Computation and Visualization for Analysis

Country Wide Car Accidents

Written By:
Venkata Sai Swathi Gattamaneni

Introduction and Research Questions

The road is a long and winding one, never staying the same, ever-changing and unpredictable. Every turn has potential for something unexpected and unknown, so it's important to be prepared for the journey ahead. We hypothesize that certain factors such as weather conditions and road conditions have a significant impact on the likelihood of car accidents.

A dataset containing details on auto accidents that have happened in various parts of a single country is referred to as "country wide vehicle accident data." The dataset often contains a variety of information about each event, including the date and time, location, cars involved, severity of the accident, number of injuries and fatalities, weather and road conditions at the time of the accident, as well as other information.

To enable policymakers, traffic engineers, and other stakeholders to make knowledgeable decisions about road safety, it is important to gather and analyze this data in order to find patterns and trends in auto accidents. By the analysis of this data, we may learn more about the factors that lead to auto accidents, spot regions that are particularly dangerous, and create efficient plans for lowering collision rates and enhancing traffic safety. For researchers, decision-makers, and other stakeholders who are interested in enhancing road safety, this dataset is an invaluable resource. We can learn more about the causes of auto accidents and create practical preventative measures by examining this data. The Nationwide Vehicle Accident Statistics is a vital resource for promoting traffic safety and building better roads for all users of the road.

The Specific questions we aim to answer through the analysis include:

- What are the typical reasons for car accidents in various parts of the nation?
- Which demographic is most frequently engaged in automobile accidents?
- Are there any hours or days of the week when there are more automobile accidents?
- How do the weather and the state of the roads impact the number and severity of auto accidents?
- Which car kinds are most frequently engaged in catastrophic accidents?
- Over time, have there been any trends or patterns in car accidents?
- What effect do traffic control and road design have on auto accidents?
- What percentage of car accidents are prevented by the present road safety policies and programs?
- What are the best methods for enhancing traffic safety and lowering car accidents?
- Can the frequency and severity of auto accidents be reduced by the employment of cutting-edge technologies, such as driverless vehicles?

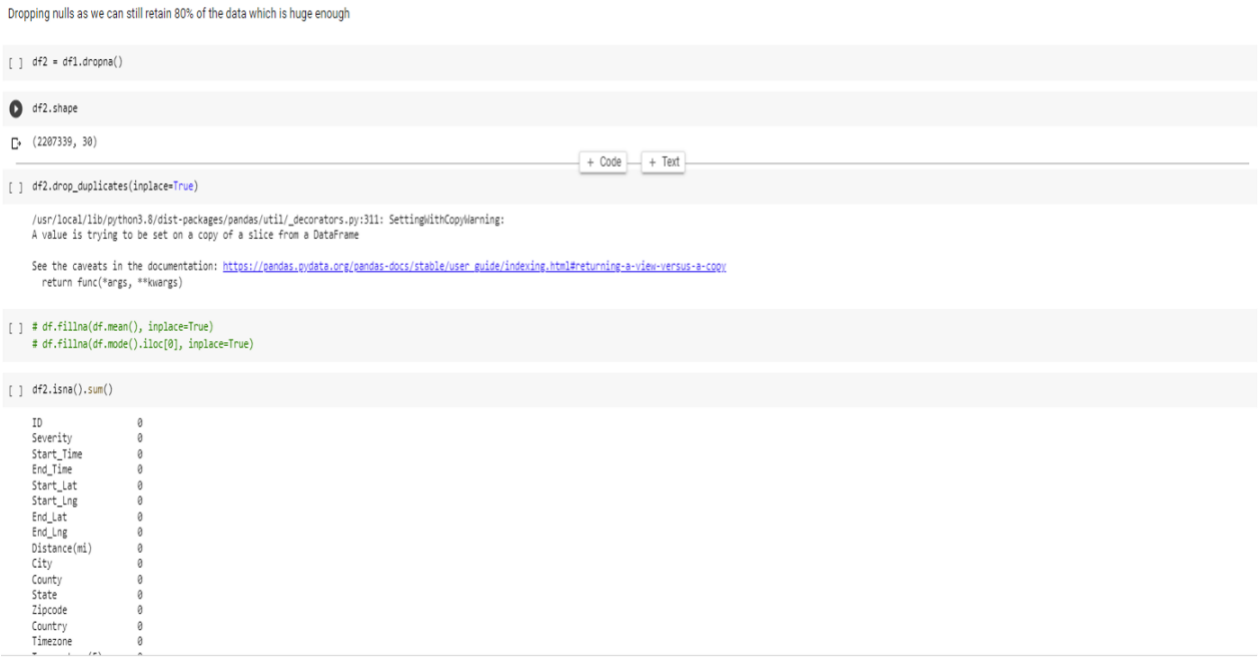
Data Sources

Most of the information used in this research came from Kaggle.
This experiment utilizes a nationwide car accident dataset that includes data from all 49 US states.

