Analyzing Police Stoppage Activity in Rhode Island, USA

CONTENTS

1. Introduction

Are the people of the US somewhat biased among few races? Due to recent happenings and overtime in various states of the United States of America, this becomes an important question in every field/aspect of life, be it job roles, arresting, community rights, voting rights and whatnot.

We in our analysis are considering Rhode Island state as the basis of our hypothesis. Unfortunately, we don't have recent years data of policing stoppage, but it is worth exploring if this bias has been continuing from the past. A Few of the major questions we want to look at are:

1. Is the Police action biased based on the driver's Gender or Race?
2. How does the time of the impact affects stoppage?
3. Does the weather impact Police actions at various stops? If yes, how?

2. Data Wrangling

Tools Used: R

2.1. Datasets

2.1.1. Traffic Stoppage data of Rhode Island, USA

Link: https://www.kaggle.com/faressayah/stanford-open-policing-project
Feature Description: See Appendix

This dataset contains various stoppage data from Rhode Island, starting from 2005 and extending till 2015. The dataset published by the team at Stanford Open Policing Project. It has ~91K rows x 15 columns with several types of data attributes.

2.1.2. Weather data of Rhode Island

Link: https://www.kaggle.com/sachinsk/weather-inprovidence-rhode-island
Feature Description: See Appendix

This dataset contains weather data collected by the National Centers for Environmental Information. It contains data that has been collected by a single station in Rhode Island. Ideally, several recording stations should be considered, but as Rhode Island is the smallest state in the U.S so the data from this station won't have much variance across the state. ~4K rows x 27 columns present.

2.2. Data Cleaning and Data Transformations

2.2.1. Traffic Stoppage data of Rhode Island

After successful reading of the dataset, take a glimpse of values by checking few initial rows. Also, check for the total numbers of columns and rows present. The output comes out to be 91741 rows and 15 columns.

Handling Missing Data

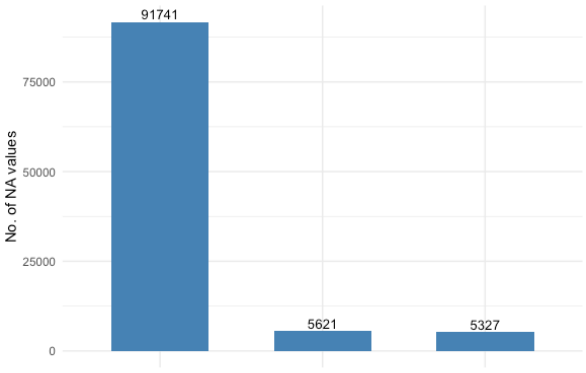


Figure 1 Count of NA values in Policing dataset

a) Check the count of NA values in each column and take appropriate action.

i. It looks like the whole *county_name* column is empty. Drop this column as we are already aware that our dataset belongs to Rhode Island.

ii. Still, 5327 and 5621 NAs in *driver_age_raw* and *driver_age* needs treatment respectively. It makes sense to drop NA rows of *driver_age_raw* and we expect that those rows will ultimately contain *driver_age* as NA as well. Few i.e., 5621-5327 = 294 rows will remain NA in *driver_age*, which on further checking showed these row's *driver_age_raw* value is 0. Discard these remaining rows as the drive_age_raw (year of born) cannot be zero.

Data Type Checking and Correcting

a) Only *driver_age_raw* and *driver_age* is by default numeric datatypes. All other remaining features/columns are character datatype by default.

b) We can go through each column unique values to get an idea about the column values. After checking we found out that most of our columns are categorized variables and few of them related to DateTime. There won't be an issue if we continue with the features as characters but converting them to category (factor in R) increases the processing and analysis speed.

c) Change to factor datatype
   i. The following columns were converted to factor using the as.factor():
      driver_gender, driver_race, violation, stop_outcome, stop_duration

d) Change to logical datatype
   i. The following columns were converted to logical using the as.logical():
      search_conducted, is_arrested, drugs_related_stop

e) Date Time column creation
   i. We already have a date column in the form of *stop_date* and *stop_time* as a time column. These two columns can be concatenated into a single column (*stop_date_time*) containing both the date and time of the stoppage.

   ii. Initially, *stop_date_time* will be of character type. This column serves us great importance for time series analysis, so it is necessary to convert this to some datetime datatype. R has POSIXct and POSIXlt, we chose POSIXct because libraries such as dplyr have some issues while manipulating and working with POSIXlt datatype.

