NYPD Motor Vehicle Collisions Analysis

Table of Contents

1. Introduction

We live in an age where human mobility is growing exponentially and people are travelling daily for work, school, pleasure or other pursuits. This has led to the ever-growing numbers of vehicles on the roads which subsequently raises the concerns of more traffic accidents. According to a WHO report, 1.25 million (WHO, 2017) people die every year because of road traffic crashes.

As a part of this project we have analyzed the past one year's historical NYPD Motor vehicle collisions data and determined the leading factors resulting in vehicle collision in New York City. Based on our analysis, we have proposed data driven informed recommendations. Based on this NYPD will be able to understand the main pain points and focus areas and accordingly proactively formulate an action plan to reduce the rate of accidents in the future.

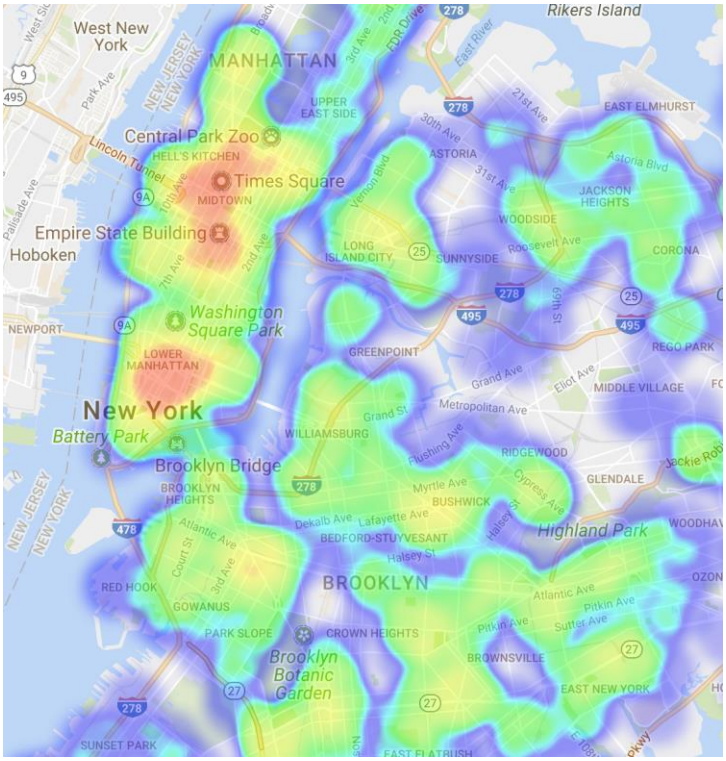
During this analysis we have attempted to ascertain if the accidents with deaths or injuries follow a certain trend, if there are certain areas that are more prone to accidents, if the accidents are concentrated during a certain time of the day so that NYPD can make informed decisions that can help them in reducing the number of accidents.

We utilized data from NYPD open data source. We initially only picked the past one year's data from 8/01/2016 to 7/31/2017, but after aggregation on the number of injuries and deaths we realized that this data is not substantial enough for an accurate prediction. As a result, we expanded the historical data used back to 07/01/2012. First, we analyzed the historical data and used descriptive analysis to get a big picture of data. Then we identified key factors causing accidents and recognized trends with time series and seasons. Subsequently, we did the multilinear with all the independent variables.

2. Project Background

In New York City, more than 200,000 motor vehicle collisions happen every year. This means about every 3 minutes a collision happens somewhere in New York City. To reduce collisions, we need to discover the key factors to help NYPD to improve the New York City road traffic conditions. Thanks to motor vehicle collision data provided by NYPD, the dataset contains detailed collisions information like: Date, Time, Borough, Location, contributing factor, Vehicle type and number of persons injured.

According to a heat map of the data (Zendrive, 2017), we noticed two areas of Manhattan seems to be the worst place in New York City. But the ongoing analysis will give us more meaningful information to help us understand the question and the way to make it better.



3. Research Questions

The following research questions have been tackled as a part of this project.

1. Is there a way to predict the traffic collisions before it happens?

2. Excluding the human factor is there a way to reduce the number of motor vehicle collisions by predicting it earlier?

Our goal is to find the statistically significant predictive model that can answer our main question and be verified by our hypothesis test.

4. Description of Data

The dataset used for this research study has been obtained from the official NYPD data base (NYPD, 2017) and consists of 1.1 million rows and 29 columns. This data set has the details of road accidents from 07/01/2012 to 08/01/2016. This source is reliable and updated on a regularly basis and the sample size is adequate enough the build a predictive model. Also, the data contains vast amount of variables that helped us to derive the relationships with the number of collisions. The following table shows the list of attributes and their type that are a part of this dataset. The format of the dataset was .csv.

