An Outcomes Based Analysis into Racial Bias for Traffic Stops in Nashville & Hartford

Alex, Conroy (08324841)

IFN704 Assessment 3, due 11:59pm Sunday 13 June 2021

Executive Summary

This projects aim is to produce an outcome bases analysis of police traffic stops within different cities in the United States of America (US). With focus on whether these outcomes are influenced by racial factors. While there have been several studies into racial bias within US policing, due to limiting factors within the data collection process, most do not take an outcomes-based approach. This study aims to take advantage of the Stanford Open Policing (SOP) dataset, along with Google API data, and US Census data to collate variables that can inform both geographic and demographic ethnic indicators. That will accompany the outcomes of each traffic stop. A multimodal logarithmic regression model was then built that determined it was possible to predict the outcome of the dataset with the same accuracy with and without the ethic-based indicators. These results seemed to indicate that once a traffic stop has been initiated, the race of the driver, or the demographic make-up of the area they were stopped, have little to no bearing on the outcome. However, there are several limiting factors to this result; as the model was based on only two cities, limiting generalisation, with indications that the model is not the best fit for the data, while almost all aspects of the data are self-reported.

Introduction

Policing is one of the many professions that has taken advantage of the fourth wave of automation and the subsequent higher levels of data analytics that this can facilitate. However, there is mounting evidence to suggest that rather than help drive out human based biases, such as those involving race. This combined with the often-complete lack of transparency around policing data in general, has left police-community relations at what appears to be all time lows.

While the ethically ambiguous but technically impressive models being used in China at present have been the focus of many news reports [1]. It is the United States of America that has grabbed headlines throughout the last year. The Black Lives Matter protests and riots mobilised a reported 8% of the total population of the US participating [2]. This has given rise to several independent attempts to collect, quantify, and analyse data in this area. Such as the Stanford Open Policing Project (SOP) [3] and Data for Black Lives [4]. These datasets were created with the hope of better understanding the outcomes and effects of the current state and realised impacts of changes to the current state of community interaction with police through a racial lens.

While studies have been conducted in this area [5], even using the SOP dataset [6], these studies often focus on the act of the stop or police interaction. It is of equal importance to understand what happens after the stop has been initiated [7]. This analysis is an important piece in the puzzle when it comes to understanding police interactions with minorities. However, few studies have aimed to assess this aspect of the issue. One such study into police shooting decisions, used a binomial logarithmic

regression do understand if there is any bias in an officer's decision to discharge their firearms [8]. This study looks to continue in the theme of this analysis but using an outcome indicator that has more than a binary option.

Literature Review

Colloquially negative perceptions of policing have been around since policing began. The US has even had a term coined after the act of police detaining a member of the public with little evidence, a 'Terry Stop' for over 50 years. Named after a supreme court case in the late 60s [9]. The public can interact with the police in many ways. One of the most frequent ways [10] these interactions do occur is through a traffic stop. As such traffic stops are considered one of the most used applications in the policing toolkit [6][8].

Interestingly research conducted by Boehme, Cann and Isom [11] suggests that the public's perception of police is more intrinsically tied to the stability and ethnical uniformity of the neighbourhood's makeup rather than actual police outcomes. Further, when considering minority communities, the negative perception of police is present and directed towards both under and over policing. Drawing conclusions that both have negative effects on these communities [12].

Research into any racial biasing in traffic stop data has generally shown that African Americans and Hispanics are both pulled over at higher rates than Caucasians [10]. This higher rate was found even after controlling for factors around neighbourhood benchmarking through the US census, or the argument that minorities commit more crime [6].

That is not to say there are not circumstances in which Caucasian driver stop rates are found to be more than other ethnic groups. A study from Cambridge University on drivers in Missouri found that instances of financially motivated or 'revenue raising' fines are more likely to be targeted towards White drivers. With the effects being found to be particularly strong in areas where over policing of Black drivers was already present. It is believing that this occurs as the Black drivers already have 'too many' fines to pay, and the White drivers are presumed to be wealthy enough to afford the cost [13].

A study by Simoiu, Corbett-Davies and Goel posit that some of these results could be attributed to the shortcomings on the benchmarking style used in determining bias. Instead an outcomes based model should be used to understand if a bias is occurring in multifaceted decision making such as that found in a traffic stop [7].

