Artificial Intelligence for Business Decisions and Transformation

CSCN8030 - Spring 2024 - Section 2

Sprint 1 - Demo

27 May 2024

Professor:
Glaucia Melo dos Santos, PhD

Group 4 - Members:

Krishna Kumar, Hemasree
Shijin, Jency
Fernandez, Arcadio

Project: US Traffic Accident Severity Prediction

US Accidents (2016 - 2023)
A Countrywide Traffic Accident Dataset (2016 - 2023)

1. Introduction: A brief introduction to the problem, libraries and obtaining the data.

+ Code + Markdown

1.1. A brief introduction to the problem

1.1.1. Why this is topic important?

There were 39,508 fatal motor vehicle crashes in the United States in 2021 in which 42,939 deaths occurred. This resulted in 12.9 deaths per 100,000 people and 1.37 deaths per 100 million miles traveled. The fatality rate per 100,000 people ranged from 5.7 in Rhode Island to 26.2 in Mississippi. The death rate per 100 million miles traveled ranged from 0.71 in Massachusetts to 2.08 in South Carolina.

1.1.2. How could this project make an impact in society?

Accident severity modeling helps understand contributing factors and develop preventive strategies. AI models, such as random forest, offer adaptability and higher predictive accuracy compared to traditional statistical models. This study aims to develop a predictive model for traffic accident severity on USA highways ML algorithm.

1.1.3. Data description

This is a countrywide car accident dataset that covers 49 states of the USA. The accident data were collected from February 2016 to March 2023, using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by various entities, including the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks. The dataset currently contains approximately 7.7 million accident records.For more information about this dataset, please visit Kaggle: Dataset

1.1.4. Indication of Reference Code

On Kaggle at US Accidents (2016 - 2023) Dataset, there are 371 codes. Two of them caught our attention because of its organization and the way the result is exposed:

https://www.kaggle.com/code/jingzongwang/usa-car-accidents-severity-prediction
https://www.kaggle.com/code/satyabrataroy/60-insights-extraction-us-accident-analysis

