622-MIDTERM
Ashley Dsilva

Using CRISP-DM methodology for the given problem

Steps in CRISP-DM

1. Business understanding
2. Data understanding
3. Data preparation
4. Modelling/Evaluation
5. Deployment

1. Business understanding:

This analysis focuses on the accidents that take place in the United States and using various plots understands various aspects that contributes to these accidents. This analysis can be very useful in understanding the areas that are accident prone, the cause of these accidents and how it can be used to predict future incidents and avoid them. It also focuses on various weather attributes and their effect on these accidents. This is a real time data which can be used by other third-party applications to assess and come up with precautions that can be carried out to minimize the frequency or the damage done.

2. Data understanding:

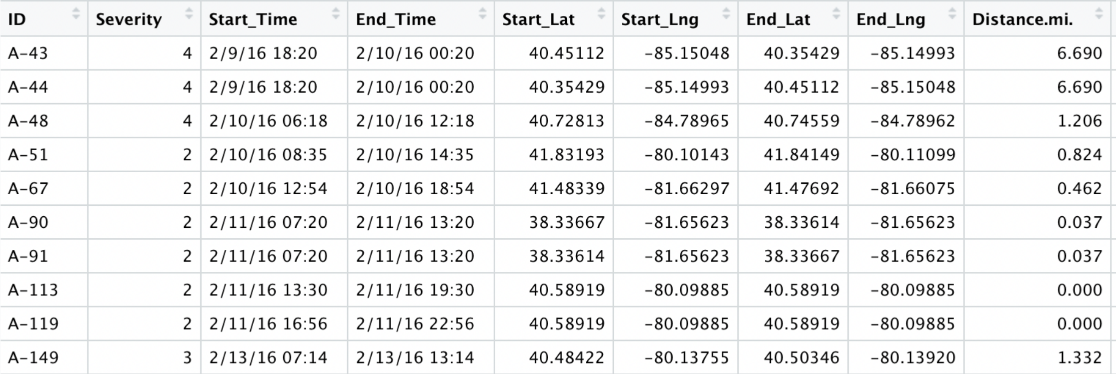
Dataset used: US accidents (2016-2021)
The dataset has 47 columns and 1.5 million rows of data. Not all the attributes are used in the analysis so ill only showcase the ones that are being used.

Attributes:

1. ID
2. Severity
3. Distance (mi)
4. Number
5. City
6. County
7. State
8. Country
9. Time zone
10. Wind chill (F)
11. Humidity (%)
12. Pressure (in)
13. Visibility (in)
14. Wind direction
15. Wind speed (mph)

16. Precipitation (in)
17. Weather condition
18. Sunrise / Sunset

Unclean dataset preview:

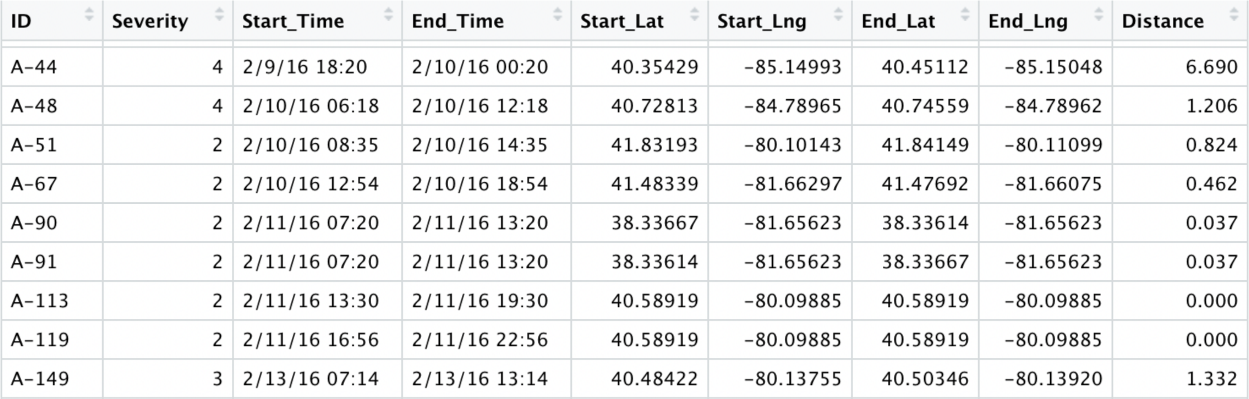


ID	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance.mi.
A-43	4	2/9/16 18:20	2/10/16 00:20	40.45112	-85.15048	40.35429	-85.14993	6.690
A-44	4	2/9/16 18:20	2/10/16 00:20	40.35429	-85.14993	40.45112	-85.15048	6.690
A-48	4	2/10/16 06:18	2/10/16 12:18	40.72813	-84.78965	40.74559	-84.78962	1.206
A-51	2	2/10/16 08:35	2/10/16 14:35	41.83193	-80.10143	41.84149	-80.11099	0.824
A-67	2	2/10/16 12:54	2/10/16 18:54	41.48339	-81.66297	41.47692	-81.66075	0.462
A-90	2	2/11/16 07:20	2/11/16 13:20	38.33667	-81.65623	38.33614	-81.65623	0.037
A-91	2	2/11/16 07:20	2/11/16 13:20	38.33614	-81.65623	38.33667	-81.65623	0.037
A-113	2	2/11/16 13:30	2/11/16 19:30	40.58919	-80.09885	40.58919	-80.09885	0.000
A-119	2	2/11/16 16:56	2/11/16 22:56	40.58919	-80.09885	40.58919	-80.09885	0.000
A-149	3	2/13/16 07:14	2/13/16 13:14	40.48422	-80.13755	40.50346	-80.13920	1.332

3. Data preparation:

The dataset had around 1.5 million rows which was very large for my R to process so I decided to only keep 10k rows and did this in excel instead of R. then when I tried to import the dataset in R it had two new columns named X and X1 which I then dropped in R.

As you can see the column Distance.mi. is poorly named so going forward with this name can cause mistakes in the code which will give rise to errors so changing the name to Distance for better readability. There are few other columns that have the same issue so doing the same for them as well.

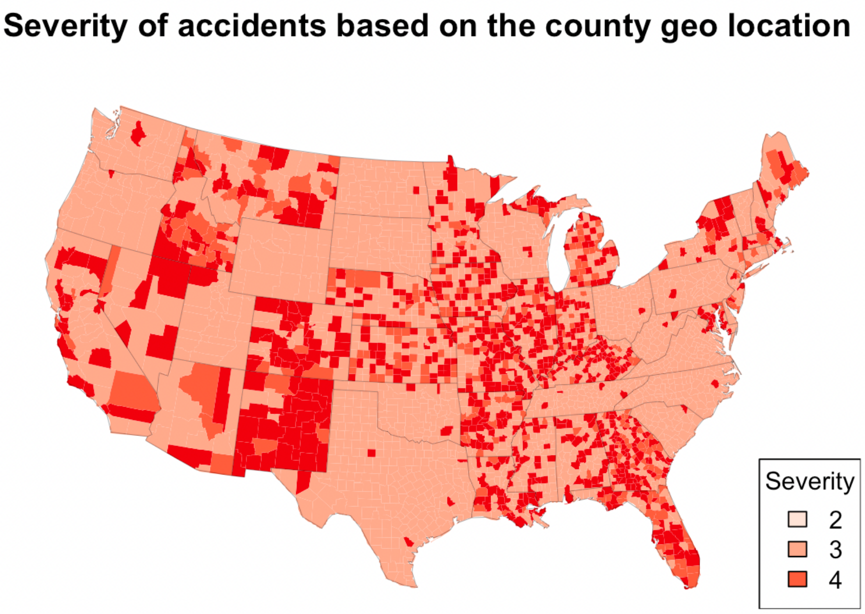


ID	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance
A-44	4	2/9/16 18:20	2/10/16 00:20	40.35429	-85.14993	40.45112	-85.15048	6.690
A-48	4	2/10/16 06:18	2/10/16 12:18	40.72813	-84.78965	40.74559	-84.78962	1.206
A-51	2	2/10/16 08:35	2/10/16 14:35	41.83193	-80.10143	41.84149	-80.11099	0.824
A-67	2	2/10/16 12:54	2/10/16 18:54	41.48339	-81.66297	41.47692	-81.66075	0.462
A-90	2	2/11/16 07:20	2/11/16 13:20	38.33667	-81.65623	38.33614	-81.65623	0.037
A-91	2	2/11/16 07:20	2/11/16 13:20	38.33614	-81.65623	38.33667	-81.65623	0.037
A-113	2	2/11/16 13:30	2/11/16 19:30	40.58919	-80.09885	40.58919	-80.09885	0.000
A-119	2	2/11/16 16:56	2/11/16 22:56	40.58919	-80.09885	40.58919	-80.09885	0.000
A-149	3	2/13/16 07:14	2/13/16 13:14	40.48422	-80.13755	40.50346	-80.13920	1.332