This scenario can be seen in the study into bias in police shooting incidents, which used an outcomes-based approach on an individualised dataset of active shooter situations in which an officer needed to make a shoot or no-shoot decision. The study used a binomial logarithmic regression analysis to understand what part the race of both the officer and perpetrator has on the outcome. The study went along way in addressing concerns raised in previous studies [8]. Unfortunately, due to the nature of collecting data in this area it did build the model using quite a small dataset.

The use of multinomial logit type models for variable types such as those of importance to this project, have been shown to be the most appropriate model to use across several university curriculum and studies [14]–[18]. With multiple sources describing in detail that an appropriate time to use a multimodal model is to analyse hypothesis that involve key variables with more than two outcomes.

As a result of the key gaps within the prior core studies discussed above, the project analysis will look to investigate all covariates within the large SOP dataset together to determine which are the most important to the outcome variable using a multimodal logarithmic model. This will be done to supplement the studies that look to understand if there is racial bias in the number of stops conducted by police. While seeking to improve on the shortcomings of the Worrall et al [8] study into Police Shooting Decisions, with a larger dataset and more complex outcome variable. This will hopefully add to the studies in this area ahead pf the larger decisions that face the US over the coming years in the wake of the widespread unrest over police conduct with minorities.

Approach

The project undertaken had four major deliverables. These were; the collation of a dataset from multiple sources to inform the creation of a logarithmic regression model relating all variables to the traffic stop outcome, the construction of multiple models using different levels of the data using the full set of variables, the analysis of the importance of these variables to the model to construct a 'lean' model that is capable of competing with the full model for accuracy, the comparison the two models to understand if there is evidence to suggest any race related variables are necessary to maintain or improve the full variable dataset accuracy.

Data Collection

The traffic stop data was collected from the SOP dataset. The SOP initiative has collected the data for many cities in the US. The SOP list was investigated with the initial idea of find a subset that would consist of both large and small cities, with at least one city having a localised population of Hispanic, African American, or Caucasian residents above that of the national average, along with outcome, location, and general stop related data. General stop related data in this instance meaning the reason for the stop, if a search conducted, etc. Unfortunately, due to time constraints only two cities were eventually used for the project's dataset: Hartford and Nashville. With Nashville having a large population of Caucasians and is a major southern city. While Hartford has a larger than average number of African Americans and is a small north-eastern city.

ZIP code data was collected using the Google Maps API reverse geocoding features, via the automatically collected latitude and longitude for each stop. As it was determined that the user entered address information were not a reliable source.

The DP05: Demographic and Housing Estimates table of the 2014 US Census [19] was collected and rearranged by ZIP code tabulated areas (ZCTAs) to produce an overall demographic profile for the code. The ZCTA was created by the census bureau to overcome the shortfalls of the ZIP code system for representing the population. While there is not a perfect match between ZCTA to ZIP code, studies have determined that the inaccuracies in matching them are contained to fringe cases [20]. For this project, it was determined that this level of inaccuracy was acceptable. This was then combined with the ZIP code produced by the reverse geocoding, to create a new racial variable. The purpose of this variable is to determine if the racial profile of the neighbourhood the stop occurred in has any determining power over the stop outcome.

The datasets were then combined and filtered down to the 2014 date range, this was done to maintain the accuracy of the demographic data found in the census. While accounting for any time-based variables that might influence the data, i.e. weather, school holiday etc.

Model Construction

The model construction was developed using both research into current practices and data exploration. Once the SOP data was collated, a short analysis showed that several variables were used as raw inputs into the meaningful variables within the data. As this would result in a double up of information to the model without any additional information gained, these were removed.

It was decided to remove any variable that was not present across all individual city dataset. This was done to allow for direct comparison and later collation of any larger models created. The de-identified officer identification variables were removed, this was done to prevent any singular officer bias or other errant behaviour skewing the results, as individual officer behaviour was not in the scope of this analysis. After this the date and time data were factorised into categorical data, to bring them in line with the handling of the major variables of interest that were already in this format. Finally, all rows that contained nan values for location, demographic, or outcome variables were removed, as these were determined to be critical to the model.

Nan values in both the searches-based variables and the outcome variable were coded to be false indicators, as the lack of response in this case was significant. Once this process was complete a nan value report was produced that showed only a handful of rows still contained nans in each dataset. These were just removed as they represented such a small subset of the overall data.

