The Impact of Factors on Vehicle Crashes

Merjem Memic, Islam Hachem, Tahamin Salam

December, 1, 2022

Introduction

Vehicle Crashes happen on a daily basis. The effect of how severe the crash was depends on many factors. Other resources similarly note that crashes happen for multiple reasons, with timing, driving skills, road design, construction zones, and numerous other factors contributing to car crashes. They explain that road crashes are the leading cause of death and serious injuries in both developed and developing countries. Of course, maintaining and constantly updating road designs or infrastructure is one aim of preventing car crashes, but exploring the numerous factors in car crashes were what motivated us to further explore our car crash dataset. We were curious to see if there were other trends on when car crashes occurred, or if there were other things we discovered independently. Interestingly, we had found that some other studies learned that the summer season is the time that the most dangerous crashes happen. In our data set, we set out to find our own trends and learn if we find anything similar or different than that of other studies. For this project, we will look at many factors that contribute to the severity of a vehicle crash. The main variable we will examine is Crash Score of each vehicle and factors that contribute to the crash. Many factors noted in this project are the type of Road Conditions, Weather, Time of day, Road classification, Road Features, and more. Environmental factors such as the type of weather affect the conditions of the road. Increase in severity level due to wet road conditions makes it much more difficult to break since it is much slipperier than when the road is dry[3]. Relating from Egypt's National Road Project journal article. Making it easier for vehicles to crash much more in the winter due to rainy season. In addition, where crashes happen can have an impact on their severity [2]. These factors need to be considered separately and as a whole to get a better idea of each factor's impact. Thus, our main focus with this data is to determine when and where crashes are most severe and/or most common.

This data-set is composed from data collected by the North Carolina Department of Transportation and provided by Cary, NC. It is observational data that gives detailed data on car crashes in Cary, North Carolina over 2014 to 2019. Since it is observational data, only association between variables can be inferred but not causation. Here, we are going to assume and treat the data as if it is population. There are 23, 137 rows of data available in the vehicle crash data set. These are all complete rows and surprisingly has no Not Available values so, no rows need to be removed. All of the variables are categorical except for `Crash_Score` which is numerical.

Data Set Description

The most interesting variable is the `Crash_Score` variable that quantifies how bad a crash was. No details on how exactly this variable is calculated are given except for the fact that many factors are considered, and the higher the crash score, the worse the crash. There are `Month` and `year` variables that give the calender month and year of the crash. This data set only includes crashes from January of 2014 to March of 2019 so there technically isn't a complete 6 years of data available. All other variables give information on the circumstances of when and where the crash happened.

The variable `Rd_Feature` gives the "special" features of the road that the crash was on. The possible options are that were none, it was on an intersection of two or more roads, it was on a ramp that was an entrance or exit of a "controlled access road," it was a driveway of a personal residence or business, or other which is a catch all for another special feature that is not specified. There is Rd_Character which gives a simple description of the road crashed on. The possibilities for Rd_Character are Straight-Level meaning no curves or hill, Straight-Grade - hills but no curves, Straight-Other so straight but not of the other categories, Curve-Level meaning curves but no hills, Curve-Grade - a curve but no hill, Curve-Other meaning a curve and a hill, and simply Other. Rd_Class tells the classification of the road. So, it can be a state highway meaning that it is maintained by the state government, a US Highway that is maintained by the federal government, or neither. The variable Rd_Configuration gives the design of the road crashed on. So, it can be recorded as a Two-Way-Protected-Median, a Two-Way-Unprotected-Median, Two-Way-No-Median, a One-Way, or Unknown. The variable Rd_Condition informs us of the condition of the road at the time of the crash. The possible observations are dry, wet, ice-snow-slush, and other. There is a variable Time_of_Day that describes when the crash happened in terms of four hour intervals. For example, a number 1 would represent that the crash happened between midnight and 4 am, and a number 3 would represent that the crash happened between 8 am and 12 pm. The variable Weather tells the whether at the time of the crash. Here, it details whether it was clear, cloudy, or if snow or rain was present, or labeled as other.

Exploratory Data Analysis and: Descriptive Statistics and Visualization

Rd_Character	Freq
CURVE-GRADE	643
CURVE-LEVEL	725
CURVE-OTHER	239
OTHER	13
STRAIGHT-GRADE	2622
STRAIGHT-LEVEL	18215
STRAIGHT-OTHER	680

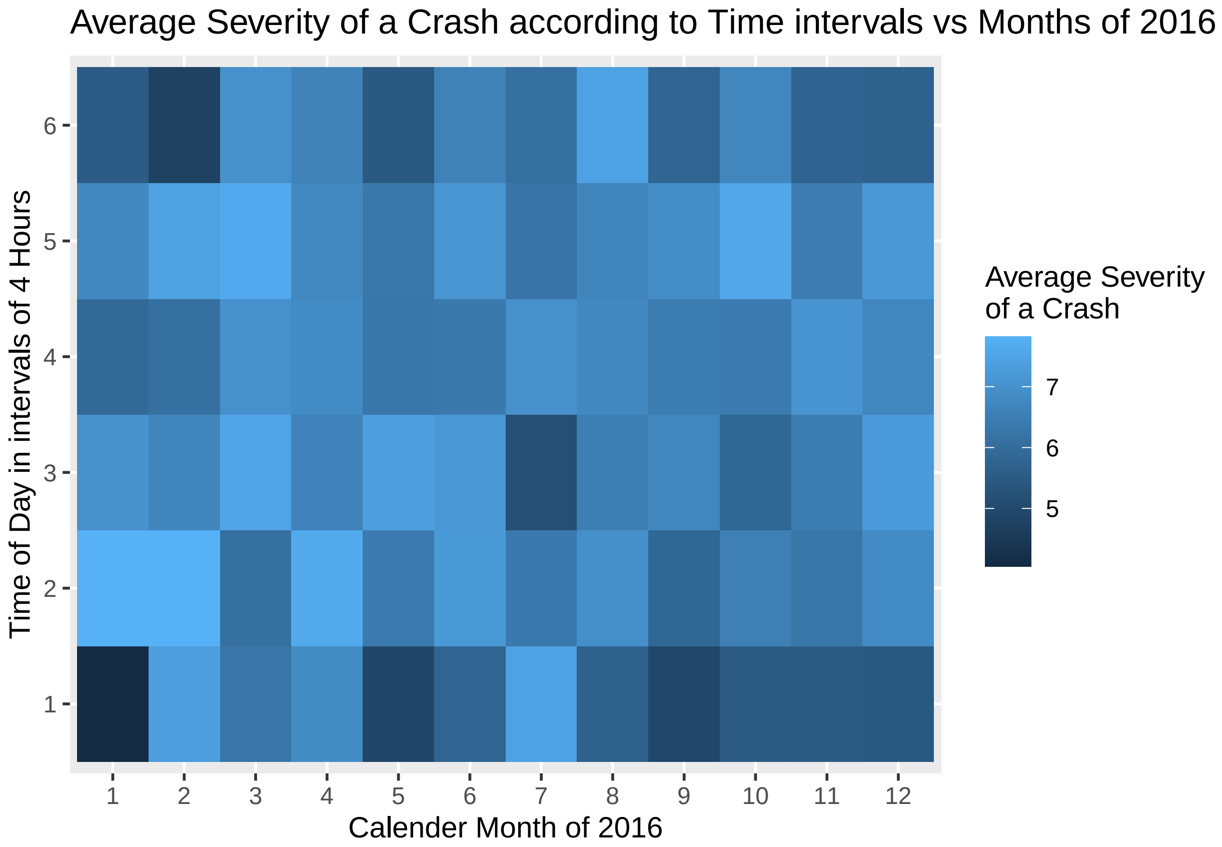
This table gives the frequency of crashes for the variable Rd_Character, which describes the main characteristic of the road crashes on. With this table, we can see that most crashes happen on roads which are straight and level. of the specified road types, roads labeled as curve-other had the least crashes.

```
##   min   Q1 median  Q3   max     mean       sd     n missing
##  0.01 3.54   5.66 8.6 53.07 6.566871 4.278949 23137       0
```

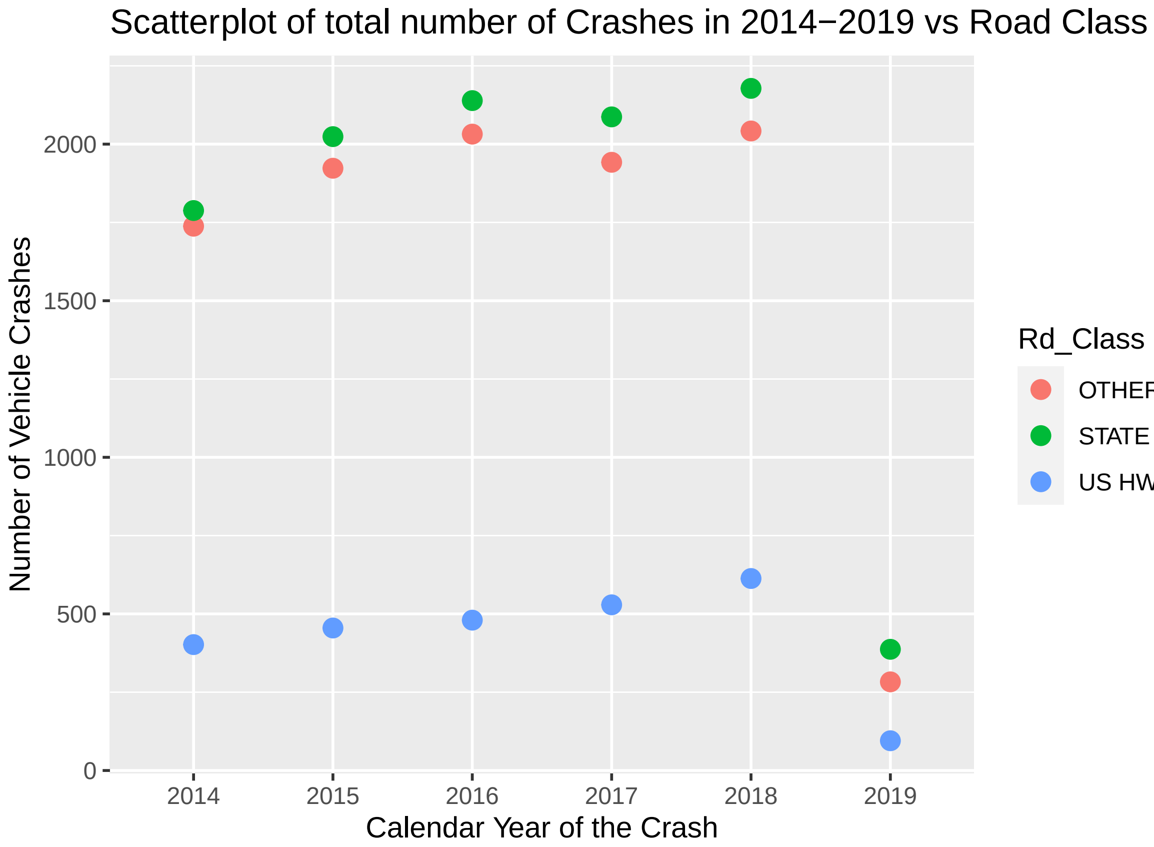
This numerical summary shows that the numerical variable Crash_Score has a max of 53.07 and a minimum of 0.01. In addition, it also gives the average Crash_Score for all crashes recorded. The total average is a modest 6.57 with a standard deviation of 4.27. The maximum score is much higher than the average which points to it possibly being an outlier.

Rd_Conditions	Freq
DRY	19262
ICE-SNOW-SLUSH	322
OTHER	134
WET	3419

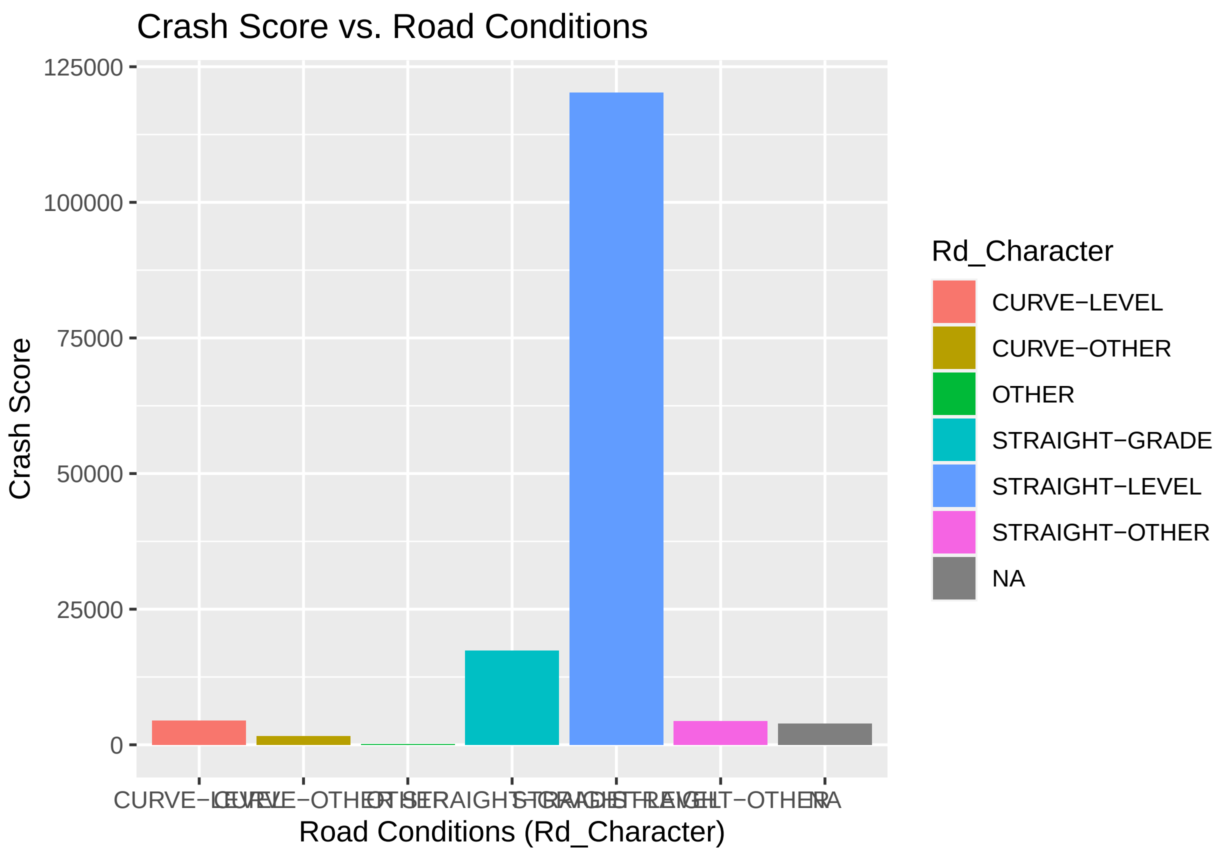
This table gives the frequency of crashes on differing road conditions. This is an important variable to look at because people often worry about their impact on their driving performance. Despite common thought, not many crashes happened on roads with ice, slush or snow. However, this table doesn't give the full picture since these crashes could be more severe than those on dry roads. Nevertheless, dry roads by far have the most crashes. There were over 15 thousand more crashes on dry roads compared to the second highest road condition.



This graph shows the average severity of a crash related to the each month of 2016 compared to the time of day. This shows that there is a relationship between the severity of a crash and the time of day. This graph shows that crashes were on average less severe between 8 pm and 4 am (represented by the integers 1 and 6). Crashes were on average the worst between 8 am and 8 pm. This coincides with usual times that people are awake and on the road driving places. This also tells us that crashes would be more dangerous between those times because the crash score would more likely be higher.



For this visualization, I used a scatterplot graph to showcase the 'year' (x variable) and the 'number of all vehicle crashes' for that specific year (y variable). I also grouped it by the 'Road Class' variable to indicate where most of the crashes occurred. Looking at the graph, you can see a correlation that the highest amount of crashes occurred in the 'State Highway' and lowest amount is the 'US Highway'.



This graph is to help compare the conditions of the road against crash score. I specified which variable I was using in the graph to avoid confusion since there are numerous variables explaining the conditions of the road. The graph appears to show that the highest crash scores happened on straight-level roads, compared to other more difficult road layouts that have a significantly lower values of crash scores.

Statistical Analysis

```
##   min   Q1 median  Q3   max     mean       sd     n missing
##  0.01 3.54   5.66 8.6 53.07 6.566871 4.278949 23137       0
```

```
## [1] 1085   14
```

bootstrap_samplemean
6.845733
7.098212
6.823152
6.732470
6.959014
6.736664

```
##     2.5%    97.5%
## 6.490807 7.048755
```

Since, we are treating our data set as if it was the population. We first needed to actually create a sample of the population. Then, a bootstrap interval was used. By using `favstats()` on our population, we can find the mean of `Crash_Score`. Comparing that mean to the confidence interval created, we can see that the population mean of 6.57 is well within the confidence interval. This means that the bootstrap interval is working as expected.

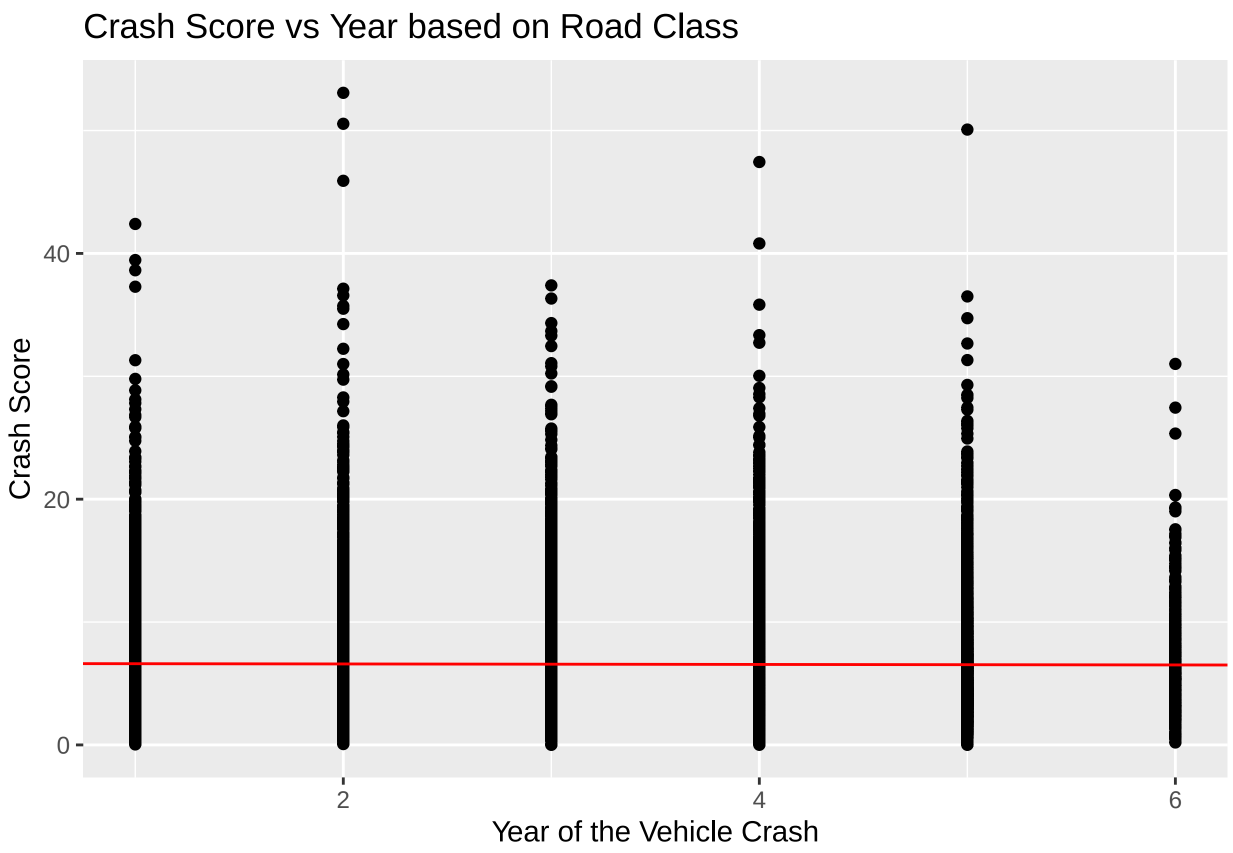
Hypothesis Test:
According to our data, our population is of the crashes in Cary, NC, from the years 2014 to 2019. Our null and alternative hypothesis are as follows:

H_0 : The difference between means is equal to 0
H_A : The difference between means is not equal to 0

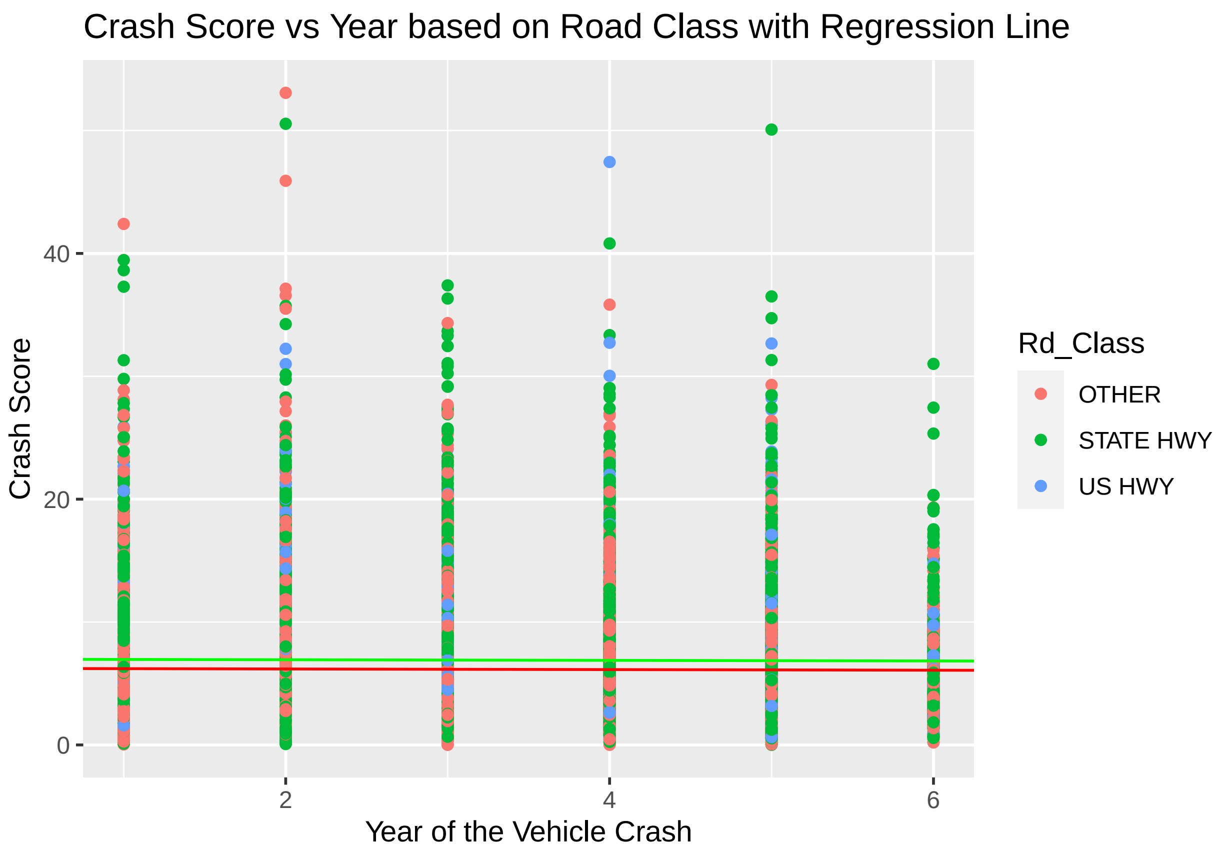
```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 10.442, df = 14019, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5189501 0.7587971
## sample estimates:
## mean of x mean of y
##  6.999202  6.360328
```

According to the hypothesis test, we can see clearly that the p-value is less than .05. Therefore, we reject the null hypothesis. The data provides convincing evidence at the .05 significance level that the difference of the mean is not equal to 0.

```
##
## Call:
## lm(formula = Crash_Score ~ as.numeric(year), data = vehicle_crash)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.571 -3.030 -0.911  2.020 46.479
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.63089    0.06705  98.888   <2e-16 ***
## as.numeric(year) -0.02011    0.01912  -1.052    0.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.279 on 23135 degrees of freedom
## Multiple R-squared:  4.782e-05,  Adjusted R-squared:  4.595e-06
## F-statistic: 1.106 on 1 and 23135 DF,  p-value: 0.2929
```



```
##
## Call:
## lm(formula = Crash_Score ~ as.numeric(year) + Rd_Class, data = vehicle_crash)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.862 -2.995 -0.892  2.008 46.890
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.23009    0.07365  84.587  < 2e-16 ***
## as.numeric(year)  -0.02480    0.01906  -1.301    0.193
## Rd_ClassSTATE HWY  0.75172    0.05950  12.634  < 2e-16 ***
## Rd_ClassUS HWY     0.64049    0.09432   6.790 1.14e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.264 on 23133 degrees of freedom
## Multiple R-squared:  0.007241,   Adjusted R-squared:  0.007113
## F-statistic: 56.25 on 3 and 23133 DF,  p-value: < 2.2e-16
```



```
## [1] 16196    14
```

```
## [1] 17.17743
```

```
## [1] 17.06643
```

Here we created two models that showcase the linear regression line for `Crash_Score` and `Year`. Since year is categorical and not a significant predictor, and there isn't any other numerical variable besides `Crash_Score`, We interpret it as as.numeric(year) to allow us to graph the scatterplot. From the first model, we can see that there is a red line above zero that is horizontally on the graph. This line tells us that its not quite a obvious relationship between the two and that the coefficient is close to zero. For `model2` we looked at the same x and y variables and added another categorical variable to the regression line. I used `Road Class` to color code the scatterplot between `US HWY`, `STATE HWY`, and `OTHER`. In which I described earlier in the file that this is important for the "where" aspect of where most vehicle crashes were severe/common. The estimate standard of the `STATE HWY` was 0.75172 while the 'US HWY' is .64049, which is not a big of a difference. Here in `model2` we see that there is another regression line in the color green which matched to the colors coefficient of `STATE HWY` that is a bit higher than the red line but the difference is not significant since the two lines are close. The test mean square error for the first model was 17.17743 and the second model was 17.0664, in which are relatively close however `model2` is best since it would be best used within the dataset.

Discussion and Conclusion

Because of the structure of our data-set, our end results focused on showcasing how our categorical variables contributed to the commonality and severity of the vehicle crashes recorded. We started off with a table on road character in terms of the number of crashes. Roads describes as being straight and level had the most crashes with 18,215 overall. This is a large increase from the second highest number of 2,622 on roads that were straight-grade. Our second table focused on road conditions and amount of crashes instead. It shows that by far the most crashes happened on dry roads since there were over 15 thousand more crashes on dry roads than wet roads which is second in terms of crashes. This begs teh question of why do people often worry about crashes much more on wet roads? For example, are those crashes on average worse? Our numerical summary of the variable `Crash_Score` shows that the average crash score for the whole data-set was 6.57 with a standard deviation of 4.27. However, the highest crash score recorded in the data was 53.07, which is very high in comparison to the average and suggests that it is an outlier.

When looking at the first graph, it can be seen that for 2016, the average severity of crashes were higher from 8pm to 8am, which correlates to when most people would be on the road. There are slight variations in severity from month to month but no considerable patterns emerge. In the future, we could look at if the severity of crashes changes during holidays or specific days when people are more of less likely to be on the road. For the second graph on road class, we were able to interpret that the crashes were more common to take place in roads labeled as being on the `STATE HWY`. For road feature, we were able to see that the crash score was much higher in the `STRAIGHT-LEVEL` category. This means that crashes on roads describes as being straight-level were more severe compared to other road with different characteristics.

For further analysis, we can look at the other variables mentioned in the dataset and see what results they provide. We can also look at each year specifically and their contributing months to note any differences or similarities with one another. For example, looking at the variable `Traffic_Control` maybe we could have incorporated it with another variable like `Work_Area` in which we can see if more crashes occurred by the area of work-area or not. To see, if there were any traffic signals or not. We could also test other variables by making graphs for them to see which areas most people should avoid in order to decrease number of crashes in North Carolina. There are many questions that arise that can't be answered with this data-set. For example, the severity of a crash is informative but vague in its calculation. Some people people might want to know whether crashes on highways are more deadly than those in intersections. Besides that, the variables that we used were able to answer many questions on the severity and commonality of the vehicle crashes in relation to certain factors. However, there is still much unknown and many questions that arise about vehicle crashes from this data and otherwise.

In conclusion, this Rmd file is about the data set of Vehicle Crashes in Cary, North Carolina between the years 2014-2019, where we looked at when and where the crashes were most common and most severe. We looked at the different variables that contributed to the numerical variable of the number of `Crash_Score`. With this information we were able to create graphs that went along with the data and compared it to the variables that contributed with the crash. We were also able to create a bootstrap confidence interval in which it compared the winter months of 2015 and gave us the mean crash score for that time. We also did the hypothesis test on the population mean for vehicle Crash Score of crashes whose `Rd_Feauture` group is `INTERSECTION`, and `DRIVEWAY` and vehicle Crash Score of crashes whose `Rd Feauture` is `OTHER` and `RAMP`. We rejected the null hypothesis since it was much smaller to the p-value of 0.05. Models with regression line showed that there is not much of a difference between the variables of `Rd_Class` for `US HWY`, `STATE HWY` in which implied that slope estimate is close to zero and the intercepts estimate are not that far different from each other. With these information, it allowed us to look at variables that contributed the most in the severity and or common areas that the crashes occurred.

Bibliography

[1] Ankin Law Office LLC. "Car Accidents More Likely During Summer." Hg.org, HG Legal Resources, Sept. 2013, https://www.hg.org/legal-articles/car-accidents-more-likely-during-summer-40867.

[2] Tay, Richard, and Shakil Mohammad Rifaat. "Factors contributing to the severity of intersection crashes." Journal of Advanced Transportation 41.3 (2007): 245-265. https://onlinelibrary.wiley.com/doi/epdf/10.1002/atr.5670410303.

[3] Zhang, Kairan, and Mohamed Hassan. “Crash Severity Analysis of Nighttime and Daytime Highway Work Zone Crashes.” PLOS ONE, vol. 14, no. 8, 2019, https://doi.org/10.1371/journal.pone.0221128.

Brief Summary: For the writing portion, Islam wrote one paragraph for Introduction and another two for the Discussion/Conclusion. Tahamin incorporated external sources. Merjem wrote the data-set description, a paragraph for the discussion, and some of the Introduction. Each person of the group provided a single graph and one element of the statistical analysis. For example, Tahamin provided the hypothesis test.