2.2.2. Weather data of Rhode Island

- Read the weather data from the weather recording station provided in the dataset. Look at the glimpse of the dataset using head() in R. Also you can check the shape of the dataset. The shape comes out to be 4017 rows x 27 columns.

- There is not much to change with the data types except for changing the *DATE* column to POSIXct i.e., a date format so as we can perform manipulations based on it in future. The Rest of the columns are auto identified to the respective classes after the read operation.

- Now, let's check for NA values and plot only the counts of the columns (if any) which have NA present.

```
> sapply(weather_df, function(x) sum(is.na(x)))
STATION    DATE    TAVG    TMIN    TMAX    AWND    WSF2    WT01    WT02    WT03    WT04    WT05
      0       0    2800       0       0       0       0    2250    3796    3793    3900    3657
   WT06    WT07    WT08    WT09    WT10    WT11    WT13    WT14    WT15    WT16    WT17    WT18
   3992    3938    3613    3948    4015    4016    2842    3442    4011    2691    4005    3672
   WT19    WT21    WT22
   4013    3999    3985
```

Figure 2 Count of NA values in weather dataset

- The features *WT01 – WT22* are encoded depiction of certain bad weather value. (For example, *WT01* can be raining). We haven't been informed what these parameters exactly mean. The values inside such columns are either NA or 1. 'NA' denotes non-prevailing of that particular type of bad weather attribute on that date, while '1' denotes the presence. So, we can simply turn all the NAs to '0' because we will require taking the sum of each row to create a rating system of weather.

- The point of concern is *TAVG* feature because it contains 2800 NAs. *TAVG* seems to be like a computed column of *TMIN* and *TMAX*. The best possible solution, which won't require dropping NAs, will be to impute the values of *(TMIN + TMAX)/2* to the NA values of *TAVG*.

```
weather_df$TAVG <- ifelse(is.na(weather_df$TAVG), ceiling((weather_df$TMIN + weather_df$TMAX)/2), weather_df$TAVG)
```

Figure 3 Imputing TAVG NA values based on condition

Here, data checking or sanitization notion arises to check if the non-null values of the *TAVG* column are correct or not. We will be checking this in our Data Checking part and handle anomalies (if any) there itself.

- Create two new features *'WIND_DIFF' (WSF2 - AWND)* and *'T_DIFF' (TMAX – TMIN*), which will be used later to gain confidence in the correctness of the dataset.

3. Data Checking

Tools Used: R & Tableau

After the operations performed during Data Wrangling, we need to check our data is trustworthy or not. This can be done by checking values of computed columns, there is no impossible value that cannot exist, checking for outliers, collinearity, and several other techniques.

3.1. Traffic Stoppage data of Rhode Island

- Values of *driver_age* can be checked by subtracting *driver_age_raw* from the year of *stoppage_date*. The resultant age of the driver can be -1 because it is possible his/her birthday

is yet to come. Taking this into consideration we calculated and found out all the values are correct and in range.

- Our dataset includes more categorical and Boolean variables and less numeric based, so we are not expecting concrete results from the collinearity plot.