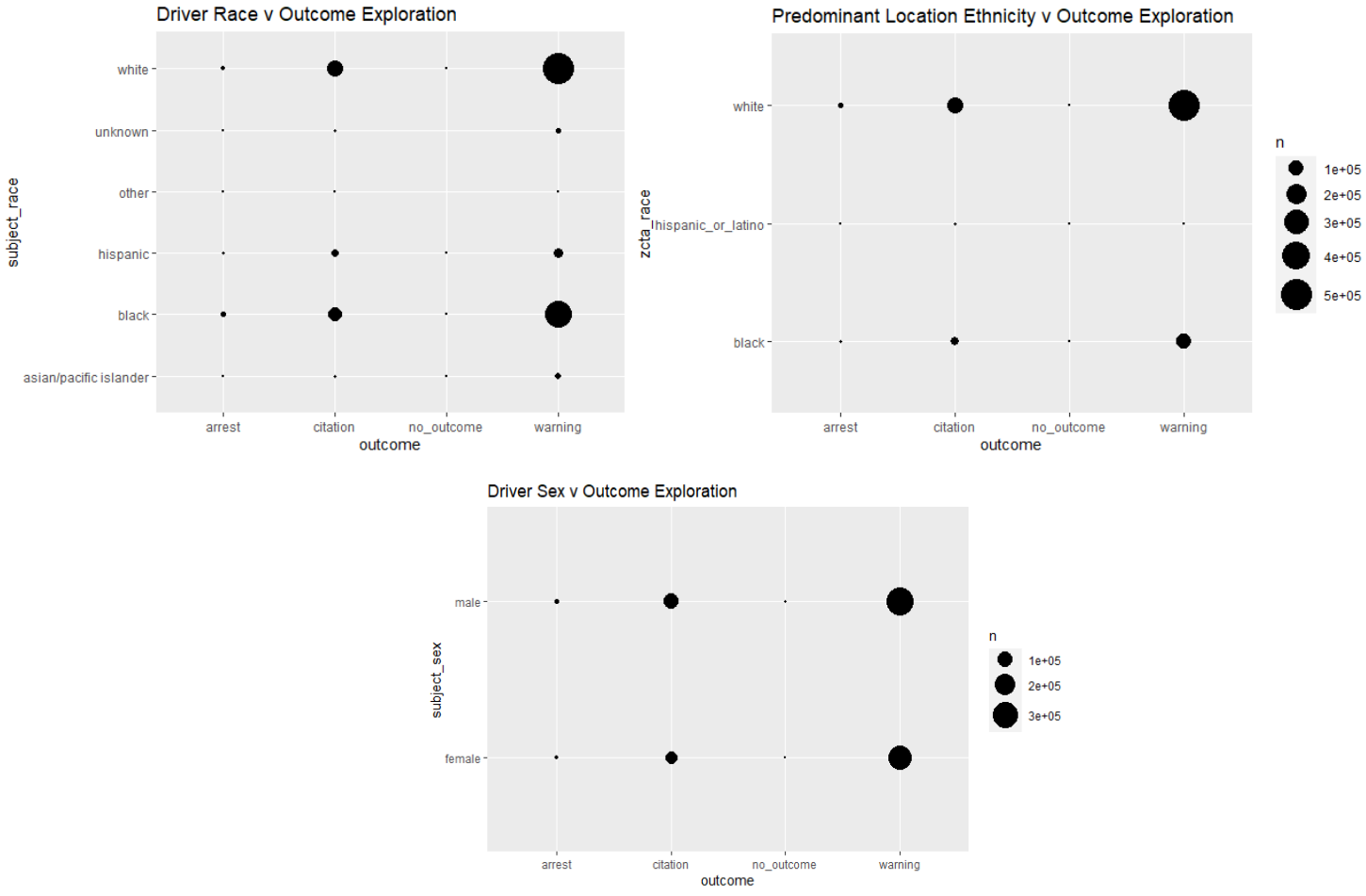


Figure 1 Data Exploration of x-variables vs Traffic Stop Outcome

Outcome was plotted against the demographic/geographic variables. This highlighted slight imbalances between the outcome and demographic/geographic variables. However, these may be accountable through the respective population percentage between the different ethnicities, and other factors. Like males being more likely to be the driver [21]. These can be seen in figure 1. Throughout the exploration it was also observed that there appeared to be additional potential imbalances within the dataset. Particularly with the outcomes observed. There were more instances of the less sever outcomes compared to the more severe, I.e., warning to arrest.

