Project

# Pattern Recognition & Machine Learning (CSL2050)

## [US Accidents (2016 - 2021)


### Abstract

The USAccidentsdataset containsinformationabout traffic accidentsthat occurredin the United StatesfromFebruary 2016 toDecember 2020. Thedataset includesfeatures suchastheseverity of theaccident, locationcoordinates, weather conditions, timeof day,andmore. Thisreport analysesthedataset topredict accident severity andlocation basedonthegivenfeatures. Weexplorethedata tounderstandtherelationshipsbetween thefeaturesandthetarget variables. Wepre-processthedata by handlingmissing values, droppingunnecessary features, andencodingcategorical variables. Wethen build machinelearningmodelstopredict accident severity andlocationbasedontheselected features. For predictingseverity, weuseclassificationalgorithmssuchasDecisionTree Classifier, RandomForest Classifier, andSupport Vector Machine. For predictingthe location, weuseregressionalgorithmssuchasLinear Regression, RandomForest Regression. Weevaluatethemodelsusingtheaccuracy scoreandf1 score. Based onthe results, weprovideinsightsintothefactorsthat contributetoaccident severity and location, suchasweather conditions, timeof day, androadfeatures. Wesuggest waysto improveroadsafety andreducethenumber of accidentsbasedontheseinsights.

# Problem Statement

**Accidents** on road are a major cause of **death** and injury worldwide. Accidents can have a significant impact on individuals, families, and society as a whole. So, predicting accident severity and location can help authorities in implementing better safety measures and infrastructure to reduce accidents and their severity. It can help **emergency services respond quickly and efficiently**, potentially reducing the severity of injuries and **loss of life**. For this, machine learning can be a great help.

# About Dataset

This is a countrywide car accident dataset, which covers **49 states of the USA**. The accident data are collected from **February 2016 to Dec 2021**, using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about **2.8 million** accident records in this dataset. Check here to learn more about this dataset.

# Exploring the Dataset

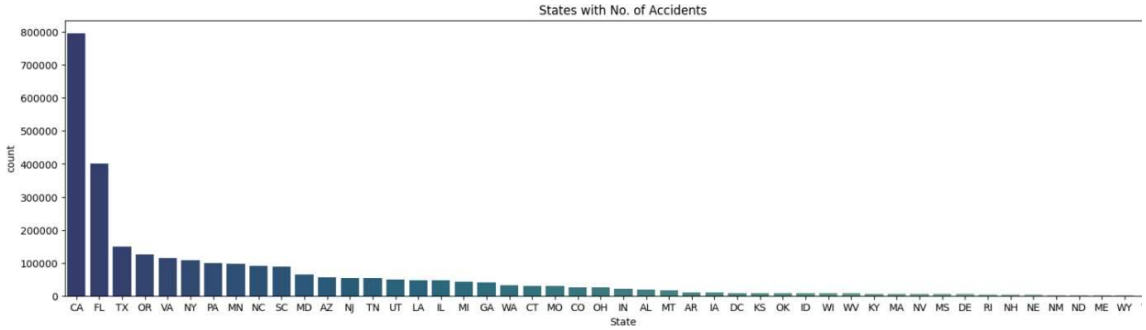
- The dataset contains 45 features that can be used to predict an accident severity and location.
- It contains 31 categorical features and 14 continuous features.
- The description of the 918 rows and 12 columns is as follows

