This Photo by Unknown Author is licensed under CC BY

SDPD Traffic: Police Action and Race

SECONDARY ANALYSIS OF POLICE VEHICLE STOPS

Matthew C. Vanderbilt | ANA625: Categorical Data Methods, Appl | March 2018

Table of Contents

Abstract

San Diego is the eight-largest city in the United States, with its current population of 1.3 million expected to grow to 1.54 million by 2020 and 1.95 million by 2050. Its population is 44.2% White (non-Hispanic), 28.8% Hispanic, 6.7% Black, 16.4% Asian, 3.2% Multi-racial, and 0.8% Other. With the neighborhood partnerships of the San Diego Police Department, San Diego has become one of the safest metropolitan cities in the country. Given the current racial climate of the United States, however, a statistical review of police action and race can identify the need for any additional outreach or training to provide for non-racially-biased policing and reduce the potential for inequalities resulting in racial tension. The objective of this study was to statistically evaluate the association of police action and race, controlling for gender, age group, San Diego residency, and time-of-day. San Diego Police Department vehicle-stop data from 2014 through 2017 was utilized to provide the review of nearly 400,000 interactions. SAS was utilized to perform univariate and logistic regression analyses, as well as provide for multivariable and multivariate analysis of interactions and investigations of confounding and multicollinearity. The model has fairly-good predictive capabilities as measured by the c-statistic, and the results indicated a statistically-significant association between race and police action, as well as between those variables and gender, age, San Diego residency status, and time-of-day. While many additional factors need to be reviewed to identify causal factors and create opportunities for correction, these results support the need for such effort to occur.

Overview

INTODUCTION

San Diego is the eight-largest city in the United States, with its current population of 1.3 million expected to grow to 1.54 million by 2020 and 1.95 million by 2050.[1] Its population is 44.2% White (non-Hispanic), 28.8% Hispanic, 6.7% Black, 16.4% Asian, 3.2% Multi-racial, and 0.8% Other.[2] San Diego is consistently ranked among the top five safest metropolitan cities within the United States. Its police department consists of 2,781 employees and 840 volunteers[3], committed to community partnerships and compassionate response. With increasing public awareness of tension between police and racial groups due to disparity in treatment, a data-driven approach to reviewing police statistics can provide objective facts to identify opportunities for additional analysis and corrective action. The objective of this analysis is to statistically investigate the association of police action at traffic stops and race[1], after controlling for gender, age group, San Diego residency status, validity of search, and time of day, within a sample of data from the City of San Diego Police Department. The formula below provides a general idea of the model utilized:

$$
\begin{aligned}
&Police\_Action(Y|N) \\
&\quad = \beta_0 + \beta_1 Race\begin{pmatrix} 1 = White \\ 2 = Black \\ 3 = Hispanic \\ 4 = Other \end{pmatrix} + \beta_2 Gender\begin{pmatrix} 1 = Male \\ 2 = Female \end{pmatrix} \\
&\quad + \beta_2 Age\_Group\begin{pmatrix} 1 = Adolescent \\ 2 = Young\ Adult \\ 3 = Middle\ Age \\ 4 = Later\ Adult \end{pmatrix} + \beta_3 SD\_Resident(Y|N) \\
&\quad + \beta_4 Search\_Conducted(Y|N) + \beta_5 Valid\_Search(Y|N) \\
&\quad + \beta_6 Time\_of\_Day\begin{pmatrix} Morning \\ Afternoon \\ Evening \\ Night \end{pmatrix} + \varepsilon
\end{aligned}
$$

where $Police\_Action$ is the variable of interest;
Independent variables or as defined in Appendix B: Data Dictionary
β_0 is the intercept, or constant offset term;
$\beta_{(1\ to\ i)}$ is the regression coefficient; and
ε is the error factor associated with each regression coefficient.

STUDY AND POPULATION DATA

This secondary data analysis was performed using San Diego Police Department *Police Vehicle Stops*[4] provided by the City of San Diego on its DataSD website. This data included all "vehicle stops made by the San Diego Police Department"[4] from 2014 through 2017 in comma-delimited (CSV) format by year, with each year comprised of two tables joined through a one-to-many relationship by [stop_id]. The CSV files were manually imported into Microsoft Access 2016 using the built-in import wizards. Due to a data-shift issue in 2017 caused by the inclusion of quotation marks within search text fields, data was manually

[1] The race and ethnicity terminology used herein is based upon the data dictionary and index descriptions provided by the dataset; no adjustment for popularly accepted, UNESCO, or U.S. Census terms was made.

corrected within the CSV prior to import. After all four years of files were imported, tables were combined into two tables: [combinedVehicleStops] ($N = 390{,}999$) and [combinedVehicleStopsSearchDetails] ($N = 436{,}538$; matched to combinedVehicleStops $N_N = 389{,}175$).

In order to provide for an analysis-optimized flat file, a custom function was written in Microsoft Access using Visual Basic for Applications (VBA) to provide for condensing of [combinedVehicleStopsSearchDetials] to a one-to-one relationship with [combinedVehicleStops] without loss of information; this involved the collapsing of multiple vehicle stop search types and descriptions by [stop_id] into a new [combiendVehicleStopsSearchDetails_Flat] table. A query was then created to combine [combinedVehicleStops] and [combinedVehicleStopsSearchDetails_Flat], as well as indexes with their descriptive values (ex. race code with race description, and to remove the 3,648 records with duplicate [stop_id][2]. Additionally, in order to compensate for data validity issues, several fields were created, including [stop_cause_clean] (normalized), age_val (numeric conversion and removal of data less than 13 or greater than 100), year (from [stop_date]), and timestamp_val (from [stop_date] and [stop_time] to correct for missing [timestamp] data).

The resultant recordset, [combinedVehicleStops_FINAL], was then imported into SAS[3] Enterprise Guide 7.13, where data was filtered for only moving and equipment violations, and additional derived fields were added. Most derived fields leveraged multiple data points to provide a final result, with "valid search" and "police action" having the most interpretation. Valid search is being defined in this study as a search for which contraband was found, property was seize, and/or an arrest was made; it is not a judgment on validity with respect to the law or police guidelines. Police Action is defined as being true if there was a citation, property seizure, or arrest; it is false if only a warning was issued or some other interaction occurred. Data was finally filtered for only those observations containing a specified race for the subject, where gender was binary[4], and where results were Boolean for San Diego residency status and Police Action. This final sample used for investigation ($n = 340{,}893$; 87.2%) contains those males and females with an identified race that had a vehicle-stop interaction with police due to an equipment or moving violation over the period from 2014 through 2017.

STATISTICAL METHODS

SAS Enterprise Guide 7.13 installed on a location machine running Windows 10 Pro, was leveraged for the performance of this study. 2014 through 2017 data was downloaded directly from the City website, modification without record loss (except for indicated duplicates), and imported to SAS. The identified sub-population of records was leveraged to perform a secondary data analysis conducted to identify the expected exposure. All original and derived variables are defined within Appendix B: Data Dictionary. Table 1 was then created to describe the univariate analysis of race, gender, age group, San Diego residency, police search, consent to search, finding of contraband, and generalized time-of-day, to determine statistical association through *Pearson's* chi-square test of categorical data. Table 2 provides the same univariate analysis, but with the addition of police-action results. Table 3 provides the logistic regression analysis to compare the adjusted odds of police action. Interactions were investigated to determine multicollinearity, confounding, and model performance.

[2] [stop_id] is indicated as the primary key; allowing for the inclusion of duplicate data would have created a discrepancy between the primary and secondary tables and resulted in compounded duplication.

[4] The data dictionary indicated Male and Female as the only valid field results, so other values were excluded as potentially invalid information. It is anticipated that non-binary gender information will be collected in the future.

Population Characteristics

The characteristics by race (White, Black, Hispanic, and Other) for the sample population ($n = 340{,}893$; 87.2%) are shown in Table 1, with 147,118 of subjects White; 38,360 Black; 101,628 Hispanic; and 53,787 Other. Of the total, 64.7% of the sample is male and 35.3% female; this data has not been updated for non-binary gender values. Proportionately more Black (70.3%) and Hispanic (67.2%) subjects are male, while proportionately less White (61.4%) subjects are male, and Other subjects match the overall proportionality. Proportionately more Black (61.8%), Hispanic (62.5%), and Other (57.8%) subjects are Young Adult than the total sample (57.4%); while proportionately less were White (52.5%). Relative to the 32.7% of middle-aged subjects, proportionately less are Black (30.9%), Hispanic (30.4%), and Other (32.4%), while proportionately more are White (34.8%). Proportionately more Black (79.6%) and Other (78.9%) subjects are San Diego residents (73.9% overall), while proportionately less White (73.6%) and Hispanic (69.7%) subjects are residents. Of the total sample, 91.1% were not searched, 7.7% were searched, and 1.2% could not be determined. Proportionately more Black (13.6%) and Hispanic (10.2%) subjects were searched; while proportionately less White (5.5%) and Other (5.1%) were subject to search. A statistically-significant association was found between these variables and race at an α level of 0.05 ($p < 0.0001$).

Data for subjects that consented to a search and where contraband was found is primarily limited to those instances where a search was conducted, although there are limited circumstances where a search was not conducted but one or the other has a value. As a search is conducted for less than a tenth of vehicle stops, the corresponding subset of consented search and contraband information is correspondingly limited. Of those interactions where a search did occur, Table 1(b) data shows that proportionately more Black (12.6%) subjects consented to a search than expected from the overall population (7.4%); proportionately less White (5.1%), Hispanic (6.9%), and Other (6.1%) subjects consented. Similarly, proportionately less Black (74.4%) subjects refused to consent to a search than within the overall population (80.5%); proportionately more White (83.7%), Hispanic (80.7%), and Other (81.8%) subjects refused to consent. Of the sample subset where a search did occur, contraband was found in 5.3% of instances, not found in 79.4%, and data was indeterminate for 15.3%. White (5.6%) and Black (6.3%) subjects had proportionately more instances of contraband being found; Hispanic (4.8%) and Other (4.5%) were proportionately less likely to have contraband. Proportionately more White (82.4%), Hispanic (79.6%), and Other (81.9%), subjects did not have contraband found than the overall (79.4%), while proportionately less Black (6.3%) subjects had no contraband. Perhaps skewing these results is that proportionately more Black subjects (13.0% and 20.9%) have indeterminate data than the overall (12.1% and 15.3%). At an α level of 0.05, the associations between these variables and race was found to be statistically significant ($p < 0.0001$).