2. Exploratory data analysis - EDA

2.1. Basic data analysis

▷ 5 cells hidden ···

2.2. Statistical analysis

▷ 1 cell hidden ···

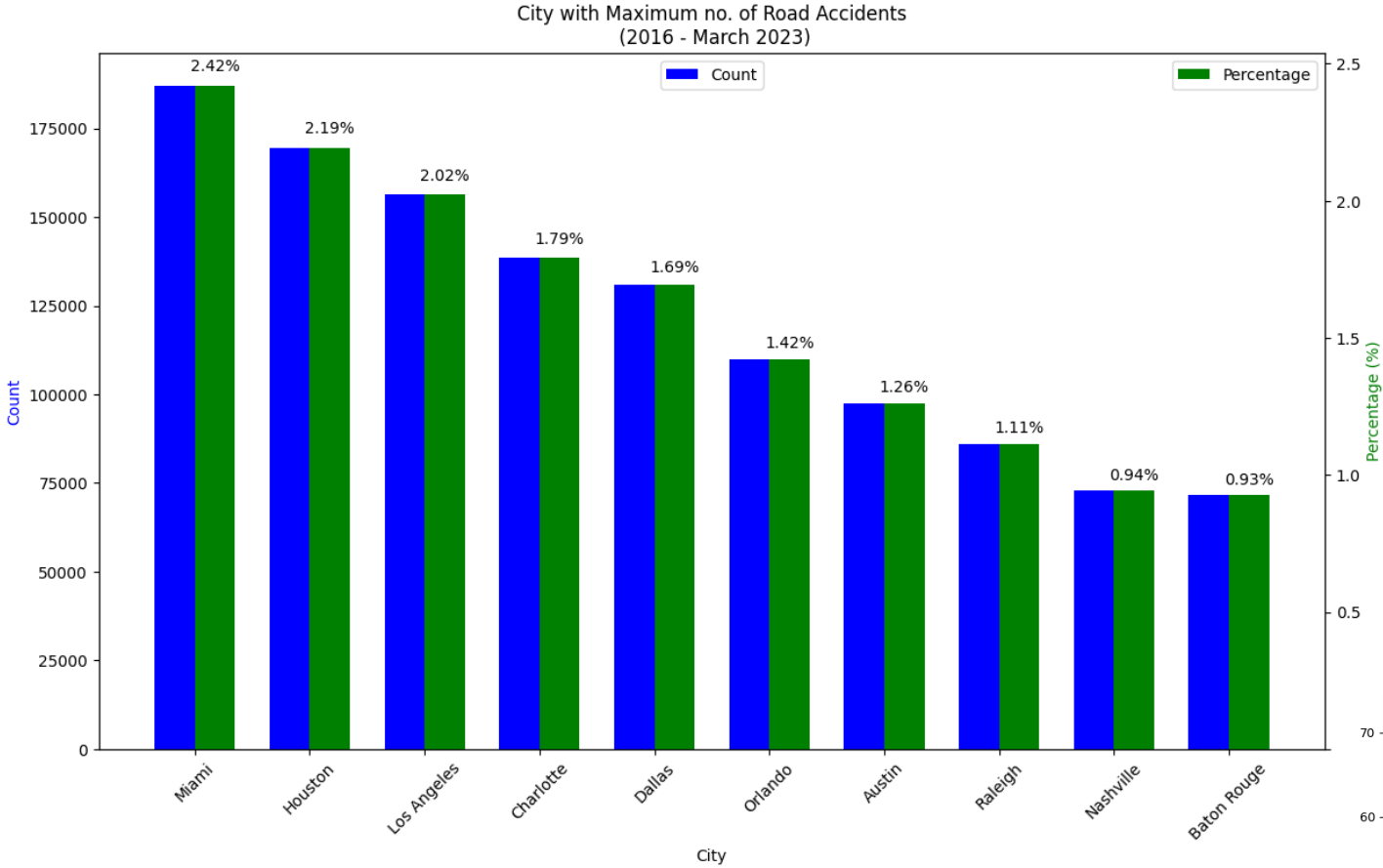
2.3. City Analysis

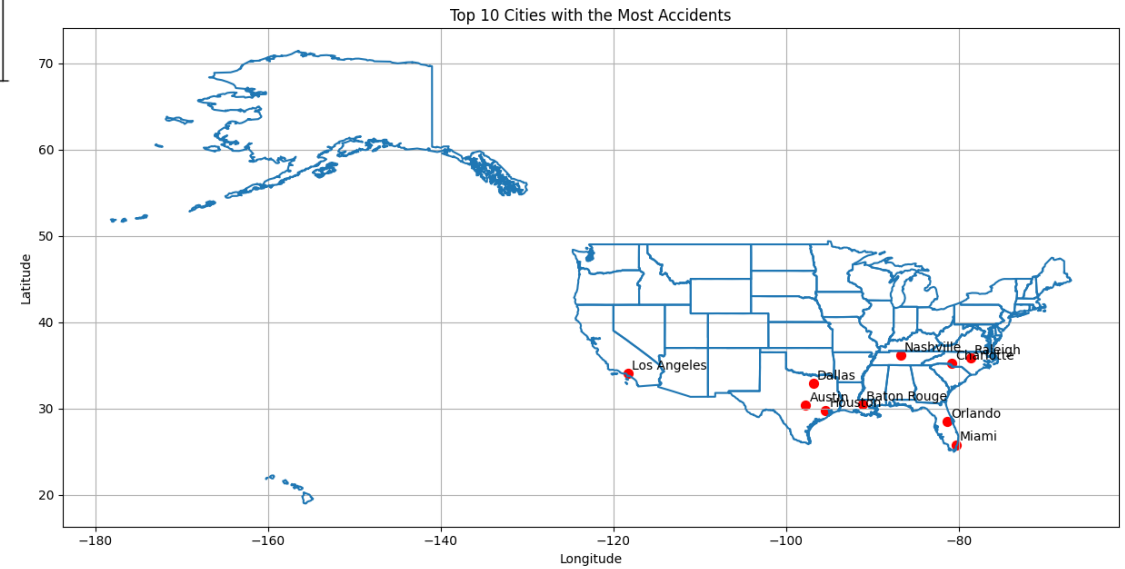
▷ 4 cells hidden ···

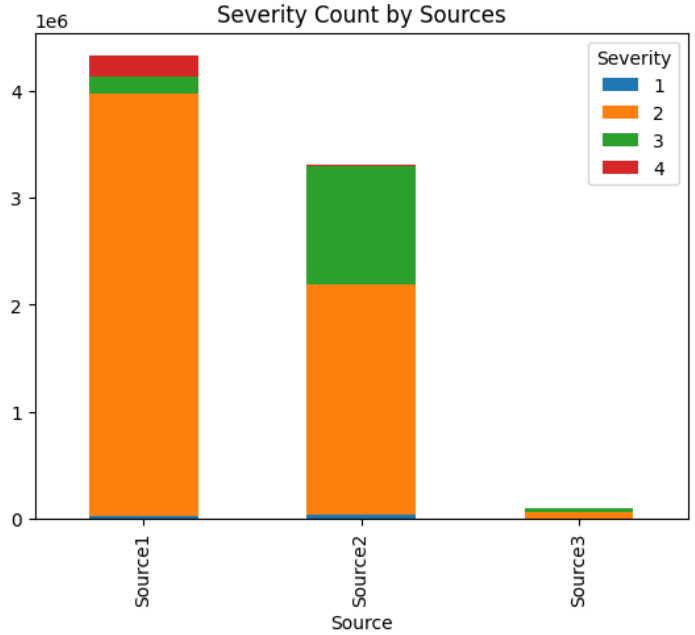
2.4. Severity Analysis

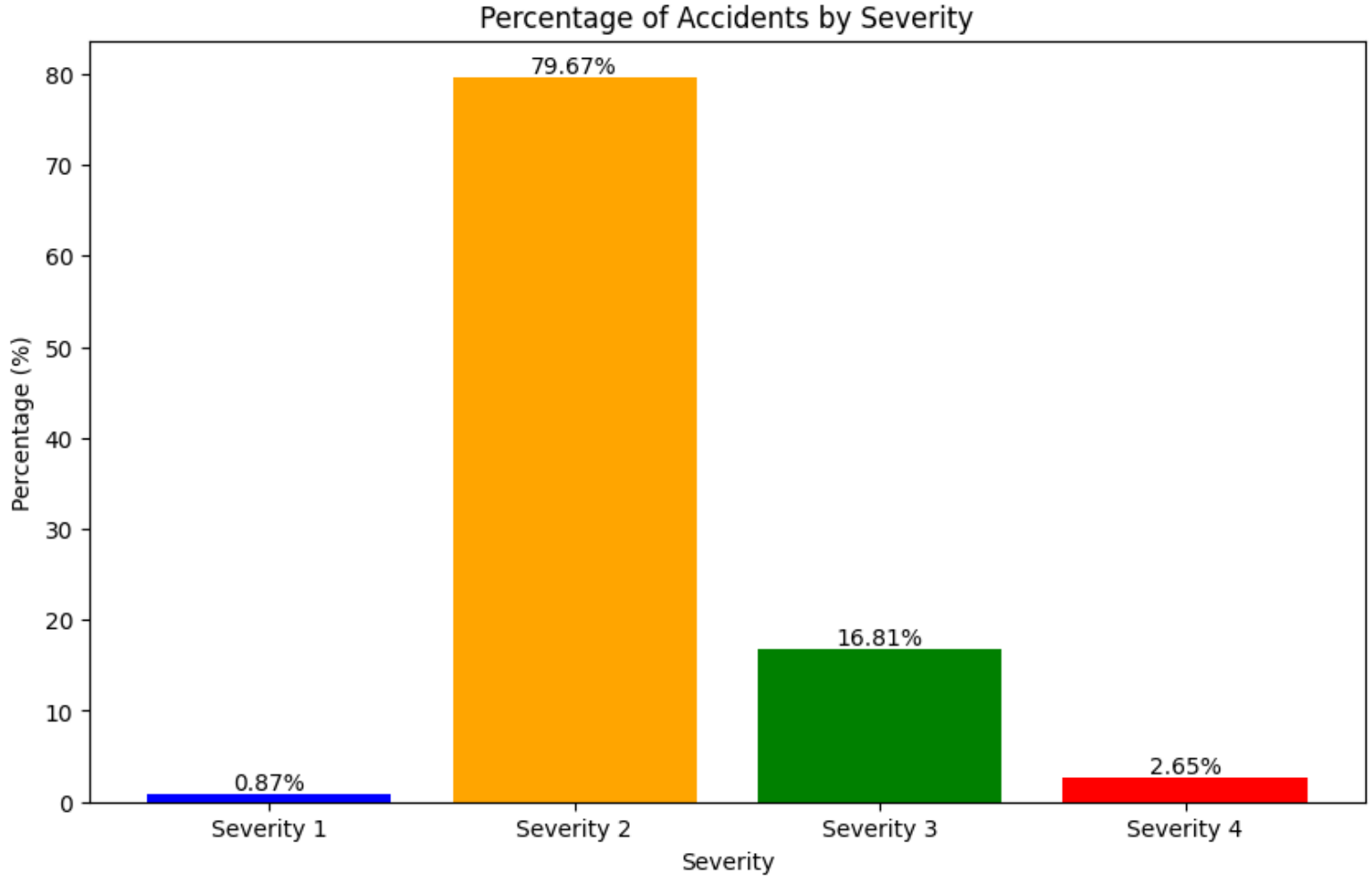
▷ 7 cells hidden ···

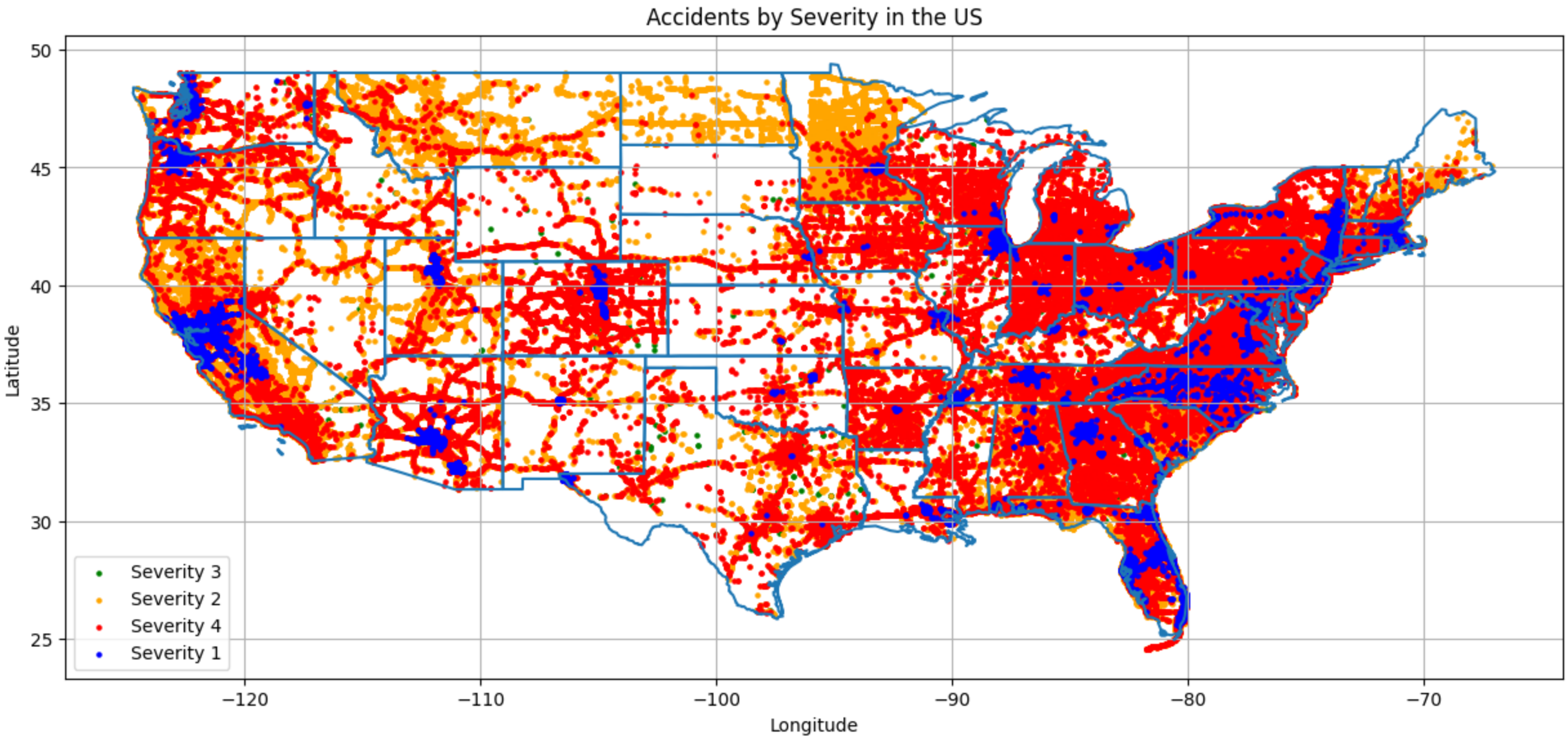
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7728394 entries, 0 to 7728393
Data columns (total 46 columns):
 #   Column                 Dtype
---  ------                 -----
 0   ID                     object
 1   Source                 object
 2   Severity               int64
 3   Start_Time             object
 4   End_Time               object
 5   Start_Lat              float64
 6   Start_Lng              float64
 7   End_Lat                float64
 8   End_Lng                float64
 9   Distance(mi)           float64
 10  Description            object
 11  Street                 object
 12  City                   object
 13  County                 object
 14  State                  object
 15  Zipcode                object
 16  Country                object
 17  Timezone               object
 18  Airport_Code           object
 19  Weather_Timestamp      object
 20  Temperature(F)         float64
 21  Wind_Chill(F)          float64
 22  Humidity(%)            float64
 23  Pressure(in)           float64
 24  Visibility(mi)         float64
 25  Wind_Direction         object
