_nu...	Continuous	0	0	None	Never	Fixed	100	509681	0	0
date	Continuous	0	0	None	Never	Fixed	99.998	509671	10	0
time	Continuous	0	0	None	Never	Fixed	99.998	509671	10	0
zone	Categorical	--	--	--	Never	Fixed	100	509681	0	0
subject_race	Categorical	--	--	--	Never	Fixed	100	509681	0	0
subject_sex	Categorical	--	--	--	Never	Fixed	100	509681	0	0
department_id	Continuous	0	0	None	Never	Fixed	99.997	509668	13	0
type	Categorical	--	--	--	Never	Fixed	100	509681	0	0
arrest_made	Continuous	0	16603	None	Never	Fixed	94.296	480608	29073	0
citation_issu...	Continuous	0	0	None	Never	Fixed	94.296	480608	29073	0
warning_iss...	Continuous	28849	0	None	Never	Fixed	94.296	480608	29073	0
outcome	Categorical	--	--	--	Never	Fixed	100	509681	0	0
contraband_f...	Categorical	--	--	--	Never	Fixed	100	509681	0	0
contraband_...	Categorical	--	--	--	Never	Fixed	100	509681	0	0
contraband_...	Categorical	--	--	--	Never	Fixed	100	509681	0	0
contraband_...	Categorical	--	--	--	Never	Fixed	100	509681	0	0
contraband_...	Continuous	0	991	None	Never	Fixed	100	509681	0	0
frisk_perform...	Continuous	0	9322	None	Never	Fixed	99.998	509671	10	0
search_cond...	Continuous	0	17762	None	Never	Fixed	99.998	509671	10	0
search_basis	Categorical	--	--	--	Never	Fixed	100	509681	0	0
reason_for_...	Categorical	--	--	--	Never	Fixed	100	509681	0	0
reason_for_...	Categorical	--	--	--	Never	Fixed	100	509681	0	0
vehicle_make	Categorical	--	--	--	Never	Fixed	100	509681	0	0
vehicle_model	Categorical	--	--	--	Never	Fixed	100	509681	0	0

