REGRESSION MODELING: THE IMPACT OF WEATHER CONDITION ON THE SEVERITY OF ROAD ACCIDENTS IN USA

Karthikeyan Shanthakumar
University of Massachusetts, Boston
K.Shanthakumar@umb.edu

Priya Padhiyar
University of Massachusetts, Boston
p.padhiyar001@umb.edu

This paper examines how the weather conditions on a given day impacts the severity of accidents on the road. For decades, there has been this conception, that road accidents are majorly due to distracted driving, fatigue, aggressive driving or driving under the influence. This research will try to reveal how the weather conditions contribute to the already existing factors of the road accidents. This paper will also study how much and how relevant those weather conditions are to the occurrence of road accidents based on severities and different weather types. We use the countrywide car accident dataset which cover 49 states of the USA. We use regression modeling to chiefly study the impact of a clear weather condition on the severity of car accidents. Our results suggest that clear weather solely have no relationship with the severity of accidents and there must be other conditions as well impacting the accidents. The future scope of this project is to study the omitted variable bias and integrate supporting datasets to overcome it and re study the results.

Keywords: Statistics, Traffic accident prediction, Logistic Regression, R, Data Analysis

1. Introduction:

Every day when we turn on the Television, Radio, Paper, on-line digital news articles/platforms we can hear or see the one story that keeps repeating is the road accidents. Be it local or national coverage, we can see them often occurring multiple times of the day, night at different times. Common discussion during the road accidents is between two parties finding who is at fault? However, it is not always true. Addition to drivers at fault, weather conditions can also contribute to the growing list of road accidents.

When we talk about weather conditions, we further delve deeper into the kind of weather conditions that can influence the road accidents.

1. Temperature
2. Precipitation
3. Wind direction
4. Wind speed
5. Visibility

These are some majorly contributing factors for the severity of road accidents. We will study them in the below research and how they act as the regressors for the paper.

The rest of the paper is organized as follows. Section 2 provides an overview of related work followed by dataset details in Section 3. The process of data cleaning and visualization is presented in section 4, Regression modeling and analyses are discussed in Section 5. Finally, Section 6 concludes the paper.

2. Related Work:

According to Statista, “The United States is one of the busiest countries in terms of road traffic with nearly 280 million vehicles in operation and more than 227.5 million drivers holding a valid driving license. The level of traffic is one of the reasons leading to more traffic accidents: In 2018, there were some 12 million vehicles involved in crashes in the United States”. Therefore, many researchers have been actively studying on topics like accident analysis and predictions. David Jaroszweski and Tom McNamara studied “The effect of rainfall on road accidents in urban area” where he demonstrates an alternative approach to weather-related accident analysis in cities and urban areas but the study is limited to only one weather condition. A similar study done by Sobhan Moosavi and his team “A countrywide accident dataset” where they perform real-time traffic accident prediction. These studies have missed on considering important weather conditions that can lead to the severity of accidents and that is what we look to cover in this paper.

3. About the Dataset:

The dataset for this research is obtained from Kaggle platform which is a massive dataset ranging from 2016 to 2020 year with a 3+ millions of the records. The data is continuously being gathered using multiple API's that provide streaming traffic event data. The data features a robust 47 attributes Detailed data description can be found on this link: USA Accident dataset.

4. Data cleaning and Visualization:

Upon studying the data, we realized that the number column has 60% of the missing values. There are also 45% missing values in the precipitation. 40% missing value in the wind column Number isn't needed for the study so we will remove that column instead of filling the missing values. Precipitation and Wind-chill will be required for the study so we will just fill the missing values here.

to define each encounter for a road accident. Some data elements of interest pulled from the metadata description is given below

