MasterClass

MASTERCLASS TEAM

PREDICTING THE FACTORS WHICH CAUSE CAR ACCIDENTS

Team Members

Name	Sec	B.N
Hazem Sherif Mohamed	2	5
Ramez Hany Fawzy Michael	2	14
Seif Selim Mohamed Selim	2	22
Abdelgawad Gomaa Abu Almajd (C)	2	26
Abdelrazek Akram Abdelrazek	2	27
Abdelrahman Ibrahim Elghonami	2	28
Abdelrahman Ahmed Hanafi	2	29
Abdellatif Mostafa Abdellatif Ragab	2	36

- **Sources**:
- Dataset Source : https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents?resource=download
- What each column refers to : https://smoosavi.org/datasets/us_accidents
- GitHub link of our project (we used python in our Analysis, statistical methods and machine learning Algorithms) : https://github.com/Megwed/Car_Accidents_Project

• Introduction :-

All countries and Egypt affect from car accidents in many levels like the Social level and Economical level in this project we will discusses the significant impact of road accidents on a global scale in terms of injuries, disabilities, deaths, and economic costs. It mentions that governments and organizations are working towards reducing road accidents through various measures. The introduction also highlights the importance of understanding factors that contribute to car accidents in order to improve road safety. It outlines a list of statistical questions that will be answered later using collected data, including the frequency, distribution, timing, weather impact, people involved, road conditions and many factors on our dataset we couldn't find an organized dataset represent the car accidents in Egypt but we found a large organized dataset related to USA and we can normalize this data to the rest of the world .

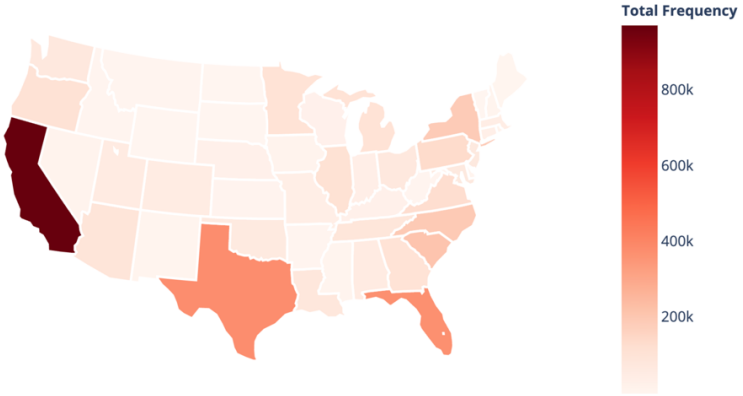
• Literature Review :-

The literature review highlights the consequences of road accidents, such as injuries, disabilities, deaths, and economic costs. It emphasizes the importance of studying road accidents to identify their root causes and develop appropriate prevention measures, such as driver education, traffic law enforcement, and road design improvements. Additionally, understanding the factors contributing to accidents can aid in the development of new safety technologies. In conclusion, road accidents are a critical issue that requires understanding and prevention measures to minimize their impact on society.

• Data Description :-

This is a countrywide traffic accident dataset, which covers 49 states of the United States. The data is continuously being collected from February 2016, using several data providers, including multiple APIs that provide streaming traffic event data. These APIs broadcast traffic events 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 3 million accident records in this dataset. Check the below descriptions for more detailed information. This is a dataset of traffic accidents in the United States that covers 49 states, with data collected continuously since February 2016 using multiple data providers including APIs from various entities such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. The dataset contains about 2.8 million accident records and includes information such as the severity of the accident, start and end times of the accident, GPS coordinates of the start and end points, length of the road extent affected by the accident, natural language description of the accident, street number and name, city, county, state, zip code, and country of the accident, time zone, closest airport-based weather station, weather conditions and observation records, and annotations for presence of various points of interest in the nearby location such as amenity, crossing, junction, railway, and traffic signal. The dataset also includes information about the period of day (i.e., day or night) based on sunrise/sunset, civil twilight, nautical twilight, and astronomical twilight.

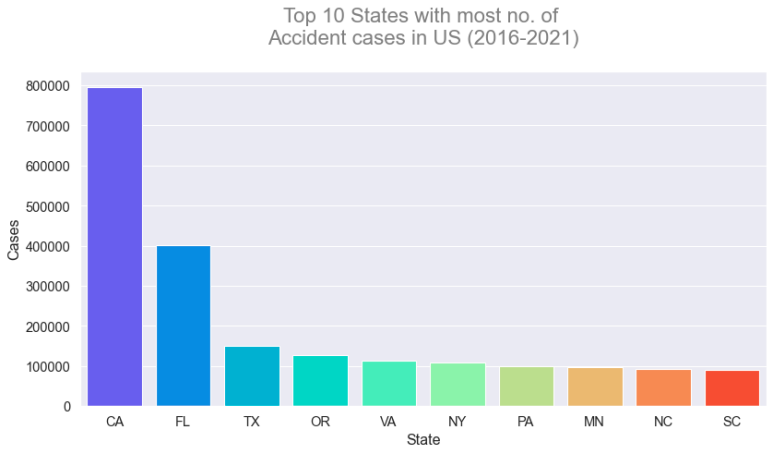
The distribution of car accidents in the United State (2016-2021)