The accident data, which covered the period from February 2016 to December 2021, was gathered through a variety of APIs that provide streaming traffic incident (or event) data.
The entities whose traffic data is broadcast over these APIs include the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors installed in the road networks, to name just a few.

The collection contains information on each accident's location, date, and circumstances as well as details on the people and vehicles that were involved.
Currently, there are about 2.8 million accident entries in this collection.

The country wide car accidents dataset preparation involved:
Data cleaning: Deal with any outliers, duplicate data, or missing data.



Dropping nulls as we can still retain 80% of the data which is huge enough

```
[ ] df2 = df1.dropna()
```

```
df2.shape
```

```
(2207339, 30)
```

+ Code + Text

```
[ ] df2.drop_duplicates(inplace=True)
```

```
/usr/local/lib/python3.8/dist-packages/pandas/util/_decorators.py:311: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  return func(*args, **kwargs)
```

```
[ ] # df.fillna(df.mean(), inplace=True)
    # df.fillna(df.mode().iloc[0], inplace=True)
```

```
[ ] df2.isna().sum()
```

```
ID                0
Severity          0
Start_Time        0
End_Time          0
Start_Lat         0
Start_Lng         0
End_Lat           0
End_Lng           0
Distance(mi)      0
City              0
County            0
State             0
Zipcode           0
Country           0
Timezone          0
```

Data transformation: Normalizing and transforming the data to enable effective analysis.

```
# Add column for day of week
df2['Day_of_Week'] = df2['Start_Time'].dt.day_name()

# Add column for hour of day
df2['Hour_of_Day'] = df2['Start_Time'].dt.hour

# Add column for month of the accident incident
df2['Month'] = df2['Start_Time'].dt.month

# Add column for year of the accident inciden
df2['Year'] = df2['Start_Time'].dt.year

# Add column for duration of accident (in minutes)
df2['Duration'] = (df2['End_Time'] - df2['Start_Time']).dt.total_seconds() / 60
```

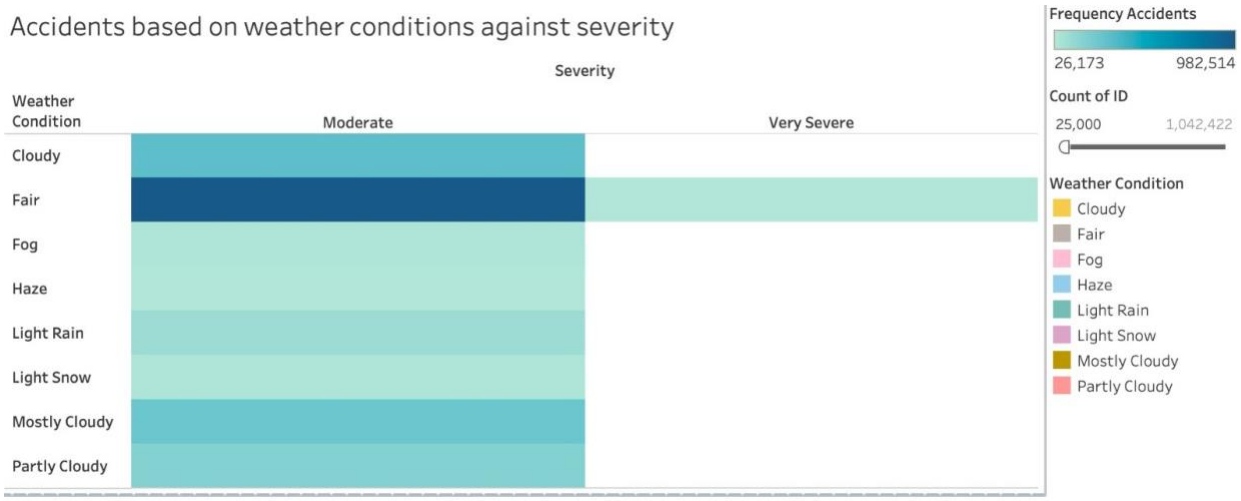
```
# Categorize Severity column
df3['Severity'] = df3['Severity'].replace({1: 'Minor', 2: 'Moderate', 3: 'Severe', 4: 'Very Severe'})
```