OK

3:09 PM 14/09/2019

Figure 2: Data Audit from SPSS Modeler

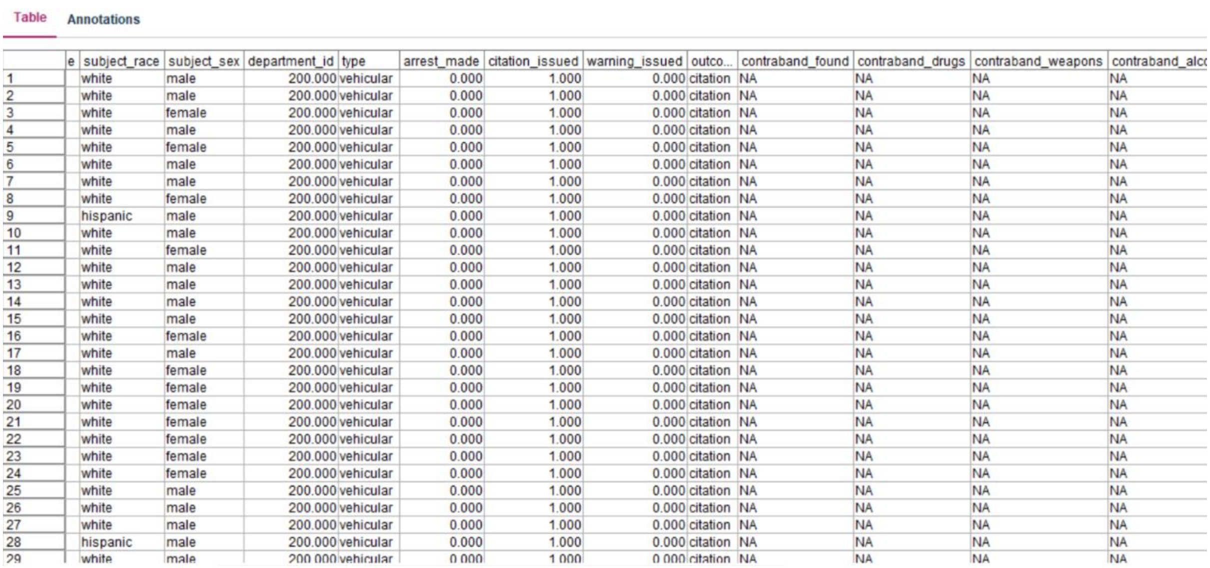


Table Annotations

	e	subject_race	subject_sex	department_id	type	arrest_made	citation_issued	warning_issued	outco...	contraband_found	contraband_drugs	contraband_weapons	contraband_alco
1		white	male	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
2		white	male	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
3		white	female	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
4		white	male	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
5		white	female	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
6		white	male	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
7		white	male	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
8		white	female	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
9		hispanic	male	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
10		white	male	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
11		white	female	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
12		white	male	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
13		white	male	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
14		white	male	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
15		white	male	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
16		white	female	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
17		white	male	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
18		white	female	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
19		white	female	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
20		white	female	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
21		white	female	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
22		white	female	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
23		white	female	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
24		white	female	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
25		white	male	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
26		white	male	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
27		white	male	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
28		hispanic	male	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA
29		white	male	200.000	vehicular	0.000	1.000	0.000	citation	NA	NA	NA	NA

Figure 3: Table Node from SPSS Indicates presence of NA Values

3. Data Preparation:

3.1. A Type node was used for instantiation because the dataset was large. As per SPSS documentation, “To optimize system resources, instantiating is a user-directed process—you tell the software to read values by specifying options on the Types tab in a source node or by running data through a Type node.”
Selection of the correct type of data involved selecting the features that might be known to the officer prior to the stop, which would be relevant to being analysed for discrimination. The key to building good models was selecting such data and not get in linearly related data for the types of models that I had in mind. Over the course of Iterative building, several features were sieved out that were yielded as insignificant by the models.

3.2. There were a lot of missing values in the dataset, which could not have been fed to the model. A Missing values filter node was run to filter out such data, and only valid values were selected, as the proportion of missing data in comparison to the data present in the report was insignificant. During the instantiation, the types of data were set, with time and date using the timestamps and other values correctly being set. For analysis of discrimination in arrests, only records where arrests were made were selected using the Select node from under the Record ops.

3.3. The date string was used to create date and timestamps for better modelling.

Table Annotations

	ntraband_weapons	contraband_alcohol	contraband_other	frisk_performed	search_conducted	search_basis	reason_for_search	reason_for_stop	vehicle_make	vehicle_model
480580	.	NA	0.000	0.000	0.000	NA	NA	Equipment/Inspecti...	CHEV	TAHOE
480581	.	NA	0.000	0.000	0.000	NA	NA	Other Traffic Violation	HOND	NA
480582	.	NA	0.000	0.000	0.000	NA	NA	Seatbelt Violation	LEXS	ES3
480583	.	NA	0.000	0.000	0.000	NA	NA	Other Traffic Violation	NISS	FRONTIER
480584	.	NA	0.000	1.000	1.000	probable cau...	Reasonable Susp...	Speeding	CHEV	MONTE CAR...
480585	.	NA	0.000	0.000	0.000	NA	NA	Speeding	DODG	RAM
480586	.	NA	0.000	0.000	0.000	NA	NA	Call for Service	HOND	CIVIC LX
480587	.	NA	0.000	0.000	0.000	NA	NA	Other Traffic Violation	ACUR	RL
480588	.	NA	0.000	0.000	0.000	NA	NA	Other Traffic Violation	TOYT	COROLLA
480589	.	NA	0.000	0.000	0.000	NA	NA	Speeding	MERZ	GL450
480590	.	NA	0.000	0.000	0.000	NA	NA	Speeding	NA	NA
480591	.	NA	0.000	0.000	0.000	NA	NA	Other Traffic Violation	NA	NA
480592	.	NA	0.000	0.000	0.000	NA	NA	Other Traffic Violation	NISS	ALTIMA
480593	.	NA	0.000	0.000	0.000	NA	NA	Speeding	TOYT	CAMRY
480594	.	NA	0.000	0.000	0.000	NA	NA	Speeding	KIA	SOL
480595	.	NA	0.000	0.000	0.000	NA	NA	Speeding	BMW	328I
480596	.	NA	0.000	0.000	0.000	NA	NA	Equipment/Inspecti...	HOND	ACCORD
480597	.	NA	0.000	0.000	0.000	NA	NA	Speeding	HOND	ACCORD
480598	.	NA	0.000	0.000	0.000	NA	NA	Speeding	ACUR	TSX
480599	.	NA	0.000	0.000	0.000	NA	NA	Speeding	TOYT	COROLL
480600	.	NA	0.000	0.000	0.000	NA	NA	Other Traffic Violation	PONT	BONNEVILLE
480601	.	NA	0.000	0.000	0.000	NA	NA	Equipment/Inspecti...	NISS	MAXIMA
480602	.	NA	0.000	0.000	0.000	NA	NA	Equipment/Inspecti...	JEEP	GRAND CHE...
480603	.	NA	0.000	0.000	0.000	NA	NA	Speeding	NA	NA
480604	.	NA	0.000	0.000	0.000	NA	NA	Equipment/Inspecti...	FORD	TAURUS
480605	.	NA	0.000	0.000	0.000	NA	NA	Speeding	VOLK	BUG