A Secondary exploration was conducted into the collinearity of the predictor variables, shown in appendix 1 due to size. For the most part it showed that there was very little correlation between all the variables. Except for the search_conducted and its related variables, i.e., search_basis and contraband_found. However, it was decided to keep these in the initial model as they represented different area of this sub-interaction of the traffic stop.

Something of note was that most predictor variables outside of those mentioned maintained a 'background level' of correlation of around a 0.01 – 0.04 correlation coefficients. However, the variable related to the ethnical makeup of the stop location ZCTA_race, driver ethnicity variable subject_race, and decision to stop the driver reason_for_stop all showed slightly higher coefficient scores. But none high enough to be considered significant, with all coefficients in the late teens. It is only of note as these variables include two critical predictor covariates to the project.

A check on the created ZCTA_race variable was conducted to confirm that the census data was not recording high levels of multiple ethnicities within one area code. This can be seen in figure 2 left, with figure 2 right showing the same analysis but only for those ZCTAs present in the SOP analysis dataset.

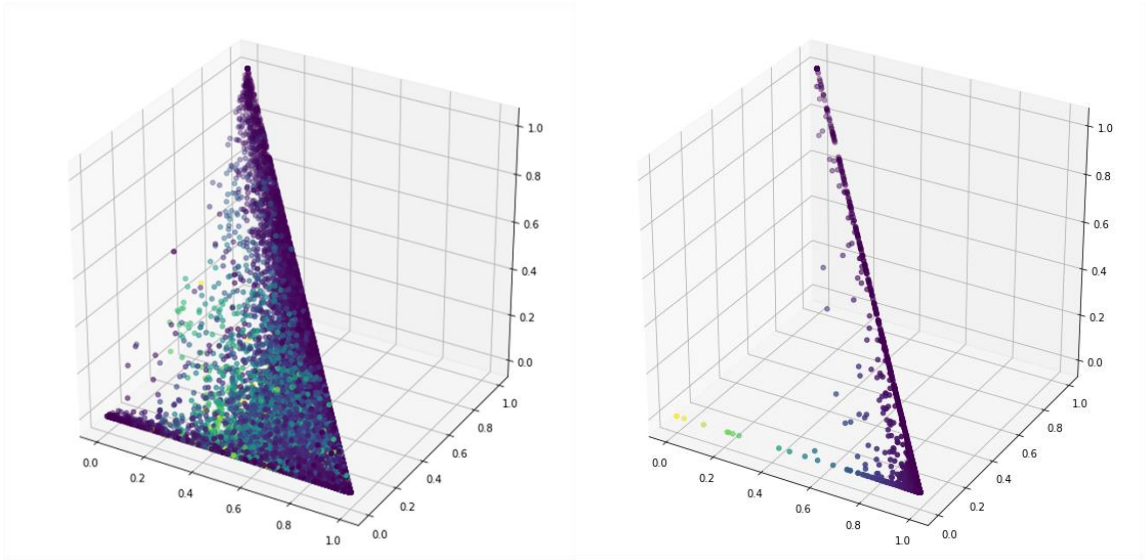


Figure 2 ZCTA plot by reported ethnical demographic for the United States.

The plot shows that this method is not particularly generalisable for the whole of the US, with significant numbers of ZCTAs showing more than 1 'dominant' ethnicity within the area. Representing more blended suburbs. However, for the purposes of this project the selected cities show a more segregated population and will be used, noting the ZCTA to zip code pairing as a potential area of inaccuracy.

The multinom model in the nnet [22] library of R was used to build all models in the analysis. This specific log-linear model was chosen after investigations into using the more traditional libraries proved to have inadequate family types to perform the required analysis. The difference between the nnet library and others, is that the nnet models are fit via neural networks that operates without the need to reshape the data beforehand. Unfortunately, this did inhibit the ability to use goodness-of-fit tests normally seen with this general model family. This appeared to be due to incompatibility between R functions. The alternatives used are discussed in the findings.

