# DATA SCIENCE 1: GROUP ASSIGNMENT


Group 11

An analysis of car accident data in the U.S. from 2016-2019

**Prepared by:**

Leo Li, Hojae Kim, Mohit Ramnani, Flora Wan

# Table of Contents

# Objectives

## Overview of Topic

According to the Association for Safe International Road Travel (ASIRT), road crashes are the leading cause of death in the U.S. for people aged 1-54. More than 38,000 people die every year in crashes on U.S. roadways and an additional 4.4 million are injured seriously enough to require medical attention.

Using data science principles and a robust dataset, an analysis of road accident data can be used to provide recommendations on how to improve traffic safety. For example, we can study hotspot locations where accidents most frequently occur, conduct casual analysis to predict which factors have the greatest effect on the rate and severity of car accidents, and look at the impact of weather and other environmental factors on accident occurrence.

## Description of Data Source

The dataset used in this analysis is posted on the online data science community Kaggle. It consists of car accident data in 49 states of the United States collected from February 2016 to December 2019. The data is sourced using several data providers, including two APIs that provide streaming traffic incident data. These APIs broadcast traffic data captured by a variety of sources, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. There are close to 3 million accident records in this dataset, making it a robust data source to conduct our analysis.

For each accident record, this dataset includes 49 attributes that can be roughly split into 6 categories. Examples of attributes within each category are provided below:

| **Category** | **Attributes** |
|---|---|
| Administration | Record ID, Source of the accident report, Traffic Message Channel (TMC) |
| Accident Description | Accident severity, Start time, End time, Natural language description |
| Location | Latitude, Longitude, Distance between start and end points, Address (Street, City, State, Zip code), Side of the street (Right/Left), Time zone, Airport code |
| Weather | Temperature, Wind Chill, Humidity, Pressure, Visibility, Wind Direction, Wind Speed, Precipitation, Weather Condition |
| Road design | Bump, Crossing, Give way sign, Junction, No exit, Railway, Roundabout, Station, Stop sign, Traffic calming features, Traffic signal, Turning loop |
| Time of day | Sunrise, Sunset, Civil Twilight, Nautical Twilight, Astronomical Twilight |

The following papers are cited as requested by the provider of the dataset:

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", 2019.
- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

## Questions and Hypotheses

The objective of this study is to conduct data analysis and report key statistical features and insights based on the recorded incidents of traffic accidents in the U.S. between 2016-2019. The goal is to come up with recommendations on how to improve road safety by exploring the predictive and descriptive factors that affect accident severity.

Some of the questions we want to explore are:

- What key factors or attributes correlate with the severity of accidents?
- How do various types of weather conditions affect accident severity?
- Which regions of the United States are more prone to car accidents? Is population density a factor in the rate and severity of accidents?
- What recommendations can be provided on improving the design of road systems to reduce the likelihood and severity of accidents?
- Are there any patterns in seasons that would predict a higher than average number of accidents?

The following are some of our initial hypotheses:

- Weather conditions will have a large effect on accident severity. For instance, accidents that occur during days with less visibility or more precipitation will be more severe than those occurring on a clear, sunny day.
- Absence of certain road and traffic safety design features such as traffic signals, all-way stops and roundabouts may contribute to a greater number of accidents.
- Cities with a denser population may have a greater percentage of high severity accidents.
- In the colder winter months (November to February), the number of accidents and severity of these accident will be higher due to snowy and icy conditions across much of the country.

# Data Preparation

## Data Cleaning

In preparation for using the data in the analysis, we explored all 49 attributes that were included in the dataset. The following table lists some examples of the type of data that was cleaned/transformed and the rationale behind the decision on how to transform the data. The full set of changes can be found in the script *data_cleaning.ipynb* in the attached supporting documents.

| **Type of change** | **Original** | **Transformed** | **Rationale** |
|---|---|---|---|
| Simplify ID field | A-1423<br>(type: string) | 1423<br>(type: int) | In the source dataset, the "ID" field is structured as "A-<number>" (ex. A-305 for the 305th record). This is a cumbersome method of naming records. Therefore, the "ID" column was transformed by removing "A-" from the string and then changing the string object into an integer to save space and enable the user to more easily locate records in the analysis. |
| Replace label values with numbers | Side = L or R | Side = 0 (Left) or 1 (Right) | In this example, the "Side" attribute which indicates which side of the road an accident occurred, is recorded as "L" (left) or "R" (right). Although this makes sense as a data label, having letters as attribute values makes it difficult to perform analysis and use model algorithms. Therefore, these types of labels are transformed into numbers so that the overall dataset can be more easily analyzed. |
| Create new attribute for "Incident Time" | None | New column = Incident_time | In this dataset, the severity of an accident is defined as the amount of impact it had on traffic flow. However, there is no attribute indicating the total amount of time from the start to end of an incident. Therefore, a new attribute was created to calculate the total accident time (in minutes) for each incident based on the recorded start and end times. |
| Categorize time zones | US/Western<br>US/Mountain<br>US/Central<br>US/Eastern | 1 = Pacific Time<br>2 = Mountain Time<br>3 = Central Time<br>4 = Eastern Time | Tens of thousands of records have missing values in the "timezone" attribute, but those records have longitude and latitude data based on the location of the accident. We leveraged the "timezonefinder" package to identify the missing time zones based on the corresponding longitude and latitude. Later on, we noticed that the attribute "timezone" uses a different time zone naming convention than the "timezonefinder" package, therefore, we transformed the original labels into numerical values from west to east in order to more easily analyze the data. |
| Impute missing values for weather | Missing value | Imputed values for weather-related attributes (ex. | We noticed some missing values related to weather attributes, such as "temperature", "wind_chill", "humidity" (%), "pressure" (in), visibility" (mi), "wind_speed" (mph) and "precipitation" (in). To fill in these missing values, we have |

