DSCI 301\_Final\_Project\_R\_Final

Milan Sapkota

12/13/2019

Motor Vehicle Crashes (New York:2014-2016)

***OVERVIEW:***

For the Final project of Data Science with R, I have done the Exploratory Data Analysis of the data from the Motor Vehicle Crashes of New York Municipality. Mainly, I have analyzed the data mostly based on what we have covered in the class. I worked through the types of data, common values, and have created diagrams for the related headers. In addition, I tried to do some predictions.

Since New York has one of the busiest roads in the United States, I had decided to work on the Motor Vehicle Crashes Case of State of New York in the three-year period 2014-2016. This data attributes about each vehicle involved in a crash as reported to NYS DMV.

***Library Used:***

library(tidyverse)

***#The Information of the file:***

data\_final <- read.csv("Motor\_Vehicle\_Crashes\_-\_Case\_Information\_\_Three\_Year\_Window.csv")  
str(data\_final)

names(data\_final)

## [1] "Year" "Crash.Descriptor"   
## [3] "Time" "Date"   
## [5] "Day.of.Week" "Police.Report"   
## [7] "Lighting.Conditions" "Municipality"   
## [9] "Collision.Type.Descriptor" "County.Name"   
## [11] "Road.Descriptor" "Weather.Conditions"   
## [13] "Traffic.Control.Device" "Road.Surface.Conditions"   
## [15] "DOT.Reference.Marker.Location" "Pedestrian.Bicyclist.Action"   
## [17] "Event.Descriptor" "Number.of.Vehicles.Involved"

summary(data\_final)

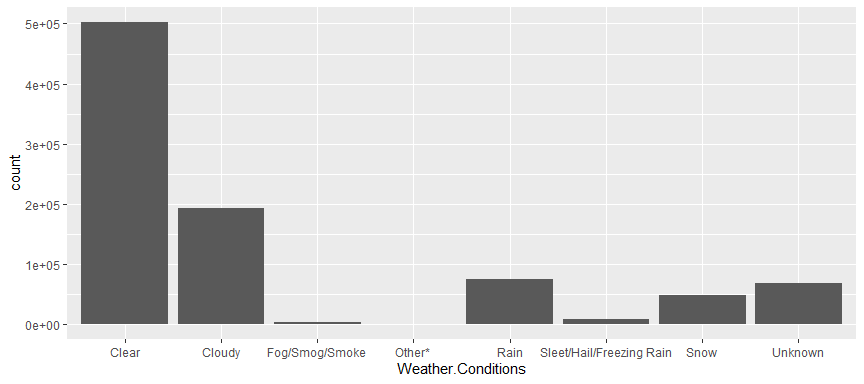
***Exploratory Data Analysis:***

In overall, most of the accidents that were reported were caused mainly in Clear Weather Conditions instead of bad weather. This is shocking that accidents may happen on the clear sky too. I guess, one reason may be that people feared to drive in bad weather days.

summary(data\_final$Number.of.Vehicles.Involved)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 1.000 2.000 1.776 2.000 50.000

ggplot(data\_final,mapping = aes(x=Weather.Conditions))+  
 geom\_bar()



***#Checkout the New York information:***

From here, I am extracting information of New York as it has one of the most crowded streets of the United States.

NY <-subset(data\_final,Municipality=="NEW YORK")  
head(NY)