| | Severity | Start_Lat | Start_Lng | End_Lat | End_Lng | Distance(mi) | Number | Temperature(F) | Wind_Chill(F) | Humidity(%) | Pressure(in) | Visibility(mi) | Wind_Speed(mph) | Precipitation(in) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 2.845342e+06 | 2.845342e+06 | 2.845342e+06 | 2.845342e+06 | 2.845342e+06 | 2.845342e+06 | 1.101431e+06 | 2.776068e+06 | 2.375699e+06 | 2.772250e+06 | 2.786142e+06 | 2.774796e+06 | 2.687398e+06 | 2.295884e+06 |
| mean | 2.137572e+00 | 3.624520e+01 | -9.711463e+01 | 3.624532e+01 | -9.711439e+01 | 7.026779e-01 | 8.089408e+03 | 6.179356e+01 | 5.965823e+01 | 6.436545e+01 | 2.947234e+01 | 9.099391e+00 | 7.395044e+00 | 7.016940e-03 |
| std | 4.787216e-01 | 5.363797e+00 | 1.831782e+01 | 5.363873e+00 | 1.831763e+01 | 1.560361e+00 | 1.836009e+04 | 1.862263e+01 | 2.116097e+01 | 2.287457e+01 | 1.045286e+00 | 2.717546e+00 | 5.527454e+00 | 9.348831e-02 |
| min | 1.000000e+00 | 2.456603e+01 | -1.245481e+02 | 2.456601e+01 | -1.245457e+02 | 0.000000e+00 | 0.000000e+00 | -8.900000e+01 | -8.900000e+01 | 1.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| 25% | 2.000000e+00 | 3.344517e+01 | -1.180331e+02 | 3.344628e+01 | -1.180333e+02 | 5.200000e-02 | 1.270000e+03 | 5.000000e+01 | 4.600000e+01 | 4.800000e+01 | 2.931000e+01 | 1.000000e+01 | 3.500000e+00 | 0.000000e+00 |
| 50% | 2.000000e+00 | 3.609861e+01 | -9.241808e+01 | 3.609799e+01 | -9.241772e+01 | 2.440000e-01 | 4.007000e+03 | 6.400000e+01 | 6.300000e+01 | 6.700000e+01 | 2.982000e+01 | 1.000000e+01 | 7.000000e+00 | 0.000000e+00 |
| 75% | 2.000000e+00 | 4.016024e+01 | -8.037243e+01 | 4.016105e+01 | -8.037338e+01 | 7.640000e-01 | 9.567000e+03 | 7.600000e+01 | 7.600000e+01 | 8.300000e+01 | 3.001000e+01 | 1.000000e+01 | 1.000000e+01 | 0.000000e+00 |
| max | 4.000000e+00 | 4.900058e+01 | -6.711317e+01 | 4.907500e+01 | -6.710924e+01 | 1.551860e+02 | 9.999997e+06 | 1.960000e+02 | 1.960000e+02 | 1.000000e+02 | 5.890000e+01 | 1.400000e+02 | 1.087000e+03 | 2.400000e+01 |


| | Severity | Start_Lat | Start_Lng | End_Lat | End_Lng | Distance(mi) | Number | Temperature(F) | Wind_Chill(F) | Humidity(%) | Pressure(in) | Visibility(mi) | Wind_Speed(mph) | Precipitation(in) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 2.845342e+06 | 2.845342e+06 | 2.845342e+06 | 2.845342e+06 | 2.845342e+06 | 2.845342e+06 | 1.101431e+06 | 2.776068e+06 | 2.375699e+06 | 2.772250e+06 | 2.786142e+06 | 2.774796e+06 | 2.687398e+06 | 2.295884e+06 |
| mean | 2.137572e+00 | 3.624520e+01 | -9.711463e+01 | 3.624532e+01 | -9.711439e+01 | 7.026779e-01 | 8.089408e+03 | 6.179356e+01 | 5.965823e+01 | 6.436545e+01 | 2.947234e+01 | 9.099391e+00 | 7.395044e+00 | 7.016940e-03 |
| std | 4.787216e-01 | 5.363797e+00 | 1.831782e+01 | 5.363873e+00 | 1.831763e+01 | 1.560361e+00 | 1.836009e+04 | 1.862263e+01 | 2.116097e+01 | 2.287457e+01 | 1.045286e+00 | 2.717546e+00 | 5.527454e+00 | 9.348831e-02 |
| min | 1.000000e+00 | 2.456603e+01 | -1.245481e+02 | 2.456601e+01 | -1.245457e+02 | 0.000000e+00 | 0.000000e+00 | -8.900000e+01 | -8.900000e+01 | 1.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| 25% | 2.000000e+00 | 3.344517e+01 | -1.180331e+02 | 3.344628e+01 | -1.180333e+02 | 5.200000e-02 | 1.270000e+03 | 5.000000e+01 | 4.600000e+01 | 4.800000e+01 | 2.931000e+01 | 1.000000e+01 | 3.500000e+00 | 0.000000e+00 |
| 50% | 2.000000e+00 | 3.609861e+01 | -9.241808e+01 | 3.609799e+01 | -9.241772e+01 | 2.440000e-01 | 4.007000e+03 | 6.400000e+01 | 6.300000e+01 | 6.700000e+01 | 2.982000e+01 | 1.000000e+01 | 7.000000e+00 | 0.000000e+00 |
| 75% | 2.000000e+00 | 4.016024e+01 | -8.037243e+01 | 4.016105e+01 | -8.037338e+01 | 7.640000e-01 | 9.567000e+03 | 7.600000e+01 | 7.600000e+01 | 8.300000e+01 | 3.001000e+01 | 1.000000e+01 | 1.000000e+01 | 0.000000e+00 |
| max | 4.000000e+00 | 4.900058e+01 | -6.711317e+01 | 4.907500e+01 | -6.710924e+01 | 1.551860e+02 | 9.999997e+06 | 1.960000e+02 | 1.960000e+02 | 1.000000e+02 | 5.890000e+01 | 1.400000e+02 | 1.087000e+03 | 2.400000e+01 |