OK

Figure 4: Missing Values Select Node Table

3.4. The filter node was used at many places, iteratively in the process of model optimisation. For analysis of discrimination in arrests, only records where arrests were made were selected using the Select node from under the Record ops. A Type Node was connected through a year generation node, which could be plotted for analysis of arrests over the years.

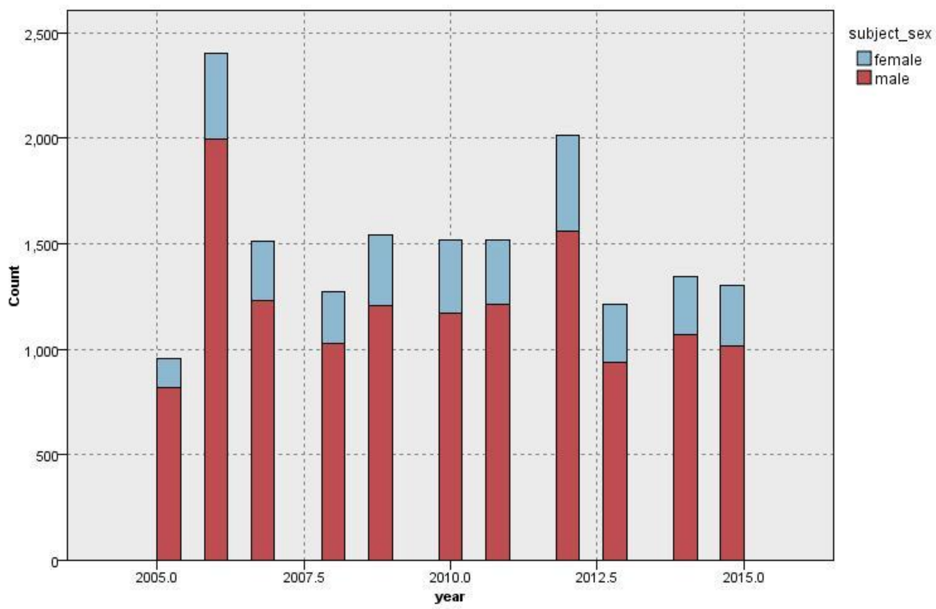


Figure 5: Histogram of arrests of males and females over the years.

4. Data Transformation

4.1. For the initial clustering iteration, the features relevant to finding discrimination, i.e., the features an officer would know before making an arrest were chosen and the rest were filtered out using a Filter Node under Field ops. The records were then fed for Associative Modeling.

