Math 748 Term Project Report

Modeling Traffic Stop Demographics with Population Estimates

Chad Kite

December 14, 2023

Contents

I. Introduction

According to the Stanford Open Policing Project (SOPP), there are over 50,000 traffic stops conducted in the United States on an average day, yet there is no prescribed means to record and report those stops. The worst outcomes, involving the death of someone involved, are often parsed in depth and at times draw national interest, headlining news coverage and sparking calls for reform, protests, and broader movements to address underlying and related problems. However, contextualizing those individual events within the broader scope of traffic stops in the United States is difficult at best.

The absence of a uniform reporting system across jurisdictions makes it challenging to compile accurate and comparable statistics on the frequency and nature of these interactions. Furthermore, disparities in data collection methods, reporting criteria, and the availability of information hinder efforts to draw meaningful conclusions about the patterns and trends associated with traffic stops. As a result, the true extent of disparities, potential biases, and the overall impact of traffic stops on various communities remains obscured, perpetuating a significant obstacle to informed policymaking and public discourse on this critical issue

Without even the most rudimentary statistics for the population of traffic stops, there is no way to develop a shared understanding of where such outcomes fit within the system as a whole. Of the many missing pieces of information which could inform such an understanding, perhaps the most important is exactly who, demographically speaking, is the subject of these traffic stops. To help bridge this gap, this project will utilize the traffic stop data accumulated by the SOPP and detailed annual demographic data for the United States to model the proportions of traffic stop subjects by age, race, and sex at the county level.

II. Data

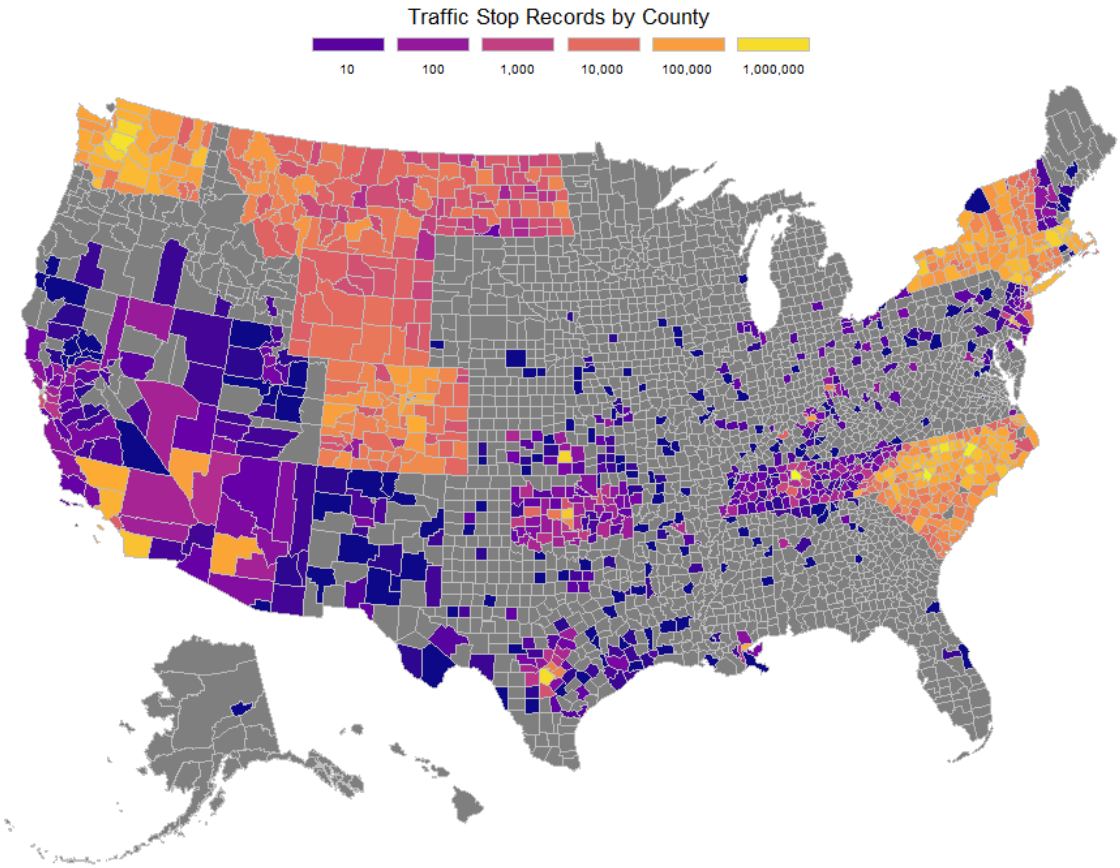
A. Traffic Stops

The SOPP has collected over 200 million traffic stop records from police departments in the United States. Of those, nearly 100 million records, spanning the period from 1999 to 2020 and including data from 21 state patrol agencies and 29 municipal police departments, were sufficiently detailed for at least some portion of their research. The processed and standardized versions of these files are available on the SOPP's webpage[1] in multiple formats. However, almost half of these records are unusable for modeling the demographic proportions of traffic stop subjects for various reasons.

Critically for this project, many lack complete demographic information on the motorist involved in the traffic stop. There are also numerous records that do not have a useable location[2], with many lacking any location information at all and others including an incomplete street address or a location identifier with unclear meaning. Excluding these records, as well as those without a date, leaves 52,644,187 traffic stop records from 11 state patrol agencies and 19 municipal police departments.

[1] https://openpolicing.stanford.edu/data/

[2] Only records with location of the stop in latitude and longitude, those that included the county in which the stop occurred and those for which the county could be reasonably inferred were used in this project. For data that was provided by cities that are entirely within one county, that county was used.



The traffic stops in these records are generally well spread throughout the country, but are clearly not randomly drawn. States for which state-level data is available are clearly visible, as are cities with a robust traffic stop data collection system that made their data available to the SOPP. These records cannot be assumed to be a representative sample of the population of traffic stops because there is potential bias introduced by the voluntary nature of the submissions.

B. Census Estimates

Demographic data is from population estimates provided by the U.S. Census Bureau[3]. They are available at the county level for every year from 1990 to 2020. The estimates for each county are by sex, race, origin and age, where age is grouped in 5 year increments. Race consists of 4 classes: Alaskan or American Indian, Asian or Pacific Islander, Black, and White, while origin has 2 classes: Hispanic and Non-Hispanic.

C. Pre-processing

In order to match the races included in the traffic stop data, origin and race were combined according to the census estimate technical descriptions into a single category with the same classes as the traffic stop data: Asian / Pacific Islander, Black, Hispanic, Other, and White. We then min-max and z-score normalized the census data so that we can compare models created

[3] https://seer.cancer.gov/popdata/download.html

using the normalized and unnormalized data as predictors. As would be expected with population demographics, the demographic categories are highly correlated, which will make model interpretation extremely difficult. However, since the intent of this project is not to identify the exact drivers of the differences between population and traffic stop demographics but to accurately model the latter given the former, correlation is only a problem insofar as it hampers the construction of an accurate model.

Processing the traffic stop records began with extracting location, date and demographic data from each of the 30 traffic stop record files that contained records that meet the requirements of this project. This process required the removal of NAs from date, location (latitude and longitude, or county), subject age, subject race and subject sex. Then, latitude and longitude values were rounded to 4 decimal digits and the date was simplified to the year of the stop to allow for matching with census estimates. Next, location information was used to identify the associated Federal Information Processing Standards (FIPS) County Code for each observation. Along with the year, the FIPS County Code allowed us to aggregate the demographics of the traffic stop data by year and county code, then match each county/year combination with its related census records.

Counties with less than 30 traffic stop records were dropped to increase the likelihood that included records constituted a representative sample of traffic stop demographics in the area they were collected. Then, a random sample of 70% of the remaining counties was drawn and used for training while the other 30% was set aside for testing. Of note, in order to accurately determine the efficacy of the chosen model, every year available for counties that were selected for testing was removed from the training set. In testing, the selected model was used to predict the demographic breakdown of traffic stops for the selected counties by year, providing a window into how the model will perform in a setting that more closely approximates an actual use case.

III. Model Selection

A. Approach and Criteria

Creating a category with classes for each combination of binned age, race, and sex would result in 30 or more classes, depending on the number of bins used for age, many with very few occurrences in the data. In testing the relationships between the three demographic categories being considered, however, we discovered that age is much more closely related to population size than race or sex. In fact, a χ^2 test showed that binned subject ages do not differ by subject race or subject sex, whereas subject sex does differ by subject race. This allowed us to separate subject age and model it individually, which reduced the maximum number of classes that needed to be considered to 10, allowed us to bin age in 8 relatively evenly distributed groups, and dramatically increased the minimum number of occurrences for any class in the data.

Since the objective of this project is to model the proportion of a set of stops that will fall into specific demographic categories, the classification of individual records is not relevant. A better measure is how closely the predicted proportions match the actual reported traffic stops at the county level. To obtain this measurement, the predicted proportions were multiplied by the known number of stops for each county and the number of correct predictions was divided by the total number of stops to produce a county-level accuracy metric. An overall accuracy was also

calculated by summing the total predictions for each demographic category and dividing by the total number of stops.

