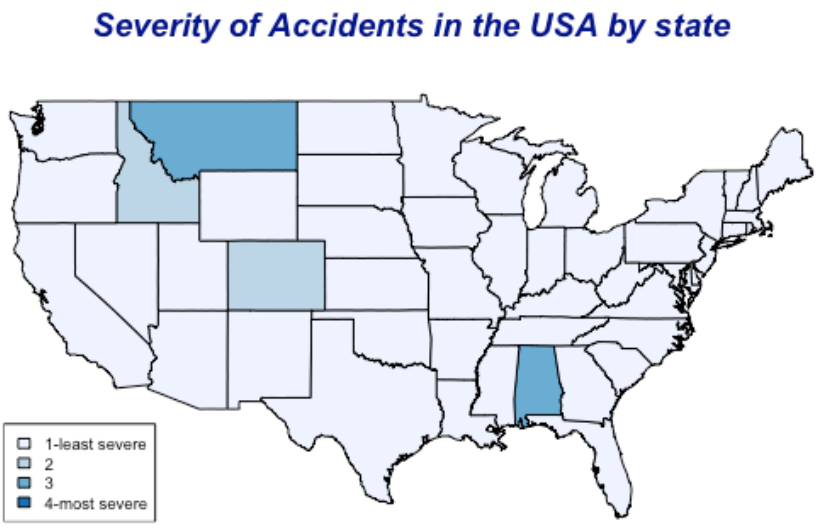
Sushmitha Mohana

CRISP-DM phases for questions 1-5
Question 1 : What is the severity of accidents based on states
Prerequisite steps:

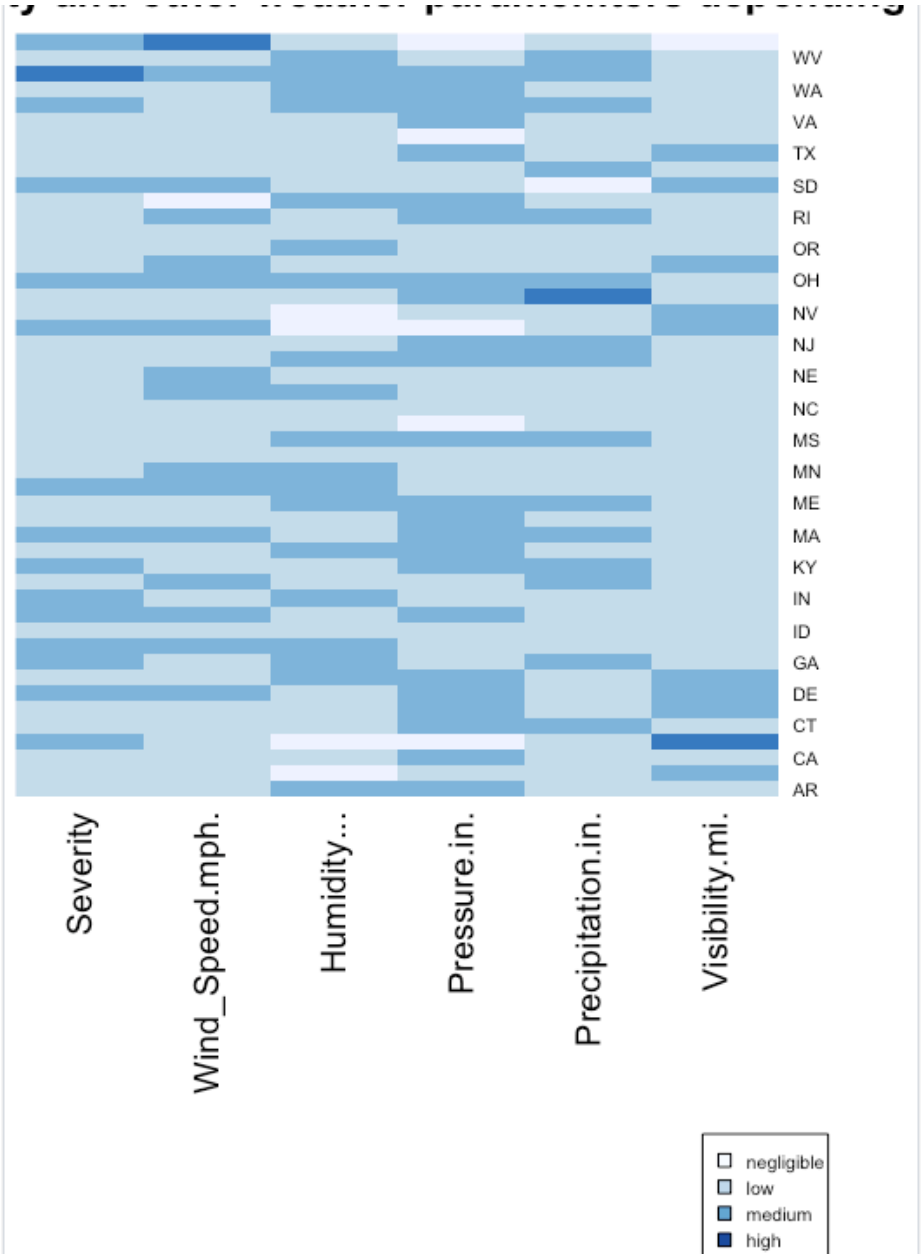
1. **Install all required packages and invoked them**
2. **Removed unnecessary columns with more than 61% missing values**
3. **Replaced missing data in certain columns with the mean of the column**

Phase 1:
-The dataset has columns that contain 61% missing data.
-The 'State' column willl need to be renamed to 'abb' to be able to join the df with data in package 'usa'
Phase 2: Data Preparation:
-Imported the data package usa
-Changed the name of column state to 'abb' to facilitate joining
- Dropped the 'Names' column
Phase 3: Modeling: Created a geographical plot of the states, filling color according to the severity of accidents
Phase 4: Evaluation: Drew conclusions from the visualization so created.States Montana and Alabama have the highest severity of accidents, followed by Idaho and Colorado. All the other states have the least severity



Question 2: What are the top 10 cities in the USA with the most number of accidents
Phase 1:
- The answer to the previous question made me wonder what the top 10 states with the most accidents are
-Importing Treemap library
Phase 2: Data Preparation:
- No additional preparation was needed
Phase 3: Modeling: Created a tree map depicting top 10 states with highest number of accidents
Phase 4: Evaluation: I was surprised to see that the states with most severe accidents did not feature in the map. This helped me frame my next question:
How does severity change in each sate according to their unique weather and visibility parameters?

Question 3: So Does severity(aka time delay) change with other weather parameters? I am doing this by reviewing trends in average severity and other average weather parameters according to state.
Phase 1:
- Decided to keep only parameters required for the analysis
Phase 2: Data Preparation:
- Discarded all column not needed for the analysis
- Segregated columns (visibility, humidity etc) based on states in the dataframe
- Created the matrix from the data frame
  Phase 3: Modeling: Created a heatmap to answer the question
  Phase 4: Evaluation:
- High average severity accidents are accompanied by medium precipitation, medium wind speed, medium humidity, and low visibility. These weather conditions promote the highest delay/wait time in traffic.
- Note: Unable to show the title of the plot her. Please check the R file for full diagram



Question 4: Plot of my Choice: How does distance affected by traffic vary with severity? What does the trend look like? How does Day/Night impact this?

Phase 1:

- need changed label names for day and night

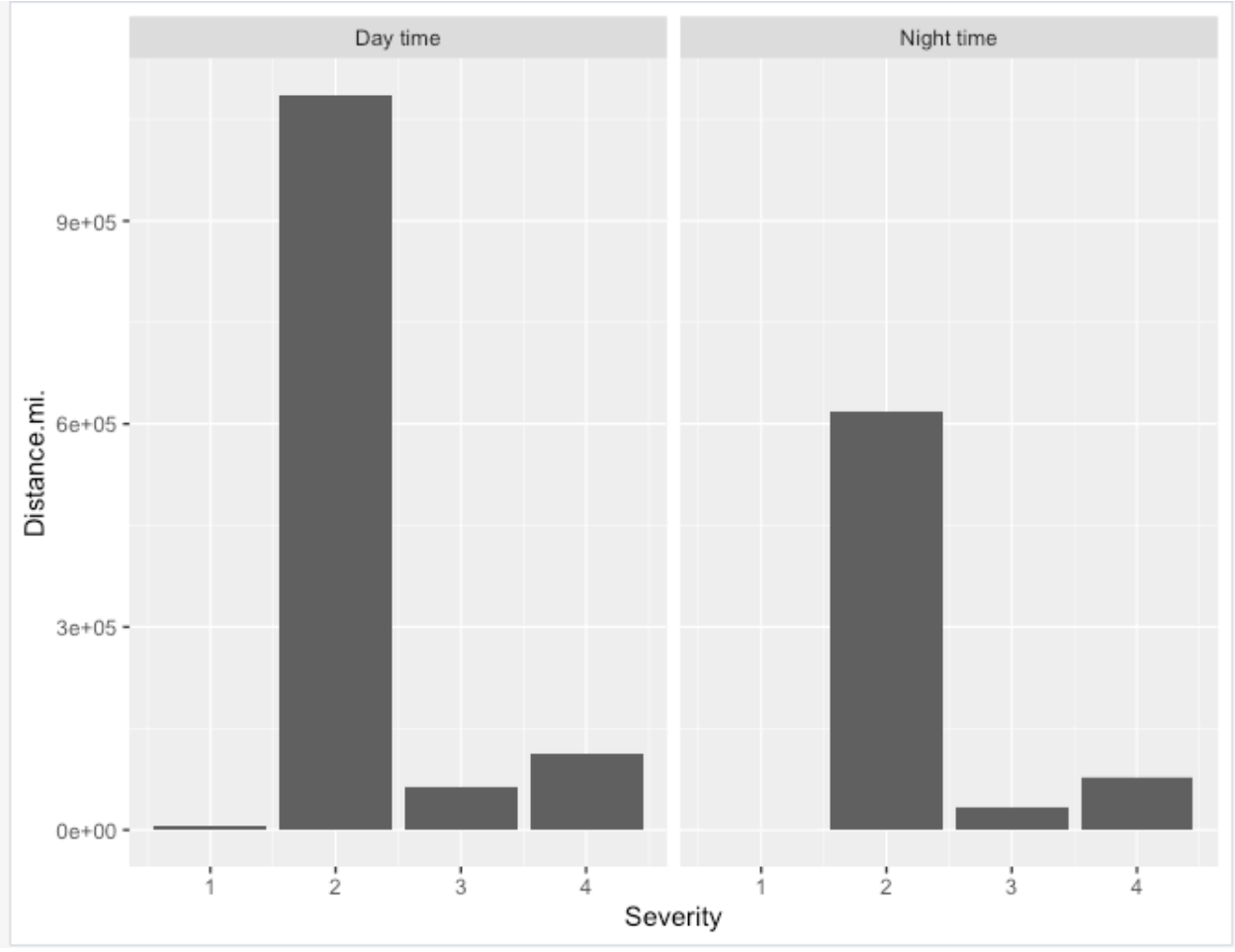
Phase 2: Data Preparation:

- Removed unknown values for night and day
- changed label names for night and day

Phase 3: Modeling: Created a bar chart to answer the question

Phase 4: Evaluation:

- The distance shows in exponential form because the miles are displayed in decimals. Since we want to observe the trend, it being in decimal form does not matter.
- The accidents with severity 2 affect the largest distance of the road, especially when accidents happen in the daytime.



Question 5:How is distance affected by accident vary with visibility in miles? How does the trend look? How does severity and Day/Night fit into this picture?

Note: Unable to change the labels of x and y axis in the plot despite putting it in the code No errors while executing but no change in the plot

Conclusion: It is observed that when the visibility is low the distance affected is higher as compared to higher visibility, irrespective of day or night and severity. Accidents with severity 2 and 3 have the highest affected distance when there is low visibility, especially during the day. When the Visibility is high, the distance affected is low, irrespective of severity.