 26  Wind_Speed(mph)        float64
 27  Precipitation(in)      float64
 28  Weather_Condition      object
 29  Amenity                bool
 30  Bump                   bool
 31  Crossing               bool
 32  Give_Way               bool
 33  Junction               bool
 34  No_Exit                bool
 35  Railway                bool
 36  Roundabout             bool
 37  Station                bool
 38  Stop                   bool
 39  Traffic_Calming        bool
 40  Traffic_Signal         bool
 41  Turning_Loop           bool
 42  Sunrise_Sunset         object
 43  Civil_Twilight         object
 44  Nautical_Twilight      object
 45  Astronomical_Twilight  object
dtypes: bool(13), float64(12), int64(1), object(20)
memory usage: 2.0+ GB
```

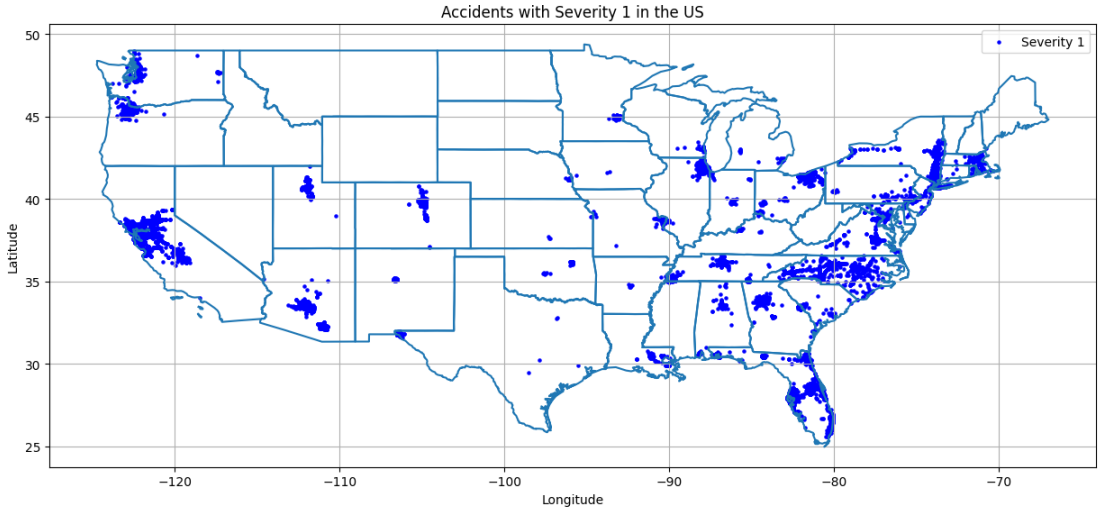

City with Maximum no. of Road Accidents
(2016 - March 2023)
Count
Percentage
Count
Percentage (%)
City
2.42%
2.19%
2.02%
1.79%
1.69%
1.42%
1.26%
1.11%
0.94%
0.93%
Miami
Houston
Los Angeles
Charlotte
Dallas
Orlando
Austin
Raleigh
Nashville
Baton Rouge

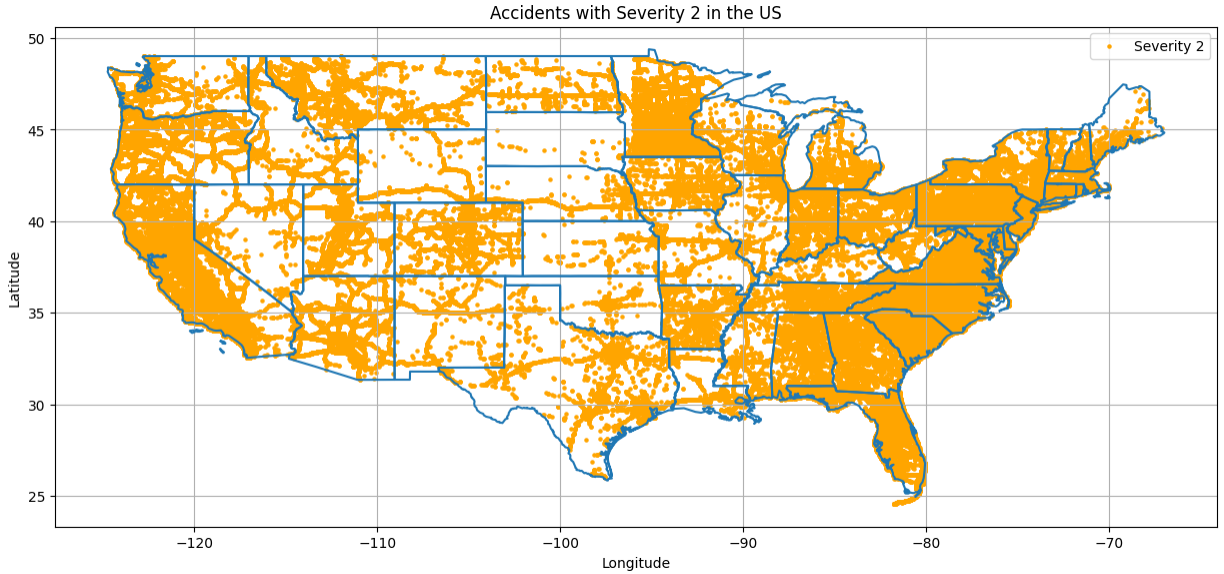

Top 10 Cities with the Most Accidents
Latitude
Longitude
Nashville
Raleigh
Charlotte
Los Angeles
Dallas
Austin
Baton Rouge
Houston
Orlando
Miami

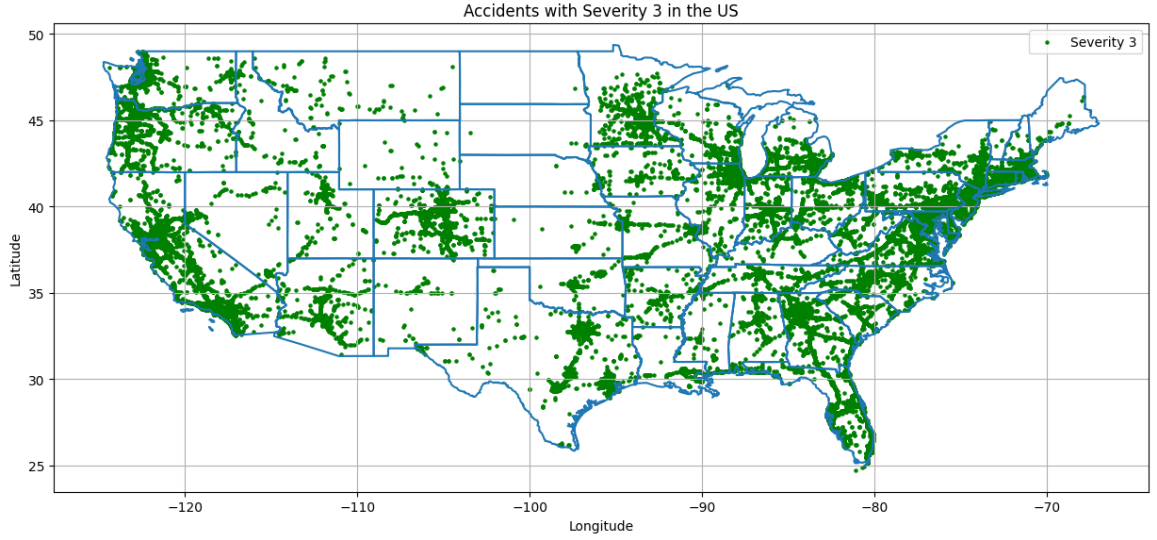

Severity Count by Sources
1e6
Severity
1
2
3
4
0
1
2
3
4
Source1
Source2
Source3
Source

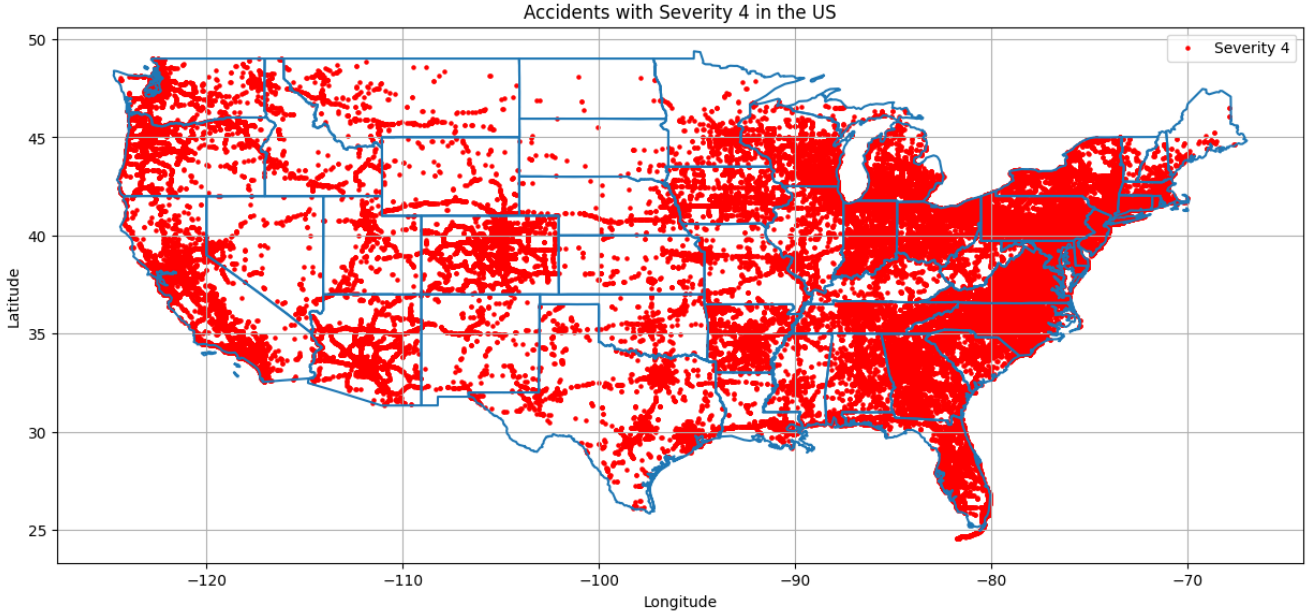

Percentage of Accidents by Severity
Percentage (%)
0
10
20
30
40
50
60
70
80
0.87%
79.67%