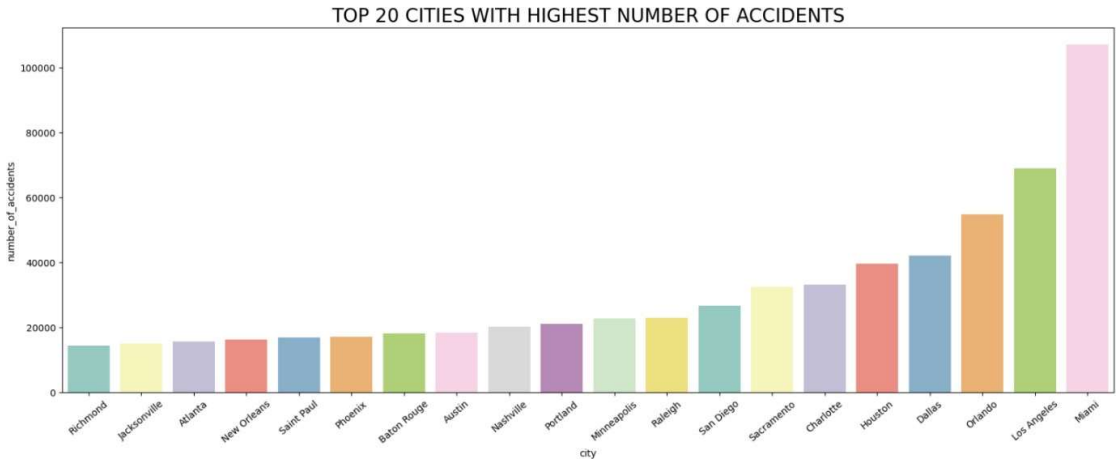

# Exploratory Data Analysis

## 1. States Analysis

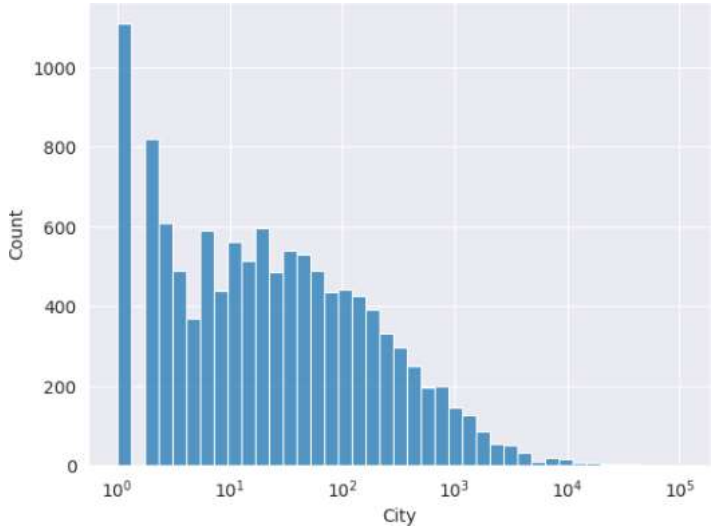



California (CA) is the 3rd most largest state of US after Texas (TX) and Alaska (AL).
Also California (CA) is the most populated among all, followed by Texas (TX)