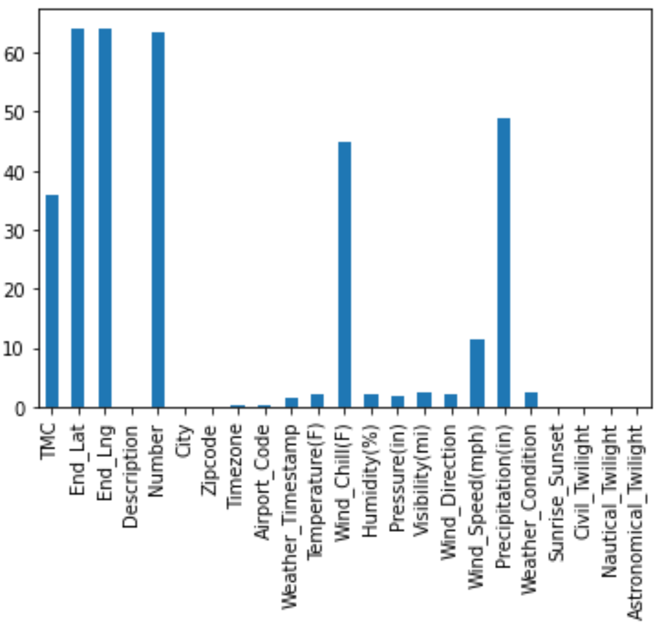
- Severity

Severity index of the accident, a designated number between 1 and 4. 1 denotes least traffic impact vs 4 being highest impact to traffic.

- Weather condition

Shows the weather condition (clear, rain, snow, thunderstorm, fog, etc.).

Also, there are some columns which are aren't required in the study and thus we will remove them like the column "country" we know we are doing analysis on only one country US. so, there is no use of that column.



There were also some data conversions that needed to be done as the original datatype isn't in the proper format. We converted the date columns from object datatypes to date datatype and extracted new columns year, month, and time. Now that the data is clean, we did some data visualizations to better understand the accidents in USA.

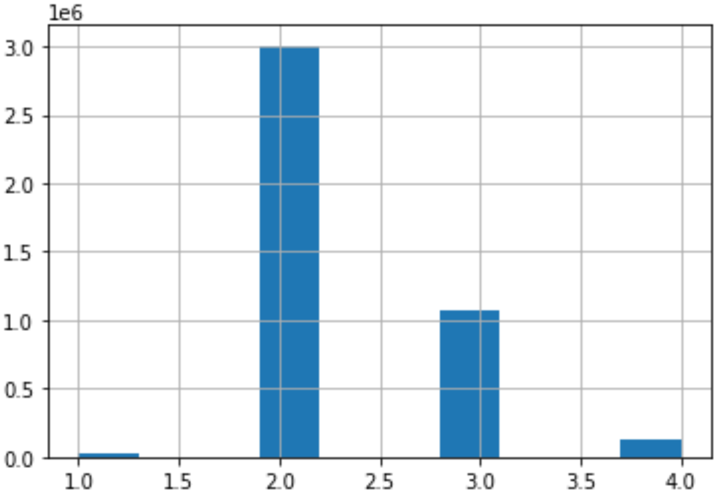
Data Visualizations:

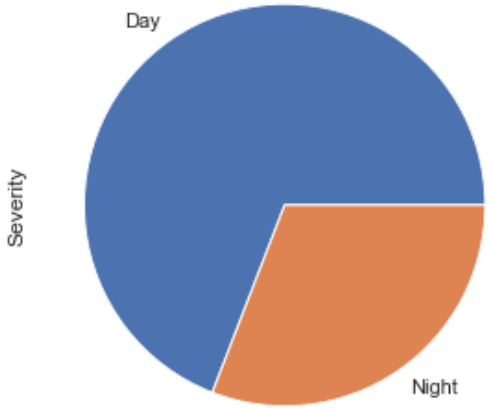
Severity is an important variable in our study and thus we started by visualizing that. This variable shows the severity of the accident, where 1 indicates the least impact on traffic (i.e., short delay because of the accident) and 4 indicates a significant impact on traffic (i.e., long delay). As we can see in the below graph most accidents have a little over short delays in traffic compared to long delays.