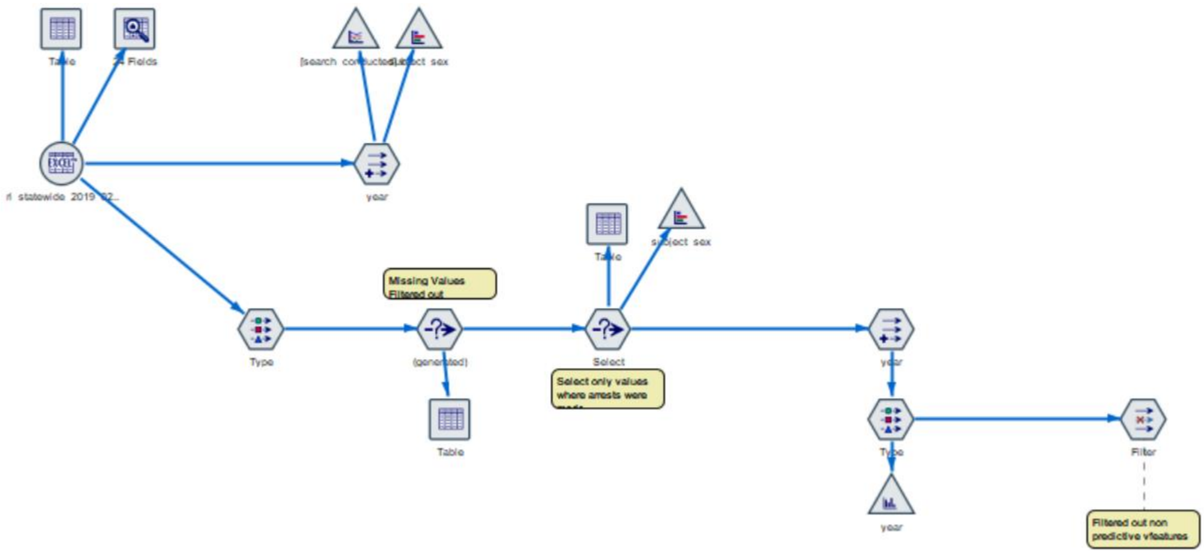


Figure 6: Data Transformation by filtering out co-related values not suitable for Logistic Regression

4.2. In some places where the normal distribution of data wasn't present, a balance (reduce) node was used to balance the distribution of data (when checked from the distribution graph node)

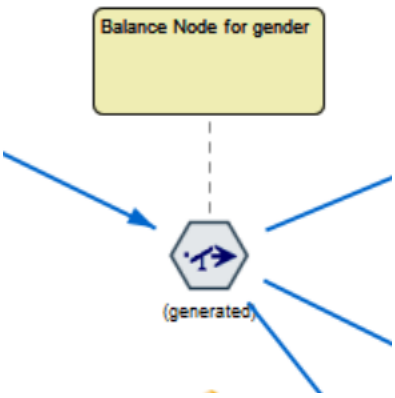


Figure 7: Transformation through balancing for gender.

5. Data Mining Method Selection:

5.1. The first kind of modeling I'd thought would yield results was Associative Modeling. Since association rules have been used historically to relate features to one another, it was common sense to proceed in this way. Associative algorithms couldn't give out results in SPSS Modeler, with some such as Apriori failing to formulate any rules. Others which made rules weren't interpretable/helpful.

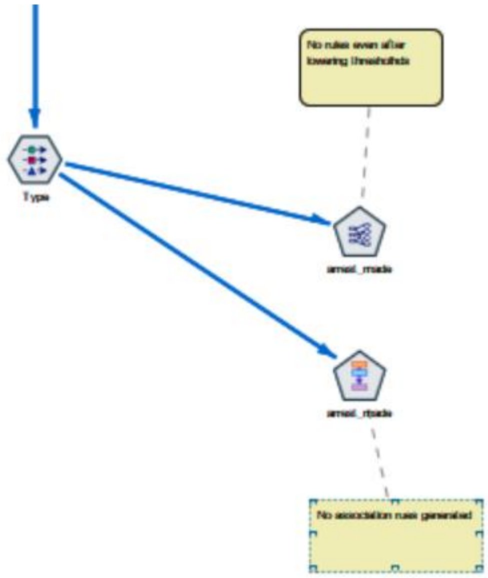


Figure 8: Associative Modeling didn't help establish any associations, even with low Thresholds

5.2. Failure of Associative Modeling led me to consider other modeling techniques that might be useful. Rather unconventional use of predictive modeling I'd found some people talking about was the use of odds ratio in discriminatory analysis(Ed, n.d.). The idea is that if a predictor is binary, odds ratio can be safely used to interpret discrimination. Evaluating from the data mining goals, this would be a good way to test out if the police officers discriminated on gender grounds, and hence very relevant to analysis for gender inequality in police stops.

6. Data Mining Algorithm Selection

6.1. The association rules and Apriori algorithms from the SPSS Modeler were used, decreasing values of confidence, lift and other parameters but to no good associations in the data. Logistic regression is the most suitable algorithm for this kind of analysis on this data (dependent variable binary, meaningful variables, independent variables independent to one another, large size dataset, not much computational resources are required, highly interpretable, not much data transformation was needed) which led me to run Logistic regression on the data. Logistic over linear regression was chosen because of no linear relationship between dependent and independent variables. Some people might also have advocated for discriminant analysis, but my research found that for not a categorization problem, Logistic Regression would be better. Logistic regression is more of a classification algorithm, unlike Linear 'regression'.

6.2. To run logistic regression, a type node was used to instantiate data. Logistic Regression was then run on the dataset to look for odds ratio, in expert mode, so advance statistics could be generated.