B. Subject Age

1. Multinomial Logistic Regression

We began modeling the relationship between subject age and population demographics by building a multinomial logistic regression model using the proportions of binned subject ages by county as the response variable with the year and binned population ages as predictors. We built three models initially using different methods of normalization to determine if normalization improved the predictive performance of the model and, if so, which version of normalization provided the most improvement. The initial set of results showed little advantage over the null model, so we experimented with changing the minimum number of traffic stops required for a county-year to be included in the data used for model building. The best results were found by retaining only county-years with a total greater than 200 traffic stops, reducing the number of county-years from 3,893 to 3,444. The results were as follows:

Normalization	AIC	χ^2 vs null model	p-value
Unnormalized	12274.03	12.85 (63 df)	1.0
Min-max	12237.26	49.61 (63 df)	0.8905
Z-score	12237.53	49.34 (63 df)	0.8957

None of the models proved significantly better than the null model. However, the unnormalized data fared much worse than either min-max or z-score normalized data, so unnormalized data was not used in further multinomial logistic regression model building.

Next, we conducted stepwise selection in both directions using AIC as the criteria. The selected models for the min-max and z-score normalized data were very similar, each including the year and one population age group. The model based on the min-max normalized data used the population between the ages of 51-60 while the model based on the z-score normalized data used the population between the ages of 61-70. Their performance follows.

Normalization	AIC	χ^2 vs null model	p-value
Min-max	12148.84	40.04 (14 df)	0.00025
Z-score	12144.75	44.13 (14 df)	0.00005

A check for interactions found no interaction terms of significance, which is unsurprising given the high correlation between age groups. The model built using z-score normalized data slightly outperformed the model built using min-max normalized data, so it was selected.

Response	(Intercept)	year	pop_61-70
stops_per_21-30	-52.1367	0.0264	0.0175
stops_per_31-40	-65.9730	0.0331	0.0074
stops_per_41-50	-93.5460	0.0467	-0.0080
stops_per_51-60	-157.3465	0.0783	-0.0337
stops_per_61-70	-184.4481	0.0914	-0.0722
stops_per_71-80	-187.3894	0.0923	-0.0994
stops_per_81-Inf	-62.3742	0.0293	-0.0308

County-level accuracy of 93.90%

2. Linear Discrimination Analysis

To use the linear discriminant analysis function in R, the aggregated traffic stop records had to be split into individual observations. The demographics for every record in a particular county and year are the same, so the data consist of the same demographic information repeated for each time a given response, labeled name in the table below, appeared in the traffic stop records for that county and year. The first 3 rows of the disaggregated data are below.

year	pop_0-20	pop_21-30	pop_31-40	pop_41-50	pop_51-60	pop_61-70	pop_71-80	pop_81-Inf	name
2000	0.0120	0.0114	0.0126	0.0132	0.0114	0.0100	0.0142	0.0124	0-20
2000	0.0120	0.0114	0.0126	0.0132	0.0114	0.0100	0.0142	0.0124	0-20
2000	0.0120	0.0114	0.0126	0.0132	0.0114	0.0100	0.0142	0.0124	0-20

We began by testing a model that included the year and all binned age groups in the demographic data across three methods of normalization to determine if normalization improved the predictive performance of the model and, if so, which version of normalization provided the most improvement. The results were as follows:

Normalization	County-level accuracy
Unnormalized	94.14%
Min-max	95.09%
Z-score	95.09%

Normalization improved the model's accuracy, but the type of normalization had no impact, so we arbitrarily chose the min-max normalized data to proceed.

Predictor	LD1	LD2	LD3	LD4	LD5	LD6	LD7
year	-0.2691	0.0470	0.0045	-0.0052	0.0471	0.0366	0.0168
`pop_0-20`	71.0608	469.9542	-163.8615	-239.0605	-294.3471	175.3673	317.8429
`pop_21-30`	9.5304	314.9625	-195.6983	-53.7806	-33.9351	111.6845	289.4849
`pop_31-40`	0.2971	-121.8927	93.4058	-127.0440	-147.8592	-31.1660	12.5604
`pop_41-50`	-17.9366	137.9225	9.6168	81.4872	316.1810	81.8148	-40.8422
`pop_51-60`	-20.2419	-75.5018	16.2295	-0.1043	-380.6436	-42.1984	172.5459
`pop_61-70`	24.8779	30.8674	3.0206	27.3932	227.6167	-52.7127	-213.2455
`pop_71-80`	-4.2586	63.4026	-38.3057	-50.0963	-107.8974	128.1945	139.5816
`pop_81-Inf`	-53.3224	-789.2829	263.4058	336.5066	406.9096	-339.5263	-664.0746

Proportion of trace:

LD1	LD2	LD3	LD4	LD5	LD6	LD7
0.7077	0.2020	0.0515	0.0349	0.0027	0.0010	0.0002

County-level accuracy of 95.10%

3. Extreme Gradient Boosting

The xgboost package in r requires disaggregated data however, we were unable to run the extreme gradient boosted model on the entire dataset. Unfortunately, downsampling proved insufficient for producing a viable model. So, we returned to the full set of disaggregated records and randomly sampled 20%. Models built using this sample were accurate enough for comparison but small enough to conduct hyperparameter tuning. Initial tests showed no change in the model's performance based on the normalization method used, so we continued to use the min-max normalized data that was used for the LDA model.

The parameters selected for tuning were maximum depth, learning rate (eta), and number of rounds. The following values were tested in all combinations: max.depth: 4, 6, 8; eta: .2, .3, .4; and nrounds: 5, 8, 10. The models produced using each set of hyperparameters were compared on county level accuracy. The top 5 results are listed in the table below.

max.depth	eta	nrounds	County-level accuracy
4	.4	10	92.66%
6	.4	10	92.17%
4	.4	8	91.93%
8	.4	10	91.91%
4	.3	10	91.66%

The model with the highest accuracy used a max depth of 4, a learning rate of .4 and 10 rounds. The top 5 models were separated by only 1% in terms of county level accuracy, but

multiple models found in the tuning process had county-level accuracies of less than 81%, so tuning was especially helpful for the boosted model.

County-level accuracy: 92.66%

4. Cross-Validation

All three models were evaluated using 10-fold cross-validation by comparing the mean of their county-level accuracies. During cross-validation, the extreme gradient boosting models were built using a random sample of 40% of the observations not in the current fold.

	MNLR	LDA	XGB
Mean county-level accuracy	94.21%	94.99%	94.94%

The three models achieved similar performance, with county-level accuracy between 94% and 95%.

C. Subject Sex and Race

1. Multinomial Logistic Regression

The obvious model for predicting the classification proportions of a multi-class problem is multinomial logistic regression, so we began there. We calculated the proportions of each demographic category in the traffic stop records for each county and used a matrix of those proportions as the response variable with the year and population demographics broken down by the same categories as predictors.

We then compared the model that includes the year and male and female sex for each category of race in the data across three methods of normalization to determine if normalization improved the predictive performance of the model and, if so, which version of normalization provided the most improvement. The results were as follows:

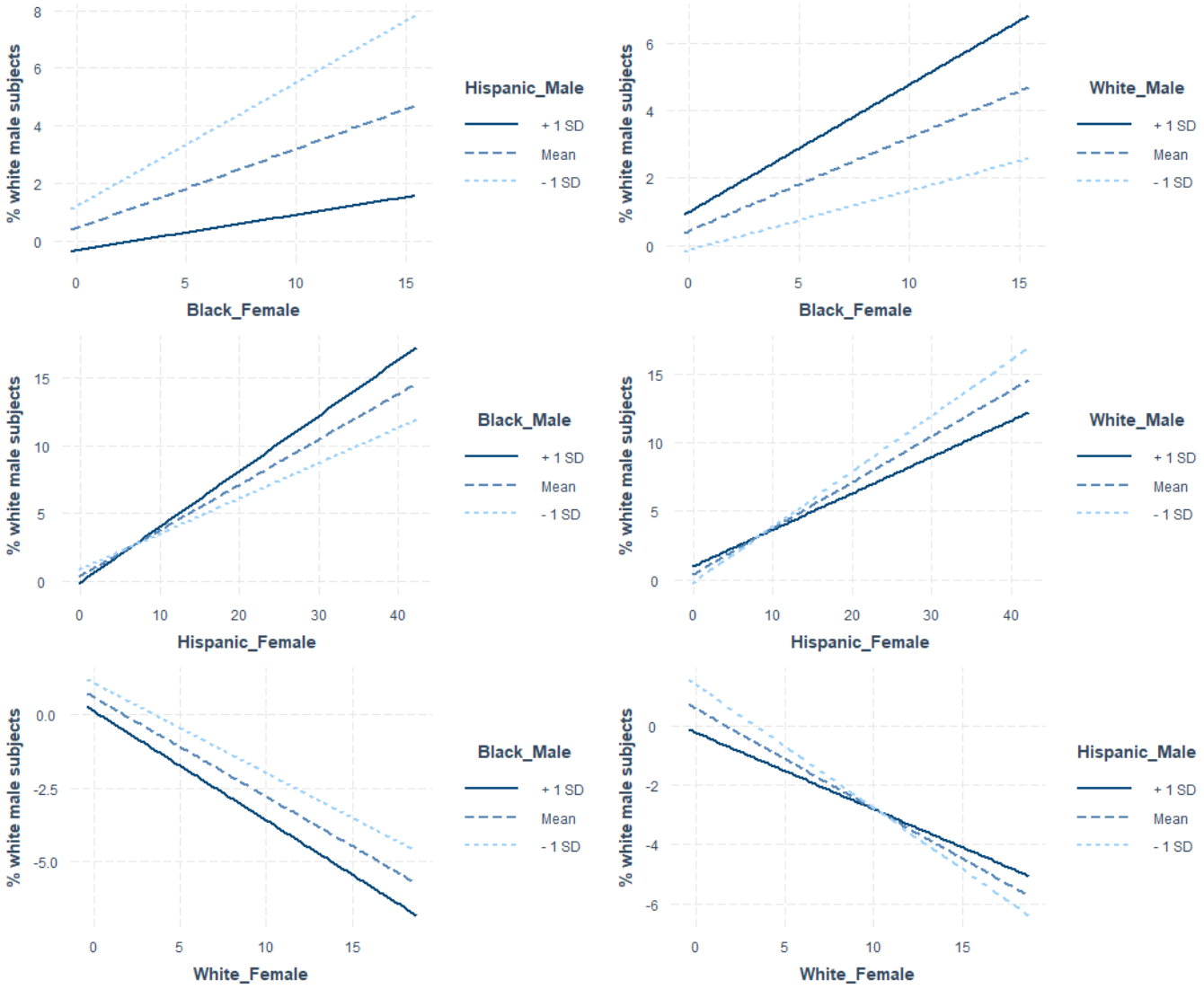
Normalization	AIC	χ^2 vs null model	p-value
Unnormalized	11398.01	346.1 (99 df)	0
Min-max	11401.76	342.35 (99 df)	0
Z-score	11397.89	346.2 (99 df)	0

The model that used z-scored normalized demographic data to predict the proportions of traffics stops by sex and race had the lowest AIC and the highest χ^2 value when compared with the null model, so that data was selected for developing a more refined multinomial model.

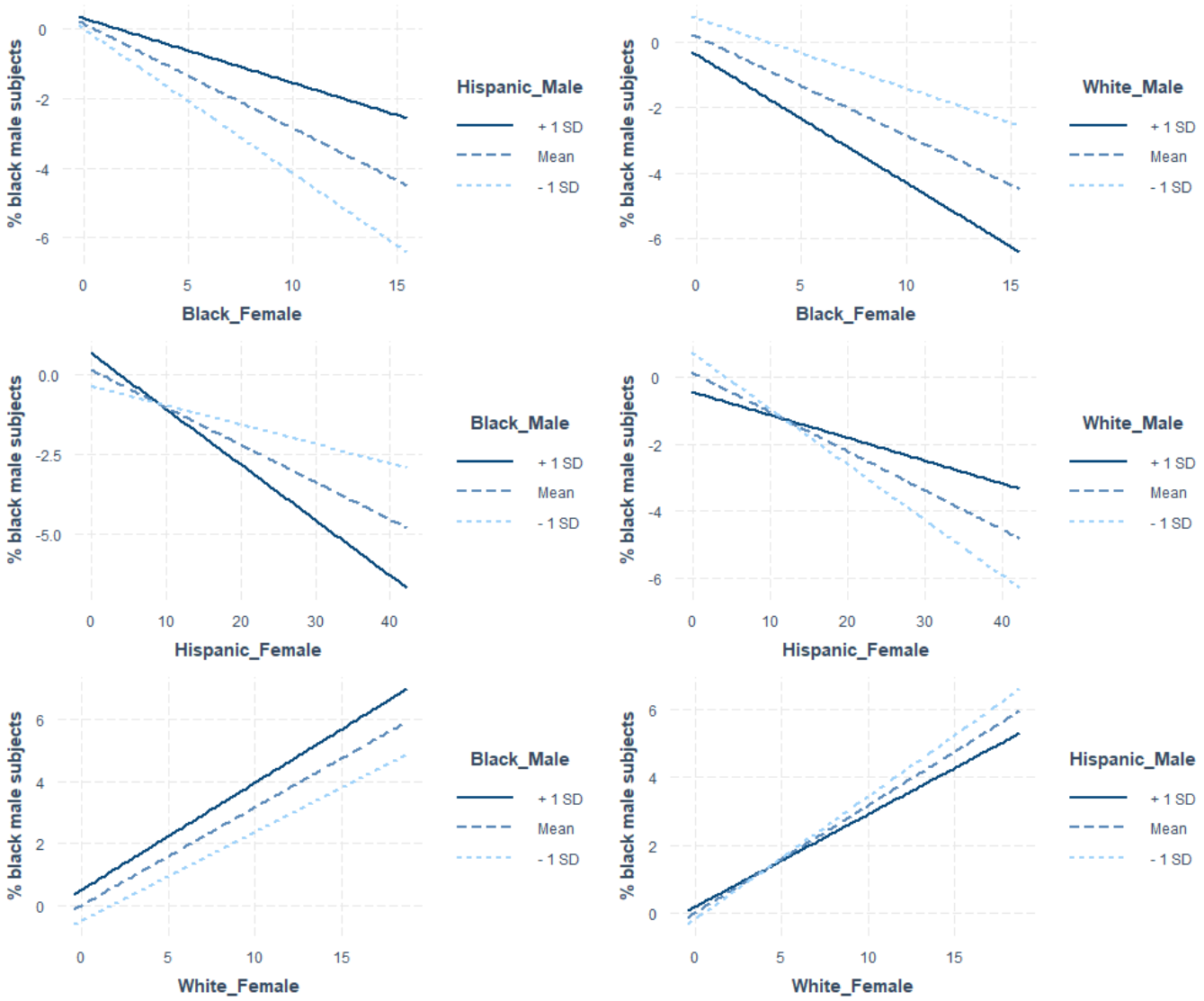
Next we attempted to identify interaction terms by creating linear models of one response variable with two terms likely to interact as predictors, then plotted the effects of those interactions. We used the proportion of stops with a white male subject as the initial response

variable as it had the highest variance of any of the response variables. To limit the number of tests needed, we first reduced the potential choices of interactions by dropping the predictors with the lowest total populations: Asian or Pacific Islander and Other. Then, we removed within-race and within-sex interactions. Finally, we modeled and compared the interactions between one across-sex pairing for each combination of the remaining races.

This produced an initial set of 6 interaction plots, which showed potential interactions between Hispanic Female and Black Male, Hispanic Female and White Male, and White Female and Hispanic Male.