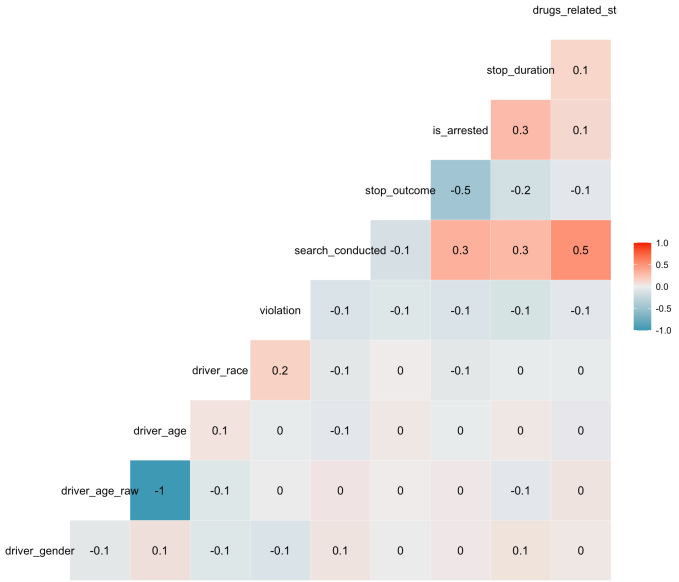


Figure 4 Collinearity plot created using ggcorr()

3.2. Weather data of Rhode Island

Wind Speed

- *AWND* is the average wind speed of the day and *WSF2* is the maximum speed noted in the frequency of 2 minutes. So *AWND* should always be less than *WSF2*. We confirmed this and there were no anomalies.

- Create a summary statistic or a box plot of wind-related columns *(AWND & WSF2)*

```
> summary(weather_df[c("AWND","WSF2")])
      AWND             WSF2
 Min.   : 0.220   Min.   : 4.90
 1st Qu.: 6.260   1st Qu.:15.00
 Median : 8.050   Median :17.90
 Mean   : 8.594   Mean   :19.27
 3rd Qu.:10.290   3rd Qu.:21.90
 Max.   :26.840   Max.   :48.10
```

Figure 5 Summary of wind columns

Note the minimum values of both wind features are greater than 0, which is correct as the value of wind speed cannot be zero. The outliers depicted do not need treatment because there can be days where wind speed is quite high, so it is feasible to trust them.

```
ggplot(melt(weather_df[c("AWND","WSF2")]),
       aes(x=variable, y=value, fill=variable)) +
  geom_boxplot() +
  xlab("") +
  ylab("Speed") +
  theme_minimal()
```

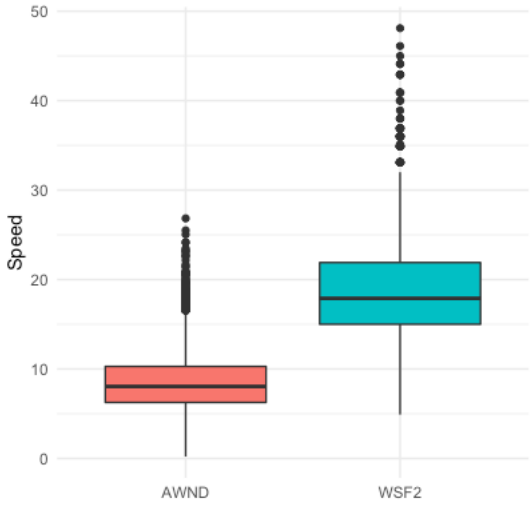
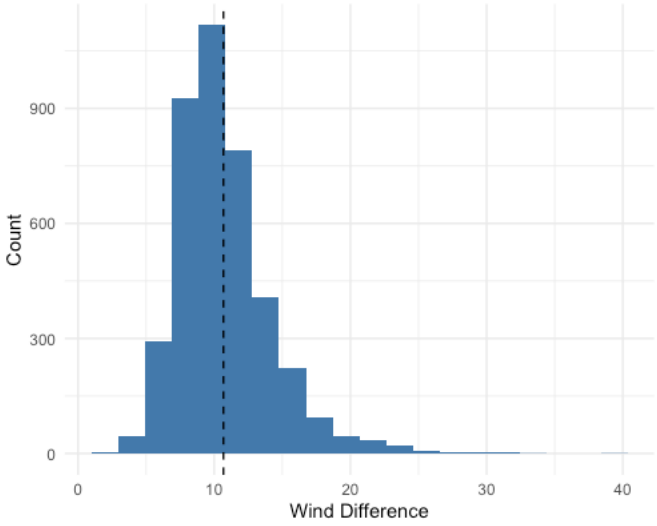