16.81%
2.65%
Severity 1
Severity 2
Severity 3
Severity 4
Severity


Accidents by Severity in the US
Latitude
Longitude
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70
Severity 3
Severity 2
Severity 4
Severity 1

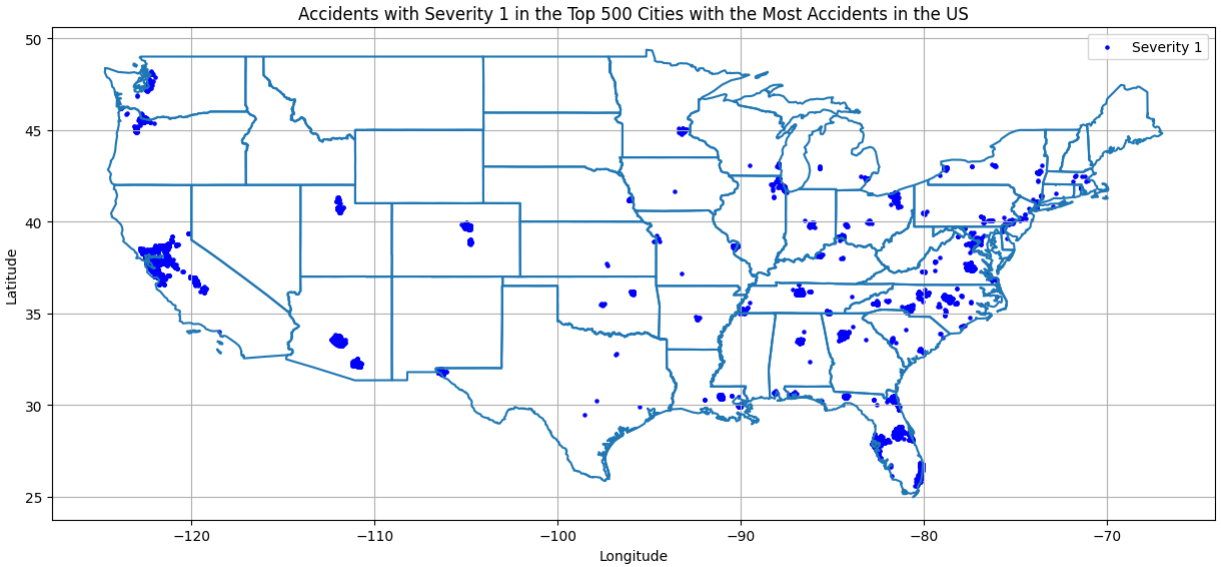

Accidents with Severity 1 in the US
Severity 1
Latitude
Longitude
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70

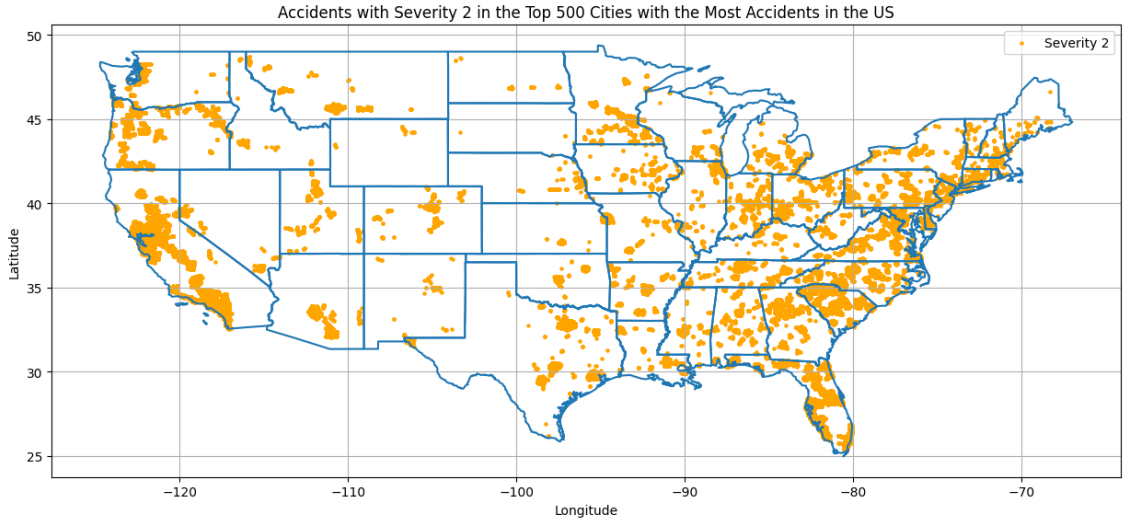

Accidents with Severity 2 in the US
Severity 2
Latitude
Longitude
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70

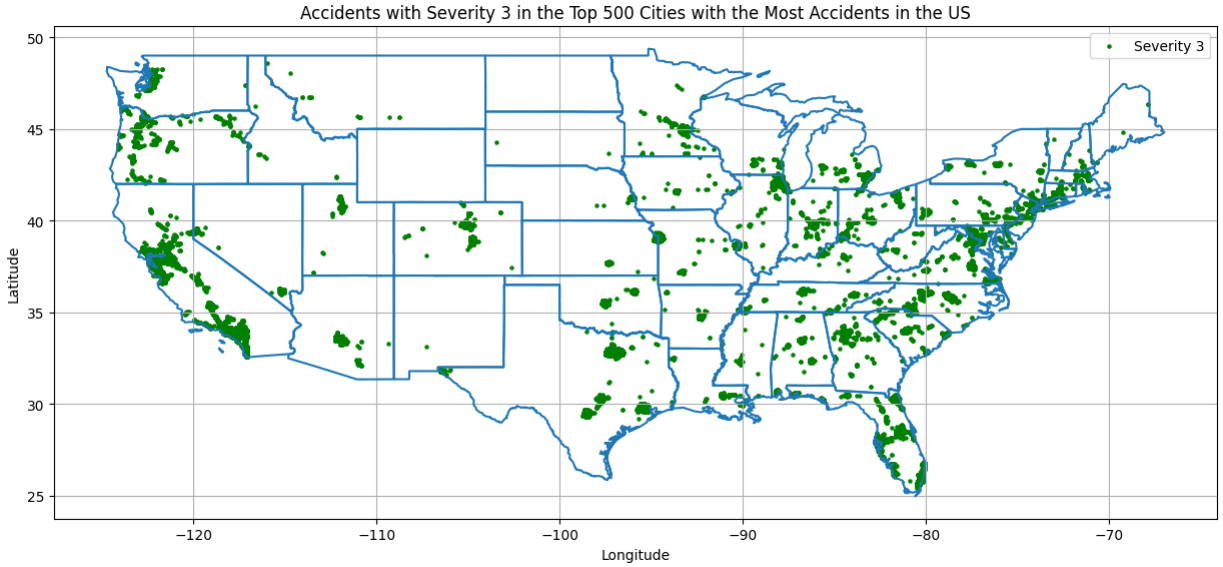

Accidents with Severity 3 in the US
Severity 3
Latitude
Longitude
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70

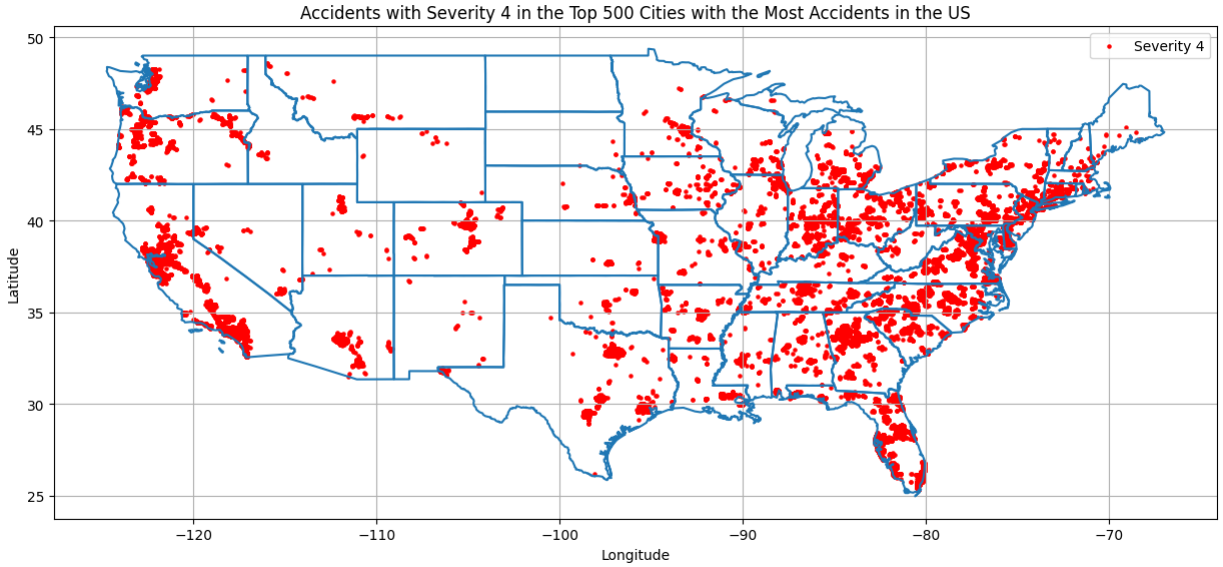

Accidents with Severity 4 in the US
Severity 4
Latitude
Longitude
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70

Accidents with Severity {severity} in the Top 500 Cities with the Most Accidents in the US'









3. Data Preprocessing

3.1. Drop some unneeded columns in the dataset

3 cells hidden

3.2. Drop duplicates in the dataset

1 cell hidden

3.3. Handle duplicate values in columns

5 cells hidden

3.4. Check for missing values in the dataset

3 cells hidden

3.5. Handling with missing values in the dataset

3 cells hidden

3.6 Add Features useful for prediction

1 cell hidden

3.7. Encode categorical variables into numerical format using techniques like one-hot encoding or label encoding.

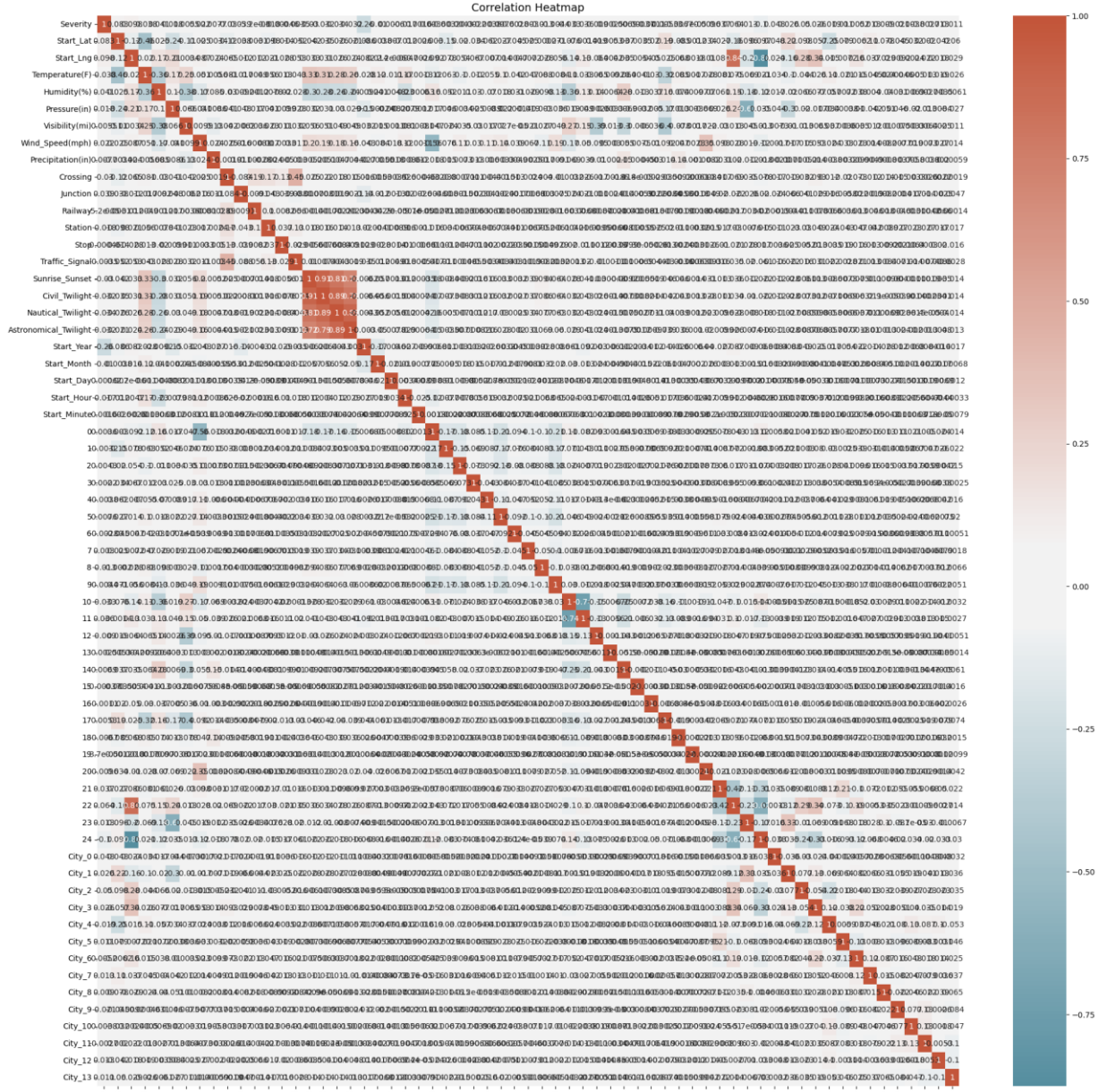