## 2. States Analysis



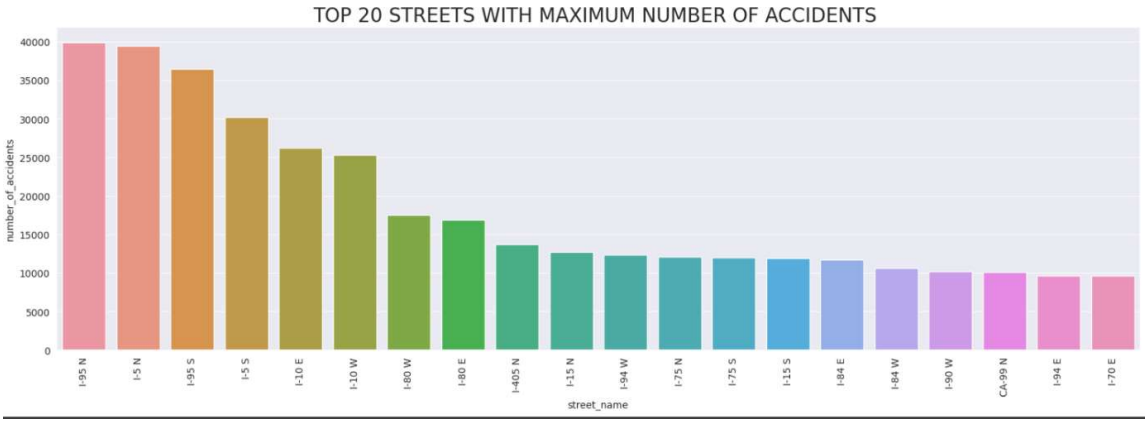
There is quite a lot of presence of cities from Florida (Miami, Orlando) followed by California (LA, Sacramento, San Diego, Jacksonville), Texas (Houston, Dallas, Austin) , and North Carolina (Charlotte, Raleigh). This is in tandem with the top states: Florida, California, Texas, and North Carolina.




Over 1200 cities have reported only one accident during the entire period. This could either be very good news or it could be a result of missing data. Let's break cities by accidents into two groups and see their respective distributions-

- High accident cities where the number of accidents is greater and equal to 1000.
- Low accident cities where the number is less than 1000.