Figure 6 Box plot of AWND and WSF2 features

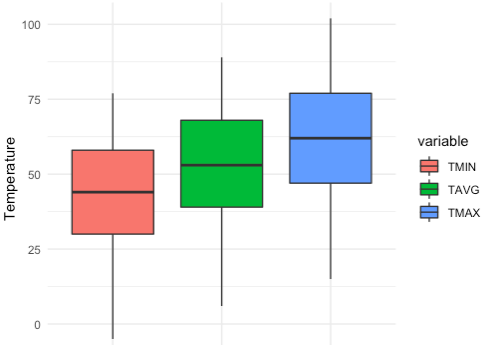
- Most of the real-world phenomena are normally distributed. Normally distributed characteristic is an important concept for doing any statistical tests or modelling. When plotted its results to a symmetric bell-shaped curve across the mean.



The curve plotted for WIND_DIFF columns is normally distributed, making sure that our data values are reliable.

Figure 7 Distribution plot of WIND_DIFF feature

Temperature



The box plot of TMIN, TAVG, and TMAX is a perfect example of an idle box plot. It has no outliers present. This increases our confidence in the dataset.

Figure 8 Box plot of Temperature Features

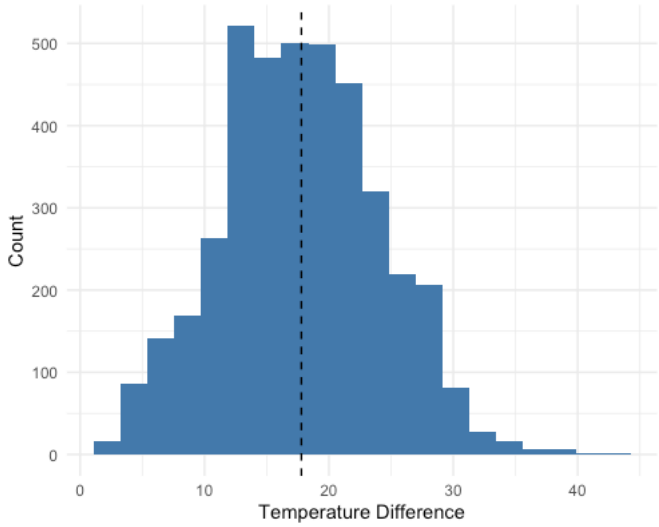


Figure 9 Distribution plot of TEMP_DIFF

- Creating histogram plot to check for normality of TEMP_DIFF column.

T_DIFF has no negative values, and its distribution is approximately normal, resulting in the conclusion our data is trustworthy.

4. Data Exploration

Tools Used: Tableau & R

4.1. Exploring the relationship between Stoppage and Driver Race

Rhode Island is the smallest state, has not much diversity or it is better to state that proportion of the population is more inclined towards a major race *. The summary dataset of the population of Rhode Island provides the following figures for the 2010 Census *[1]:*

* we do not intend to hurt the sentiments of any individual, community, or race. Black Lives Matter.

Total Population of 2010 = 1052567

Race	Proportion of Total Population (2010)
White	76.4%
Black	5.7%
Asian	2.9%
Hispanic	12.4%
Others * * other races are combined, to use with our other datasets efficiently	2.6%

Let's start exploring the relation step by step before we conclude.

```
  driver_race     n prop_race
1       Asian  2253    0.0262
2       Black 12197    0.1416
3    Hispanic  9477    0.1101
4       Other   239    0.0028
5       White 61949    0.7194
```

- We started with getting the count of each race present in policing activity dataset.
- The above would not make much sense, so added another factor of the proportion of each race we have.

- Let's look at how the number of policing stoppage changed for races with year.

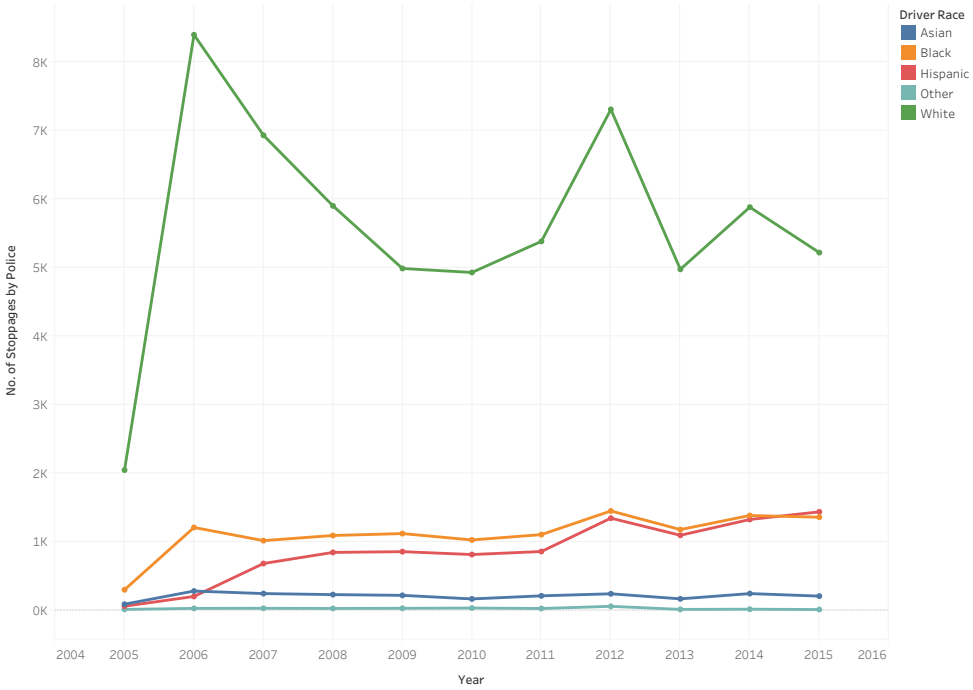


Figure 10 Count of stoppage of each race by year

The line plot shows that the number of stoppages of "White" drivers increased drastically from 2005-2006 and then again, a small increase of numbers was seen from 2011-2012. "Black" driver's stoppage count had a very steady trend over the years with not much of increase and fall. "Hispanic" driver's stoppage count remained lower than "Black" for all the years until it crossed than "Black" numbers by a small margin in 2015.

We cannot make any firm conclusion right now because this trend does not take into account each race population in the state. It's time to bring our Census 2010 population numbers into play.

- Since the policing dataset we have is last updated till 2015 and we have the population dataset of Census 2010, we will be analyzing the rate of stoppage among races for the year 2010.

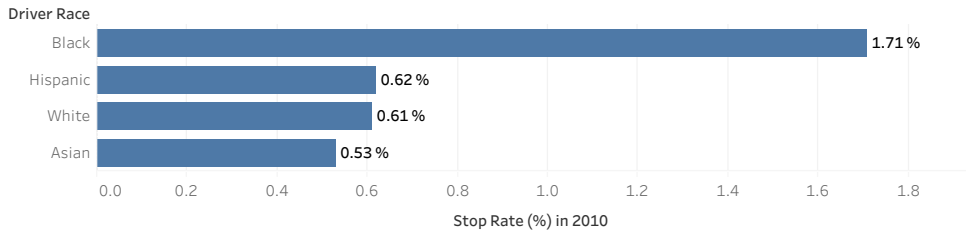
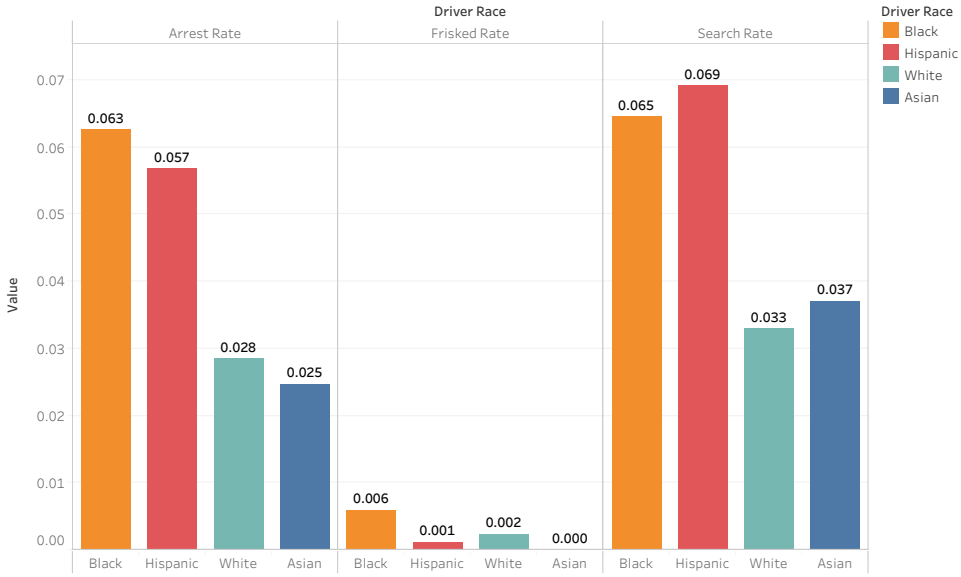


Figure 11 Stop Rate of each Race in 2010

From the visualisation (Fig 11) it can be concluded that the Black drivers are stopped 2.8 times more than the White drivers, relative to their proportion of the state's population in 2010.

- Till now we have done the analysis based on the stop rates. We can also look at the other factors and check if the trend/bias persists or not. Again, we will be considering the Census 2010 dataset.



There is bias in arrest rate and search rate as well. Frisked Rate shows bias too, but the numbers depend upon several other aspects.

These are interesting results because having a bias in a state with comparatively few demographics of Black or Hispanic race as compared to the other states of United States, gives us somewhat perspective of the bias that has been prevailing from the past in the country. This gives a sense of racial inequalities but is not evidence of discrimination performed by the police. There are several other aspects we should consider before stating it as evidence (discussed in 6. Reflection)

4.2. Exploring the relationship between Stoppage and Driver Gender

<u>Speeding violation outcomes by gender</u>

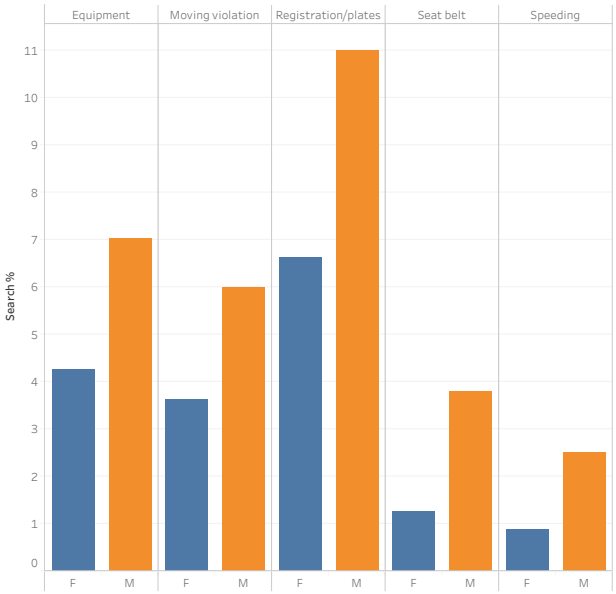
Stop Outcome of Speeding	Female Driver (%)	Male Driver (%)
Citation	95.34	94.72
Warning	3.91	3.42
Arrest Driver	0.54	1.52
Arrest Passenger	0.08	0.12
No Action	0.05	0.11
N/D	0.08	0.11

The figures are similar for males and females. About 95% of the stops resulted in Citation (ticket) for a speed violation. Also, the warning was given equally to both genders. This rejects our hypothesis of bias in speed violation.

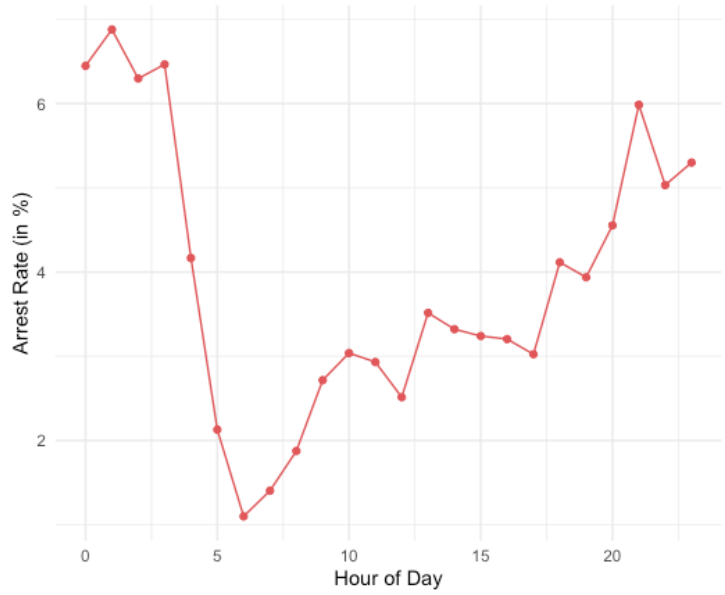
Figure 12 Gender vice search rate w.r.t Violation

<u>Search Rate by gender</u>

All the kinds of violation have a search rate for male almost twice that for females, so there is bias.



4.3. Time Impact on Arrest Rate



From the line visualization, it is observable that the arrest rate is less in daylight. Maximum arrests happen around 12:00 AM - 3:00 AM.

Figure 13 Arrest rate by hour of day

4.4. Weather Impact on Police Activity

We are unsure that any relationship or trend will be observed. Before commencing we will be constructing the rating system for our weather.

Sequential ordinal based rating system for the weather (WT01 – WT22)

Our policing dataset is based more on categorical data due to which it is better to create a Sequential ordinal based rating system for our weather dataset. This rating system will be based on the sum of all the bad weather boolean parameters on each day and categorized into Good<Bad<Worse based on the following:

- If the sum for the particular date is 0, the weather is considered Good.
- If the sum for the particular date is 2,3 or 4, the weather is considered Bad.
- If the sum for the particular date is greater than 4, the weather is considered Worse.

```
rating_map <- c(`0`="good",`1`="bad", `2`="bad", `3`="bad",`4`="bad",`5`="worse",`6`="worse",`7`="worse",`8`="worse",`9`="worse")
weather_df$weather_rating <- recode(weather_df$bad_weather, !!!rating_map)
weather_df['weather_rating'] = as.factor(weather_df$weather_rating)
weather_df$weather_rating <- factor(weather_df$weather_rating, levels = c("good","bad","worse"), ordered = TRUE)
```

Figure 14 Weather Rating Procedure

Preparing the Merged Dataset

```
policing_weather_df <- merge(
  x=policing_activity,
  y=weather_df[c("DATE", "weather_rating")],
  by.x = "stop_date",
  by.y = "DATE",
  all.x = TRUE)
```

Figure 15 Merging the tables

We will be needing a merged dataset of police activity and weather's *weather_rating* column. To do this we will use left join on policing dataset with the weather dataset.

