Introduction

The main point of this project was to explore New York City's 5-car-accidents and figure out the who, what, where and why by using R. I used several new libraries, and the main focus of this product was data cleaning; there were several transformations that I have not done before that I found interesting to explore. I also visualized much of the data.

The main questions I intended to answer were:

1. Where are 5 car accidents occurring in NYC?
2. Who, under which circumstances, are in accidents?
3. What kind of cars are wrecking the most?
4. When (which season, Summer/Winter) has more crashes?
5. Why are crashes happening the way they are? Can a predictive model be formed from some attributes and is the model significant?

The following were my hypotheses for the questions above:

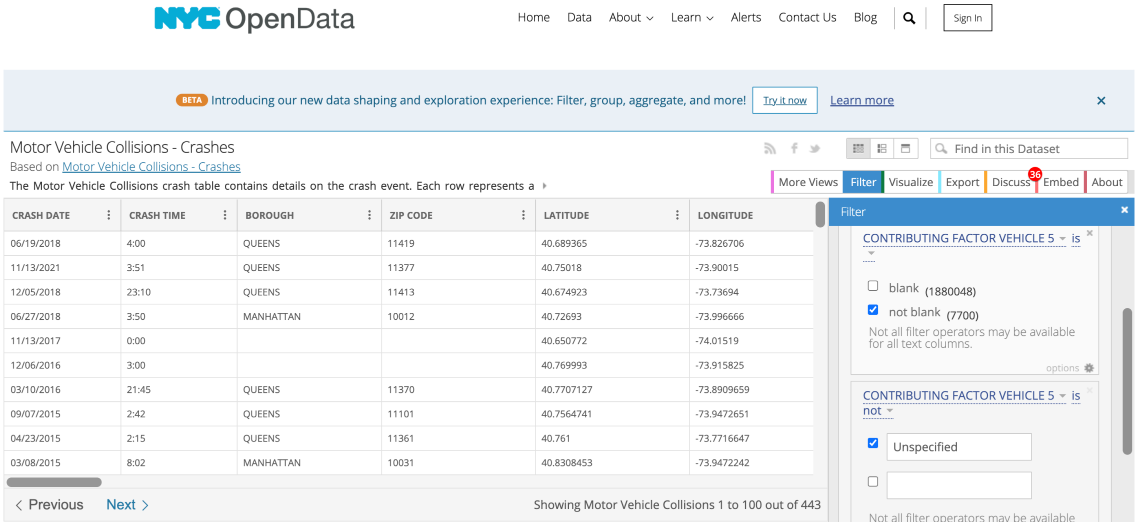
1. Roads leading in/out of NYC
2. Impaired drivers, usually the drivers' faults
3. Taxis/Ubers
4. Winter because of slippery roads
5. Impaired drivers in the winter that are going on road-trips are the most at risk of an accident because of these characteristics. I think a predictive model can be formed, but it will be very weak given the low sample size.

The following are my conclusions for the questions above:

1. 5-car-accidents occur most in south-central Manhattan, on average at city hall
2. People that drive too quickly are most at risk for 5-car-accidents
3. Sedans - not taxis – account for the most accidents
4. Winters are when the most accidents are
5. Predictive models from this dataset are weak- there need to be more observations or better variables

Brief Data Description

This section describes the data, where it came from, and some brief information to help paint a full picture of the dataset before it was transformed.



Introducing our new data shaping and exploration experience: Filter, group, aggregate, and more!

Motor Vehicle Collisions - Crashes

Based on Motor Vehicle Collisions - Crashes

The Motor Vehicle Collisions crash table contains details on the crash event. Each row represents a

CRASH DATE	CRASH TIME	BOROUGH	ZIP CODE	LATITUDE	LONGITUDE
06/19/2018	4:00	QUEENS	11419	40.689365	-73.826706
11/13/2021	3:51	QUEENS	11377	40.75018	-73.90015
12/05/2018	23:10	QUEENS	11413	40.674923	-73.73694
06/27/2018	3:50	MANHATTAN	10012	40.72693	-73.996666
11/13/2017	0:00			40.650772	-74.01519
12/06/2016	3:00			40.769993	-73.915825
03/10/2016	21:45	QUEENS	11370	40.7707127	-73.8909659
09/07/2015	2:42	QUEENS	11101	40.7564741	-73.9472651
04/23/2015	2:15	QUEENS	11361	40.761	-73.7716647
03/08/2015	8:02	MANHATTAN	10031	40.8308453	-73.9472242

Showing Motor Vehicle Collisions 1 to 100 out of 443

Filter

CONTRIBUTING FACTOR VEHICLE 5 is

□ blank (1880048)
☑ not blank (7700)

Not all filter operators may be available for all text columns.

CONTRIBUTING FACTOR VEHICLE 5 is not

☑ Unspecified
□

Not all filter operators may be available

- Data was sourced from: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/data
- n = 443 instances of 5 car crashes

- N = 1,887,748 crashes in total
- 29 variables, mostly categorical
- No deaths in 5 car accidents in NYC for entirety of dataset

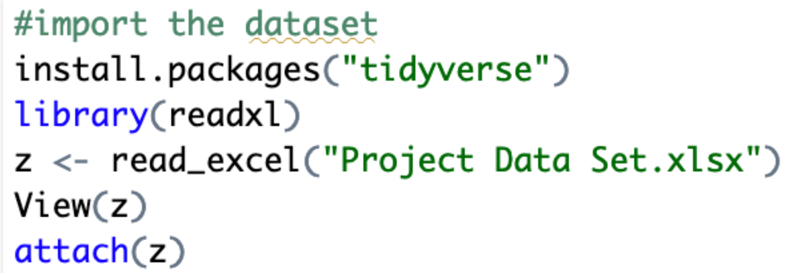
Analysis/Results

The following explains what I did throughout this project and key results.

The two new libraries that I used are:

- StringR to convert text to proper case
- Lubridate (to break down the dates into just the month) for the first time.

1. First, I imported the dataset. I saved the dataset as "z" to make the early transformations easy (since all I would need to do is type in z$ and then the column I need to change). In retrospect, I should have chosen df so that I could have ran dataframe related functions from Stack Overflow to see how they work.