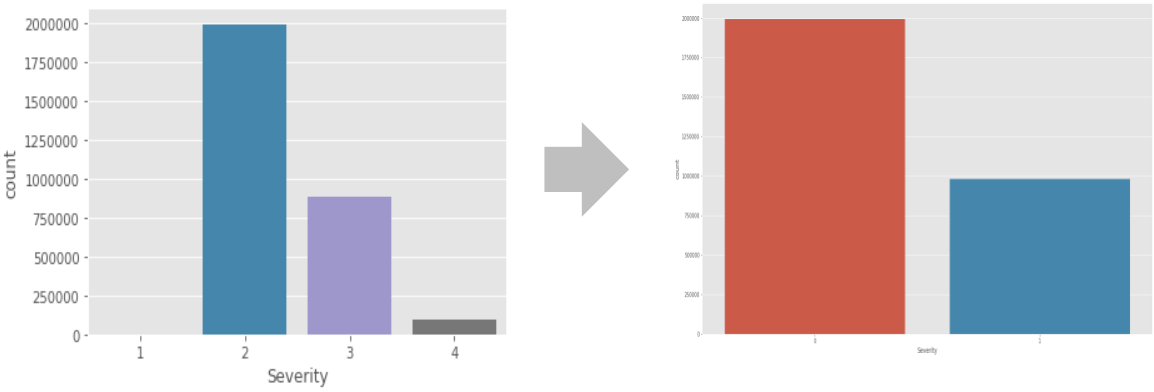
| | | | |
|---|---|---|---|
| attributes | | temperature) | imputed it with the median value of the attribute during that week, as weather patterns should be similar within the same week in the same State. The particular week in question can be determined by the using the start time of the recorded incident. |
| Replace True/False labels with numbers | Traffic_Signal = True<br><br>(type: Boolean) | Traffic_Signal = 1<br><br>(type: int) | All road design attributes (ex. "Traffic_Signal", "Bump", "Crossing", etc.) are labelled as True or False in the dataset depending on whether that particular road feature is present at the location of the accident. In order to perform analysis and use model algorithms, these True or False values were transformed into 0=False and 1=True. |
| Remove some irrelevant attributes | Country, number, street, airport code | Columns dropped | After reviewing the attributes available in the full dataset, some attributes were deemed to be irrelevant to the overall analysis of the data. For example, since the dataset is from the U.S., the value for "country" will always be "US". The street name and number of the exact address is also unimportant as our analysis will be focusing on general regions. Therefore, these types of attributes were dropped from the dataset. |

## Defining "Severity"

Since much of our analysis focuses on factors that affect the severity of accidents, it is important to first understand and agree on how severity is to be defined. In the description of the dataset, severity is defined as follows:

***Severity:*** *Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).*

A count of the number of records for each level of severity shows that the majority of accidents are rated at severity level 2 (moderate delay as a result of the accident) and the remaining accidents are rated at severity levels 3 and 4 (significant impact on traffic).




Therefore, for the purpose of the analysis, severity levels 1 and 2 will be grouped together as "low severity" = 0 and severity levels 3 and 4 will be grouped together as "high severity" = 1.

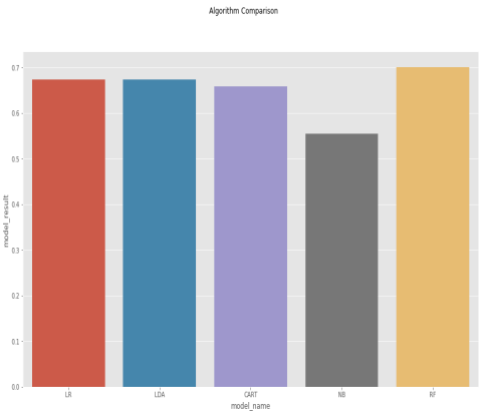
# Analysis

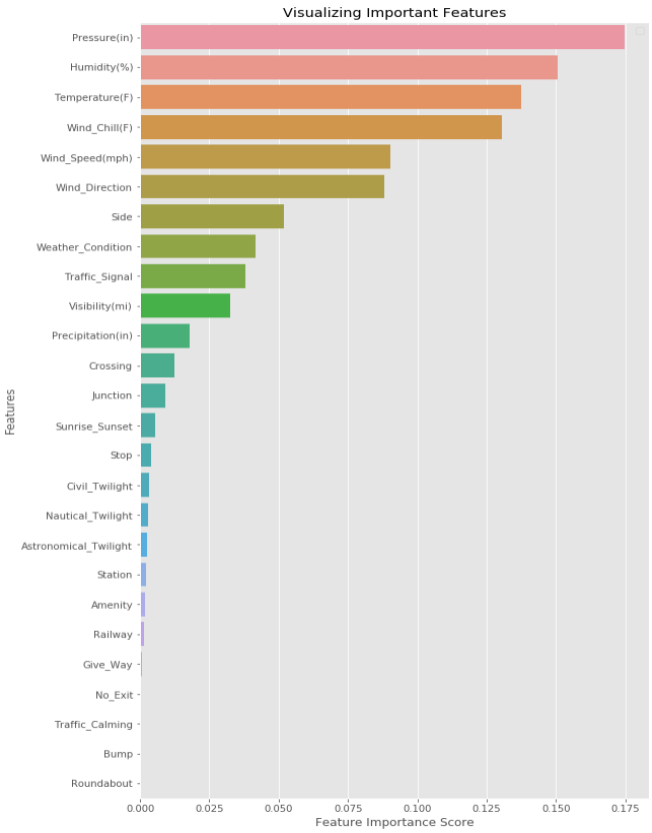
The analysis section discusses the trends, correlations and patterns we uncovered as we explored the dataset. We have separated the analysis results into four sections:

1. **Data Exploration**: Initial exploration of the data to determine feature importance.
2. **Predictive Insights**: Analysis focused on the factors that may predict the likelihood and severity of accidents, such as weather conditions and road design.
3. **Descriptive Insights**: Analysis focused on patterns related to the frequency and severity of accidents based on geographic location.
4. **Time Series Analysis**: Exploring the trends based on accident rates as it relates to seasonality.