Table 1. Characteristics of Study Population by Race

	Population	**Race: White**	**Race: Black**	**Race: Hispanic**	**Race: Other**	
	***N* (%)**	***n* (%)**	***n* (%)**	***n* (%)**	***n* (%)**	
Variable	**(N=340,893)**	**(n=147,118)**	**(n=38,360)**	**(n=101,628)**	**(n=53,787)**	***p* value***
Gender						<0.0001
Male (1)	220,467 (64.7%)	90,356 (61.4%)	26,953 (70.3%)	68,335 (67.2%)	34,823 (64.7%)	
Female (2)	120,426 (35.3%)	56,762 (38.6%)	11,407 (29.7%)	33,293 (32.8%)	18,964 (35.3%)	
Age Group						<0.0001
Adolescent 13-18 (1)	8,072 (2.4%)	3,702 (2.5%)	657 (1.7%)	2,389 (2.4%)	1,324 (2.5%)	
Young Adult 19-39 (2)	195,530 (57.4%)	77,248 (52.5%)	23,694 (61.8%)	63,512 (62.5%)	31,076 (57.8%)	
Middle Age 40-64 (3)	111,341 (32.7%)	51,198 (34.8%)	11,850 (30.9%)	30,845 (30.4%)	17,448 (32.4%)	
Later Adult ≥65 (4)	14,829 (4.4%)	9,064 (6.2%)	1,035 (2.7%)	2,551 (2.5%)	2,179 (4.1%)	
Unavailable (5)	11,121 (3.3%)	5,906 (4.0%)	1,124 (2.9%)	2,331 (2.3%)	1,760 (3.3%)	
San Diego Resident						<0.0001
Yes (1)	252,045 (73.9%)	108,240 (73.6%)	30,551 (79.6%)	70,804 (69.7%)	42,450 (78.9%)	
No (0)	88,848 (26.1%)	38,878 (26.4%)	7,809 (20.4%)	30,824 (30.3%)	11,337 (21.1%)	
Search Conducted						<0.0001
Yes (1)	26,375 (7.7%)	8,012 (5.5%)	5,202 (13.6%)	10,403 (10.2%)	2,758 (5.1%)	
No (0)	310,505 (91.1%)	137,483 (93.5%)	32,647 (85.1%)	89,896 (88.5%)	50,479 (93.9%)	
Indeterminate (5)	4,013 (1.2%)	1,623 (1.1%)	511 (1.3%)	1,329 (1.3%)	550 (1.0%)	
Consented to Search						<0.0001
Yes (1)	1,950 (0.6%)	408 (0.3%)	656 (1.7%)	718 (0.7%)	168 (0.3%)	
No (0)	21,228 (6.2%)	6,704 (4.6%)	3,870 (10.1%)	8,397 (8.3%)	2,257 (4.2%)	
No Search (2)	310,504 (91.1%)	137,483 (93.5%)	32,646 (85.1%)	89,896 (88.5%)	50,479 (93.9%)	
Indeterminate (5)	7,211 (2.1%)	2,523 (1.7%)	1,188 (3.1%)	2,617 (2.6%)	883 (1.6%)	
Contraband Found						<0.0001
Yes (1)	1,398 (0.4%)	451 (0.3%)	328 (0.9%)	494 (0.5%)	125 (0.2%)	
No (0)	20,930 (6.1%)	6,605 (4.5%)	3,786 (9.9%)	8,281 (8.2%)	2,258 (4.2%)	
No Search (2)	310,505 (91.1%)	137,483 (93.5%)	32,647 (85.1%)	89,896 (88.5%)	50,479 (93.9%)	
Indeterminate (5)	8,060 (2.4%)	2,579 (1.8%)	1,599 (4.2%)	2,957 (2.9%)	925 (1.7%)	
Valid Search Conducted						<0.0001
Yes (1)	5,665 (1.7%)	1,670 (1.1%)	1,095 (2.9%)	2,428 (2.4%)	472 (0.9%)	
No (0)	18,485 (5.4%)	5,899 (4.0%)	3,377 (8.8%)	7,127 (7.0%)	2,082 (3.9%)	
No Search (2)	310,505 (91.1%)	137,483 (93.5%)	32,647 (85.1%)	89,896 (88.5%)	50,479 (93.9%)	
Indeterminate (5)	6,238 (1.8%)	2,066 (1.4%)	1,241 (3.2%)	2,177 (2.1%)	754 (1.4%)	
Generalized Time of Day						<0.0001
Morning 05:00 - 11:59 (1)	121,210 (35.6%)	56,024 (38.1%)	11,184 (29.2%)	35,928 (35.4%)	18,074 (33.6%)	
Afternoon 12:00 - 16:59 (2)	85,126 (25.0%)	38,770 (26.4%)	8,684 (22.6%)	25,091 (24.7%)	12,581 (23.4%)	
Evening 17:00 - 20:59 (3)	50,013 (14.7%)	19,572 (13.3%)	6,550 (17.1%)	15,809 (15.6%)	8,082 (15.0%)	
Night 21:00 - 04:59 (4)	84,544 (24.8%)	32,752 (22.3%)	11,942 (31.1%)	24,800 (24.4%)	15,050 (28.0%)	

* *p* values based on Pearson chi-square test of association.

Table 1(b). Characteristics of Study Population by Race where a Search was Performed

	Population	**Race: White**	**Race: Black**	**Race: Hispanic**	**Race: Other**	
	***N* (%)**	***n* (%)**	***n* (%)**	***n* (%)**	***n* (%)**	
Variable	**(N=340,893)**	**(n=147,118)**	**(n=38,360)**	**(n=101,628)**	**(n=53,787)**	***p* value***
Consented to Search						<0.0001
Yes (1)	1,949 (7.4%)	408 (5.1%)	655 (12.6%)	718 (6.9%)	168 (6.1%)	
No (0)	21,228 (80.5%)	6,704 (83.7%)	3,870 (74.4%)	8,397 (80.7%)	2,257 (81.8%)	
Indeterminate (5)	3,198 (12.1%)	900 (11.2%)	677 (13.0%)	1,288 (12.4%)	333 (12.1%)	
Contraband Found						<0.0001
Yes (1)	1,398 (5.3%)	451 (5.6%)	328 (6.3%)	494 (4.8%)	125 (4.5%)	
No (0)	20,930 (79.4%)	6,605 (82.4%)	3,786 (72.8%)	8,281 (79.6%)	2,258 (81.9%)	
Indeterminate (5)	4,047 (15.3%)	956 (11.9%)	1,088 (20.9%)	1,628 (15.7%)	375 (13.6%)	
Valid Search						<0.0001
Yes (1)	5,665 (21.5%)	1,670 (20.8%)	1,095 (21.1%)	2,428 (23.3%)	472 (17.1%)	
No (0)	18,485 (70.1%)	5,899 (73.6%)	3,377 (64.9%)	7,127 (68.5%)	2,082 (75.5%)	
Indeterminate (5)	2,225 (8.4%)	443 (5.5%)	730 (14.0%)	848 (8.2%)	204 (7.4%)	

* *p* values based on Pearson chi-square test of association.

Associations of Arrest

Race, gender, age group, San Diego residency, whether a search was conducted, if the search was consented to, if contraband was found, if the search was valid (i.e. search led to action/finding), and generalized time of day, are detailed by police action in Table 2. For the purposes of this analysis, police action is defined as citation, property seizure, and arrest, as opposed to verbal warning, written warning, and other. Data shows that 37.4% ($n = 127,536$) of the sample received a warning / other (i.e. no police action), while 62.6% ($n = 213,357$) received a citation, property seizure, or arrest. The association between all variables was statistically significant at an α level of 0.05 (search conducted $p = 0.0002$; all other $p < 0.0001$).

Of those subjects that had no police action, the proportion of Black subjects was higher at 14.3% ($n = 18,212$) than the overall sample at 11.3% ($N = 38,360$); the proportion of White (42.0% vs 43.2%; $n = 53,568$), Hispanic (28.0% vs 29.8%; $n = 35,752$), and Other (15.7% vs 15.8%;$n = 20,004$), subjects were all lower. Of those stops that culminated in police action, the proportion of subjects that were Black (9.4%; $n = 20,148$) were lower than the overall population (11.3%); the proportion of White (43.9% vs 43.2%; $n = 93,550$) and Hispanic (30.9% vs 29.8%; $n = 65,876$) subjects were lower. The sample population showed that 64.7% (; $N = 220,467$) were male and 35.3% ($N = 120,426$) were female. Of those incidents that had no police action, proportionately more subjects were male (66.7%; $n = 85,052$) and proportionately less were female (33.3%; $n = 42,484$). This is slightly reversed in those incidents that had police action, where proportionately less subjects were male (63.5%; $n = 135,415$) and proportionately more were female (36.5%; $n = 77,942$). Age groups were split 2.4% adolescents aged 13 through 18 ($N = 8,072$), 57.4% young adults aged 19 through 39 ($N = 195,530$), 32.7% middle ages 40 through 64 ($N = 111,341$), and 4.4% later adults at least 65 ($N = 14,829$); data was unavailable or suspect for 3.3% ($N = 11,121$). In those incidents with no police action, proportionately more subjects were adolescent (2.6%; $n = 3,371$), young adult (58.2%; $n = 74,197$), and later adults (32.7%; $n = 41,739$). Proportionately less subjects were adolescent (2.2%; $n = 4,701$), young adult (56.9%; $n = 121,333$), middle aged (32.6%; $n = 69,602$), or later adult (4.1%; $n = 8,725$), in those incidents that culminated in police action. While the sample was split 73.9% ($N = 252,045$) San Diego residents and 26.1% ($N = 88,848$) non-residents, proportionately more of the non-police-action data were non-residents (27.1%; $n = 34,510$) and proportionately more of the police-action data were residents (74.5%; $n = 159,019$). While the sample was split 35.6% ($N = 121,210$) morning, 25.0% ($N = 85,126$) afternoon, 14.7% N=50,013) evening, and 24.8% N=84,544) night, proportionately more incidents culminating in police action occurred in the morning (39.4%; $n = 84,081$) and afternoon (27.2%; $n = 58,126$), while proportionately more incidents not culminating in police action occurred in the evening (17.7%; $n = 22,559$) and night (32.0%; $n = 40,848$)[5].

Although statistically significant, the proportional differences by police-action category from the overall sample differed by no more than 0.4 points for searches being conducted, consent to search, and contraband being found. Across the valid-search categorization, however, the overall sample was split 1.7% ($N = 5,665$) valid, 5.4% ($N = 18,485$) not valid, 91.1% ($N = 310,505$) without search, and 1.8% ($N = 6,238$) indeterminate. The proportion of valid searches culminating in police action was higher at 2.6% ($n = 5,517$), while the proportion of invalid searches not culminating in police action was higher at 6.4% ($n = 8,171$). Given that validity of search is a compound field necessitating police action, this is expected.