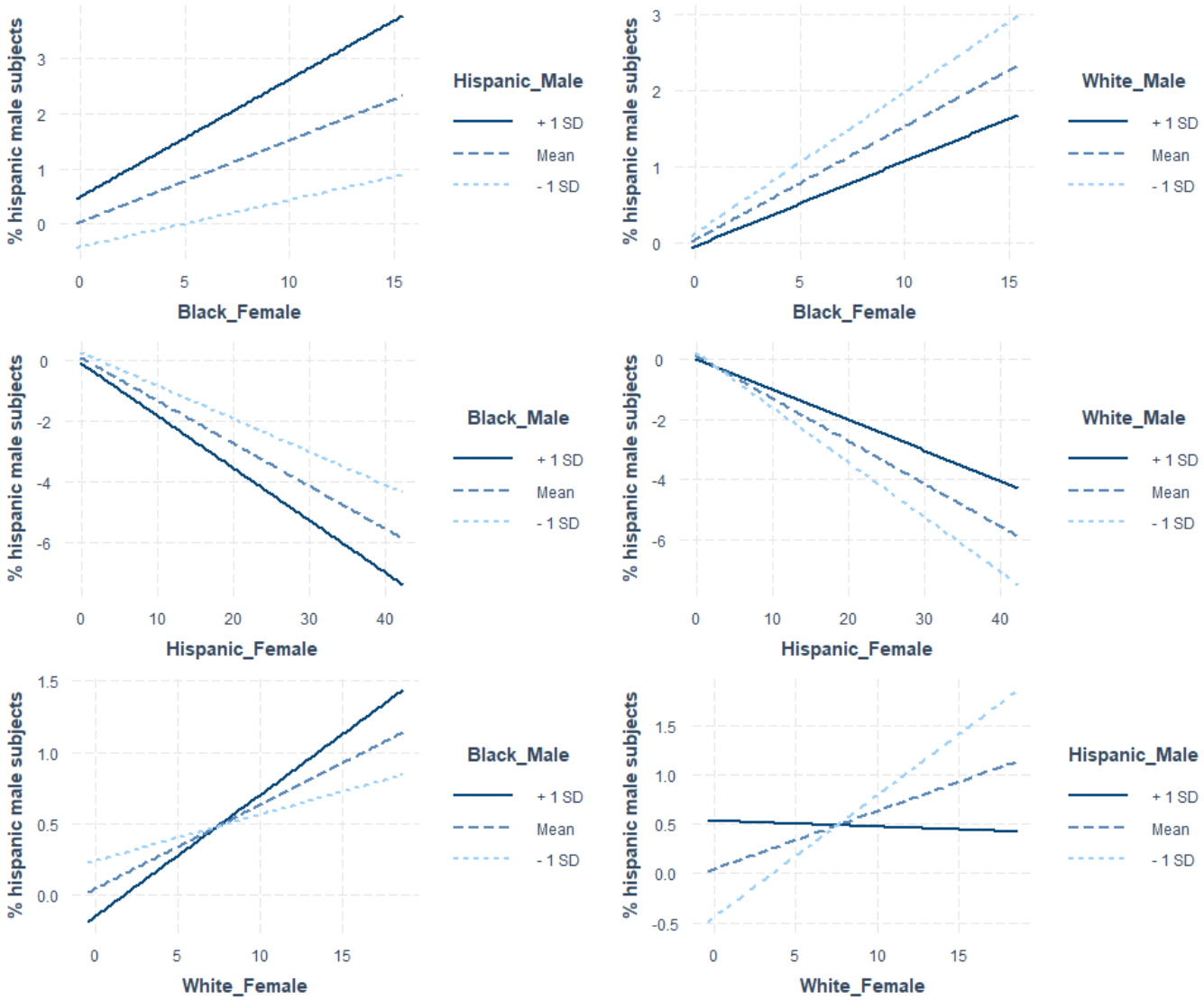


Next we changed the response in the linear model to the response variable with the next highest variance, the proportion of stops with a black male subject and plotted the interactions again. This set of plots showed the same set of potential interactions.



Finally, we selected the response variable reflecting a race which had the next highest variance that had not yet been used as the response, the proportion of stops with a hispanic male subject, and plotted the interactions again.

This set of plots also showed potential interactions between Hispanic Female and White Male, and White Female and Hispanic Male, but also showed a potential interaction between White Female and Black Male.



Since we did not test the race and gender combinations in each direction, we included both directions in the next step of testing, where we created individual multinomial models with each potential interaction term as the sole predictor. We then compared these models to the null model to determine how relevant the included potential interaction term was to predicting our response variables. The results follow.

Term 1	Term 2	χ^2 vs null model	p-value
Hispanic Female	Black Male	16.17	0.0635
Hispanic Male	Black Female	15.81	0.0710
Hispanic Female	White Male	19.37	0.0222*
Hispanic Male	White Female	19.48	0.0214*
White Female	Black Male	22.36	0.0078*
White Male	Black Female	21.67	0.0100*

* Result is significant at the $\alpha = .05$ level

Next, we created a model that included each demographic category and the four interaction terms that appeared to be significant and conducted stepwise selection in both directions using AIC as the criteria. This resulted in the following model and results:

Response	(Intercept)	year	Black_Male	Hispanic_Male	White_Female	Other_Male	Black_Male: White_Female	Hispanic_Male: White_Female
stops_per_api_m	-71.8421	0.0362	-0.1501	-0.0093	0.0043	-0.0291	0.0258	-0.0066
stops_per_b_f	126.0057	-0.0613	1.1397	-0.7966	-0.2656	0.0570	-0.2591	0.1246
stops_per_b_m	150.8361	-0.0734	1.0439	-0.6810	-0.2406	0.0517	-0.2283	0.1082
stops_per_h_f	-108.7589	0.0548	0.1093	0.1885	-0.1608	-0.0066	-0.0176	-0.0002
stops_per_h_m	-7.9211	0.0053	0.2025	0.1000	-0.1840	0.0098	-0.0368	0.0101
stops_per_o_f	-46.4323	0.0233	-0.4871	0.2241	-0.5696	0.2837	0.0387	-0.0049
stops_per_o_m	-77.2702	0.0390	0.1134	0.0879	-0.4828	0.2397	-0.0276	0.0138
stops_per_w_f	31.8108	-0.0137	-0.0828	-0.1703	-0.2376	0.0622	0.0061	0.0116
stops_per_w_m	35.3436	-0.0151	-0.2660	-0.1114	-0.2871	0.0460	0.0639	-0.0088

AIC: 11291.29

χ^2 versus null model: 380.82

p-value: 0

County-level accuracy: 86.35%

2. Linear Discriminant Analysis

As with modeling subject age, the aggregated data had to be split into individual observations for use with the linear discriminant analysis function in R. The demographics for every record in a particular county and year are the same, so the data consist of the same demographic information repeated for each time a given response, labeled name in the table below, appeared in the traffic stop records for that county and year.The first three rows of the disaggregated data set are below.

year	API Female	API Male	Black Female	Black Male	Hispanic Female	Hispanic Male	Other Female	Other Male	White Female	White Male	name
2000	0.000799	0.00088	0.0176	0.0178	0.00153	0.00215	0.00487	0.00507	0.0322	0.0298	api_female
2000	0.000799	0.00088	0.0176	0.0178	0.00153	0.00215	0.00487	0.00507	0.0322	0.0298	api_female
2000	0.000799	0.00884	0.0176	0.0178	0.00153	0.00215	0.00487	0.00507	0.0322	0.0298	api_female

We began by testing a model that included the year and male and female sex for each category of race in the data across three methods of normalization to determine if normalization improved the predictive performance of the model and, if so, which version of normalization provided the most improvement. The results were as follows:

Normalization	County-level accuracy
Unnormalized	79.16%
Min-max	83.97%
Z-score	83.97%

Normalization improved the model's accuracy, but the type of normalization did not have an impact, so we arbitrarily chose to use the min-max normalized data to proceed. Finally, due to the similarities between LDA and logistic regression, we adjusted the model built using min-max normalized data to include the same predictors as the final multinomial model, which produced the following results:

Predictor	LD1	LD2	LD3	LD4	LD5	LD6	LD7
year	0.0478	0.0020	-0.0382	-0.0099	-0.1383	0.1517	-0.1650
Black_Male	-28.8218	0.7935	-0.2386	7.7065	-27.6336	-11.4443	0.8167
Hispanic_Male	15.9097	-33.1562	11.0740	-1.0335	24.7219	15.6093	10.2733
White_Female	-0.5205	1.0062	-1.3461	-9.6367	-10.0834	-8.4135	1.7814
Other_Male	0.0745	-0.4344	-9.6929	8.3292	0.5684	2.2313	4.2278
Black_Male:White_Female	100.8618	-2.6899	-8.2938	-48.9489	197.1665	93.8036	-10.8167
Hispanic_Male:White_Female	-63.2992	32.3274	-4.8040	35.8047	-121.5915	-71.0841	-19.1821

Proportion of trace:

LD1	LD2	LD3	LD4	LD5	LD6	LD7
0.5587	0.3312	0.0576	0.0452	0.0048	0.0022	0.0003

County-level accuracy of 84.53%

3. Random Forest (Probability Forest)

The available Random Forest functions in R also require disaggregated data, however, we were unable to run the random forest model with the entire dataset. Unfortunately, downsampling the data set proved insufficient for producing a viable prediction model. So, we returned to the full set of disaggregated records and randomly sampled 20%. This sample of traffic stop records produced significant improvements in accuracy while also remaining small enough to tune the hyperparameters of the random forest model.

The parameters selected for tuning were the number of trees, the number of variables to consider for splitting at each node and the minimum size of a terminal node. The following values were tested in all combinations: num.trees: 200, 500, 1000; mtry: 2, 3, 4; and min.bucket: 400000, 500000, 600000.

Since out-of-bag error calculation is accomplished by predicting the classification of observations, it is not an appropriate measure for this project. Instead, the models produced using each set of hyperparameters were compared on county level accuracy. The top 5 results are listed in the table below.

num.trees	mtry	min.bucket	County-level accuracy
200	2	400000	83.67%
500	2	400000	83.62%
1000	2	400000	83.61%
500	3	400000	83.60%
1000	3	400000	83.60%

The model with the highest accuracy used 200 trees, considered 2 variables at each node and had a minimum terminal node size of 400000. The top 5 models were separated by less than 0.1% in terms of county level accuracy, and the worst model found in the tuning process still had an accuracy of nearly 81.5%, so tuning did not drastically alter the efficacy of the model.

County-level accuracy: 83.67%

4. Cross-Validation

All three models were evaluated using 10-fold cross-validation by comparing the mean of their county-level accuracies.

	MNLR	LDA	RF
Mean county-level accuracy	86.32%	84.53%	85.03%

The three models achieved similar performance, with county-level accuracy between 84% and 87%.