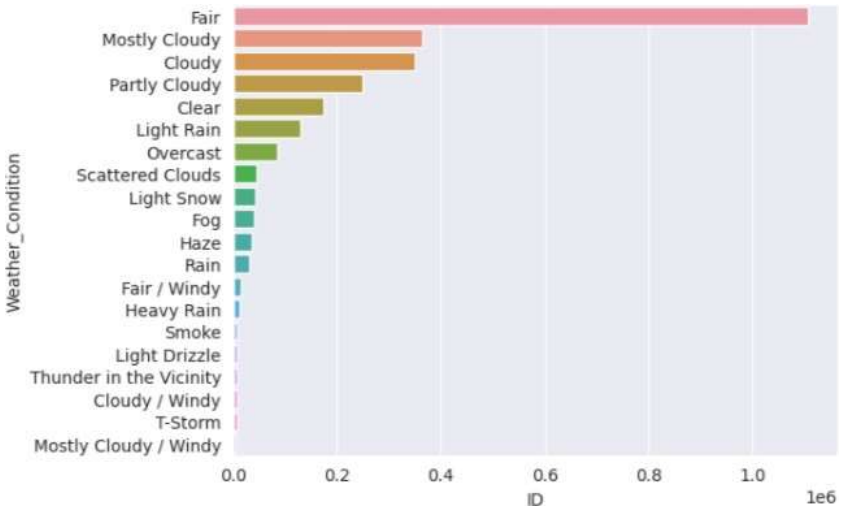
## 3. Street Analysis



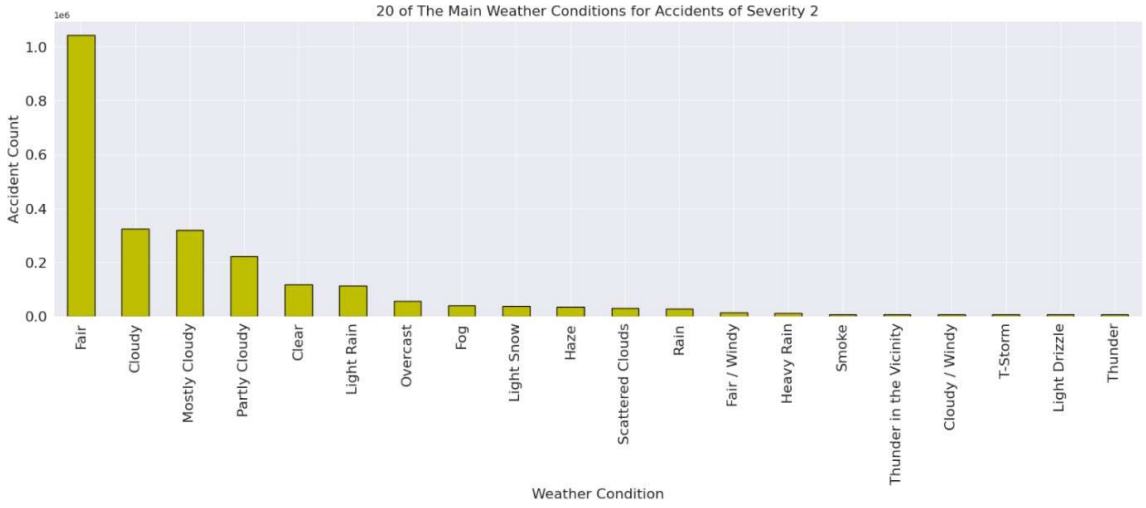

- In last 5 years (2016-2020) Street No. I-5 N is having the highest road accidents records.
- In Street No. I-5 N, daily 14 accidents occurred in average.

- There are 36,441 Streets (39%) in US which have only 1 accident record in past 5 years.
- 98% Streets of US, have less than 100 road accident cases.

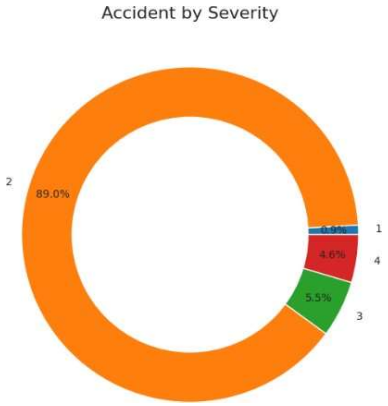
## 4. Weather Analysis




- From graph and top 20 dataframe, most of the accidents happened in Fair weather condition.
- It is very surprising that most of the accidents happened in Fair weather.
- Weather analysis for severity 2.