4. Modelling and evaluation:

For the modelling various graphs were used to plot and obtain insights to help answer desired questions. For me Severity of the accidents is a major concern and understanding as much as I can about it to better predict the future and incorporate changes with regards to it was of high priority.

- For geographic heat map: to find the severity of accidents based on the state wise geolocation



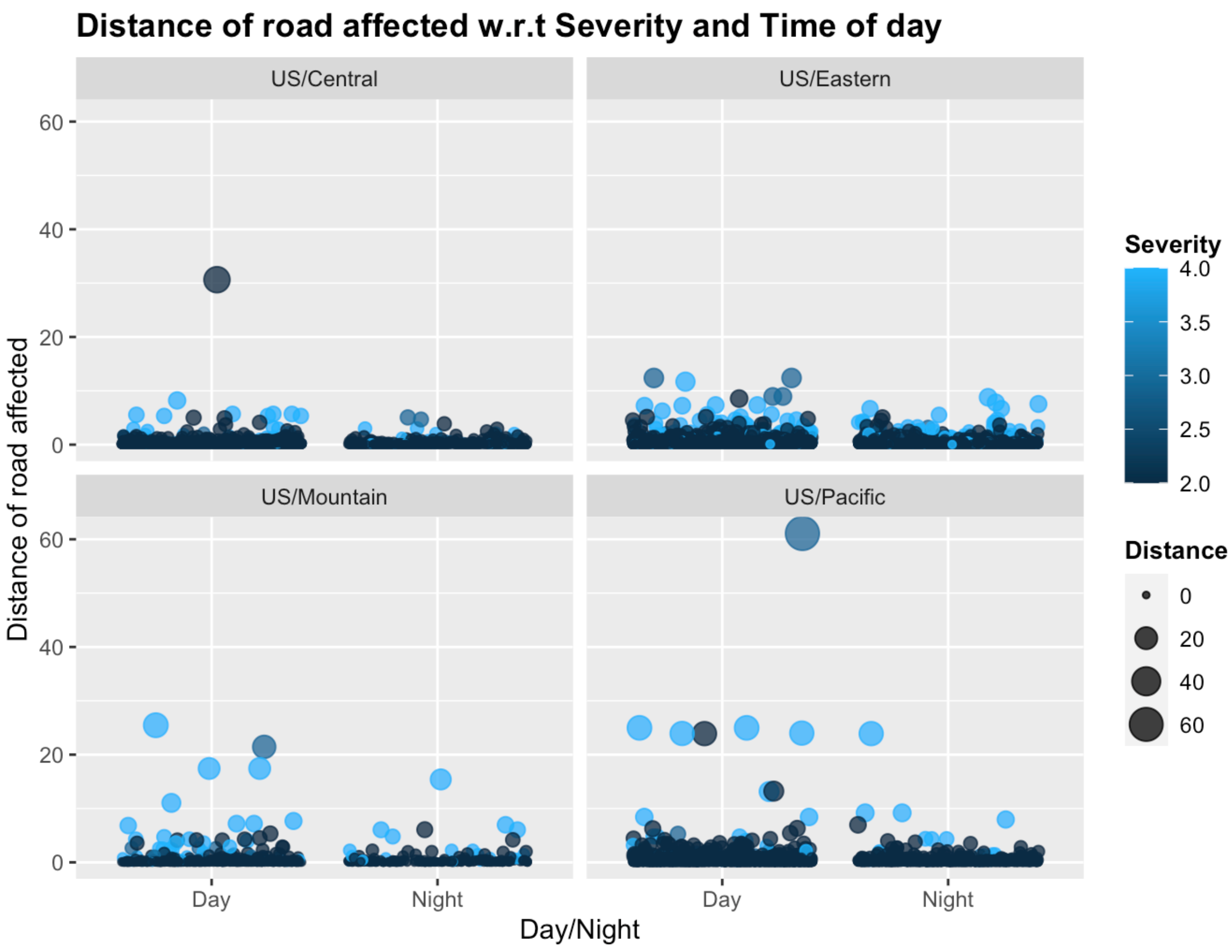
Severity of accidents based on the county geo location

(illustraion of heat map)

From the plot it can be seen that the severe accidents frequency is less than that compared to the less severe accidents. In New Mexico majority of the accidents that occur are severe. in Texas it's the opposite as there are only mild accidents that take place throughout the state. States like Florida, Kansas, Missouri has a mix of mild and severe accidents.