D. Assessment

1. Age

The accuracies of all three models are very similar. The linear discriminant analysis model had the best accuracy, so it will be the model used for testing. It is important to note, though, that the extreme gradient boosting model was trained on much less data than the other two models due to limitations in time and computational power. It is very possible that had this model been trained on more of the data, it would have performed even better.

2. Subject Race and Sex

Though the accuracies measured during cross-validation of the three models are similar, the multinomial logistic regression model had the highest accuracy, so it will be the model used for testing. As with the extreme gradient boosting model used to model subject age, the random forest model was trained on much less of the data than was used to train the other two models and likewise, it is very possible that a random forest trained on more of the data would have performed even better.

E. Aggregating the Selected Models

To combine the model for race and sex with the model for age, we made predictions based on the training data and then distributed the percentages predicted for age by county and year across the percentages predicted by race and sex. This produced an 82 column data frame with the first two columns representing the year and county and the next 80 columns representing the percentage of stops in a given county-year for a specific combination of subject race, sex and binned age. The application of this technique for one race and sex in a single county year is below.

Predicted proportion of subjects that will be black males: 0.2004

Predicted proportions of subject by age group:

0-20	20-30	30-40	40-50	50-60	60-70	70-80	81+
0.1789	0.3324	0.2239	0.1528	0.0747	0.0269	0.0083	0.0020

So, the predicted proportions of the traffic stops in this county-year that will be black male by age group are:

Black male age 0-20:	0.0359
Black male age 21-30:	0.0666
Black male age 31-40:	0.0449
Black male age 41-50:	0.0306
Black male age 51-60:	0.0150
Black male age 61-70:	0.0054
Black male age 71-80:	0.0017
Black male age 81+:	0.0004

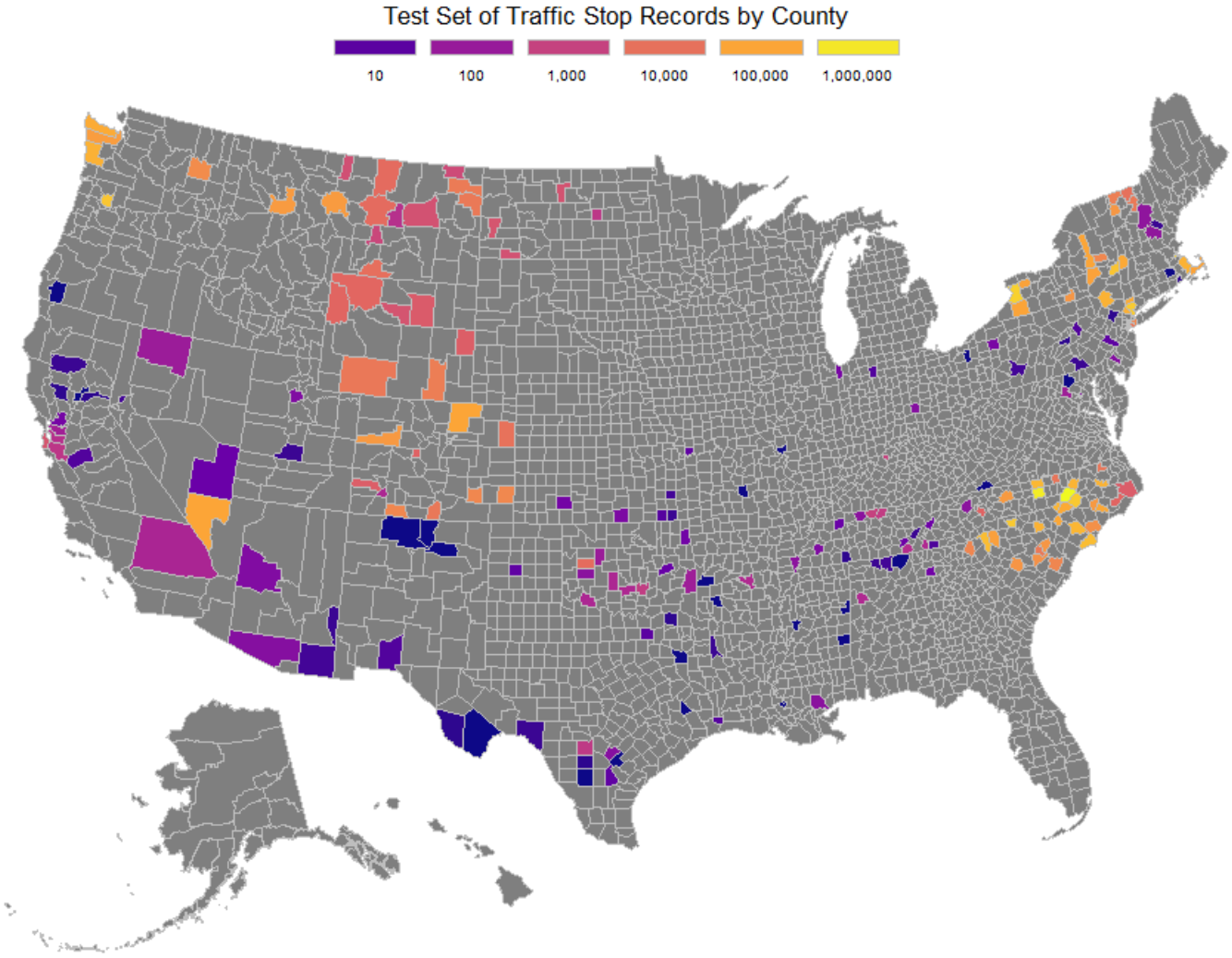
As an initial check on the viability of this approach, the accuracy of this combined model was assessed on the training data, which produced the following results:

County-level accuracy: 83.58%
Overall accuracy: 94.32%

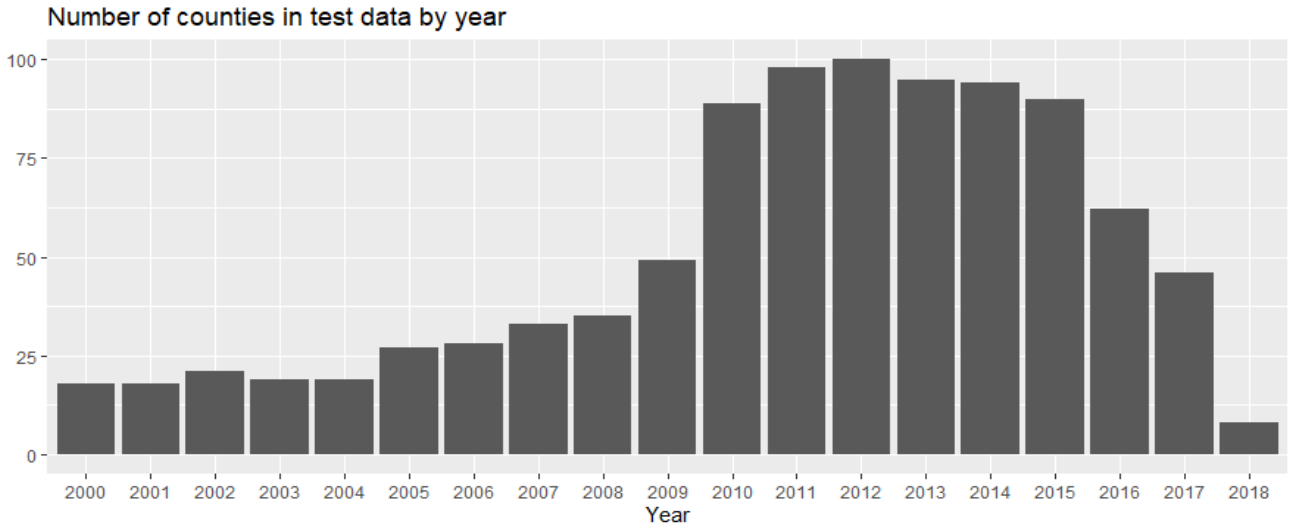
IV. Testing

A. Data

The geographic distribution of the data in the test set is similar to the training set



The distribution of counties by year in the test data is as follows:



The distribution of the test data by year is also very similar to the training set, with a concentration of data in the six years from 2010 to 2015. There are at least 75 counties in each of those years, which should serve as a large enough sample to evaluate the predictive power of the models. We considered the accuracy by county and overall for each of the individual models and the combined model for each of these years and the entire six year period.