## Data Exploration

The first step of our analysis focuses on determining which of the attributes provided in the dataset most correlate to accident severity. Using a random split of train and test cases, five different predictive models were compared (*Logistic Regression, Linear Discriminant Analysis, Decision Tree Classifier, Gaussian Naïve Bayes, Random Forest Classifier*) to determine which algorithm provides the most accurate prediction of severity based on the attributes in the dataset. Results of the comparison shows that the Random Forest Classifier model (represented by the yellow bar in the graph on the right) was the most accurate prediction model, hence it was used to generate a bar graph showing the relative importance of features in determining accident severity.




Feature importance is a method of data exploration that outputs a score for each feature or attribute of the data. The higher the score, the more important or relevant is the feature towards the output variable.

In this case, we are exploring the predictive variables that contribute to the level of accident severity. The top 10 variables or features that most affect accident severity are:

1. Pressure
2. Humidity
3. Temperature
4. Wind Chill
5. Wind Speed
6. Wind Direction
7. Side of the Road (Right/Left)
8. Weather Condition
9. Presence of Traffic Signal
10. Visibility

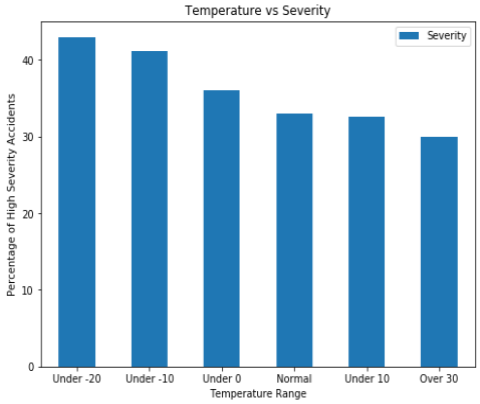
These factors will be explored in greater detail in the next section of this report.

# Predictive Insights

The analysis in this section focuses on the impact of weather patterns and road/traffic design on accident rate and severity.

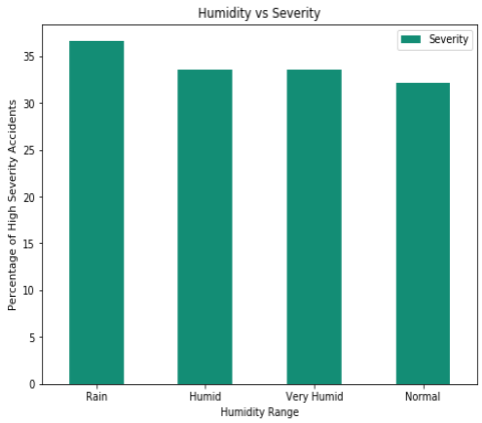
## Weather Patterns

As there are many features related to weather patterns, we have decided to focus our analysis on the following five specific attributes: Temperature, Humidity, Precipitation, Visibility and Wind Speed. We will first explore each attribute on its own and then combine several attributes to show the aggregate impact on accident severity.



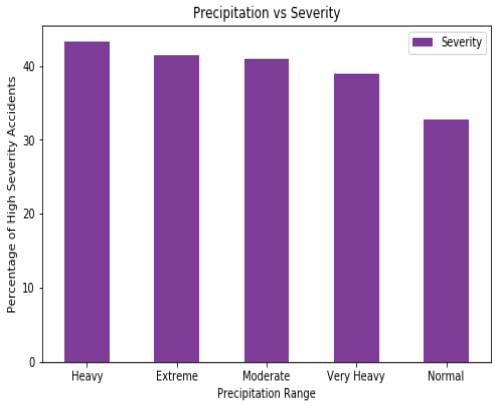

**Temperature:** In order to analyze the effect of temperature on accident severity, the range of temperatures were divided into 6 categories, by degrees in Celsius: 'Under -20', 'Under -10', 'Under 0', 'Under 10', 'Normal', 'Over 30'.

As expected, as temperature gets below 0, severity of accident increases. Surprisingly, during normal temperature (10~30 degree Celsius), accident severity is still quite high. A possible explanation is that people tend to drive faster during normal temperature and as such, result in more severe accidents.



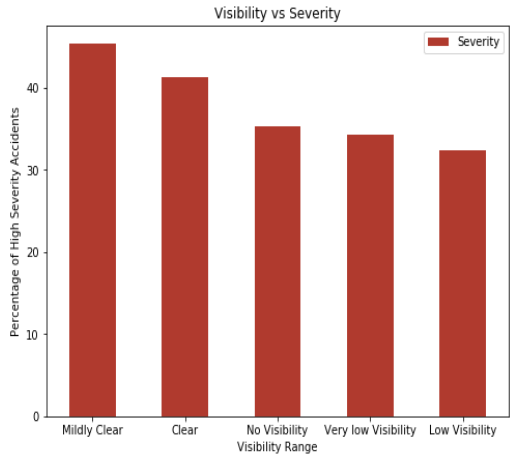

**Humidity:** Similar to the analysis of temperature, humidity is also a numerical scale ranging from 0-100 (when humidity is close to 100%, it is most likely to rain). Therefore, the range of humidity percentages were divided into 4 categories: 'Normal' (0-64%), 'Humid' (65-74%), 'Very Humid' (75-99%), 'Rain' (100%)

Even though the number of data points is smaller for 'Rain' compared to the other categories, it appears that when it rains, the relative severity of the accident also increases. However, as humidity is also dependent on the location (high humidity if close to shore), it is difficult to conclusively say whether humidity is correlated to severity.