DATE	Date & Time	NUMBER OF CYCLIST KILLED	Number
TIME	Plain Text	NUMBER OF MOTORIST INJURED	Number
BOROUGH	Plain Text	NUMBER OF MOTORIST KILLED	Number
ZIP CODE	Plain Text	CONTRIBUTING FACTOR VEHICLE 1	Plain Text
LATITUDE	Number	CONTRIBUTING FACTOR VEHICLE 2	Plain Text
LONGITUDE	Number	CONTRIBUTING FACTOR VEHICLE 3	Plain Text
LOCATION	Location	CONTRIBUTING FACTOR VEHICLE 4	Plain Text
ON STREET NAME	Plain Text	CONTRIBUTING FACTOR VEHICLE 5	Plain Text
CROSS STREET NAME	Plain Text	UNIQUE KEY	Number
OFF STREET NAME	Plain Text	VEHICLE TYPE CODE 1	Plain Text
NUMBER OF PERSONS INJURED	Number	VEHICLE TYPE CODE 2	Plain Text
NUMBER OF PERSONS KILLED	Number	VEHICLE TYPE CODE 3	Plain Text
NUMBER OF PEDESTRIANS INJURED	Number	VEHICLE TYPE CODE 4	Plain Text
NUMBER OF PEDESTRIANS KILLED	Number	VEHICLE TYPE CODE 5	Plain Text
NUMBER OF CYCLIST INJURED	Number		

5. Testing and Analysis

5.1 Exploratory Data Analysis

The first step of this research was to format the data and upload it to R-Studio. There were a few missing fields in the data and they were removed during the formatting phase. After uploading the data, the variables were observed using the str () function. This was a means to understand the type of variables in the dataset.

We then obtained the descriptive statistics of all explanatory and dependent variables to get a more holistic picture for the dataset. This helped us direct our analysis to focus on variables showing significant patterns that were further exhibited through histograms.

First load data into R

```{r}
df<-read.csv("C:/USERS/KUN HU/DESKTOP/NYPD.csv",stringsAsFactors=F,header=T)
str(df)
```

```
'data.frame':   1089265 obs. of  29 variables:
 $ DATE                          : chr  "08/04/2017" "08/04/2017" "08/04/2017" "08/04/2017" ...
 $ TIME                          : chr  "0:00" "0:00" "0:00" "0:00" ...
 $ BOROUGH                       : chr  "QUEENS" "" "" "" ...
 $ ZIP.CODE                      : int  11436 NA NA NA NA NA NA 10460 11101 10001 ...
 $ LATITUDE                      : num  40.7 40.7 40.7 40.8 40.8 ...
 $ LONGITUDE                     : num  -73.8 -74 -74 -74 -73.8 ...
 $ LOCATION                      : chr  "(40.666885, -73.790405)" "(40.71995, -74.00859)" "(40.718666, -73.9635)" "(40.754677, -73.975815)"
...
 $ ON.STREET.NAME                : chr  "NORTH CONDUIT AVENUE            " "HUDSON STREET                    " "KENT AVENUE
" "PARK AVENUE                    " ...
 $ CROSS.STREET.NAME             : chr  "149 STREET" "" "" "" ...
 $ OFF.STREET.NAME               : chr  "" "" "" "" ...
 $ NUMBER.OF.PERSONS.INJURED     : int  0 0 0 0 0 0 1 2 0 0 ...
 $ NUMBER.OF.PERSONS.KILLED      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NUMBER.OF.PEDESTRIANS.INJURED: int  0 0 0 0 0 0 0 2 0 0 ...
 $ NUMBER.OF.PEDESTRIANS.KILLED : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NUMBER.OF.CYCLIST.INJURED     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NUMBER.OF.CYCLIST.KILLED      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NUMBER.OF.MOTORIST.INJURED    : int  0 0 0 0 0 0 1 0 0 0 ...
 $ NUMBER.OF.MOTORIST.KILLED     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CONTRIBUTING.FACTOR.VEHICLE.1: chr  "Unspecified" "Unsafe Lane Changing" "Turning Improperly" "Unspecified" ...
 $ CONTRIBUTING.FACTOR.VEHICLE.2: chr  "Unspecified" "Unsafe Lane Changing" "Following Too Closely" "Unspecified" ...
 $ CONTRIBUTING.FACTOR.VEHICLE.3: chr  "" "" "" "" ...
 $ CONTRIBUTING.FACTOR.VEHICLE.4: chr  "" "" "" "" ...
 $ CONTRIBUTING.FACTOR.VEHICLE.5: chr  "" "" "" "" ...
 $ UNIQUE.KEY                    : int  3725017 3725047 3725533 3724870 3725662 3725120 3725209 3725174 3724308 3725543 ...
 $ VEHICLE.TYPE.CODE.1           : chr  "PASSENGER VEHICLE" "PICK-UP TRUCK" "SPORT UTILITY / STATION WAGON" "SPORT UTILITY / STATION WAGON"
...
 $ VEHICLE.TYPE.CODE.2           : chr  "PASSENGER VEHICLE" "SPORT UTILITY / STATION WAGON" "PASSENGER VEHICLE" "TAXI" ...
 $ VEHICLE.TYPE.CODE.3           : chr  "" "" "" "" ...
 $ VEHICLE.TYPE.CODE.4           : chr  "" "" "" "" ...
 $ VEHICLE.TYPE.CODE.5           : chr  "" "" "" "" ...
```

```{r}
summary(df)
```

```
     DATE               TIME             BOROUGH            ZIP.CODE        LATITUDE        LONGITUDE         LOCATION
 Length:1089265     Length:1089265     Length:1089265     Min.   :10000   Min.   : 0.00   Min.   :-201.36   Length:1089265
 Class :character   Class :character   Class :character   1st Qu.:10128   1st Qu.:40.67   1st Qu.: -73.98   Class :character
 Mode  :character   Mode  :character   Mode  :character   Median :11205   Median :40.72   Median : -73.93   Mode  :character
                                                          Mean   :10811   Mean   :40.72   Mean   : -73.92
                                                          3rd Qu.:11236   3rd Qu.:40.77   3rd Qu.: -73.87
                                                          Max.   :11697   Max.   :41.13   Max.   :   0.00
                                                          NA's   :297136  NA's   :207066  NA's   :207066
 ON.STREET.NAME     CROSS.STREET.NAME  OFF.STREET.NAME    NUMBER.OF.PERSONS.INJURED NUMBER.OF.PERSONS.KILLED
 Length:1089265     Length:1089265     Length:1089265     Min.   : 0.0000           Min.   :0.000000
 Class :character   Class :character   Class :character   1st Qu.: 0.0000           1st Qu.:0.000000
 Mode  :character   Mode  :character   Mode  :character   Median : 0.0000           Median :0.000000
                                                          Mean   : 0.2556           Mean   :0.001198
                                                          3rd Qu.: 0.0000           3rd Qu.:0.000000
                                                          Max.   :43.0000           Max.   :5.000000

 NUMBER.OF.PEDESTRIANS.INJURED NUMBER.OF.PEDESTRIANS.KILLED NUMBER.OF.CYCLIST.INJURED NUMBER.OF.CYCLIST.KILLED
 Min.   : 0.00000              Min.   :0.0000000            Min.   :0.00000           Min.   :0.0e+00
 1st Qu.: 0.00000              1st Qu.:0.0000000            1st Qu.:0.00000           1st Qu.:0.0e+00
 Median : 0.00000              Median :0.0000000            Median :0.00000           Median :0.0e+00
 Mean   : 0.05247              Mean   :0.0006647            Mean   :0.02047           Mean   :7.8e-05
 3rd Qu.: 0.00000              3rd Qu.:0.0000000            3rd Qu.:0.00000           3rd Qu.:0.0e+00
 Max.   :28.00000              Max.   :2.0000000            Max.   :4.00000           Max.   :1.0e+00

 NUMBER.OF.MOTORIST.INJURED NUMBER.OF.MOTORIST.KILLED CONTRIBUTING.FACTOR.VEHICLE.1 CONTRIBUTING.FACTOR.VEHICLE.2
 Min.   : 0.0000            Min.   :0.000000          Length:1089265                 Length:1089265
 1st Qu.: 0.0000            1st Qu.:0.000000          Class :character               Class :character
 Median : 0.0000            Median :0.000000          Mode  :character               Mode  :character
 Mean   : 0.1862            Mean   :0.000454
 3rd Qu.: 0.0000            3rd Qu.:0.000000
 Max.   :43.0000            Max.   :5.000000

 CONTRIBUTING.FACTOR.VEHICLE.3 CONTRIBUTING.FACTOR.VEHICLE.4 CONTRIBUTING.FACTOR.VEHICLE.5   UNIQUE.KEY      VEHICLE.TYPE.CODE.1
 Length:1089265                Length:1089265                 Length:1089265                 Min.   :     22   Length:1089265
 Class :character              Class :character               Class :character               1st Qu.: 274242   Class :character
 Mode  :character              Mode  :character               Mode  :character               Median :3180753   Mode  :character
                                                                                             Mean   :2200695
                                                                                             3rd Qu.:3453075
                                                                                             Max.   :3726256

 VEHICLE.TYPE.CODE.2 VEHICLE.TYPE.CODE.3 VEHICLE.TYPE.CODE.4 VEHICLE.TYPE.CODE.5
 Length:1089265      Length:1089265      Length:1089265      Length:1089265
 Class :character    Class :character    Class :character    Class :character
 Mode  :character    Mode  :character    Mode  :character    Mode  :character
```

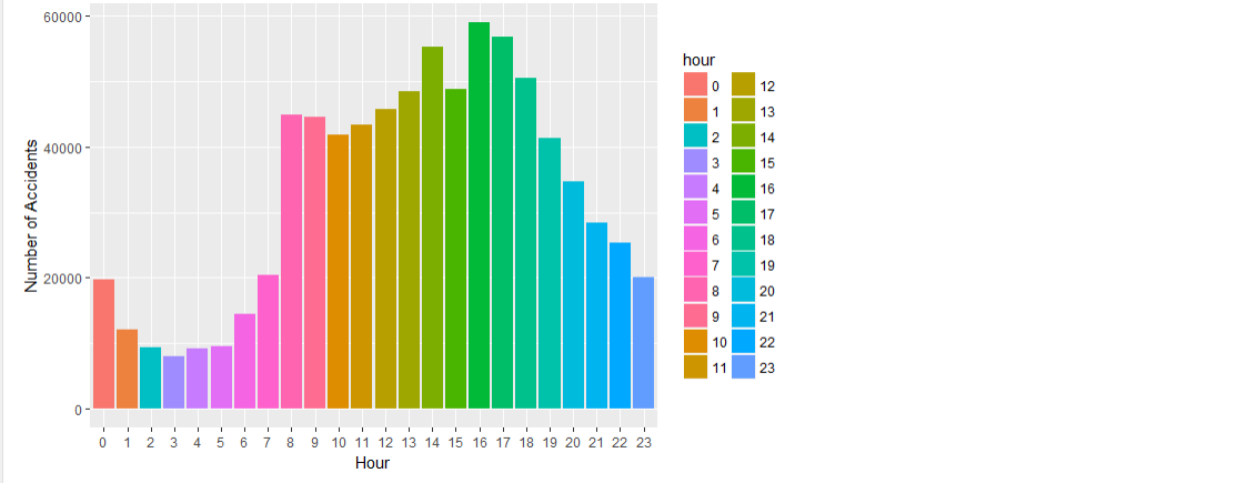
We created histograms for all these variables to better view their impact on predicting traffic collisions. The histogram for collision type showed a lot of skewness based on which we did not choose it as one of

the dependent variables. Accordingly, based on the variables capturing the significant variations in the data set, we identified the independent variables as follows:

1. Time of day
2. Month of the year
3. Borough
4. Day of the week
5. Year

Time of the Day

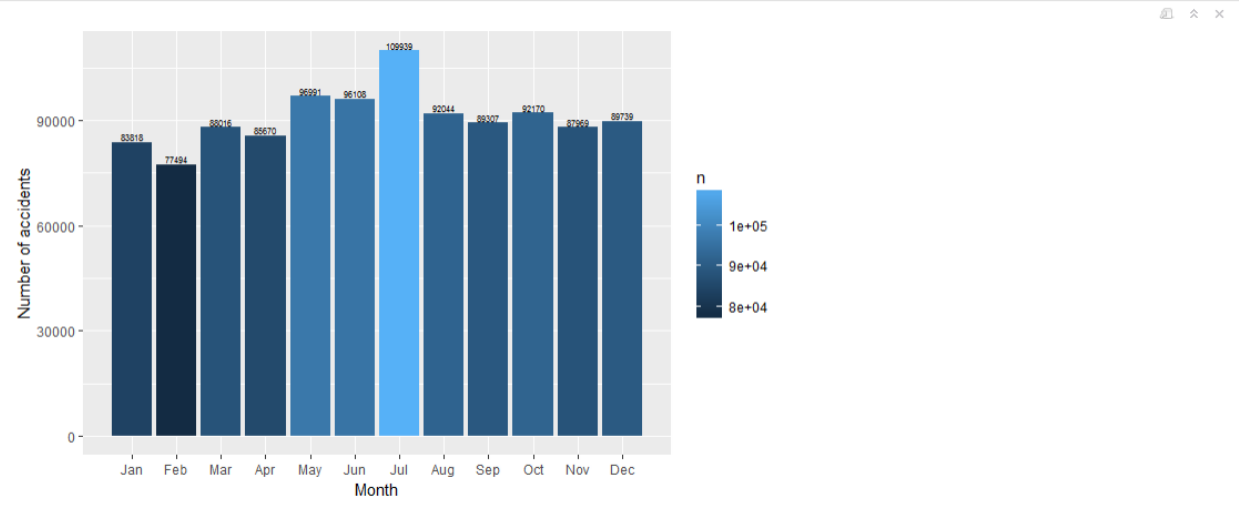
```r
df %>% filter(BOROUGH!="") %>% group_by(hour) %>% summarise(n=sum(n())) %>% na.omit() %>%
ggplot(aes(x=hour,y=n,fill=hour))+geom_bar(position="dodge",stat="identity")+scale_x_discrete(limits=c("0","1","2","3","4","5","6","7","8","9","10","11","12","13","14","15","16","17","18","19","20","21","22","23"))+scale_fill_discrete(breaks=c("0","1","2","3","4","5","6","7","8","9","10","11","12","13","14","15","16","17","18","19","20","21","22","23"))+xlab("Hour")+ylab("Number of Accidents")
```



Month of the year

Number of collisions per month

```{r}
df %>% group_by(month) %>% summarise(n=sum(n())) %>%  ggplot(aes(x=month,y=n,fill=n))+geom_bar(position="dodge",stat =
"identity")+scale_x_discrete(limits=c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec"))+xlab("Month")+ylab("Number
of accidents")+geom_text(aes(label=n),
position=position_dodge(1),vjust=-0.2,size=2)
```

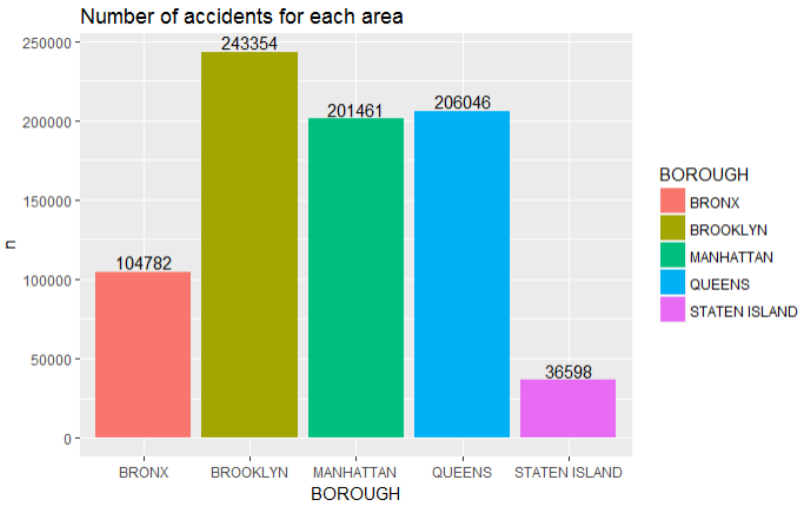


Borough

Number of accidents for each area

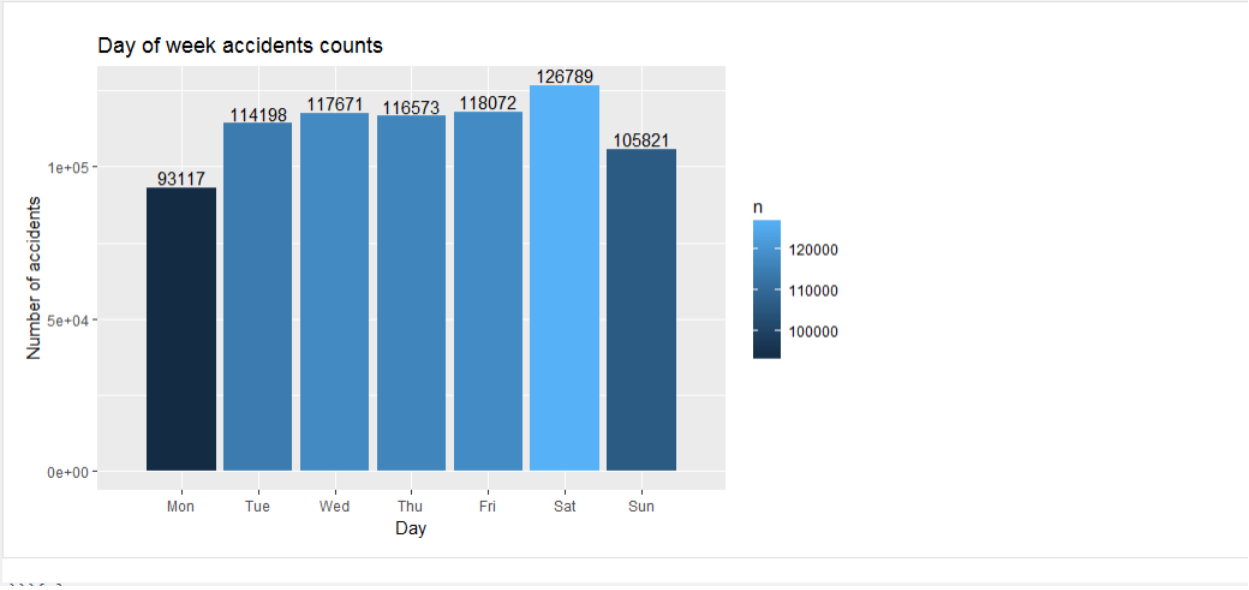
```{r}
df %>% filter(BOROUGH!="") %>%group_by(BOROUGH) %>% summarise(n=sum(n())) %>%
ggplot(aes(x=BOROUGH,y=n,fill=BOROUGH))+geom_bar(position="dodge",stat = "identity")+ggtitle("Number of accidents for each
area")+geom_text(aes(label=n),position=position_dodge(1),vjust=-0.2,size=4)

```



Day of the week

```r
df %>% filter(BOROUGH!="") %>% group_by(day) %>% summarise(n=sum(n())) %>%
  ggplot(aes(x=day, y=n,fill=n)) +
geom_bar(position="dodge",stat = "identity")+geom_text(aes(label=n),
position=position_dodge(1),vjust=-0.2,size=4)+ggtitle("Day of week accidents
counts")+scale_x_discrete(limits=c("Mon","Tue","Wed","Thu","Fri","Sat","Sun"))+ylab("Number of accidents")+xlab("Day")
```



5.2 Best Fit Model

We then divided the data set into two parts with similar Mean, STD and CI.

1. Test Population – 80% of the data, used for model creation
2. Holdout Population – 20% of the data, used for model validity

As all the independent variables in our case are categorical, we created dummy variables to capture the relationship between them and the dependent variable. Considering a total of K variables, we created K-1 dummy variables. The description of the dummy variables is below.

1. Year – As there are 6 years in our dataset, we created 5 (6-1) dummy variables namely y1, y2, y3, y4 and y5 for the years 2012 to 2016.
2. Month – As there are 12 months in a year, we created 11 (12-1) dummy variables m1, m2, m3, m4, m5, m6, m7, m8, m9, m10, m11 for the months February to December.
3. Borough – As there are 5 boroughs, we created 4 (5-1) dummy variables namely b1, b2, b3 and b4.
4. Day – As there are 7 days in a week, we created 6 (7-1) dummy variables d1, d2, d3, d4, d5 and d6 for the days Monday to Saturday.
5. Hour – As there are 24 hours in a day, we created 23 (24-1) dummy variables from h1 to h23 from 1:00 to 23:59.

We created Train and Test Data for conducting first level multilinear regression. Based on this we calculated the estimated test predicted values and the difference between these values and the actual values. Then we used the following equation to generate next levels of regression equations to finally achieve our optimized regression model.

Considering we had 49 dummy variables, we decided to run a step function to drop statistically insignificant independent variables. We ran this function both step forward and step backwards to eliminate as many variables based on t-sat and p-value of t-stat with significance level of 0.05. It was determined that the backward step function was more effective as it had lesser number of variables. The independent variables with mod T-stat value of less than 1 and p-value greater than 0.05 were considered statistically insignificant and thus dropped from one run to the next. We ran 4 runs and reduced the dummy variables from 49 in the first run to 45 in the second, 43 in the third and finally 40 dummy variables in the final run.

The result of the final run for the multilinear regression model is depicted below. The R-Squared value 68.74% with Adjusted R-Squared being 68.71%. Basically, our model could explain 68.74% variation in the dependent variables i.e. number of accidents based on our independent variables.

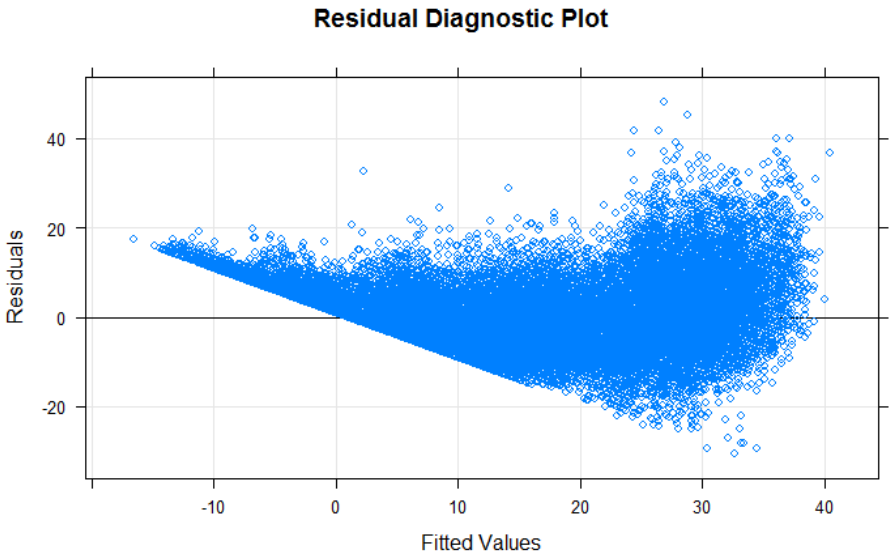
```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.289 on 39677 degrees of freedom
Multiple R-squared:  0.6874,    Adjusted R-squared:  0.6871
F-statistic:  2181 on 40 and 39677 DF,  p-value: < 2.2e-16
```

Considering there were 40 variables, the final regression equation obtained as a result of this analysis was complex and is as follows:

No. of Accidents = 1.46353 + 13.48537b1 +9.42171b2 +9.95945b3 -7.52846b4 +2.81701d1 +3.12942d2 +2.94808d3 +3.12335d4 +4.40326d5 +1.72549d6 -4.11946h1 -5.61456h2 -6.36711h3 -5.86049h4 -5.63161h5 -2.81339h6 +11.83022h8 +11.54842h9 +10.25786h10 +11.05205h11 +12.01617h12 +13.36296h13 +16.54731h14 +13.35255h15 +18.42619h16 +17.33349h17 +14.42466h18 +10.01112h19 +6.9282h20 +4.0239h21 +2.61902h22 -1.65294m1 +0.47578m2 +1.79823m4 +1.36904m5 +0.64073m6 -0.92347m7 + 0.64894m9 + 0.79644y3 -2.39983y5

We then created a Residual vs Fit Diagnostic plot to check our model and the spread of observations. Some portion of the residuals shows linear pattern, but majority (right hand side) seems random and without heteroscedasticity problem. There are some issues but they can be primarily attributed to collinearity between month year and days. The model can still be considered a good representation considering the diversity of variables considered. Also, it is homoscedastic around the mean and we see variations equally.



5.3 Model Testing and Results

Confidence Interval

We then tested our model on the test data to ascertain its validity. We used the equation to compute estimated values on the test holdout population. We then computed the difference between actual and estimated values. Finally, we then calculated the 95% Confidence Interval of the mean of the difference in actual and estimated values which was -0.22 and 0.06. Since, the confidence interval includes 0 as one of the values, it indicates that our model does a good job estimating the total number of accidents.

We then conduct a hypothesis test at 0.05 significance level to see if the mean is significantly different than 0.

Hypothesis Testing

We also conducted hypothesis testing to see if the mean is significantly different than zero. Our hypothesis was formulated as follows:

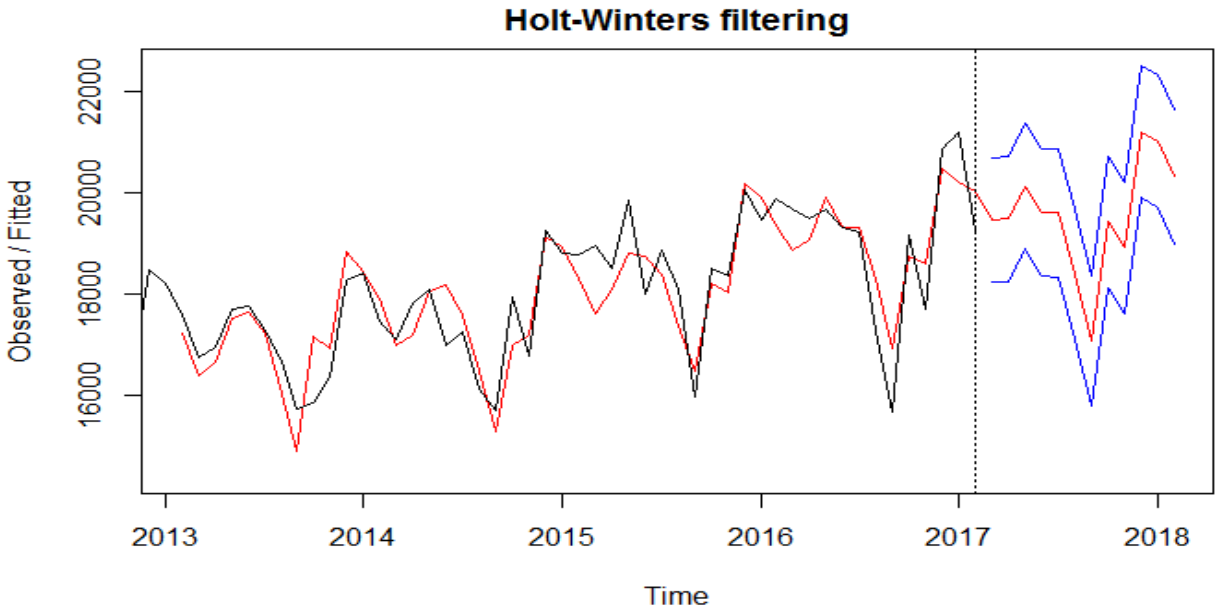
H_0: Mean of the test data = 0

H_1: Mean of the test data <> 0

We obtained a p-value of 0.14 which is greater than 0.05 hence we failed to reject the null hypothesis. Thus, the mean of the difference between estimated and actual values is 0. In other words, mean of estimated values is equal to the mean of actual values

5.4 Forecasting

We had seasonality in our data so we used Holt-Winters filtering to arrive at some predictions about the occurrence of accidents.Below is the prediction of number of accidents from Mar 2017- Feb 2018



	fit	upr	lwr
Mar 2017	19468.88	20706.24	18231.53
Apr 2017	19486.14	20731.89	18240.39
May 2017	20139.18	21393.26	18885.10
Jun 2017	19617.25	20879.61	18354.88
Jul 2017	19606.91	20877.50	18336.33
Aug 2017	18385.48	19664.24	17106.72
Sep 2017	17072.47	18359.35	15785.59
Oct 2017	19429.34	20724.29	18134.39
Nov 2017	18918.63	20221.60	17615.66
Dec 2017	21214.15	22525.10	19903.21
Jan 2018	21016.62	22335.49	19697.76
Feb 2018	20329.51	21656.25	19002.77

6. Conclusion

The model created in this research gives us valuable takeaways regarding the relationship between the number of accidents with location and date.

When it comes to the boroughs, Brooklyn has the strongest positive linear relationship with the number of collisions whereas Staten Island is the only borough that has a negative linear relationship with the number of collisions. Subsequently when we consider the time of the day, the hours between 1am and 6am have a negative linear relationship with the number of collisions. After 6am all the hours have a positive linear relationship with the number of collisions with the peak being between 4 to 5pm. Moving on, days have positive linear relationship with the number of collisions with number peaking on Friday and lowest point occurring on Saturday. Finally, looking at the months, majority of them have small and negligible positive linear relationship with the number of collisions. February and August are the only months with a negative linear relationship with the number of collisions.

Using this model, NYPD can develop strategies to deploy officers on roads to control traffic more efficiently. Another use case can be creating route scheduler applications, such as mapping services to provide effective route navigation to drivers away from accident prone areas and times of the day.

Overall, given the data available, we believe that the model is robust and incorporates some key factors that can help in predicting the likelihood of accidents in New York City. This work can certainly be built on by including other significant independent variables such as historical weather data, construction activity, ongoing road work data. This can help improve the adjusted R-square which in turn means an improvement in the regression model.

7. References

1. NYPD. 2017. Retrieved from https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95

2. WHO.2017. Retrieved from http://www.who.int/mediacentre/factsheets/fs358/en/

3. Zendrive, NYU. 2017. Retrieved from https://s3-us-west-2.amazonaws.com/zd-website-assets/casestudies/NYU+Zendrive+data+review+-+May2017.pdf