Multiple models were then produced that ran an 80/20 train and test of each individual city as well as a combined dataset for all predictor variables. These model variables were then analysed to extract the key variables, using coefficient magnitude, p-value, and accuracy to create a 'lean' model.

Findings

The final model produced an output with the summarised predictor variables shown in table 1, for the full set see appendix 2. The variables were selected based on a mixture of goodness-of-fit tests and comparison of test accuracies. Both as an absolute value and when compared to the 'full' model accuracy.

Variable	Coefficients	STD. Error	Z-stat	p-value
reason_for_stop	-2.03	0.91	-9.5x10^12	0.02
contraband_found	5.08	0.39	19.85	0.00
search_basis	-12.58	0.34	-4.6x10^12	0.00

Table 1 Averaged Final Predictor Variable Levels

The variables in the model account for ~80% of the variance in traffic stop outcome. The results are interesting, but not generally surprising. As they represent what would be considered natural steps within a traffic stop that would result in different outcomes. With the inclusion of the contraband_found variable a potential indicator of the unique 'war' on drugs laws found in the US [23]. The most surprising aspects of this variable set can mostly be found in the predictor covariates that are excluded. Specifically, those of original interest to the project, involving either the race of the driver or the geographic dominant ethnicity.

However, it should be noted that driver demographic indicators were present in the individual city 'lean' models that did achieve close to the same levels of accuracy. But these models included almost identical variable sets to the final model, just with the added inclusion of driver race and/or driver age. This seems to indicate that the more generalised the model becomes, the less significant the ethnicity-based covariates become.

As discussed, the implementation of the multimodal model proved troublesome when it came to running the more traditional visual checks such as QQ plots and residuals. In lieu of this the McFadden pseudo R^2, general R^2, deviance, AIC, McNemar's Test and Predicted Confusion Matrix were used. Shown below in table 2 and 3.

Test	Score
McFadden	0.15
R2ML	0.17
R2CU	0.24
Deviance	621603.4
AIC	621933.4
McNemar's	< 2.2x10^-16

Table 2 Model Goodness of fit Test Scores

McFadden's pseudo-R squared value is calculated using the maximum likelihood of the current fitted model and the corresponding value of the null model. The interpretation of the McFadden pseudo-R squared value indicate that an excellent fit should be within the range of 0.2 – 0.4 [24]. While those observed in the project's models, are again reasonably close, still indicated that there are improvements that could be made. This is further asserted with the other two R squared values indicate that the model fit is fair/medium but by means a high level. Looking at the comparison of the deviance with the AIC, the AIC is slightly higher which indicates a positive outcome. However, it is only slightly which continues the assertion that there is room for improvement for the model. Lastly the McNemar's test looks to compare the 'lean' and 'full' models, to accept or reject the hypothesis that both models operate equally. In this instance the score indicated that the 'lean' model does operate unequally to that of the 'full' model.

Looking to the confusion matrix outlined in table 2. It appears that the model does a reasonably good job of interpreting the different possible outcomes. Particularly with the warning outcome. This matrix does however seem to reinforce that the model is suffering from the imbalance of the different outcomes. As discussed in the exploration, most prevalent of all outcomes was a warning. Ways to potentially improve the model would be to investigate weighting, augmenting, or focusing on collecting additional datapoints for the other outcome types. To try and bring more balance to the dataset.

		Reference			
Prediction		arrest	citation	No outcome	warning
	arrest	922	147	2	22
	citation	622	2401	175	836
	No outcome	11	8	107	15
	warning	902	29303	45	121481

Table 3 Final Model Confusion Matrix

There are a few limitations to this study. The study did not properly investigate the use of alternative models for this dataset, instead rely on the literature that indicated that it was the appropriate choice. This limitation was predominately driven by the specific choice in R library used to complete the analysis. Another limitation is in the small city type variance used, as there are 3 main ethnicities in the US, as well as variability in the size of the populations. The project is not able to accurately generalise for the full US population, that would be possible with the inclusion of more city types. Lasty, the project was limited in the same way other studies of this subject matter are [8]. In that the data is hard to collect and verify. While the SOP project has done a lot to provide a large volume of data, the dataset almost entirely consists of self-reported data. Making the verification of data accuracy effectively impossible.