Descriptive Statistics:

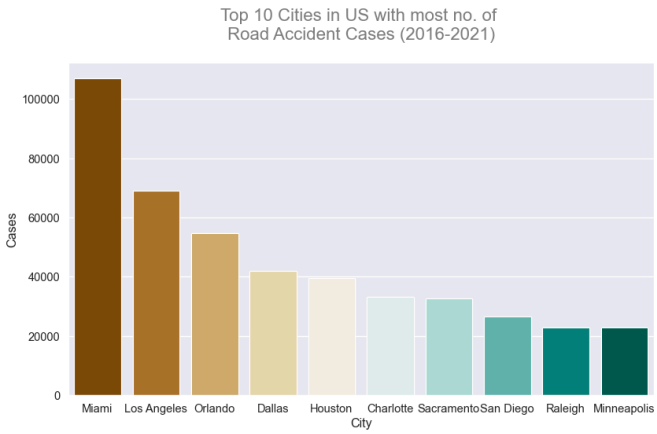
1- Location Condition Analysis:

What is the most state in car accidents?



- that most state in the number of car accidents is California with 795868 cases which is 27% of the Total number of accidents.

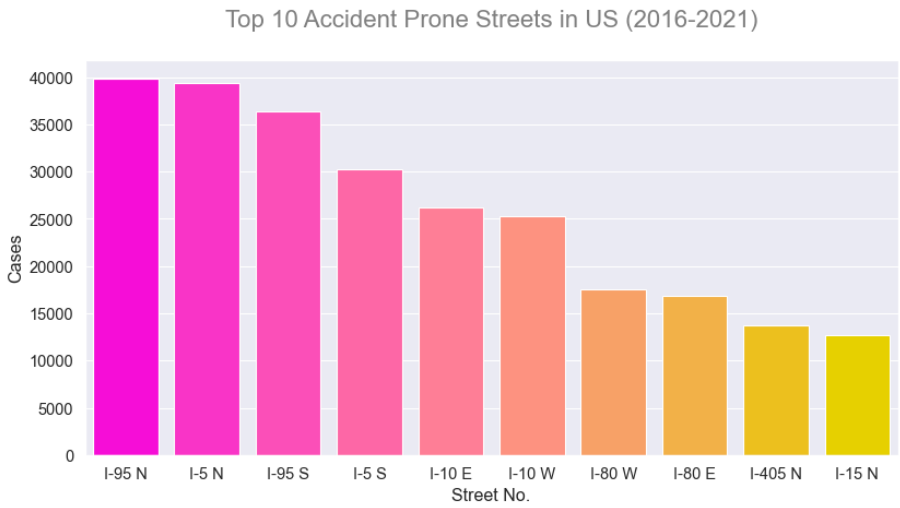
What are the cities with the most accidents ?



Insights:
The top three cities are.
Miami: 106966
Los Angeles: 68956
Orlando: 54691

- If we sum the number of accidents of the top ten cities and divide it by the total number of accidents, we will discover that 15% of total numbers of accidents occurred in only 10 cities.

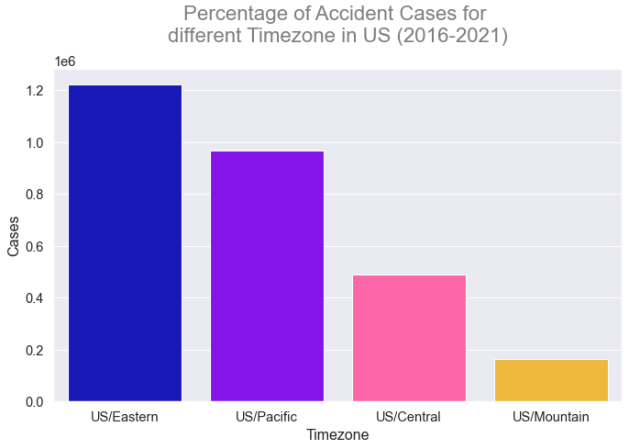
What are the top streets in car accidents?



- Street No. I-95 N has the highest road accidents records.
- 40.18% of the streets in the US have only one accident in the past 6 years
- 97.94% of the street in US have less than one hundred accidents in the past 6 years
- 99.8.18% of the street in US have less than one thousand accidents in the past 6 years
- 0.2% of the street in US have more than one thousand accidents in the past 6 years
- 0.04% of the street in US have more than five hundred accidents in the past 6 years

