Eddieb Sadat
Professor Zadbood
EM622-WS Decision Making via Data Analysis
11/2/22

Midterm Project

Table of Contents

1. Introduction

In past EM-622 assignments, it was clear that data analysis and visualization is imperative to generate meaningful information out of raw data. Meaningful data is particularly important for companies or organizations who are aiming to increase their utility and decrease their costs. To achieve this, the raw data must be analyzed and prepared, and based upon the stakeholder needs, a list of actionable questions must be generated. Then, using proper graphical tools, visualizations of the relationships in the data are produced to help answer the actionable questions, and hopefully, provide the stakeholders with useful information to move forward with.

This report, created for the EM-622 midterm exam, utilizes the tools and knowledge acquired in the course to evaluate, process, and visualize data from a large repository of traffic accidents in the United States[1], and documents the process through the CRISP-DM methodology.

2. Business Understanding

2.1 Business Overview

As stated in the midterm requirements, the goal is to use the traffic accident data[1] to summarize the data and provide important insights by producing visuals (scatterplot, heatmap, geographic map, treemap, and a graph-of-choice). As such, the following steps were developed to successfully complete this project:

1. Preliminary Analysis
    a. Brief Overview of Data
    b. Identify the Potential Stakeholder(s)
    c. Develop Questions that are useful to the Stakeholder(s)
2. Data Preparation
    a. Filter out Data that is Irrelevant to the Question(s)
    b. Address Missing Values if Applicable
3. Visualization
    a. Scatterplot, Heatmap, Geographic map, Treemap, graph of choice.
4. Analysis and Conclusions

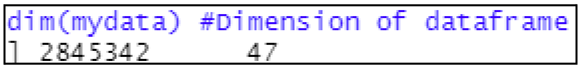
2.2 Stakeholders

Identifying potential stakeholders is important for developing meaningful results. By knowing the stakeholders, the project can be formulated to provide conclusions that pertain specifically to their needs.

Traffic information can be useful for any person or organization who relies on road conditions. For example, navigation apps aim to give customers the most reliable routes and time estimations, and thus may want to use traffic information to develop their AI. Taxi drivers may want to avoid certain times of day where accidents and traffic due to accidents are more likely to cause big delays for their customers. To narrow the scope, the specific target stakeholder for this project will be organizations who have an interest in keeping safe road conditions for citizens. As such, some important questions these stakeholders might want answered are “What causes the most severe accidents?” and “Do current measures alleviate accidents?”.

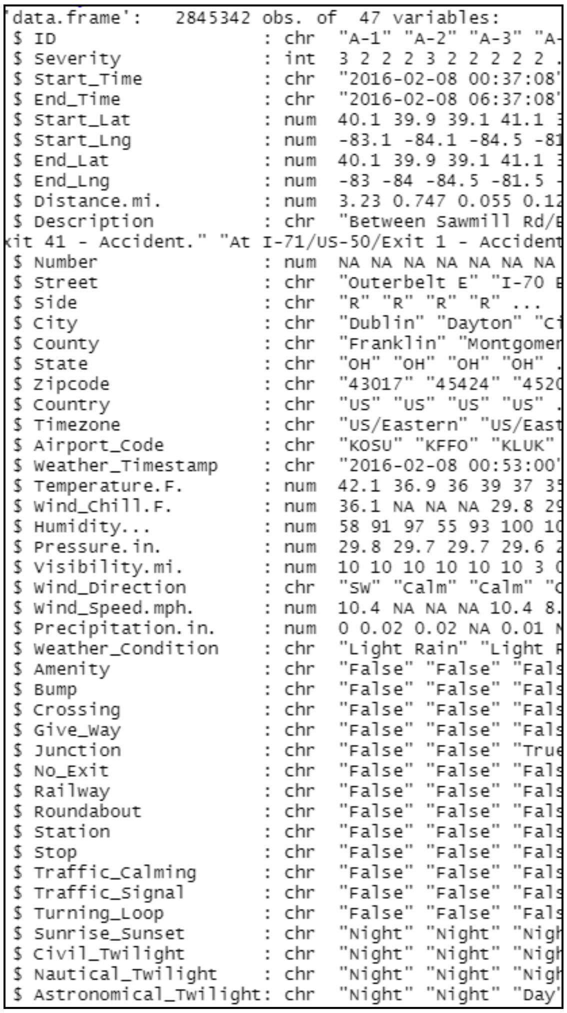
3. Data Understanding

After downloading the dataset and uploading it into R as a dataframe, the dimension is explored. In total, as of the date of this report, there are 47 columns and 2,845,342 rows.



```
dim(mydata) #Dimension of dataframe
] 2845342      47
```

On the download page for the traffic data is a detailed explanation of each variable[1]. Alongside the detailed explanations, the structure of the data frame is used to get a better understanding of what each variable represents.