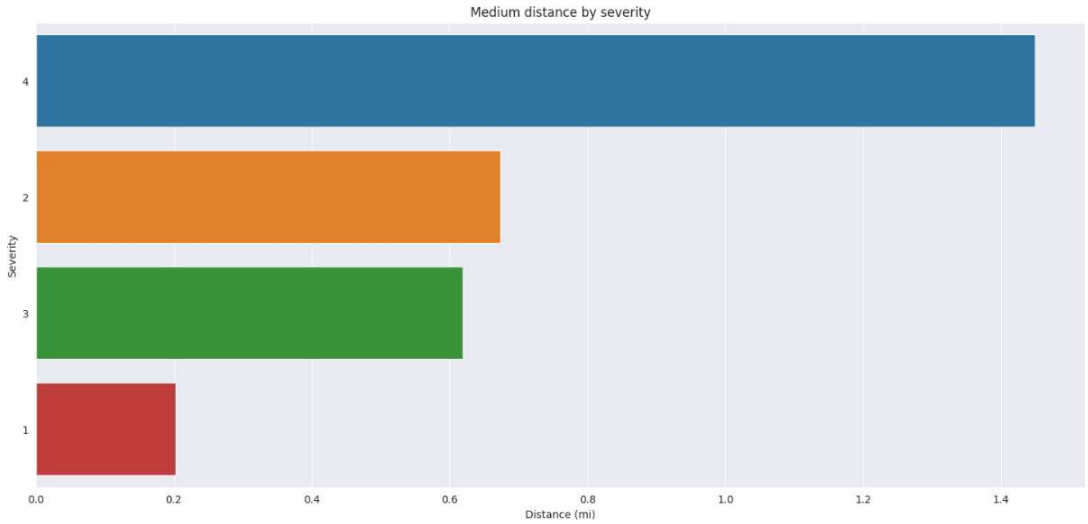



## 5. Severity Analysis



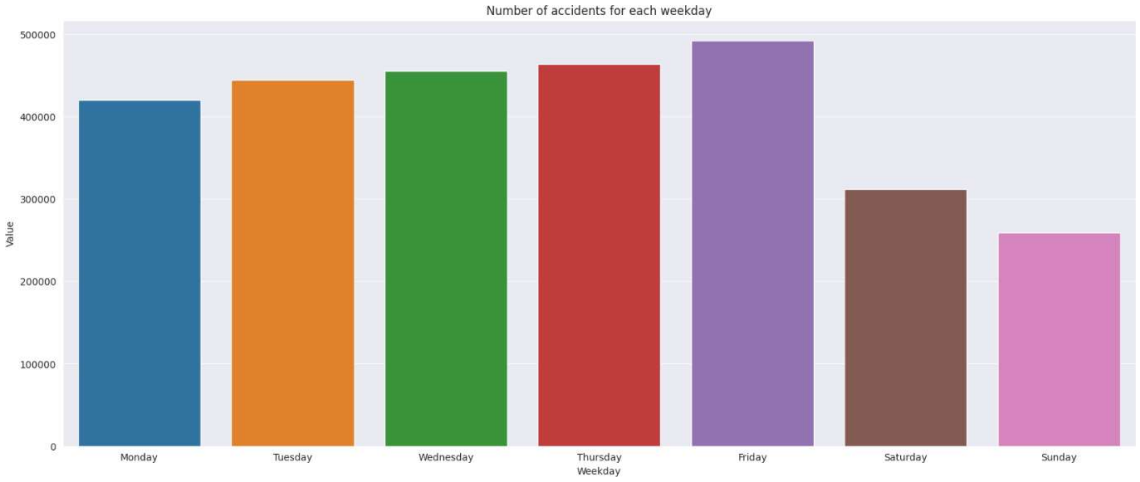

4.6% of the accidents recorded were found to be very severe.

## 6. Distance Analysis



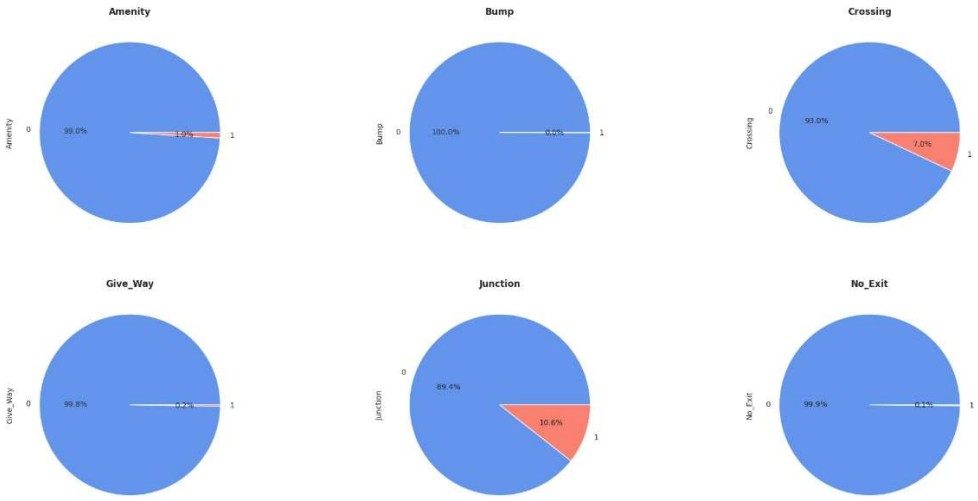

In this graph we can see that the distance of the accident is more or less proportional to the severity, and in fact accidents with severity 4 have the longest distance.

## 7. Start Analysis




As we can see from the plot above, the days with the most accidents are working days, while in the weekend we have a frequency of at least 2/3 less. This may be due to the fact that during the weekend there are fewer cars on the road.