[5] While different than anticipated, this association seems reasonable given that morning and afternoon incidents would cover rush-hour traffic (i.e. people potentially running late to work).

Table 2. Associations of Police Action by Race and Other Characteristics

	Population	**Warning / Other**	**Citation / Arrest**	
	***N* (%)**	***n* (%)**	***n* (%)**	
Variable	**(N=340,893)**	**(n=127,536)**	**(n=213,357)**	***p* value***
Race				<0.0001
White (1)	147,118 (43.2%)	53,568 (42.0%)	93,550 (43.9%)	
Black (2)	38,360 (11.3%)	18,212 (14.3%)	20,148 (9.4%)	
Hispanic (3)	101,628 (29.8%)	35,752 (28.0%)	65,876 (30.9%)	
Other (4)	53,787 (15.8%)	20,004 (15.7%)	33,783 (15.8%)	
Gender				<0.0001
Male (1)	220,467 (64.7%)	85,052 (66.7%)	135,415 (63.5%)	
Female (2)	120,426 (35.3%)	42,484 (33.3%)	77,942 (36.5%)	
Age Group				<0.0001
Adolescent 13-18 (1)	8,072 (2.4%)	3,371 (2.6%)	4,701 (2.2%)	
Young Adult 19-39 (2)	195,530 (57.4%)	74,197 (58.2%)	121,333 (56.9%)	
Middle Age 40-64 (3)	111,341 (32.7%)	41,739 (32.7%)	69,602 (32.6%)	
Later Adult ≥65 (4)	14,829 (4.4%)	6,104 (4.8%)	8,725 (4.1%)	
Unavailable (5)	11,121 (3.3%)	2,125 (1.7%)	8,996 (4.2%)	
San Diego Resident				<0.0001
Yes (1)	252,045 (73.9%)	93,026 (72.9%)	159,019 (74.5%)	
No (0)	88,848 (26.1%)	34,510 (27.1%)	54,338 (25.5%)	
Search Conducted				0.0002
Yes (1)	26,375 (7.7%)	9,633 (7.6%)	16,742 (7.9%)	
No (0)	310,505 (91.1%)	116,318 (91.2%)	194,187 (91.0%)	
Indeterminate (5)	4,013 (1.2%)	1,585 (1.2%)	2,428 (1.1%)	
Consented to Search				<0.0001
Yes (1)	1,950 (0.6%)	1,092 (0.9%)	858 (0.4%)	
No (0)	21,228 (6.2%)	7,896 (6.2%)	13,332 (6.3%)	
No Search (2)	310,504 (91.1%)	116,317 (91.2%)	194,187 (91.0%)	
Indeterminate (5)	7,211 (2.1%)	2,231 (1.8%)	4,980 (2.3%)	
Contraband Found				<0.0001
Yes (1)	1,398 (0.4%)	148 (0.1%)	1,250 (0.6%)	
No (0)	20,930 (6.1%)	8,231 (6.5%)	12,699 (6.0%)	
No Search (2)	310,505 (91.1%)	116,318 (91.2%)	194,187 (91.0%)	
Indeterminate (5)	8,060 (2.4%)	2,839 (2.2%)	5,221 (2.5%)	
Valid Search Conducted				<0.0001
Yes (1)	5,665 (1.7%)	148 (0.1%)	5,517 (2.6%)	
No (0)	18,485 (5.4%)	8,171 (6.4%)	10,314 (4.8%)	
No Search (2)	310,505 (91.1%)	116,318 (91.2%)	194,187 (91.0%)	
Indeterminate (5)	6,238 (1.8%)	2,899 (2.3%)	3,339 (1.6%)	
Generalized Time of Day				<0.0001
Morning 05:00 - 11:59 (1)	121,210 (35.6%)	37,129 (29.1%)	84,081 (39.4%)	
Afternoon 12:00 - 16:59 (2)	85,126 (25.0%)	27,000 (21.2%)	58,126 (27.2%)	
Evening 17:00 - 20:59 (3)	50,013 (14.7%)	22,559 (17.7%)	27,454 (12.9%)	
Night 21:00 - 04:59 (4)	84,544 (24.8%)	40,848 (32.0%)	43,696 (20.5%)	

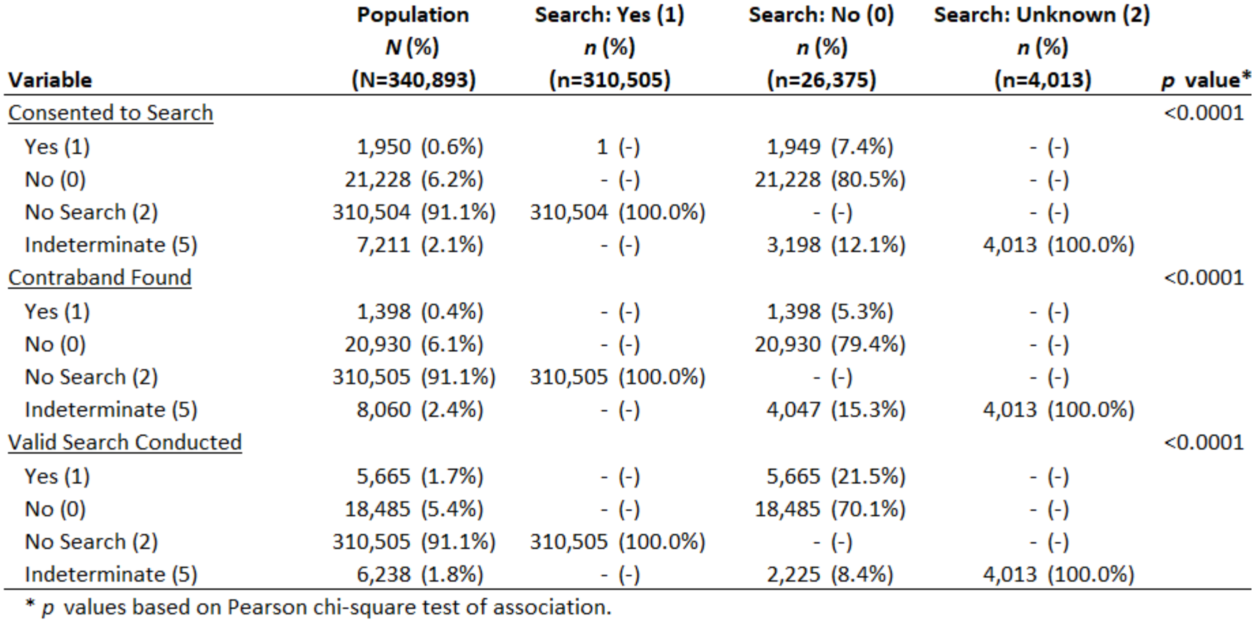
* *p* values based on Pearson chi-square test of association.

Variable Interaction Evaluation and Adjustment

VARIABLE ASSOCIATION

Table 3(a) provides the association of search consent, contraband found, and search validity, with whether or not a search occurs. As may be seen and supplemented with the respective Pearson chi-square statistics of 507,587 ($p < 0.0001$), 484,577 ($p < 0.0001$), and 541,694 ($p < 0.0001$), there is a statistically-significant association at the α level of 0.05. Given that the data is designed for consent and contraband to only be completed if a search occurs, and given that search validity necessitates a search, the strong association is as expected. Inclusion of the Boolean search field with the other variables could result in erroneous modeling results due to the interaction of variables.

Table 1(a). Association of Search and Related Fields



	Population	Search: Yes (1)	Search: No (0)	Search: Unknown (2)	
	N (%)	*n* (%)	*n* (%)	*n* (%)	
Variable	**(N=340,893)**	**(n=310,505)**	**(n=26,375)**	**(n=4,013)**	***p* value***
Consented to Search					<0.0001
Yes (1)	1,950 (0.6%)	1 (-)	1,949 (7.4%)	- (-)	
No (0)	21,228 (6.2%)	- (-)	21,228 (80.5%)	- (-)	
No Search (2)	310,504 (91.1%)	310,504 (100.0%)	- (-)	- (-)	
Indeterminate (5)	7,211 (2.1%)	- (-)	3,198 (12.1%)	4,013 (100.0%)	
Contraband Found					<0.0001
Yes (1)	1,398 (0.4%)	- (-)	1,398 (5.3%)	- (-)	
No (0)	20,930 (6.1%)	- (-)	20,930 (79.4%)	- (-)	
No Search (2)	310,505 (91.1%)	310,505 (100.0%)	- (-)	- (-)	
Indeterminate (5)	8,060 (2.4%)	- (-)	4,047 (15.3%)	4,013 (100.0%)	
Valid Search Conducted					<0.0001
Yes (1)	5,665 (1.7%)	- (-)	5,665 (21.5%)	- (-)	
No (0)	18,485 (5.4%)	- (-)	18,485 (70.1%)	- (-)	
No Search (2)	310,505 (91.1%)	310,505 (100.0%)	- (-)	- (-)	
Indeterminate (5)	6,238 (1.8%)	- (-)	2,225 (8.4%)	4,013 (100.0%)	

* *p* values based on Pearson chi-square test of association.

An evaluation of age group and time-of-day was also performed, resulting in a Pearson chi-squared statistic of 8,643.1 with *p*-value of less than 0.0001, indicating a statistically-significant association. While it may be reasonable for a slight association between minors and adults, it is doubtful that a significant association could exist. Therefore, it is expected that the quantity of observations being reviewed is resulting in a skewed chi-square value that is limiting the usefulness of the *p*-value reviewed in absence of other statistics.

MULTIVARIABLE LOGISTIC REGRESSION

Multivariable logistic regression was performed modeling police action by race, gender, age group, San Diego residency, whether a search occurred, whether consent for a search was obtained, whether contraband was found during search, whether the search was valid, and time-of-day. Additionally, multivariable interactions of search and consent, contraband, and validity; validity and consent and contraband; and age group and time-of-day, as also reviewed. This identified an extremely low Wald Chi-Square value for contraband and search validity (0.0001; $p = 0.9926$). Given that the *p*-value is above the α level of 0.05, the null hypothesis fails to be rejected, meaning that the regression coefficient for the multivariable fields is not statistically different from zero in estimating police action given the other predictors modeled. Therefore, no interaction appears to be occurring between these variables.