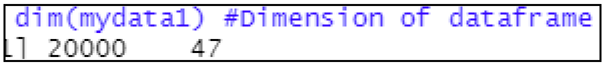


```
'data.frame':   2845342 obs. of  47 variables:
 $ ID                   : chr  "A-1" "A-2" "A-3" "A-
 $ Severity             : int  3 2 2 2 3 2 2 2 2 2 .
 $ Start_Time           : chr  "2016-02-08 00:37:08"
 $ End_Time             : chr  "2016-02-08 06:37:08"
 $ Start_Lat            : num  40.1 39.9 39.1 41.1 3
 $ Start_Lng            : num  -83.1 -84.1 -84.5 -81
 $ End_Lat              : num  40.1 39.9 39.1 41.1 3
 $ End_Lng              : num  -83 -84 -84.5 -81.5 -
 $ Distance.mi.         : num  3.23 0.747 0.055 0.12
 $ Description          : chr  "Between Sawmill Rd/E
xit 41 - Accident." "At I-71/US-50/Exit 1 - Accident
 $ Number               : num  NA NA NA NA NA NA NA 
 $ Street               : chr  "Outerbelt E" "I-70 E
 $ Side                 : chr  "R" "R" "R" "R" ...
 $ City                 : chr  "Dublin" "Dayton" "Ci
 $ County               : chr  "Franklin" "Montgomer
 $ State                : chr  "OH" "OH" "OH" "OH" .
 $ Zipcode              : chr  "43017" "45424" "4520
 $ Country              : chr  "US" "US" "US" "US" .
 $ Timezone             : chr  "US/Eastern" "US/East
 $ Airport_Code         : chr  "KOSU" "KFFO" "KLUK" 
 $ Weather_Timestamp    : chr  "2016-02-08 00:53:00"
 $ Temperature.F.       : num  42.1 36.9 36 39 37 35
 $ Wind_Chill.F.        : num  36.1 NA NA NA 29.8 29
 $ Humidity...          : num  58 91 97 55 93 100 10
 $ Pressure.in.         : num  29.8 29.7 29.7 29.6 2
 $ Visibility.mi.       : num  10 10 10 10 10 10 3 0
 $ Wind_Direction       : chr  "SW" "Calm" "Calm" "C
 $ Wind_Speed.mph.      : num  10.4 NA NA NA 10.4 8.
 $ Precipitation.in.    : num  0 0.02 0.02 NA 0.01 N
 $ Weather_Condition    : chr  "Light Rain" "Light R
 $ Amenity              : chr  "False" "False" "Fals
 $ Bump                 : chr  "False" "False" "Fals
 $ Crossing             : chr  "False" "False" "Fals
 $ Give_Way             : chr  "False" "False" "Fals
 $ Junction             : chr  "False" "False" "True
 $ No_Exit              : chr  "False" "False" "Fals
 $ Railway              : chr  "False" "False" "Fals
 $ Roundabout           : chr  "False" "False" "Fals
 $ Station              : chr  "False" "False" "Fals
 $ Stop                 : chr  "False" "False" "Fals
 $ Traffic_Calming      : chr  "False" "False" "Fals
 $ Traffic_Signal       : chr  "False" "False" "Fals
 $ Turning_Loop         : chr  "False" "False" "Fals
 $ Sunrise_Sunset       : chr  "Night" "Night" "Nigh
 $ Civil_Twilight       : chr  "Night" "Night" "Nigh
 $ Nautical_Twilight    : chr  "Night" "Night" "Nigh
 $ Astronomical_Twilight: chr  "Night" "Night" "Day"
```

There are clearly variables that will be useful in answering “What causes the most severe accidents?” and “Do current measures alleviate accidents?”, namely the location information (City, State, County…), Weather Condition, and traffic control systems (Bump, Crossing, Give_Way, Junction, No_Exit…), and Severity. There are also clearly variables that will be useless in answering the questions, such as ID and Country (all of the data is in the United States, so this variable is redundant).

4. Data Preparation:

Because there are 47 columns over 2.8 million rows, R is very slow to run every time it calls back to the dataframe. As such a subset of data was created by randomly selecting 20,000 rows from the 2.8million. A seed was set (to 0) so that results stay consistent throughout the project, and so that results can be replicable.

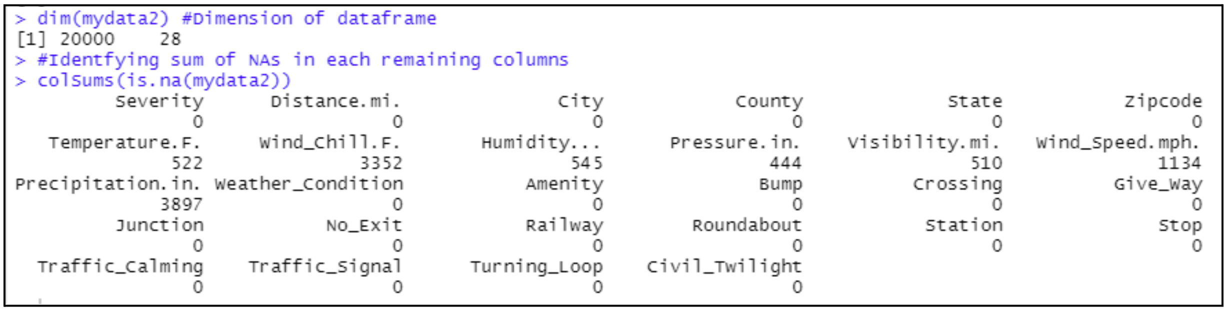


```
 dim(mydata1) #Dimension of dataframe
1] 20000    47
```

Next, unnecessary columns were removed, and are listed below:

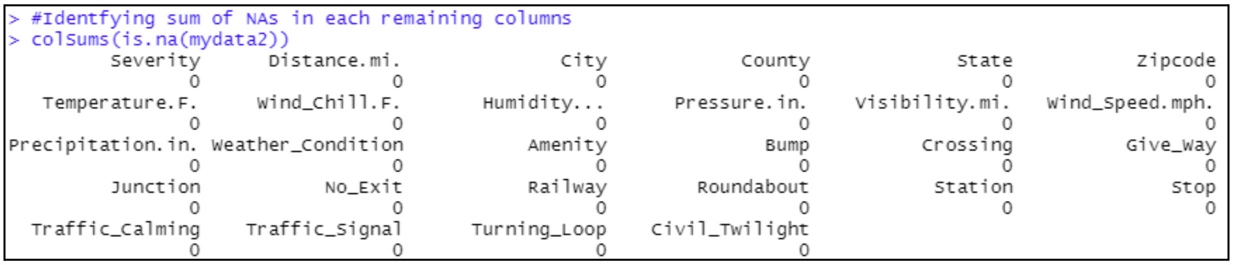
1. ID: Removed because ID of traffic accident is irrelevant for this project.
2. State_Time, End_Time: These data/time rows are not very meaningful for the questions → Can be replaced by Civil_Twilight.
3. Number, Street, Side: Street information is unneeded → Far too specific for this scope.
4. Country: Redundant, all data points are from the U.S.
5. Timezone: Irrelevant information.
6. Description: Too difficult to break down reliably for automated analysis.
7. Start_Lat, Start_Lng, End_Lat, End_Lng: Related to where the traffic starts/stops and is too specific → can be replaced by County location.
8. Airport_Code: Nearby Airports irrelevant, too specific.
9. Weather_Timestamp: Time when weather occurs can be useful, but it is too specific and thus unreliable to use for this scope.
10. Sunrise_Sunset, Nautical_Twilight, Astronomical_Twilight: Not useful, Civil_Twilight gives insight into natural light and when artificial lighting is needed → can replace the other three columns.
11. Wind_Direction: Seems irrelevant to the scope of this project.

After removing all unwanted rows and columns, the remaining data frame contains 20,000 rows and 28 columns. Next, the sum of NAs is calculated for each column to determine if further action is needed.



```
> dim(mydata2) #Dimension of dataframe
[1] 20000    28
> #Identfying sum of NAs in each remaining columns
> colSums(is.na(mydata2))
         Severity       Distance.mi.               City             County              State            Zipcode 
                0                  0                  0                  0                  0                  0 
    Temperature.F.       Wind_Chill.F.        Humidity...        Pressure.in.      Visibility.mi.     Wind_Speed.mph. 
              522               3352                545                444                510               1134 
 Precipitation.in.  Weather_Condition            Amenity               Bump           Crossing           Give_Way 
             3897                  0                  0                  0                  0                  0 
         Junction            No_Exit            Railway         Roundabout            Station               Stop 
                0                  0                  0                  0                  0                  0 
  Traffic_Calming     Traffic_Signal       Turning_Loop      Civil_Twilight 
                0                  0                  0                  0 
```

Temperature.F., Wind_Chill.F., Humidity(%), Pressure.in., Visibility.mi., Wind_Speed.mph., and Precipitation.in. are all missing values. Luckily, all are numerical columns, which means a simple solution is to replace the NAs with the means of each column.



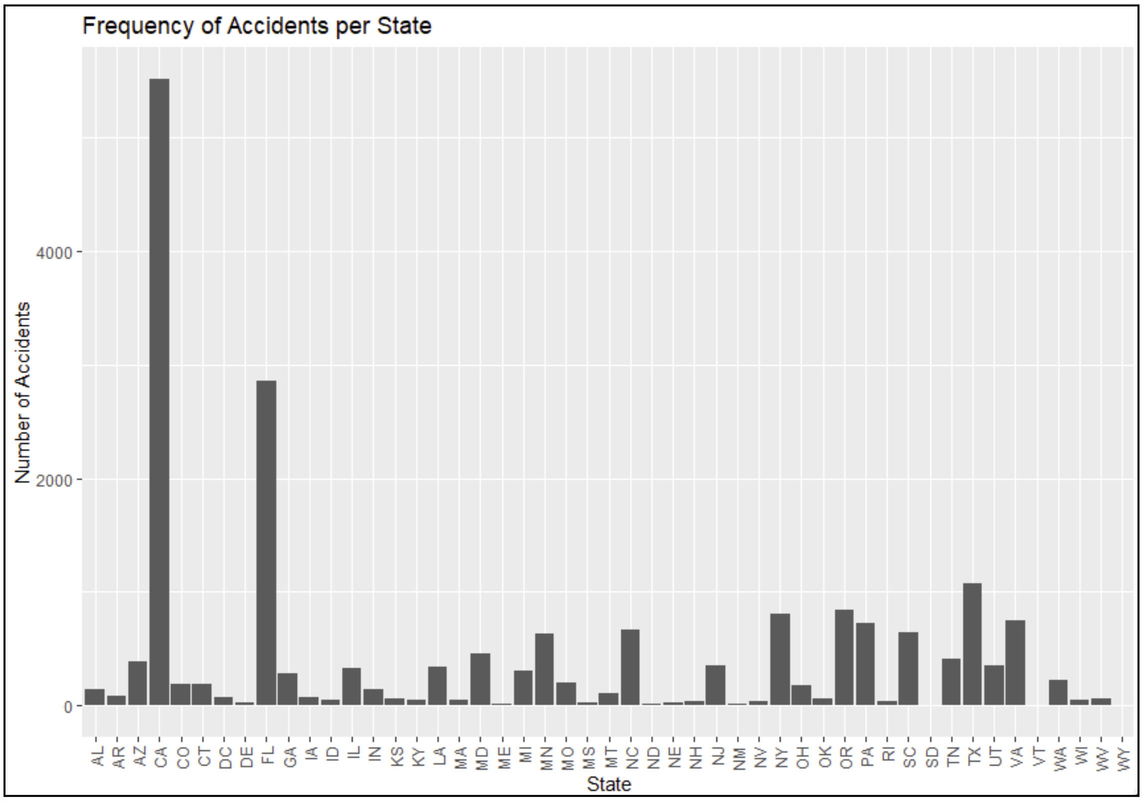
```
> #Identfying sum of NAs in each remaining columns
> colSums(is.na(mydata2))
         Severity       Distance.mi.               City             County              State            Zipcode 
                0                  0                  0                  0                  0                  0 
    Temperature.F.       Wind_Chill.F.        Humidity...        Pressure.in.      Visibility.mi.     Wind_Speed.mph. 
                0                  0                  0                  0                  0                  0 
 Precipitation.in.  Weather_Condition            Amenity               Bump           Crossing           Give_Way 
                0                  0                  0                  0                  0                  0 
         Junction            No_Exit            Railway         Roundabout            Station               Stop 
                0                  0                  0                  0                  0                  0 
  Traffic_Calming     Traffic_Signal       Turning_Loop      Civil_Twilight 
                0                  0                  0                  0 
```

After this, no more rows contain NAs, and as such, no more data preparation is needed. There may be additional data manipulation during the modeling phase, but those are not global changes, and thus, will not be mentioned in this section.