9 cells hidden

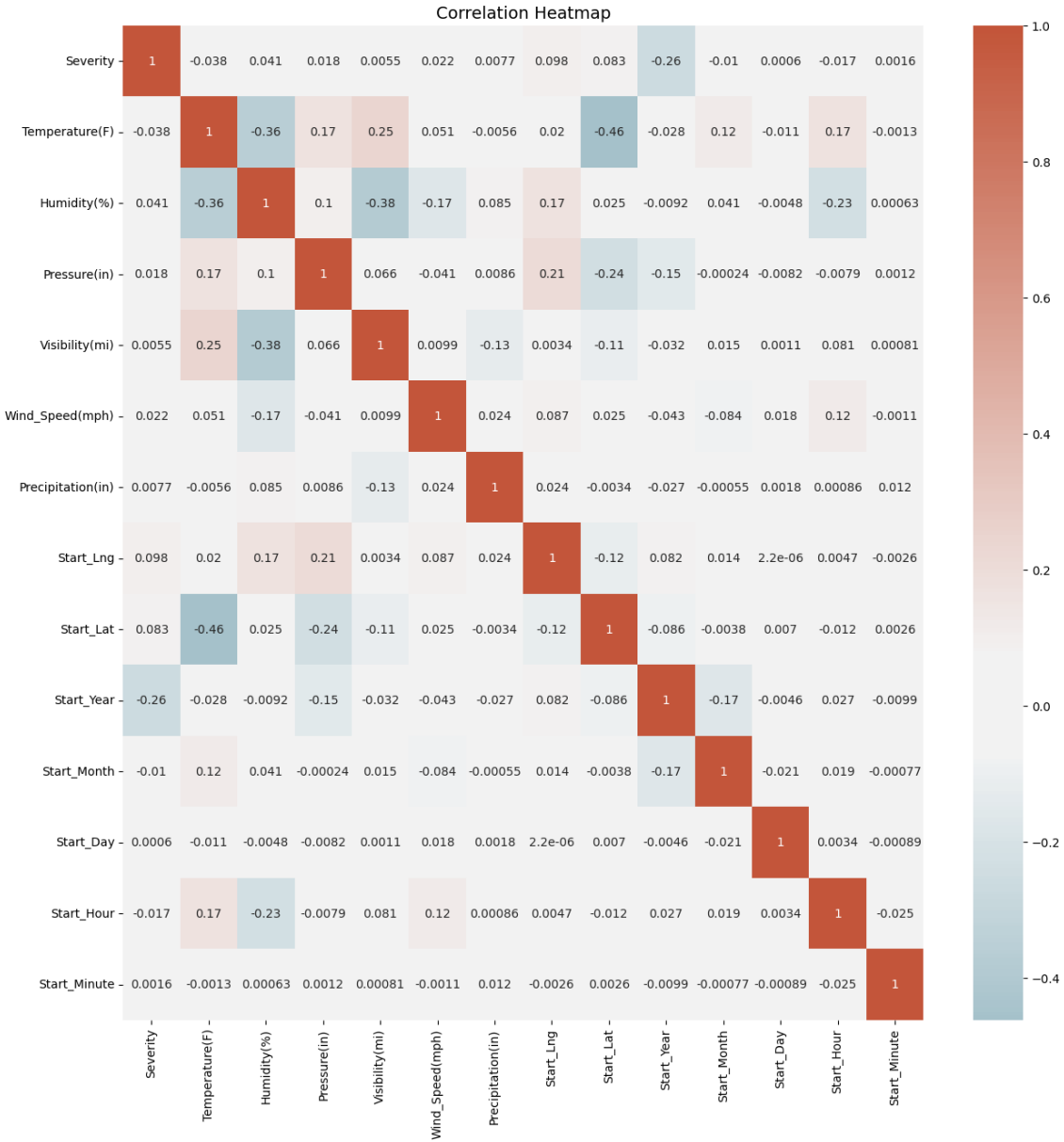
3.8. Correlation matrix

3 cells hidden

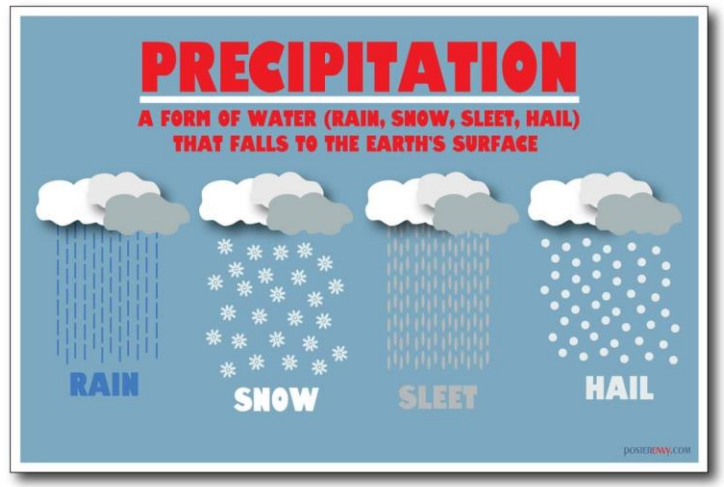
3.9. Scale and normalize the features

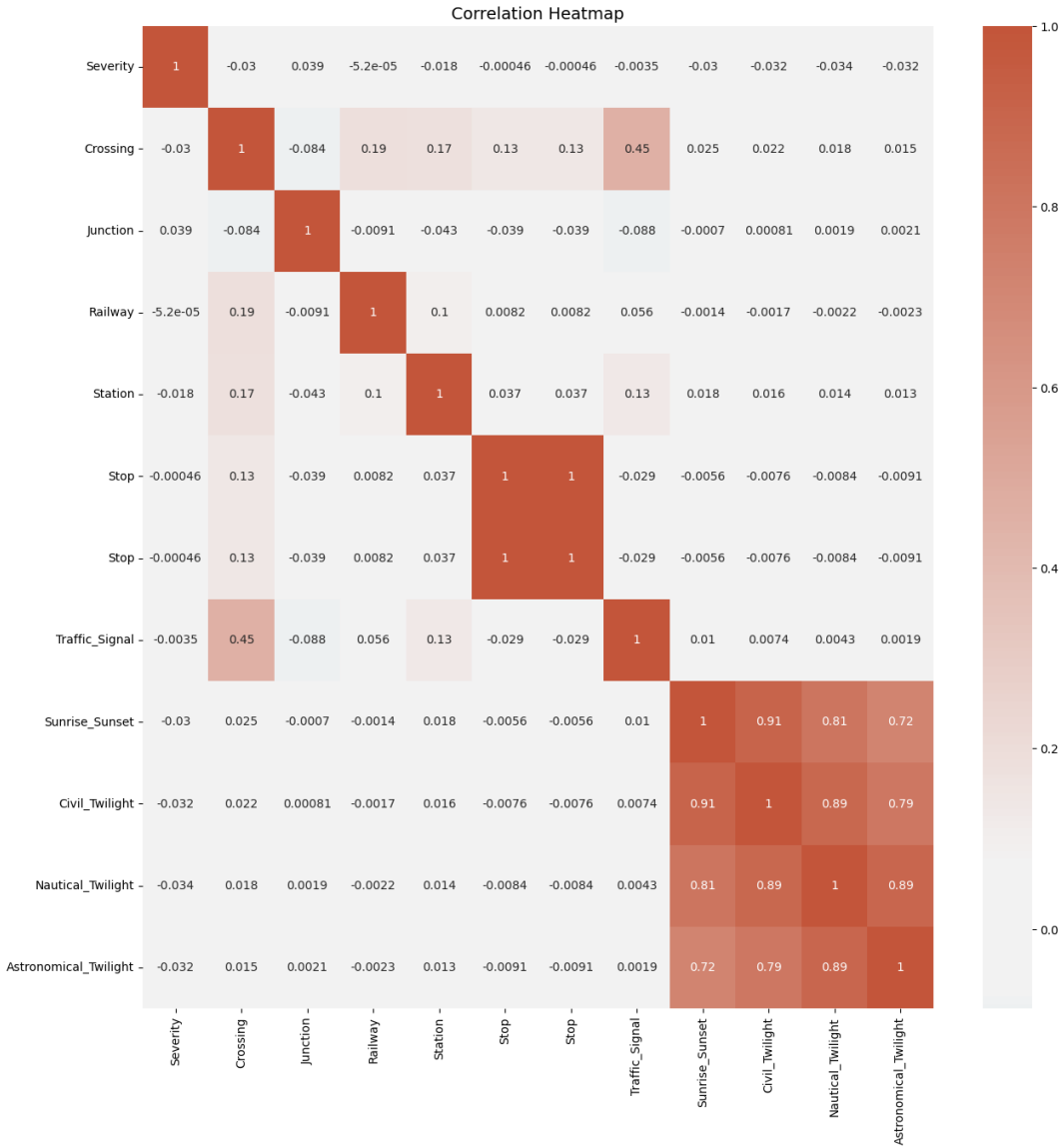
5 cells hidden


Correlation Heatmap
Severity
Start_Lat
Start_Lng
Temperature(F)
Humidity(%)
Pressure(in)
Visibility(mi)
Wind_Speed(mph)
Precipitation(in)
Crossing
Junction
Railway
Station
Stop
Traffic_Signal
Sunrise_Sunset
Civil_Twilight
Nautical_Twilight
Astronomical_Twilight
Start_Year
Start_Month
Start_Day
Start_Hour
Start_Minute
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
City_0
City_1
City_2
City_3
City_4
City_5
City_6
City_7
City_8
City_9
City_10
City_11
City_12
City_13
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

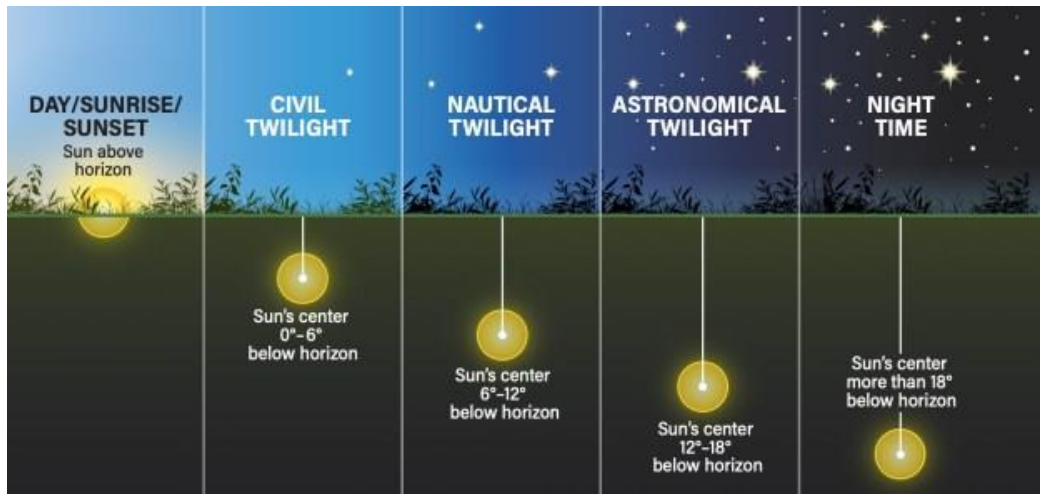

Correlation Heatmap
Severity 1 -0.038 0.041 0.018 0.0055 0.022 0.0077 0.098 0.083 -0.26 -0.01 0.0006 -0.017 0.0016
Temperature(F) -0.038 1 -0.36 0.17 0.25 0.051 -0.0056 0.02 -0.46 -0.028 0.12 -0.011 0.17 -0.0013
Humidity(%) 0.041 -0.36 1 0.1 -0.38 -0.17 0.085 0.17 0.025 -0.0092 0.041 -0.0048 -0.23 0.00063
Pressure(in) 0.018 0.17 0.1 1 0.066 -0.041 0.0086 0.21 -0.24 -0.15 -0.00024 -0.0082 -0.0079 0.0012
Visibility(mi) 0.0055 0.25 -0.38 0.066 1 0.0099 -0.13 0.0034 -0.11 -0.032 0.015 0.0011 0.081 0.00081
Wind_Speed(mph) 0.022 0.051 -0.17 -0.041 0.0099 1 0.024 0.087 0.025 -0.043 -0.084 0.018 0.12 -0.0011
Precipitation(in) 0.0077 -0.0056 0.085 0.0086 -0.13 0.024 1 0.024 -0.0034 -0.027 -0.00055 0.0018 0.00086 0.012
