**New York Police Department Stop-and-Frisk Data Analysis**

Logo

Description automatically generated

**Northeastern University**

**College of Professional Studies**

**ALY6015 - 21454: Intermediate Analytics**

**Professor Wada Roy**

**Group Members**

Amryta Panda

Bennett Furman

Durga Bhanu Nayak

Date of Submission: Apr 02, 2022

**INTRODUCTION**

The stop & frisk policy was initially introduced in 1968 to check the crime rates of suspects showing criminal behavior in New York City. The policy went through some changes in the 1990’s. In the year 2002, a report found that among the people who were stopped, 82% were innocent. After this, as years passed, the number of stops grew dramatically. The majority of the targeted population were Black followed by Hispanic individuals. During the year 2011, a population of more than 600,000 people stopped. The data from the year 2003-2011 shows that 82-90% of the population who were stopped were innocent, and more than half the population were Black and 27-34% were Hispanic. The Hispanic race is divided into 2 categories, which include White Hispanic and Black Hispanic. During the year 2012, complaints were raised against the ‘Stop and Frisk’ policy. This policy was flagged for racial profiling, harassing minorities, and violating the constitutional rights of citizens. This case was heard by Judge Shira Scheindlin, who concluded that the ‘Stop and Frisk’ policy had become a method of indirect racial profiling. Also, the policy objectified the people belonging to certain races. The policy ended in January 2014, when the litigation was settled, and the program was closed.

In this document, we will be discussing the “Stop and Frisk” data from the year 2003-2015. The dataset provides precinct level data and not individual level data. A total number of 77 precincts are present in the dataset, which has 31 variables and 12,012 observations. As the data is grouped by precinct, yearly totals of stops by New York City precinct can be calculated. We will look into the worst affected races from the ‘Stop and Frisk’ policy and determine whether the aforementioned claims are true or false.

Keywords*: Black, racial, stop and frisk, precinct, subset, variables, population, regression, correlation, chi-square, ANOVA, LASSO, stepwise selection*

**ANALYSIS**

1. **Exploratory Data Analysis**

As we are having a summarized data of NYPD stop and frisk, we will start our analysis by loading and exploring the data to our R Studio environment.

**Loading the dataset**

*## Importing dataset*

nypd <- read.csv('NYPD\_SQF.csv')

After successfully loading the dataset, we explored the dataset by checking its dimensions, structure of the dataset.

Then we looked for any missing values in the dataset across every column

*## Checking for any missing values*

colSums(is.na(nypd))

Here, we could not find any missing values in the dataset.

Now, we converted some of the required categorical variables as factor which includes, year, pct (precinct), sex and race.

*## Converting the required variables into factors*

nypd <- nypd %>%

mutate(year = as.factor(year),

pct = as.factor(pct),

sex = as.factor(sex),

race = as.factor(race))

After performing the above sanitary checks, we are now interested to check the insights from the data on how and what information can be extracted from the exploration of the given dataset.

*# Total Stops by Year*

nypd %>% group\_by(year) %>%

summarise(Total\_Stopped = sum(stopped))%>%

ggplot() +

geom\_line(aes(x = year, y = Total\_Stopped, group = 1), col = '#440DFA', lwd = 1.5) +

geom\_point(aes(x = '2011', y = max(Total\_Stopped), size = 3, col = 'red'), show.legend = F) +

annotate('text', label = '659,079', x = '2010', y = 659079) +

geom\_vline(xintercept = '2012', linetype = 'dashed') +

annotate('text', label = 'Petition Filed', x = '2013', y = 540000) +

ggtitle('NYPD stops plummetted after Petition filed in 2012') +

xlab('Year') + ylab('Total Stops by NYPD') +

theme\_classic(base\_size = 12)

Chart, line chart

Description automatically generated

*Fig 1. Total number of stops by NYPD during year 2003 - 2015*

**Observations:**

From the above chart, it is shown that since 2003, the process of stopping individuals by NYPD was increasing each year. There was a peak of more than 650K cases during 2011 which raised an alarming situation among residents of New York and a petition was filed later the year as per the New York Times. After the petition got filed, we can see a sudden drop in the total stops by NYPD till the year 2015. During this stop and frisk process, many citizens accused this policy as a racial bias act toward people belonging from different races.

