**Exploratory Data Analysis: US - Accidents**

Dataset:

This dataset is a comprehensive collection of US countrywide accident data that covers 49 states. The original dataset ranged from February 2016 to March 2023, but due to its size, only data up to December 2020 is included in this EDA project. The dataset comprises over 40,000 rows and 46 columns, including 16 string, 13 Boolean, 10 decimal, and 7 other data types. According to Sobhan Moosavi on Kaggle, this dataset has many potential applications, such as real-time accident prediction, hotspot location analysis, casualty analysis, and studying the causal factors of accidents. Additionally, it can be used to investigate the impact of precipitation or other environmental factors on accident occurrence.

Methods used:

I started my EDA project by downloading the data. Using pandas, I loaded the data and looked at the initial data frame to get an idea of the data I was using. I used preliminary information and summary statistics to understand the data better. To do this, I used the following methods: df.info(), df.summary(), df.describe, and df.isna(). With the help of df.info() I understood the type of variables (columns) I was working with. Using df.describe(), I understood the number of non-null values, the minimums, the maximum mean values and percentiles. The final step I took to understand my data was to look for missing values. Firstly I used the df.isna() method, which converted everything into boolean and returned false for columns that did not have missing values. However, that was not very useful because I had many columns, and it was hard for me to understand which columns had the most missing values, so I sorted the values in ascending order to find out the number of missing values in each column. Then, I found the percentage share of missing values for each column in the entire dataset and sorted the percentages again to ensure I had a clear understanding of the columns with the highest percentage of missing values. I found that the columns ‘Number’, ‘Precipitation(in)’, and ‘Wind\_Chill(F)’ had the most missing values (40–60% of the data). While it would be wise to remove these columns, I believe that precipitation might play a key role in accidents, so I haven't removed it yet because I may need it in the next part of my project. However, it would be wise to either remove the columns altogether or exclude them from the ensuing analysis since the percentage of missing values is so high.

Discussion of Dataset:

I have analyzed the accident data for each state and plotted a bar chart to see the number of accidents. The analysis revealed that California has the highest number of accidents, followed by Florida. Moving on to the cities, I plotted another bar chart to visualize the top 30 cities. I was curious to see how the cities were distributed by the number of accidents, so I also created a histogram. The histogram showed that over 1200 cities had only one accident during the entire period. This could either be good news or a result of missing data. To identify the cities with high and low accident rates, I plotted separate charts for both.

For the start time, I converted the data to timestamp and plotted the distribution of accidents by hour. The plot showed two peak times, one in the morning between 6am to 9am and another between 3 pm and 6 pm. I also created separate plots to see if this trend was consistent on weekdays and weekends. Additionally, I plotted a chart for the months to see if certain months had more accidents than others. The analysis showed that the second half of the year had more accidents, which could be due to summer break and the new school term around August-September.

There is still more analysis to be done on the dataset to explore accidents in the US further. I will continue to explore and analyze the dataset to ensure that there are no misleading columns due to significant missing data.

Insights and conclusion:

I have found some interesting facts related to accidents in the US. Among all the states, the top 5 with the highest number of accidents are Los Angeles, Houston, and Florida, which are densely populated states. Between February 2016 and December 2020, less than 5% (~ 4.35%) of cities had more than 1000 accidents, which suggests that most cities had fewer accidents than that. So, some cities, especially the populous ones, need to take more strict measures than others. New York, which is also a populous city, had less than 7500 accidents during this period.

Moreover, out of over 1200 cities, some had only one accident in four years, which seems unusual and requires further investigation. Most cities had between 10 to 100 accidents during this period. Additionally, the distribution of accidents by cities follows an exponentially decreasing pattern.

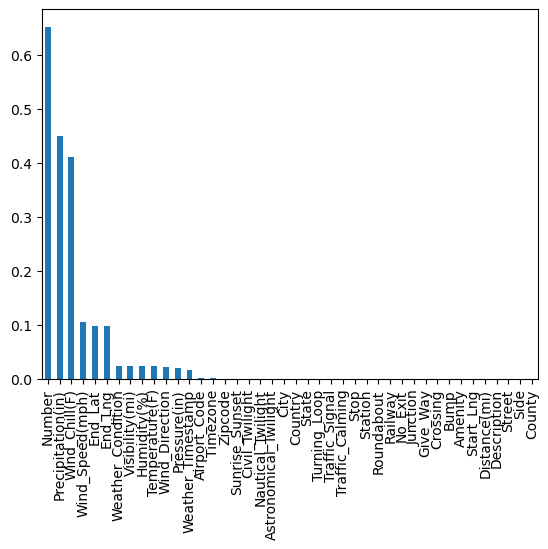
Lastly, the data supports the hypothesis that weekdays see more accidents during morning and evening rush hours, while weekends have a peak during the afternoon.

Future Analysis Possible:

I haven't decided yet what I'd like to predict for the next step of the project. However, I could analyze year-on-year trends of accidents. Perhaps I could attempt to predict accidents in real-time or investigate how weather conditions, such as precipitation, visibility, or other environmental factors, impact accident occurrence.

Below you can find all the plots:

PLOT – Percentage of Missing Values



PLOT – Accidents Per State

A graph of a number of states

Description automatically generated

PLOT #3 – Accidents per city

A graph of the number of cities

Description automatically generated

PLOT #4 – low Accident Cities

A graph of a number of blue bars

Description automatically generated

PLOT #5 – Start Time

A graph of a graph

Description automatically generated

PLOT #7 – Start Time on Monday

A graph showing a number of columns

Description automatically generated

PLOT #8 – Start Time on Sunday

A graph with blue bars

Description automatically generated

PLOT #9 – Start Time Month

A graph of a graph

Description automatically generated with medium confidence

Link to the dataset: https://www.kaggle.com/datasets/pateluttam/usaccidents/data