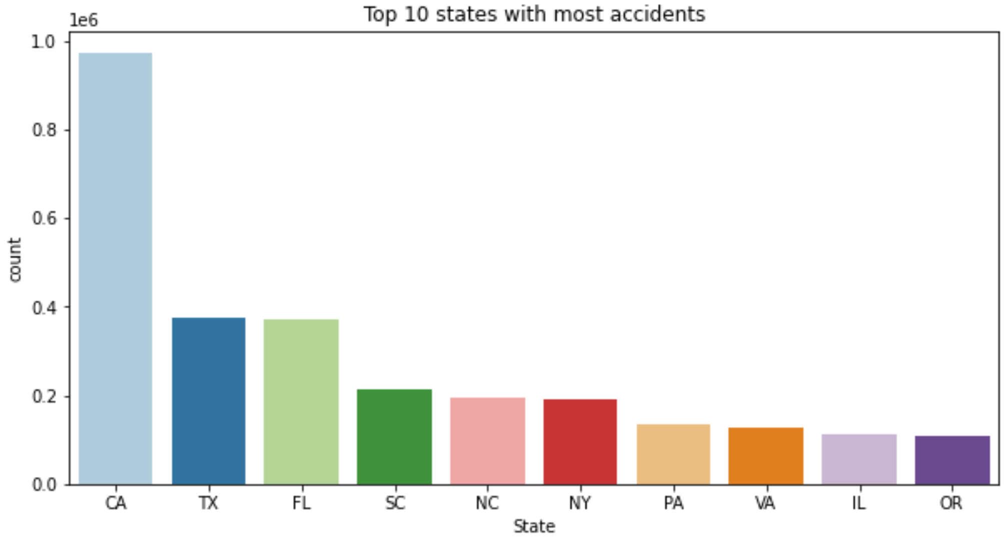
Another graph represents the top 10 states and cities with respect to most accidents.

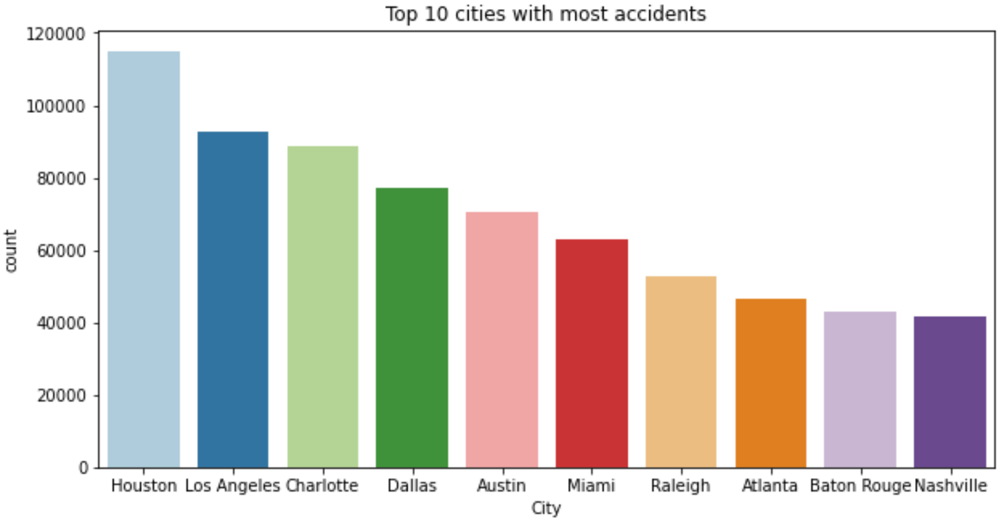
Also, we see the severity of accident is more during the day then night.