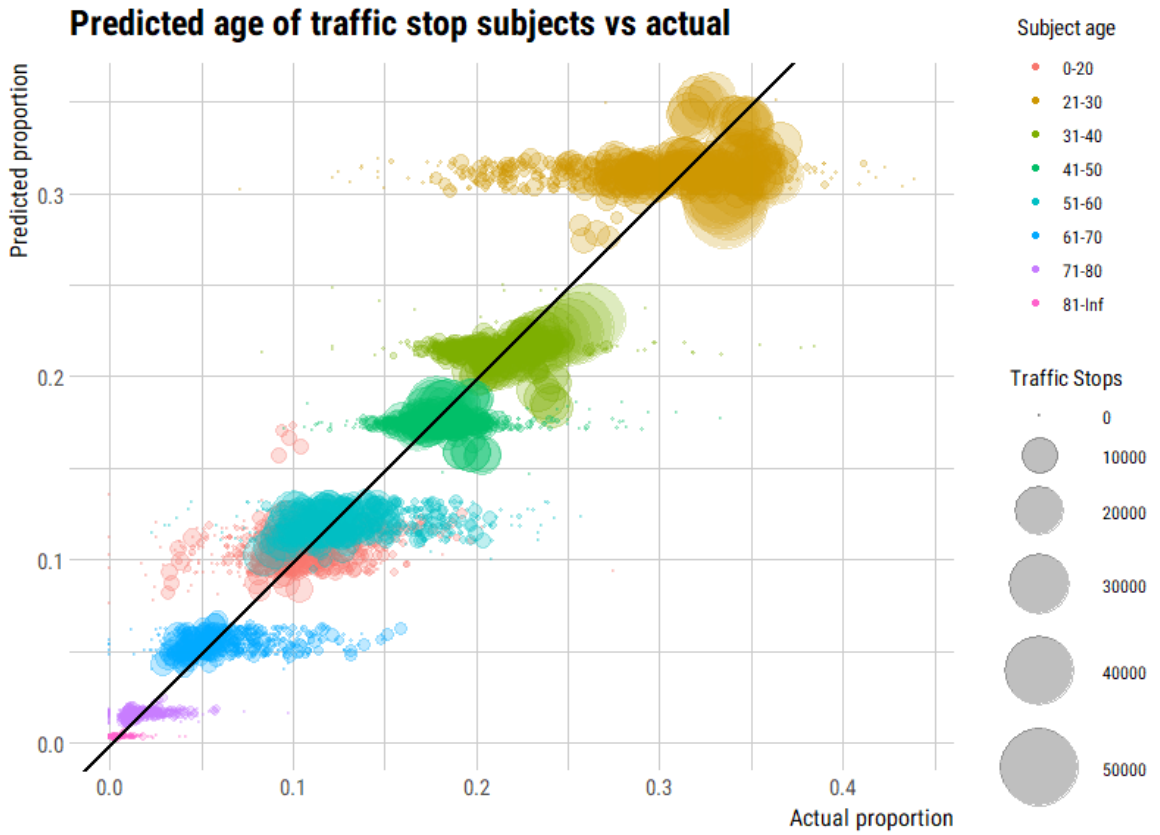
B. Subject Age

Using the linear discriminant analysis model to predict the proportions of traffic stops in the test data for the years 2010 to 2015 by binned subject age produced the following results:

Year	County-level accuracy	Total accuracy
2010	94.45%	98.40%
2011	94.58%	97.95%
2012	94.62%	98.20%
2013	94.76%	98.63%
2014	94.85%	98.90%
2015	94.59%	98.05%

Combined county-level accuracy: 94.64%
Combined total accuracy: 98.83%

A comparison of the model's predictions versus the actual proportions shows just how dominant class means are within the model.



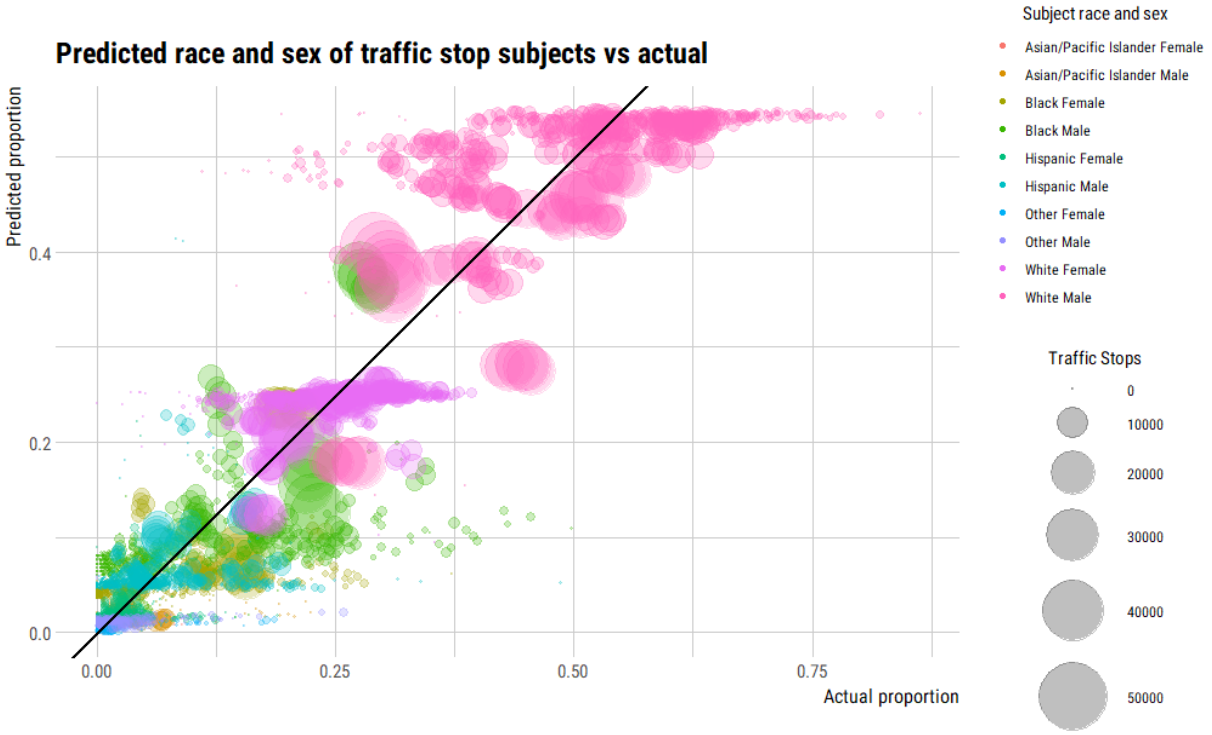
C. Subject Race and Sex

Using the multinomial logistic regression model to predict the proportions of traffic stops in the test data for the years 2010 to 2015 by subject race and sex produced the following results:

Year	County-level accuracy	Total accuracy
2010	86.49%	97.23%
2011	86.36%	97.91%
2012	85.94%	97.95%
2013	85.80%	97.88%
2014	85.94%	96.74%
2015	84.75%	95.38%

Combined county-level accuracy: 85.91%
Combined total accuracy: 98.26%

A comparison of the model's predictions versus the actual proportions by subject race and sex gives an intuitive sense of why the model performs so much better in aggregate than on individual county-years. The model misses frequently, and sometimes quite badly, but its misses are generally evenly split, both in terms of frequency and magnitude, above and below the actual proportions.



D. Subject Race, Sex and Age

Combining the two models' predictions produced the following results:

Year	County-level accuracy	Total accuracy
2010	84.02%	94.02%
2011	84.08%	94.28%
2012	83.73%	94.11%
2013	83.61%	93.86%
2014	83.82%	94.11%
2015	82.60%	93.36%

Combined county-level accuracy: 83.67%
Combined total accuracy: 94.42%

V. Discussion

A. Process

Available computational resources limited several aspects of this project. Processing latitudes and longitudes to identify counties was an extremely costly process in terms of time, taking nearly 60 hours to complete. The boosted and random forest models also suffered, as neither were able to train on the entire training set. Both of these issues could be resolved with the use of web-based data and analytic platforms such as Kaggle, an option that may be pursued in the future.

Normalizing the data was extremely important to building more accurate models. Planning for normalization before pre-processing to ensure the numerous datasets created would not get confused was also very important. In a similar vein, it would have been useful to completely map out the models to be used before pre-processing to ensure that the end result could be piped directly into the functions, rather than preparing the data for one or two functions then adapting on the fly for other models. As a whole, though, the approach to modeling taken here, from analyzing which factors should be modeled together, through the modeling process, to combining the model predictions to create a single result that could be evaluated, was very successful.

B. Results

It is tempting to over-simplify results such as these into spurious conclusions, like “police stop black subjects at a higher rate than their representation in the county where the stop occurs, so police traffic stop decisions are biased against black subjects”. Though true that this data shows police stop black subjects at a higher rate than their representation in the county where the stops occur, there are numerous other potential causes in addition to individual bias, e.g. access to vehicles, need to drive/ability to live close to work, police resourcing and allocation decisions, etc. Any or all of these, and many more, are likely to be causal factors in the disparity. Pinning the results on any one factor elides the role each of the others plays. Furthermore, the correlation of the predictors used in these models makes interpreting their coefficients and log-odds unreliable.

