Big Data - Project Deliverable 2

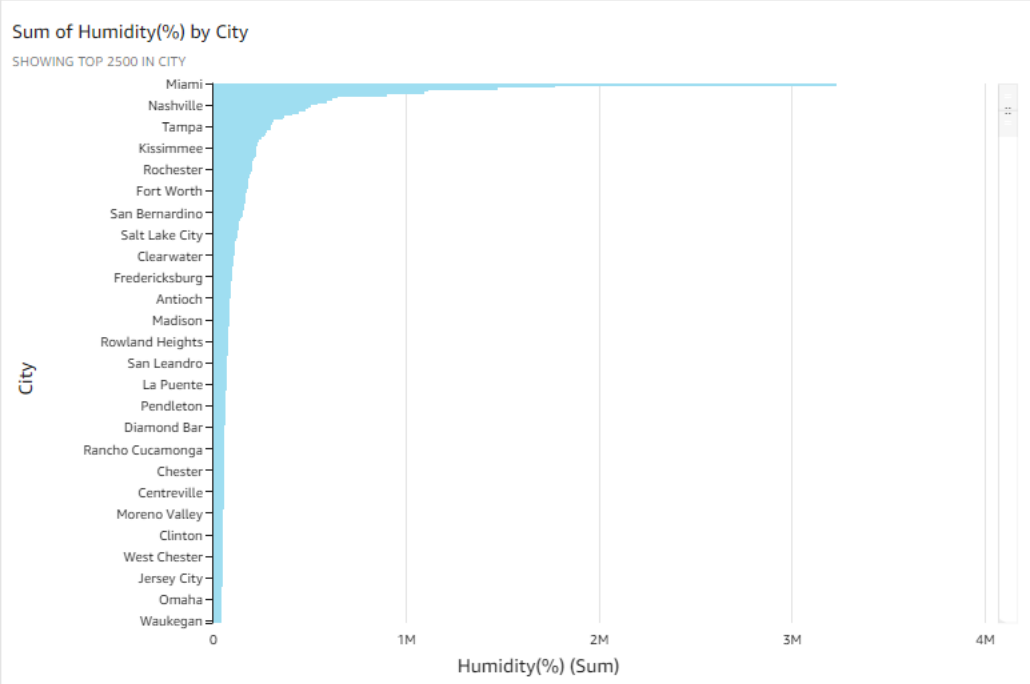
Group 21

5. Data Understanding:

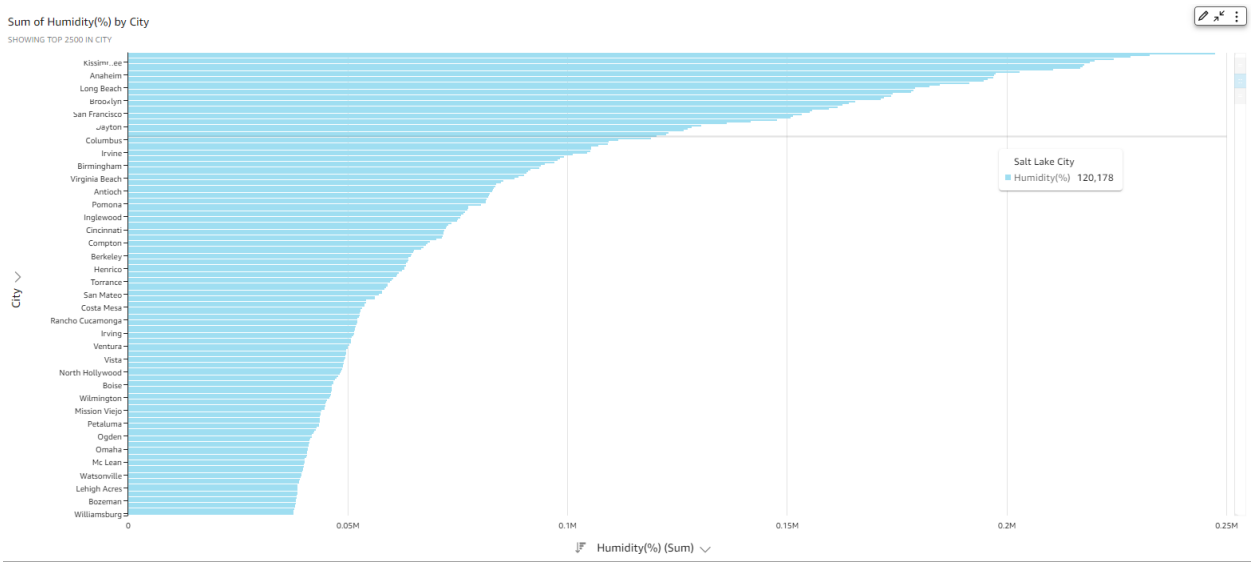
a. Exploratory Data Analysis: We have used the AWS Quick Sight for our Exploratory Data Analysis. We analyzed how much the weather conditions can cause lead to the road accidents. The weather parameters like humidity, visibility and wind speed are considered for the analysis. Also, we analyzed the effect of traffic signs in causing the road accidents. The following are the visualizations that we generated on US accidents dataset.

1. Humidity effects on road accidents in every city:

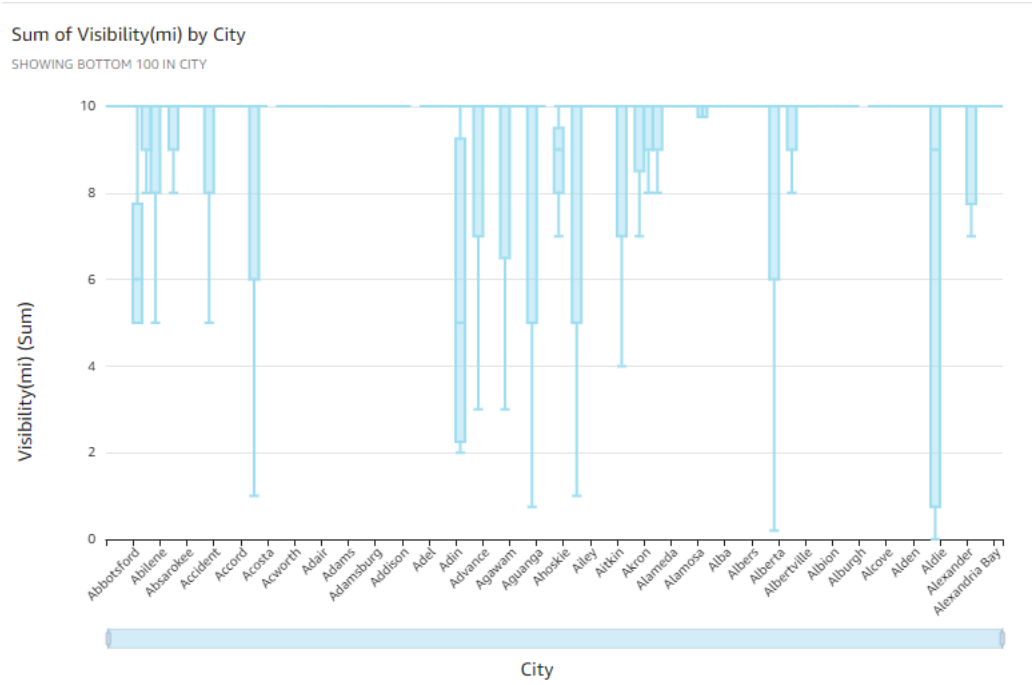
The below graph shows the number of accidents that was caused by the humidity in different cities. We can observe that Miami is recorded highest number of accidents that occurred because of highest humidity level.

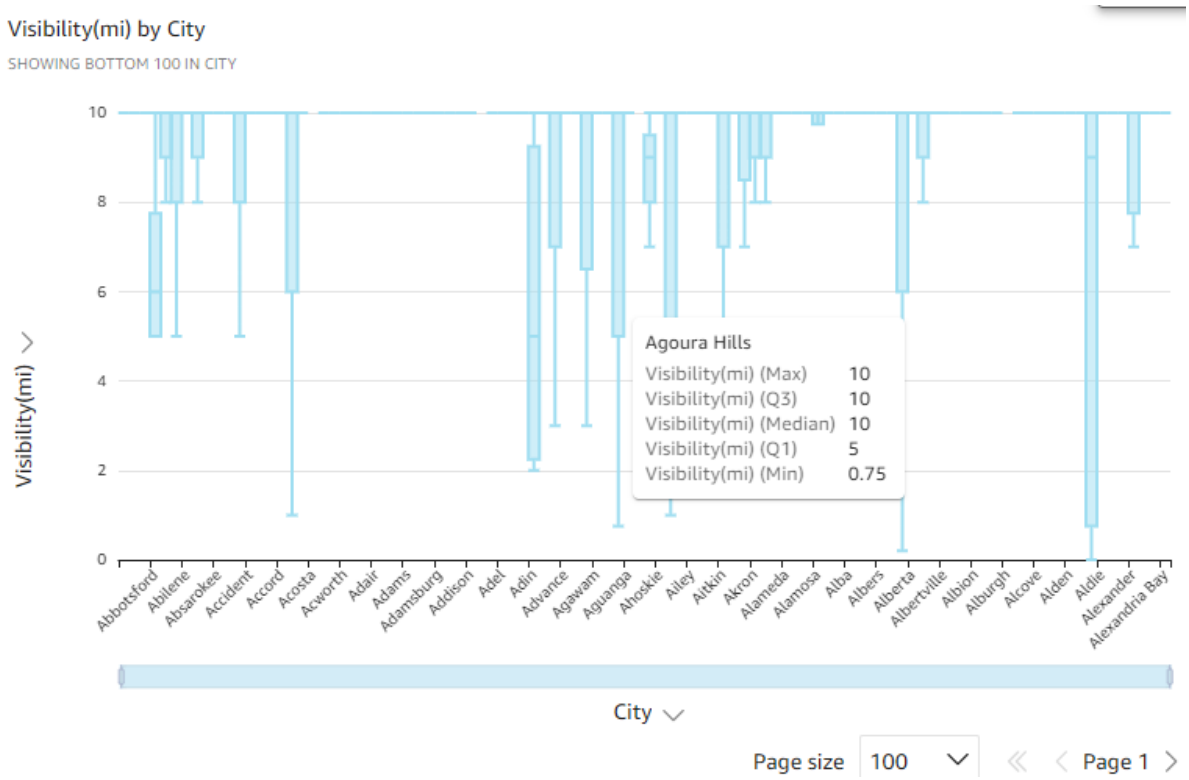


The below graph is another version of showing the humidity effect on causing road accidents in different cities.

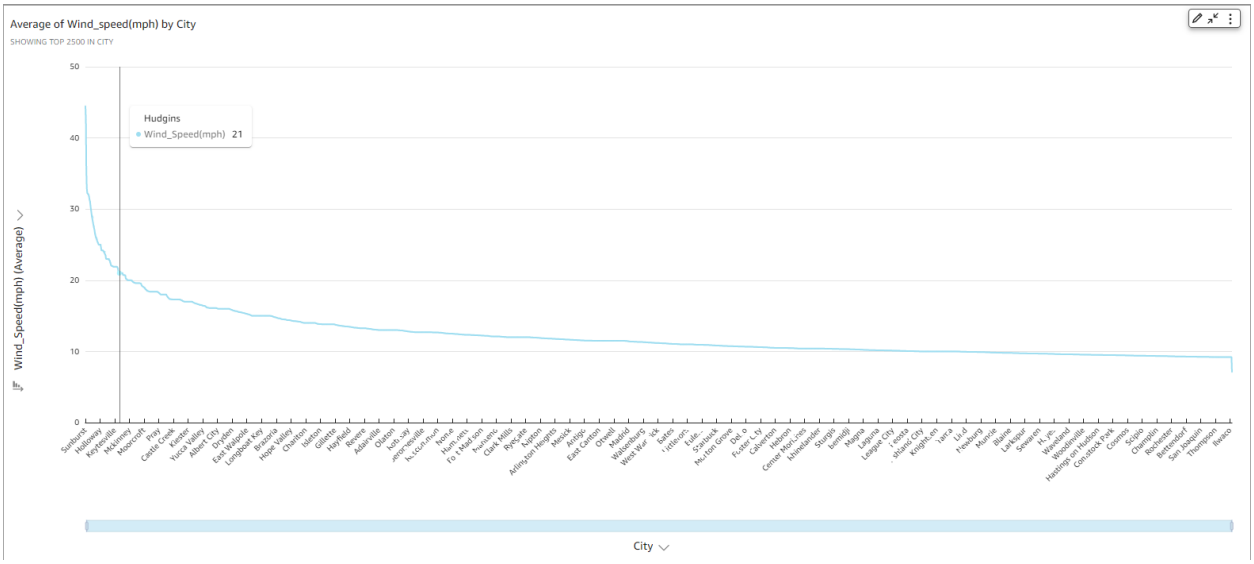


2. **Visibility effect on road accidents in every city:** The below box plot gives the information about the road accidents in every city that was caused because of low visibility. The box plot also identifies the median, max, min, Q1 and Q3 values for every city.

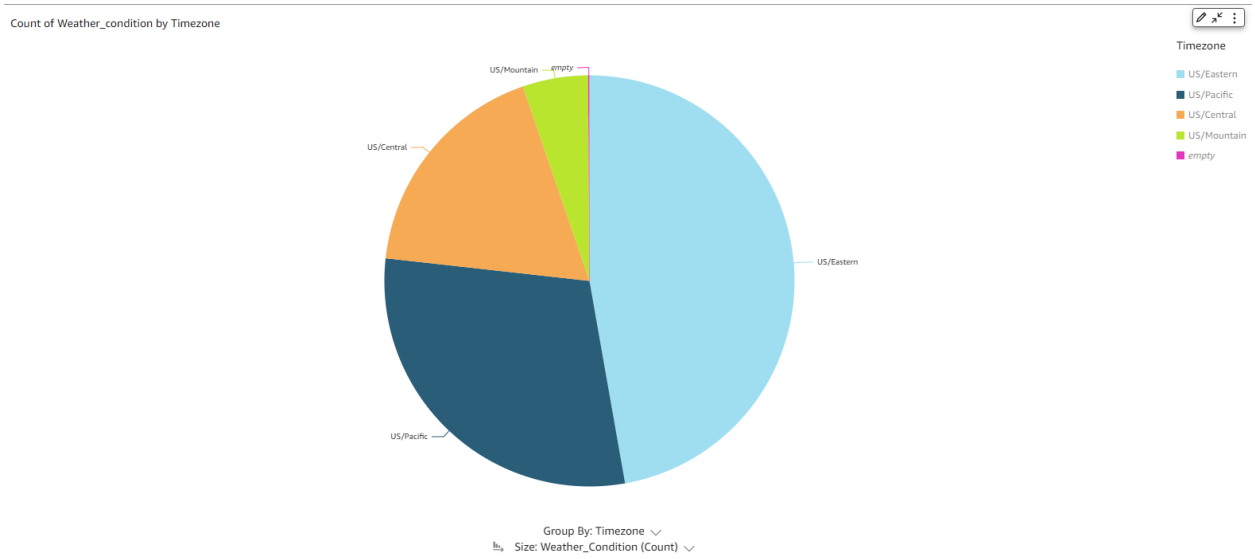




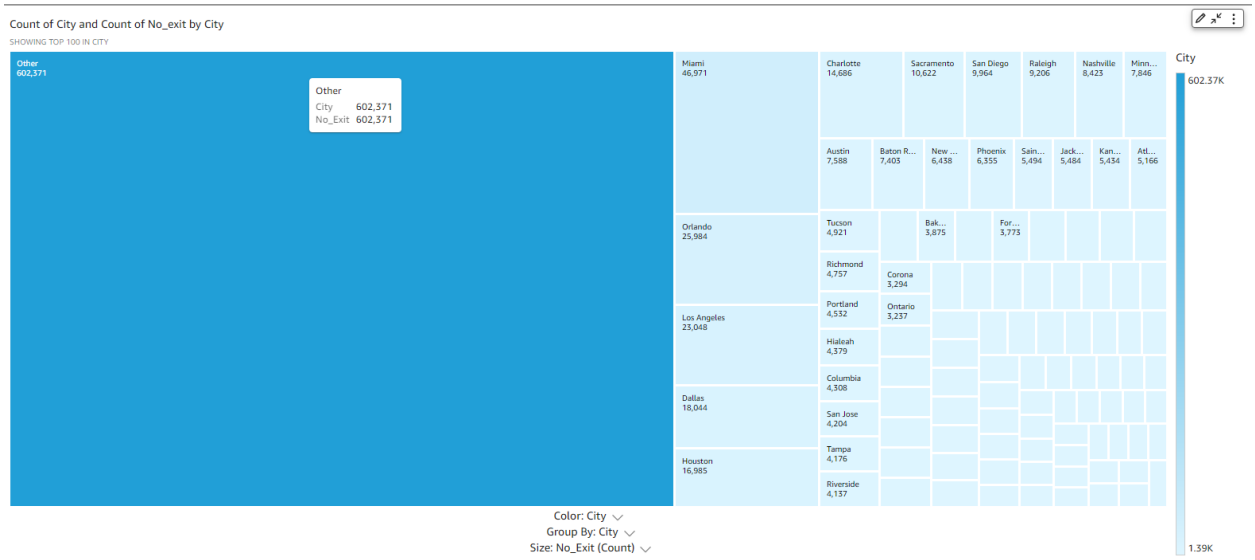
3. **Wind speed by city:** The below graph indicates the average number of accidents that was caused because of wind speed for every city. Also, the graph gives the average wind speed in every city.



4. **Weather condition by time zone**: The below graph indicates the effect of weather condition in every time zone. This shows that US/ Eastern time zone has recorded highest number of accidents that was caused by bad weather conditions.



5. **Accidents because of No exit sign in every city:** The below pivot table gives the information of number of accidents that was caused by the no exit traffic signal in each city.

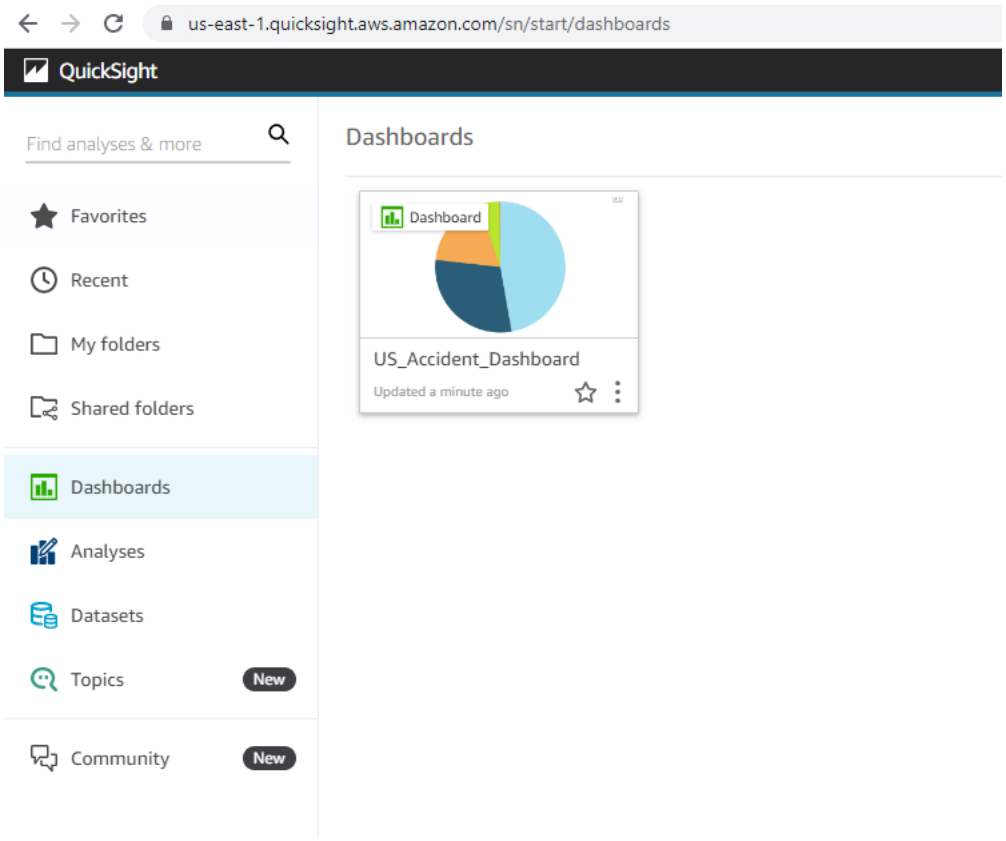


b. Dashboard

We created the dashboard in AWS Quick Sight. The below is the image showing the dashboard that we created.

GitHub:

https://github.com/developer-rohith/bigdata_project/blob/main/BDA_project.ipynb



6. Data Preparation:

We used AWS Sage Maker for or Data Preparation part. We have created a Sage Maker notebook instance for our project work and worked in Data preparation and EDA part in that instance.

GitHub:

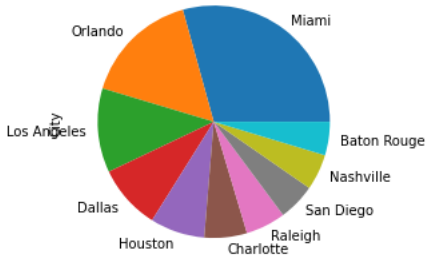
https://github.com/developerrohith/bigdata_project/blob/main/BDA_project.ipynb

Step 1: Read the data into data frame.

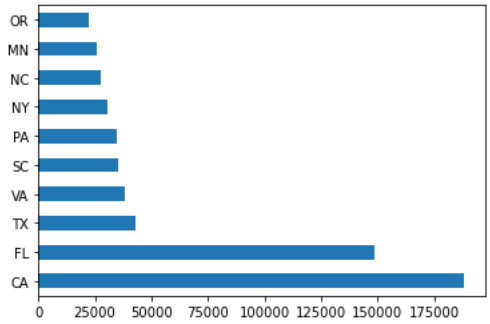
Step 2: Remove the unwanted columns, checking for null values and deleting them.

Step 3: Visualizations

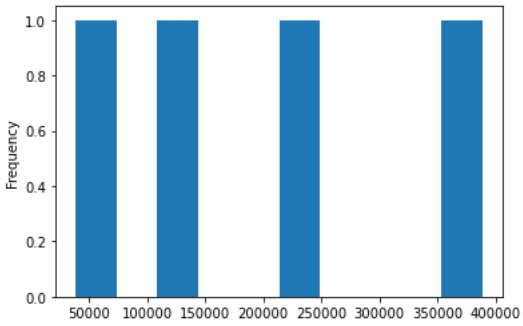
Number of accidents on each city:



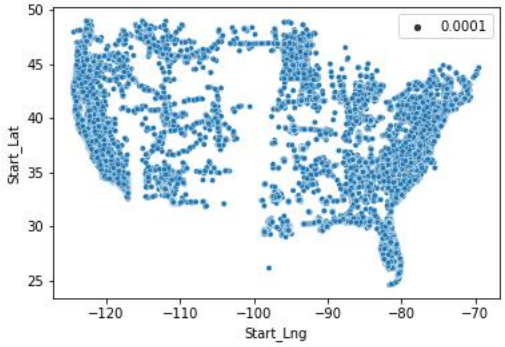
Number of accidents on each state:



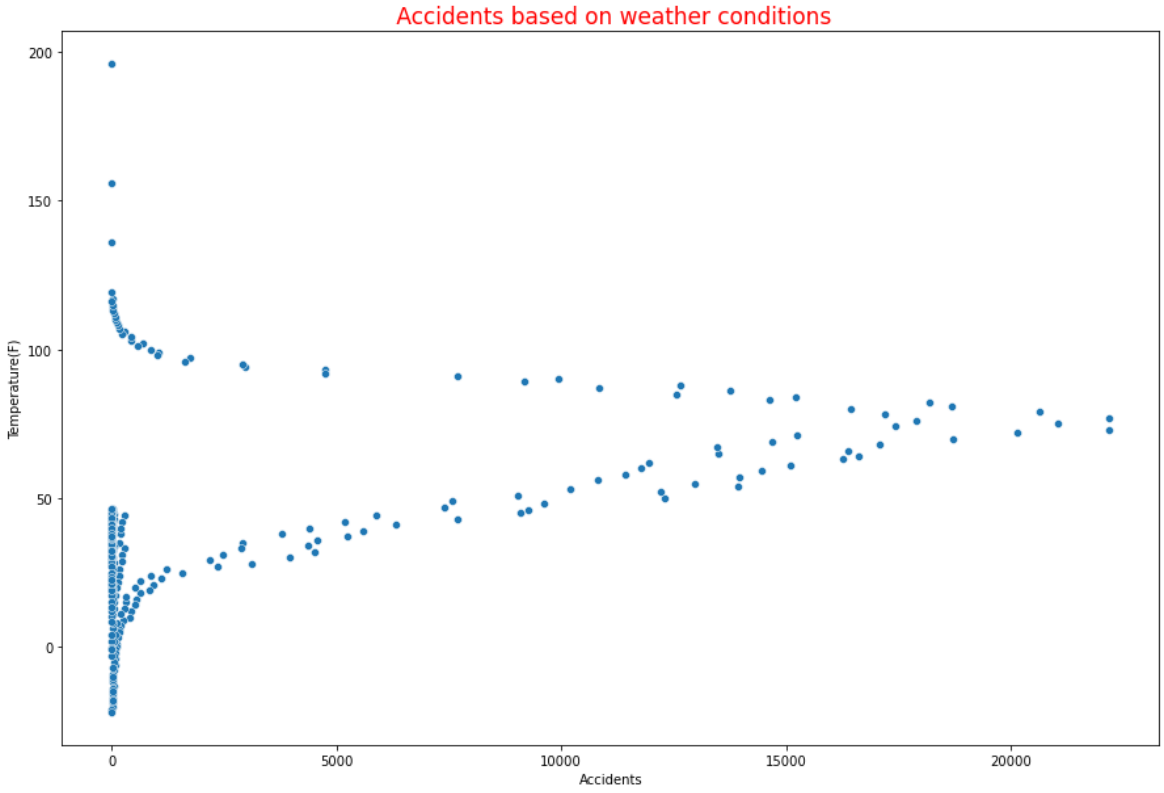
Number of accidents on each time zone:



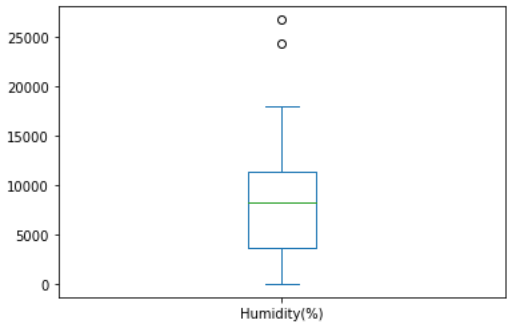
Number of accidents on start latitude and longitude:



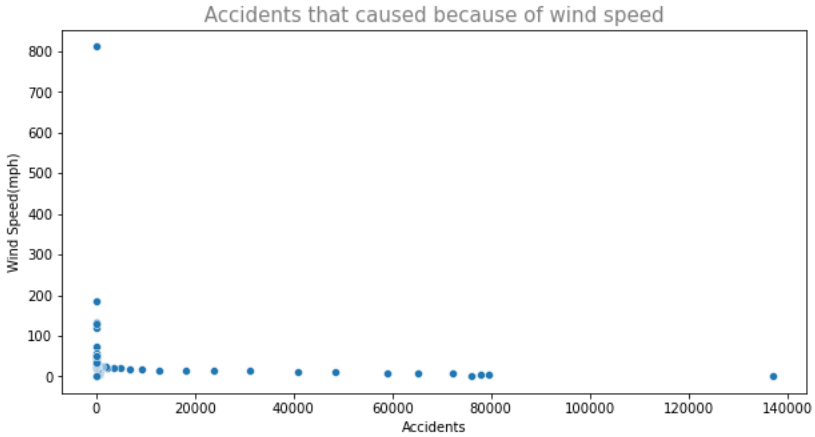
Accidents based on Weather Conditions:



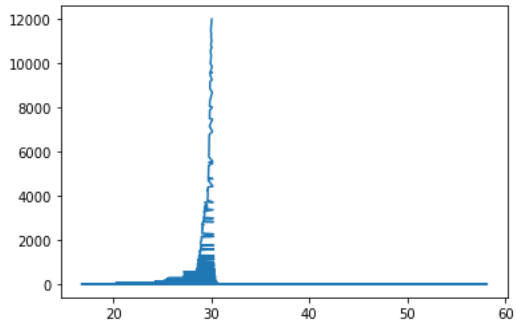
Accidents caused by Humidity:



Accidents caused by Wind Speed:



Accidents caused by Pressure:



Step 4: Analysis/Observations

1. The top 5 cities that has recorded highest number of accidents are Miami, Orlando, Los Angeles, Dallas, and Houston.
2. The top 5 states that has highest accidents recorded are CA, FL, TX, VA, and SC.
3. The US/Eastern time zone has recoded the highest number of accidents i,e 388572 accidents and least is US/Mountain region.
4. Most of the accidents are happening when the temperature is higher than 77F.
5. When the humidity of higher than 65 has greater chance of collision happening.
6. Maximum number of accidents occurred when the wind speed range is between 5 - 10 miles per hour.