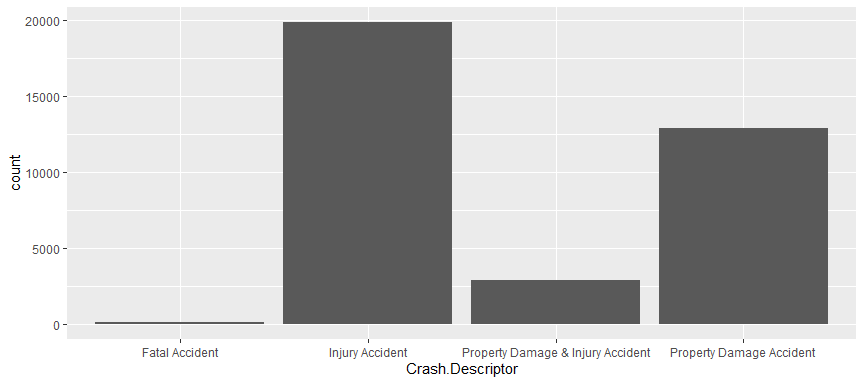
***#Total no. of accident by its Crash descriptions:***

NY %>%   
 count(Crash.Descriptor)

## # A tibble: 4 x 2  
## Crash.Descriptor n  
## <fct> <int>  
## 1 Fatal Accident 117  
## 2 Injury Accident 19863  
## 3 Property Damage & Injury Accident 2893  
## 4 Property Damage Accident 12887

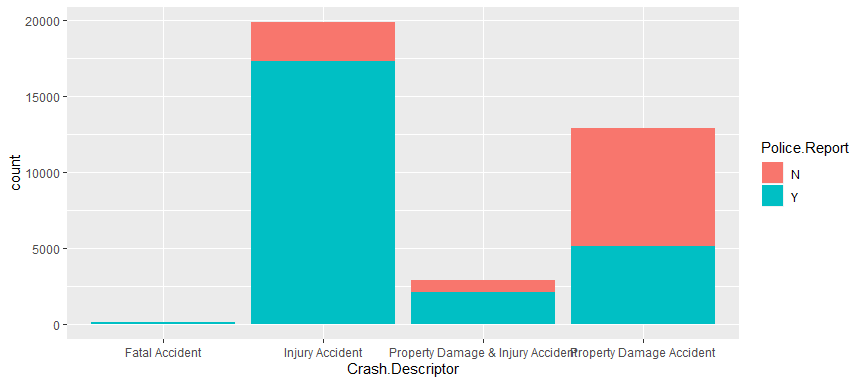
#It doesn’t sound good about the accident, but fortunately there were very low number of fatal incidents.

ggplot(data = NY) +  
 geom\_bar(mapping = aes(x=Crash.Descriptor))