What can be taken from these results is not the models themselves, but the broader picture their predictions paint of the landscape of traffic stop demographics. Subject ages do differ from population ages, but they do so consistently across subject race and sex, indicating subject age is not affected by at least some of the causal factors that drive the disparity between the demographics of traffic stop subjects and the population. Subject sexes differ from population sexes, but do not do so consistently across subject races. Subject races appear to have the most variable relationship with their population demographic counterpart, indicating that of these three categories, subject race is affected the most severely by the afore-mentioned causal factors.

Zeroing in on those factors will require more research. Of particular interest would be modeling traffic stop subject demographics baked on a data source more closely related to who drives and how much, such as the population of drivers or car owners.

VI. Further Research

In addition to acquiring a data set more closely related to the population of motorists, it would also be extremely helpful to model the number of traffic stops that occur in a given county-year. Population demographics may be a useful place to start, but other factors, such as population density, will likely be relevant as well. Those factors may also be able to improve the results of the models created here.

There are also several other potential ways to improve these models. First, the data set is heavily weighted to the years 2010-2015, which resulted in model accuracy that reached its peak in those years and fell off on either side. Balancing the representation of years in the training day would likely create a more robust set of models. Another easy potential improvement is the filtering of demographic data to remove groups that are not relevant to predicting traffic stop demographics, such as those under the age of 16 or older than some predefined limit. Lastly, during this project, the SOPP made more data available which could potentially improve these models or serve as additional data for testing.

VII. Conclusion

The intent of this project was to model the demographics of traffic stop subjects in order to provide insight into the composition of the population of traffic stops in the United States. To do so, we used detailed population demographic data from U.S. Census estimates at the county level, which proved to be a viable set of predictors for traffic stop subject age, race, and sex. The resulting models were reasonably accurate at the individual county level for a given year, but grew significantly more accurate when predicting for multiple counties and years, indicating that the models not only scale well, but likely will perform better the larger the scale on which they are applied.

These models do not explain the differences between population and traffic stop demographics, but do provide a window into just how significant and consistent the differences between the two are. There are numerous potential factors that contribute to this gap, but this research shows that they do not apply evenly across the three demographic categories for which models were created. The differences between the two sets of demographics and the way those differences manifest in each demographic category both warrant far more research to define and explain.

APPENDIX A: Data

1. Traffic stop records summary tables and graphs

Stop records by year

Year	# of Stops	% of Stops
2000	252,343	0.48%
2001	570,413	1.08%
2002	1,287,163	2.45%
2003	1,071,016	2.03%
2004	992,787	1.89%
2005	995,976	1.89%
2006	1,290,312	2.45%
2007	2,117,018	4.02%
2008	2,270,066	4.31%
2009	3,050,840	5.80%
2010	5,223,562	9.92%
2011	5,272,328	10.00%
2012	5,370,608	10.20%
2013	5,188,463	9.86%
2014	5,489,812	10.40%
2015	5,459,726	10.40%
2016	2,897,259	5.50%
2017	2,640,829	5.02%
2018	987,398	1.88%
2019	169,517	0.32%
2020	46,751	0.09%

The largest number of stops occurred in 2014, with the number of stops decreasing both earlier and later than that year.

Percentage of stops by subject race

Subject Race	# of Stops	% of Stops
Asian/Pacific Islander	1,500,866	2.85%
Black	11,183,670	21.24%
Hispanic	4,976,955	9.45%
White	34,084,360	64.74%
Other	898,336	1.70%

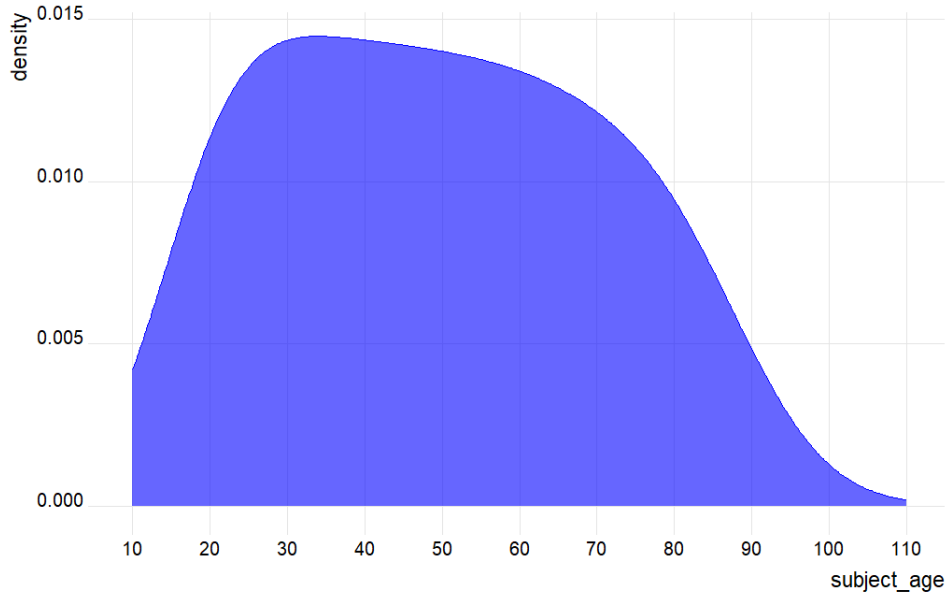
The largest racial demographic in the stop data is white, followed by black and hispanic.

Stop records by subject sex

Subject Sex	# of Stops	% of Stops
Female	18,074,879	34.33%
Male	34,569,308	65.67%

The largest subject sex in the stop data is male, with nearly twice as many results as female.

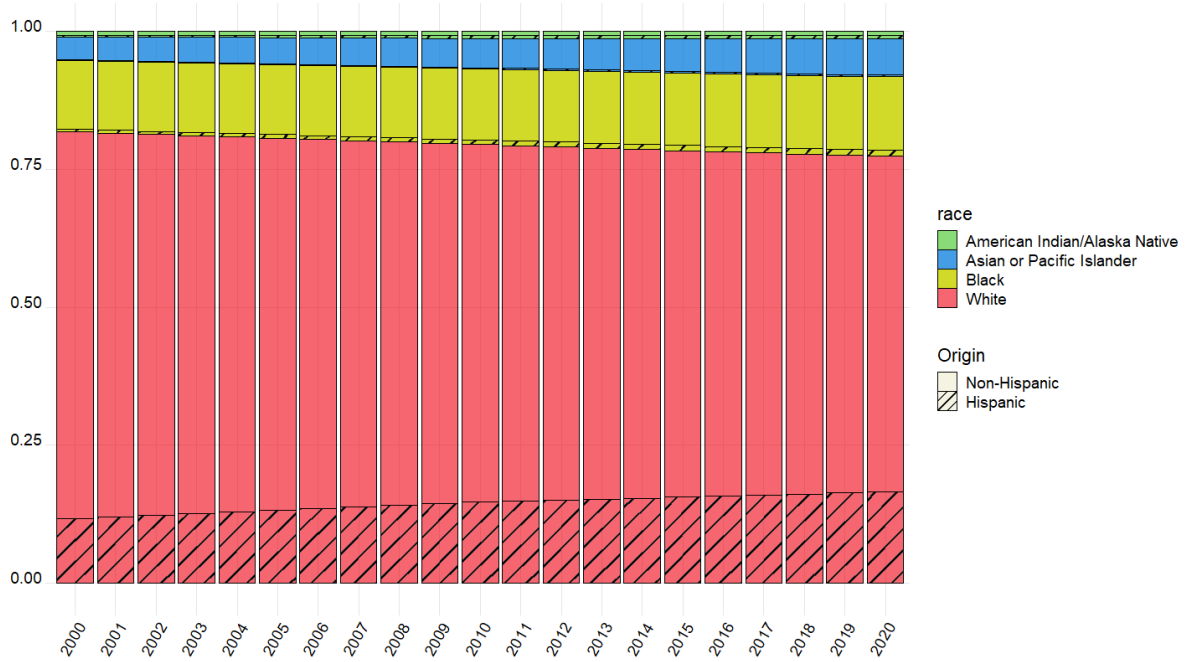
Smoothed density plot of subject age



Age of traffic stop subjects tends to be younger with a slight decrease in the distribution from a peak just above 30 until 60. After age 60, the distribution decreases dramatically.