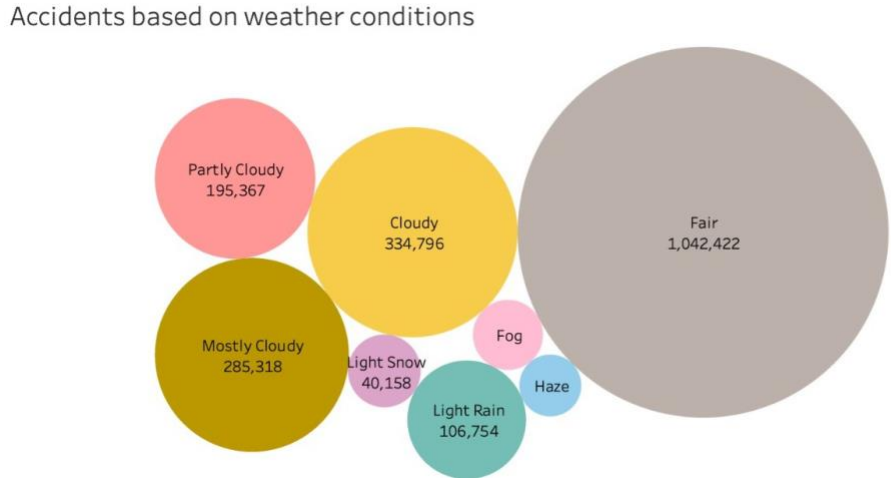
```
<ipython-input-21-fe45aa46f501>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df3['Severity'] = df3['Severity'].replace({1: 'Minor', 2: 'Moderate', 3: 'Severe', 4: 'Very Severe'})
```

Data visualization: To better comprehend the distribution and trends in the data, use data visualization.

developing an analytics dashboard with visuals.





Results and Methods:

The data we have obtained from Kaggle contained large amounts of null values and redundant columns which needed removal to obtain the data for visualizations and dashboards. For the initial EDA, we chose to use Python to remove null values, drop unnecessary columns like date recorded and unit ID, and finally export the data frame for easy import into Tableau. In fact, the entire file contains our data cleaning and feature extraction.

With the simple idea of being easily viewable and understandable to the end user/viewer we decided to have four dashboards that would serve as a tool in a Country Wide Car Accidents.

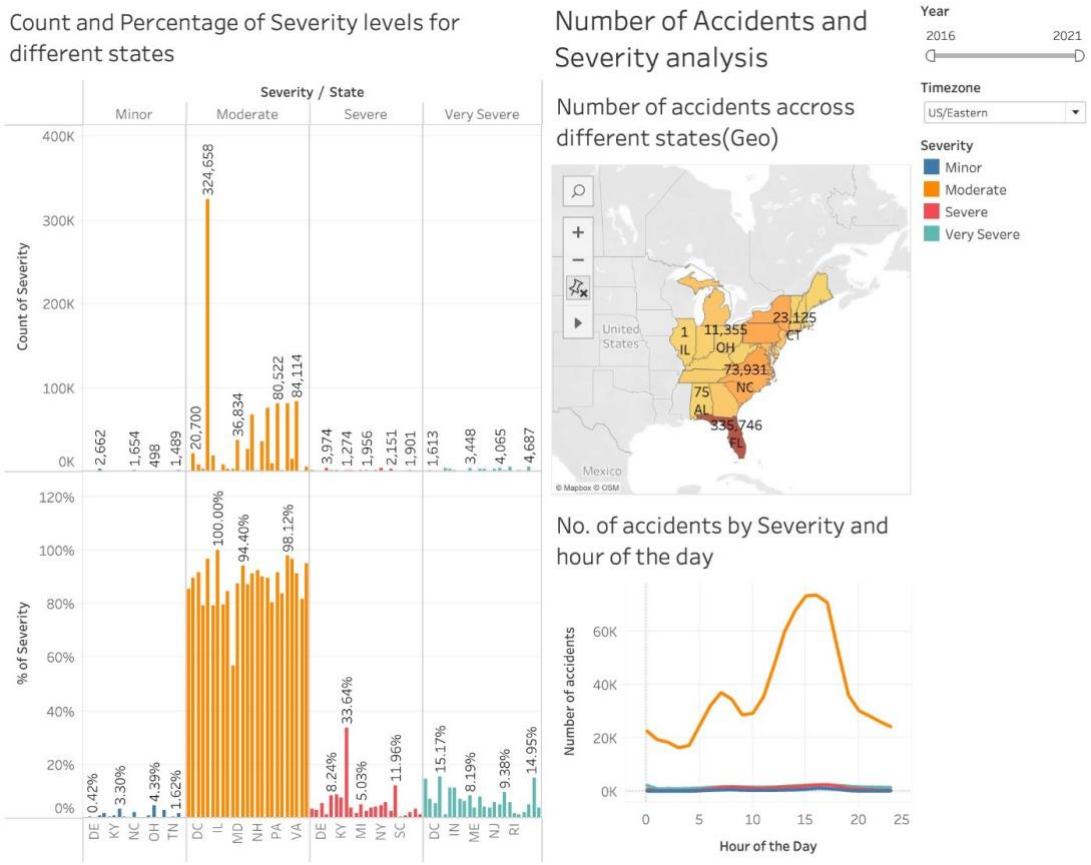
The first dashboard includes a generic and easy to follow visualizations of accidents happening across different states. The second dashboard involves a direct comparison between the days, on which days we have more accidents. The third dashboard is about the severity of the accidents and comparing whether the severity is more during daytime or nighttime. The fourth dashboard states the accidents based on weather conditions against severity. Our idea is that users will be able to use the first dashboard to get a high-level overview and then utilize the second dashboard to perform a more in-depth analysis into what each city can readily offer them.

