Introduction: San Diego State about traffic

This report presents a comprehensive analysis of traffic stop data spanning from December 2013 to March 2017. The dataset consists of 383,027 records and 21 variables, encompassing a range of attributes, including the date, driver demographics, traffic issues, and regulatory compliance. One key regulation is that drivers must be at least 16 years old, highlighting the importance of data cleaning as a crucial step in preparing the dataset for analysis. Additionally, this report explores various visualizations, hypotheses, and a predictive model from different perspectives to uncover meaningful insights that can inform effective traffic management strategies in San Diego.

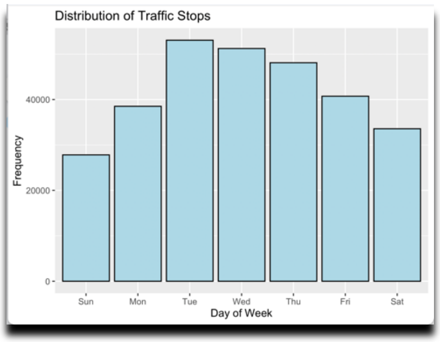
The Valuables of Interest

1. Date
2. Service_area
3. Subject_age
4. Subject_race
5. Subject_sex
6. Type
7. Arrest_made
8. Citation_issued
9. Warning_issued
10. Outcome
11. Search_person
12. Reason_for_stop

The Data Cleaning

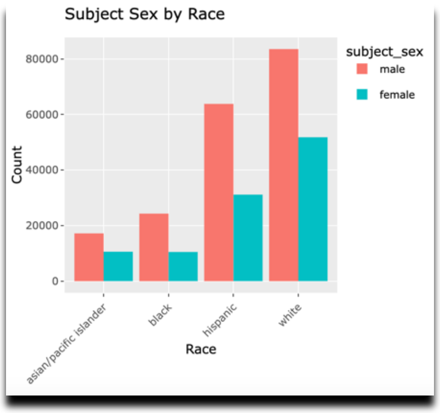
1. Select only interested variables, you can see 21 variables at first. After selecting, there are only 12 variables. This makes data looks more clear and easy to use.
2. We can also change data type with function such as “as.Date”, “factor”, “as.logical”, “as.numeric” to manipulate appropriately.
3. Create a new variable, “dayofweek”, with function “mutate” to see the day trend of traffic stop. Creating new variable is useful when we need to see other dimensions of this dataset. It could help us find something new.
4. Create the aggregated data to find the correlation heatmap and regression model because the data must be numeric such as number_of_arrest, number_for_stop (reason_for_stop = “Moving Violation”), subject_age_mean, and number_of_person.

The Results



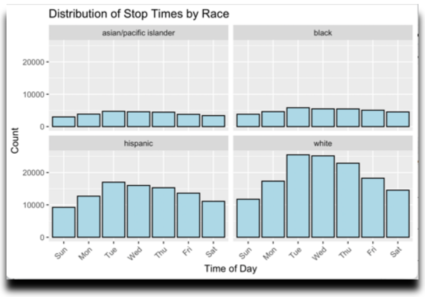
The histogram of Distribution of traffic stop by day

The histogram shows that the distribution of traffic stop by day is quite different. The lowest stop is Sunday that might be because Sunday is a family day, so most people stay home. On the other hand, Tuesday is the highest. Tuesday, Wednesday, and Thursday are quite similar. The trend is gradually decline from Tuesday to Sunday.



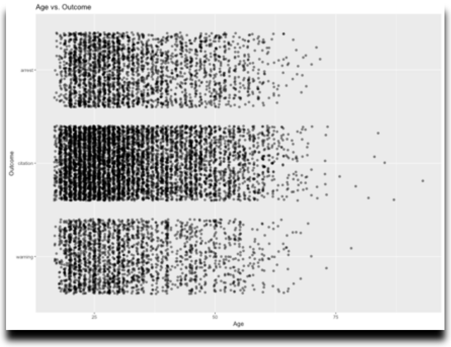
The bar graph of subject_sex by race with “ggplot”

Across all races, the results show that males are subjected to traffic stops approximately twice as often as females.