## 8. Road Condition Analysis




## 9. Summary of data Analysis

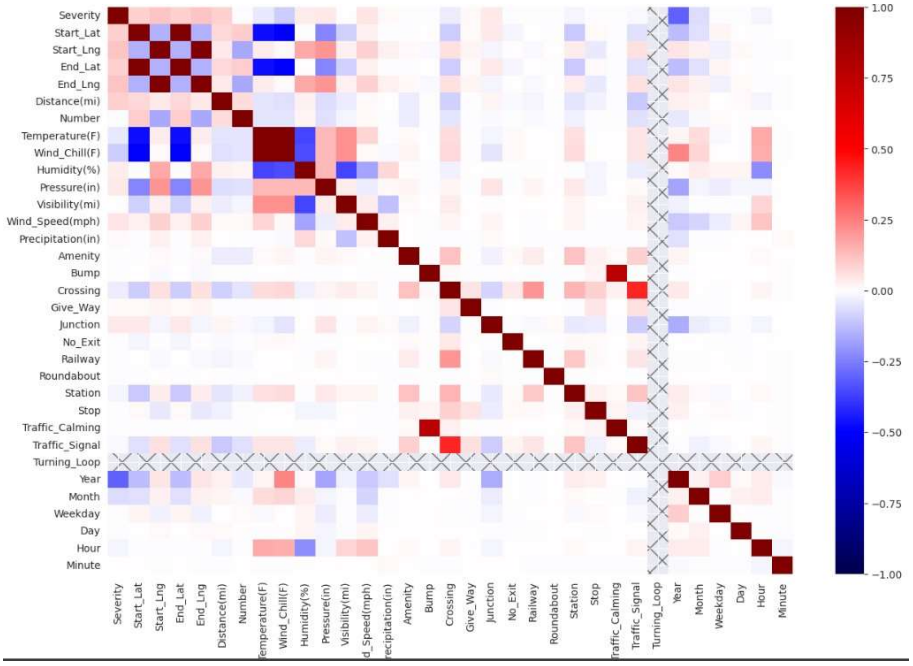
- The top 5 states by accidents include populous ones like California, Florida, Texas, Oregon, Virginia.
- Less than 5% of cities have more than 1000 accidents in the period between February 2016 and December 2021.
- The majority of them have witnessed between 10–100 accidents during the period.
- Accidents by cities follow an exponentially decreasing distribution.
- The hypothesis that weekdays see more accidents during morning and evening rush hours is corroborated by the data. On working days of week most of the accidents

happened from 7am to 9am, may be because of many people goes to office at this time. Also, number of accidents are more from 4pm to 6pm, may be because of it is time of returning from office.

- But on weekend days, distribution of number of accidents is pretty much different from working days of week. (12 pm - 6 pm).
- All the working week days had approx same number of accidents.
- Most of the accidents happened in Fair weather condition
- Upto 80% humidity, number of accidents increased (approx) uniformly with increase in humidity. Above 80% humidity there is no such relation between number of accidents & humidity.
- Number of accidents increased from 2016 to 2021 & Year 2021 had most accidents.
- Majority of accidents have severity ~2 means not much higher impact on traffic.
- Most-deadliest accident hour is 5:00PM
- Maximum no of cases occurred between temperature range: 50-80 F.
- Wind speed is not the reason for accidents.
- Weather condition was Fair in most of the cases hence it is not a major cause behind the accidents.
- Weather Conditions like 'Fair' and 'Cloudy' has most occurring cases of Severity type 2.

# Data Preprocessing

## 1. Find the Correlation Matrix




- From the matrix we can see that the start and end GPS coordinates of the accidents are highly correlated.

- In fact, from the medium distance shown before, the end of the accident is usually close to the start, so we can consider just one of them for the machine learning models.
- Moreover, the wind chill (temperature) is directly proportional to the temperature, so we can also drop one of them.
- We can also see that the presence of a traffic signal is slightly correlated to the severity of an accident meaning that maybe traffic lights can help the traffic flow when an accident occurs.
- From the matrix we can also note that we couldn't compute the covariance with Turning_sLoop, and that's because it's always False.

## 2. Feature Selection

From the observations made with the correlation matrix, we are going to drop the following features:
End_Lat and End_Lng, Wind Chill

- Moreover, we are going to drop the following features:
- ID, Source: since they don't carry any information for the severity
- TMC: because it could already contains information about the accident severity
- Start_Time: because it was decomposed by the time features added before (day, month, weekday)
- End_Time: beause we cannot know in advance when the traffic flow will become regular again
- Description: most description only report the name of the road of the accident, and so we decided to omit this feature for semplicity
- Number, Street, County, State, Zipcode, Country: because we just focus on the City where the accident happened
- Timezone, Airport_Code, Weather_Timestamp: because they are not useful for our task
- Turning_Loop: since it's always False
- Sunrise_Sunset, Nautical_Twilight, Astronomical_Twilight: because they are redundant

## 3. Drop the duplicates value

## 4. Handle erroneous and missing values

If we analyze the weather conditions, we can see that there are lots of them, so it's better to reduce the number of unique conditions.

## 5. Outlier Treatment

Remove the outliers from the following features:

Wind_Speed,Distance,Temperature,Pressure,Visibility

## 6. Combining Weather conditions

If we analyze the weather conditions, we can see that there are lots of them, so it's better to reduce the number of unique conditions.

So combining 126 weather conditions into these weather conditions,
['Rain' 'Cloudy' 'Snow' nan 'Clear' 'Fog' 'Thunderstorm' 'Smoke' 'Windy' 'Hail' 'Sand' 'Tornado']

Also combine the 25 different wind directions into 12 wind directions.