As we observed that people were accusing the New York Police Department of the racial bias they were conducting in the name of law to harass people from races other than their own. We were curious if the data reflects the same for the people getting stopped and frisked by NYPD even though they are innocent.

Chart, histogram

Description automatically generated

*Fig 2. Total number of frisks by NYPD across races*

Here, we can clearly observe that average number of people from ‘Black’ race have been frisked the highest among all he races which was followed by the ‘Black Hispanic’ race. The average ‘White’ race frisks are almost 6 times lesser than the ‘Black’ race frisks. This somewhat shows that a clear racial bias was being enacted by the NYPD against the people of color.

Now that we observed that there was a racial bias going on based on which many innocent people were stopped and were subjected to frisking actions which was clearly indicating the violation of their privacy and unnecessary mental harassment was imposed over them.

Hence, we selected top three racial groups which were getting affected by this policy, which include races as, ‘Black’, ‘Black-Hispanic’ and ‘White’. To study them further, we made subsets of these groups.

*## Creating subgroups according to Races*

race.B <- subset(nypd, race == 'B') *## Black race*

race.Q <- subset(nypd, race == 'Q') *## Black-Hispanic race*

race.W <- subset(nypd, race == 'W') *## White race*

Before going further, we also wanted to check the data about the top police precincts where the maximum number of stopping happened. For this we selected, top 5 precincts in New York.

Chart, bar chart

Description automatically generated

*Fig 3. Top Police precincts with Highest number of stops in New York*

Here, we can clearly observe that, police precinct number **75** - East New York is recorded for having highest number of stops. This is followed by police precinct **73** – Ocean Hill Brownsville and then precinct number **79** – Bedford-Stuyvesant. This suggests people travelling around these areas are affected by the stop and frisk policy the most.

Now that we observed, many people from Black race is getting frisked than the other races, we wanted to check if according to the requirement of the policy what is the proportion of contrabands and weapons getting found across races.

*## Proportion of Illegal weapons and contraband found among races*

nypd\_stats <- nypd\_dummy %>% pivot\_longer(!c(race,total\_frisked), names\_to = 'Illegal\_Stuff',values\_to = 'Percent\_found')

ggplot(nypd\_stats[which(nypd\_stats$race %in% c('B','W')),]) +

geom\_bar(aes(x= reorder(race, -Percent\_found), y=Percent\_found, fill = Illegal\_Stuff), stat = 'identity', color='white', position = 'dodge') +

scale\_x\_discrete(breaks=c('W','B'), labels = c('White', 'Black'))+

xlab('Race of people')+ ylab('Illegal Stuff found in %') +

ggtitle('Rate of Illegal Stuff is found higher with Whites')+

scale\_fill\_manual(values = c('#F37B93','#7BCFF3'))+

theme\_classic(base\_size = 14)

Chart, bar chart

Description automatically generated

*Fig 4. Comparison of proportion of weapons and contrabands found among races*

From the above plot, we can observe that even though Black race people are getting frisked higher than the white race people, the proportion of contrabands and weapons confiscated from the white people is almost double the rate than black people. This proves that many of the black race people are getting frisked on the basis of racial bias even though majority of them are innocent. This why a sanction might be needed over the strict use of this stop and frisk policy.

After observing the number of stops and frisks happened across the racial groups and observing the racial bias, we have selected three groups to be analyzed further and confirm if any actual racial bias being enacted under the name of stop and frisk policy.

To study this, we checked the descriptive statistics of the races (‘Black’, ‘Black-Hispanic’ and ‘White’) and compared the mean and standard deviations for these race groups.

*Table 1. Descriptive statistics for three selected races*

| ***(n = 2002)*** | *Black* | | *Black Hispanic* | | *White* | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variables** | **mean** | **sd** | **mean** | **sd** | **mean** | **sd** |
| stopped | 1,297.81 | 2,570.65 | 608.42 | 1,131.64 | 152.96 | 315.27 |
| arrested | 76.96 | 127.09 | 37.89 | 63.68 | 10.37 | 19.90 |
| frisked | 704.39 | 1,456.63 | 327.20 | 733.49 | 87.67 | 200.23 |
| searched | 108.07 | 196.86 | 55.29 | 110.52 | 14.09 | 28.55 |
| summoned | 78.28 | 196.31 | 39.10 | 83.55 | 9.80 | 24.73 |
| contrabn | 23.61 | 42.78 | 10.92 | 18.59 | 3.02 | 6.47 |
| weapnfnd | 14.39 | 28.39 | 7.97 | 15.00 | 2.02 | 4.67 |
| pf\_weapn | 8.91 | 14.90 | 3.46 | 6.08 | 1.03 | 2.42 |
| pf\_hcuff | 48.07 | 77.13 | 21.96 | 36.87 | 6.04 | 11.06 |