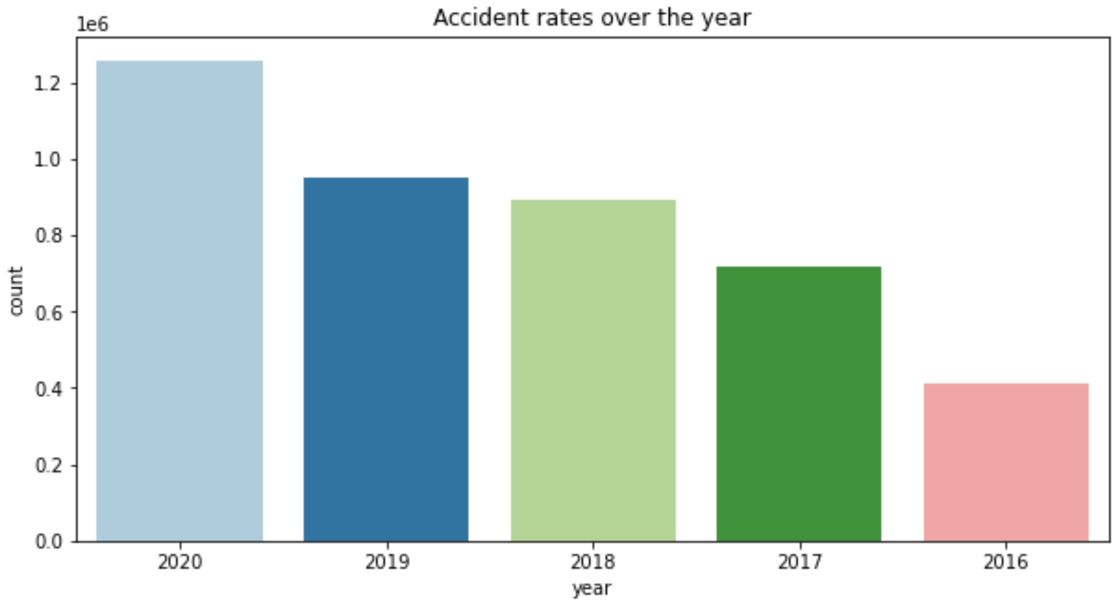
Lastly, the bar chart shows year-on-year increase of road accidents occurred in the USA. Despite 2020 being a covid -19 hit pandemic year has highest number of accidents.

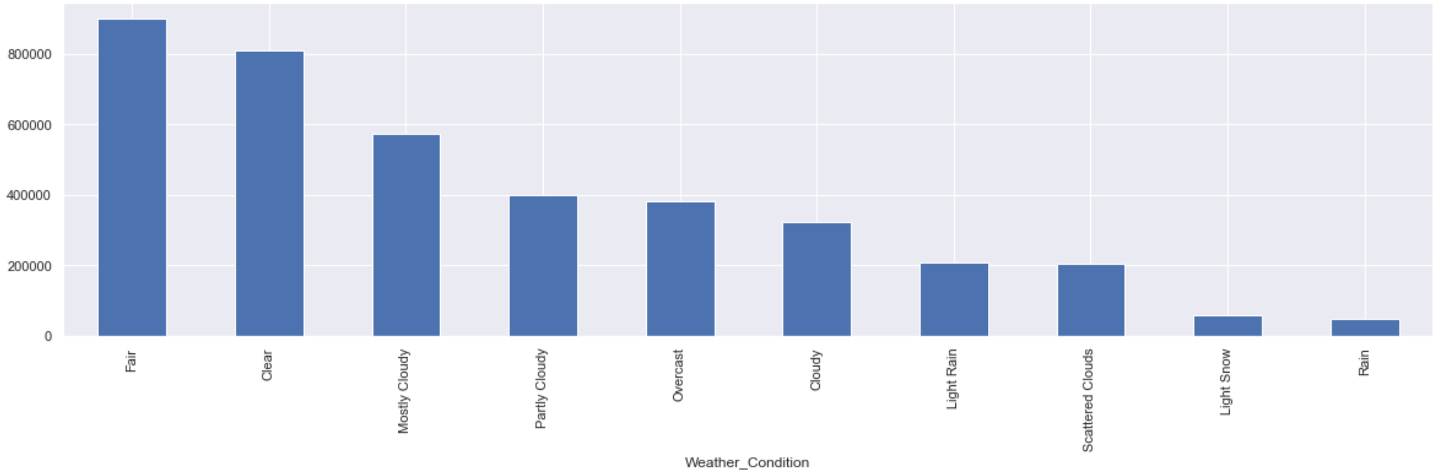

1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Day
Severity
Night


Top 10 states with most accidents
1e6
1.0
0.8
0.6
0.4
0.2
0.0
count
CA
TX
FL
SC
NC
NY
PA
VA
IL
OR
State


Top 10 cities with most accidents
120000
100000
80000
60000
40000
20000
0
count
Houston
Los Angeles
Charlotte
Dallas
Austin
Miami
Raleigh
Atlanta
Baton Rouge
Nashville
City





The above chart reveals the fact that, a greater number of accidents occurred during a clearer and fair-weather condition when compared to other critical weather conditions like rains, snows etc. It’s strange, but that’s the fact of the nature of accidents.