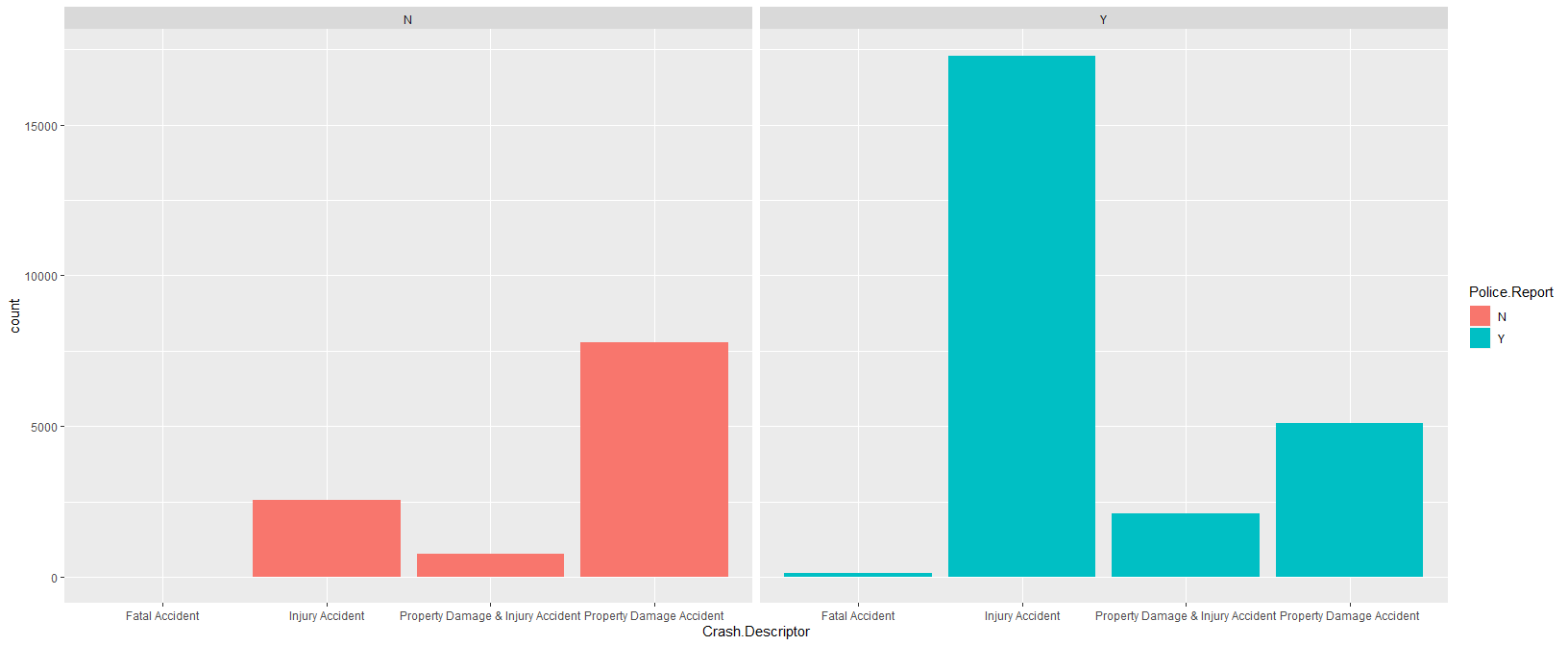


#I saw that most of the Property Damage were not reported to Police than that of the Injury accidents.

ggplot(data = NY) +  
 geom\_bar(mapping = aes(x=Crash.Descriptor, fill=Police.Report))



ggplot(data = NY) +  
 geom\_bar(mapping = aes(x=Crash.Descriptor, fill=Police.Report)) +  
 facet\_grid(~Police.Report)

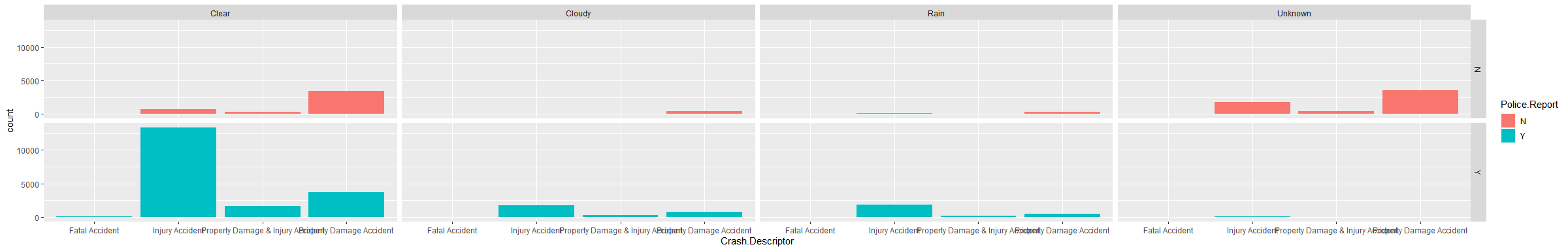


# Then I decided to look to over Crashes descriptor on the weather conditions with the Police Report on them. Obviously, much accidents were on snowing conditions; however, it also seems shocking that more of the accidents were under clear weather too.