5. Modeling

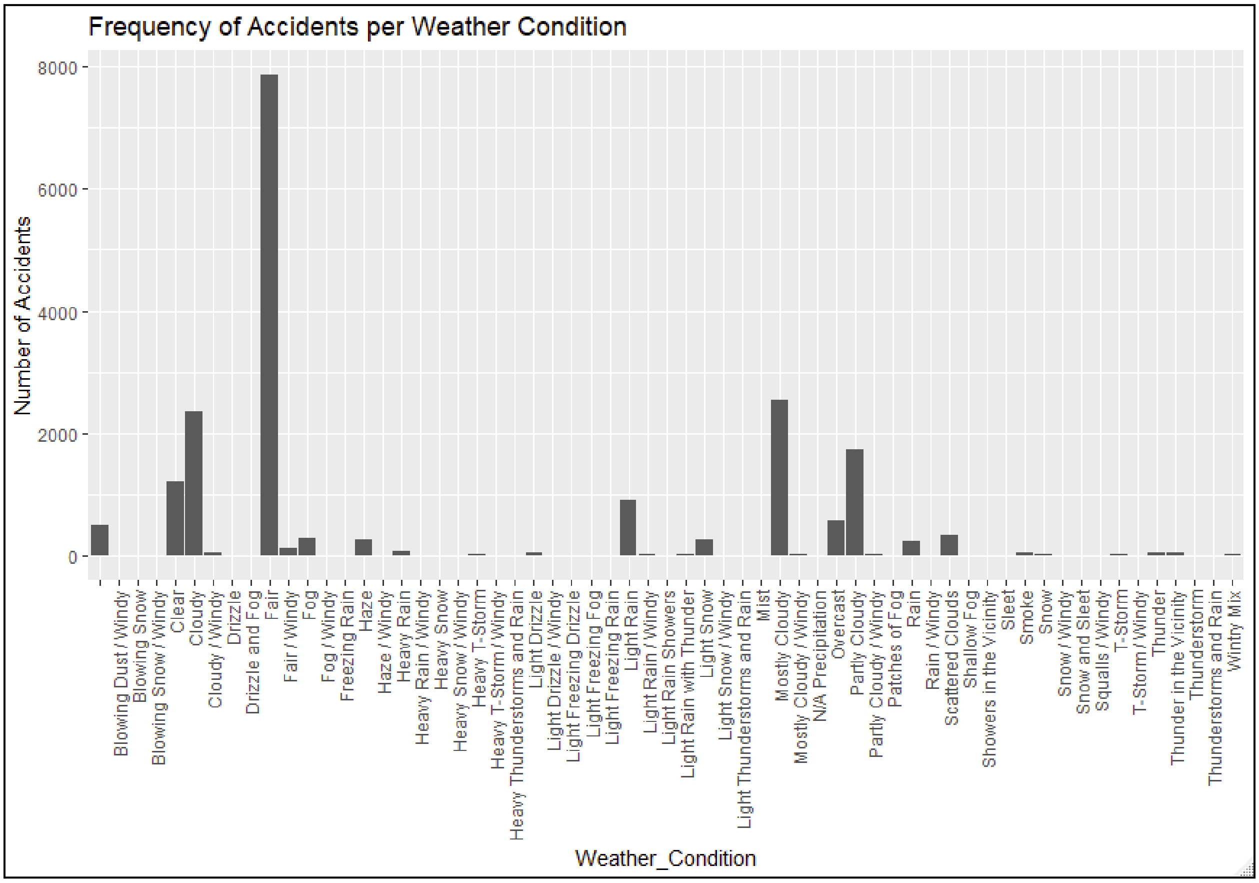
5.1 Graph of Choice: Barplot

The frequency of accidents may be linked to particular states, each of whom have different factors, such as varying road laws and population. Generating a barplot of states can be useful in determining accident frequencies per state.



Clearly, California has the largest number of accidents, well over 400% of the majority of other states. Florida as well has an abnormally large number of accidents, with about 200% more than the majority of states. At first glance, this may make sense because both California and Florida have very large populations compared to other states, but New York and Texas, who both have comparable population sizes, do not have anywhere near the frequency as California and Florida. This means that California and Florida could have unique factors that are causing so many accidents compared to other states, indicating that population may be less significantly correlated with accidents. However, it could also be due to the methods the accident data were collected on. Perhaps California and Florida have more robust accident and traffic detection systems than other states, causing a data bias. More research into how the data was collected will be needed to develop a sure conclusion.