However, relatively low Wald Chi-Square values were also identified for consent and search validity (41.52; $p < 0.0001$) and age-group and time-of-day (61.34; $p < 0.0001$). Given that the p-values are above the α level of 0.05, the null hypothesis is accepted, meaning that the regression coefficient for the multivariable fields is statistically different from zero in estimating police action given the other predictors modeled. To compensate for this, either consent or search validity will need to be removed from the final model. Given that age group and time-of-day are reasonably expected to only have a loose correlation, those are not being removed at this time from the final model.

Table 3(b). Variable Interaction Between Related Fields (with singular duplication)

Effect	DF	Wald Chi-Square	Pr > ChiSq
race_val	3	1564.3742	<.0001
gender_val	1	89.9803	<.0001
age_group_val	4	368.9703	<.0001
resident_val	1	115.2678	<.0001
searched_val	2	60746.6116	<.0001
obtained_consent_val	2	54.8642	<.0001
contraband_found_val	2	0.1012	0.9507
valid_search	2	1.2540	0.5342
searched_*obtained_c	0		
searched_*contraband	0		
searched_*valid_sear	0		
obtained_*valid_sear	5	41.5158	<.0001
contraban*valid_sear	1	0.0001	0.9926
time_of_day	3	5066.1040	<.0001
age_group*time_of_da	12	61.3378	<.0001

Table 3(c) was created to separately identify potential interactions absent singular forms of the variables. The multivariable searched and contraband found, searched and search validity, and contraband and search validity, resulted in low Wald Chi-Square values (0.10, 1.31, and 0.0001, respectively) with p-values above the α level of 0.05 (0.9531, 0.5199, and 0.9926), indicating that compound association does not exist for these multivariables. Searched and consent, consent and search validity, and age group and time of day, however, have elevated Wald Chi-Square values (57.69, 172.74, and 4,046.24, respectively) with p-values of less than 0.0001 all below the α. This indicates that there is a level of linear correlation between those multivariables that could be corrected by the removal of the consent variable.

Table 3(c). Variable Interaction Between Related Fields (without singular duplication)

Effect	DF	Wald Chi-Square	Pr > ChiSq
race_val	3	1788.2041	<.0001
gender_val	1	230.0151	<.0001
resident_val	1	138.6264	<.0001
searched_*obtained_c	3	57.6892	<.0001
searched_*contraband	2	0.0962	0.9531
searched_*valid_sear	2	1.3082	0.5199
obtained_*valid_sear	6	172.7368	<.0001
contraban*valid_sear	1	0.0001	0.9926
age_group*time_of_da	12	4046.2411	<.0001

MULTICOLLINEARITY INVESTIGATION

The model of police action as predicted by race, gender, age group, San Diego residency, search, consent to search, finding of contraband, validity of search, and time of day, was run with SAS "PROC REG …/ vif tol" to investigate the variance inflation factor (VIF). As shown in Table 4(d), the variables with the highest VIF were consent-to-search at 3.44, contraband found at 7.07, and search validity at 4.40. Given that consent to search was identified as having an interaction with other variables, it was removed from the model, resulting in the data shown in Table 3(e). As shown, this still resulted in high VIFs for contraband and search validity, even though the multivariable regression did not indicate a concern. Given that search validity was created as a compound variable using the finding of contraband, it was removed and the model re-run.

Table 3(d). Regression Parameter Estimates with Variance Inflation Factor (VIF)

Variable	DF	Param Est.	Std. Err.	t Value	Pr > |t|	Tolerance	VIF
Intercept	1	0.70477	0.00544	129.44	<.0001		0
Race	1	0.00610	0.00071229	8.56	<.0001	0.99180	1.00827
Gender	1	0.01856	0.00172	10.78	<.0001	0.98633	1.01386
Age Group	1	0.01299	0.00109	11.92	<.0001	0.97039	1.03051
SD Resident	1	0.01432	0.00186	7.68	<.0001	0.99767	1.00234
Searched	1	0.00660	0.00142	4.65	<.0001	0.93823	1.06584
Consent	1	0.01178	0.00228	5.16	<.0001	0.29037	3.44390
Contraband	1	0.05033	0.00321	15.70	<.0001	0.14151	7.06640
Valid Search	1	-0.07219	0.00273	-26.41	<.0001	0.22739	4.39778
Time of Day	1	-0.06227	0.00070123	-88.80	<.0001	0.96229	1.03918

Table 3(e). Regression Parameter Estimates with Variance Inflation Factor (VIF)

Variable	DF	Param Est.	Std. Err.	t Value	Pr > |t|	Tolerance	VIF
Intercept	1	0.70875	0.00539	131.49	<.0001		0
Race	1	0.00607	0.00071230	8.52	<.0001	0.99184	1.00823
Gender	1	0.01878	0.00172	10.91	<.0001	0.98693	1.01324
Age Group	1	0.01309	0.00109	12.01	<.0001	0.97071	1.03017
SD Resident	1	0.01418	0.00186	7.61	<.0001	0.99788	1.00212
Searched	1	0.00688	0.00142	4.85	<.0001	0.93960	1.06428
Consent	-	-	-	-	-	-	-
Contraband	1	0.06047	0.00253	23.87	<.0001	0.22673	4.41048
Valid Search	1	-0.07279	0.00273	-26.66	<.0001	0.22780	4.38975
Time of Day	1	-0.06235	0.00070109	-88.93	<.0001	0.96275	1.03869

As may be seen in Table 3(f), with the consent and contraband variables removed, no variable has a VIF above 1.06, which is well below the 3.0 to 4.0 concern levels and the 10.0 maximum. Searched has the highest VIF at 1.06, followed by search validity at 1.05, time of day at 1.04, and age group at 1.03. Given composition of search and search validity, as well as the potential linear association previously identified between age group and time-of-day, these variables will need to be monitored in the final regression to determine if they are resulting in interaction issues. From this test, however, since no VIF exceed the threshold or is significantly higher than others, none of the variables appears to be a linear combination of other, independent, variables.

Table 3(f). Regression Parameter Estimates with Variance Inflation Factor

Variable	DF	Param Est.	Std. Err.	t Value	Pr > |t|	Tolerance	VIF
Intercept	1	0.71604	0.00539	132.95	<.0001		0
Race	1	0.00611	0.00071289	8.57	<.0001	0.99184	1.00822
Gender	1	0.01868	0.00172	10.84	<.0001	0.98694	1.01323
Age Group	1	0.01295	0.00109	11.88	<.0001	0.97074	1.03014
SD Resident	1	0.01447	0.00187	7.76	<.0001	0.99793	1.00208
Searched	1	0.00979	0.00142	6.92	<.0001	0.94655	1.05647
Consent	-	-	-	-	-	-	-
Contraband	-	-	-	-	-	-	-
Valid Search	1	-0.01597	0.00134	-11.92	<.0001	0.94829	1.05453
Time of Day	1	-0.06217	0.00070164	-88.61	<.0001	0.96285	1.03858

INTERACTION CONCLUSION

Based upon the Wald chi-square statistics and VIF results, consent and contraband are removed from the final model. Additionally, these tests indicated potential interaction between a subject being searched and the validity of searches, as well as the age group of the subject and the time-of-day of the vehicle stop. As it is reviewed, the final regression model will need to be closely analyzed to determine if any of these variables are resulting in potentially-erroneous results and warrant removal from the model.

Full Regression Analysis

INITIAL MODEL

Table 4(a) provides the logistic regression analysis comparing the adjusted odds of police action when compared to race after controlling for gender, age group, San Diego residency, whether a search was conducted, whether the search was valid, and the generalized time of day. Black subjects were 1.5 times as likely to have a vehicle stop for moving/equipment violation culminate in arrest than White subjects (AOR=1.50; 95% CI=1.47-1.54). Based upon the confidence interval (CI), this difference was statistically significant. Hispanic subjects were at slightly decreased risk (AOR=0.92; 95% CI=0.91-0.94) to have a traffic stop culminate in police action than White subjects; again, the CI showed this to be statistically significant. Other races were also identified to have a slightly decreased risk of police action; however, the CI showed this to not be a statistically-significant difference (AOR=0.98; 95% CI=0.96-1.00).

Within the classified data, men show as 1.1 times as likely to have a vehicle stop culminate in police action than women (AOR=1.08; 95% CI=1.06-1.09). While not the expected results, this may align more closely with overall statistics regarding gender and crime than be suggestive of any police bias; the CI indicated the results to be statistically significant. Adolescents, middle-aged adults, and later adults, were 1.1 (AOR=1.14; 95% CI=1.08-1.19), 1.1 (AOR=1.05; 95% CI=1.04-1.07), and 1.3 (AOR=1.31; 95% CI=1.26-1.35), times as likely to have a traffic stop culminate in police action than young adults; the CI indicated the difference to be statistically significant. Unsurprisingly, San Diego residents are at higher risk (AOR=1.08; 95% CI=1.07-1.11) of a traffic stop ending in police action than non-residents; the CI range is indicative of a statistically-significant difference. Risk of traffic stops culminating in police action increase from morning to night, with afternoon stops being 1.1 times as likely to culminate in police action (AOR=1.05; 95% CI=1.02-1.06), evening stops 1.8 times as likely (AOR=1.82; 95% CI=1.78-1.86), and night stops more than twice as likely (AOR=2.11; 95% CI=2.07-2.15); all figures represent statistically-significant differences based upon the CI. Traffic stops with a valid search were more than 33 times as likely (AOR=33.02; 95% CI=27.96-38.99) to result in police action than those without a search; the figure is statistically significant based upon the CI.

While interaction was not indicated, the full regression provided unexpected results indicative of a modeling error with regards to whether or not a search was conducted and whether the search was valid. Traffic stops without a search showed 29 (AOR=28.98; 95% CI=24.60-34.14) times as likely to result in police action than those with a search, while invalid searches were more than 33 times as likely (AOR=33.02; 95% CI=27.96-38.99) to result in police action. Both figures appeared to be statistically-significant differences based upon the CI. Reasonably, this relationship does not make sense, since police action (i.e. arrest) necessitates a search, and a valid search should result in police action. Therefore, a revision to the model to exclude the search parameters is indicated.

Ignoring the validity issue, the predictive capabilities of this model are fairly good based upon the c-statistic result of 0.62, although additional data points to provide an improved c-statistic would be ideal. The Hosmer and Lemeshow Goodness-of-Fit Test (HL-Test) provides a *p*-value below 0.0001, meaning the null hypothesis that the model is a good fit is rejected (i.e. invalid model); however, this is most likely being influence improperly by the number of observations, so the c-statistic is instead used to evaluate the model.