Dashboard 1:

As previously stated, the first dashboard was created with the aim to provide the user high level insights on each of the cities individually as well as providing a comparative analysis across all cities. We have created a graph which gives the number of accidents happening across different states. The X-axis represents the states present in the US and the Y-axis represents the count. The highest state which has a greater number of accidents is Illinois and the state that has less number of accidents is Ohio. Similar, we have a Geo representation of number of accidents across different states.

The graph can help identify states with higher or lower numbers of accidents, allowing for comparisons between different states. It may reveal trends or patterns in the data, such as

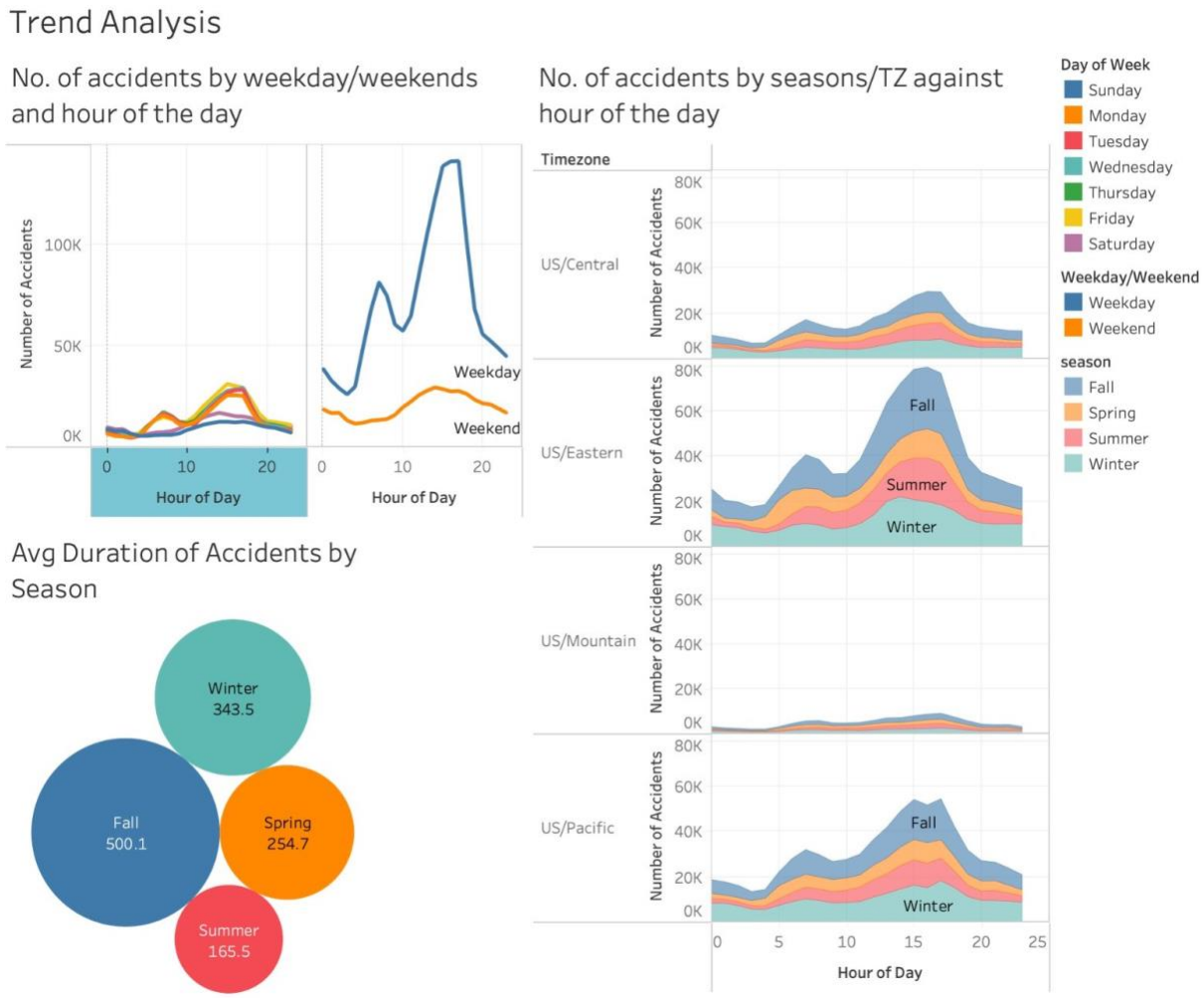
states with higher populations or higher traffic volumes having higher numbers of accidents. It may also highlight states with particularly low or high accident rates compared to the national average.