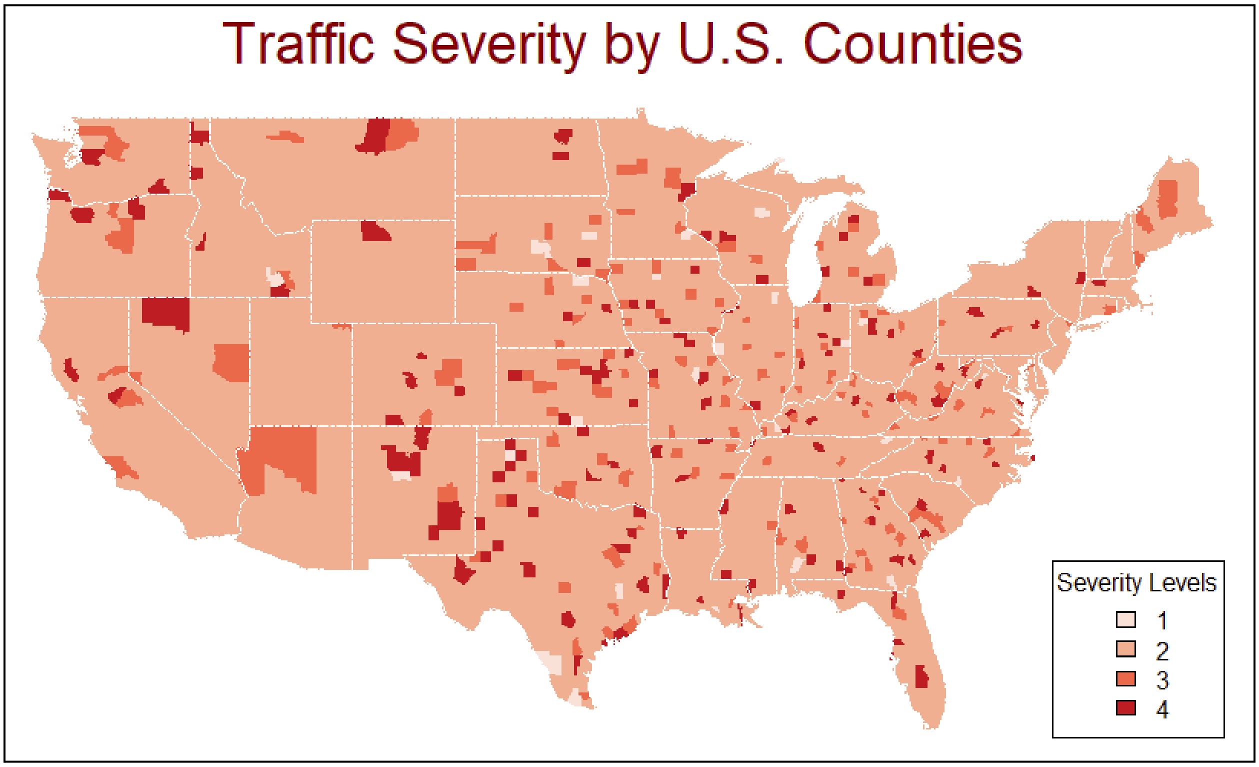
Another factor that can have an influence on the frequency of accidents are weather conditions. It is reasonable to assume that certain weather conditions, such as rain and snow, will increase the frequency of accidents due to increased hazards and risks. Another barplot was generated, except using weather conditions as bins.



Clearly the largest number of accidents occur on days with ‘Fair’ weather conditions, which may indicate that accidents are likely to occur in tamer/favorable weather conditions. The other notable weather conditions are ‘Clear’, ‘Cloudy’, ‘Light Rain’, ‘Mostly Cloudy’, and ‘Partly Cloudy’. These results are also surprising, as these weather conditions are not generally considered to be very hazardous. However, this makes sense, as most people generally do not travel during hazardous weather conditions unless necessary. With a higher number of people driving in better weather conditions, the frequency of accidents in these conditions will also be larger.

5.2 Geographic Heatmap

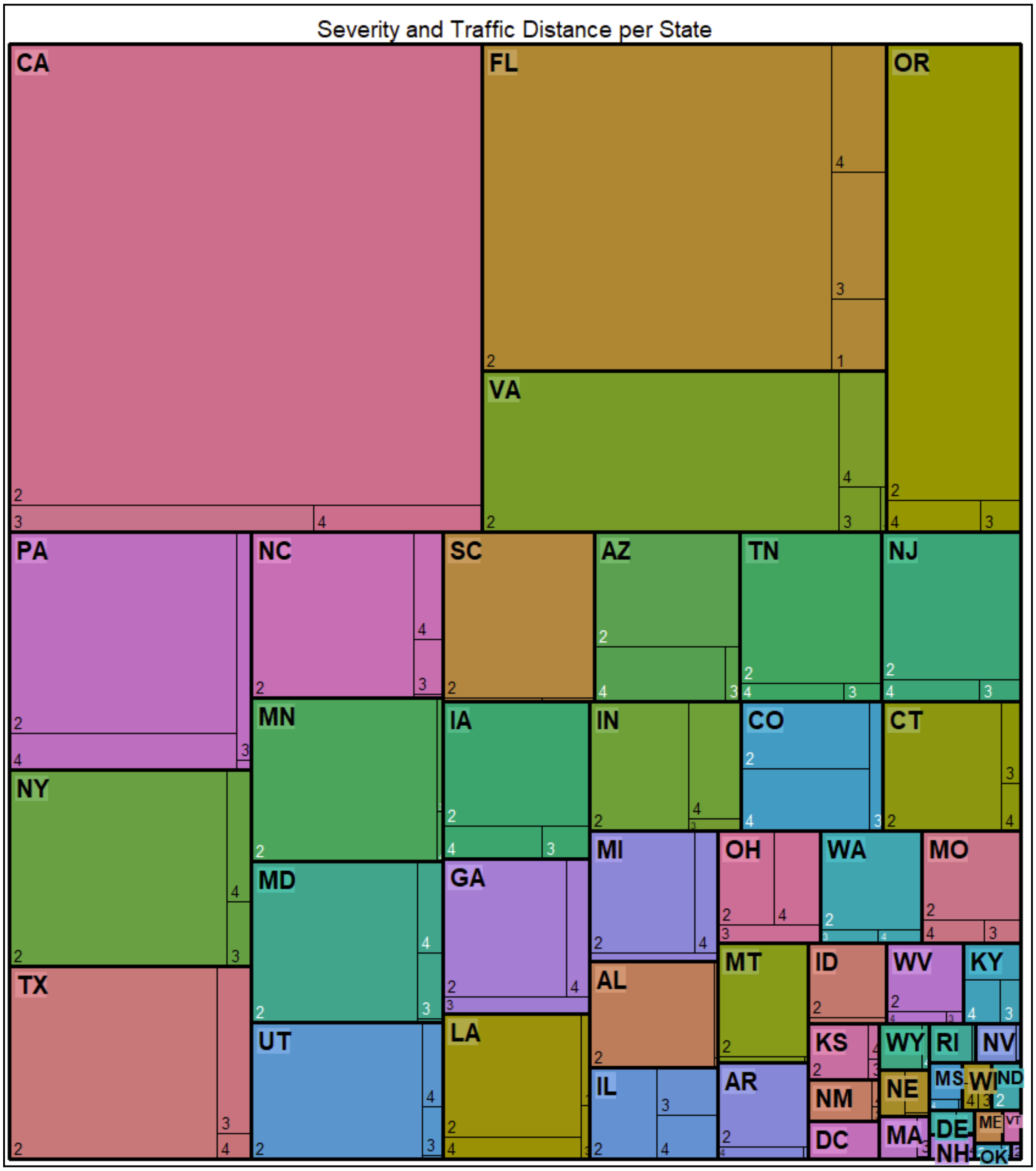
Instead of the frequency of accidents per state, the severity of accidents per state (by county) can give insight into how accident severity correlates to each state. Since these values rely on the average severity per county, it can provide good comparison between states regardless of the accident records available per state. In other words, each state can be compared to each other more effectively.



