Midterm Exam EM 622

Name: Harsh Shah

Professor: Amineh Zadbood, Ph.D.

Topic: US-accidents Analysis Report

Introduction

Here I have used a country wide traffic accident dataset, which covers 49 states of the United States. As accidents should be avoided and even after it traffic is largely affected which is also a major problem. The data is continuously being collected from February 2016, using several data providers, including two APIs which provide streaming traffic event data. These APIs broadcast traffic events captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. The dataset contains almost 3 million entries of accident and traffic affected related to it. In this report various trends for the traffic affected areas due to accident and even trends related to temperature, wind speed etc. for each state during the time of accident. Dataset also have some of NA values which we will try to ignore for better results. At the end this report will be useful to answer many questions related to this dataset and get a clear view of this dataset. Main goal is to provide detailed EDA for the dataset. Whole analysis is done by using R studio.

Business Understanding

This report project is prepared for EM 622 Midterm exam. The main objective is to provide a detailed EDA for the dataset. Various trends and questions will be presented with the help of graphs for better visualizations. Some of the major questions answered in this report are as follows:

1. What are distributions of accidents according to states in United States?

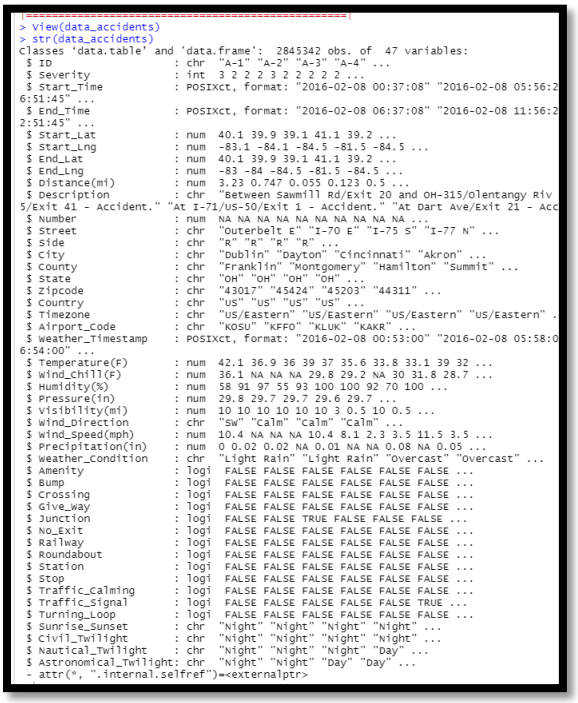
2. What are the distributions of variables like temperature, wind speed etc. according to the different states?

3. Does weather condition is related to accident or does visibility factor plays any role?

4. How severity of traffic affected among cities of states?

Data Understanding

Here analysing the structure of data. There are total 47 columns with unique variables and there are in total 2845342 rows of data. Below is attached the type distribution for each of the variable among the 47. Our dataset is divided in types of “num”, “chr”, “logi” and “int”. Description of each variable is taken from dataset reference website. Here dataset is taken from Kaggle and continuously being updated since 2016 till 2021.



```
> View(data_accidents)
> str(data_accidents)
Classes 'data.table' and 'data.frame':  2845342 obs. of  47 variables:
 $ ID                    : chr  "A-1" "A-2" "A-3" "A-4" ...
 $ Severity              : int  3 2 2 2 3 2 2 2 2 2 ...
 $ Start_Time            : POSIXct, format: "2016-02-08 00:37:08" "2016-02-08 05:56:2
6:51:45" ...
 $ End_Time              : POSIXct, format: "2016-02-08 06:37:08" "2016-02-08 11:56:2
2:51:45" ...
 $ Start_Lat             : num  40.1 39.9 39.1 41.1 39.2 ...
 $ Start_Lng             : num  -83.1 -84.1 -84.5 -81.5 -84.5 ...
 $ End_Lat               : num  40.1 39.9 39.1 41.1 39.2 ...
 $ End_Lng               : num  -83 -84 -84.5 -81.5 -84.5 ...
 $ Distance(mi)          : num  3.23 0.747 0.055 0.123 0.5 ...
 $ Description           : chr  "Between Sawmill Rd/Exit 20 and OH-315/Olentangy Riv
5/Exit 41 - Accident." "At I-71/US-50/Exit 1 - Accident." "At Dart Ave/Exit 21 - Acc
 $ Number                : num  NA NA NA NA NA NA NA NA NA NA ...
 $ Street                : chr  "Outerbelt E" "I-70 E" "I-75 S" "I-77 N" ...
 $ Side                  : chr  "R" "R" "R" "R" ...
 $ City                  : chr  "Dublin" "Dayton" "Cincinnati" "Akron" ...
 $ County                : chr  "Franklin" "Montgomery" "Hamilton" "Summit" ...
 $ State                 : chr  "OH" "OH" "OH" "OH" ...
 $ Zipcode               : chr  "43017" "45424" "45203" "44311" ...
 $ Country               : chr  "US" "US" "US" "US" ...
 $ Timezone              : chr  "US/Eastern" "US/Eastern" "US/Eastern" "US/Eastern" .
 $ Airport_Code          : chr  "KOSU" "KFFO" "KLUK" "KAKR" ...
 $ Weather_Timestamp     : POSIXct, format: "2016-02-08 00:53:00" "2016-02-08 05:58:0
6:54:00" ...
 $ Temperature(F)        : num  42.1 36.9 36 39 37 35.6 33.8 33.1 39 32 ...
 $ Wind_Chill(F)         : num  36.1 NA NA NA 29.8 29.2 NA 30 31.8 28.7 ...
 $ Humidity(%)           : num  58 91 97 55 93 100 100 92 70 100 ...
 $ Pressure(in)          : num  29.8 29.7 29.7 29.6 29.7 ...
 $ Visibility(mi)        : num  10 10 10 10 10 10 3 0.5 10 0.5 ...
 $ Wind_Direction        : chr  "SW" "Calm" "Calm" "Calm" ...
 $ Wind_Speed(mph)       : num  10.4 NA NA NA 10.4 8.1 2.3 3.5 11.5 3.5 ...
 $ Precipitation(in)     : num  0 0.02 0.02 NA 0.01 NA NA 0.08 NA 0.05 ...
 $ Weather_Condition     : chr  "Light Rain" "Light Rain" "Overcast" "Overcast" ...
 $ Amenity               : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Bump                  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Crossing              : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Give_Way              : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Junction              : logi  FALSE FALSE TRUE FALSE FALSE FALSE ...
 $ No_Exit               : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Railway               : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Roundabout            : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Station               : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Stop                  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Traffic_Calming       : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Traffic_Signal        : logi  FALSE FALSE FALSE FALSE FALSE TRUE ...
 $ Turning_Loop          : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Sunrise_Sunset        : chr  "Night" "Night" "Night" "Night" ...
 $ Civil_Twilight        : chr  "Night" "Night" "Night" "Night" ...
 $ Nautical_Twilight     : chr  "Night" "Night" "Night" "Day" ...
 $ Astronomical_Twilight: chr  "Night" "Night" "Day" "Day" ...
 - attr(*, ".internal.selfref")=<externalptr>
```

1. ID: unique identifier for the accident
2. Severity: range from 1 to 4 (1 means low impact on traffic)
3. Start time: shows start time of the accident
4. End time: When impact on traffic was over
5. Start_Lat: latitude point in Gps coordinate
6. Start_Lng: shows longitude of start of accident
7. End_Lat: shows latitude in in Gps
8. End_Lng: Shows end longitude
9. Distance(mi): length of road affected by accident
10. Description: Human provided description of accident
11. Number: Shows the street number in address field.
12. Street: Shows the street name in address field.
13. Side: Shows the relative side of the street (Right/Left) in address field.
14. City: Shows the city in address field.
15. County: Shows the county in address field.
16. State: Shows the state in address field.
17. Zipcode: Shows the zipcode in address field.
18. Country: Shows country which is US

19. Timezone: location timezone
20. Airport_Code
21. Weather_Timestamp: Show when weather report was recorded
22. Temperature(F)
23. Wind_Chill(F)
24. Humidity(%)
25. Pressure(in): air pressure
26. Visibility(mi): shows visibility in miles
27. Wind_Direction: Shows wind direction whether it is calm, SW etc.
28. Wind_speed: indicating the speed of wind
29. Precipitation(in)
30. Weather_Condition: Shows the weather condition like rain, snow, thunderstorm, fog, etc.
31. Amenity: whether amenity is present or not (True or False)
32. Bump: bump present or not
33. Crossing: present or not
34. Give_way
35. Junction: whether there is junction or not.
36. No_exit
37. Railway
38. Roundabout
39. Station: whether there is a station in nearby region or not.
40. Stop: whether there is a stop sign
41. Traffic_calming
42. Traffic_signal
43. Turning_loop
44. Sunrise_sunset: Whether it was night or day at time of accident
45. Civil twilight: showing day or night according to civil twilight
46. Nautical twilight
47. Astronomical twilight

Data Preparation

Here data was clean and not much of cleaning was required. Though we will analyse whether the column State and county have missing values or not. These both are our main column to work. Missing values in all other columns can be dropped while analysing as it will not affect our results as there are multiple entries for each state. Rest all columns missing value can be dropped according to the need for analysis.

```
> sum(is.na(data_accidents$State))
[1] 0
> sum(is.na(data_accidents$County))
[1] 0
```

There are no missing values in any of the columns and our data is fine for working. Analysis of missing value was done for two main Column State and county and it was found there was not a single missing value. As missing value on one of this column would have question on the data entry that whether it was dummy or by mistake. Here we will check for any duplicates in the “Id” column. The main goal is to check that there are no duplicate entries.

```
> length(unique(data_accidents$ID))
[1] 2845342
```

Here there are 2845342 unique values which match our row count hence there are no duplicate values. Hence we can start our analysis.

Here some of the column name has also being changed according to the analysis as in of the names there were some brackets. Major transformation was done in the weather condition column as light rain, heavy rain all were considered as rain same goes with many other weather conditions.

```
data_accidents$weather_updated <- ifelse(grepl("Cloudy",data_accidents$Weather_Condition),"Cloudy",ifelse(grepl("Rain",data
data_accidents$weather_updated <- ifelse(grepl("Snow",data_accidents$weather_updated),"Snow",data_accidents$weather_updated
data_accidents$weather_updated <- ifelse(grepl("Fair",data_accidents$weather_updated),"Fair",data_accidents$weather_updated
data_accidents$weather_updated <- ifelse(grepl("T-Storm",data_accidents$weather_updated),"Thunderstorm",data_accidents$weat
data_accidents$weather_updated <- ifelse(grepl("Thunder",data_accidents$weather_updated),"Thunderstorm",data_accidents$weat
data_accidents$weather_updated <- ifelse(grepl("Thunderstorm",data_accidents$weather_updated),"Thunderstorm",data_accidents
data_accidents$weather_updated <- ifelse(grepl("Fog",data_accidents$weather_updated),"Fog",data_accidents$weather_updated)
#view(data_accidents)
```

Also for scatter plot and tree map random samples were generated for better analysis and smooth run of R studio. NA values were dropped for some of the columns

Modelling

Here starting our modelling process we have first installed necessary libraries as shown in Figure 1. As our dataset is pretty large we have used “data.table” library for easing running process of R studio.

```
library(data.table)
library(maps)
library(RColorBrewer)
library(mapproj)
library(ggplot2)
library(tidyverse)
library(dplyr)
library(stringr)
library(treemap)
```

Figure 1

Coming to our first question **what are distributions of accidents according to states in United States?** For this here advanced geographic plot is shown with the help of “maps” library. No NA values were dropped for this as we for this we just need a record belonging to an accident to get the over view of accidents distribution. Figure 2 indicates the advanced geographical map.

Here from the figure 2 we can see that there are more than 200 accidents for every county of California and moving further Florida, Oklahoma, Wisconsin, Michigan, Ohio and coming to east, New york and New Jersey are also have more than 200 accidents. On the other hand states like North Dakota, South Dakota, Nebraska, and Kansas have very less accidents compare to other states over a period for almost 4 years. Even Oregon also has least accidents probably even less than 20 in major counties over period of 4 years.

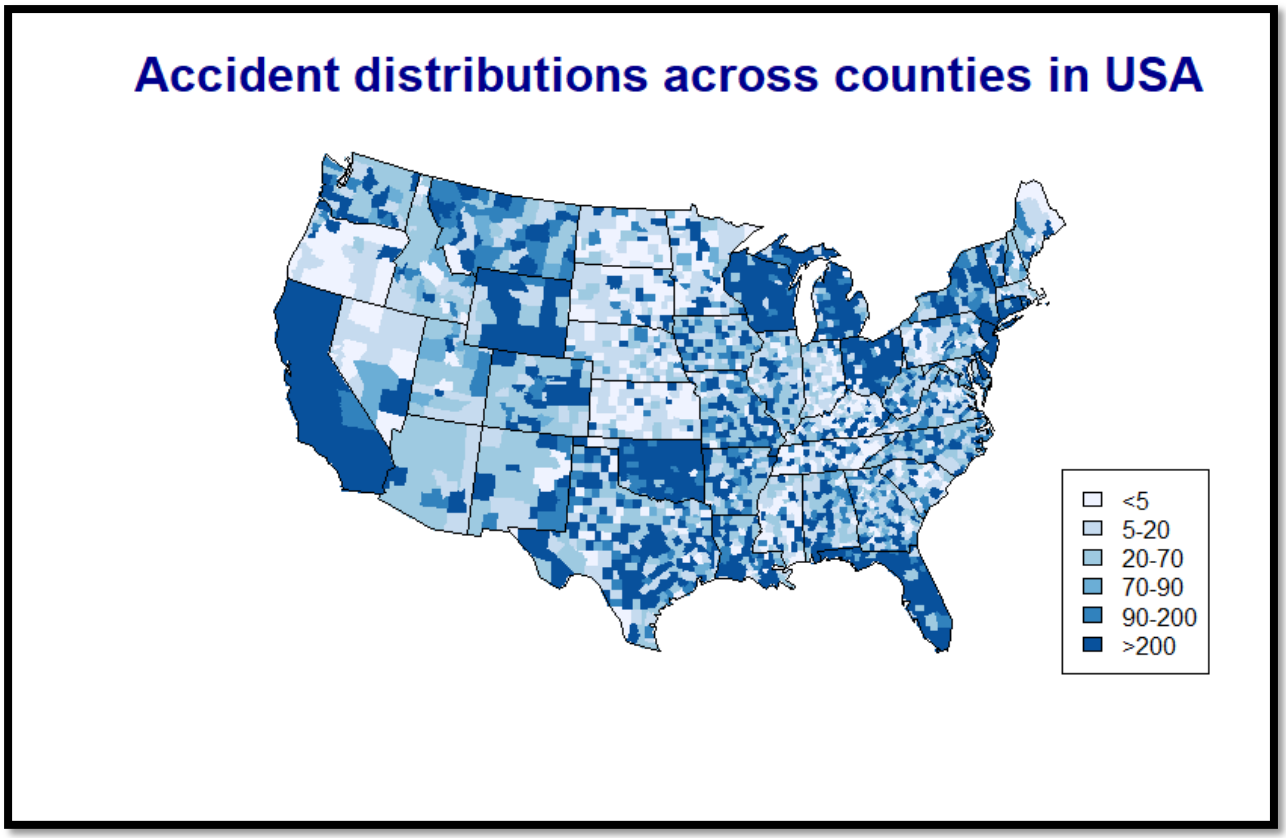


Figure 2

Coming to our second question **what are the distributions of variables like temperature, wind speed etc. according to the different states?** Here average for variable temperature, wind chill, humidity, severity, wind speed, pressure and visibility are plotted over a Heat map for all 49 states. Here we have dropped all missing values from the dataset and took average for generalized approach and have a general approach for what was the weather distribution in comparison to other states during the time of accident.

Here severity and visibility were also plotted in comparison to other states as how traffic was affected in comparison to other states and also the visibility in compare to other state.

Here as seen in Figure 3 we can observe that for temperature variable LA, Florida, Texas, Arizona have the highest temperature which is obvious. As seen from the wind chill variable was also highest among this states. However we see that severity was pretty low in comparison to other states which is, there was very less impact on traffic in this states. On the other side in terms of severity states like Wisconsin, Illinois etc. had the greatest impact on traffic among all other states. Also humidity in this states were pretty low. In terms of visibility Texas, Arizona, Los angles were having highest visibility as this maybe the reason of weather there. In terms of the pressure lowest was for the state Wyoming, Colorado etc.

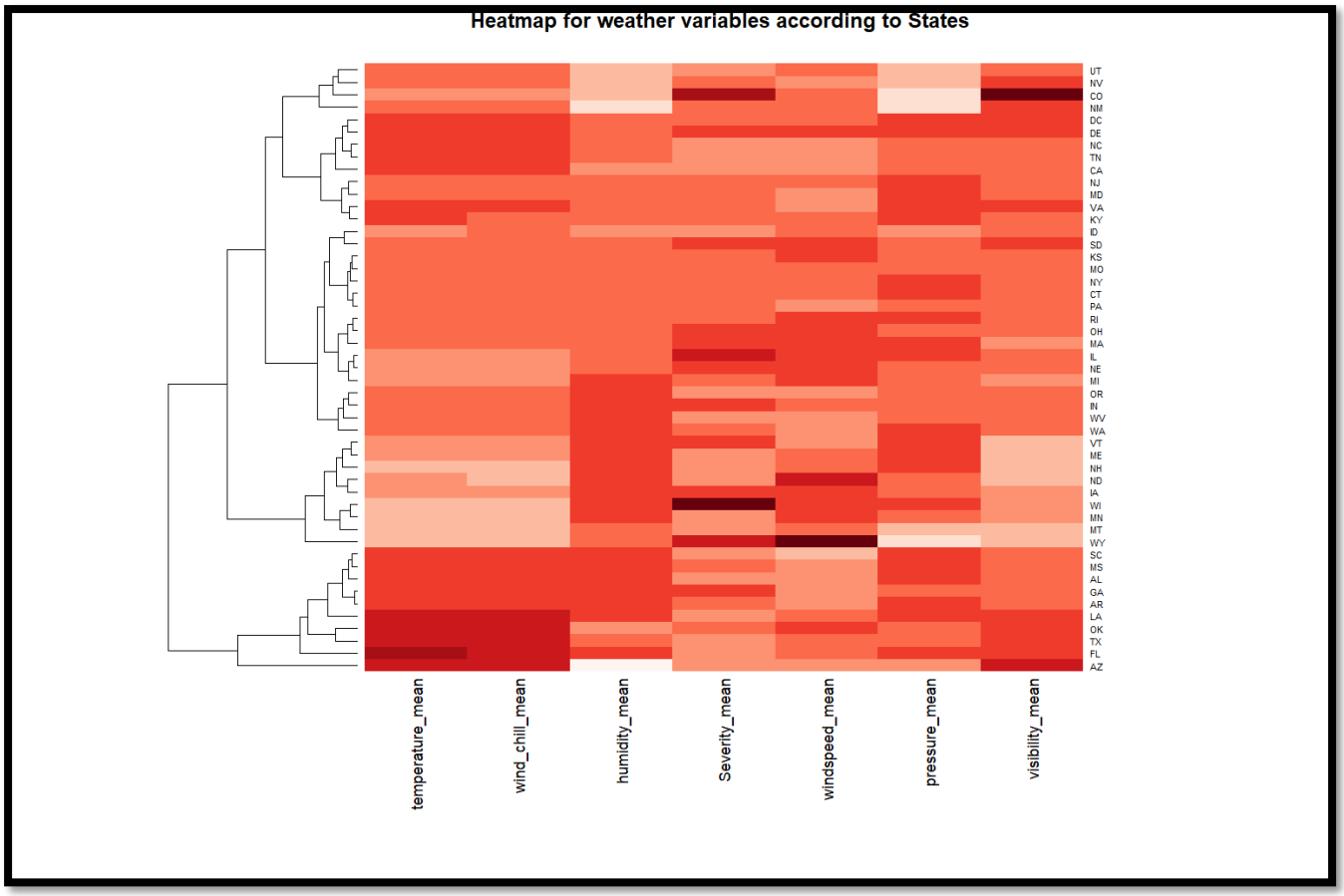


Figure 3

Here addressing our third question **does weather condition is related to accident or does visibility factor plays any role?** Figure 4 shows here the scatter plot with weather conditions visibility and side of the road and even whether it was day or night during the time of accident. Here we have dropped all missing values and took a random sample of 20000 for better analysis and to run our R studio more efficiently.

Here some data transformation is done for the weather column. For example three weather conditions like light rain, heavy rain or moderate rain all are considered as rain. Same with T-storm, thunderstorm is considered as thunderstorm. Same goes with some other weather. This is done for better analysis and to get an overall picture of the dataset. Due to big size of dataset it is hard to get a clear picture for so many variables.

As from the figure it is clearly visible that visibility doesn't have any trends with the accidents as even with the visibility of 10 there is large amount of accidents. Majority of accidents have severed effect on traffic. Also day and night factor is also almost equal. Though it seems that there is more number of accidents in right lane. Also there is cluster for weather condition like snow, rain which has low visibility.

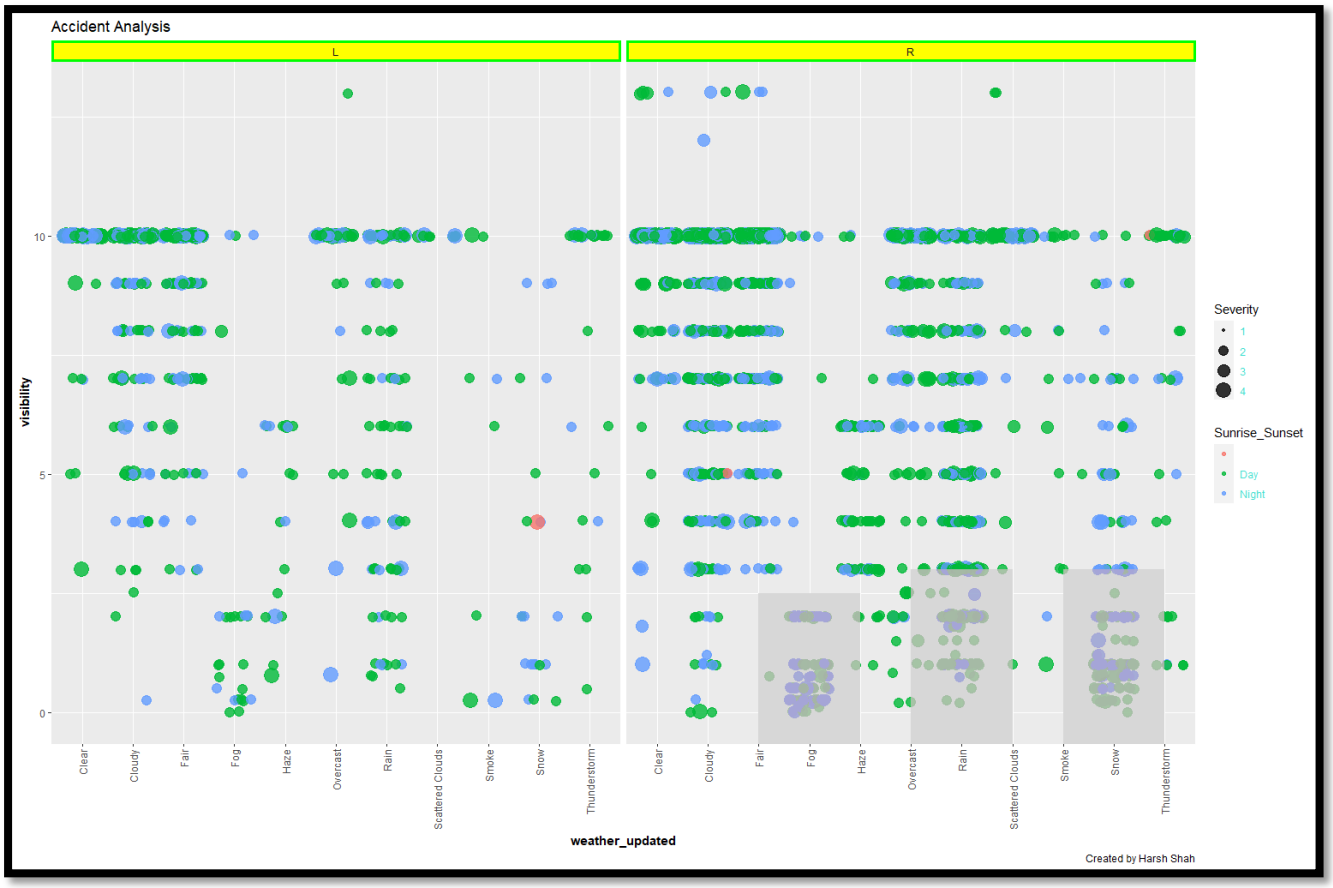


Figure 4

Now analysing the question **how severity of traffic affected among cities of states?** Here a tree map for the following is being plotted. Here severity for each case for each county is being added, resulting the maximum sum would be the most affected city. Here also the random data which is 10000 observations is taken for better analysis and easy running of R studio. This will give a major idea for the trend. It is clear from the figure that Los Angeles, Miami, Orlando, Dallas, and Portland were some of the major cities that were affected. Figure 5 shows here detailed tree map.

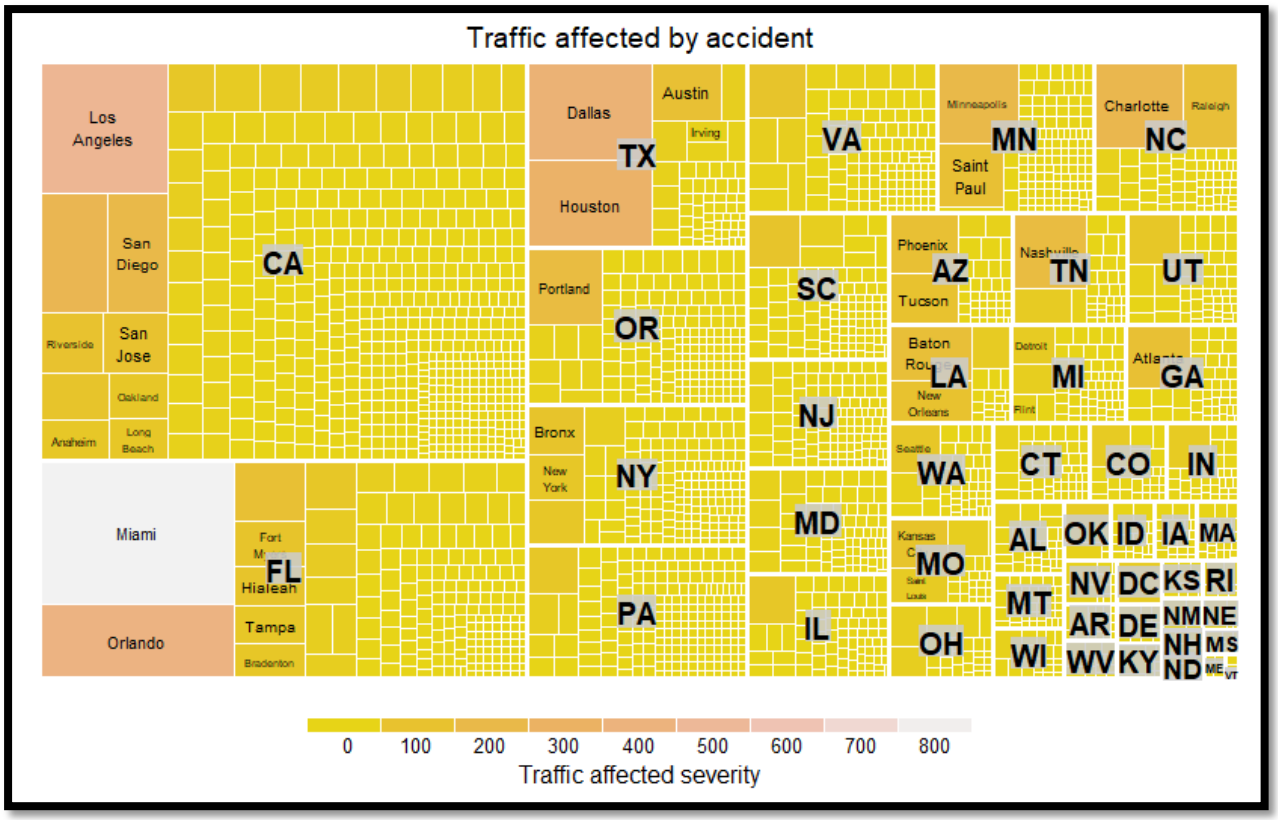


Figure 5

Conclusions

- From the basic visualizations it seems California, Florida were some the states with maximum number of accidents.
- In terms of severity also cities from California like Los Angeles, San Diego, San Jose etc. and cities from Florida like Miami, Orlando were the top most affected cities in terms of traffic.
- For Texas accidents were comparatively low in comparison to other states but cities were highly affected due to accident in terms of accident.
- By analysing the scatter plot it seems that even when there was good visibility there were plenty of accidents and weather conditions like rain , snow, thunderstorm play some role in accidents.
- It was also observed there were more accidents on right side of the lane.

References

https://smoosavi.org/datasets/us_accidents