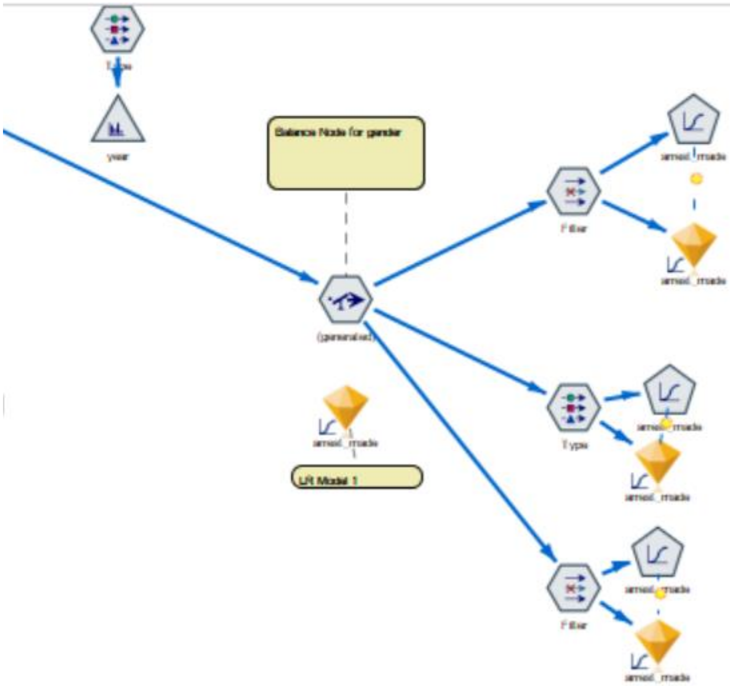


Figure 9: LR being run with different parameters and fields for best results

6.3. Multiple Iterations of Logistic Regression were run on the dataset, with parameters, features, being tuned each time, like filtering out features playing an insignificant role in modeling, balancing for gender, balancing for arrest_made (also balancing for both!), etc. Reduce nodes were used from analysis of distribution of these features. Refer to the DMAS stream to see the different branches with different parameters.

7. Data Mining:

7.1. After running multiple iterations with different parameters, a Logistic Regression Model was built. Train and test were 70/30 (universally accepted ratios of train and test sizes so that enough data is available for training and you also have enough data for validating your results).

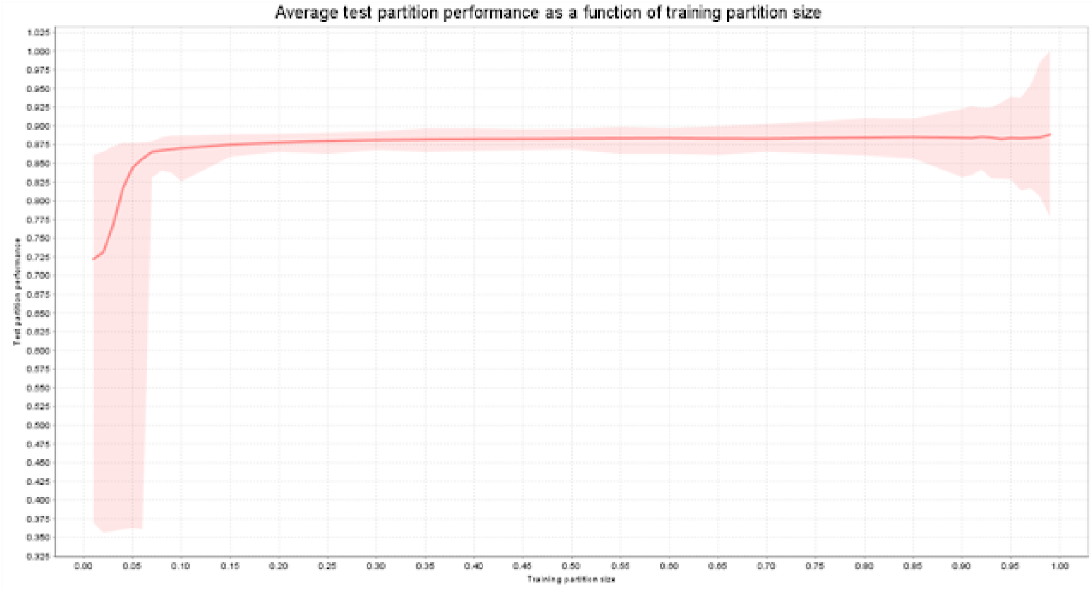


Figure 10: An evaluation of test train splits, showing saturation after a point in accuracy even if test data is increased. Link: http://information-gain.blogspot.com/2012/07/why-split-data-in-ratio-7030.html