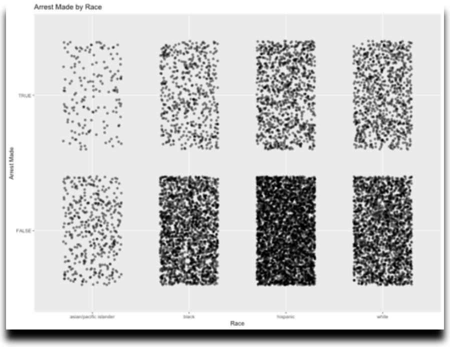
The bar chart of distribution of stop each day by separated race in the different areas with “ggplot”

This visualization is so powerful because I can compare the distribution among races and among days of week easily with just one picture. The most variation in traffic stops throughout the day is White drivers compared to other races.



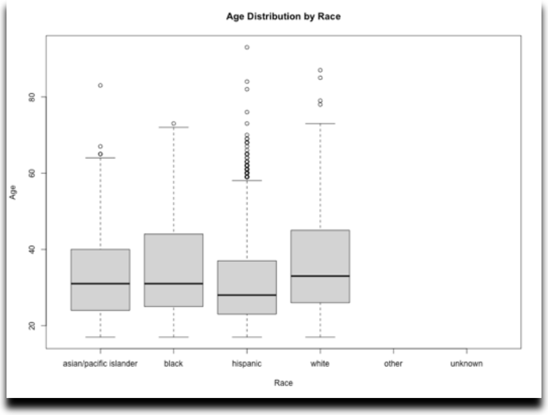
Scatter Plot

- The distribution of traffic stops of arrest and warning are similar that occurs in age 20-40 years, and lighter in age over 60 years.
- Traffic stops from citation are higher frequent than arrest and warning. The younger drivers are arrested more often than older drivers, because of the high density of the spots.



Jitter Plot

- TRUE = drivers were arrested, FALSE = drivers were not arrested. This plot shows that most drivers are arrested less than are not arrested across the races.
- In terms of TRUE, Asian/Pacific islander is the lowest density due to the lightest distribution of the spots.



Boxplot

- This result shows the IQR for White and Black drivers are higher than Asian/Pacific islander and Hispanic races, indicating that White and Black drivers are distributed across a wider age range, while the rest two groups are more concentrated in narrower age ranges.

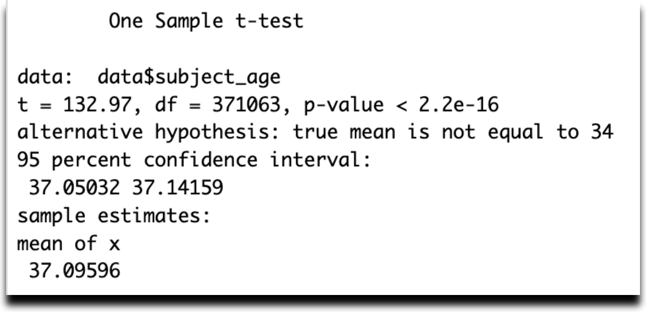
- Most Hispanic drivers are young compared to other races because of their lowest median, minimum age, maximun age, and narrowest IQR
- The outliers vary slightly across races, for Hispanic is 60 years, for Asian/Pacific islander is 65, Black and White are 70 years
- There are many outliers in Hispanic group compared to other groups significantly.

Creating Hypothesis Statements for San Diego

1. Hypothesis 1:

Null Hypothesis (H_0): The mean age of individuals stopped is equal to 34 years.

Alternative Hypothesis (H_1): The mean age of individual stopped is not equal to 34 years.



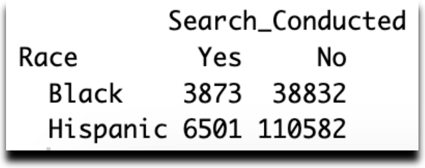
```
        One Sample t-test

data:  data$subject_age
t = 132.97, df = 371063, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 34
95 percent confidence interval:
 37.05032 37.14159
sample estimates:
mean of x
 37.09596
```

Interpretation: The one-sample t-test compares the sample mean age to a hypothesized population mean of 34 years. The degree of freedom shows how large of this dataset is. The null hypothesis assumes that the true mean age is equal to 34 years. The result depicts the p-values is < $2.2e^{-16}$ or 2.2 x 10^{-16} that is less than our significant level (0.05), so we reject the null hypothesis and conclude that there is evidence to suggest that the mean age of individual stopped is significantly different from 34 years.