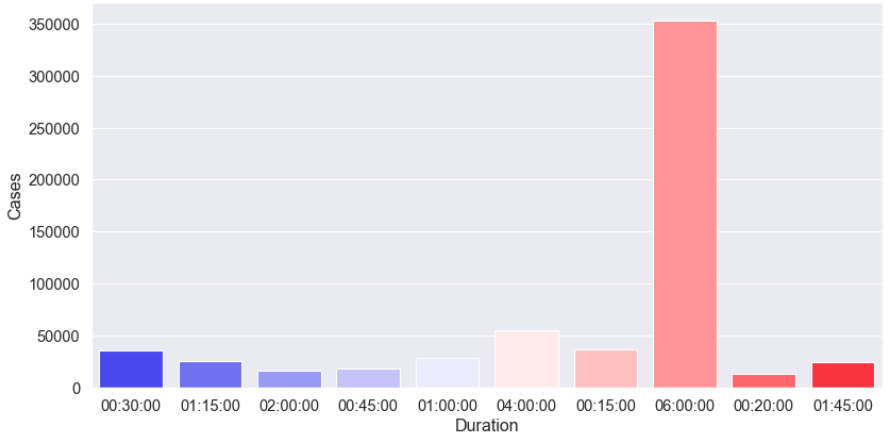
2- Time Condition Analysis

Which Time zone has the most car accidents?



- The Eastern time zone region of the US has the highest no. of road accident cases (43%) in the past 6 years.
- The mountain time zone region of the US has the lowest no. of road accident cases (5.7%) in the past 6 years.

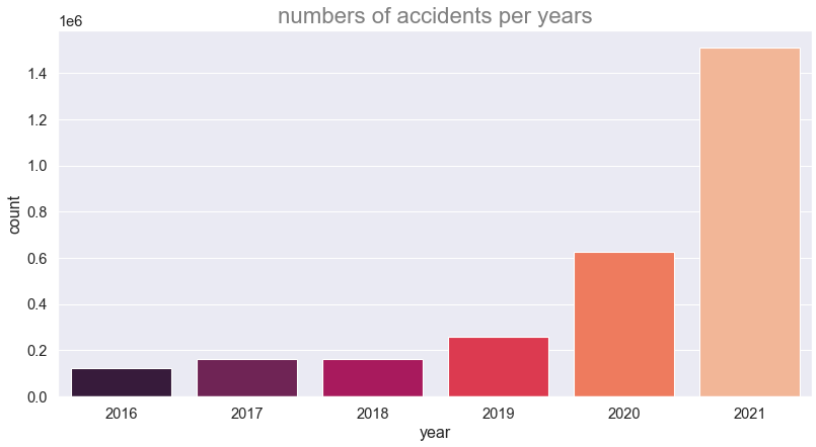
For How long did car accident affect the traffic flow?



So, most accidents affect the traffic flow by 6 Hours.

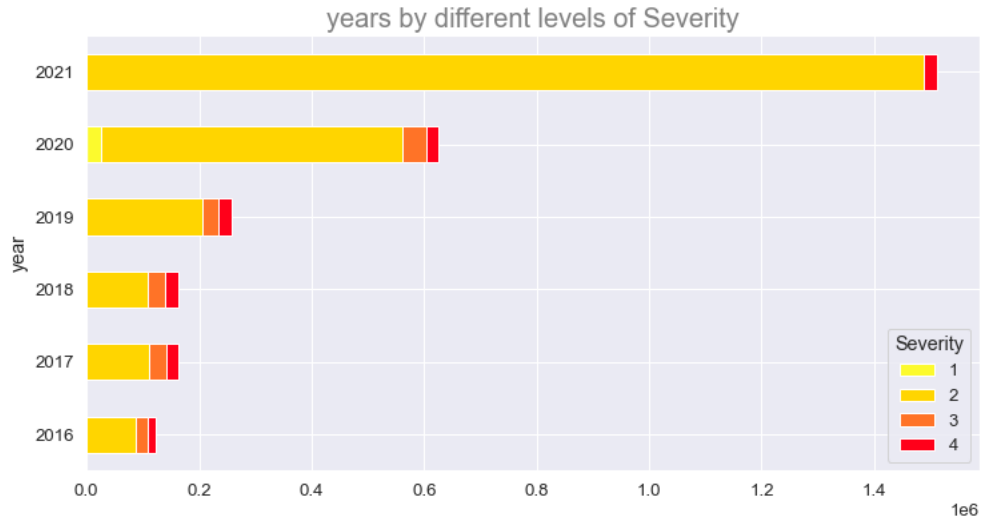
• Year Analysis

How many car accidents occur each year?