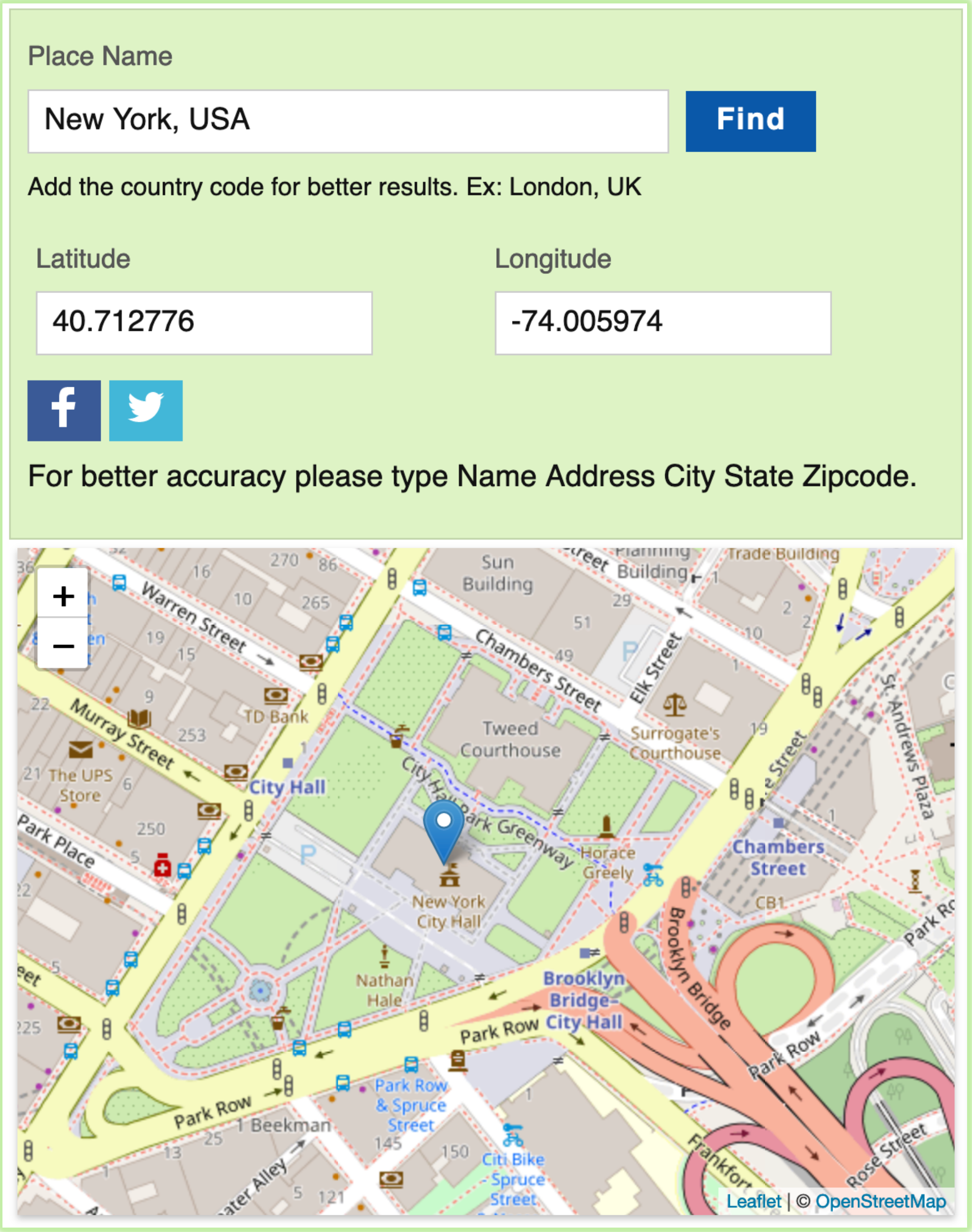


```
#import the dataset
install.packages("tidyverse")
library(readxl)
z <- read_excel("Project Data Set.xlsx")
View(z)
attach(z)
```

2. I then replaced all N/A values so that there are no nulls. This allows for predictive modeling to take place and is overall the initial goal that I had to make this dataset usable.

- I replaced the null zip codes with 0, since this would indicate no zip code at all while still allowing zip codes to be used for predictive modeling. Replacing the nulls with an average would not make sense.
- I replaced all null street names with ‘UNSPECIFIED.’ This follows the format of other categorical variables in this dataset, and allows for nulls to be represented in charts without creating errors.
- I then converted all text columns to proper case using stringr. This makes the text more homogenous across the board and more legible. The str_to_title() function was quite simple to implement after reading its documentation. Prior to using this function, some text was in ALL CAPS, some text was in Proper Case, and some text was in all lowercase. Having all text be in the same format makes the data look cleaner overall.
- Finally, I imputed null latitudes and longitudes with the averages of all non-null values. These were 40.74 and -73.91 for latitudes and longitudes, respectively. These values were taken from Excel since the mean() function in excel can function while nulls are present (unlike R without any other functions). To keep things simple, I took the mean of each column in excel and copied the answer over to R.

3. I then decided to google the average latitude and longitude. As it turns out, the average latitude and longitude of 5-car-accidents in New York City is directly over NYC City Hall (in south-central Manhattan). This is an interesting coincidence, and may suggest that the roads in NYC all revolve around City Hall, since most accidents occur somewhat close to this area. Below shows what came up:



Place Name

New York, USA

Find

Add the country code for better results. Ex: London, UK

Latitude

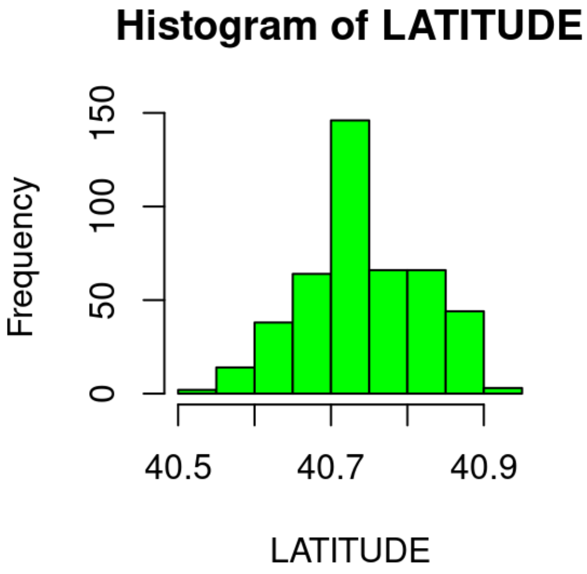
40.712776

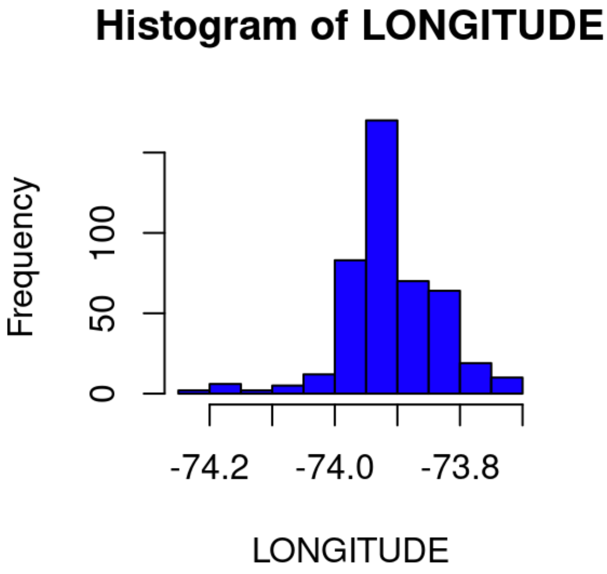
Longitude

-74.005974

For better accuracy please type Name Address City State Zipcode.

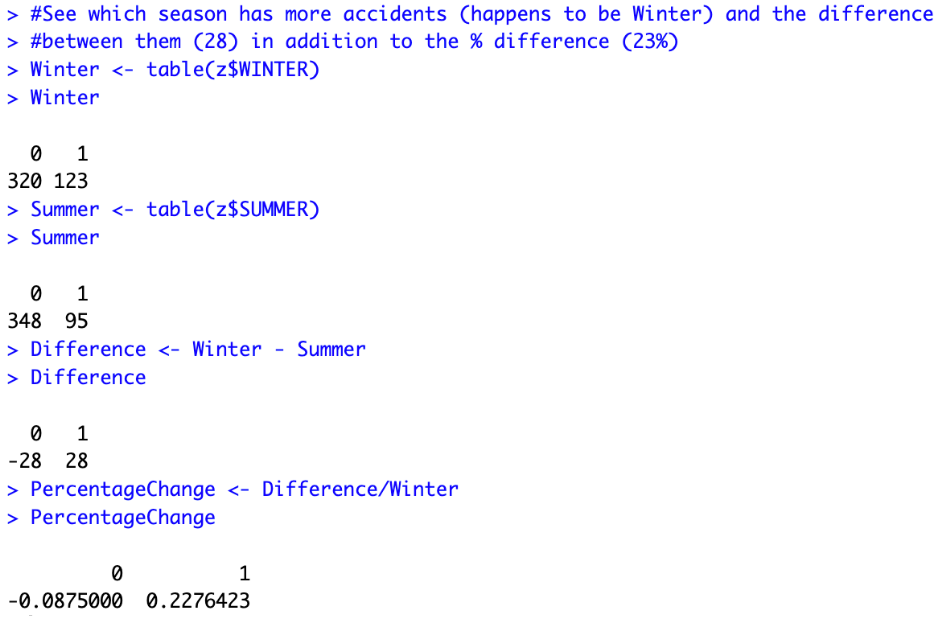
4. After seeing the location of the average 5-car-crash, I decided to picture the distribution of latitudes and longitudes using histograms. These histograms show normal distributions following gaussian functions. The average’s frequency is overrepresented, however, since all nulls were imputed with the mean. The rest of the distribution still reflects gaussian tendencies with both central-eastward and central-southern tendencies, which shows that most observations occur in South-Central Manhattan.



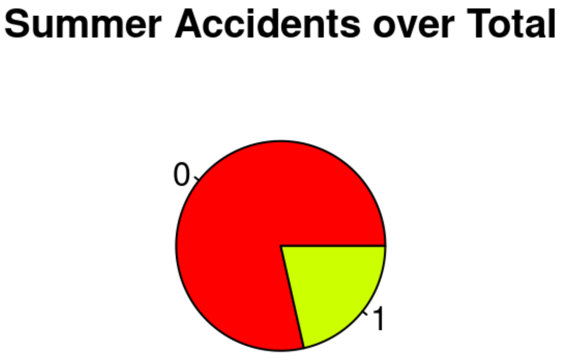


5. After this I decided to create dummy variables for accidents being in the Winter and the Summer. To do this I used the month() function and a new lubridate library function that I hadn’t used called as.POSIXlt() which made it possible to extract the ‘Date’ column into just a column indicating the month (from 1 to 12). I made two separate variables by dummy coding all observations that indicated either Summer or Winter (for the 2 new columns) as 1 and every other month as 0. This set up these 2 new columns for my next stage of analysis.

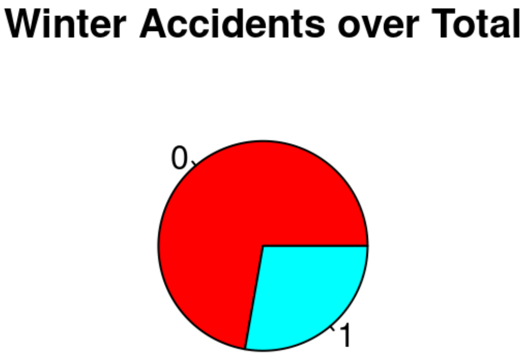
6. After creating the 2 dummy variables, I sought to compare the 2. I created tables of each that counted the instances of each. I then subtracted the summer values from the winter values into a difference table (28). Finally, I divided the winter table by the difference table to result in the percentage change (23%). I then visualized both Summer and Winter in pie charts to demonstrate these differences.



```
> #See which season has more accidents (happens to be Winter) and the difference
> #between them (28) in addition to the % difference (23%)
> Winter <- table(z$WINTER)
> Winter

  0   1 
320 123 
> Summer <- table(z$SUMMER)
> Summer

  0   1 
348  95 
> Difference <- Winter - Summer
> Difference

  0   1 
-28  28 
> PercentageChange <- Difference/Winter
> PercentageChange

         0          1 
-0.0875000  0.2276423 
```

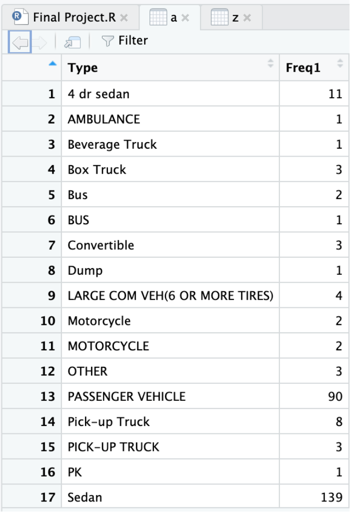


Summer Accidents over Total



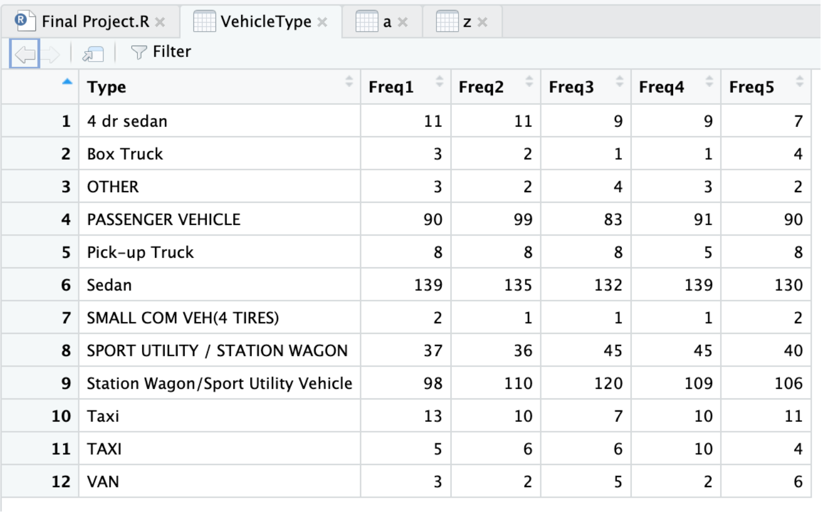
Winter Accidents over Total

7. I then wanted to dissect the frequency with which each vehicle type got into 5-car-accidents. This was difficult considering there were 5 separate columns that described 5 vehicles for each row, so there were 2,215 values for 443 observations. To get around this, I had to create tables of each column's values, merge them into one table, and then take the sum of each vehicle's frequency into a calculated column. I used the table(), rename() and merge() functions to accomplish this. I also dropped the unneeded columns after they were summed, and converted the frequency to be numeric. Below demonstrates the process flow that I am describing here:



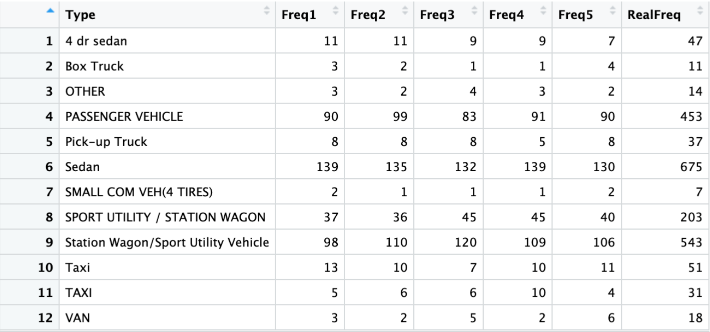
	Type	Freq1
1	4 dr sedan	11
2	AMBULANCE	1
3	Beverage Truck	1
4	Box Truck	3
5	Bus	2
6	BUS	1
7	Convertible	3
8	Dump	1
9	LARGE COM VEH(6 OR MORE TIRES)	4
10	Motorcycle	2
11	MOTORCYCLE	2
12	OTHER	3
13	PASSENGER VEHICLE	90
14	Pick-up Truck	8
15	PICK-UP TRUCK	3
16	PK	1
17	Sedan	139

→



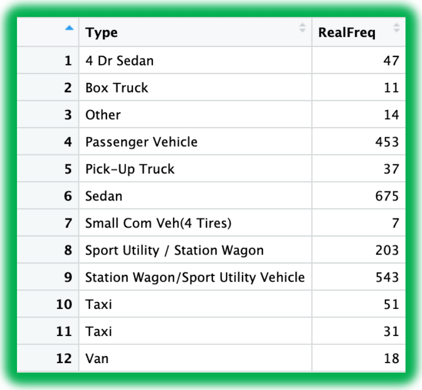
	Type	Freq1	Freq2	Freq3	Freq4	Freq5
1	4 dr sedan	11	11	9	9	7
2	Box Truck	3	2	1	1	4
3	OTHER	3	2	4	3	2
4	PASSENGER VEHICLE	90	99	83	91	90
5	Pick-up Truck	8	8	8	5	8
6	Sedan	139	135	132	139	130
7	SMALL COM VEH(4 TIRES)	2	1	1	1	2
8	SPORT UTILITY / STATION WAGON	37	36	45	45	40
9	Station Wagon/Sport Utility Vehicle	98	110	120	109	106
10	Taxi	13	10	7	10	11
11	TAXI	5	6	6	10	4
12	VAN	3	2	5	2	6

→



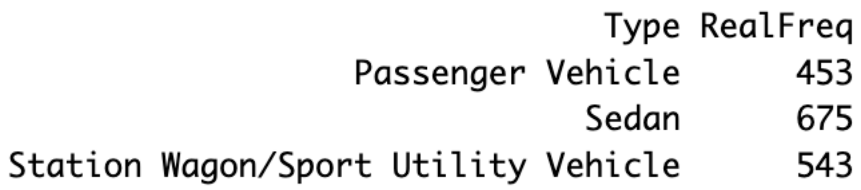
	Type	Freq1	Freq2	Freq3	Freq4	Freq5	RealFreq
1	4 dr sedan	11	11	9	9	7	47
2	Box Truck	3	2	1	1	4	11
3	OTHER	3	2	4	3	2	14
4	PASSENGER VEHICLE	90	99	83	91	90	453
5	Pick-up Truck	8	8	8	5	8	37
6	Sedan	139	135	132	139	130	675
7	SMALL COM VEH(4 TIRES)	2	1	1	1	2	7
8	SPORT UTILITY / STATION WAGON	37	36	45	45	40	203
9	Station Wagon/Sport Utility Vehicle	98	110	120	109	106	543
10	Taxi	13	10	7	10	11	51
11	TAXI	5	6	6	10	4	31
12	VAN	3	2	5	2	6	18

→

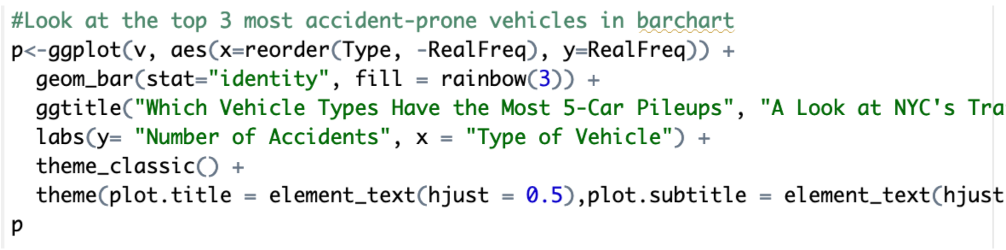


	Type	RealFreq
1	4 Dr Sedan	47
2	Box Truck	11
3	Other	14
4	Passenger Vehicle	453
5	Pick-Up Truck	37
6	Sedan	675
7	Small Com Veh(4 Tires)	7
8	Sport Utility / Station Wagon	203
9	Station Wagon/Sport Utility Vehicle	543
10	Taxi	51
11	Taxi	31
12	Van	18

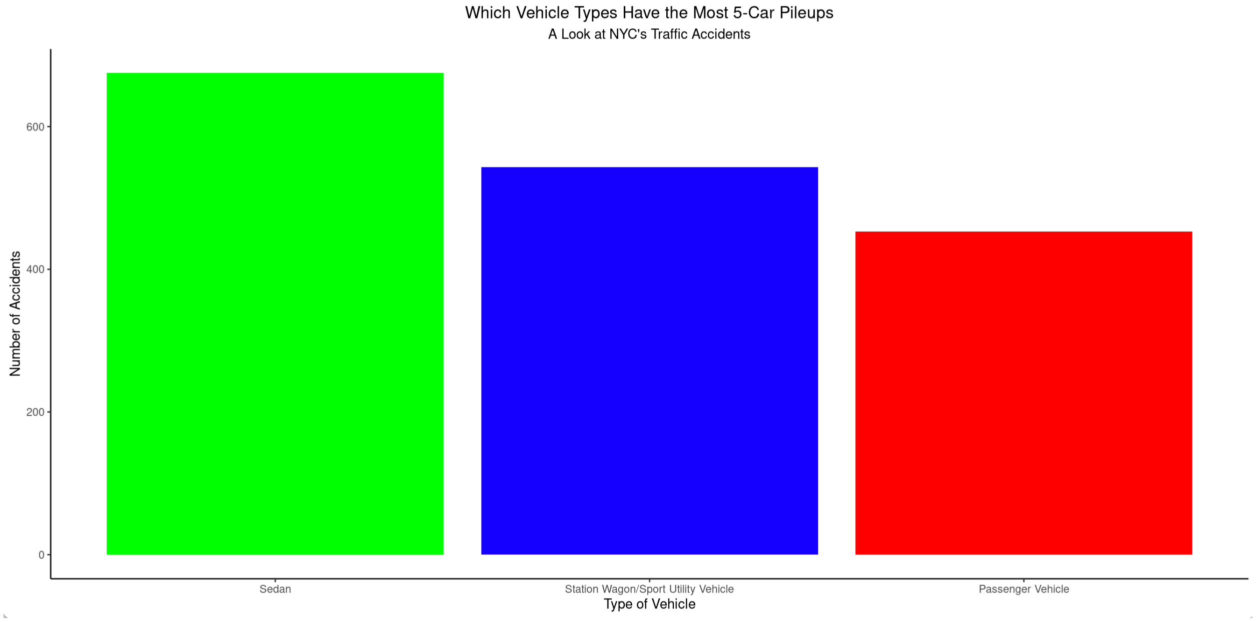
8. I then created a variable to hold the top 3 vehicle types by number of accidents. I used this variable to create a geometric bar chart using ggplot, and it demonstrates that sedans make up a large portion of the 5-car-accidents in NYC. This may be explained by sedans being the most popular car in general, or this may be indicative that sedan drivers simply crash more; more data is needed to determine the accurate answer to this. Below are the top 3 vehicle types by number of accidents as well as the geometric bar chart.



```
                               Type RealFreq
                  Passenger Vehicle      453
                              Sedan      675
 Station Wagon/Sport Utility Vehicle      543
```



```
#Look at the top 3 most accident-prone vehicles in barchart
p<-ggplot(v, aes(x=reorder(Type, -RealFreq), y=RealFreq)) +
  geom_bar(stat="identity", fill = rainbow(3)) +
  ggtitle("Which Vehicle Types Have the Most 5-Car Pileups", "A Look at NYC's Traffic Accidents") +
  labs(y= "Number of Accidents", x = "Type of Vehicle") +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 0.5))
p
```



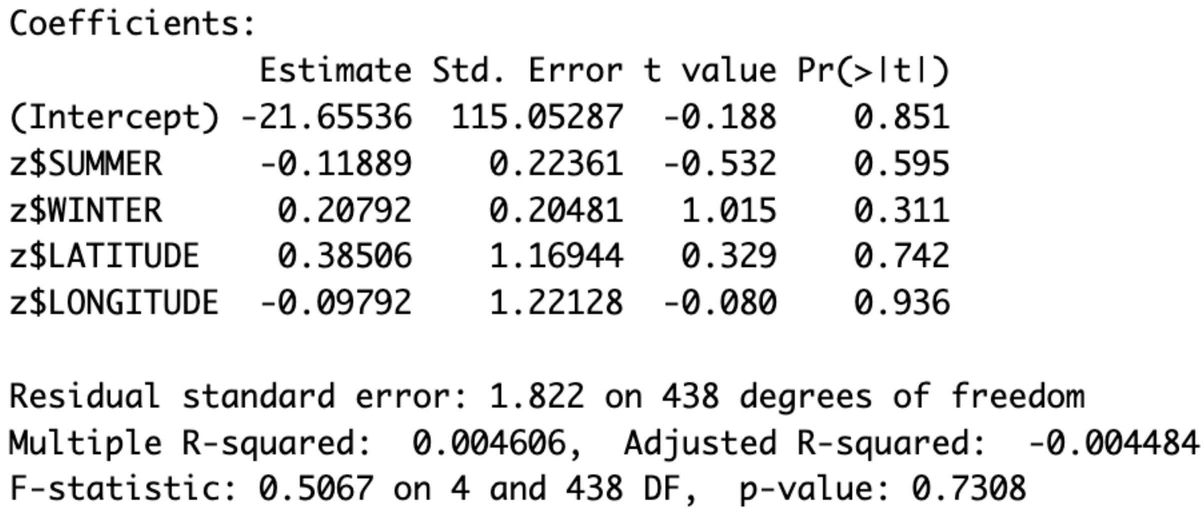
9. I ended the project with several predictive models. First, I created a linear regression model with summer, winter, longitude and latitude as variables; it was not significant and surprisingly had a negative r-squared (which means that the model is worse than a horizontal line). The resulting model is pictured in Appendix A.

10. Then, I dummy coded the first contributing factor column (which shows the reason given for causing [the first driver recorded by the police] to crash; I used the fastDummy library to do this given the fact that there were 29 different observations. I used these variables to construct a predictive model. The resulting model is pictured in Appendix B. Most variables were significant, but r-squared was 0.06 which is abysmal; but as I trimmed variables away the model got worse. The overall model is insignificant as a result, and I would need better data to create a better predictive model (perhaps with more observations or better variables). The residuals also fell very far from the model itself, meaning the model fit the data poorly (Ap. C). However, the model is informative in that its factors indicate increasing/decreasing probabilities of 5-car-accidents if any or multiple factors are present (such as speeding leading to accidents).

Discussion/Conclusion

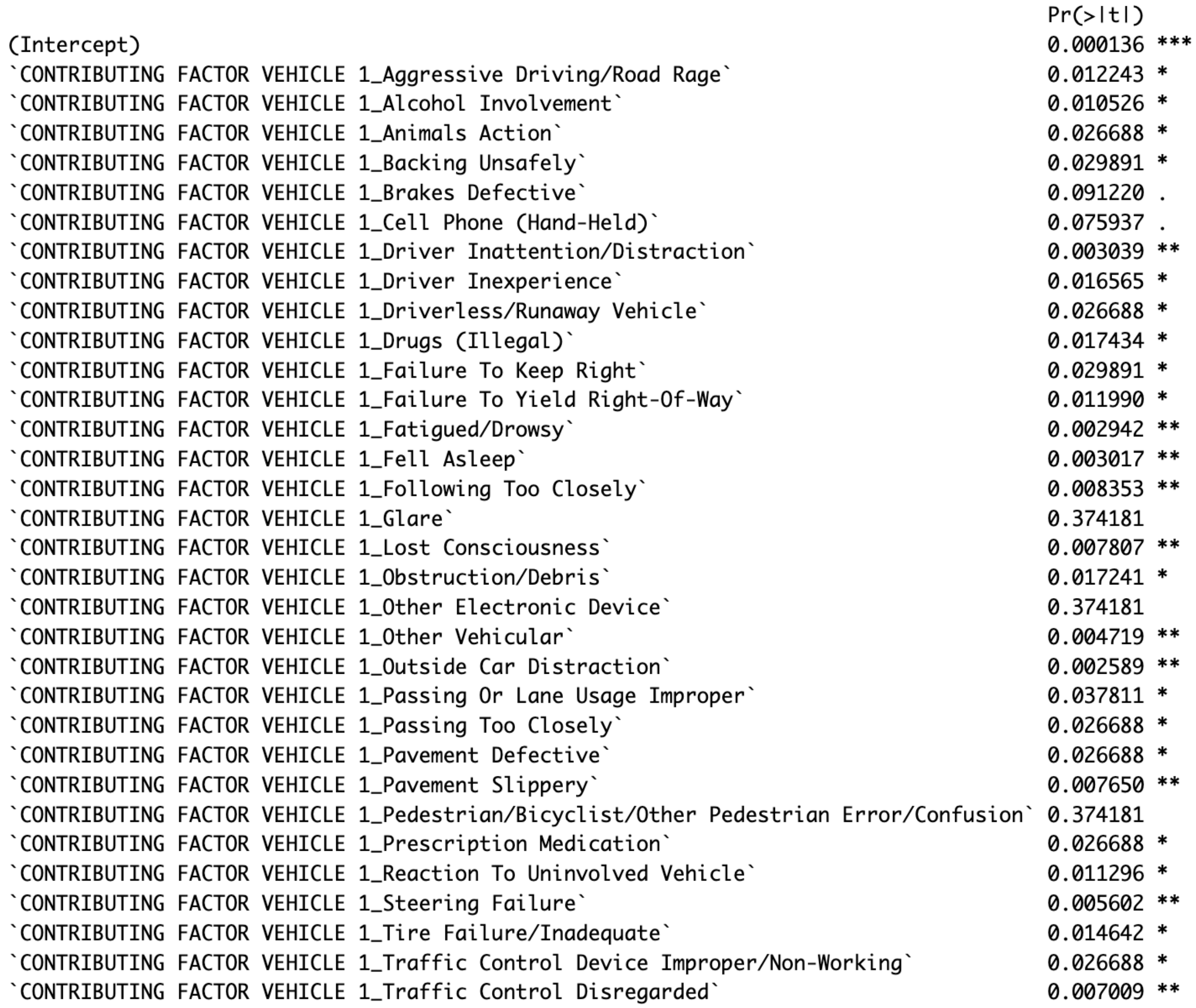
The data mostly answers the questions I had in the introduction. To improve the study, I would try different machine learning models to see if one fits the data better than simple regression. I would also use the full data set, but RAM constraints led me to use just this sample. I am likely victim of confirmation bias in that I combed through the data expecting my hypotheses to be true, but this ultimately gave structure to the project. In conclusion, this project helped me develop and showcase my R skills, and the data was interesting to look over!

Appendix A: Model 1

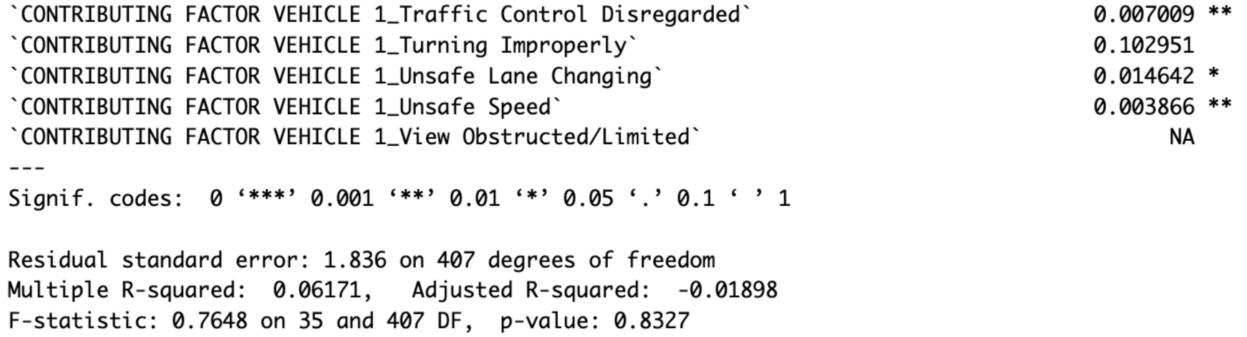


```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -21.65536  115.05287  -0.188    0.851
z$SUMMER     -0.11889    0.22361  -0.532    0.595
z$WINTER      0.20792    0.20481   1.015    0.311
z$LATITUDE    0.38506    1.16944   0.329    0.742
z$LONGITUDE  -0.09792    1.22128  -0.080    0.936

Residual standard error: 1.822 on 438 degrees of freedom
Multiple R-squared:  0.004606,  Adjusted R-squared:  -0.004484
F-statistic: 0.5067 on 4 and 438 DF,  p-value: 0.7308
```

Appendix B: Model 2

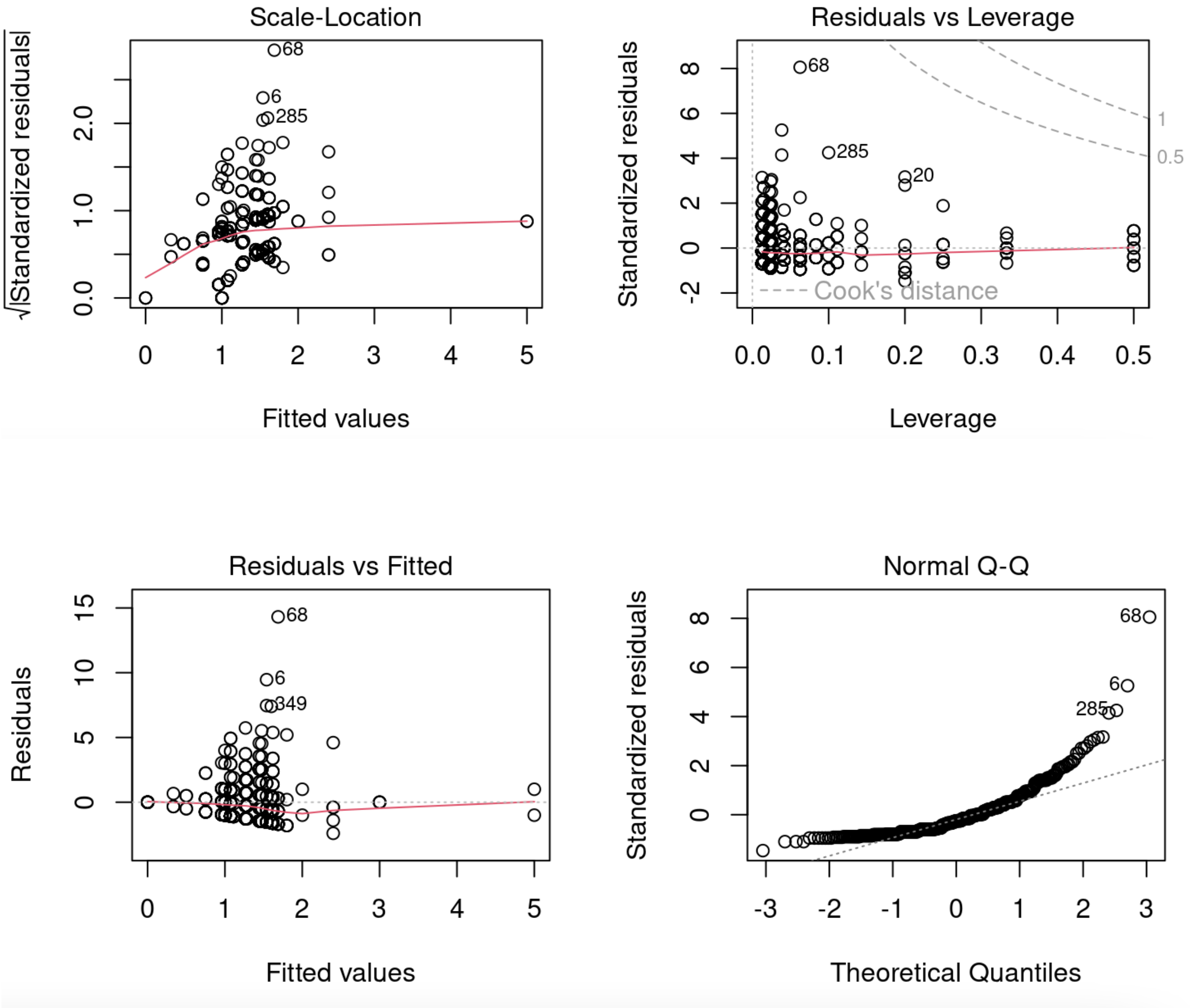


```
                                                                                                   Pr(>|t|)
(Intercept)                                                                                        0.000136 ***
`CONTRIBUTING FACTOR VEHICLE 1_Aggressive Driving/Road Rage`                                       0.012243 *
`CONTRIBUTING FACTOR VEHICLE 1_Alcohol Involvement`                                                0.010526 *
`CONTRIBUTING FACTOR VEHICLE 1_Animals Action`                                                     0.026688 *
`CONTRIBUTING FACTOR VEHICLE 1_Backing Unsafely`                                                   0.029891 *
`CONTRIBUTING FACTOR VEHICLE 1_Brakes Defective`                                                   0.091220 .
`CONTRIBUTING FACTOR VEHICLE 1_Cell Phone (Hand-Held)`                                             0.075937 .
`CONTRIBUTING FACTOR VEHICLE 1_Driver Inattention/Distraction`                                     0.003039 **
`CONTRIBUTING FACTOR VEHICLE 1_Driver Inexperience`                                                0.016565 *
`CONTRIBUTING FACTOR VEHICLE 1_Driverless/Runaway Vehicle`                                         0.026688 *
`CONTRIBUTING FACTOR VEHICLE 1_Drugs (Illegal)`                                                    0.017434 *
`CONTRIBUTING FACTOR VEHICLE 1_Failure To Keep Right`                                              0.029891 *
`CONTRIBUTING FACTOR VEHICLE 1_Failure To Yield Right-Of-Way`                                      0.011990 *
`CONTRIBUTING FACTOR VEHICLE 1_Fatigued/Drowsy`                                                    0.002942 **
`CONTRIBUTING FACTOR VEHICLE 1_Fell Asleep`                                                        0.003017 **
`CONTRIBUTING FACTOR VEHICLE 1_Following Too Closely`                                              0.008353 **
`CONTRIBUTING FACTOR VEHICLE 1_Glare`                                                              0.374181
`CONTRIBUTING FACTOR VEHICLE 1_Lost Consciousness`                                                 0.007807 **
`CONTRIBUTING FACTOR VEHICLE 1_Obstruction/Debris`                                                 0.017241 *
`CONTRIBUTING FACTOR VEHICLE 1_Other Electronic Device`                                            0.374181
`CONTRIBUTING FACTOR VEHICLE 1_Other Vehicular`                                                    0.004719 **
`CONTRIBUTING FACTOR VEHICLE 1_Outside Car Distraction`                                            0.002589 **
`CONTRIBUTING FACTOR VEHICLE 1_Passing Or Lane Usage Improper`                                     0.037811 *
`CONTRIBUTING FACTOR VEHICLE 1_Passing Too Closely`                                                0.026688 *
`CONTRIBUTING FACTOR VEHICLE 1_Pavement Defective`                                                 0.026688 *
`CONTRIBUTING FACTOR VEHICLE 1_Pavement Slippery`                                                  0.007650 **
`CONTRIBUTING FACTOR VEHICLE 1_Pedestrian/Bicyclist/Other Pedestrian Error/Confusion` 0.374181
`CONTRIBUTING FACTOR VEHICLE 1_Prescription Medication`                                            0.026688 *
`CONTRIBUTING FACTOR VEHICLE 1_Reaction To Uninvolved Vehicle`                                     0.011296 *
`CONTRIBUTING FACTOR VEHICLE 1_Steering Failure`                                                   0.005602 **
`CONTRIBUTING FACTOR VEHICLE 1_Tire Failure/Inadequate`                                            0.014642 *
`CONTRIBUTING FACTOR VEHICLE 1_Traffic Control Device Improper/Non-Working`                        0.026688 *
`CONTRIBUTING FACTOR VEHICLE 1_Traffic Control Disregarded`                                        0.007009 **
```



```
`CONTRIBUTING FACTOR VEHICLE 1_Traffic Control Disregarded`                                        0.007009 **
`CONTRIBUTING FACTOR VEHICLE 1_Turning Improperly`                                                 0.102951
`CONTRIBUTING FACTOR VEHICLE 1_Unsafe Lane Changing`                                               0.014642 *
`CONTRIBUTING FACTOR VEHICLE 1_Unsafe Speed`                                                       0.003866 **
`CONTRIBUTING FACTOR VEHICLE 1_View Obstructed/Limited`                                                  NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.836 on 407 degrees of freedom
Multiple R-squared:  0.06171,   Adjusted R-squared:  -0.01898
F-statistic: 0.7648 on 35 and 407 DF,  p-value: 0.8327
```

Appendix C: Residuals for Model 2


Scale-Location
√|Standardized residuals|
0.0
1.0
2.0
0
1
2
3
4
5
Fitted values
68
6
285
Residuals vs Leverage
Standardized residuals
-2
0
2
4
6
8
0.0
0.1
0.2
0.3
0.4
0.5
Leverage
68
285
20
1
0.5
Cook's distance
Residuals vs Fitted
Residuals
0
5
10
15
0
1
2
3
4
5
Fitted values
68
6
349
Normal Q-Q
Standardized residuals
0
2
4
6
8
-3
-2
-1
0
1
2
3
Theoretical Quantiles
68
6
285