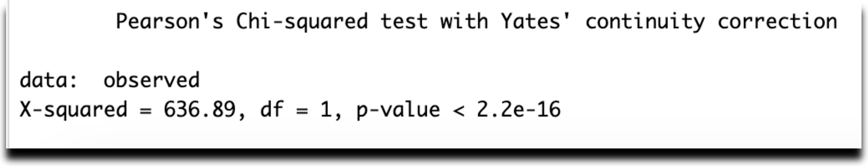
2. Hypothesis 2:

Null Hypothesis (H_0): There is no association between races (Black vs. Hispanic) and being search conducted during a traffic stop.

Alternative Hypothesis (H_1): There is an association between races (Black vs. Hispanic) and being search conducted during a traffic stop. In other words, from contingency table below, Hispanic drivers are searched more than Black.

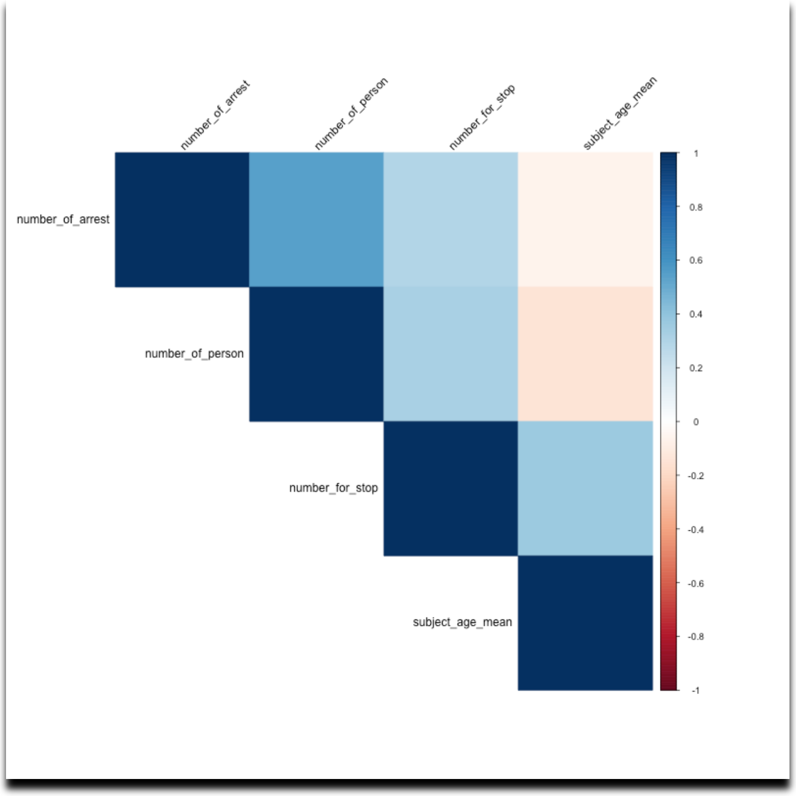


```
          Search_Conducted
Race        Yes     No
  Black     3873  38832
  Hispanic 6501 110582
```