**Precipitation:** Precipitation indicates inches of rain/snow that are falling per hour. According to "Weather Shack", 0.1~0.3 inches of rainfall is generally considered moderate. If the rainfall amount is more than 0.3 inches per hour, it is considered heavy. Using this information, precipitation was divided into 5 categories: 'Normal' (less than 0.1 inches/hr), 'Moderate' (0.1-0.2 inches/hr), 'Heavy' (0.3-0.9 inches/hr), 'Very Heavy' (1-4 inches/hr), 'Extreme' (5 inches/hr or above).

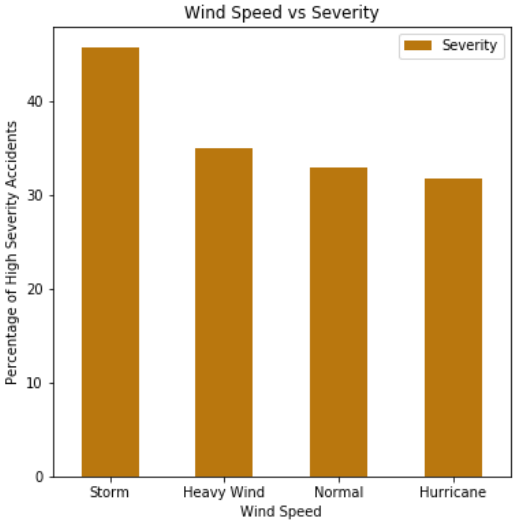
Even though the number of data points is small for precipitation that is not normal, it appears that as precipitation increase, so does the percentage of more severe accidents.




**Visibility:** According to an article in Time magazine, visibility is one of the most important factors that affect severity when accident happens.

On a clear day, visibility is about 18.6 miles. Using this information, the visibility range was divided into 5 categories: 'No Visibility' (less than 4 miles),'Very Low Visibility' (4-7 miles),'Low Visibility' (8-11 miles), 'Mildly Clear' (12-17 miles),'Clear' (18 miles and above).

Surprisingly, although a smaller number of accidents occur on days with clear visibility, there is actually a greater percentage of higher severity accidents. A possible explanation is that on mildly clear or clear days, people might tend to drive faster, resulting in less frequent but more severe accidents.

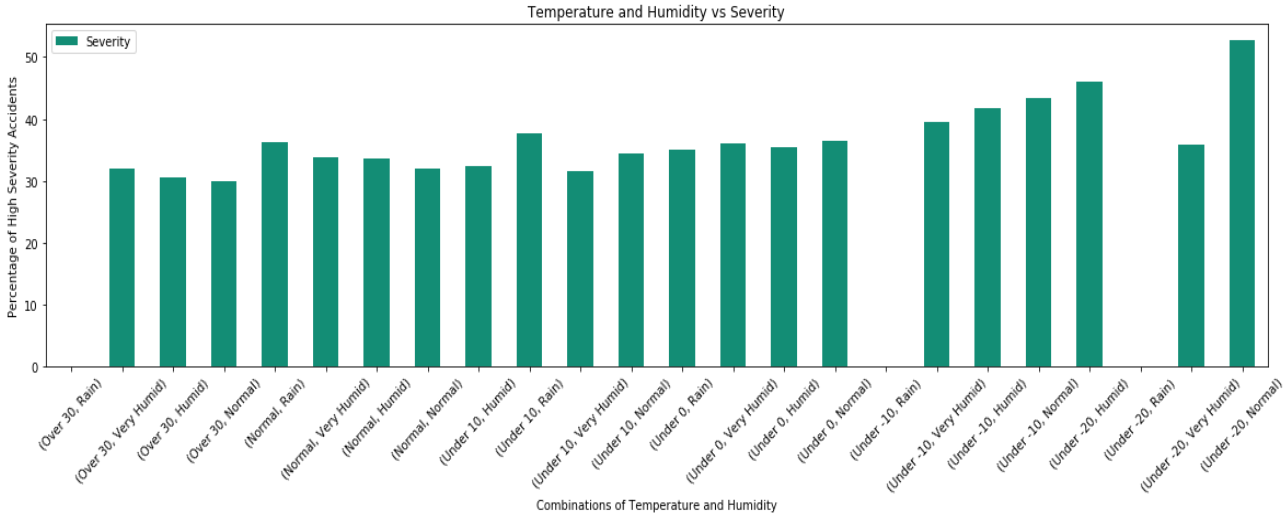



**Wind Speed:** The last attribute related to weather patterns we will explore is wind speed, as it will cause disruptions to car movement when the wind speed is high. According to The Northeastern Regional Association of Coastal Ocean Observing Systems, when wind speed exceeds 24mph, it is considered strong wind. Based on this information, the range of wind speed was divided into 4 categories: 'Normal' (less than 24mph), 'Heavy Wind' (24-45mph), 'Storm' (46-72mph), 'Hurricane'. (73mph and above).