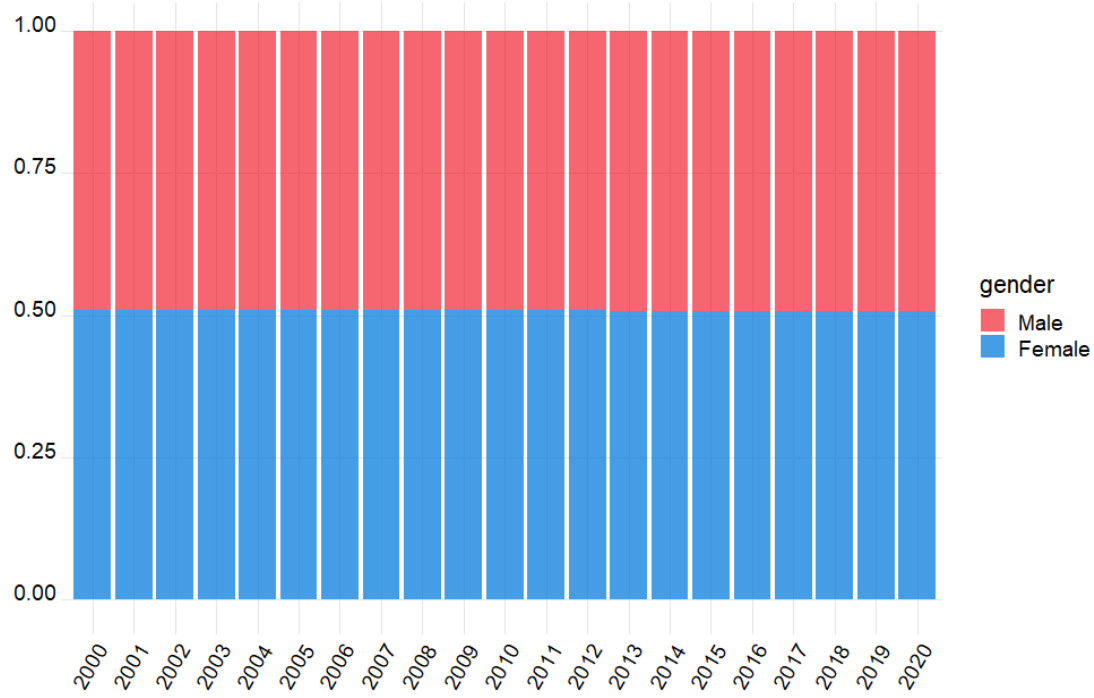
2. Census estimate summary tables and graphs

Percentage of population by race and origin (hispanic or non-hispanic) between 2000 and 2020



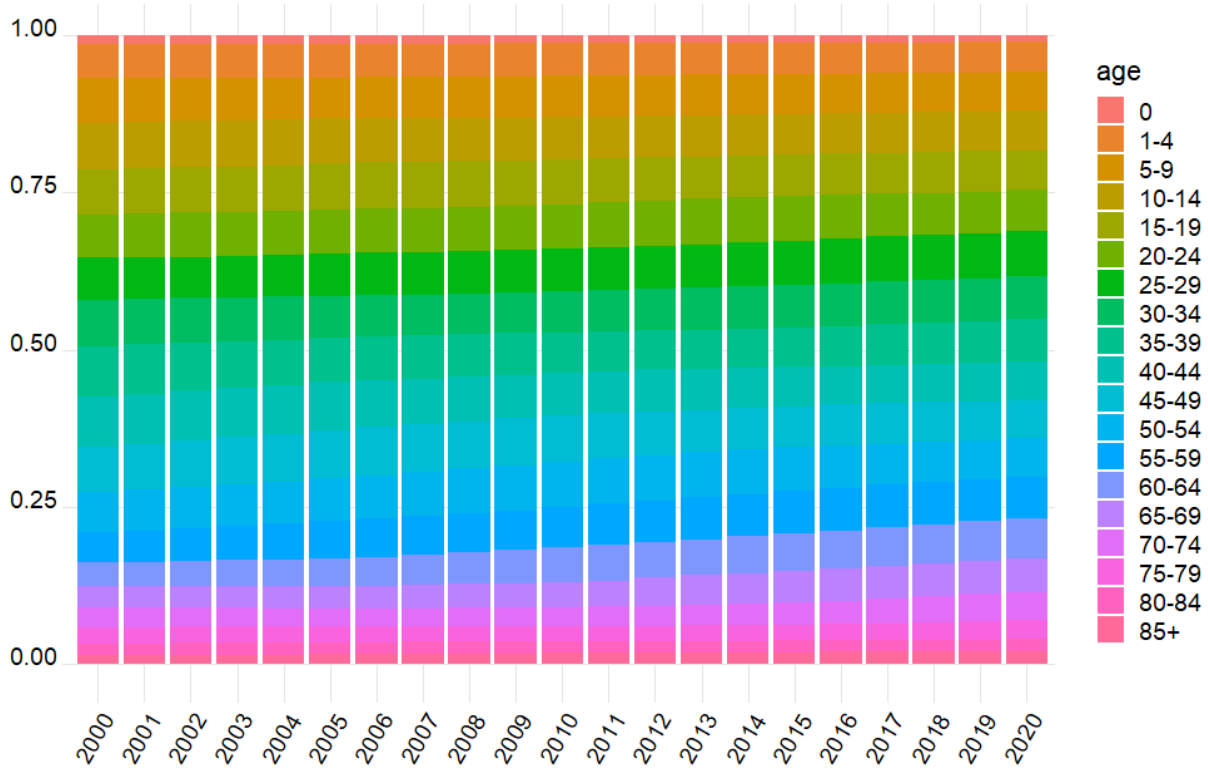
The chart shows that the percentage of the population of hispanic origin grew and the percentage of the population that identifies as non-hispanic, particularly white non-hispanic, decreased.

Percentage of population by gender between 2000 and 2020



The chart shows no significant change in the percentage of population by gender.

Percentage of population by age between 2000 and 2020



The chart shows that the population became slightly older between 2000 and 2020.

3. Traffic stop record files

All traffic stop record files retrieved from https://openpolicing.stanford.edu/data/.

wb225bk3255_az_mesa_2023_01_26.rds
wb225bk3255_co_aurora_2023_01_26.rds
wb225bk3255_ks_wichita_2023_01_26.rds
wb225bk3255_ky_louisville_2023_01_26.rds
wb225bk3255_mt_statewide_2023_01_26.rds
wb225bk3255_ok_oklahoma_city_2023_01_26.rds
wb225bk3255_tx_san_antonio_2023_01_26.rds
wb225bk3255_vt_burlington_2023_01_26.rds
yg821jf8611_ar_little_rock_2020_04_01.rds
yg821jf8611_ca_bakersfield_2020_04_01.rds
yg821jf8611_ca_long_beach_2020_04_01.rds
yg821jf8611_ca_san_diego_2020_04_01.rds
yg821jf8611_ca_san_francisco_2020_04_01.rds
yg821jf8611_co_statewide_2020_04_01.rds
yg821jf8611_ct_hartford_2020_04_01.rds
yg821jf8611_ct_statewide_2020_04_01.rds
yg821jf8611_ky_owensboro_2020_04_01.rds

yg821jf8611_la_new_orleans_2020_04_01.rds
yg821jf8611_ma_statewide_2020_04_01.rds
yg821jf8611_nc_statewide_2020_04_01.rds
yg821jf8611_nd_statewide_2020_04_01.rds
yg821jf8611_nj_camden_2020_04_01.rds
yg821jf8611_nv_henderson_2020_04_01.rds
yg821jf8611_ny_statewide_2020_04_01.rds
yg821jf8611_pa_philadelphia_2020_04_01.rds
yg821jf8611_sc_statewide_2020_04_01.rds
yg821jf8611_tn_nashville_2020_04_01.rds
yg821jf8611_vt_statewide_2020_04_01.rds
yg821jf8611_wa_statewide_2020_04_01.rds
yg821jf8611_wy_statewide_2020_04_01.rds

4. Dataset variables

Relevant variables found in the traffic stop records

Variable Name	**Description**
date	Date traffic stop occurred
lat	Latitude of traffic stop
lng	Longitude of traffic stop
county_name	Name of county
subject_age	Age of subject involved in traffic stop
subject_race	Race of subject involved in traffic stop
subject_sex	Sex of subject involved in traffic stop

The demographic data are stored in a text file with each line representing a population group for a specific combination of location and demographics. Using the following example, 2020WY56045994020800000026, the relevant information contained in each line is explained in the table below.

Variable Name	**Description**
Year (1-4)	The first four digits represent the year of the population estimate or census collection ex. 2020

State (5-6)	The fifth and sixth digits are the State postal abbreviation ex. WY □ Wyoming
FIPS (7-11)	The seventh - eleventh digits represent the five digit FIPS code for the state and county ex. 56045 □ Weston County, WY
Race (14)	The fourteenth digit represents the race of the population group 1 = White 2 = Black 3 = American Indian/Alaska Native 4 = Asian or Pacific Islander ex. 4 □ Asian or Pacific Islander
Origin (15)	The fifteenth digit represents the origin of the population group 0 = Non-Hispanic 1 = Hispanic ex. 0 □ Non-Hispanic
Gender (16)	The sixteenth digit represents the gender of the population group 1 = Male 2 = Female ex. 2 □ Female
Age (17-18)	The seventeenth and eighteenth digits represent the age group of the population group 00 = 0 years 01 = 1-4 years 02 = 5-9 years 03 = 10-14 years 04 = 15-19 years … 17 = 80-84 years 18 = 85+ years ex. 08 □ 35-39 years old
Population (19-26)	The nineteenth - twenty-sixth digits represent the size of the population group ex. 00000026 □ 26 people