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

data:  observed
X-squared = 636.89, df = 1, p-value < 2.2e-16
```

Interpretation: From the result, p-value < $2.2e^{-16}$ or 2.2 x 10^{-16} that is less than 0.05, so we reject the null hypothesis. It means that there is evidence to suggest that Hispanic drivers are more likely to be searched than Black drivers.

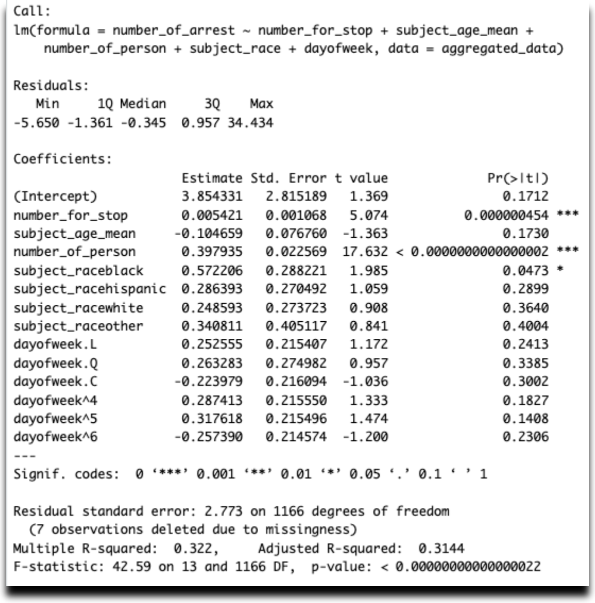


Correlation heatmap

I am focusing on the relationship between number_of_arrest and other variables, as these will be relevant to the regression model I will develop next.

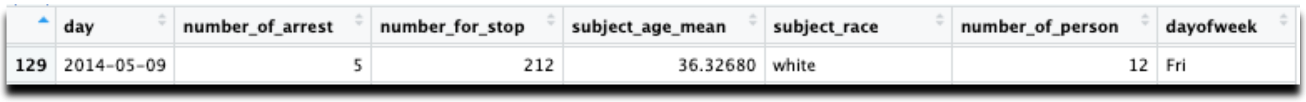
- Number_of_arrest shows a positive correlation with both number_of_person and number_of_stop. As the number_of_arrest increases, both the number_of_person and number_of_stop tend to increase as well.
- The relationship between number_of_arrest and number_of_stop appears to be weak, as indicated by the pale blue color, suggesting that the correlation is far from 1. Similarly, the connection between number_of_arrest and number_of_person is moderately weak, sitting somewhere in the middle but still showing a lack of strong association.
- There is a negative relationship between number_of_arrest and subject_age_mean, meaning that as the number_of_arrest increases, the subject_age_mean tends to decrease. However, the pale red color indicates that the correlation is relatively weak, as it is far from -1.

Linear regression model

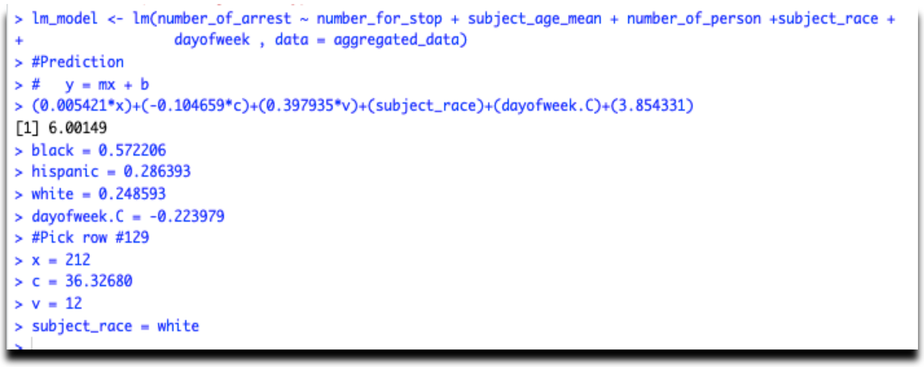


```
Call:
lm(formula = number_of_arrest ~ number_for_stop + subject_age_mean + 
    number_of_person + subject_race + dayofweek, data = aggregated_data)

Residuals:
   Min     1Q Median     3Q    Max 
-5.650 -1.361 -0.345  0.957 34.434 