Although most accidents happen during normal wind speed, results of the analysis show that during a storm or at heavy winds, the highest percentage of severe accidents occur. Unexpectedly, the severity rate at hurricane wind speeds is actually lower than all the other categories. We speculate that it may be because there are less drivers on the road aside from emergency vehicles when the weather is at such an extreme level, but that is certainly a surprising result.

## Combining temperature and humidity

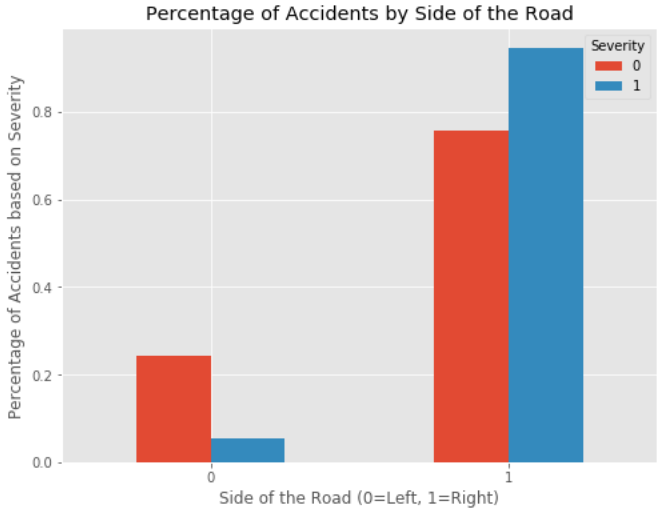
To further explore the effects of weather patterns on the severity of accidents, we will combine two of the weather-related attributes to see if it impacts the analysis. The following graph combines **temperature** and **humidity** and plots it against the percentage of high severity accidents:



The graph above shows that accident severity for the combination of temperature lower than -20 degrees Celsius with humidity at a 'Normal' level has the greatest percentage of high severity accidents. We speculate that this could be because the temperature in this scenario is very cold at under -20 degree Celsius. At this temperature, ice can form even at 30% humidity, so a 'Normal' humidity level of up to 65% would likely cause ice formation but appear weather-wise to be less dangerous as it is not raining or snowing. Therefore, unsuspecting drivers may not be as cautious when driving and accidentally hit ice patches at high speeds, causing greater rates of severe accidents.

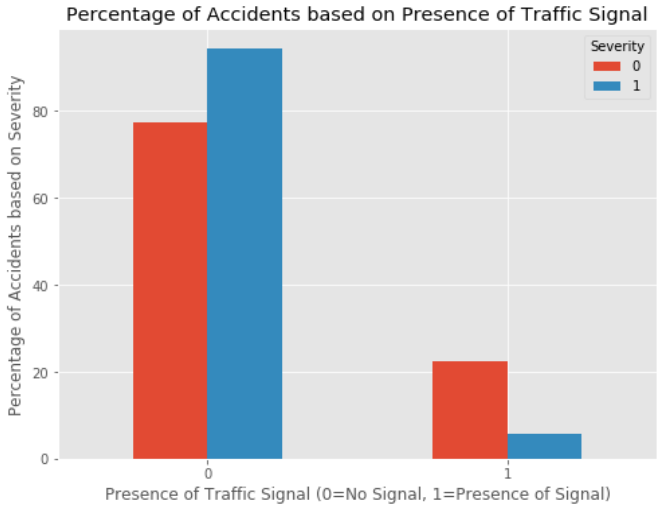
## Road and Traffic Design

Finally, we explore the two attributes related to road and traffic design that are rated highly on the feature importance graph: 1) Side of the road (Right/Left), and 2) Presence of a traffic signal. For these two attributes, both the low severity (0) and high severity (1) accidents are plotted to analyze the overall correlation of road design features and number of accidents.




**Side of the Road (Left/Right):**

The graph on the left shows that regardless of severity, more accidents occur on the right side of the road (0=Left, 1=Right). This makes sense as North American vehicles drive on the right side of the road, so accidents are expected to occur more often on the same side. Therefore, although there is a significant correlation, the result itself is not a factor that can be used as part of our recommendations in how to reduce the number of accidents.




**Presence of Traffic Signals:**

The graph on the left shows that, also regardless of severity, there are more accidents when a traffic signal is *not* present (0=Not Present, 1=Present). An interesting area to explore in future studies to determine whether there are specific locations or intersections that can potentially have a large reduction in accident occurrences if traffic signals or other road and traffic safety features are implemented.