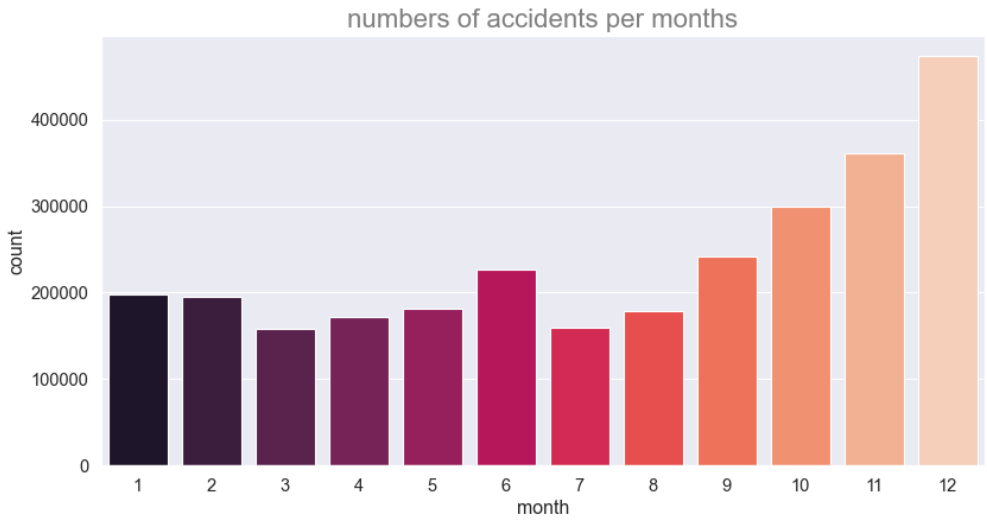


- first insight: car accidents increase year by year in an exponential way what is abnormal.
- Second insight: year 2021 is more than other years with 53% of total accidents number a side NOTE:The increase in accidents number comes from the modern ways of collecting data.

car accidents per years by level of severity

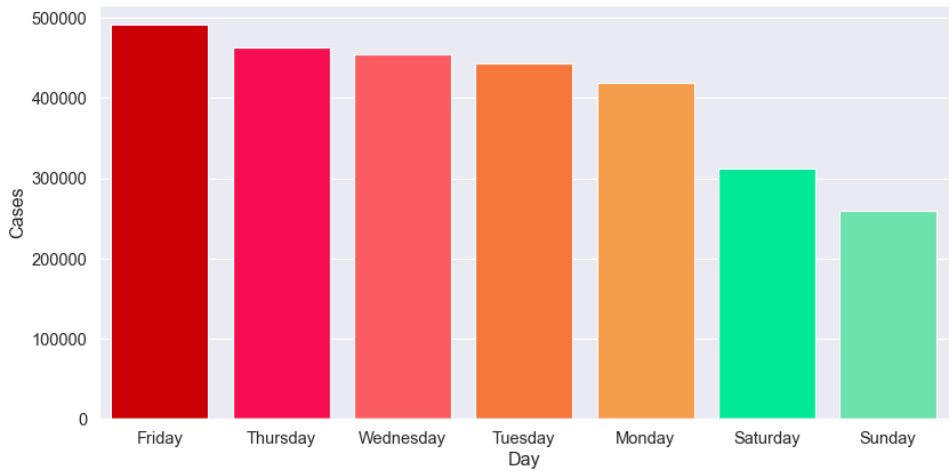


- # Month Analysis



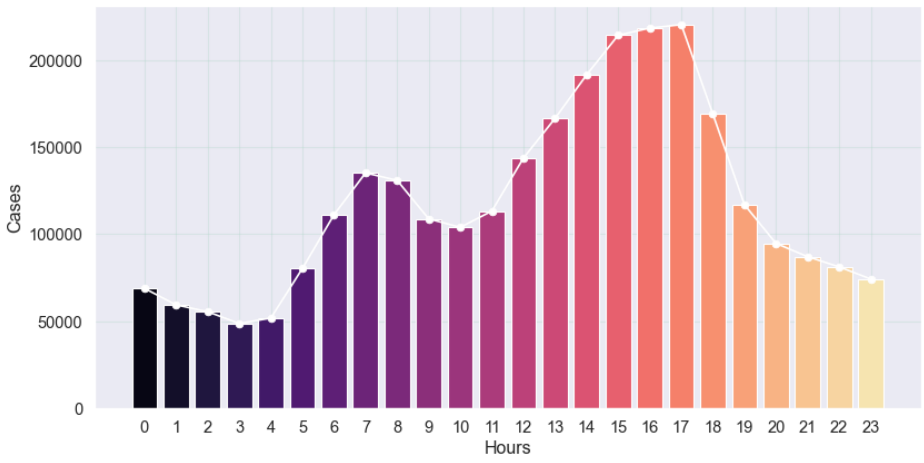
- The number of car accidents is almost constant in the firs 6 months then it starts to increase

Which days have more accidents?



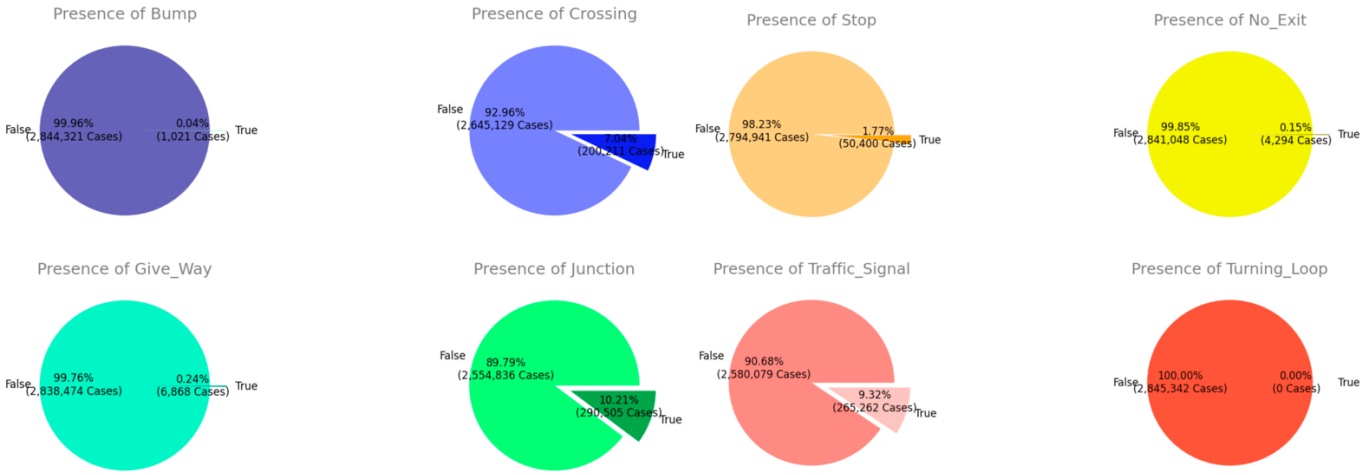
- Accidents occur in working day more than weekend (almost working days are 1.5 times weekend) Only 20% of total accidents occur in weekend because maybe people are staying home, and they are likely to be stress-free.