Coefficients:
                      Estimate Std. Error t value            Pr(>|t|)    
(Intercept)           3.854331   2.815189   1.369              0.1712    
number_for_stop       0.005421   0.001068   5.074         0.000000454 ***
subject_age_mean     -0.104659   0.076760  -1.363              0.1730    
number_of_person      0.397935   0.022569  17.632 < 0.0000000000000002 ***
subject_raceblack     0.572206   0.288221   1.985              0.0473 *  
subject_racehispanic  0.286393   0.270492   1.059              0.2899    
subject_racewhite     0.248593   0.273723   0.908              0.3640    
subject_raceother     0.340811   0.405117   0.841              0.4004    
dayofweek.L           0.252555   0.215407   1.172              0.2413    
dayofweek.Q           0.263283   0.274982   0.957              0.3385    
dayofweek.C          -0.223979   0.216094  -1.036              0.3002    
dayofweek^4           0.287413   0.215550   1.333              0.1827    
dayofweek^5           0.317618   0.215496   1.474              0.1408    
dayofweek^6          -0.257390   0.214574  -1.200              0.2306    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.773 on 1166 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.322,	Adjusted R-squared:  0.3144 
F-statistic: 42.59 on 13 and 1166 DF,  p-value: < 0.00000000000000022
```

- The coefficient indicates positive relationships between number_of_arrest and both number_of_stop and number_of_person. For example, a one-unit increase in number_of_arrest is associated with a 0.005421 increase in number_of_stop, which is statistically significant with a P-value of 0.000000454, highlighting a strong relationship marked by three asterisks.
- The analysis reveals negative relationships between number_of_arrest and both subject_age_mean and dayofweek.C. For instance, a one-unit increase in number_of_arrest is associated with a decrease of 0.104659 in subject_age_mean, though the P-value of 0.1730 indicates that this relationship is not statistically significant. This finding aligns with the heatmap, which suggests a weak negative relationship.
- The R-squared value is 0.322, meaning that these variables collectively explain 32.2% of the variation in number_of_arrest.
- For the prediction model, I chose dayofweek.C because the cubic transformation (C) captures the fluctuating pattern across the week. Based on my earlier analysis, particularly the bar chart of distribution of stop each day by separated race in the different areas with “ggplot”, I observed significant variations in patterns across different days and races. This fluctuation makes the cubic model a good fit for my dataset, as it can better capture these dynamic changes.



	day	number_of_arrest	number_for_stop	subject_age_mean	subject_race	number_of_person	dayofweek
129	2014-05-09	5	212	36.32680	white	12	Fri



```
> lm_model <- lm(number_of_arrest ~ number_for_stop + subject_age_mean + number_of_person +subject_race +
+                  dayofweek , data = aggregated_data)
> #Prediction
> #   y = mx + b
> (0.005421*x)+(-0.104659*c)+(0.397935*v)+(subject_race)+(dayofweek.C)+(3.854331)
[1] 6.00149
> black = 0.572206
> hispanic = 0.286393
> white = 0.248593
> dayofweek.C = -0.223979
> #Pick row #129
> x = 212
> c = 36.32680
> v = 12
> subject_race = white
> 
```

- After comparing the predicted and actual values for row #129, I found that the regression model predicted a number_of_arrest of 6.00149, while the actual value was 5.

Conclusion

The analysis shows that traffic stops are lowest on Sundays, suggesting the traffic team could reduce staffing on that day to optimize their budget. Males are stopped more frequently than females, indicating the need for closer monitoring of male drivers. Traffic stop patterns vary by race. Asian/Pacific Islander and Black drivers have similar distributions, while White and Hispanic drivers show notable differences. Most drivers receive citations rather than arrests, aligning with the expectation that citations are more common than arrests in less severe cases.

Drivers over 60 years old are outliers, with fewer traffic stops, we should scrutinize to understand why they got traffic stop. Arrests are positively correlated with both the number of people and stops, but negatively with age, suggesting younger drivers are arrested more often. While these correlations are weak, they provide useful insights for targeting younger drivers in real-world monitoring.

Although the model explains 32.2% of the variance in arrests, with a predicted value of 6.00149 vs. the actual 5, it offers a rough estimate. The visualization highlights areas for further exploration, such as understanding traffic stops for older and Hispanic drivers. The model can also forecast arrest trends, helping with staffing and resource planning in different areas.

Citation

San Diego Personal Injury Lawyer. Mission Personal Injury Lawyers. (2024, October 2). https://missionlegalcenter.com/