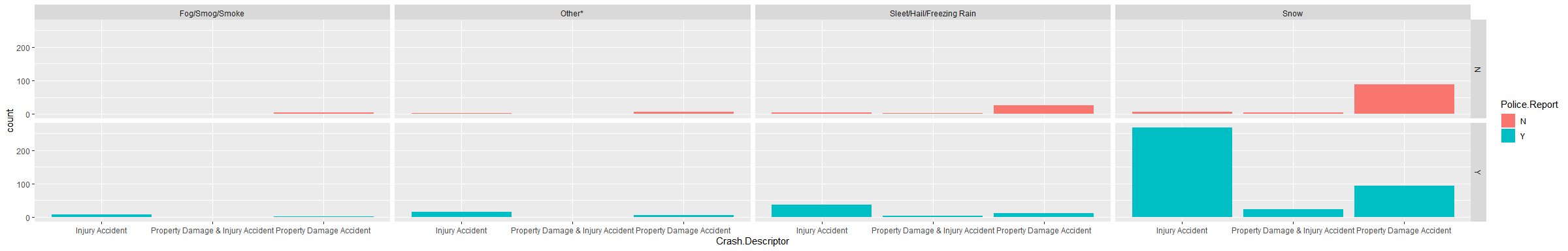
NY$Weather.Conditions[1]

## [1] Clear  
## 8 Levels: Clear Cloudy Fog/Smog/Smoke Other\* ... Unknown

NY %>%   
 filter(Weather.Conditions %in% c("Clear", "Cloudy","Rain","Unknown")) %>%   
 ggplot(mapping = aes(x=Crash.Descriptor, fill=Police.Report)) +  
 geom\_bar() +   
 facet\_grid(Police.Report~Weather.Conditions)



NY %>%   
 filter(Weather.Conditions %in% c("Fog/Smog/Smoke", "Other\*","Sleet/Hail/Freezing Rain", "Snow")) %>%   
 ggplot(mapping = aes(x=Crash.Descriptor, fill=Police.Report)) +  
 geom\_bar() +   
 facet\_grid(Police.Report~Weather.Conditions)

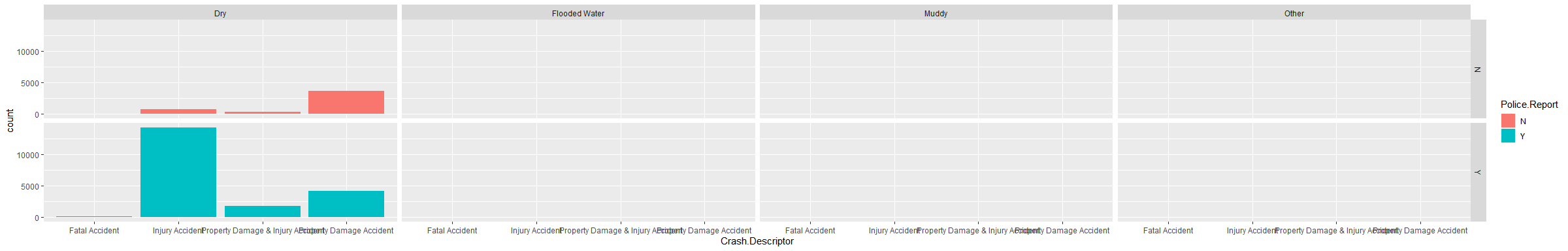


# Then I looked over Crashes descriptor on the Road surface conditions with the Police Report on them. Obviously, much accidents were on wet condition; however, it also seems shocking that more of the accidents were under dry surface.

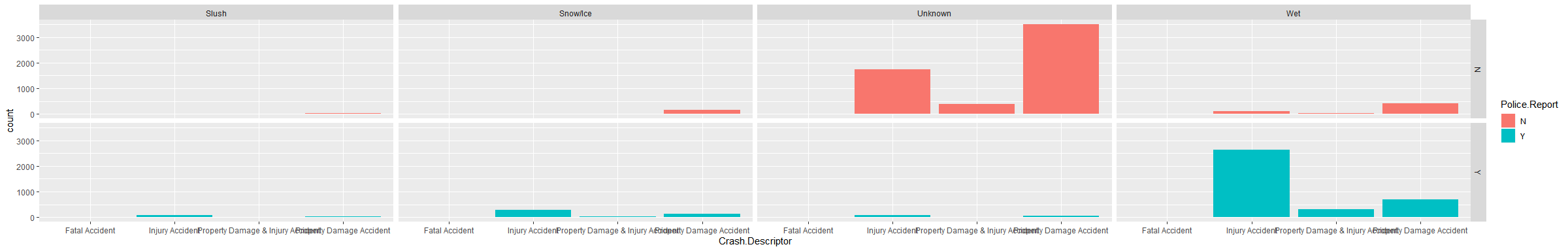
NY$Road.Surface.Conditions[1]

## [1] Dry  
## Levels: Dry Flooded Water Muddy Other Slush Snow/Ice Unknown Wet

NY %>%   
 filter(Road.Surface.Conditions %in% c("Dry","Flooded Water","Muddy","Other")) %>%   
ggplot(mapping = aes(x=Crash.Descriptor, fill=Police.Report)) +  
 geom\_bar() +  
 facet\_grid(Police.Report~Road.Surface.Conditions)



NY %>%   
 filter(Road.Surface.Conditions %in% c("Slush","Snow/Ice", "Unknown", "Wet")) %>%   
ggplot(mapping = aes(x=Crash.Descriptor, fill=Police.Report)) +  
 geom\_bar() +  
 facet\_grid(Police.Report~Road.Surface.Conditions)



***Linear Model***

I tried to some modeling on my data set. So, I chose number of vehicles involved as a target variable and year as a predictor for my linear modeling.

modu <- lm(Number.of.Vehicles.Involved~Year, data=NY)  
modu

##   
## Call:  
## lm(formula = Number.of.Vehicles.Involved ~ Year, data = NY)  
##   
## Coefficients:  
## (Intercept) Year   
## -17.911654 0.009763

summary(modu)

##   
## Call:  
## lm(formula = Number.of.Vehicles.Involved ~ Year, data = NY)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.7705 -0.7509 0.2295 0.2491 9.2295   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -17.911654 8.308894 -2.156 0.0311 \*  
## Year 0.009763 0.004123 2.368 0.0179 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6428 on 35758 degrees of freedom  
## Multiple R-squared: 0.0001567, Adjusted R-squared: 0.0001288   
## F-statistic: 5.606 on 1 and 35758 DF, p-value: 0.01791

coef(modu)

## (Intercept) Year   
## -17.911654190 0.009762956

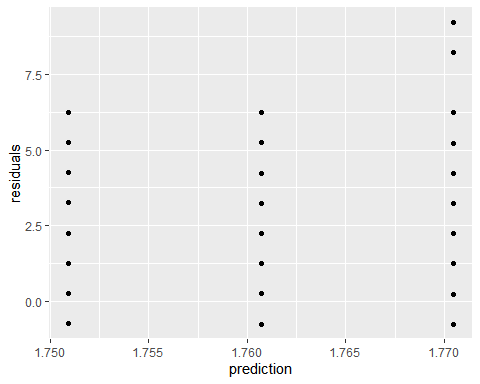
confint(modu)

## 2.5 % 97.5 %  
## (Intercept) -34.197339162 -1.62596922  
## Year 0.001680803 0.01784511

predict(modu, data.frame(Year=c(2020,2030,2050)))

## 1 2 3   
## 1.809517 1.907146 2.102405

DataFrame <- data.frame(prediction= modu$fitted.values, residuals=modu$residuals)  
DataFrame %>%   
 ggplot(aes(x=prediction, y= residuals)) +  
 geom\_point()



#Since the variables were categorical, my analysis through linear modeling did not work.

***CONCLUSION:***

From the dataset taken from the NYS DMV, most of the accidents, especially Injury accidents reported to Police, of the New York Municipality were occurred on the clear sky as well as on the Dry surface road. This could have happened due to the high number of traffics on the clear and dry days of the New York.