7.2. After running multiple iterations with different types of instantiated and different values of partitions another stream was built with missing values filtered out and a Logistic Regression model was built

Figure 11: Train, Test Validation for Accuracy and Model run with improved parameters

7.3. The model was successfully fit into the data and gave the following metrics of performance:

i. The model with only the subject sex as an input yielded insignificant results.

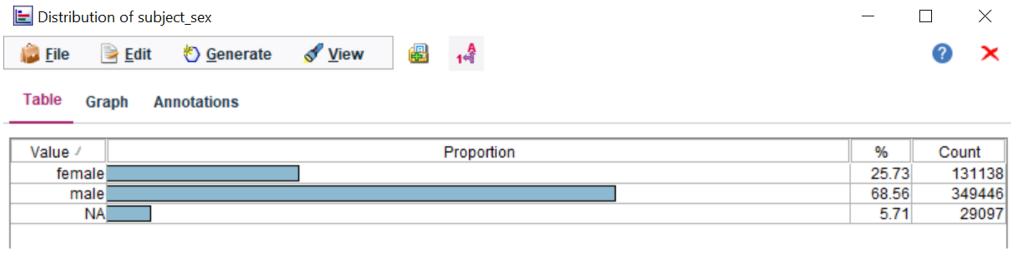
Equation For 0
Equation For 1
11.84 * [subject_sex=female] +
12.23 * [subject_sex=male] +
+ -15.47

Figure 12: model with only subject_sex

ii. The model with vehicle make couldn't yield significant results, so the unique values of vehicles were grouped into fields such as Hatchback, SUV, other, but this also yielded insignificant results. The most interpretable model is hence indicated in the stream and insights are highlighted in the section 8

8. Interpretation:

8.1. In general, the plots made of the data show that there are possibilities that cops in Rhode Island are discriminatory when it comes to the gender of the person they are arresting.



Value	Proportion	%	Count
female		25.73	131138
male		68.56	349446
NA		5.71	29097

Figure 13: Distribution of subject_sex

8.2. In all the models run, the confidence that a person stopped will be arrested was higher if the person was male, which indicates one might be at a higher likelihood of being arrested if one is a male driving through Rhode Island.

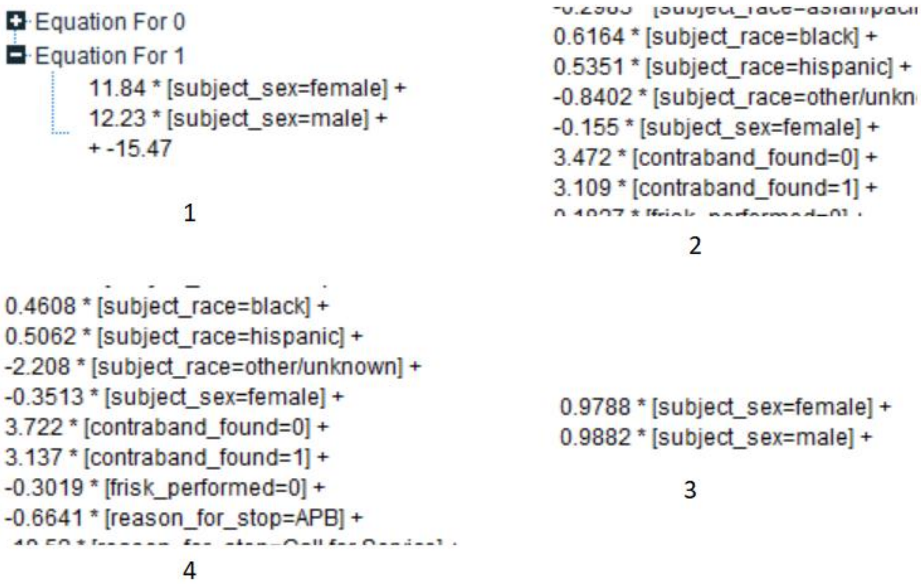


Figure 14: Confidence values in different models

8.3. The most interpretable results, according to me, were achieved when the dataset was cleaned of records, balanced for gender, and the input features to the model were the date, time, zone, sex, race, the reason for stopping, whether or not search was conducted and whether or not contraband was found. Intuitively speaking, logical distribution of the sigmoid values of confidence in the Logistic Regression was found, in which the Odds ratio of subject_sex = male was 4.45933655285 (calculated manually as $e^{1.495}$). A value greater than 1 of the odds ratio indicates positive relationship, which means that one is more likely to be arrested if one is male.