## Descriptive Insights

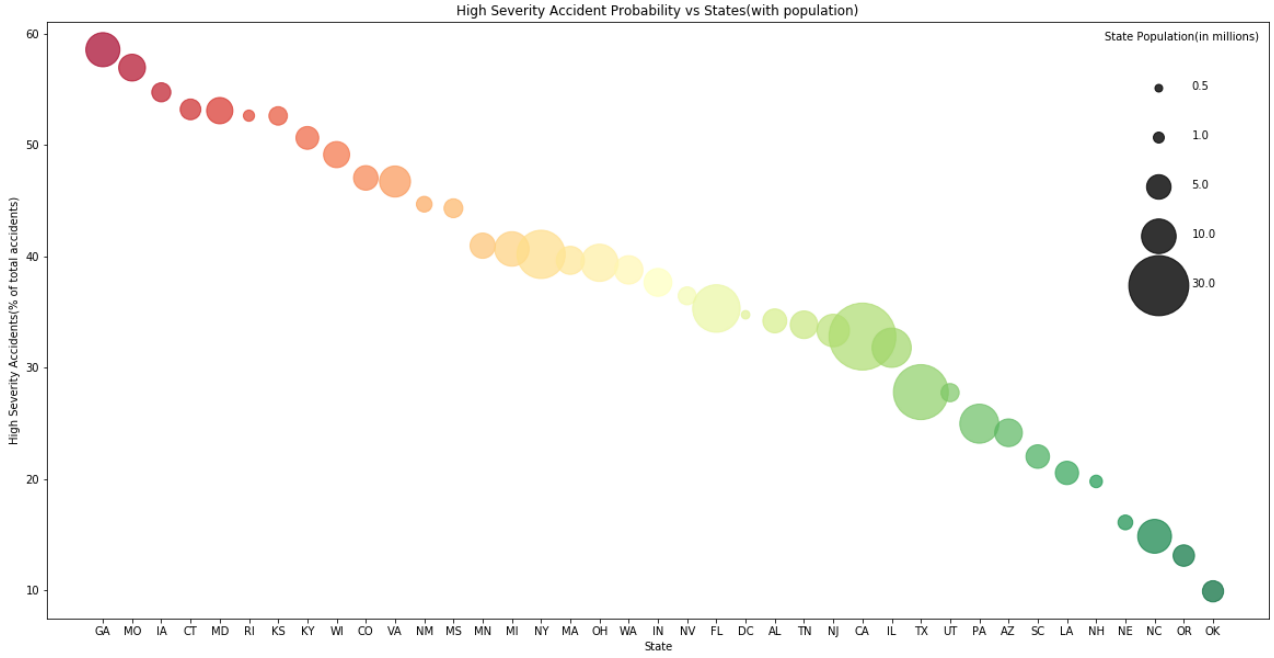
The analysis in this section focuses on the geographic factors that correlate to the highest rate and severity of accidents. Although geographic location is not a predictor of accidents, it is nonetheless interesting to explore *where* accidents occur, the results of which can inform our recommendations on overall strategies to reduce accidents in the country as a whole.

### Geographic Patterns

In order to begin our analysis, we looked at some of the geographical location related data in the dataset and came up with one feature our entire analysis can be based upon, **State.**

First, we looked at all the accidents recorded irrespective of dates and found out that there were three major states that shared a significant proportion of all accidents recorded in the United States namely, California, Texas and Florida. This observation can be further supported by the fact that these three states have among the largest populations.

Then we grouped Severity levels 3 and 4 to get 'High Severity' Accidents and plotted a scatter plot to determine which states had the highest probability of a high severity accident.

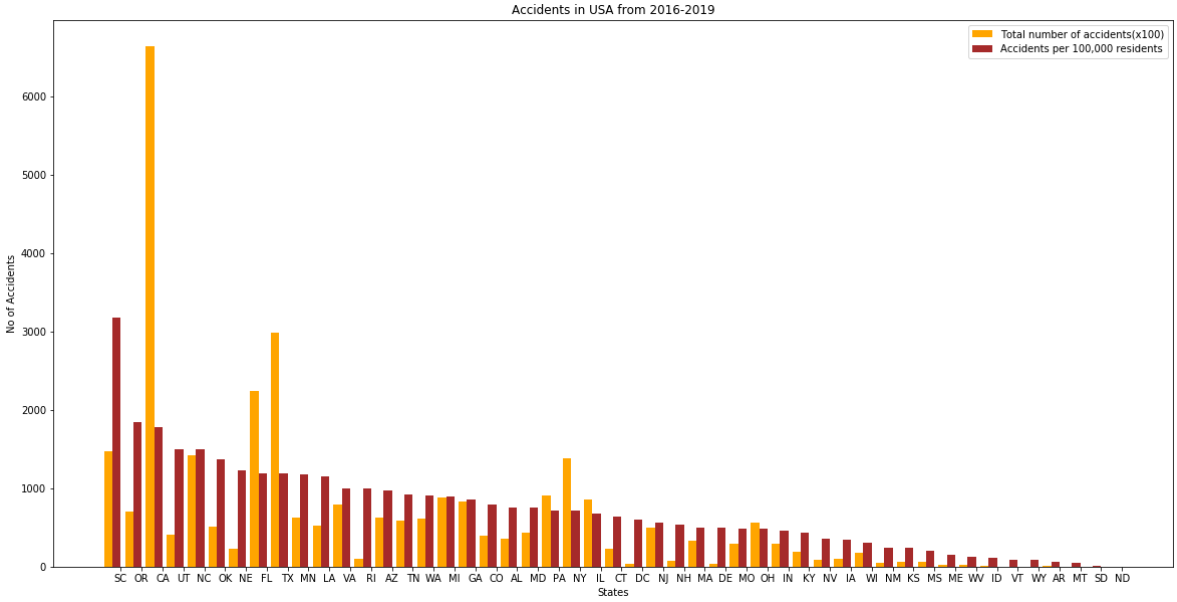



From the image above, we can see that, surprisingly, states with a comparatively lower population were more likely to record high severity accidents as compared to states like NY, CA, TX and FL with a higher population.

However, this wasn't true in all cases. As seen in the graph, there are still some states with much lower populations with a lower probability of high severity accidents.

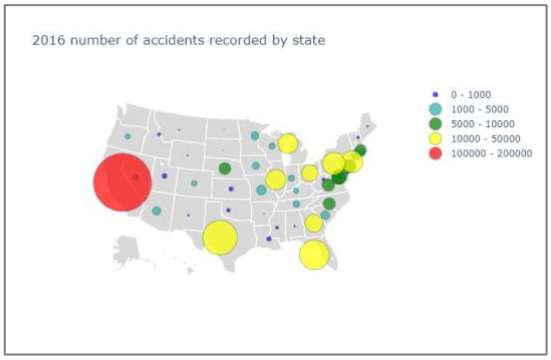
Next, we looked at state populations and number of accidents for each state to figure out the rate of accidents (accidents per 100k residents) in each state.