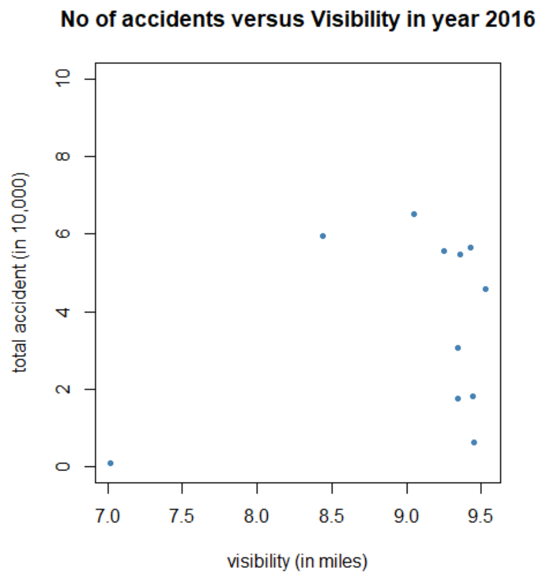
Plotting some statistical charts in R Studio

Let’s take a subset of the USA road accidents dataset for year 2016 to study the correlation between Average visibility and Total accidents.

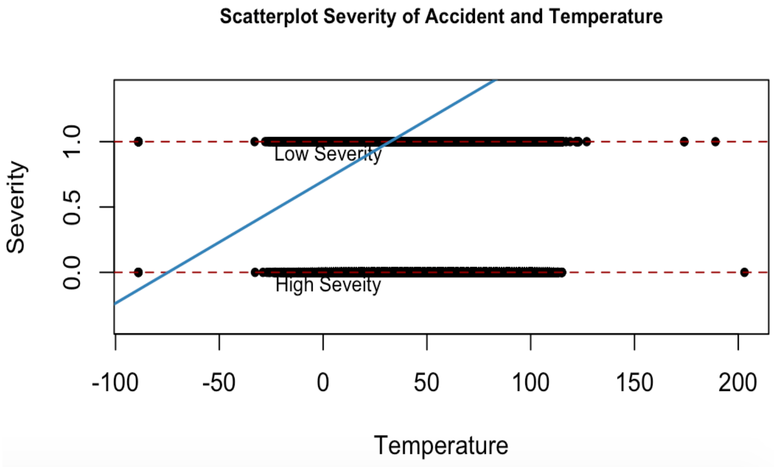
The below graph shows, despite clear visibility to a scale of 1 to 10, 1 being lower visibility index and 10 being highest visibility index, the total number of accidents occurred during the high visibility index for the given year 2016.

```
>df2016<-subset(Accidents_AggregationByYearByMonth, YEAR==2016)

>plot(x = df2016$AVG_VISIBILITY,
    y = df2016$TOTAL_ACCIDENT/10000,
    xlab = "visibility (in miles)",
    ylab = "total accident (in 10,000)",
    main = "No of accidents versus Visibility in
year 2016",
    ylim = c(0, 10),
    pch = 20,
    col = "steelblue")
>cor(df2016$TOTAL_ACCIDENT ,
df2016$AVG_VISIBILITY)
```



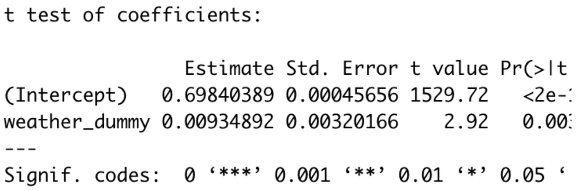
It shows, despite clear visibility to a scale of 1 to 10, 1 being lower visibility index and 10 being highest visibility index, the total number of accidents occurred during the high visibility index for the given year 2016.