Table 4(a). Logistic Regression Analysis Comparing the Adjusted Odds by Variable

	Warning / Other		**Citation / Arrest**			
	***n* (%)**		***n* (%)**		**AOR[A]**	**95% CI[B]**
Variable	**(n=127,536)**		**(n=213,357)**			
Race						
White (1)	53,568	(42.0%)	93,550	(43.9%)	1.00	--
Black (2)	18,212	(14.3%)	20,148	(9.4%)	1.50	1.47,1.54
Hispanic (3)	35,752	(28.0%)	65,876	(30.9%)	0.92	0.91,0.94
Other (4)	20,004	(15.7%)	33,783	(15.8%)	0.98	0.96,1.00
Gender						
Male (1)	85,052	(66.7%)	135,415	(63.5%)	1.08	1.06,1.09
Female (2)	42,484	(33.3%)	77,942	(36.5%)	1.00	--
Age Group						
Adolescent 13-18 (1)	3,371	(2.6%)	4,701	(2.2%)	1.14	1.08,1.19
Young Adult 19-39 (2)	74,197	(58.2%)	121,333	(56.9%)	1.00	--
Middle Age 40-64 (3)	41,739	(32.7%)	69,602	(32.6%)	1.05	1.04,1.07
Later Adult ≥65 (4)	6,104	(4.8%)	8,725	(4.1%)	1.31	1.26,1.35
Unavailable (5)	2,125	(1.7%)	8,996	(4.2%)	0.45	0.43,0.48
San Diego Resident						
Yes (1)	93,026	(72.9%)	159,019	(74.5%)	1.09	1.07,1.11
No (0)	34,510	(27.1%)	54,338	(25.5%)	1.00	--
Search Conducted						
Yes (1)	9,633	(7.6%)	16,742	(7.9%)	1.00	--
No (0)	116,318	(91.2%)	194,187	(91.0%)	28.98	24.60,34.14
Indeterminate (5)	1,585	(1.2%)	2,428	(1.1%)	0.52	0.47,0.58
Valid Search Conducted						
Yes (1)	148	(0.1%)	5,517	(2.6%)	1.00	--
No (0)	8,171	(6.4%)	10,314	(4.8%)	33.02	27.96,38.99
Indeterminate (5)	2,899	(2.3%)	3,339	(1.6%)	56.11	46.64,67.50
Generalized Time of Day						
Morning 05:00 - 11:59 (1)	37,129	(29.1%)	84,081	(39.4%)	1.00	--
Afternoon 12:00 - 16:59 (2)	27,000	(21.2%)	58,126	(27.2%)	1.04	1.02,1.06
Evening 17:00 - 20:59 (3)	22,559	(17.7%)	27,454	(12.9%)	1.82	1.78,1.86
Night 21:00 - 04:59 (4)	40,848	(32.0%)	43,696	(20.5%)	2.11	2.07,2.15
A Adjusted *Odds Ratio*						
B *95% confidence intervals are for reported odds ratios.*						

FINAL MODEL

Based upon the results of the Initial Model, the Final Model was created where the independent variable, police action at a moving/equipment violation traffic stop, is derived from the linear relationship with race, gender, age group, San Diego residency, and time of day. The logistic regression analysis comparing the adjusted odds of police action when compared to race after controlling for the specified variables is provided in Table 4(b). Based upon the c-statistic of 0.61, this model has fairly good predictive capabilities; additional data points are certainly necessary to improve the model, however. The Hosmer and Lemeshow Goodness-of-Fit Test (HL-Test) provides a *p*-value below 0.0001, meaning the null hypothesis that the model is a good fit is rejected (i.e. invalid model); however, this is most likely being influence improperly by the number of observations, so the c-statistic is instead used to evaluate the model.

While the model does not provide statistically-significant information for subjects racially classified as Other (not White, Black, or Hispanic, for purposes of this model; AOR=0.99; 95% CI=0.99-1.01; *p*-value=0.25), all other results do represent statistically-significant differences based upon the CI. As shown in Table 4(b), subjects racially classified as Black are almost 1.5 times as likely (AOR=1.48; 95% CI=1.44-1.51) to have a traffic stop culminate in police action than subjects racially classified as White. Subjects classified as Hispanic are at decreased likelihood of police action (AOR=0.92; 95% CI=0.90-0.93). Men are at slightly increased odds than women to be subject to police action (AOR=1.07; 95% CI=1.05-1.08). Adolescents, middle-aged adults, and later-aged adults, are all more likely to have police action than young adults (Adolescent AOR=1.14; 95% CI=1.09-1.19 | Middle-Aged AOR=1.06; 95% CI=1.04-1.08 | Later-Aged AOR=1.33; 95% CI=1.28-1.37). San Diego residents are 1.1 times as likely to have police action than non-residents (AOR=1.10; 95% CI=1.08-1.12). Finally, the odds of having police action increase from Morning to Night, with afternoon stops at slightly increased risk (AOR=1.04; 95% CI=1.02-1.06), evening stops at 1.8 times likelihood (AOR=1.80; 95% CI=1.77-1.84), and night stops more than double the likelihood of police action (AOR=2.05; 95% CI=2.01-2.09).

Table 4(b). Logistic Regression Analysis Comparing the Adjusted Odds by Variable

	Warning / Other		**Citation / Arrest**			
	***n* (%)**		***n* (%)**		**AOR[A]**	**95% CI[B]**
Variable	**(n=127,536)**		**(n=213,357)**			
Race						
White (1)	53,568	(42.0%)	93,550	(43.9%)	1.00	--
Black (2)	18,212	(14.3%)	20,148	(9.4%)	1.48	1.44,1.51
Hispanic (3)	35,752	(28.0%)	65,876	(30.9%)	0.92	0.90,0.93
Other (4)	20,004	(15.7%)	33,783	(15.8%)	0.99	0.97,1.01
Gender						
Male (1)	85,052	(66.7%)	135,415	(63.5%)	1.07	1.05,1.08
Female (2)	42,484	(33.3%)	77,942	(36.5%)	1.00	--
Age Group						
Adolescent 13-18 (1)	3,371	(2.6%)	4,701	(2.2%)	1.14	1.09,1.19
Young Adult 19-39 (2)	74,197	(58.2%)	121,333	(56.9%)	1.00	--
Middle Age 40-64 (3)	41,739	(32.7%)	69,602	(32.6%)	1.06	1.04,1.08
Later Adult ≥65 (4)	6,104	(4.8%)	8,725	(4.1%)	1.33	1.28,1.37
Unavailable (5)	2,125	(1.7%)	8,996	(4.2%)	0.46	0.44,0.48
San Diego Resident						
Yes (1)	93,026	(72.9%)	159,019	(74.5%)	1.10	1.08,1.12
No (0)	34,510	(27.1%)	54,338	(25.5%)	1.00	--
Generalized Time of Day						
Morning 05:00 - 11:59 (1)	37,129	(29.1%)	84,081	(39.4%)	1.00	--
Afternoon 12:00 - 16:59 (2)	27,000	(21.2%)	58,126	(27.2%)	1.04	1.02,1.06
Evening 17:00 - 20:59 (3)	22,559	(17.7%)	27,454	(12.9%)	1.80	1.77,1.84
Night 21:00 - 04:59 (4)	40,848	(32.0%)	43,696	(20.5%)	2.05	2.01,2.09
A Adjusted *Odds Ratio*						
B *95% confidence intervals are for reported odds ratios.*						

Confounding Analysis

Table 5, below, shows the change in the parameter estimate for the police-action variable as other, independent, variables are removed. The removal of the time-of-day variable provides the highest change estimate 6.4% for the comparison of Black subjects to White, 1.5% for Hispanic subjects to White, and 4.6% for Other subjects to White. All other changes are less than 1.0%. Based on this information, none of the variables within the model are viewed as confounders.

Table 5. Confounding Analysis of Point Estimates After Variable Removals

Model	Point Estimate	Change from Overall	c-Statistic
Measure Overall:	1.48, 0.92, 0.99	-	0.61
w/o Race	*	-	-
***w/o* Gender**	1.49, 0.92, 0.99	0.5%, 0.3%, 0.2%	0.61
***w/o* Age Group**	1.47, 0.92, 0.98	0.7%, 0.3%, 0.5%	0.61
***w/o* San Diego Resident**	1.46, 0.91, 0.98	1.0%, 0.3%, 0.4%	0.60
***w/o* Time of Day**	1.57, 0.93, 1.03	6.4%, 1.5%, 4.6%	0.55

* Not removed; this is the independent variable of interest.

Results of Statistical Analysis

SUMMARY

The dependent variable, police action (citation, property seizure, or arrest, at a traffic stop for moving/equipment violation) analyzed in the study was found to have a statistically-significant association with race. Univariate analyses were conducted using Pearson's chi-square test for independence in SAS to compare that association independent of gender, age group, San Diego residency, and time of day; search-related variables were removed due to interactions with other, independent, variables. Results of this analysis indicated statistically significant associations between race and each of those variables, as well as between police action, race, gender, age group, San Diego residency, and time of day. Through logistic-regression analysis, which was classified with fairly-good predictive capabilities (c=0.61), the model identified statistically significant increases in risk of police action for subjects racially identified as Black over those identified as White. Increased risks were also identified after controlling for other variables for men; adolescents, middle-aged adults, and later-adults; and San Diego residents. It was also identified that the risk of police action increases from morning to night, after controlling for other variables. Removal of all factors except race and time-of-day classifications resulted in a fairly-good model with c-statistic of 0.60 to 0.61. Removal of the time-of-day classification resulted in a poor model with c-statistic of 0.55.

STRENGTHS & LIMITATIONS

Based upon the c-statistic, this model has fairly-good predictive capabilities. The large number of observations were available for review and limited data issues were identified. In some instances where data was missing, secondary table data was utilized to improve the available data. However, this dataset is limited in the fields defining a police interaction, and it is only the traffic-related subset of police interactions. There is no information regarding non-traffic interactions, the outcome of judicial action, or greater insight into the parameters of each interaction. Identifying any potential racial bias would necessitate a broader data review to truly identify any potential issues.

Conclusion

This model contains statistically-significant variables, does not include confounders or have issues with multicollinearity, and has a fairly-good c-statistics. It indicates the statistical association between race and police actions at traffic stops, resulting in increase risk of action for individuals racially identified as Black over White. While additional data is necessary to provide a stronger model and gain greater insight into each incident, this does provide a troubling statistic. Whether associated with negative bias towards Black members of the community, positive bias towards other races, or an underlying cultural disconnect and inherent inequality, such statistics highlight the need for additional analysis.