From the geographic heatmap, it is clear that an overwhelming number of counties have accidents that have a severity level of 2 (slightly severe). Some counties have higher severities of 3 or 4, but they do not seem to be solely related to population or congestion. For example, LA county in California seems to have a severity of 3, but other locations with comparable population, like NYC, still only has a severity of 2. This indicates that population may have some role in accident severity (especially since many states have higher severity in their more populated counties), but there are clearly other factors that affect accident severity.

5.3 Treemap

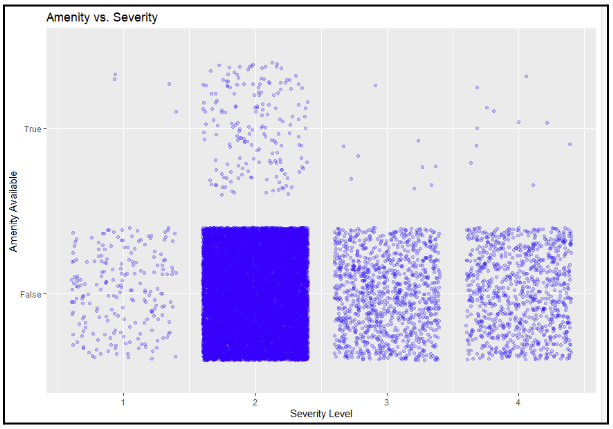
A treemap can help identify the relationship between severity, traffic distance and civil twilight between all mapped states.

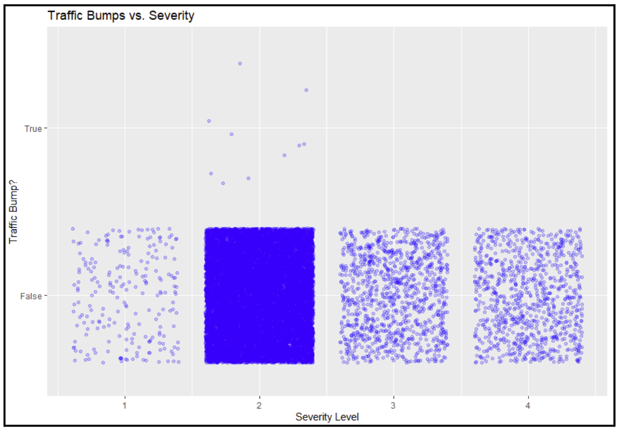


Firstly, because civil twilight indicated that most accidents occurred during the day, there are no color changes within each state (civil twilight was used as vColor). From the treemap, it is clear that California and Florida had the most traffic distances. Furthermore, each state shows a majority of severities of 2, further supporting the results from the Geographical Heatmap. To prevent label overlap, the labels were aligned to the top and bottom left[3].