Start_Lng 0.098 0.02 0.17 0.21 0.0034 0.087 0.024 1 -0.12 0.082 0.014 2.2e-06 0.0047 -0.0026
Start_Lat 0.083 -0.46 0.025 -0.24 -0.11 0.025 -0.0034 -0.12 1 -0.086 -0.0038 0.007 -0.012 0.0026
Start_Year -0.26 -0.028 -0.0092 -0.15 -0.032 -0.043 -0.027 0.082 -0.086 1 -0.17 -0.0046 0.027 -0.0099
Start_Month -0.01 0.12 0.041 -0.00024 0.015 -0.084 -0.00055 0.014 -0.0038 -0.17 1 -0.021 0.019 -0.00077
Start_Day 0.0006 -0.011 -0.0048 -0.0082 0.0011 0.018 0.0018 2.2e-06 0.007 -0.0046 -0.021 1 0.0034 -0.00089
Start_Hour -0.017 0.17 -0.23 -0.0079 0.081 0.12 0.00086 0.0047 -0.012 0.027 0.019 0.0034 1 -0.025
Start_Minute 0.0016 -0.0013 0.00063 0.0012 0.00081 -0.0011 0.012 -0.0026 0.0026 -0.0099 -0.00077 -0.00089 -0.025 1
Severity
Temperature(F)
Humidity(%)
Pressure(in)
Visibility(mi)
Wind_Speed(mph)
Precipitation(in)
Start_Lng
Start_Lat
Start_Year
Start_Month
Start_Day
Start_Hour
Start_Minute
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


30-60%
25-30%
60-70%
<25%
>70%
HUMIDITY METER


PRECIPITATION
A FORM OF WATER (RAIN, SNOW, SLEET, HAIL)
THAT FALLS TO THE EARTH'S SURFACE
RAIN
SNOW
SLEET
HAIL



Road junction



4. Splitting the Dataset

[52]
```
X = df_accidents_sparse_encoded.drop('Severity', axis=1)  # Features
y = df_accidents_sparse_encoded['Severity']  # Target variable
```

[53]
```
X.columns = X.columns.astype(str)
```

[54]
```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report
```

[55]
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

[56]
```
X_train.shape
```