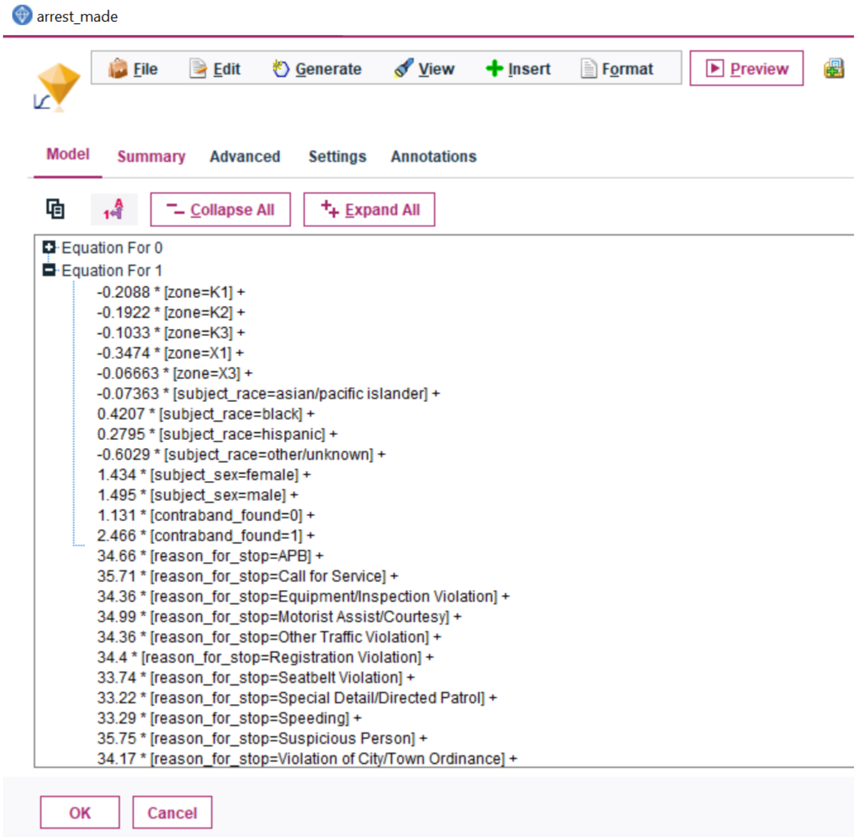
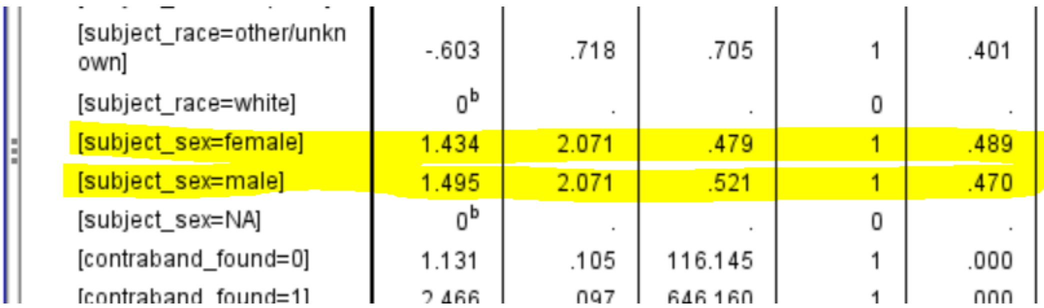


Figure 15: Most Interpretable model, with logical values of confidence for all features.

8.4. However, there are limitations to our analysis. One, there is no comparison with a base population of Rhode Island. Even when compared with a base population, finding a dataset that includes information of male vs female drivers in Rhode Island would be difficult. The patterns found in the analysis of gender bias in policing, but just like all other discriminatory analysis, there is a limitation to what can be concluded from statistics alone! Another big limitation of the interpretation is that it is necessary for the significance value of a predictor to be low in for us to safely make a prediction about odds ratio. The significance value in all the models built was high and therefore even though the odds ratio derivation still holds, it is not possible to say that with huge confidence that men are being discriminated against in arrests.