Dashboard 2:

Once the user has determined which two cities have a greater number of accidents. An additional visualization we were interested in seeing was the number of accidents based on the timeline. In the first graph we have number of accidents by hour of the day , where the X-axis represents hours of day and Y-axis represents the count .Initially the accidents are less during night time but the accidents are very high between 2pm to 5pm .Similarly if we are considering the accidents on day of week and hour of the day ,each day has different color representation .If we compared between the days Friday has more number of accidents when compared to the remaining days and Sunday serves to be less number of accidents taking place. On the other side

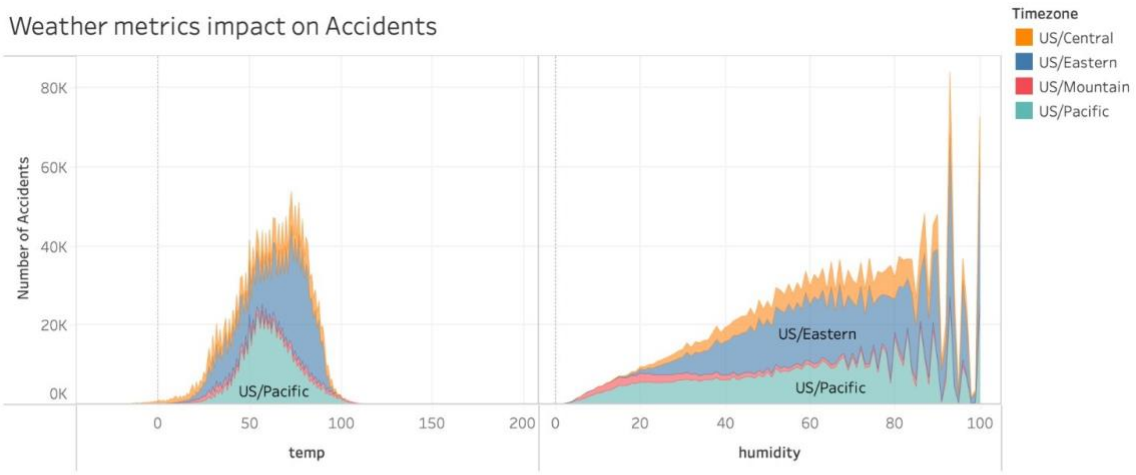
we can also see the accidents by season, in winter, summer, fall how many accidents are taking place.