**Observations**

From the above descriptive statistics table, we can observe, average stops, arrests, and frisks were comparably way higher in the Black race as compared to the others. Even the highest data variations were being recorded for the Black people than other races. Compared to ‘white’ race, people from ‘black’ race are being stopped, frisked, searched by around 8 times higher. This includes the physical forces being applied to them using guns or handcuffs.

Using the above data, we wanted to check statistically if the frisk to stop ratio among the above-mentioned races is significantly different or not. For this we created a new variable for frisked to stop ratio and performed ANOVA test to compare this variable for these groups.

To perform this ANOVA test, we first defined the Null and Alternate Hypothesis, as:

***Claim, H­0*** = Average frisk-to-stop ratio is similar among all the selected races

***H1*** = Average frisk-to-stop ratio of at least one race is significantly different from the others

***Performing ANOVA***

*# ANOVA test for frisk to stop ratio among top affected races*

f.anova <- aov(frisk\_stop\_ratio ~ race, data = filter(nypd, nypd$race %in% c('B', 'Q', 'W')))

*# Saving summary to an object*

f.summary <- summary(f.anova)

*Table 2. Summary Statistics of ANOVA test on different races*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Df** | **Sum Sq** | **Mean Sq** | **F value** | **p - value** |  |
| race | 2 | 7.67 | 3.835 | 91.05 | <0.0000000000000002 | \*\*\* |
| Residuals | 6003 | 252.82 | 0.042 |  |  |  |

For a significant value (α) of 0.05, we will also calculate the critical value as,

*# Critical Value*

qf(p=0.05, df1 = df.numerator, df2 = df.denominator, lower.tail = F)

Here, we found the Critical value as **2.997228**

***Tukey Test***

*# Tukey Test*

TukeyHSD(f.anova, conf.level = 0.95)

*Table 3. Tukey Test Statistics*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **diff** | **lwr** | **upr** | **p - value** |
| **Q-B** | -0.04 | -0.05 | -0.01993 | 0.0000002 |
| **W-B** | -0.09 | -0.102 | -0.07179 | 0 |
| **W-Q** | -0.05 | -0.067 | -0.03665 | 0 |

From the above Tukey test result, we can observe the p-value which is much lower than the significant value (0.05).

***Result***

Since, our F-test value is much higher than the critical value, we have enough evidence to reject the Null Hypothesis / Claim.

Hence, from the above ANOVA test, we can clearly state that, at a significance level of 5%, average frisk-to-stop ratio is significantly different among the Black, Black-Hispanic, and White races.

**Observing difference before and after the petition filed**

Now that we have proved that there is a statistically significant difference among the frisking and stopping activities based on races, we also wanted to check if there was any difference found after filing the petition again the stop and frisk act. To perform the same, we will be checking the distribution of frisks happened across the races in the year of 2011 and 2012 when the petition got filed.

Below is the table depicting the average frisking frequencies across all races for year 2011 and 2012.

*Table 4: Average frisk distribution for year 2011 and 2012 race wise*

| **race** | **Year\_2011** | **Year\_2012** |
| --- | --- | --- |
| A | 73 | 52 |
| B | 1,311 | 1,055 |
| I | 9 | 7 |
| P | 182 | 135 |
| Q | 641 | 482 |
| W | 177 | 143 |

Significance Level, **α** = 0.05

***Stating Hypothesis***

***Claim, H0*** = Average people frisked by year is independent of races

***H1*** = Average people frisked by year is dependent on the races

*# Chi-square test*

cs1 <- chisq.test(race\_mtrix)

*Table 5: Chi-squared test statistic*

| **statistic** | **p.value** | **parameter** | **method** |
| --- | --- | --- | --- |
| 1.465169 | 0.917054 | 5 | Pearson's Chi-squared test |

As the p-value is higher than the significant alpha level, we do not have enough evidence to reject the null hypothesis. Hence, here we conclude that average people getting frisked in the year 2011 and 2012 is not dependent over the races.