[subject_race=other/unknown]	-.603	.718	.705	1	.401
[subject_race=white]	0^b	.	.	0	.
[subject_sex=female]	1.434	2.071	.479	1	.489
[subject_sex=male]	1.495	2.071	.521	1	.470
[subject_sex=NA]	0^b	.	.	0	.
[contraband_found=0]	1.131	.105	116.145	1	.000
[contraband_found=1]	2.466	.097	646.160	1	.000

Figure 16: High value of significance!

8.5. Multiple iterations were done for the results, filtering out insignificant features, grouping vehicle_make amongst other steps.

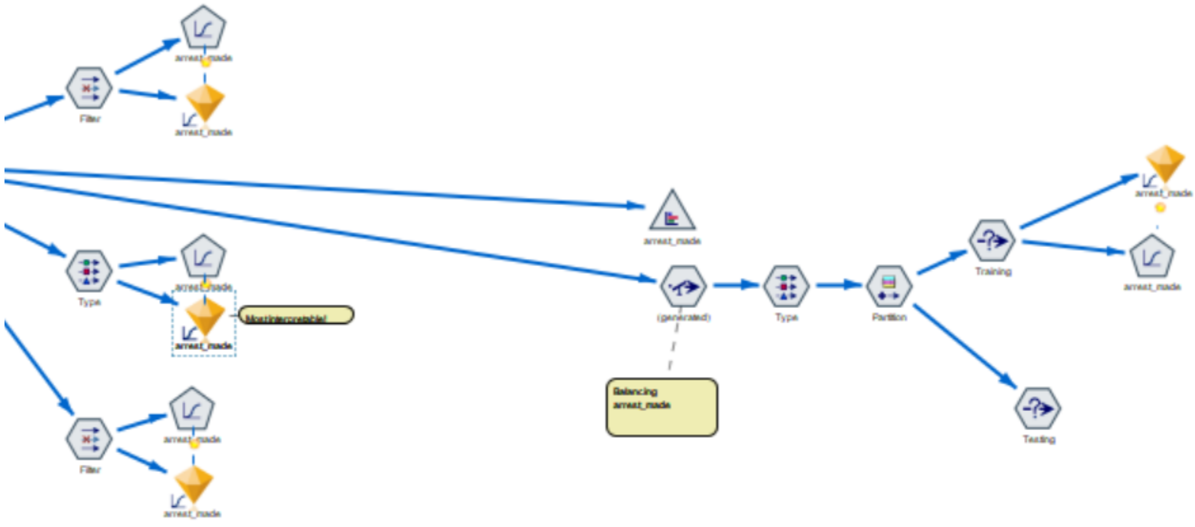


Figure 17: Multiple Iterations in SPSS

9. Future Work:

A better test for discrimination will be built and run using the open-source software for the OSAS. Better visualizations of the data will be possible in Python, by the support of its libraries. Features will be selected appropriately with appropriate targets to check if there is an effect on significance.

10. Acknowledgement

I acknowledge that the submitted work is my own original work in accordance with the University of Auckland guidelines and policies on academic integrity and copyright. (See: https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.html).

I also acknowledge that I have appropriate permission to use the data that I have utilised in this project. (For example, if the data belongs to an organisation and the data has not been published in the public domain then the data must be approved by the rights holder.) This includes permission to upload the data file to Canvas. The University of Auckland bears no responsibility for the student's misuse of data.

11. References

Ed. (n.d.). There's Nothing Odd about the Odds Ratio: Interpreting Binary Logistic Regression. Retrieved from https://www.statisticssolutions.com/theres-nothing-odd-about-the-odds-ratio-interpreting-binary-logistic-regression/

Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Ramachandran, V., Phillips, C., & Goel, S. (2017). *A large-scale analysis of racial disparities in police stops across the United States*. 1–10. Retrieved from http://arxiv.org/abs/1706.05678

Sharad, Goel, C., & Philips. (2017). Police Data Suggests Black and Hispanic Drivers Are Searched More Often Than Whites. Retrieved from https://slate.com/technology/2017/06/statistical-analysis-of-data-from-20-states-suggests-evidence-of-racially-biased-policing.html

Corbett-Davies, S., Pierson, E., And, A. F., & Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. Retrieved from https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-analgorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

Julia Angwin, Jeff Larson, S. M. and L. K. (2016). Machine Bias. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing