Project Part C Report

Andrew Alsop, Alan Raj, Onaopemipo Olagoke

2022-12-06

```
car_crash = read.csv("Vehicle_crash_data.csv")
vehicle_crashes<- read.csv("Vehicle_crash_data.csv")
```

```
library(ggplot2)
library(dplyr)
library(mosaic)
```

Introduction

In the country of the United States of America, nearly everyone drives. Driving is usually seen as a mandatory activity, the only option for many people going about their daily lives. In July 2018, Gallup Polls reported that 83% of Americans drive frequently, a staggering majority [1]. With the sheer volume of drivers using the road system, the safety of that road system becomes incredibly important. Investing in safe road infrastructure positively impacts the lives of these drivers. Lowering the severity of a crash when it occurs is a key component of this puzzle.
Throughout the years of 2014-2019, the Department of Transportation for Cary, North Carolina recorded car crashes that occurred on their roads. For each crash, a crash score was recorded, this crash score corresponds to the severity of the crash. In addition, many road factors were recorded, such as weather, traffic control, and road features. From this data set, a general picture of an individual crash can be drawn. By investigating the larger trends over many crashes, an idea of which road features are most dangerous can begin to form.
This report will research which factors contribute to a higher crash score. We will explore many of the factors present in the data set, and compile those that are associated with a higher average crash score. The isolation of these factors would be useful for guiding road infrastructure investments by targeting the factors with the highest risk.
In order to determine which factors contribute to a higher crash score, we will conduct exploratory data analysis on the data set and visualize the results. The focus of this section will be to determine average crash score against different variables, examining how those variables may affect crash score. Following that, we will conduct statistical analysis on the data including: two bootstrap confidence intervals, one hypothesis test, and two linear regression models.

Dataset Description

The observational data set was obtained from **The North Carolina Department of Transportation, Crashes in Cary, North Carolina (NC) 2014-2019.** In this data set no `NA` values were present.

Variables of Interest: We will be focusing on the following variables: `Crash_score`: It measures the extent of the crash using factors such as number of injuries, fatalities, the number of vehicles involved, and other factors.

`Rd_conditions`: Condition of the road. It tells if the road is dry, wet, ice-snow, or other.

`Light`: Lightning on the road. It tells if the road has street lamps, dark, dusk, dawn or other.

`Weather`: Weather Conditions. It tells if its cloudy, snowy , or clear.

`Rd_Character`: Description of the road where the crash occurred. It tells if the road straight, has a grade, a curve or other.

`Rd_Class`: Classification of the road type. It tells if the road is a State highway or a Federal highway.

`Rd_Configuration`: Design of the road. It tells if the road has a protected median, unprotected median, two-way or one-way.

`Time_of_day`: Time of Day.

`Month`: Month of the year.

All variables are categorical variables except `crash_score`.

This data is observational data.

Data Transformation

Because all variables except `Crash_Score` are categorical, I convert all variables except `Crash_Score` to factors(categorical variables).

```
vehicle_crashes$Month = as.factor(vehicle_crashes$Month)
vehicle_crashes$year = as.factor(vehicle_crashes$year)
vehicle_crashes$Time_of_Day = as.factor(vehicle_crashes$Time_of_Day)
vehicle_crashes$Rd_Feature  = as.factor(vehicle_crashes$Rd_Feature )
vehicle_crashes$Rd_Character = as.factor(vehicle_crashes$Rd_Character)
vehicle_crashes$Rd_Class = as.factor(vehicle_crashes$Rd_Class)
vehicle_crashes$Rd_Configuration = as.factor(vehicle_crashes$Rd_Configuration)
vehicle_crashes$Rd_Surface = as.factor(vehicle_crashes$Rd_Surface)
vehicle_crashes$Rd_Conditions = as.factor(vehicle_crashes$Rd_Conditions)
vehicle_crashes$Light = as.factor(vehicle_crashes$Light)
vehicle_crashes$Weather = as.factor(vehicle_crashes$Weather)
vehicle_crashes$Traffic_Control = as.factor(vehicle_crashes$Traffic_Control)
vehicle_crashes$Work_Area   = as.factor(vehicle_crashes$Work_Area  )

str(vehicle_crashes)
```

```
## 'data.frame':    23137 obs. of  14 variables:
##  $ Crash_Score     : num  6.56 6.53 1.58 7.15 9.57 8.14 6.06 2.11 4.37 9 ...
##  $ year            : Factor w/ 6 levels "2014","2015",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Month           : Factor w/ 12 levels "1","2","3","4",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ Time_of_Day     : Factor w/ 6 levels "1","2","3","4",..: 2 3 5 3 6 3 5 4 3 4 ...
##  $ Rd_Feature      : Factor w/ 5 levels "DRIVEWAY","INTERSECTION",..: 3 3 3 3 3 3 2 3 5 3 ...
##  $ Rd_Character    : Factor w/ 7 levels "CURVE-GRADE",..: 6 6 6 6 6 6 6 5 7 6 ...
##  $ Rd_Class        : Factor w/ 3 levels "OTHER","STATE HWY",..: 2 1 2 1 1 1 2 2 3 2 ...
##  $ Rd_Configuration: Factor w/ 5 levels "ONE-WAY","TWO-WAY-NO-MEDIAN",..: 3 2 2 2 2 2 2 4 1 2 ...
##  $ Rd_Surface      : Factor w/ 5 levels "COARSE ASPHALT",..: 5 1 5 5 1 5 5 5 5 5 ...
##  $ Rd_Conditions   : Factor w/ 4 levels "DRY","ICE-SNOW-SLUSH",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Light           : Factor w/ 6 levels "DARK-LIT","DARK-NOT-LIT",..: 4 4 2 4 1 4 4 4 4 4 ...
##  $ Weather         : Factor w/ 5 levels "CLEAR","CLOUDY",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Traffic_Control : Factor w/ 5 levels "NONE","OTHER",..: 1 1 1 1 1 1 3 1 3 1 ...
##  $ Work_Area       : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 1 1 1 1 2 ...
```

Exploratory Data Analysis and Descriptive Statistics and Visualization

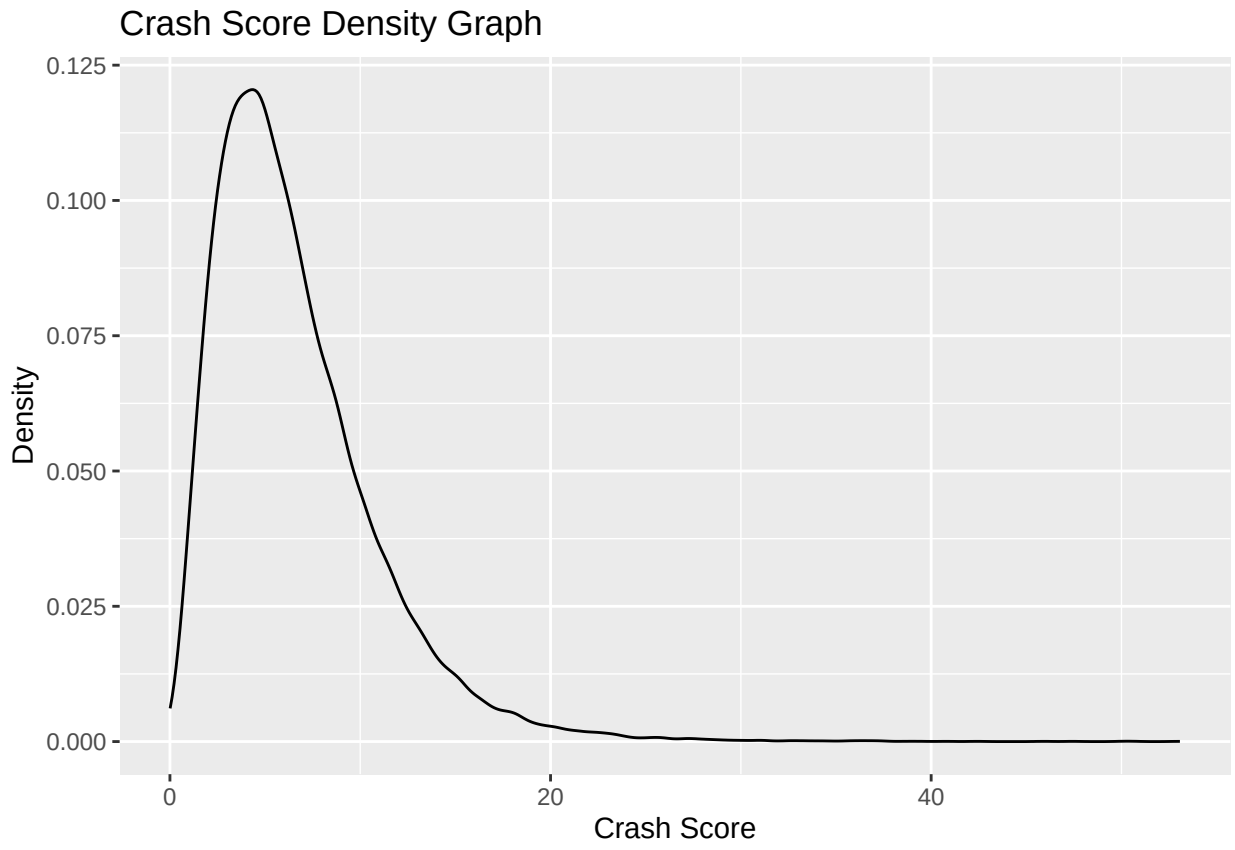
Table of summary statistics:

```
favstats(~Crash_Score, data = vehicle_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
```

From the numerical summary above, the highest crash score in the given data is 53.07, the minimum is 0.01 and the average crash score is 6.566871.

```
vehicle_crashes%>%
ggplot(aes(x=Crash_Score))+
geom_density()+
ggtitle("Crash Score Density Graph")+
xlab("Crash Score")+
ylab("Density")
```



The density graph above shows that the highest density of `Crash_Score` is found between approximately 0-12, peaking at around 5.

```
knitr::kable(
vehicle_crashes%>%
select(Crash_Score, Rd_Conditions, Light, Weather, year,Work_Area )%>%
```

```
group_by(Rd_Conditions,year)%>%
summarize(count_of_accidents=n())
)
```

```
## `summarise()` has grouped output by 'Rd_Conditions'. You can override using the
## `.groups` argument.
```

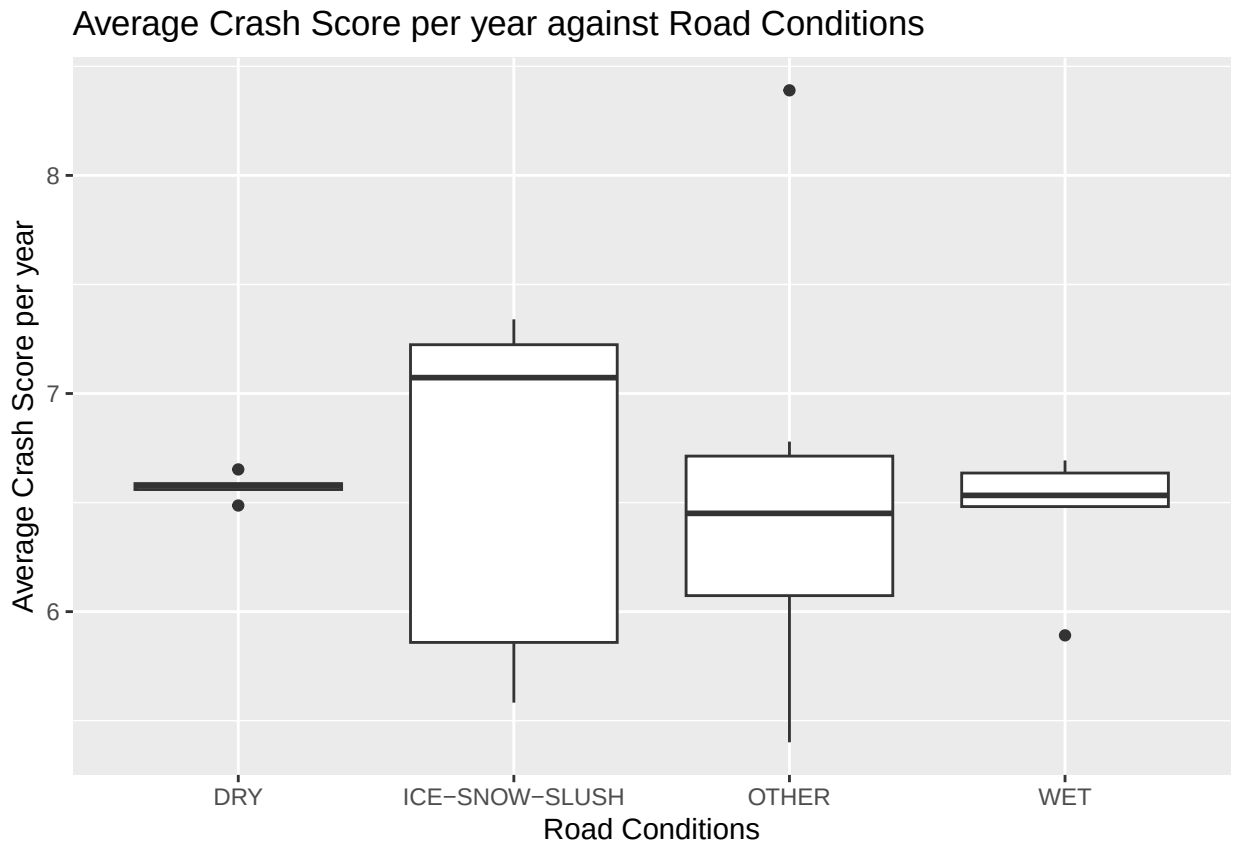
Rd_Conditions	year	count_of_accidents
DRY	2014	3230
DRY	2015	3488
DRY	2016	4032
DRY	2017	3977
DRY	2018	3968
DRY	2019	567
ICE-SNOW-SLUSH	2014	74
ICE-SNOW-SLUSH	2015	98
ICE-SNOW-SLUSH	2016	42
ICE-SNOW-SLUSH	2017	25
ICE-SNOW-SLUSH	2018	83
OTHER	2014	22
OTHER	2015	28
OTHER	2016	25
OTHER	2017	23
OTHER	2018	32
OTHER	2019	4
WET	2014	602
WET	2015	788
WET	2016	552
WET	2017	533
WET	2018	750
WET	2019	194

From the numerical summary above, we observe that more accidents occur when it is dry. However, this does not necessarily mean that the accidents occur because it is dry. We can infer that because many people drive when it is dry hence more accidents occur especially when compared to the number of accidents that occur when the road condition is not dry.

```
n<-vehicle_crashes%>%
select(Crash_Score, Rd_Conditions,year, Light, Weather )%>%
group_by(Rd_Conditions, year)%>%
summarize(average_Crash_score=mean(Crash_Score))
```

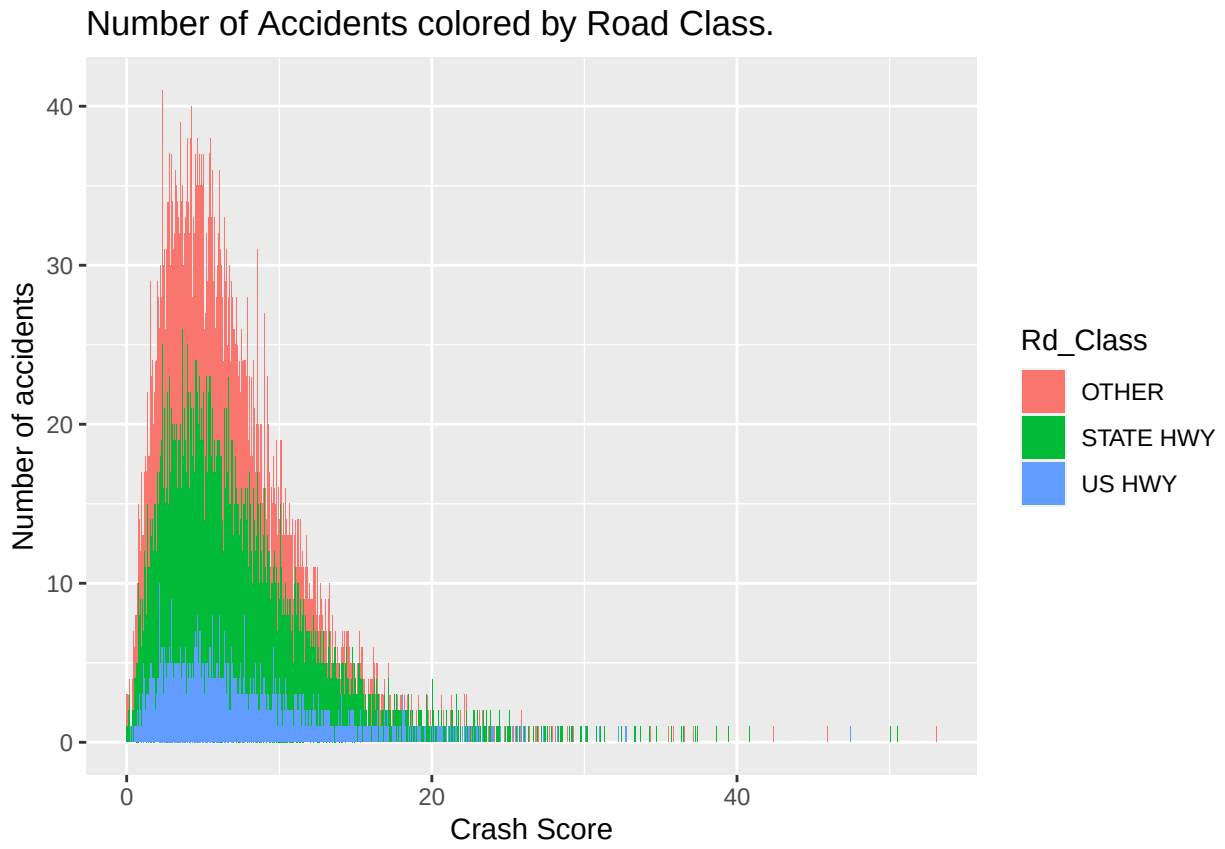
```
## `summarise()` has grouped output by 'Rd_Conditions'. You can override using the
## `.groups` argument.
```

```
ggplot(n, aes(Rd_Conditions, average_Crash_score))+
geom_boxplot()+
ggtitle('Average Crash Score per year against Road Conditions ')+
ylab("Average Crash Score per year")+
xlab("Road Conditions")
```



From the box plot above, The highest average crash score per year is found when the road condition is ice-snow slush. With an approximate median of about 7.2. From this observation we can infer that although more accidents happen when its dry, the most severe crashes occur when the road condition is ice snow slush.

```
vehicle_crashes%>%
  select(Crash_Score, Rd_Class, Traffic_Control)%>%
  ggplot(mapping = aes(x = Crash_Score, fill = Rd_Class))+
geom_bar()+
ggtitle("Number of Accidents colored by Road Class.")+
xlab("Crash Score")+
ylab("Number of accidents")
```



From the bar graph above we observe that the highest number of accidents occurred on state highways, we use the numerical summary below to confirm this.

```
knitr::kable(
vehicle_crashes%>%
select(Crash_Score, Rd_Class, Light, Weather, year,Work_Area )%>%
group_by(Rd_Class)%>%
summarize(count_of_accidents=n())
)
```

Rd_Class	count_of_accidents
OTHER	9960
STATE HWY	10603
US HWY	2574

```
summaryDF = vehicle_crashes %>% select(Rd_Class ,year,Crash_Score) %>%
  group_by(Rd_Class) %>%
  summarize("Average Crash Score" = mean(Crash_Score))
knitr::kable(summaryDF)
```

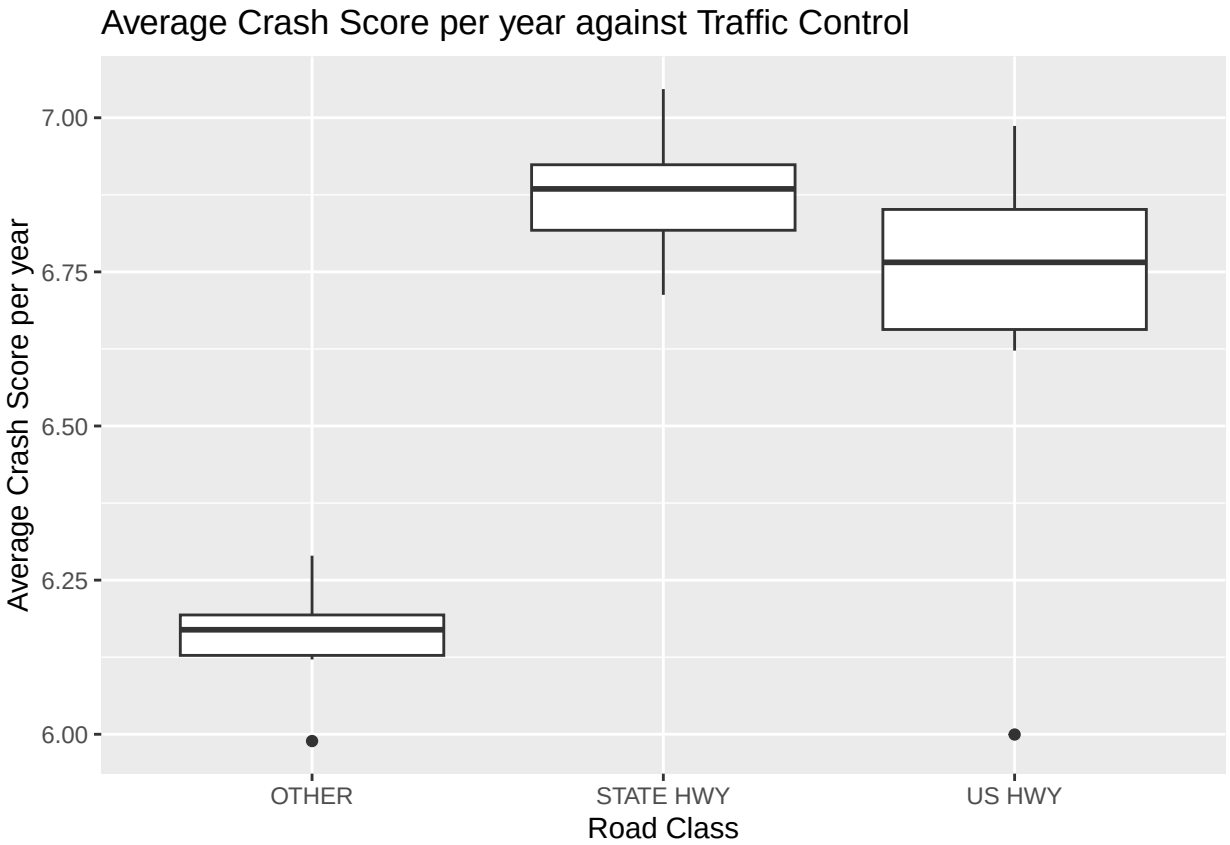
Rd_Class	Average Crash Score
OTHER	6.152018

Rd_Class	Average Crash Score
STATE HWY	6.902724
US HWY	6.788660

```
n<-vehicle_crashes%>%
select(Crash_Score, Rd_Conditions,year, Light, Weather,Rd_Class )%>%
group_by(Rd_Class, year)%>%
summarize(average_Crash_score=mean(Crash_Score))
```

```
## `summarise()` has grouped output by 'Rd_Class'. You can override using the
## `.groups` argument.
```

```
ggplot(n, aes(Rd_Class, average_Crash_score))+
geom_boxplot()+
ggtitle('Average Crash Score per year against Traffic Control ')+
ylab("Average Crash Score per year")+
xlab("Road Class")
```



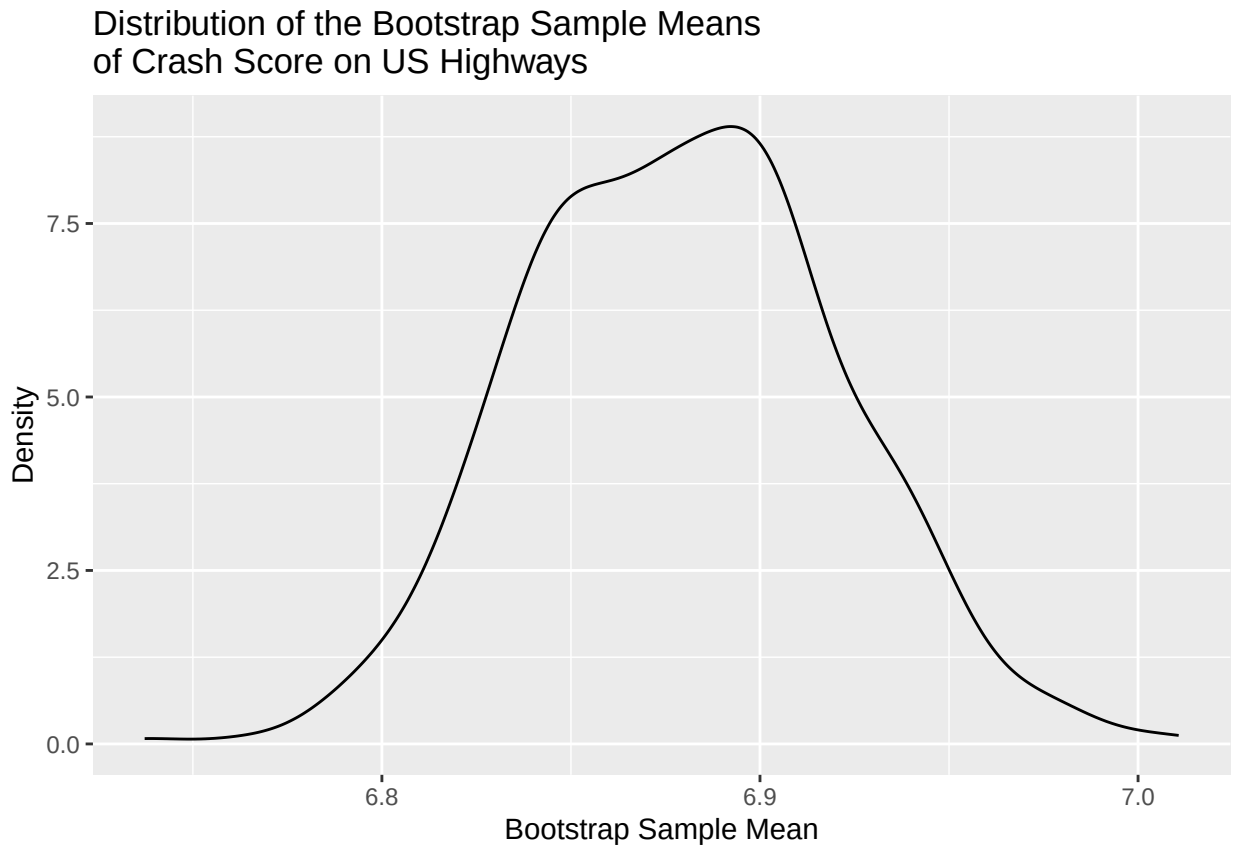
From the box plot and numerical summary above we can infer that more accidents occur on state highways and the accidents that happen on state highways have higher crash scores.

Statistical Analysis

Bootstrap Confidence Interval

```
#Performing a bootstrap confidence interval
HighwayDF = car_crash %>%
  filter(Rd_Class %in% c("US HWY", "STATE HWY")  )
set.seed(101)
bootstrap_samplemeans = do(500)*mean(~Crash_Score, data = sample_n(HighwayDF,
                                                                   size=nrow(HighwayDF),
                                                                   replace=TRUE))
#Displaying the head of the confidence interval
bootstrap_samplemeans = bootstrap_samplemeans %>%
  rename(bootstrap_samplemean = mean)

#Testing if the bootstrap confidence interval has a normal distribution
ggplot(bootstrap_samplemeans, aes(x=bootstrap_samplemean)) +
  geom_density() +
  labs (
    title = "Distribution of the Bootstrap Sample Means \nof Crash Score on US Highways",
    x = "Bootstrap Sample Mean",
    y = "Density"
    )
```



```
confidenceinterval = quantile(bootstrap_samplemeans$bootstrap_samplemean,
                              probs = c(.025, .975))
knitr::kable(head(bootstrap_samplemeans))
```

bootstrap_samplemean
6.922620
6.898783
6.891486
6.931267
6.981896
6.802765

```
knitr::kable(confidenceinterval)
```

	x
2.5%	6.803166
97.5%	6.958291

We are 95% confident that the true population crash score mean lies between 6.803 and 6.958

```
NOTHWY_CAR_CRASH = car_crash %>%
  filter(Rd_Class == "OTHER")
  set.seed(101)

bootstrap_samplemeans =
  do(500)*mean(~Crash_Score, data = sample_n(NOTHWY_CAR_CRASH,
                                        size=nrow(NOTHWY_CAR_CRASH),
                                        replace=TRUE))

bootstrap_samplemeans = bootstrap_samplemeans %>%
  rename(bootstrap_samplemean = mean)
knitr::kable(head(bootstrap_samplemeans))
```

bootstrap_samplemean
6.133466
6.094900
6.127943
6.124794
6.053231
6.179523

```
confidenceintervals = quantile(bootstrap_samplemeans$bootstrap_samplemean,
                              probs = c(.025, .975))
knitr::kable(confidenceintervals)
```

	x
2.5%	6.080609
97.5%	6.227545

We are 95% confident that the true population crash score mean lies between 6.080 and 6.227

We compared two bootstrap confidence intervals, one for Highways and one for non-highways. Looking at the results, we are 95% confident that the true population crash score mean is higher for highways (true population crash score mean is between 6.801 and 6.956) and lower for non-highways (true population crash score mean is between 6.074 and 6.232).

Hypothesis Testing

In the hypothesis test below, the total population is the number of accidents that occurred when it was snowing and when it was clear with their respective crash scores. Two samples `vehiclecrashes_snow_sample` and `vehiclecrashes_clear_sample` were made from the original population with respect to the two weather conditions (SNOW, CLEAR) . The two samples are of size 150. This lead to the hypothesis questions below:
H_0 : : The average crash score when its snowing is the same as the average crash score when its clear
H_A : : The average crash score when its snowing is not the same as the average crash score when its clear.

```
set.seed(101)
vehiclecrashes_snow_sample<- car_crash%>%
filter(Weather%in%c('SNOW'))%>%
sample_n(size=150, replace=FALSE)

vehiclecrashes_clear_sample<-car_crash%>%
filter(Weather%in%c('CLEAR'))%>%
sample_n(size=150, replace=FALSE)

t.test(vehiclecrashes_snow_sample[,'Crash_Score'],vehiclecrashes_clear_sample[,'Crash_Score'])
```

```
##
##  Welch Two Sample t-test
##
## data:  vehiclecrashes_snow_sample[, "Crash_Score"] and vehiclecrashes_clear_sample[, "Crash_Score"]
## t = -1.2946, df = 290.16, p-value = 0.1965
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.6257835  0.3356501
## sample estimates:
## mean of x mean of y
##  6.461200  7.106267
```

Conclusion: The data do not provide convincing evidence at the .05 significance level, then we do not have enough evidence to reject the null hypothesis. The data does not provide enough evidence to infer that the average crash scores are different.

Linear Regression Model

We fitted a linear regression model for the crash score based on the time of the day along with a categorical variable `Work Area`.

```
# Comparing 2 Linear regression models for Crash Score based on Time of day in a work area and outside
WA_model = lm(Crash_Score ~ Time_of_Day + Work_Area, data = car_crash)


ggplot(car_crash, aes(x = Time_of_Day, y = Crash_Score, color=Work_Area)) +
  geom_point() +
  ggtitle("Linear Regression models for Time of Day vs Crash Score")+
  xlab("Time of Day")+
  ylab("Crash Score")+
  geom_abline(intercept = WA_model$coefficients[1],
              slope = WA_model$coefficients[2],
              color="red") +
  geom_abline(intercept = WA_model$coefficients[1] + WA_model$coefficients[3],
              slope = WA_model$coefficients[2],
              color="#00BFC4")
```

Observation: When using the regression line to make a prediction about crash score based on the type of Area (work area or non-work area), we can see a slight difference in the regression lines. This suggests an association between work area and higher crash scores.

Second Linear Regression Model

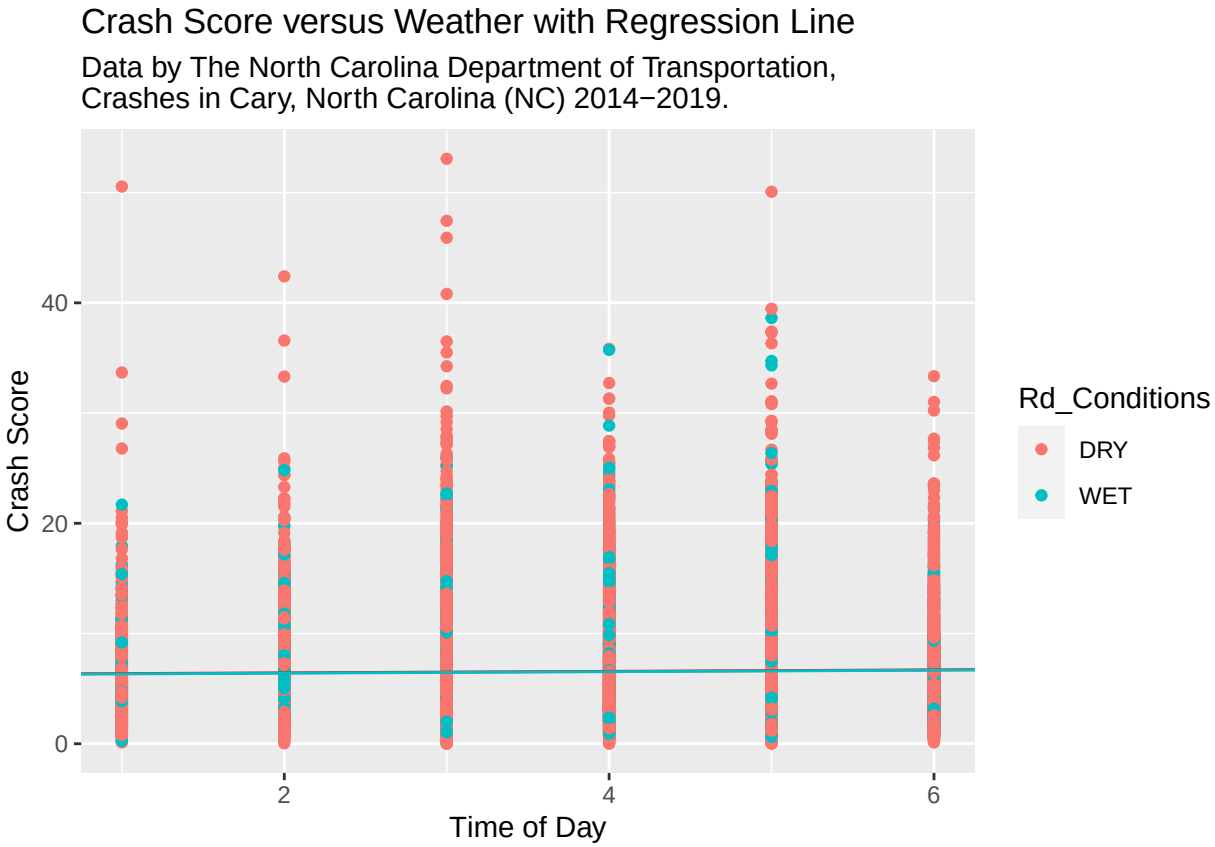
We fitted a linear regression model for the crash score based on the time of the day along colored by two factors of the `Road Conditions` variable, wet and dry.

```
Rd_Conditions_pro <- car_crash%>%
filter(Rd_Conditions%in%c('WET', 'DRY'))

model3 = lm( Crash_Score ~ as.numeric(Time_of_Day) + Rd_Conditions, data = Rd_Conditions_pro)
summary(model3)
```

```
##
## Call:
## lm(formula = Crash_Score ~ as.numeric(Time_of_Day) + Rd_Conditions,
##     data = Rd_Conditions_pro)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.628 -3.026 -0.913  2.022 46.567
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.30059    0.09923  63.496  < 2e-16 ***
## as.numeric(Time_of_Day)   0.06749    0.02332   2.894  0.00381 **
## Rd_ConditionsWET         -0.03213    0.07939  -0.405  0.68567
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.278 on 22678 degrees of freedom
## Multiple R-squared:  0.0003766,  Adjusted R-squared:  0.0002884
## F-statistic: 4.272 on 2 and 22678 DF,  p-value: 0.01397
```

```
ggplot(Rd_Conditions_pro, aes(x = as.numeric(Time_of_Day), y =Crash_Score , color=Rd_Conditions)) +
geom_point() +
labs(title = "Crash Score versus Weather with Regression Line",
subtitle = "Data by The North Carolina Department of Transportation, \nCrashes in Cary, North Carolina
xlab("Time of Day")+
ylab("Crash Score")+
geom_abline(intercept = model3$coefficients[1],
slope = model3$coefficients[2],
color="red") +
geom_abline(intercept = model3$coefficients[1] + model3$coefficients[3],
slope = model3$coefficients[2],
color="#00BFC4")
```



Observation: When using the regression line to make a prediction about crash score based on time of day and road conditions- which include only when the road is wet and dry- the road conditions don't make a big difference on average crash score

Test error values

To compare the two linear models, we compute the mean squared error. The mean squared error gives us a good idea of how close a regression line is to a set of data points.

```
#knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(echo = FALSE,
                      results = "markup",
                      warning = FALSE,
                      message = FALSE
                      )
total = dim(car_crash)[1]
n = round(.7*total)
m = dim(car_crash[1]) - n

# randomly select 70% of the rows to be training data
# and the remaining rows to be the test data
set.seed(101)
trainingrownumbers = sample(1:total, size=n)
trainingdata = car_crash[trainingrownumbers, ]
testdata = car_crash[-trainingrownumbers, ]
dim(trainingdata)
```

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

```
model1 = lm(Crash_Score ~ Time_of_Day + Work_Area, data = trainingdata)
model2 = lm( Crash_Score ~ as.numeric(Time_of_Day) + Rd_Conditions, data = trainingdata)



# predict on the test data and find the mean square error
mse1 = mean( (testdata$Crash_Score - predict(model1, newdata  = testdata))^2  )
mse2 = mean( (testdata$Crash_Score - predict(model2, newdata = testdata))^2 )


mse1
```

```
## [1] 17.15787
```

```
mse2
```

```
## [1] 17.18014
```

To find the test error, we fit the models on approximately 70% of the data and then computed the mean squared error for the previously unseen rows. The test error values are pretty high (17.157 and 17.180), meaning that the error could be high. This could be due to the fact that Crash score is the only continuous variable for the models.

Discussion and Conclusion

In conclusion, our research of the data set has revealed many factors that are associated with a higher average crash score. Using bootstrap confidence intervals, we revealed an association between highways and higher crash scores. This means that whether a crash occurred on a highway or on a surface road is a factor that affects crash score. Intuitively this makes sense because the highway system has higher speed limits, so when a crash occurs there could be more damage. This conclusion is also supported by the graph that compares average crash score per year against road class which can be found in the EDA section. Experts in the field of road safety at NATCO agree that high speed limits on highways significantly increase the risk of death when involved in an accident. [2]
Another factor influencing crash scores is whether construction work is ongoing at the time. This is supported by the linear regression model comparing work zone and non-work zone crashes throughout the day. The model shows an association between work zone and increased crash score. A third factor that can influence crash score is the road conditions, shown by the graph comparing average crash score across different weathers: dry, wet, snow, etc. Not all factors significantly affected crash score. For example, the hypothesis test did not provide enough evidence to reject the null hypothesis. The null hypothesis was that there isn't a significant difference between crashes that occurred in various weather conditions. This does not contradict earlier findings, as the data set records weather conditions and actual road conditions differently.
Further research on the questions raised by this report would be needed before any action could be taken to lower crash scores. Determining why these factors influence crash score is beyond the scope of this report, as well as how to lower crash score for these factors. Additionally, there are many more factors that can influence crash score outside of the ones listed in this data set. The human factor in an accident is also very important, "The driver ... They are the backbone of the system and the most important factor that affects road traffic safety." [3] If the data set recorded information about the driver, other conclusions could have been reached.

Bibliograpy

[1] Brenan, Megan (2018, July 9th) “83% of U.S. Adults Drive Frequently; Fewer Enjoy It a Lot”
Retrieved from: https://news.gallup.com/poll/236813/adults-drive-frequently-fewer-enjoy-lot.aspx
[2] National Association of City Transportation Officials “City Limits”
Retrieved from: https://nacto.org/publication/city-limits/the-need/speed-kills/
[3] Xue-Jing, Du (2018, December 22) “Highway safety influencing factors in cold regions based on attribute recognition theory”
Retrieved from: https://journals.sagepub.com/doi/10.1177/1687814018818337#:~:
text=The%20driver%2C%20as%20a%20user,impact%20on%20road%20traffic%20safety

Individual Contributions

Alan Raj compiled graphs and wrote descriptions for the statistical analysis section.
Onaopemipo Olegoke compiled graphs and wrote descriptions for the exploratory data analysis and descriptive statistics and visualization section.
Andrew Alsop wrote the introduction, conclusion, and bibliography sections.