This also suggests that the petition that was filed against the NYPD stop and frisk policy in 2012 did made a significant difference in the pattern of frisk among the races.

***LASSO Regression***

# Creating new variables

nypd <- nypd %>% group\_by(year, pct, sex) %>% mutate(total\_stopped = sum(stopped), prop\_frisked = frisked/total\_stopped\*100)

nypd$prop\_frisked <- ifelse(is.na(nypd$prop\_frisked),0,nypd$prop\_frisked)

First, we used the pipe operator to group our dataset by year, precinct, and sex. Then, we piped in the mutate function and created two new variables, ‘total\_stopped’ and ‘prop\_frisked’. The first variable sums all the stops for a given year, precinct, and sex. The second variable takes the number of frisks by race and divides it by the total number of stops. Finding the proportion of frisks to total stops will help us determine if one race is getting frisked more than another. Next, we checked if there were any null values created in the ‘prop\_frisked’ column. If there were zero total stops for a given year, precinct, and sex, a zero divided by zero would result when calculating the frisked proportion. This could result in errors later. Next, we used the ‘regsubsets’ function to find the best five predictor linear model. We specified that ‘prop\_frisked’ was the outcome variable. This function recommended using the Black and Black-Hispanic races, ‘stopped’, ‘frisked’, and ‘cs\_descr’ as predictors. We chose to use the entire race variable instead of just two races.

# regsubsets feature selction

best\_subset <- regsubsets(prop\_frisked ~ ., data = nypd, nvmax = 5, really.big=T)

summary(best\_subset)

# basic linear regression

model1 <- lm(formula = prop\_frisked ~ race + stopped + frisked + cs\_descr, data = nypd)

summary(model1)

# fitting LASSO model

lasso.1se <- glmnet(train\_x, train\_y, alpha = 1, lambda = lasso$lambda.1se)

Shape

Description automatically generated

*Fig 5. LASSO model variable selection curve for minimum and 1se value*

Then, we ran an ordinary least square (OLS) linear regression with race and the other three variables mentioned. All coefficients and the p-value were significant. The Adjusted R-squared value was .64, which is not a strong value. The purpose of this basic linear model was to serve as a comparison to the least absolute shrinkage and selection operator (LASSO) model, which we run next. We checked the variance inflation factor (VIF) values for this basic linear model to determine if there was a problematic amount of multicollinearity. According to the article below called ‘Multicollinearity Essentials and VIF in R’, VIF values above 5 or 10 are considered to indicate a harmful amount of multicollinearity. The only value above five was for the ‘stopped’ predictor. Next, we created training and test sets with a seven to three split. We created model matrices as well, which serve as the inputs for the ‘cv.glmnet’ and ‘glmnet’ functions. Next, we ran the ‘glmnet’ cross validation function to determine the one standard error lambda value, which was .062. Subsequently, we fit a LASSO model with this lambda value. Next, our goal was to compare the root mean square error (RMSE) values for the basic linear model and our LASSO model. We made predictions on the training and test sets with these two models. Please find a table of these values below.

|  |  |  |
| --- | --- | --- |
| *Table 5: RMSE Values For Simple Linear Regression and LASSO* | | |
|  | **ols\_rmse** | **lasso\_rmse** |
| train predictions | 6.632 | 6.05 |
| test predictions | 6.404 | 5.789 |

The first row is for the predictions on the training set, and the second is for the test set. The first column is for the basic linear model, and the second is for the LASSO model. For both models, the test set RMSE values were slightly better. When comparing the two models, the RMSE values for the LASSO model were better than both values for the OLS model. The lasso model eliminated forty-six variables when all factors are listed out. The LASSO model helps balance under-fitting and over-fitting. It also provides a simple and interpretable model. Additionally, using the one standard error lambda value produces the most parsimonious model, which is a simple model with great predictive abilities.