Rush Hours of car accidents

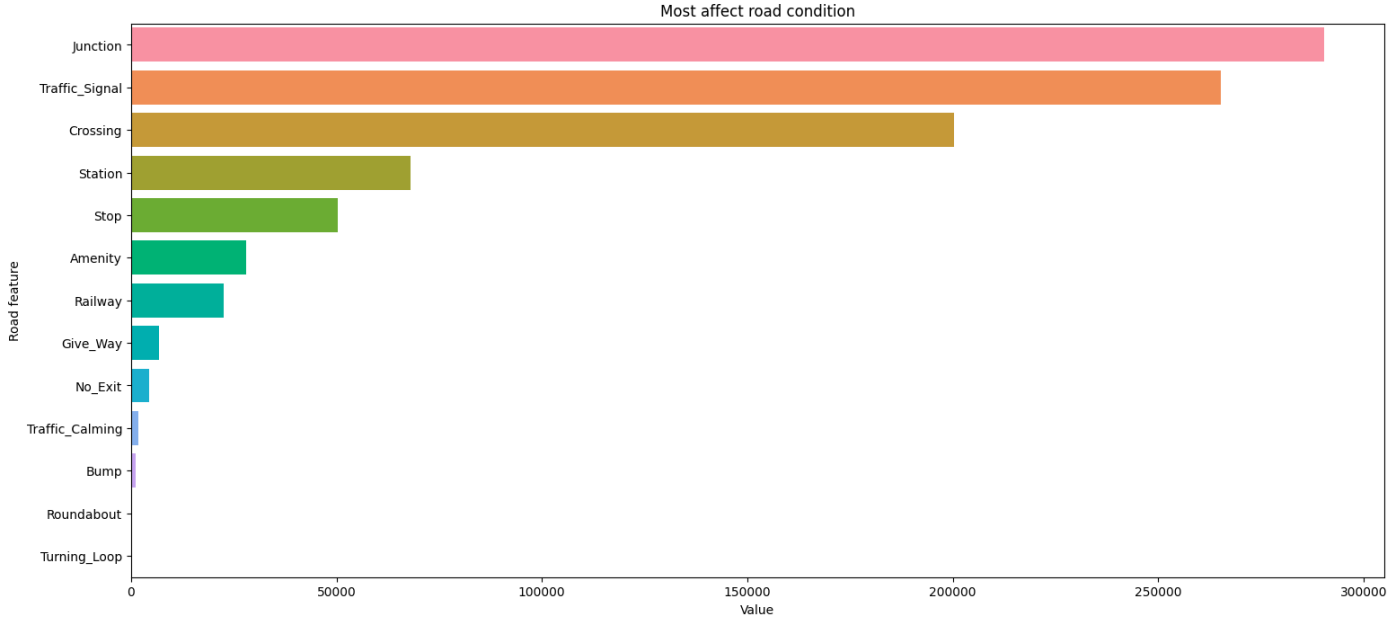


- This Shape called Bimodal/Multimodal Distribution Rush Hours of accidents are 3pm, 4pm and 5pm maybe they are the time of Evening Office-Returning Hours There is another Peak point on 7pm maybe it's the time of Morning Office-Going Hours Around 17% of the oad accidents occurred in between 6:00AM to 9:00AM In evening, around 29% of the road accidents occurred in between 3:00PM to 6:00PM.

3- Road condition analysis



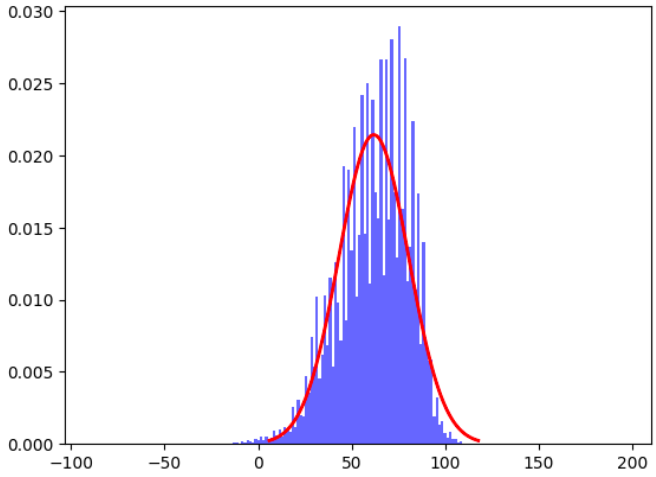
The most affect road condition



- As we can see, most of the accidents occurred near a junction
- The fourth most common road features instead, was the presence of a nearby station probably because of the high presence of vehicles.