5. Regression Modeling and Analyses:

Having cleaned up the data and done some visuals we now move to the regression modeling part. We will perform logistic regression on the traffic dataset to predict the severity of the accident. Through this study we will try to answer the following question: Is there any impact of weather condition on the probability of road accidents?
Since linear regression is easier to interpret, we will first run a simple lm model on the dataset to read the results.

- ***Code for the model in R:***
  test=lm(Severity_new~weather_dummy, data=accident_us)
  coeftest(test,vcov. = vcovHC)



```
t test of coefficients:

              Estimate Std. Error t value Pr(>|t
(Intercept)   0.69840389 0.00045656 1529.72   <2e-
weather_dummy 0.00934892 0.00320166    2.92   0.00
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘
```

- ***Interpretation of the model:***

Holding constant the P/I ratio, being an unclear weather increases the probability of severity of accident by 0.93%.
There might be an element of omitted variable bias so this could be a premature conclusion.

- ***The linear probability model has a major flaw***:
  - It assumes the conditional probability function to be linear.
  - This does not restrict dependent variable to lie between 0 and 1.

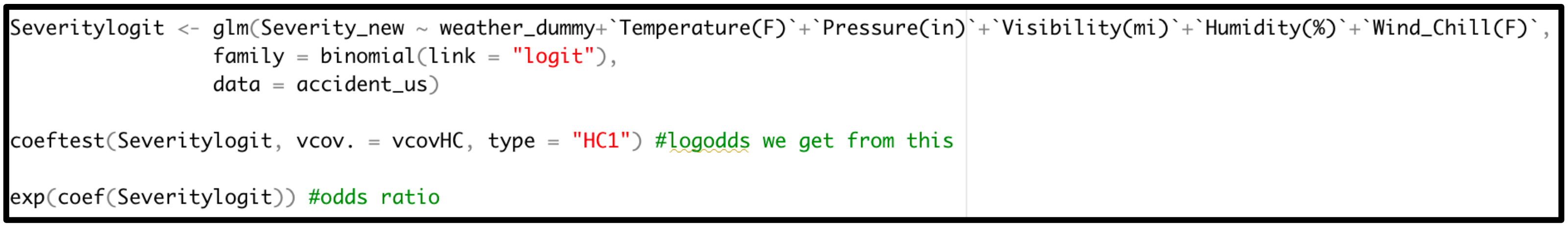
So, we will instead run a logistic regression model for more accurate analysis.

<u>Important points to note before running the logistic regression model:</u>

- For logistic regression the dependent variable must be binary. In our case Severity will be the dependent variable and to make it binary we have combined Severity 1 and 2 as low severity and made it a binary variable “1” and Severity of 3 and 4 have been combined as a high Severity and assigned the binary variable “0”.
- We must keep in mind that the model is suffering from omitted variable bias as the severity might be affected by the location i.e., the severity in one state might be high due to a natural calamity. A hurricane must have hit CA and not the eastern part of USA that is why severity of New York might be lower compared to CA.
- Another case of omitted variable bias must be that alcohol tax in Florida must

- be lower than the tax in MA and that is why we see more severe accident happening in Florida. Extra data needs to be studied for these.
- Additionally, to avoid overfitting or underfitting of the model all the significant variables like `Temperature(F)` + `Humidity(%)` + `Pressure(in)` + `Visibility(mi)` + `Wind_Speed(mph) have been included.
- Convert the weather_Condition categorical variable into a dummy variable since it has 113 different types of weather conditions mentioned. Variable "1" means a clear weather and "0" means all other type of weather conditions.
- Check the correlation between different variable to avoid multi-collinearity by running the cor.test function.

After following all the above steps, we can now finally run the logistic regression model using the glm function to study the prediction results.