Reflection

Future work into this area would be able to improve upon this study in several ways. A more verifiable dataset could be one way to improve accuracy of findings. While on the opposite end of the size spectrum, the inclusion of more data across various demographics would allow for more in-depth modelling and comparisons between the different sub-sections within the US. Furthermore, additional analysis into the interactions of driver race and location ethnicity should be explored in future studies.

Putting the above discussion about future work aside, this project has allowed me to understand the skill required and difficulty in undertaking a real-world modelling problem. Particularly when the dataset has not been provided in a complete way for analysis to begin at the outset of the undertaking. The creations of data-pipelining have allowed me to grow as an analyst. Another interesting learning point for me was having to juggle a full-time professional job alongside advanced project/research work. Given that one has quite hard and immediate deadlines and deliverables, while the other was over a much longer timeframe. This allowed me to further develop my time management and scheduling skills to cope with the rigours of a professional career.

References

[1] 'China: Police "Big Data" Systems Violate Privacy, Target Dissent', *Human Rights Watch*, Nov. 19, 2017. https://www.hrw.org/news/2017/11/19/china-police-big-data-systems-violate-privacy-target-dissent (accessed Jun. 11, 2021).

[2] A. G. Ferguson, *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*. NYU Press, 2019.

[3] 'The Stanford Open Policing Project', *openpolicing.stanford.edu*. https://openpolicing.stanford.edu/ (accessed Mar. 28, 2021).

[4] 'Data 4 Black Lives'. https://d4bl.org/ (accessed Mar. 28, 2021).

[5] S. Goel, M. Perelman, R. Shroff, and D. A. Sklansky, 'Combatting Police Discrimination in the Age of Big Data', *New Crim. Law Rev.*, vol. 20, no. 2, pp. 181–232, May 2017, doi: 10.1525/nclr.2017.20.2.181.

[6] A. Chohlas-Wood, S. Goel, A. Shoemaker, and R. Shroff, 'An Analysis of the Metropolitan Nashville Police Department's Traffic Stop Practices', p. 10.

[7] C. Simoiu, S. Corbett-Davies, and S. Goel, 'The problem of infra-marginality in outcome tests for discrimination', *Ann. Appl. Stat.*, vol. 11, no. 3, Sep. 2017, doi: 10.1214/17-AOAS1058.

[8] J. L. Worrall, S. A. Bishopp, S. C. Zinser, A. P. Wheeler, and S. W. Phillips, 'Exploring Bias in Police Shooting Decisions With Real Shoot/Don't Shoot Cases', *Crime Delinquency*, vol. 64, no. 9, pp. 1171–1192, Aug. 2018, doi: 10.1177/0011128718756038.

[9] A. Gelman, J. Fagan, and A. Kiss, 'An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias', *J. Am. Stat. Assoc.*, vol. 102, no. 479, pp. 813–823, Sep. 2007, doi: 10.1198/016214506000001040.

[10] E. Pierson *et al.*, 'A large-scale analysis of racial disparities in police stops across the United States', *Nat. Hum. Behav.*, vol. 4, no. 7, pp. 736–745, Jul. 2020, doi: 10.1038/s41562-020-0858-1.

[11] H. M. Boehme, D. Cann, and D. A. Isom, 'Citizens' Perceptions of Over- and Under-Policing: A Look at Race, Ethnicity, and Community Characteristics', *Crime Delinquency*, p. 0011128720974309, Dec. 2020, doi: 10.1177/0011128720974309.

[12] G. Ben-Porat, 'Policing multicultural states: lessons from the Canadian model', *Polic. Soc.*, vol. 18, no. 4, pp. 411–425, Dec. 2008, doi: 10.1080/10439460802094686.

[13] A. P. Harris, E. Ash, and J. Fagan, 'Fiscal Pressures and Discriminatory Policing: Evidence from Traffic Stops in Missouri', *J. Race Ethn. Polit.*, vol. 5, no. 3, pp. 450–480, Nov. 2020, doi: 10.1017/rep.2020.10.

[14] 'Multinomial Logistic Regression | R Data Analysis Examples'. https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/ (accessed Jun. 11, 2021).