4- Weather Analysis Condition

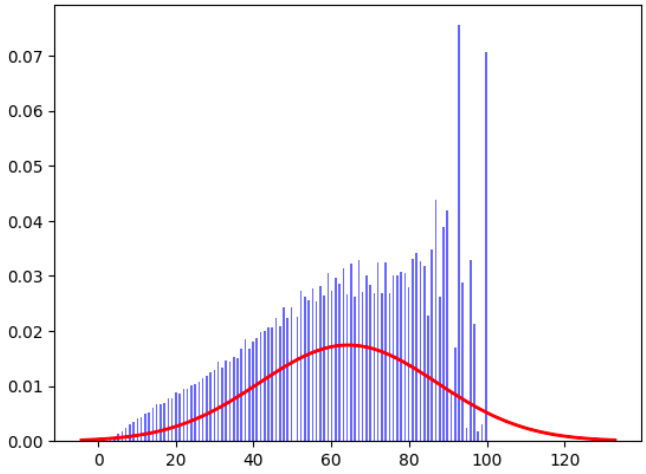
Temperature :-



Mean= 61.79, std = 18.6

As it's a narrow distribution
the probabilities are higher that values won't fall far from the mean. As you increase the spread of the bell curve, the likelihood that observations will be further away from the mean also increases.

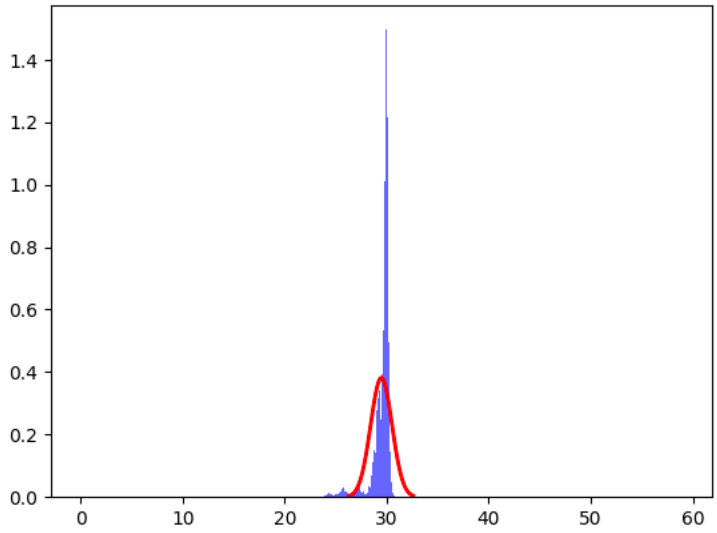
Humidity :-



Mean = 64.36, std = 22.87

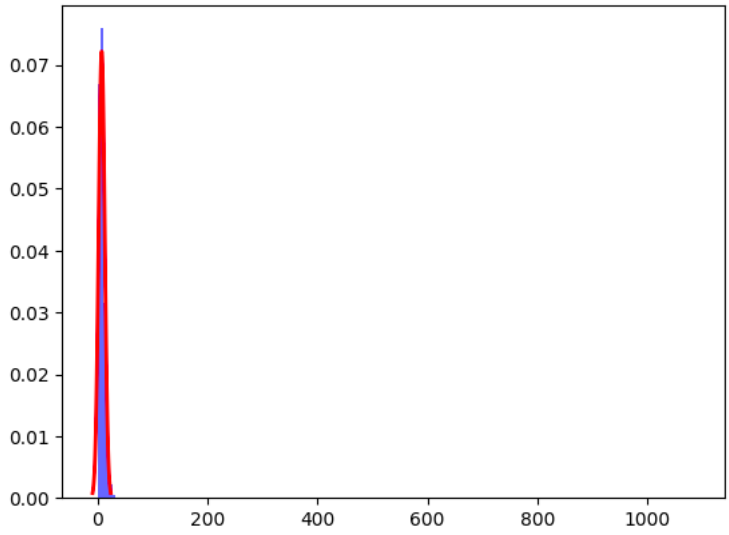
As it's a wide normal distribution
the probabilities are higher that values will fall far from the mean. As you increase the spread of the bell curve, the likelihood that observations will not be further away from the mean also increases.