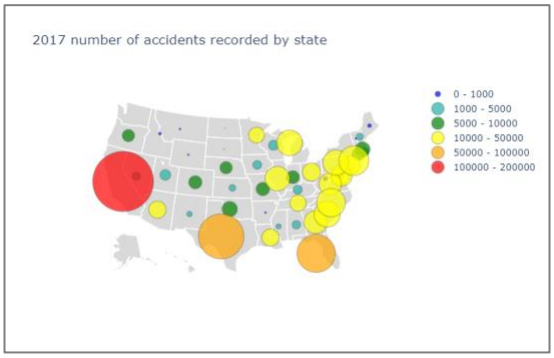
The graph below is sorted by rate of accidents (in red) from highest to lowest:

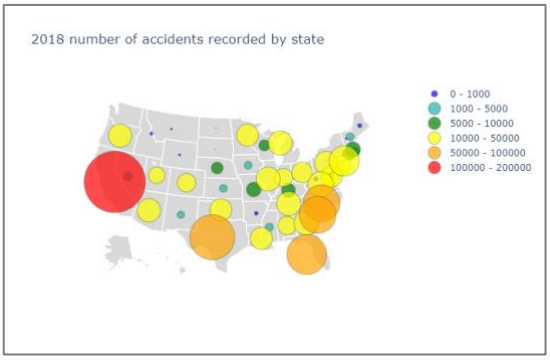


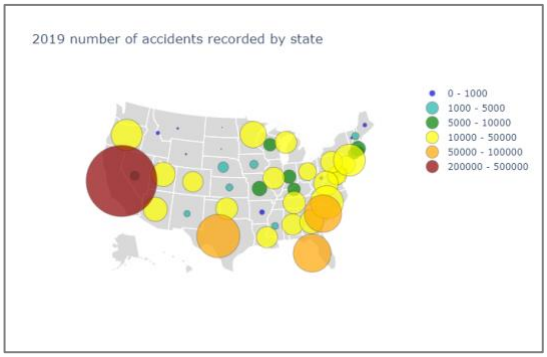

Here we can see that the state of South Carolina and Oregon, despite having a low number of total accidents, top the chart in terms of rate of accidents. However, apart from some anomalies, we can see that rate of accidents decline with lower total number of accidents.

We then looked at the number of accidents in each state for different years to see if we can plot out and notice the increase or decrease in total number of accidents across the country.



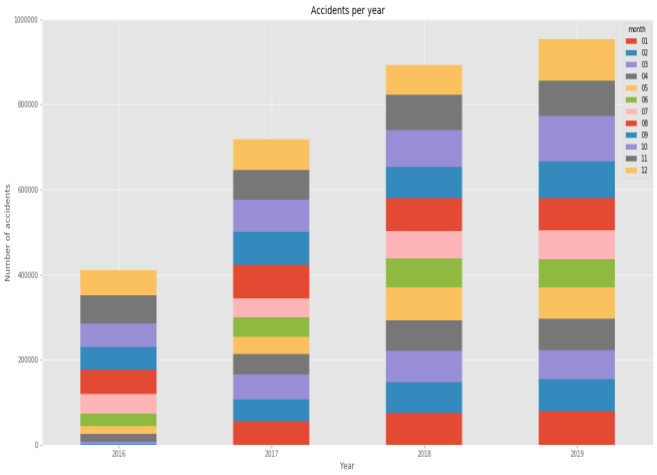




From the images above, we can clearly see a rise in the number of accidents, especially in the coastal regions and states with high population. Although there seems to be a steady increase from 2016 to 2018, if we look closely, on the east coast, there is a visible decline in the number of accidents from 2018 to 2019, however, on the west coast, the trend of increase remained.

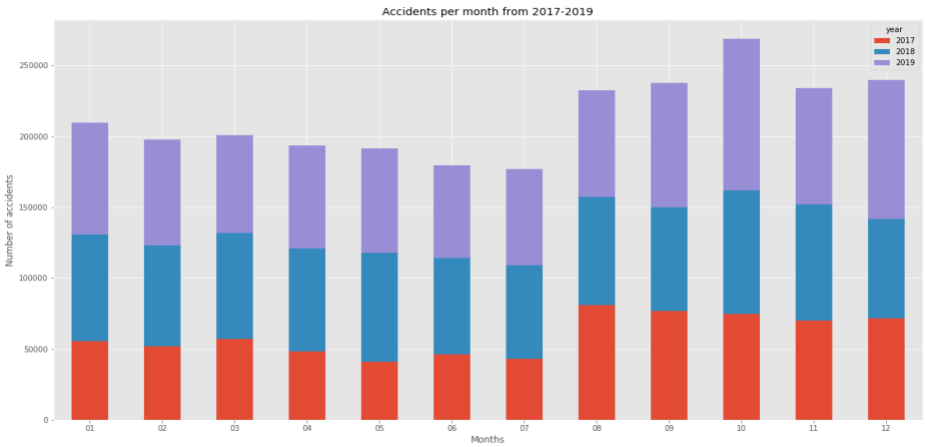
## Time Series Analysis

The final part of our analysis focuses on whether there are any seasonal trends within the accident data. *Seasonality* is a characteristic of a time series in which the data experiences regular and predictable changes that recur every calendar year. Any predictable fluctuation or pattern that recurs or repeats over a one-year period is said to be seasonal.