"The American Anthropological Association recommends the elimination of the term 'race'…as [it] has been scientifically proven to not be a real, natural phenomenon… Yet the concept of race has become thoroughly – and perniciously – woven into the cultural and political fabric of the United States… [T]hese classifications must be transcended and replaced by more non-racist and accurate ways of representing the diversity of the U.S. population."[2] As one of the safest metropolitan areas in the United States, with a police department focused on neighborhood partnerships rather than reactionary policing, gaining additional insight into these statistics and real-time monitoring of associated risk trends, could identify outreach and training opportunities that could fundamentally change racial dynamics within the city.

Appendix A: SAS Code

1. DB IMPORT

```
/*       --------------------------------------------------------------------
         NATIONAL UNIVERSITY
         ANA625: Categorical Data Methods, Appl (February 2018)
         Matthew C. Vanderbilt

         OBJECTIVE:
         Import database of San Diego Police Department Vehicle Stop data.
         -------------------------------------------------------------------- */

/*       --------------------------------------------------------------------
         By task:     Import Data Wizard

         Source file:
         C:\Users\rdy2d\OneDrive\Documents\Education\Matthew\National
         University\ANA625\Project\ANA625.accdb
         Server:      Local File System

         Output data: WORK.ANA625
         Server:      Local

         PROC IMPORT reads the data directly from the Microsoft Access
         database.
         -------------------------------------------------------------------- */
LIBNAME ANA625 'C:\Users\rdy2d\OneDrive\Documents\Education\Matthew\National University\ANA625\Project';

PROC IMPORT
        TABLE="combinedVehicleStops_FINAL"
        OUT=ANA625.RawData
        REPLACE
        DBMS=ACCESS;
    DATABASE="C:\Users\rdy2d\OneDrive\Documents\Education\Matthew\National University\ANA625\Project\ANA625.accdb";
RUN;

/*       --------------------------------------------------------------------
         PROC DATASETS modifies the attributes of the columns within the
         output data set.
         -------------------------------------------------------------------- */
PROC DATASETS LIBRARY=ANA625 NOLIST;
    MODIFY RawData;
        FORMAT
            Index            $CHAR26.
            stop_id          BEST12.
            stop_cause       $CHAR38.
            service_area     $CHAR7.
            subject_race     $CHAR1.
            subject_sex      $CHAR1.
            subject_age      $CHAR6.
            timestamp        $CHAR19.
            stop_date        $CHAR10.
            stop_time        $CHAR255.
            sd_resident      $CHAR1.
            arrested         $CHAR1.
            searched         $CHAR1.
            obtained_consent $CHAR1.
            contraband_found $CHAR1.
            property_seized  $CHAR1.
            search_details_type_flat $CHAR72.
            search_details_description_flat $CHAR113.
            stop_cause_clean $CHAR38.
            age_val          BEST12.
            year_val         BEST12.
            timestamp_val    DATETIME21.2
            race             $CHAR16. ;
        INFORMAT
            Index            $CHAR26.
            stop_id          BEST12.
            stop_cause       $CHAR38.
            service_area     $CHAR7.
            subject_race     $CHAR1.
            subject_sex      $CHAR1.
            subject_age      $CHAR6.
            timestamp        $CHAR19.
            stop_date        $CHAR10.
            stop_time        $CHAR255.
            sd_resident      $CHAR1.
            arrested         $CHAR1.
            searched         $CHAR1.
            obtained_consent $CHAR1.
            contraband_found $CHAR1.
            property_seized  $CHAR1.
            search_details_type_flat $CHAR72.
            search_details_description_flat $CHAR113.
            stop_cause_clean $CHAR38.
            age_val          BEST12.
            year_val         BEST12.
            timestamp_val    DATETIME21.
            race             $CHAR16. ;
RUN;
```

2. CREATE POPULATION

```
/*       --------------------------------------------------------------------
         NATIONAL UNIVERSITY
         ANA625: Categorical Data Methods, Appl (February 2018)
         Matthew C. Vanderbilt

         OBJECTIVE:
         Create analytical population with derived fields for analysis
         -------------------------------------------------------------------- */

TITLE '-- SAMPLE TABLE CONTENTS';
PROC CONTENTS DATA=ANA625.RawData;
RUN;

/*       Creation of Population Sub-Sample */
DATA ANA625.Population (KEEP=stop_id
                                             stop_cause
                                             service_area
                                             subject_race
                                             subject_sex
                                             subject_age
                                             sd_resident
                                             arrested
                                             searched
                                             obtained_consent
                                             contraband_found
                                             property_seized
                                             search_details_type_flat
                              search_details_description_flat
                              stop_cause_clean
                                             age_val
                                             year_val
                                             timestamp_val
                                             race
                                             race_val
                                             gender_val
                                             age_group_val
                                             resident_val
                                             arrested_val
                                             searched_val
                                             obtained_consent_val
                                             contraband_found_val
                                             property_seized_val
                                             warning_val
                                             citation_val
                                             valid_search
                                             outcome
                                             policeaction
                                             time_of_day);
         SET ANA625.RawData        (WHERE=(stop_cause IN('Moving Violation','Equipment Violation')
                                                      ));

         LABEL stop_id = 'unique stop identifier';
         LABEL stop_cause = 'reason for the stop';
         LABEL service_area = 'police service area';
         LABEL subject_race = 'race code';
         LABEL subject_sex = 'sex code (M,F)';
         LABEL subject_age = 'age';
         LABEL timestamp = 'ISO8601 timestamp';
         LABEL stop_date = 'date (mm/dd/yy)';
         LABEL stop_time = 'time (24hrs format)';
         LABEL sd_resident = 'if subject is a resident of the City of San Diego (Y|N)';
         LABEL arrested = 'if subject was arrested (Y|N)';
         LABEL searched = 'if a search was conducted (Y|N)';
         LABEL obtained_consent = 'if a search was conducted, if consent was obtained (Y|N)';
         LABEL contraband_found = 'if a search was conducted, if contraband was found (Y|N)';
         LABEL property_seized = 'if a search was conducted, if property was seized (Y|N)';
         LABEL search_details_type_flat = 'type of search details record (Action
taken|actionTakenOther|SearchBasis|SearchBasisOther|SearchType)';
         LABEL search_details_description_flat = 'search details description';
         LABEL stop_cause_clean = 'reason for the stop normalized for key data';
         LABEL age_val = 'numeric subject_age';
         LABEL year_val = 'numeric stop year';
         LABEL timestamp_val = '[stop_date]&[stop_time]';
         LABEL race = 'Vehicle Stop Race Code Description';
         LABEL race_val = 'race (1=White|2=Black|3=Hispanic|4=Other)';
```

```
        LABEL gender_val = 'gender (1=Male|2=Female|3=Other)';
        LABEL age_group_val = 'rn.com primary age group';
        LABEL resident_val = 'San Diego resident (1=Yes|0=No)';
        LABEL arrested_val = 'arrested (1=Yes|0=No)';
        LABEL searched_val = 'search appears to have been conducted (1=Yes|0=No)';
        LABEL obtained_consent_val = 'assumed consent (1=Yes|0=No/BLANK)';
        LABEL contraband_found_val = 'contraband appears to have been found (1=Yes|0=No)';
        LABEL property_seized_val = 'property appears to hvae been seized (1=Yes|0=No)';
        LABEL citation_val = 'citation issued (1=Yes|0=No)';
        LABEL warning_val = 'warning issued (1=Yes|0=No)';
        LABEL valid_search = 'search with contraband or seized property (1=Yes|0=No)';
        LABEL outcome = 'Highest Outcome: 1=Arrest, 2=Citation, 3=Prop Seized, 4=Warning, 5=Other';
        LABEL policeaction = 'Arrest or Citation (1=Yes|0=No)';
        LABEL time_of_day = '1:5<=Morning<12;2:12<=Afternoon<17;3:17<=Evening<21;4:(Night>=21 | Night<5)';

        IF subject_race IN(' ','') OR MISSING(subject_race) THEN subject_race='5';
        IF subject_sex IN(' ','') OR MISSING(subject_sex) THEN subject_sex = '5';
        IF age_val IN(' ','') OR MISSING(age_val) THEN age_val = 5;
        IF sd_resident IN(' ','') OR MISSING(sd_resident) THEN sd_resident = '5';
        IF arrested IN(' ','') OR MISSING(arrested) THEN arrested = '5';
        IF searched IN(' ','') OR MISSING(searched) THEN searched = '5';
        IF obtained_consent IN(' ','') OR MISSING(obtained_consent) THEN obtained_consent = '5';
        IF contraband_found IN(' ','') OR MISSING(contraband_found) THEN contraband_found = '5';
        IF property_seized IN(' ','') OR MISSING(property_seized) THEN property_seized = '5';
        IF race IN(' ','') OR MISSING(race) THEN race = 'NOT RECORDED';

        race_val = 4;
        IF UPCASE(subject_race) = 'W' THEN race_val = 1;
        IF UPCASE(subject_race) = 'B' THEN race_val = 2;
        IF UPCASE(subject_race) = 'H' THEN race_val = 3;

        gender_val = 5;
        IF UPCASE(subject_sex) = 'M' THEN gender_val = 1;
        IF UPCASE(subject_sex) = 'F' THEN gender_val = 2;

        age_group_val = 5;
        IF age_val >= 13 AND age_val <= 18 THEN age_group_val = 1; *adolescent;
        IF age_val >  18 AND age_val <  40 THEN age_group_val = 2; *young adult;
        IF age_val >= 40 AND age_val <  65 THEN age_group_val = 3; *middle aged;
        IF age_val >= 65 AND age_val < 111 THEN age_group_val = 4; *later adult;

        resident_val = 5;
        IF UPCASE(sd_resident) = 'N' THEN resident_val = 0;
        IF UPCASE(sd_resident) = 'Y' THEN resident_val = 1;

        arrested_val = 5;
        IF UPCASE(arrested) = 'N' THEN arrested_val = 0;
        IF UPCASE(arrested) = 'Y' THEN arrested_val = 1;
        IF UPCASE(arrested) NOT IN('Y','N') AND FIND(search_details_description_flat,'Arrest','i') > 0 THEN
arrested_val = 1;

        searched_val = 5;
        IF UPCASE(searched) = 'N' THEN searched_val = 0;
        IF UPCASE(searched) = 'Y' THEN sarched_val = 1;
        IF obtained_consent IN('Y','N')
                OR contraband_found IN('Y','N')
                OR property_seized IN('Y','N')
                OR FIND(search_details_type_flat,'Search','i') > 0
                THEN searched_val = 1;

        obtained_consent_val = 5;
        IF searched_val = 0 THEN obtained_consent_val = 2;
        IF UPCASE(obtained_consent) = 'N' THEN obtained_consent_val = 0;
        IF UPCASE(obtained_consent) = 'Y' THEN obtained_consent_val = 1;
        IF UPCASE(obtained_consent) NOT IN('Y','N') AND (FIND(search_details_description_flat,'Consent','i') > 0

        OR FIND(search_details_description_flat,'4th Waiver','i') > 0)

        THEN obtained_consent_val = 1;

        contraband_found_val = 5;
        IF searched_val = 0 THEN contraband_found_val = 2;
        IF UPCASE(contraband_found) = 'N' THEN contraband_found_val = 0;
        IF UPCASE(contraband_found) = 'Y' THEN contraband_found_val = 1;
        IF UPCASE(contraband_found) NOT IN('Y','N') AND FIND(search_details_description_flat,'Contraband
Visible','i') > 0 THEN contraband_found_val = 1;
```

```
         property_seized_val = 5;
         IF searched_val = 0 THEN property_seized_val = 2;
         IF UPCASE(property_seized) = 'N' THEN property_seized_val = 0;
         IF UPCASE(property_seized) = 'Y' THEN property_seized_val = 1;
         IF UPCASE(property_seized) NOT IN('Y','N') AND (FIND(search_details_description_flat,'Tow','i') > 0

         OR FIND(search_details_description_flat,'Impound','i') > 0)

         THEN property_seized_val = 1;

         citation_val = 0; * default - no citation;
         IF FIND(search_details_description_flat,'Citation','i') > 0 THEN citation_val=1;
         IF FIND(search_details_description_flat,'Ticket','i') > 0 THEN citation_val=1;

         warning_val = 0; * default - no warning;
         IF (FIND(search_details_description_flat,'Warning','i') > 0 AND citation_val IN(0,2,5) AND arrested_val
IN(0,5) AND property_seized_val IN(0,2,5)) THEN warning_val=1;

         valid_search = 5; * default - not a valid search;
         IF searched_val = 0 THEN valid_search = 2;
         IF searched_val = 1 AND (arrested_val = 1 OR contraband_found_val = 1 OR property_seized_val = 1) THEN
valid_search = 1;
         IF searched_val = 1 AND (arrested_val = 0 AND contraband_found_val = 0 AND property_seized_val = 0) THEN
valid_search = 0;

         outcome = 5; * default - no action / no data;
         IF warning_val = 1 THEN outcome = 4;
         IF property_seized_val = 1 THEN outcome = 3;
         IF citation_val = 1 THEN outcome = 2;
         IF arrested_val = 1 THEN outcome = 1;

         policeaction = 5;
         IF outcome IN(1,2,3) THEN policeaction = 1;
         IF outcome IN(4,5) THEN policeaction = 0;

         time_of_day = 0;
         IF HOUR(timestamp_val) >=  5 AND HOUR(timestamp_val) < 12 THEN time_of_day = 1;
         IF HOUR(timestamp_val) >= 12 AND HOUR(timestamp_val) < 17 THEN time_of_day = 2;
         IF HOUR(timestamp_val) >= 17 AND HOUR(timestamp_val) < 21 THEN time_of_day = 3;
         IF HOUR(timestamp_val) >= 21 OR  HOUR(timestamp_val) <  5 THEN time_of_day = 4;

RUN;

TITLE '-- SAMPLE TABLE CONTENTS';
PROC CONTENTS DATA=ANA625.Population;
RUN;

TITLE '-- CHECK RECODING OF VARIABLES';
PROC FREQ DATA=ANA625.Population;
         TABLES race*race_val subject_sex*gender_val sd_resident*resident_val arrested*arrested_val
searched*searched_val obtained_consent*obtained_consent_val contraband_found*contraband_found_val
property_seized*property_seized_val outcome*policeaction;

RUN;
```