Pressure :-



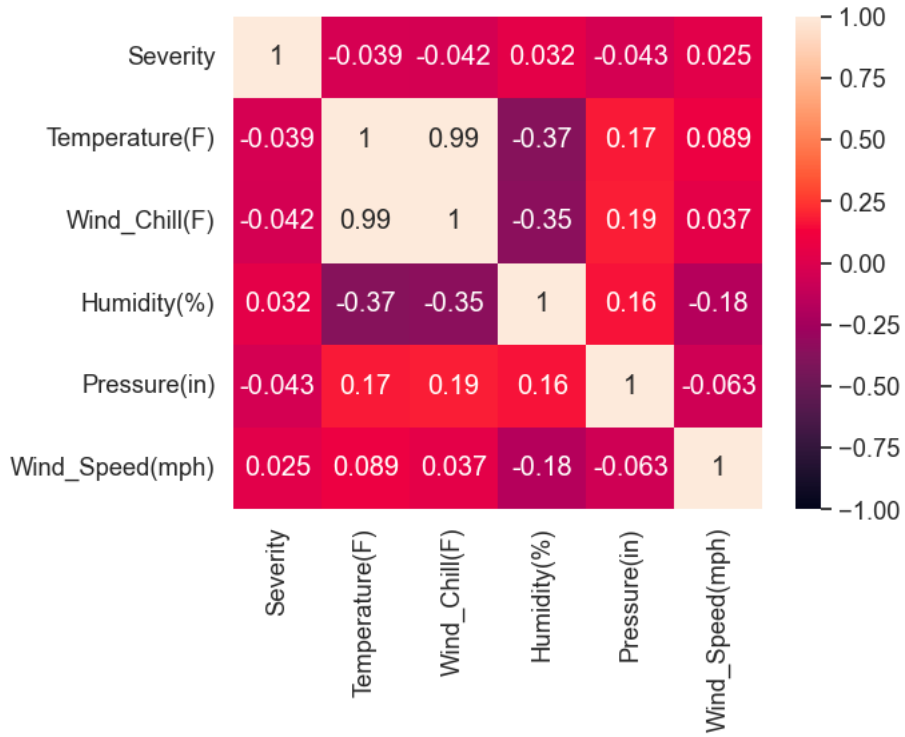
mean = 29.47, std = 1.04
the spread in this curve is very narrow.
The most of the data points concentrated in a very small range

wind speed :-



mean = 7.39, std = 5.52
like the pressure data distribution the spread is very narrow and the most of data points concentrated in a very small range

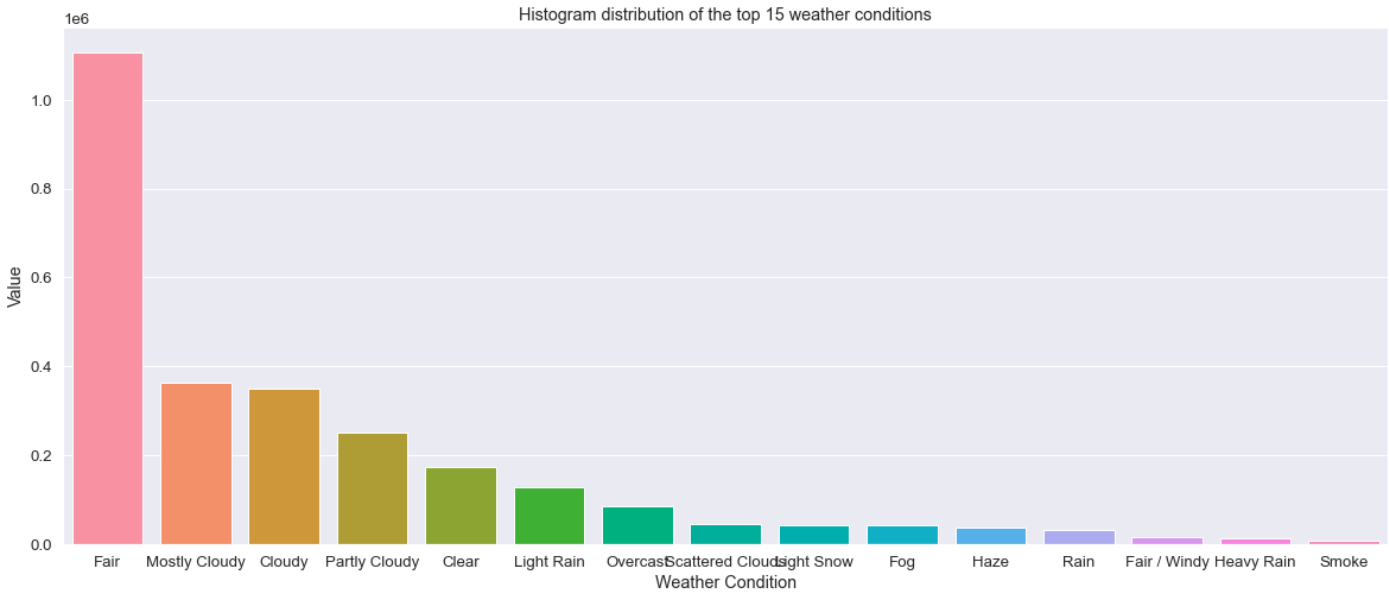
This weather heat map shows the correlation between each different weather items.



This weather heat map shows the correlation between each different weather items.