## 7. Feature scaling

To improve the performance of our models, we normalized the values of the these continuous features

[Temperature(F),Distance(mi),Humidity(%),Pressure(in),Visibility,Wind_Speed, Precipitation(in),Start_Lng,Start_Lat,Year,Month,Weekday,Day,Hour,Minute]

## 8. Feature encoding

Finally, in this section we are going to encode these categorical features.
[Side, City, Wind_Direction, Weather_Condition, Civil_Twilight]

# Result and analysis for severity prediction

| | Decision Tree | KNN | Logistic Regression | SVM | Random Forest | XGBoost | AdaBoost |
|---|---|---|---|---|---|---|---|
| **Accuracy** | 0.8875 | 0.8873 | 0.8860 | 0.8872 | 0.8874 | 0.9056 | 0.7814 |
| **F1 score** | 0.8346 | 0.8873 | 0.8457 | 0.8348 | 0.88745 | 0.8861 | 0.8153 |
| **MSE** | 0.2602 | 0.2598 | 0.8457 | 0.2616 | 0.2604 | 0.2119 | 0.5067 |

From the above observation table we can see that we got the **Highest Accuracy** value 0.9056 for the **XGBoost** model and the **Lowest Accuracy** value for **AdaBoost** model.

- F1 score is Highest for **Random Forest** which is 0.8874.
- MSE is Lowest for **XGBoost** model which is 0.2219

So we can say that **XGBoost** is the best model for severity prediction of the given dataset.

## Result and analysis for location(start latitude) prediction

| | Decision Tree | KNN | Logistic Regression | SVM | Random Forest | XGBoost | AdaBoost |
|---|---|---|---|---|---|---|---|
| Accuracy | 0.7399 | 0.7824 | 0.7440 | 0.7985 | 0.7127 | 0.8629 | 0.7374 |
| F1 score | 0.7399 | 0.7824 | 0.7440 | 0.7985 | 0.7127 | 0.8629 | 0.7272 |
| MSE | 0.2601 | 0.3067 | 0.2839 | 0.2160 | 0.2905 | 0.1427 | 0.2783 |

From the above observation table we can see that we got the **Highest Accuracy** value 0.8629 for the **XGBoost** model and the **Lowest Accuracy** value for **Random Forest** model.

- F1 score is Highest for **XGBoost** which is 08629.
- MSE is Lowest for **XGBoost** model which is 0.1427

So we can say that **XGBoost** is the best model for start latitude prediction of the given dataset.

## Result and analysis for location(start longitude) prediction

| | Decision Tree | KNN | Logistic Regression | SVM | Random Forest | XGBoost | AdaBoost |
|---|---|---|---|---|---|---|---|
| Accuracy | 0.6996 | 0.7732 | 0.6837 | 0.8002 | 0.6911 | 0.8506 | 0.6634 |
| F1 score | 0.6996 | 0.7732 | 0.6837 | 0.8002 | 0.6911 | 0.8629 | 0.6593 |
| MSE | 0.9271 | 0.6664 | 1.0193 | 0.5842 | 0.9708 | 1.3637 | 1.0809 |

From the above observation table we can see that we got the **Highest Accuracy** value 0.8506 for the **XGBoost** model and the **Lowest Accuracy** value for **AdaBoost** model.

- F1 score is Highest for **XGBoost** which is 0.8629.
- MSE is Lowest for **SVM** model which is 0.5842

So we can say that **XGBoost** is the best model for start longitude prediction of the given dataset.

# Conclusion

In conclusion, the analysis of the US accident dataset provides insights into the causes, frequency, and severity of accidents that occur on roads. The analysis shows that most accidents are caused by human error, with factors such as speeding, distracted driving, and drunk driving being the primary culprits.

Moreover, the analysis also indicates that the severity of accidents varies based on different factors such as weather conditions, road surface, and location. Therefore, the prediction of accident severity and location is crucial in improving road safety and reducing the number of accidents on roads.

The machine learning models, including Logistic Regression, Decision Trees, XGBoost and Random Forests, can accurately predict accident severity and location based on various features such as weather condition, visibility, time, and location. The models can help in identifying high-risk areas and taking preventive measures to reduce the occurrence of accidents.

In conclusion, the analysis of the US accident dataset provides valuable insights into the causes, frequency, and severity of accidents on roads. The prediction of accident severity and location can be a useful tool for improving road safety, reducing the number of accidents, and saving lives.