3. CREATE SAMPLE

```
/*       -----------------------------------------------------------------------
         NATIONAL UNIVERSITY
         ANA625: Categorical Data Methods, Appl (February 2018)
         Matthew C. Vanderbilt

         OBJECTIVE:
         Create analytical sample for study
         ----------------------------------------------------------------------- */

/*       Creation of Population Sub-Sample */
DATA ANA625.Sample  (KEEP=stop_id
                                               stop_cause
                                               service_area
                                               subject_race
                                               subject_sex
                                               subject_age
                                               sd_resident
                                               arrested
                                               searched
                                               obtained_consent
                                               contraband_found
                                               property_seized
                                               search_details_type_flat
                               search_details_description_flat
                               stop_cause_clean
                                               age_val
                                               year_val
                                               timestamp_val
                                               race
                                               race_val
                                               gender_val
                                               age_group_val
                                               resident_val
                                               arrested_val
                                               searched_val
                                               obtained_consent_val
                                               contraband_found_val
                                               property_seized_val
                                               warning_val
                                               citation_val
                                               valid_search
                                               outcome
                                               policeaction
                                               time_of_day);
         SET ANA625.Population (WHERE=(      subject_race ^= '5'
                                                                                     AND race NOT IN('NOT RECORDED')
                                                                                     AND gender_val IN(1,2) /*data
dictionary indicates field is only binary*/
                                                                                     AND resident_val IN(0,1)
                                                                                     AND policeaction IN(0,1)
                                                                                     AND time_of_day IN(1,2,3,4)
                                               ));

RUN;

TITLE '-- SAMPLE TABLE CONTENTS';
PROC CONTENTS DATA=ANA625.Sample;

RUN;
```

4. TABLES 1 & 2

```
/*      --------------------------------------------------------------------
        NATIONAL UNIVERSITY
        ANA625: Categorical Data Methods, Appl (February 2018)
        Matthew C. Vanderbilt

        OBJECTIVE:
        Characteristics of modeled sample.
        -------------------------------------------------------------------- */

TITLE '-- TABLE 1: Descriptive & Bivariate Statistics';
PROC FREQ DATA=ANA625.Sample;
        TABLES (gender_val age_group_val resident_val searched_val obtained_consent_val contraband_found_val
valid_search time_of_day)*race_val / CHISQ;
RUN;

TITLE '-- TABLE 1: Descriptive & Bivariate Statistics (Subset)';
PROC FREQ DATA=ANA625.Sample (WHERE=(searched_val = 1));
        TABLES (obtained_consent_val contraband_found_val valid_search)*race_val / CHISQ; *REMOVED valid_search;
RUN;

TITLE '-- TABLE 2: Descriptive & Univariate Statistics';
PROC FREQ DATA=ANA625.Sample;
        TABLES (race_val gender_val age_group_val resident_val searched_val obtained_consent_val
contraband_found_val valid_search time_of_day)*policeaction / CHISQ;

RUN;
```

5. INTERACTIONS

```
/*      -------------------------------------------------------------------
        NATIONAL UNIVERSITY
        ANA625: Categorical Data Methods, Appl (February 2018)
        Matthew C. Vanderbilt

        OBJECTIVE:
        Testing of interactions between model variables.
        ------------------------------------------------------------------- */

TITLE '-- VARIABLE ASSOCIATION: Testing Interactions with Search';
PROC FREQ DATA=ANA625.Sample;
        TABLES (obtained_consent_val contraband_found_val valid_search)*searched_val / CHISQ;
RUN;

TITLE '-- VARIABLE ASSOCIATION: Testing Interactions with Search';
PROC FREQ DATA=ANA625.Sample;
        TABLES (age_group_val)*time_of_day / CHISQ;
RUN;

TITLE '-- MULTIVARIABLE REGRESSION: Testing Interactions with Search';
PROC LOGISTIC DATA=ANA625.Sample;
        CLASS policeaction(REF='0') race_val(REF='1') gender_val(REF='1') age_group_val(REF='2')
resident_val(REF='1') searched_val(REF='0') obtained_consent_val(REF='0') contraband_found_val(REF='0')
valid_search(REF='0') time_of_day(REF='2') / PARAM=REFERENCE;
        MODEL policeaction =  race_val gender_val age_group_val resident_val searched_val obtained_consent_val
contraband_found_val valid_search searched_val*obtained_consent_val searched_val*contraband_found_val
searched_val*valid_search valid_search*obtained_consent_val valid_search*contraband_found_val time_of_day
age_group_val*time_of_day;
RUN;

TITLE '-- MULTIVARIABLE REGRESSION: Testing Interactions with Search';
PROC LOGISTIC DATA=ANA625.Sample;
        CLASS policeaction(REF='0') race_val(REF='1') gender_val(REF='1') age_group_val(REF='2')
resident_val(REF='1') searched_val(REF='0') obtained_consent_val(REF='0') contraband_found_val(REF='0')
valid_search(REF='0') time_of_day(REF='2') / PARAM=REFERENCE;
        MODEL policeaction =  race_val gender_val resident_val searched_val*obtained_consent_val
searched_val*contraband_found_val searched_val*valid_search valid_search*obtained_consent_val
valid_search*contraband_found_val age_group_val*time_of_day;
RUN;

TITLE '-- MULTICOLLINEARITY INVESTIGATION #1: Testing Interactions';
PROC REG DATA=ANA625.Sample;
        MODEL policeaction = race_val gender_val age_group_val resident_val searched_val obtained_consent_val
contraband_found_val valid_search time_of_day / VIF TOL;
RUN;

TITLE '-- MULTICOLLINEARITY INVESTIGATION #2: Testing Interactions';
PROC REG DATA=ANA625.Sample;
        MODEL policeaction = race_val gender_val age_group_val resident_val searched_val /*obtained_consent_val*/
contraband_found_val valid_search time_of_day / VIF TOL;
RUN;

TITLE '-- MULTICOLLINEARITY INVESTIGATION #3: Testing Interactions';
PROC REG DATA=ANA625.Sample;
        MODEL policeaction = race_val gender_val age_group_val resident_val searched_val /*obtained_consent_val
contraband_found_val*/ valid_search time_of_day / VIF TOL;

RUN;
```