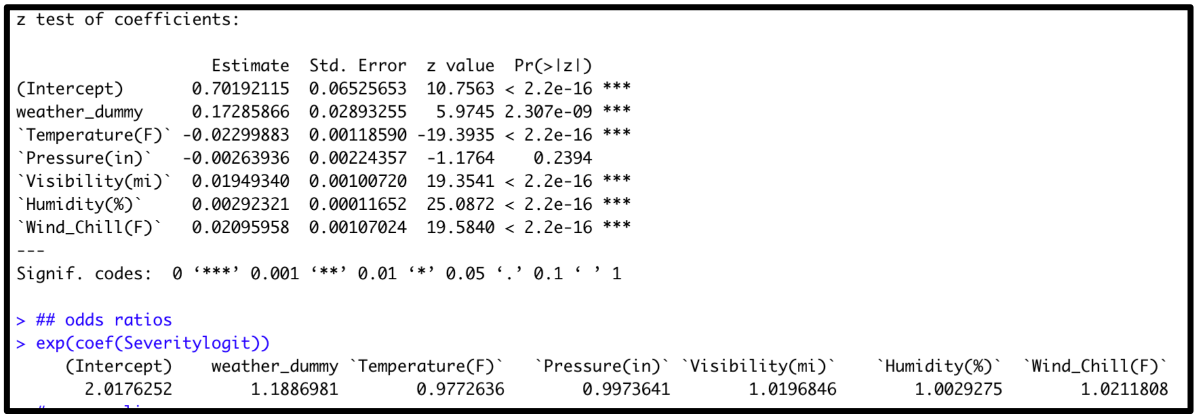
```
Severitylogit <- glm(Severity_new ~ weather_dummy+`Temperature(F)`+`Pressure(in)`+`Visibility(mi)`+`Humidity(%)`+`Wind_Chill(F)`,
                family = binomial(link = "logit"),
                data = accident_us)

coeftest(Severitylogit, vcov. = vcovHC, type = "HC1") #logodds we get from this

exp(coef(Severitylogit)) #odds ratio
```

__Analyzing the results:__

The log odds value for weather_dummy is 1.19. Thus, we can conclude: Clear weather has no significant impact on the Severity of the accidents. But had it been saying hypothetically 3.28 then we would conclude that Clear weathers are 3.28 times more likely than other weathers to result into a more severe accident.



```
z test of coefficients:

                   Estimate  Std. Error  z value  Pr(>|z|)    
(Intercept)       0.70192115  0.06525653  10.7563 < 2.2e-16 ***
weather_dummy     0.17285866  0.02893255   5.9745 2.307e-09 ***
`Temperature(F)` -0.02299883  0.00118590 -19.3935 < 2.2e-16 ***
`Pressure(in)`   -0.00263936  0.00224357  -1.1764    0.2394    
`Visibility(mi)`  0.01949340  0.00100720  19.3541 < 2.2e-16 ***
`Humidity(%)`     0.00292321  0.00011652  25.0872 < 2.2e-16 ***
`Wind_Chill(F)`   0.02095958  0.00107024  19.5840 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> ## odds ratios
> exp(coef(Severitylogit))
     (Intercept)    weather_dummy `Temperature(F)`   `Pressure(in)` `Visibility(mi)`    `Humidity(%)`  `Wind_Chill(F)` 
       2.0176252        1.1886981        0.9772636        0.9973641        1.0196846        1.0029275        1.0211808 
```

6. Conclusion:

The current paper helps us study the nature of road accidents in the USA and understand how the different variables contribute to a low severity or high severity accident. Our belief was that since highest number of accidents happen during clear weather condition it is an important variable in predicting the severity but on running the model the results concluded that there is no significant impact of clear weather on the severity of accident. There is a scope of improvement in the current paper. To name a few we further need to combine clear and fair weather as clear weather to get a better model result. Also, need to conduct a vaf test for checking correlation between the different variables and lastly to overcome the omitted variable bias find extra datasets that can help in a more accurate model result. Also, in future we can additionally run a prediction model to predict the time and the day of the week when accidents severity is the highest.

Overall, it was a good learning on utilizing regression models to do predictions on the dataset and moreover learn about the factors that can be a hindrance in reaching to a more accurate model result

Reference:

- https://www.statista.com/topics/3708/road-accidents-in-the-us/#dossierSummary
- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", arXiv preprint arXiv:1906.05409 (2019).
- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.
- https://www.kaggle.com/sobhanmoosavi/us-accidents