```
(3045011, 62)
```

[57]
```
X_test.shape
```

```
(761253, 62)
```

5. Model Training

5.1. Training

[58]
```
model = RandomForestClassifier(n_estimators=100, random_state=42)
```

[59]
```
model.fit(X_train, y_train)
```

```
c:\Users\arcad\AppData\Local\Programs\Python\Python311\Lib\site-packages\sk
  warnings.warn(
```

RandomForestClassifier

```
RandomForestClassifier(random_state=42)
```

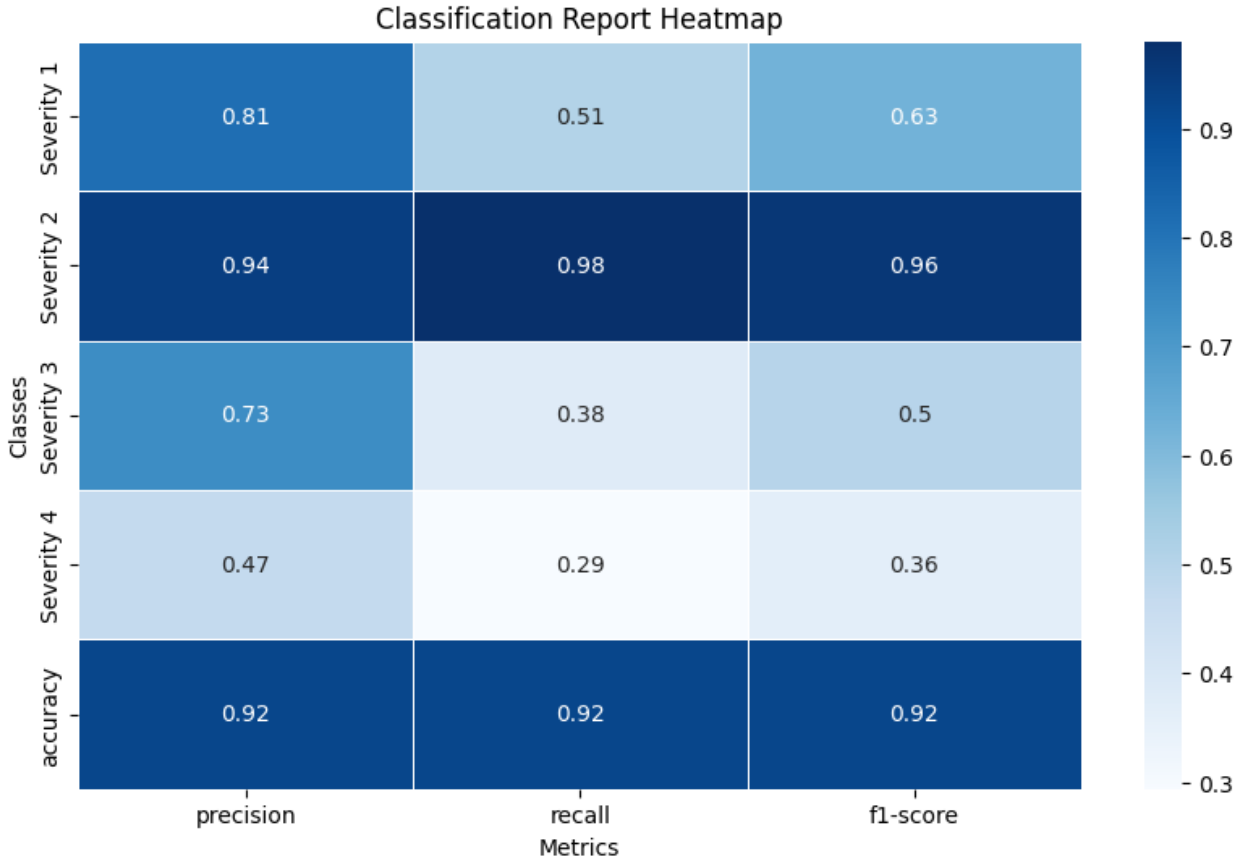
5.2. Saving the Model

[60]
```
import joblib
```

[61]
```
joblib.dump(model, 'random_forest_model.joblib')
```

```
['random_forest_model.joblib']
```

6. Model Evaluation

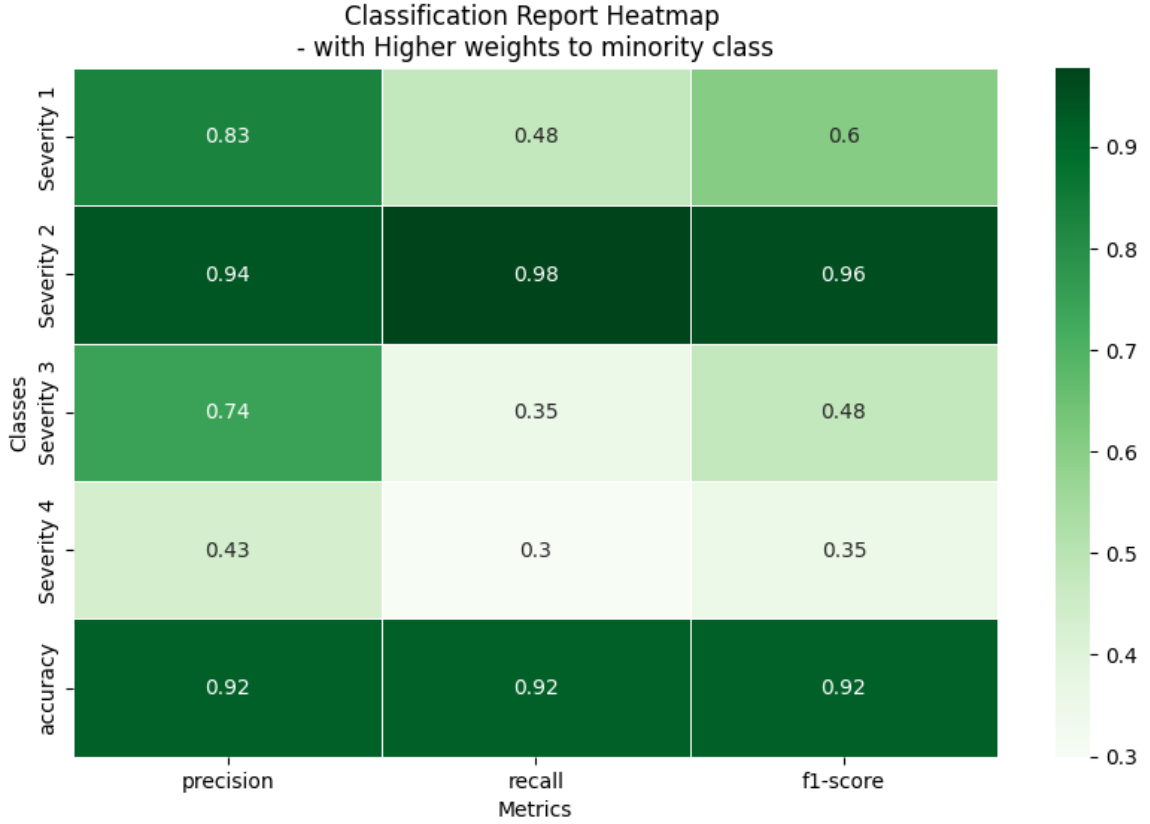


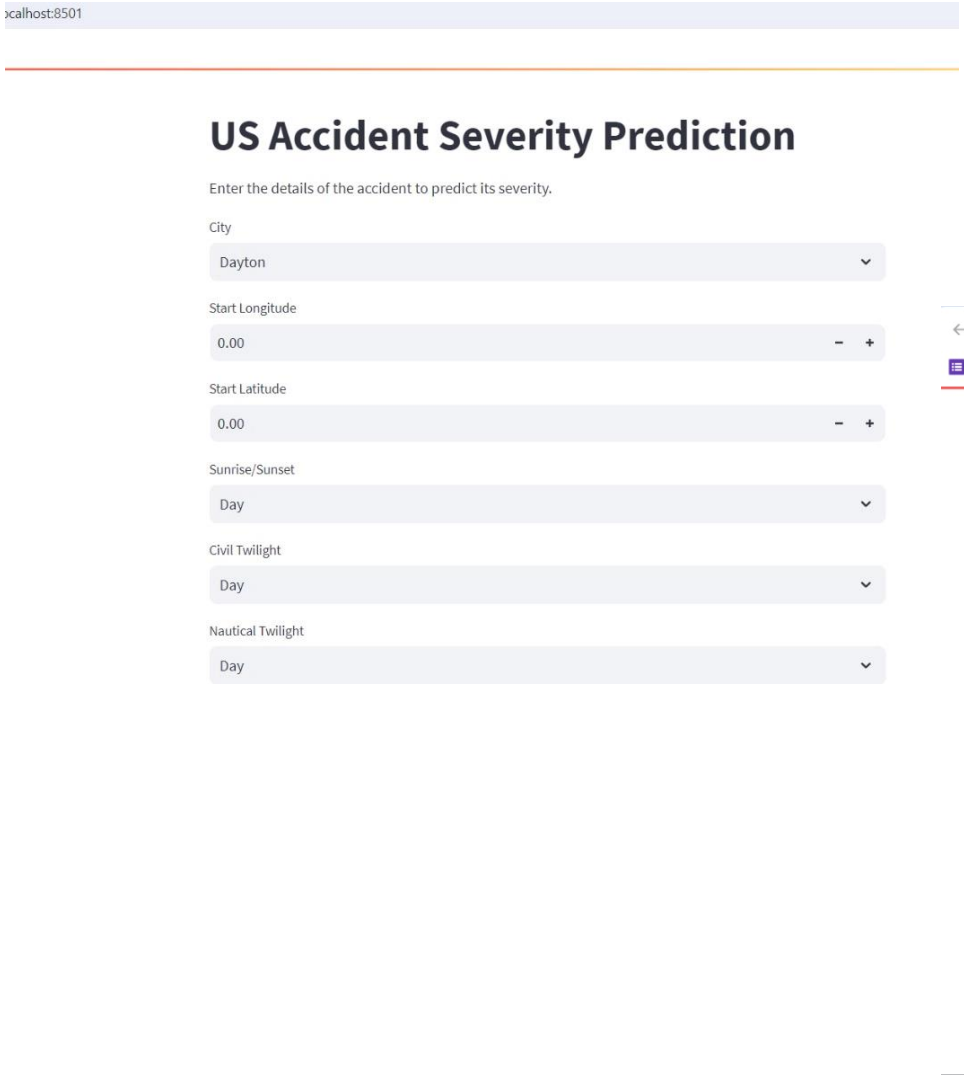
6.2. Handling Class Imbalance by training a Model by assigning Higher weights to minority class