**Determining the High or Low frisking rate using Logistic Regression**

As of now we have explored the dataset for the possibility of racial bias in the stop and frisk policy and performed Lasso regression to figure out the possible rate of frisk for a group of people. Now we are also interested to classify if for the particular group of people, the frisking rate is high or low. For this, we pre-classified our training data as ‘High’ frisking category where ratio is higher than 0.5 and ‘Low’ if its below.

***Confusion Matrix over the Testing Data:***

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Observed | |
|  |  | **0** | **1** |
| Prediction | **0** | 2349 | 397 |
| **1** | 130 | 727 |

Using the values obtained from the confusion matrix, we can observe that our logistic model has True Positive and True Negative values as 727 and 2349 respectively.

Also, the False positive and False negative values are 130 and 397 respectively.

Further, we have extracted the performance of the model by evaluating the accuracy, precision, recall as below for both training and testing dataset:

*Table 5: Logistic regression model metrics*

| **Metrics** | **Training\_set** | **Test\_set** |
| --- | --- | --- |
| Accuracy | 0.8596 | 0.8537 |
| Specificity | 0.6659 | 0.6468 |
| Precision | 0.8532 | 0.8554 |
| Recall | 0.9548 | 0.9476 |

From the above metrics we have observed that our model has an accuracy of 85.37% over the testing data which is decent enough to predict the ‘High’ or ‘Low’ frisking rate for a data. In addition to this, the model is capable of predicting the High Frisking rate around 94% if it’s actually a high frisking scenario. However, prediction of actual low frisking rate is lower as approximately 64% of times.

We also, tried to plot the Receiver Operator Characteristics curve for our model where we have received the Area under ROC (AUC) as **0.7971**.

Chart, line chart

Description automatically generated

*Fig 6. ROC curve for Logistic Regression model*

With the above analysis, we can conclude that based on the number of stops made and frisks happened, we can categorize the frisking rate of groups using the model to High and Low frisking rate. This will help us track in which precinct and race the rate is higher than the others. And how should the community deal with it.

**CONCLUSION**

* Analyzing the data set we found that the data suggested that the “Stop and Frisk” policy in New York City was majorly abused by profiling people of color.
* The total population of New York city was 8,175,133 according to United States Census Bureau among which 23.8% were Black of African American origin, 28.9% belong from Hispanic, while 41.3% were White.
* Although, Black being the minority population, but half of the population stopped at a precinct belonged from the African American origin.
* In the year 2013, with the election coming forth, the candidate was selected by the city because he stood on abrogating the “Stop and Frisk” policy.
* The policy with was mentioned as “racial incongruity” stop, with a committee who took over to analyze and present on court about the findings on the data.
* January 2014, when the policy was abrogated for biases of the officer based on the appearance of a person, in other words showing “racial disparity” and targeting the minority race such as Black African American and Hispanics.
* Using the ANOVA hypothesis test, we confirmed that there was a significant difference in the frisk to stop ratio among Black, Black-Hispanic, and White races.
* From the Chi-squared test results we verified that the frisking pattern that was being followed back in 2011 was changed in 2012 after the petition was filed against the policy.
* With the help of linear regression, we tried to predict the frisk proportion and we selected the optimum variables using the LASSO regularization resulting in a better model.
* And finally, to classify the high and low frisking rate, we first classified the training set based on more or less than 0.5 factor, then we utilized the logistic regression to classify the test data.
* The model is having approximately 85% accuracy with AUC as 0.7971

**REFERENCES**

* Multicollinearity Essentials and VIF in R. (2018, March 11). STHDA. Retrieved March 19, 2022, from http://www.sthda.com/english/articles/39-regression-model-diagnostics/160-multicollinearity-essentials-and-vif-in-r/
* Parsimonious Model: Definition, Ways to Compare Models. (2018, June 5). Statistics How To. Retrieved March 19, 2022, from https://www.statisticshowto.com/parsimonious-model/#:%7E:text=The%20most%20parsimonious%20model%20will,from%20that%20poor%2Dquality%20set.
* Bluman, A. G. (2018). Elementary statistics: A step by step approach. New York, NY: McGraw-Hill Education.
* Robert I, Kabacoff, (2015). “R in Action: Data Analysis and graphics with R”, Published by Manning Publications Co.
* *Publications, Reports - NYPD*. (2022). NYC. https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page

**APPENDIX**

Below is the R script file attached which is used for the above analysis purposes.