[15] H. A. Hamid, Y. B. Wah, X.-J. Xie, and O. S. Huat, 'Investigating the power of goodness-of-fit tests for multinomial logistic regression', *Commun. Stat. - Simul. Comput.*, vol. 47, no. 4, pp. 1039–1055, Apr. 2018, doi: 10.1080/03610918.2017.1303727.

[16] J. J. Faraway, *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, Second Edition*. CRC Press, 2016.

[17] J. J. Goeman and S. le Cessie, 'A Goodness-of-Fit Test for Multinomial Logistic Regression', *Biometrics*, vol. 62, no. 4, pp. 980–985, 2006.

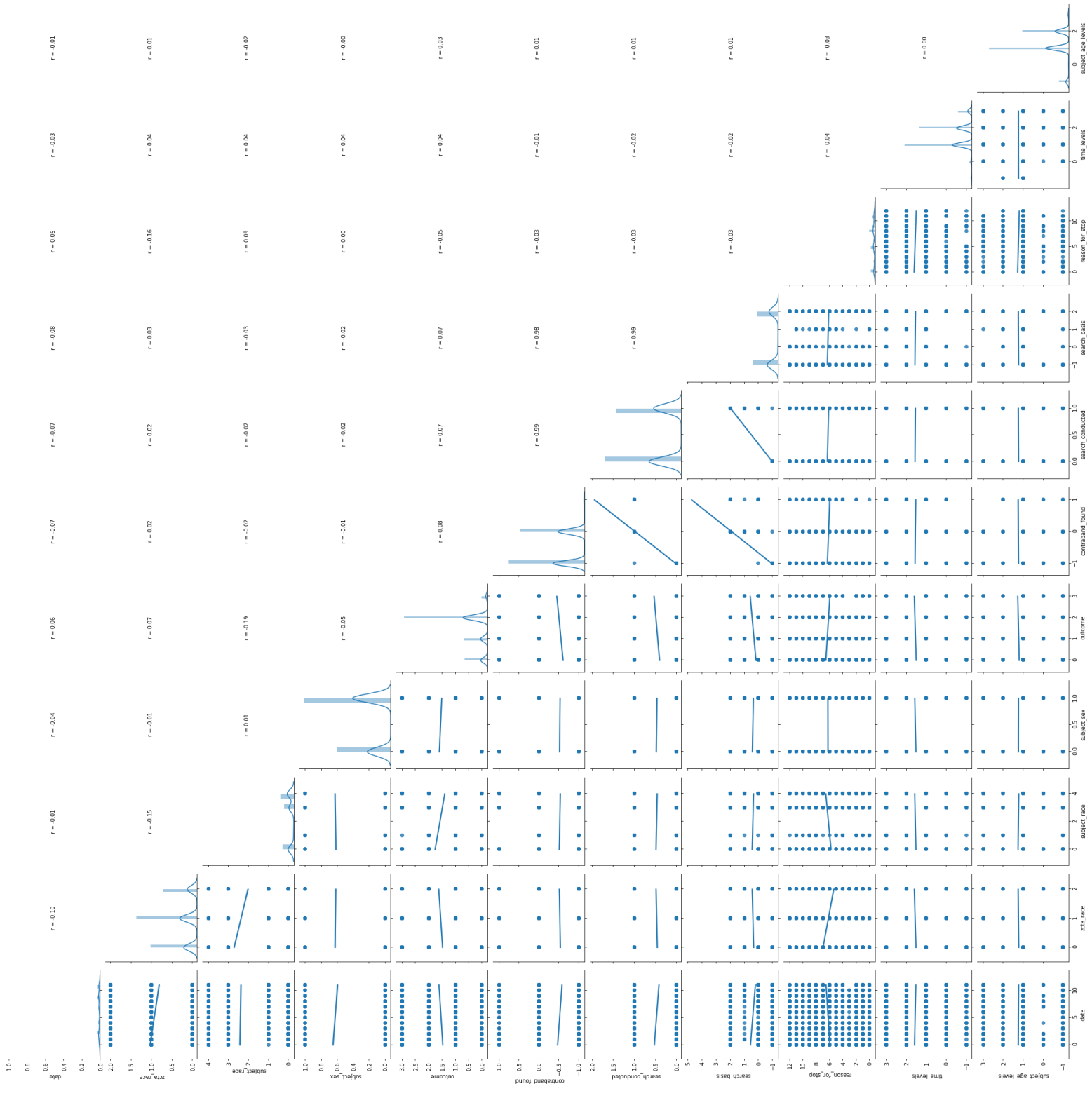
[18] W. W. Stroup, *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC Press, 2012.