- Scatter plot: to find Distance of road affected w.r.t Severity and Time of day based on the time zone

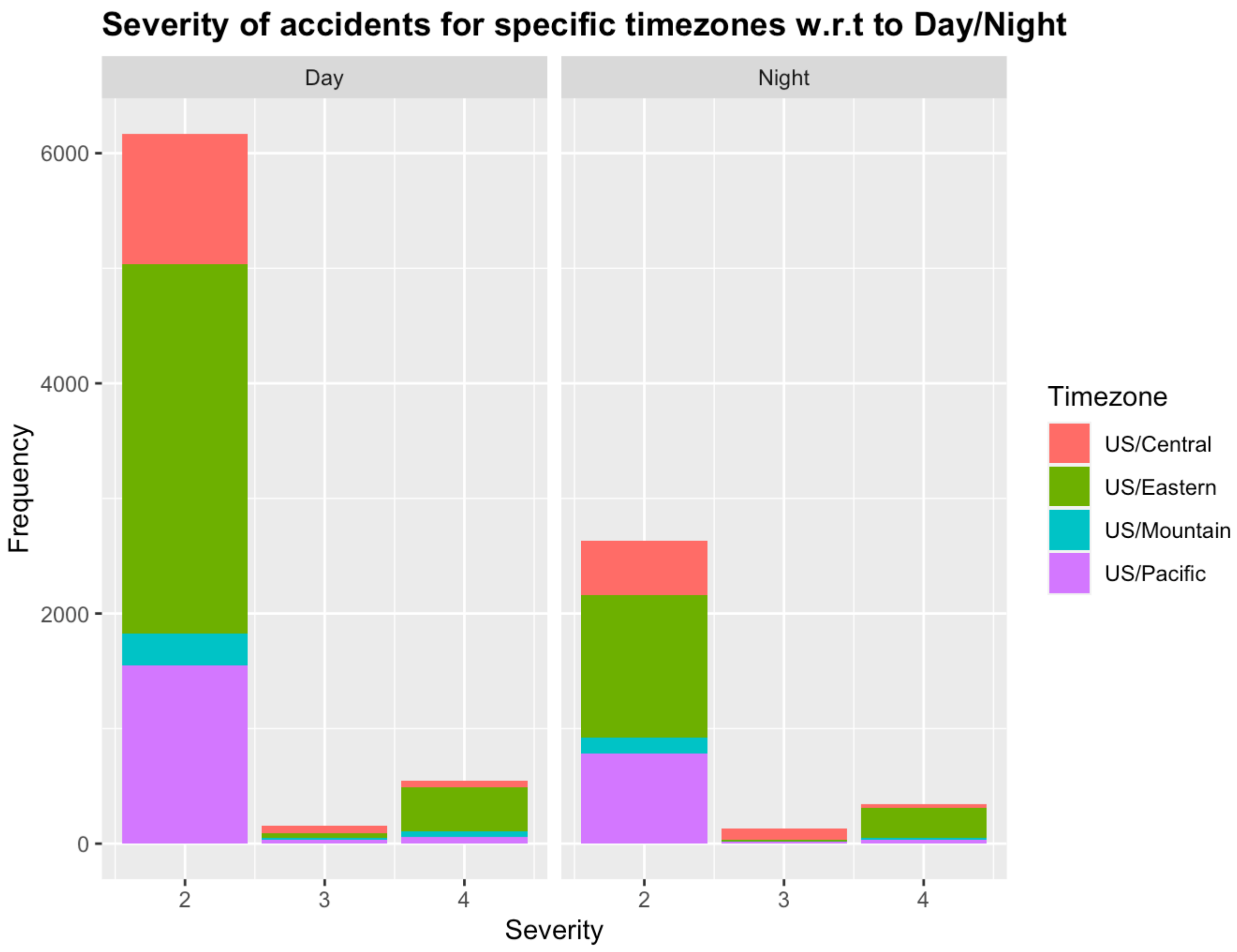
From the plot it is seen that the number of severe accidents take place during the daytime. It is even observed that much severe accidents tend to do more damage to the roads as one can expect.



- For the graph of our choice, I went for bar plot: to find out the frequency of accidents based on the time zone and Severity

A bar plot would be best suited to answer my question as it can provide lots of information in less space and easily readable to gain insights, so I went for it.

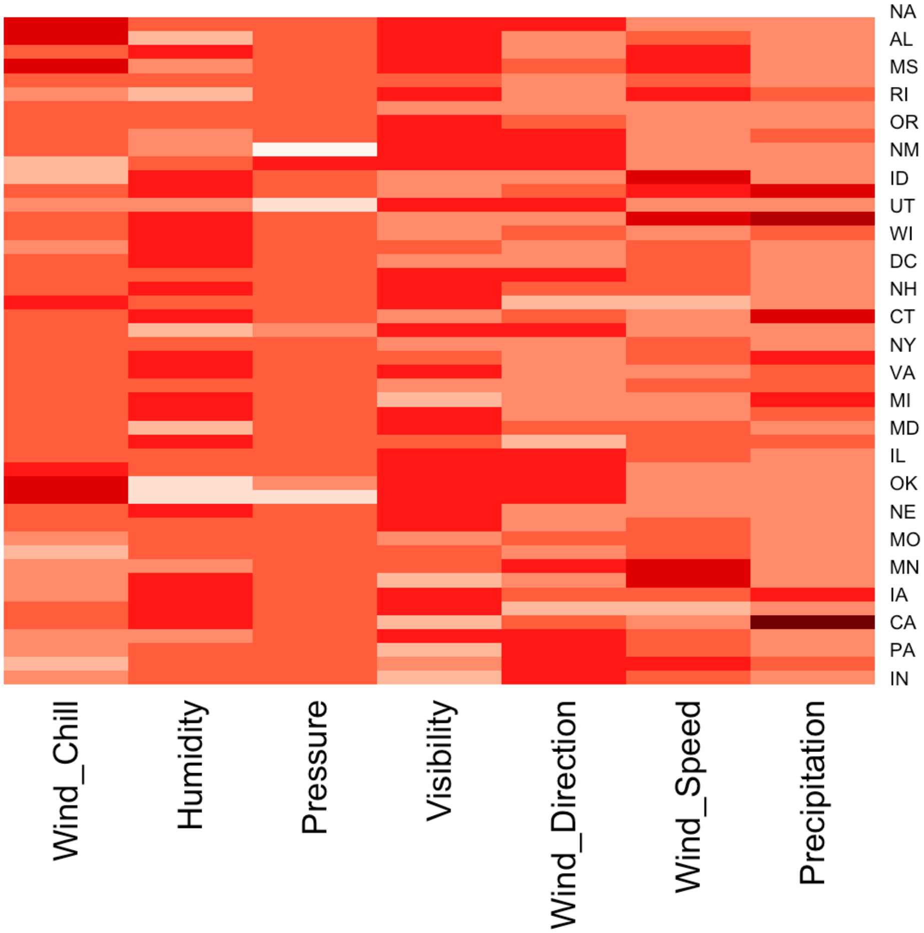
A breakdown of the frequency of accidents and the severity of it with respect to duration of day was found out. It can be seen that the number of accidents is more during the daytime than that compared to the nighttime. And the frequency of accidents in the US/Eastern time zone is much higher than the other time zones.



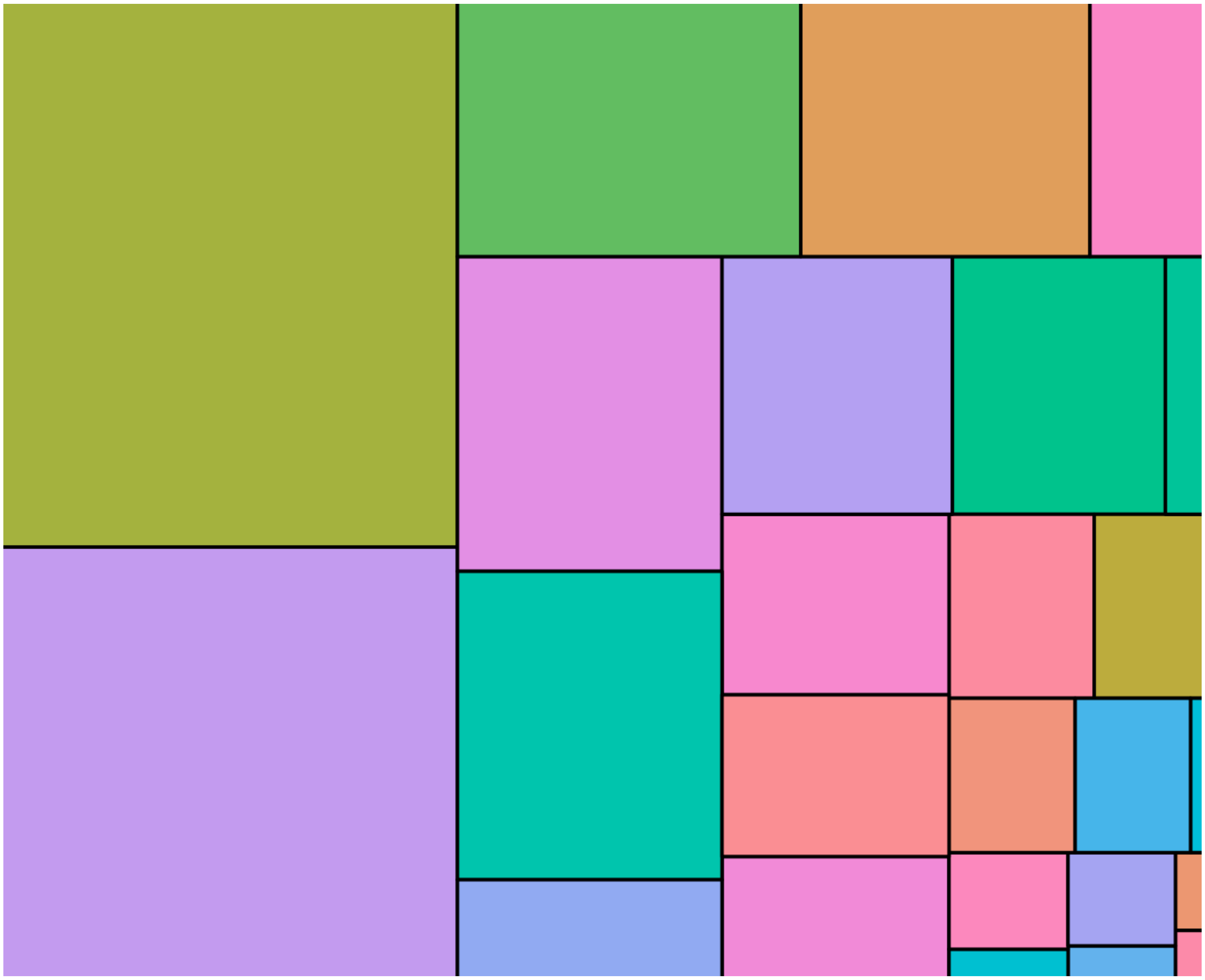
- Table heatmap: to find state wise distributions of various weather aspects

From the below heatmap we can draw some insights about the weather and its attributes according to different states. The shading provides us with a more better insight than other line or bar plots for this scenario.
Wind is chilly in the state of Alabama, Mississippi, Oklahoma, Nebraska, on the other hand the humidity is less for Oklahoma, Nebraska.
Visibility is low in the states of Indiana, California, this might lead to increase in accidents. Other various insights can be drawn and then compared with analysis to support the question using heatmaps.

Statewise Heatmap of various Weather attributes



- Tree map: to find the frequency of accidents that occur state wise:
  Due to some error I can’t read the labels and am not able to make insights.



5. Results:

In conclusion, from the analysis we did, and the graphs plotted, it is observed that the State of Florida has the highest number of accidents followed by the state of California. And the number of accidents that occur during the day is far more than that compared to night.
As the visibility was seen lower in the State of California it sees higher number of accidents as per our previous analysis which stands robust. And the states that had much better visibility had less accidents. Comparing these graphs, we can derive more insights and find out what other factors leads to accidents in the United States of America, which could help provide us with predictions and come up with precautionary measures to lower these rates.