```
model2 = RandomForestClassifier(class_weight={1: 1, 2: 1, 3: 1, 4: 10}, random_state=42)
model2.fit(X_train, y_train)
```

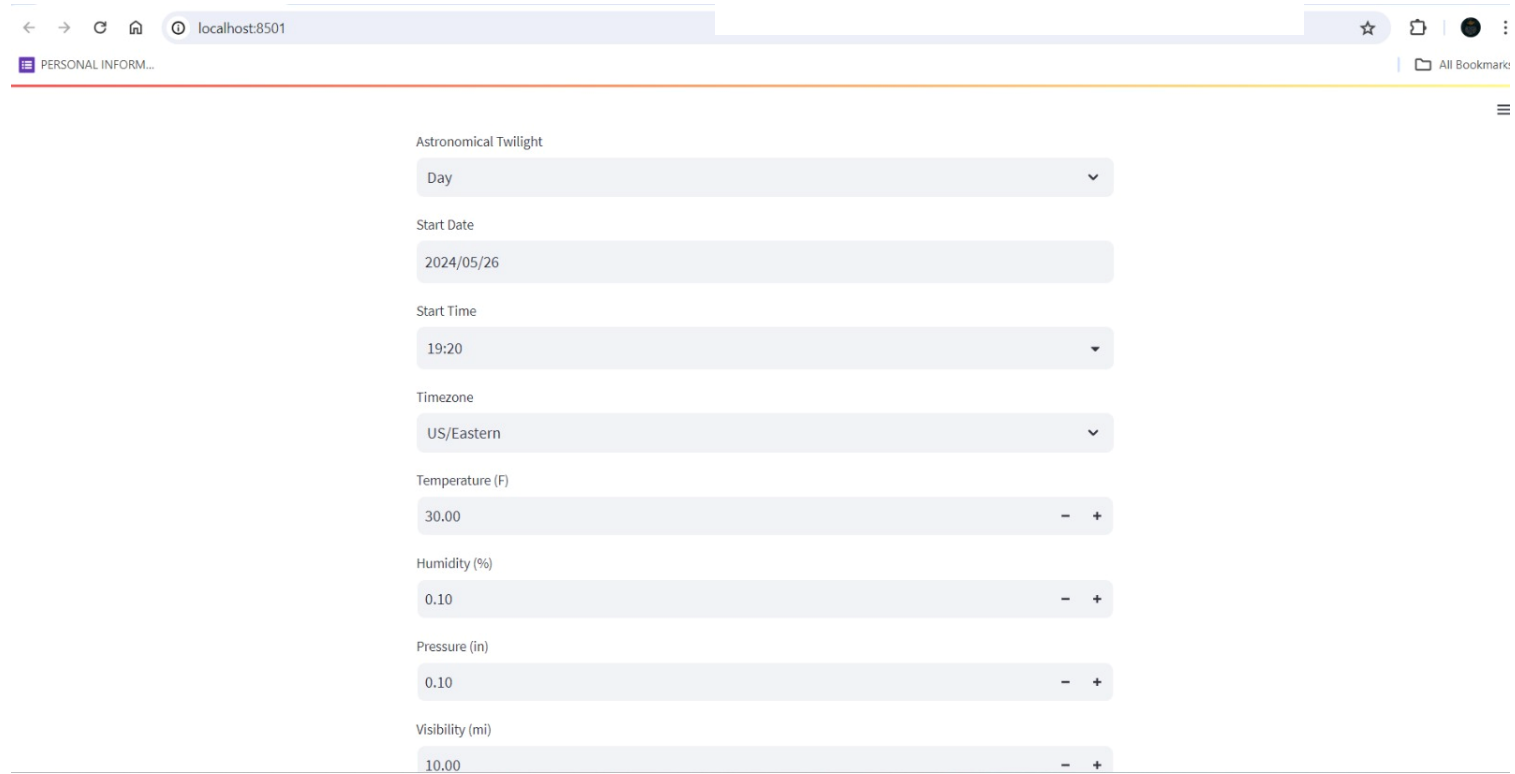
RandomForestClassifier

```
RandomForestClassifier(class_weight={1: 1, 2: 1, 3: 1, 4: 10}, random_state=42)
```




localhost:8501
US Accident Severity Prediction
Enter the details of the accident to predict its severity.
City
Dayton
Start Longitude
0.00
Start Latitude
0.00
Sunrise/Sunset
Day
Civil Twilight
Day
Nautical Twilight
Day

Streamlit


localhost:8501
PERSONAL INFORM...
All Bookmarks
Astronomical Twilight
Day
Start Date
2024/05/26
Start Time
19:20
Timezone
US/Eastern
Temperature (F)
30.00
Humidity (%)
0.10
Pressure (in)
0.10
Visibility (mi)
10.00

Wind Speed (mph)

80.00 − +

Precipitation (in)

1.00 − +

Wind Direction

W

Weather Condition

Rain

Crossing

- Yes
- No

Junction

- Yes
- No

Railway

- Yes
- No

PERSONAL INFORM... | All Bookr

Junction

- Yes
- No

Railway

- Yes
- No

Station

- Yes
- No

Stop

- Yes
- No

Predict Severity

The predicted severity of the accident is: Severity 2

impact

A total of 42,939 people died in motor vehicle crashes in 2021. The U.S. Department of Transportation's most recent estimate of the annual economic cost of crashes is $340 billion.

https://www.iihs.org/topics/fatality-statistics/detail/yearly-snapshot

In 2022, property and casualty insurance premiums written in the United States amounted to 715.9 billion U.S. dollars, while life and annuity premiums stood at 635.7 billion U.S. dollars.

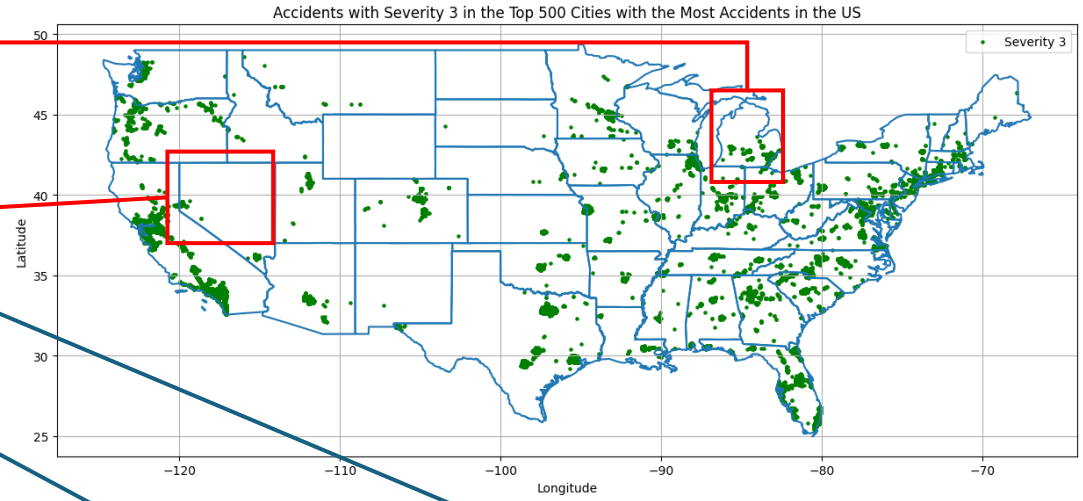
https://www.statista.com/statistics/1102810/insurance-market-size-usa-by-type/

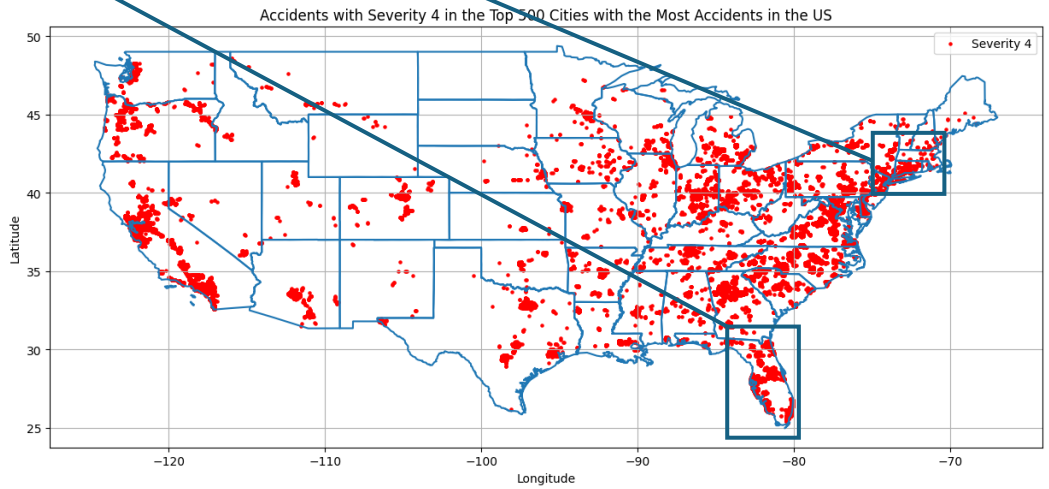
- Nevada →
- factors responsible increased premiums include inflation and the increased costs for repairs and parts, and more drivers in Nevada engaging in riskier behavior behind the wheel.
- Labor shortage

Estimated annual car insurance premiums in the United States from 2021 to 2023, by state*(in U.S. dollars)*

Search: Records: 13

Characteristic	2021	2022	2023
Michigan	5,740	4,386	2,352
Rhode Island	1,375	1,197	1,200
Nevada	1,033	1,138	1,164
Florida	2,361	2,072	1,092
New Jersey	812	979	1,032
Delaware	1,200	1,183	1,008
Connecticut	1,165	1,041	960
Oregon	1,050	996	948
New York	1,373	1,085	924
Maryland	1,081	1,044	900
Kentucky	1,549	1,027	876
Louisiana	1,128	1,002	876
Utah	909	793	792





Why it varies state by state

The huge variance in premiums between states is due to the difference in state laws, the percentage of uninsured drivers in the state, the frequency of natural disasters and claim rates. For instance, Michigan has a no-fault car insurance system, which means that claims are more common. This drives up the cost of insurance for all drivers, because insurers need to pay out more money in claims.

https://www.statista.com/statistics/675367/annual-auto-insurance-premiums-usa-by-state/

Thank you for your attention.

Any questions?