With this in mind, we first looked at the distribution of accident data over 2016-2019 to determine whether the entire dataset should be included in the analysis. The graph to the left shows the number of records in each year. We noticed that the year 2016 has much fewer records and the records for certain months were missing (i.e. Jan, Feb and March), possibly because it was the initial year where data is collected and therefore the data collection system had not accounted for all the sources that can provide streaming traffic incident data. As such, we have decided to exclude the year 2016 from the time series analysis and focused instead on years 2017-2019.

Using the data from year 2017-2019, we plotted the aggregate number of accidents in each month of the year, as shown in the graph below:




Results show that there is a steady decrease of accidents starting in January until July, then a sudden increase in August followed by a peak number of accidents in October. The remaining months of November and December also see a higher than average number of accidents.

We speculate that the peak number of accidents occurring in the month of October may be caused by the first snowfall in certain parts of the country, or the reduced daylight hours in the afternoon and evening due to Daylight Saving Time. As people acclimatize to the winter weather from November to February, they will have likely changed their driving styles to be more cautious and thus reducing the total number of accidents in those winter months. As warmer weather comes in March to July, accidents are further reduced as temperature increases and visibility is improved.

August is an interesting observation, as it is difficult to speculate why there is a sudden increase in the number of accidents since the weather is still warm and children are not yet back in school. A possibility is that families are taking road trips so there may be more cars on the highways – this would be an interesting future study to explore whether there are other factors causing such a large increase in the number of accidents in August.

# Conclusions

## Results

Through our analysis of the accident data, we came across many interesting and sometimes surprising results:

- **Important predictive features:** Our initial exploration of the attributes that are most predictive of accident severity showed that the majority of important features (8/10) are related to weather conditions, such as humidity and temperature.
- **Expected results from weather patterns:** A deeper analysis of certain weather conditions shows some patterns that are expected and match our hypotheses, such as an increase in accident severity when the temperature is at its lowest (-20 degrees Celsius) or when the precipitation is most severe (heavy or extreme amounts of rainfall).
- **Some surprising observations:** However, we also encountered some unexpected results in the data. For example, with regard to visibility, the most severe accidents actually occurred on mildly clear or clear days. We speculate that this might be because drivers are not as cautious on days where visibility is not affected, leading to faster driving speeds that would result in more severe accidents.
- **Effect of road and traffic design:** On the other hand, the design of the road and traffic, such as whether a traffic signal is present, is mostly predictive of whether accidents are *more likely to occur* instead of *how severe* an accident will be when it does occur.
- **Geographic factors:** An exploration of geographic factors shows that states with the highest number of population (ex. California, Texas and Florida) also have the highest number of accidents. Interestingly, when we looked instead at the severity of the incidents, some states with a comparatively lower population (ex. Georgia, Missouri, Iowa) actually recorded a greater percentage of high severity accidents.
- **Seasonal patterns:** Finally, because our data spanned several years (2016-2019), we were able to explore patterns related to seasonality. The results showed that as expected, there were more accidents when the temperature starts to decrease and daylight is reduced, beginning in October. However, it was surprising to note that there was also a large increase in the number of accidents in August compared to the previous months – we have provided some speculations, but further study will need to be done to determine what factors are involved in this unexpected observation.

## Recommendations

The goal of our study is to provide recommendations on how to reduce the number of high severity accidents. Here are some of our recommendations that can be implemented right away as well as suggestions for future follow-up studies that will provide more data for analysis:

1. **Increase official warnings based on weather conditions:** As weather patterns are such an important predictor of accident severity, states should implement more prominent warnings especially on days with combinations of dangerous weather conditions that drivers may not be aware of (ex. ice that can form on clear but very cold days).
2. **Add traffic signals:** Add traffic signals and other road safety features to reduce the overall number of accidents in major intersections. Further studies will need to be done to pinpoint which specific locations with the highest number of accidents will benefit the most from improved road and traffic design features.

3. **Conduct follow-up study on high severity states:** Another interesting follow-up study can be done for states with a comparatively low population that have a greater than average percentage of high severity accidents. Some factors that can be explored is whether these states can benefit from increased weather warnings, improved road and traffic design, or the addition of more public transit options to reduce the number of drivers on the road.
4. **Issue reminders to be cautious in fall and winter months:** In the months leading to fall and winter, accidents may be reduced if officials put out more reminders for drivers to be cautious of sudden weather changes or a reduction in daylight. An interesting related study can be conducted to determine why there is an unexpected increase in accidents in the month of August in order to further implement plans that may decrease the overall number of accidents throughout the year.

## Challenges and Limitations

Overall, considering the size of our dataset (~3 million records, 49 attributes), the data that was provided is for the most part fairly complete and usable. In some instances where data is missing, we tried to impute these values based on other data that is available (ex. if a temperature value was missing, we used the median of that particular week's temperature to fill in the missing value). For future studies, the data in this dataset can be cross-referenced with other datasets such as records of reasons given by drivers on the circumstances of their accidents to provide more detailed analyses and better customized recommendations.

# Appendix

## List of Documents

The following documents are included as part of the project submission:

- 1 Group Assignment Final Report (.pdf)
- 1 Zip file containing scripts and supporting documents – *note that .ipynb and .csv files should be placed in the same folder*:
    - 3 .ipynb files:
        - data_cleaning.ipynb (data cleaning and model comparison script)
        - data_analysis_weather.ipynb (data analysis for weather patterns)
        - data_analysis_geography.ipynb (data analysis for geographic locations)
    - 7 .csv files:
        - US_Accident_Dec19.csv (source dataset from Kaggle)
        - cleaned_ads.csv (dataset after data cleaning)
        - features.csv, features2.csv and coordinate2timezone.csv (supporting files for data cleaning script)
        - statelatlong.csv and statePopulation.csv (supporting files for data analysis geography script)