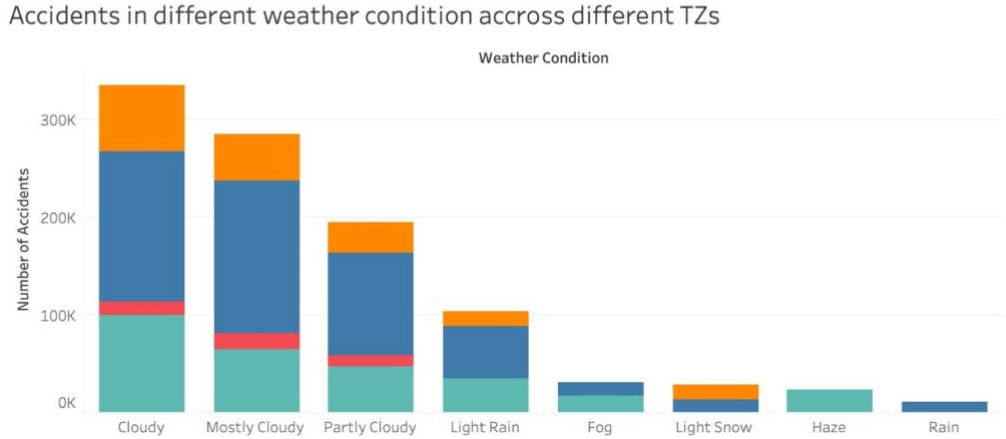


Dashboard 3:

In this dashboard we will investigate the severity of accidents based on weather conditions. There are more accidents taking place when the weather is cloudy followed by Mostly cloudy, partly cloudy (195,367). The graph may reveal trends or patterns in the data. For example, it may show that accidents tend to be more severe during certain weather conditions, such as high severity accidents being more common during foggy weather. It may also show that certain weather conditions are associated with low severity accidents,

such as clear weather having a higher percentage of low severity accidents compared to other weather conditions.





Limitations and Future Work:

The Country Wide Car Accidents project, which aims to analyze and understand the occurrence of car accidents across different states in the US, may have certain limitations and areas for future work.

One potential limitation of the project could be the accuracy and reliability of the data used. The analysis and conclusions drawn from the project would depend on the quality and completeness of the data available on car accidents, which may vary across states and jurisdictions. Incomplete

or inconsistent data, reporting biases, or data errors could impact the accuracy of the findings and limit the project's scope.

Another limitation could be the lack of detailed contextual information related to the car accidents. The graph showing the number of accidents across states may not capture other important factors such as weather conditions, road infrastructure, driver behavior, or traffic regulations that could significantly influence the occurrence of accidents. Further research could involve incorporating additional data and variables to gain a more comprehensive understanding of the factors contributing to car accidents.

In terms of future work, the project could explore more advanced statistical techniques or machine learning algorithms to analyze the data and identify patterns or trends that may not be readily apparent from a simple bar chart. For example, time-series analysis could be used to examine temporal patterns in accident occurrences, or spatial analysis could be used to investigate geographical clusters or hotspots of accidents.

Additionally, the project could benefit from conducting comparative studies to benchmark the accident data against established safety standards or performance indicators. This could help in identifying states or regions that deviate from the norm and require specific interventions or policy measures to improve road safety.

Finally, the project could also consider incorporating real-time or near real-time data on accidents to provide more up-to-date and dynamic insights. This could involve leveraging technologies such as Internet of Things (IoT) devices, telematics, or social media analytics to collect and analyze real-time data on accidents, which could potentially enhance the accuracy and timeliness of the findings.

Summary:

In summary, while the Country Wide Car Accidents project provides valuable insights into the occurrence of car accidents across states in the US, it may have limitations related to data accuracy and contextual information. Future work could involve advanced statistical techniques, comparative studies, and incorporation of real-time data to further enhance the project's findings and applicability for road safety interventions.

Using the Pandas package in Python, the data was thoroughly cleaned and processed before being connected to Tableau. The project consists of four dashboards. The first dashboard provides comparisons between various states by highlighting those with greater or lower accident rates. It could highlight patterns or trends in the data, such as a higher rate of accidents in states with larger populations or more traffic. In the second dashboard we compared between the days Friday has a greater number of accidents when compared to the remaining days and Sunday serves to be a smaller number of accidents taking place. In the third dashboard, after we were aware of the days that have the most accidents, we could look at the third dashboard to see how serious the accidents are. We have count and the severity categories of moderate, minor, severe, and very severe on the X-axis. There are 2,057,089 accidents with moderate severity, compared to 23,556 accidents with mild severity. In the fourth dashboard We will examine the seriousness of accidents dependent on weather conditions in this dashboard.