[19] 'Census - Table Results'. https://data.census.gov/cedsci/table?tid=ACSDP1Y2018.DP05&hidePreview=true (accessed Jun. 11, 2021).

[20] T. H. Grubesic and T. C. Matisziw, 'On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data', *Int. J. Health Geogr.*, vol. 5, no. 1, p. 58, Dec. 2006, doi: 10.1186/1476-072X-5-58.

[21] F. Valham, M. Eriksson, B. Stegmayr, and K. A. Franklin, 'Snoring men with daytime sleepiness drive more than others: A population-based study', *Sleep Med.*, vol. 10, no. 9, pp. 1012–1015, Oct. 2009, doi: 10.1016/j.sleep.2008.09.020.
[22] 'multinom function - RDocumentation'. https://www.rdocumentation.org/packages/nnet/versions/7.3-16/topics/multinom (accessed Jun. 11, 2021).
[23] D. Wodak, 'Mandatory Minimums and the War on Drugs', in *The Palgrave Handbook of Philosophy and Public Policy*, D. Boonin, Ed. Cham: Springer International Publishing, 2018, pp. 51–62. doi: 10.1007/978-3-319-93907-0_5.
[24] 'Urban Travel Demand: A Behavioral Analysis, by Tom Domencich and Daniel McFadden, 1975, North-Holland'. https://eml.berkeley.edu/~mcfadden/travel.html (accessed Jun. 11, 2021).

Appendix 1: Large Correlation Plot



Appendix 2 Full Final Predictor Variables

(Intercept)	Coefficient	Std. Errors	z stat	p value
reason_for_stopCell Phone	-0.6047985	0.4861075	-1.2441662	0.2134384
reason_for_stopchild restraint	2.4016514	0.5519911	4.3508881	0.0000136
reason_for_stopDefective Lights	-2.26E+01	0.00E+00	-2.39E+10	0.00E+00
reason_for_stopDisplay of Plates	3.3548572	0.6797487	4.9354372	0.0000008
reason_for_stopEquipment Violation	5.0932387	0.5776997	8.8164124	0
reason_for_stopinvestigative stop	13.4274466	0.4638356	28.9487173	0
reason_for_stopmoving traffic violation	-8.172414	7.2234	-1.13138	0.257895
reason_for_stopMoving Violation	-2.4549462	0.5005139	-4.9048511	0.0000009
reason_for_stopOther	3.342514	0.574375	5.819394	0
reason_for_stopparking violation	3.6053443	0.5037507	7.1570012	0
reason_for_stopregistration	-4.50E+01	0.00E+00	-1.99E+14	0.00E+00
reason_for_stopRegistration	-4.8971748	1.2806609	-3.8239433	0.0001313
reason_for_stopsafety violation	2.6264767	0.518706	5.0635176	0.0000004
reason_for_stopSeatbelt	-3.5678696	0.820762	-4.3470209	0.0000138
reason_for_stopseatbelt violation	3.3599821	0.7636373	4.3999713	0.0000108
reason_for_stopSpeed Related	-3.3018772	0.8548487	-3.862528	0.0001122
reason_for_stopStop Sign	4.2671057	0.6945827	6.1434092	0
reason_for_stopSuspended License	3.591742	0.571645	6.283168	0
reason_for_stopTraffic Control Signal	4.5727537	0.5216349	8.7661959	0

reason_for_stopvehicle equipment violation	3.4644813	0.5650013	6.1318108	0
reason_for_stopWindow Tint	-5.0197086	0.8600731	-5.8363746	0
contraband_foundFalse	3.1362678	0.5642497	5.5582975	0
contraband_foundTrue	7.0305674	0.2059485	34.1375081	0
search_basisconsent	5.2858425	0.4645268	11.3789829	0
search_basisother	-9.4498588	0.2720461	-34.7362357	0
search_basisplain view	-12.1725443	0.6148845	-19.7964743	0
search_basisprobable cause	-3.40E+01	0.00E+00	-1.83E+13	0

Table 4 Final Predictor Variables

Appendix 3 Various Analysis Excerpts