6. REGRESSION

```
/*       --------------------------------------------------------------------
         NATIONAL UNIVERSITY
         ANA625: Categorical Data Methods, Appl (February 2018)
         Matthew C. Vanderbilt

         OBJECTIVE:
         The objective of this analysis is to statistically investigate the
         association of police action at traffic stops and race, after
         controlling for gender, age group, San Diego residency status,
         validity of search, and time of day, within a sample of data from the
         City of San Diego Police Department.  The formula below provides a
         general idea of the model utilized:

                  Police Action (Y|N) =
                           race (White|Black|Hispanic|Other) +
                           gender (Male|Female) +
                           age (1=Adolescent:  13 <= subject_age <= 18 |
                                     2=Young Adult: 18 <  subject_age <  40 |
                                     3=Middle-Aged: 40 <= subject_age <  65 |
                                     4=Later Adult:       subject_age >= 65) +
                           resident (Y|N) +
                           valid search (Y|N)
                           time of day

         Original Population Size: 390,999 observations from 2014 - 2017
         - After Removal of Duplicate Records by [stop_id]: 387,351 (99.1%)
         -------------------------------------------------------------------- */

TITLE '-- TABLE 3: MULTIVARIABLE LOGISTIC REGRESSION';
PROC LOGISTIC DATA=ANA625.Sample;
         CLASS race_val(REF='1') gender_val(REF='2') age_group_val(REF='2') resident_val(REF='1')
searched_val(REF='1') valid_search(REF='1') time_of_day(REF='1') / PARAM=REFERENCE;
         MODEL policeaction = race_val gender_val age_group_val resident_val searched_val valid_search time_of_day /
LACKFIT;
RUN;

TITLE '-- TABLE 3: MULTIVARIABLE LOGISTIC REGRESSION';
PROC LOGISTIC DATA=ANA625.Sample;
         CLASS race_val(REF='1') gender_val(REF='2') age_group_val(REF='2') resident_val(REF='1')
/*searched_val(REF='1')*/ valid_search(REF='1') time_of_day(REF='1') / PARAM=REFERENCE;
         MODEL policeaction = race_val gender_val age_group_val resident_val /*searched_val*/ valid_search
time_of_day / LACKFIT;
RUN;

TITLE '-- TABLE 3: MULTIVARIABLE LOGISTIC REGRESSION';
PROC LOGISTIC DATA=ANA625.Sample;
         CLASS race_val(REF='1') gender_val(REF='2') age_group_val(REF='2') resident_val(REF='1')
/*searched_val(REF='1') valid_search(REF='1')*/ time_of_day(REF='1') / PARAM=REFERENCE;
         MODEL policeaction = race_val gender_val age_group_val resident_val /*searched_val valid_search*/
time_of_day / LACKFIT;
RUN;

/* CONFOUNDING TESTING through Manual Backward Stepwise Regression Analysis */
TITLE '-- CONFOUNDING TEST: MULTIVARIABLE LOGISTIC REGRESSION / remove gender_val';
PROC LOGISTIC DATA=ANA625.Sample;
         CLASS race_val(REF='1') gender_val(REF='2') age_group_val(REF='2') resident_val(REF='1')
/*searched_val(REF='1') valid_search(REF='1')*/ time_of_day(REF='1') / PARAM=REFERENCE;
         MODEL policeaction = race_val /*gender_val*/ age_group_val resident_val /*searched_val valid_search*/
time_of_day / LACKFIT;
RUN;

TITLE '-- CONFOUNDING TEST: MULTIVARIABLE LOGISTIC REGRESSION / remove age_group_val';
PROC LOGISTIC DATA=ANA625.Sample;
         CLASS race_val(REF='1') gender_val(REF='2') age_group_val(REF='2') resident_val(REF='1')
/*searched_val(REF='1') valid_search(REF='1')*/ time_of_day(REF='1') / PARAM=REFERENCE;
         MODEL policeaction = race_val gender_val age_group_val /*resident_val*/ /*searched_val valid_search*/
time_of_day / LACKFIT;
RUN;

TITLE '-- CONFOUNDING TEST: MULTIVARIABLE LOGISTIC REGRESSION / remove resident_val';
PROC LOGISTIC DATA=ANA625.Sample;
         CLASS race_val(REF='1') gender_val(REF='2') age_group_val(REF='2') resident_val(REF='1')
/*searched_val(REF='1') valid_search(REF='1')*/ time_of_day(REF='1') / PARAM=REFERENCE;
         MODEL policeaction = race_val gender_val /*age_group_val*/ resident_val /*searched_val valid_search*/
time_of_day / LACKFIT;
```

```
RUN;

TITLE '-- CONFOUNDING TEST: MULTIVARIABLE LOGISTIC REGRESSION / remove time-of-day';
PROC LOGISTIC DATA=ANA625.Sample;
        CLASS race_val(REF='1') gender_val(REF='2') age_group_val(REF='2') resident_val(REF='1')
/*searched_val(REF='1') valid_search(REF='1')*/ time_of_day(REF='1') / PARAM=REFERENCE;
        MODEL policeaction = race_val gender_val age_group_val resident_val /*searched_val valid_search*/
/*time_of_day*/ / LACKFIT;
RUN;

QUIT;
```

Appendix B: Data Dictionary

RAW DATA

The data dictionary shown here is provided within DataSD – Police Vehicle Stops. It is licensed under the Open Data Commons Public Domain Dedication and License, which provides for sharing (copying, distribution, and use), creation (production of works), and adaptation (modification, transformation, and building-upon). The following is provided as summary description of the data[4]:

- Vehicle stops made by the San Diego Police Department. **Vehicle Stops** files contain all vehicle stops for a given year.
- Field descriptions for this data are available in the Vehicle Stops Dictionary, and race codes are documented in the Vehicle Stops Race Codes file.
- In certain cases a search is conducted following a vehicle stop. Details about these searches are available in the **Vehicle Stops Search Details** files. Field descriptions for this data are available in the Vehicle Stops Search Details Dictionary. The stop outcomes are also listed in the **Vehicle Stops Search Details** files.
- Both **Vehicle Stops** and **Vehicle Stops Search Details** datasets can be joined using the common [stop_id] field. There could be one [or] more than one [search_id] per [stop_id].

Vehicle Stops Dictionary[4]

Variable	**Description**	**Possible Values**
stop_id	unique stop identifier	
stop_cause	reason for the stop	
service_area	police service area	
subject_race	race code	See race code dictionary
subject_sex	sex code	M, F
subject_age	Age	
timestamp	ISO8601 timestamp	
stop_date	date (mm/dd/yy)	
stop_time	time (24hrs format)	
sd_resident	if subject is a resident of the City of San Diego	Y, N
arrested	if subject was arrested	Y, N
searched	if a search was conducted	Y, N
obtained_consent	if a search was conducted, if consent was obtained	Y, N
contraband_found	if a search was conducted, if contraband was found	Y, N
property_seized	if a search was conducted, if property was seized	Y, N

Vehicle Stops Race Codes[4]

Race Code	Description
A	Other Asian
B	Black
C	Chinese
D	Cambodian
F	Filipino
G	Guamanian
H	Hispanic
I	Indian
J	Japanese
K	Korean
L	Laotian
O	Other
P	Pacific Islander
S	Samoan
U	Hawaiian
V	Vietnamese
W	White
Z	Asian Indian

Vehicle Stops Search Details Dictionary[4]

Variable	Description	Possible Values
stop_id	unique stop identifier	
search_details_id	unique search details identifier	
search_details_type	type of search details record	ActionTaken, ActionTakenOther, SearchBasis, SearchBasisOther, SearchType
search_details_description	search details description	see search details description list

Vehicle Stops Search Details Description List[4]

Search Details Description	Description
ActionTaken	Citation, Written Warning, Verbal Warning, Field Interview, other
ActionTakenOther	Additional text if other is selected for ActionTaken
SearchBasis	Contraband visible, canine alert, 4th waiver search, inventory search, observed evidence related to criminal activity, odor of contraband, consent search, search incident to arrest, other
SearchBasisOther	Additional text if other is selected for SearchBasis
SearchType	Vehicle, Driver, Passenger

DERIVED / CALCULATED FIELDS

Variable	Description	Source
search_details_type_flat	all search details record types	populated from [search_details_type]
search_details_description_flat	all search details descriptions	populated from [search_details_description]
stop_cause_clean	normalized reason for the stop	converted from [stop_cause]
age_val	normalized subject age	converted from [subject_age]
year_val	incident year	derived from [stop_date]
timestamp_val	corrected timestamp	converted from [stop_date] and [stop_time]
race	race code description	joined from [subject_race] and race code descriptions
race_val	numeric race where 1 = W, 2 = B, 3 = H, and 4 = *else*	derived from [subject_race]
gender_val	numeric gender where 1 = Male and 2 = Female, and 5 = *else*	derived from [subject_sex]
age_group_val	generalized age groups	derived from [age_val]
resident_val	Boolean residency	derived from [sd_resident]
arrested_val	Boolean subject arrested	derived from [arrested] and [serach_details_description_flat]
searched_val	Boolean subject searched	derived from [searched], [obtained_consent], [contraband_found], [property_seized], and [search_details_type_flat] (Search)
obtained_consent_val	numeric consent-to-search, where 0 = No, 1 = Yes, and 5 = *else*	derived from [obtained_consent] and [search_details_description_flat] (Consent | 4th Waiver)
contraband_found_val	numeric contraband found, where 0 = No, 1 = Yes, and 5 = *else*	Derived from [contraband_found] and [search_details_description_flat] (Contraband Visible)
property_seized_val	numeric property seized, where 0 = No, 1 = Yes, and 5 = *else*	derived from [property_seized] and [search_details_description_flat] (Tow | Impound)
citation_val	Boolean citation issued	derived from [search_details_description_flat] (Citation | Ticket)
warning_val	Boolean warning provided	derived from [search_details_description_flat] (Warning)

valid_search	Boolean search valid (search with arrest, contraband, or property seizure)	derived from [searched_val], [arrested_val], [contraband_found_val], and [property_seized_val]
outcome	Maximum outcome of the incident, where 1 = arrested, 2 = citation, 3 = property seizure, 4 = warning, and 5 = *else*	derived from [warning_val], [property_seizedd_val], [citation_val], and [arrested_val]
policeaction	Boolean police action, defined as a citation, property seizure, or arrest	derived from [outcome]
time_of_day	Generalized time at which stop occurred	derived from [timestamp_val]

References

1. City of San Diego. *Population*. Economic Development [cited 2018 March 4, 2018]; Available from: https://www.sandiego.gov/economic-development/sandiego/population.
2. Statistical Atlas. *Race and Ethnicity in San Diego, California (City)*. Wikipedia: The Free Encyclopedia [cited 2018 March 4, 2018]; Available from: https://statisticalatlas.com/place/California/San-Diego/Race-and-Ethnicity.
3. *San Diego Police Department*. Wikipedia: The Free Encyclopedia [cited 2018 March 4, 2018]; Available from: https://en.wikipedia.org/wiki/San_Diego_Police_Department.
4. San Diego Police Department, *Police Vehicle Stops*, City of San Diego, Editor., DataSD: San Diego, CA.