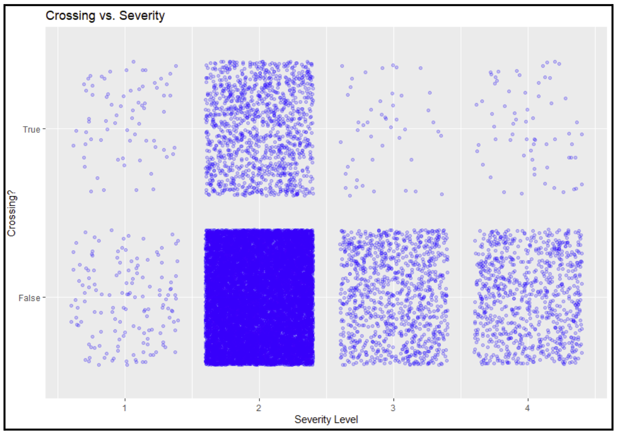
5.4 Scatterplot

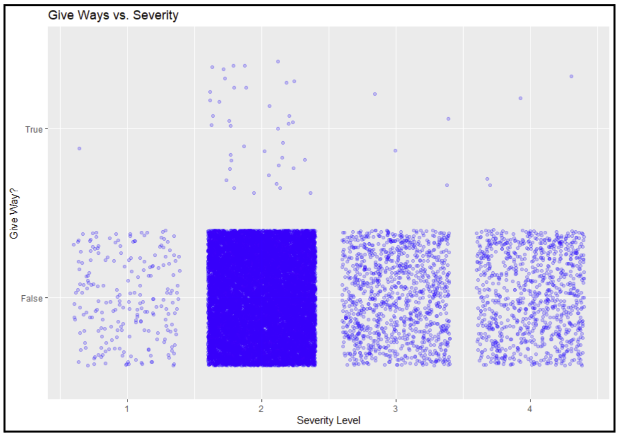
Scatterplots can be useful for displaying relationships between discrete variables, especially when jittering and alpha are applied.



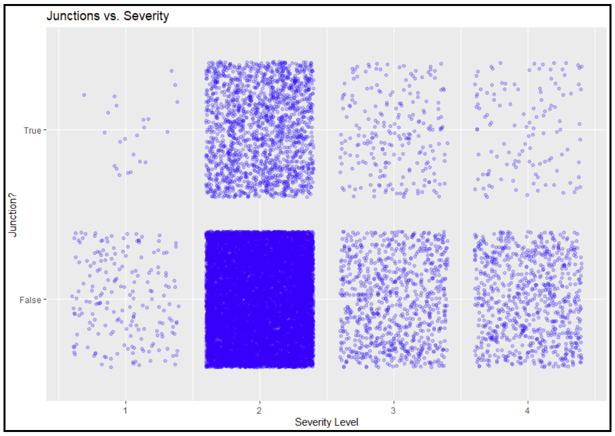


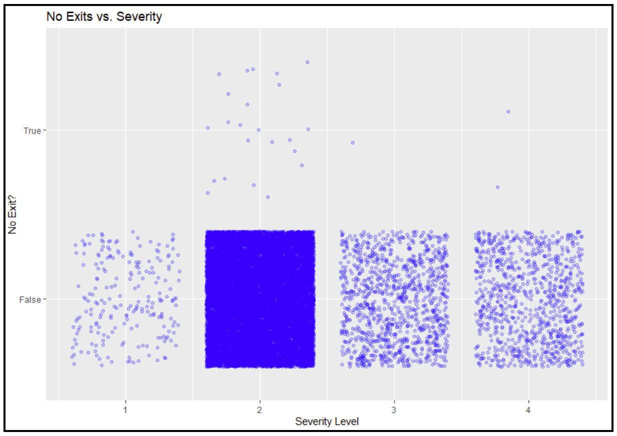
From the graph on the left, accidents are likely to occur where amenities are not nearby, and are likely to be accidents with a severity value of 2. From the graph on the right, accidents are likely to occur where there are no traffic bumps, and are likely to be an accident severity of 2.



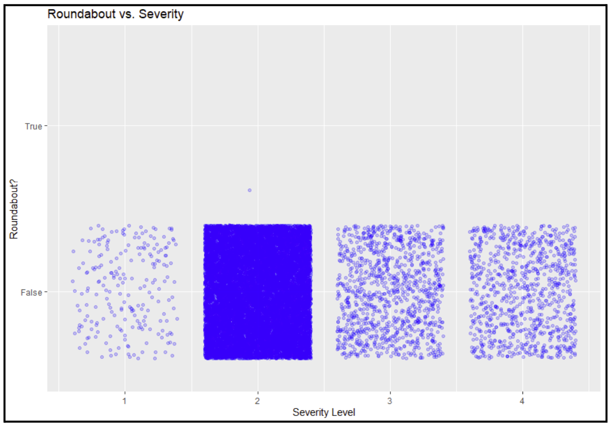


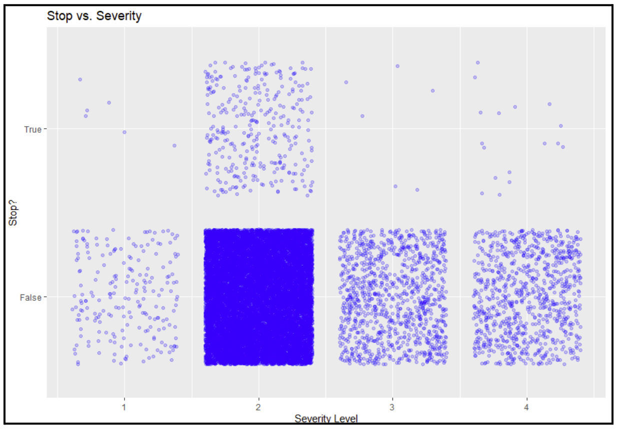
From the graph on the left, accidents are likely to occur where there are no crossings for accidents with severities 2, 3, and 4, whereas severity of 1 has no correlation to whether a crossing exists or not. Accidents with or without crossings are more likely to occur with a severity of 2. From the graph on the right, accidents are likely to occur where there are no give ways, and are likely to have a severity of 2.