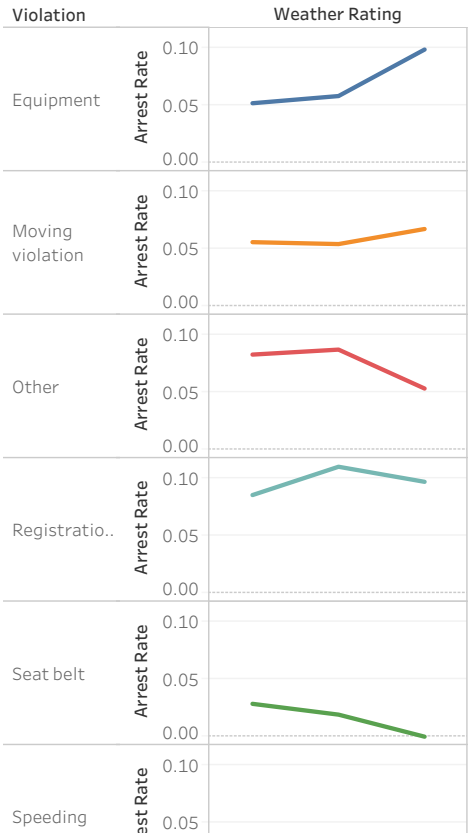


Figure 16 Weather impact on arrest rate for each violation

The arrest rate increases as the weather get eviler and the same trend continued in most of the violations. "Equipment" violation has the steepest increasing trend. Arrest rate due to "Seat belt" violation decreases as the weather worsens, which is an accurate outcome because visibility decreases, resulting in less stoppage of cars for seat belt violation.

These results are thought-provoking findings, but it doesn't prove a causative connection.

5. Conclusion

The exploration gives us a sense of racial inequalities but is not evidence of discrimination performed by the police.

All the kinds of violation have a search rate for male almost twice that for females, so our hypothesis of bias with respect to gender is proved correct.

It is observed that the arrest rate increases as the daylight diminishes.

The arrest rate increases as the weather get eviler and the same trend continued in most of the violations. This result is thought-provoking findings, but it doesn't prove a causative relationship.

6. Reflection

As we are marking the end of our exploration project a lot of takeaways are there to take from this project. This project enabled us to follow a procedure of analysis, from data collection to data exploring. During the execution of the process, I levelled up my R and Tableau skills.

There are few things I would have done differently. The first one would be to choose a similar dataset of U.S state that will include spatial attributes as well. Such spatial parameters can help us answer our question in-depth with cluster analysis. Hopefully, I will be able to choose it while executing the interactive dashboard. Dataset like race vice driving population from the *Rhode Island Department* of Transportation could have helped us to turn the results of racial bias shown by policing activity as evidence.

7. Bibliography

[1] United States Census Bureau. (2017). Providence (city), Rhode Island. Retrieved from https://web.archive.org/web/20130127054813/http://quickfacts.census.gov/qfd/states/44/4459000.html
[2] Tableau Community. Build a Box Plot. Retrieved from https://help.tableau.com/current/pro/desktop/en-us/buildexamples_boxplot.htm
[3] Tableau Community. Show, Hide, and Format Mark Labels. Retrieved from https://help.tableau.com/current/pro/desktop/en-us/annotations_marklabels_showhideworksheet.htm
[4] Harshita Mekala. (2018). Dealing with Missing Data using R. Retrieved from https://medium.com/coinmonks/dealing-with-missing-data-using-r-3ae428da2d17
[5] Nicola. (2019). POSXlt and filter() in R. Retrieved from https://stackoverflow.com/questions/56850331/posxlt-and-filter-in-r
[6] http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization
[7] https://stackoverflow.com/questions/39903376/if-column-contains-string-then-enter-value-for-that-row

APPENDIX

R Code Repository
Link: https://gitlab.com/abhigambhir97/r_ri_police_activity

Dataset: Traffic Stoppage data of Rhode Island, USA

stop_date: date of stoppage
stop_time: time of stoppage
county_name: county name, though empty
driver_gender: gender of the driver
driver_age_raw: driver dob year
driver_age: age of the driver at the time of stoppage
driver_race: driver's race
violation_raw: raw form of violation, can include several violations separated by comma
violation: categorized violation i.e., the final violation for which the driver is charged
search_conducted: Boolean, whether the driver's car or driver was searched
search_type: reason for which search was conducted
stop_outcome: outcome of the stoppage
is_arrested: Boolean, whether driver was arrested or not
stop_duration: Range of time for which the driver was stopped
drugs_related_stop: Boolean, whether the stop was made for drugs checking or consumption

Dataset: Weather data of Rhode Island

STATION: weather station code which recorded the row
DATE: date
TAVG: average temperature of the day
TMAX: maximum temperature noted during the day
AWND: average wind speed of the day
WSF2: maximum speed noted in the frequency of 2 minutes during the day
WT01 – WT22: encoded Boolean value of certain bad weather attribute (not mentioned which exact factor it is)