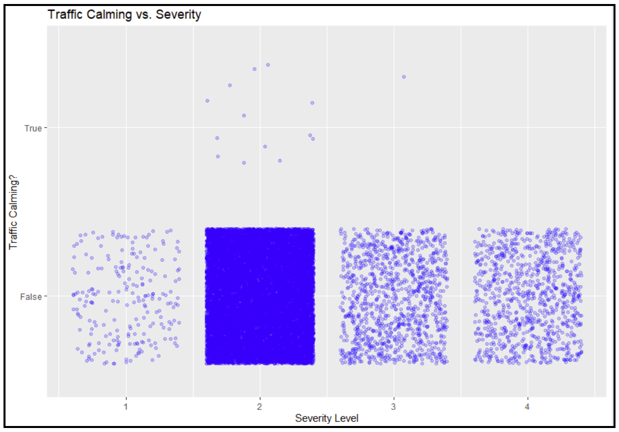


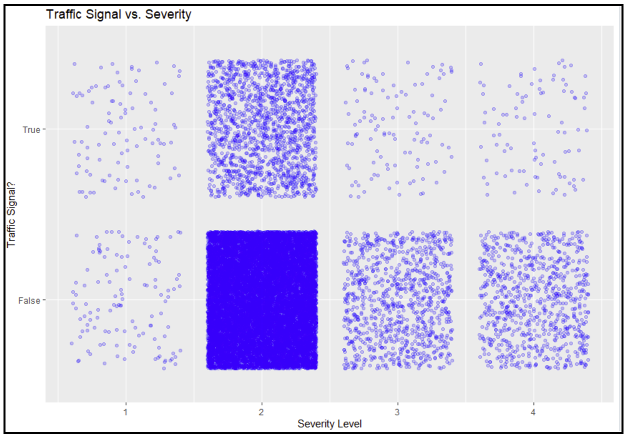
From the graph on the left, accidents are more likely to occur where there are no junctions, and the accident severity is likely to be 2 whether or not there is a junction. From the graph on the right, accidents are more likely to occur where there are no No Exits, and are likely to be a severity of 2.





From the graph on the left, accidents are more likely to occur when there are no roundabouts, and the accidents are likely to be a severity of 2. From the graph on the right, accidents are more likely to occur where there are no stops, and are likely to be a severity of 2.

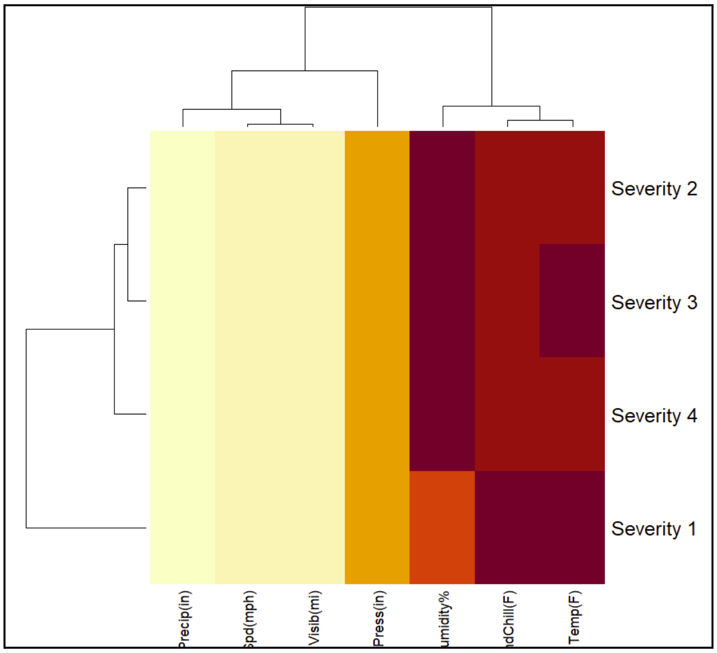




From the graph on the left, accidents are likely to occur when there is no traffic calming, and are likely to be a severity of 2. From the graph on the right, accidents are more likely to occur when there is no traffic signal, though there are noticeably increased numbers of accidents even with traffic signals. Both with or without traffic signals will likely yield severity of 2.

5.5 Table Heatmap

Thus far, no visuals were produced for the numerical variables. A table heatmap is perfect for visualizing relationships between numerical values in data. After some data modification, a heatmap was produced to compare Severity levels with the numerical column. One note, I tried everything I could source on how to make the x labels more visible, but unfortunately nothing amounted to anything (they seemed to also pose an issue on the canvas example). The numerical values are means of their columns.



From the heatmap, it seems that each column value is heavily related in relation to Severity levels. There is also very little variability inside all the columns, which makes sense considering the graphs above, which showed that generally weather conditions did not seem to affect Severity. However, it clear that Wind Chill, Temperature, Humidity, and Pressure have decent correlation to causing accidents in general, since they are clearly darker than the other three columns.

6. Evaluation

Surprisingly, it seems that population size of states and bad weather conditions are not as strongly, positively correlated to traffic accidents and accident severity as one may normally, and reasonably, think. Most accidents occur during better weather conditions because more people are driving, which will naturally increase accident frequency. Furthermore, the Geographical heat map and Treemap both show that despite population sizes, accident severity is very likely to be 2, which indicates that it is relatively safe to drive, and traffic caused by accidents will clear up relatively quickly most of the time.

One positive note is that there is a clear indication from the scatterplots that the current traffic systems are very clearly decreasing the number of accidents. Very clearly in all of the scatterplot of different traffic control methods (e.g. speed bumps, stop signs…), show lower frequencies of accidents when they are deployed. However, it also may be important to focus on improving Crossing, Junction, and Traffic Signal systems, as they indicate higher accidents when they are deployed, as compared to other traffic control methods.

7. Deployment

The results of this report are confident enough to be deployed to a real-life stakeholder. Of course, the results should also be taken with some skepticism, especially since we are not 100% of the data collection methods, and thus, cannot verify the integrity of the data collected. However, from an analysis standpoint, this report should provide useful, brief insight into traffic accidents across the United States.

Sources:

[1] - https://smoosavi.org/datasets/us_accidents

[2] - https://statisticsglobe.com/r-sample-random-rows-in-data-frame

[3] - https://stackoverflow.com/questions/18128370/how-do-i-change-the-position-of-labels-in-the-r-treemap