- We notice there is a high positive correlation between Temperature and Wind chill.
- Humidity with both Temperature and wind chill has a large negative correlation.
- Humidity is the only factor of weather factors in a positive correlation with severity means that when the humidity increase the severity also increase.
- Pressure is the most Negative correlated factor with severity.

What is the most common weather conditions ?



- In the most frequent cases the weather is clear.

Hypothesis Test :-

In our project, we use hypothesis testing as a statistical method to examine and test specific assumptions (like we assumed that the severity of car accidents in the USA is 2.5 as we think if it 2.5 or above it will be dangerous and other assumption) or hypotheses about a population parameter based on the data we have collected from a sample. It allows us to draw conclusions and make inferences about the larger population using the available sample information. Within hypothesis testing, the p-value plays a crucial role. It represents the probability of observing a test statistic that is as extreme as, or more extreme than, the one we obtained from our sample data, assuming that the null hypothesis is true. In simpler terms, it quantifies the likelihood of obtaining the observed results purely by chance. By analysing the p-value, we can assess the strength of evidence against the null hypothesis. If the p-value is small (typically below a pre-determined significance level like 0.05), it suggests that the observed data is unlikely to have occurred by random chance alone. In such cases, we reject the null hypothesis and consider the alternative hypothesis, indicating that there is a significant relationship or effect present in the population. Therefore, in our project, we utilize hypothesis testing and examine the p-value to make informed decisions about the validity of our assumptions and the significance of the relationships or effects we observe in the data and these are our results.

- **Null hypothesis (H_0):** There's **no effect** in the population.
- **Alternative hypothesis (H_a or H_1):** There's an **effect** in the population.

The formula of calculating P-Value :

$$Z = \frac{\hat{p} - p0}{\sqrt{\frac{p0(1 - p0)}{n}}}$$

Where,

$\hat{p}$ = Sample proprotion

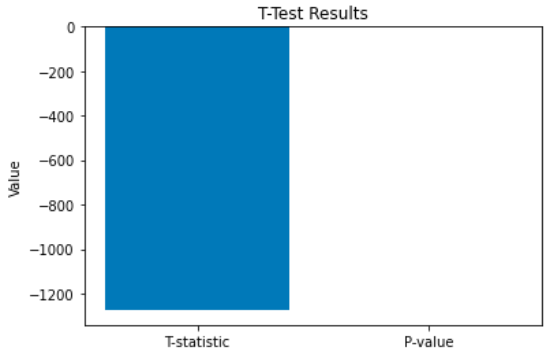
p0 = Assumed population proportion in the null hypothesis

n = Sample size

First Hypothesis :

We assumed that the average Severity of Car Accidents in the USA is above 2.5 .

The Hypothesis can be rejected as the P-value is equal = 0
And T-Test = -1277.0459

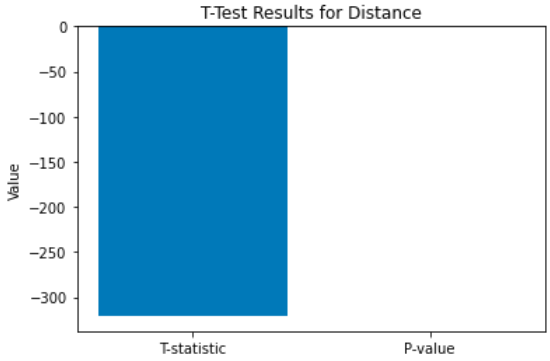


- That's Indicate that the most of car accidents in USA isn't dangerous

Second Hypothesis :

We assumed that Average distance that an accident affects the road would be one mile.

This Hypothesis Also can be rejected as the P-value is equal = 0
And T-Test = -321.4174

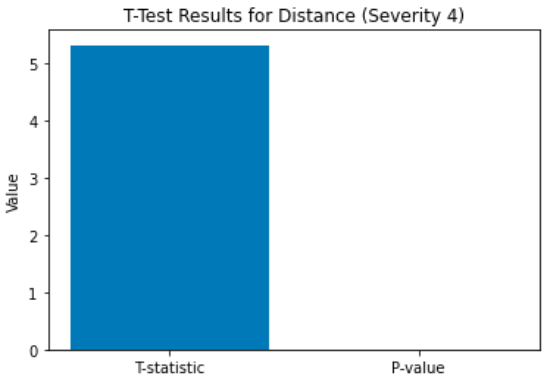


- That's Indicate that the most of car accidents don't exceed one mile

Third Hypothesis :

We assumed that severity of four result accidents that are 1.4 miles long

This Hypothesis Also can be rejected as the P-value is equal = 1e^-07
Which very close to ZERO And T-Test = -321.4174



- That's indicate that the most of car accidents with high severity don't effect the road a lot.

Pearson Correlation coefficient

The correlation coefficient, r, is a summary measure that describes the extent of the statistical relationship between two interval or ratio level variables. The correlation coefficient is scaled so that it is always between -1 and +1. When r is close to 0 this means that there is little relationship between the variables and the farther away from 0 r is, in either the positive or negative direction, the greater the relationship between the two variables.

In our project we used Pearson Correlation coefficient and this it's formula

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where n = Quantity of Information

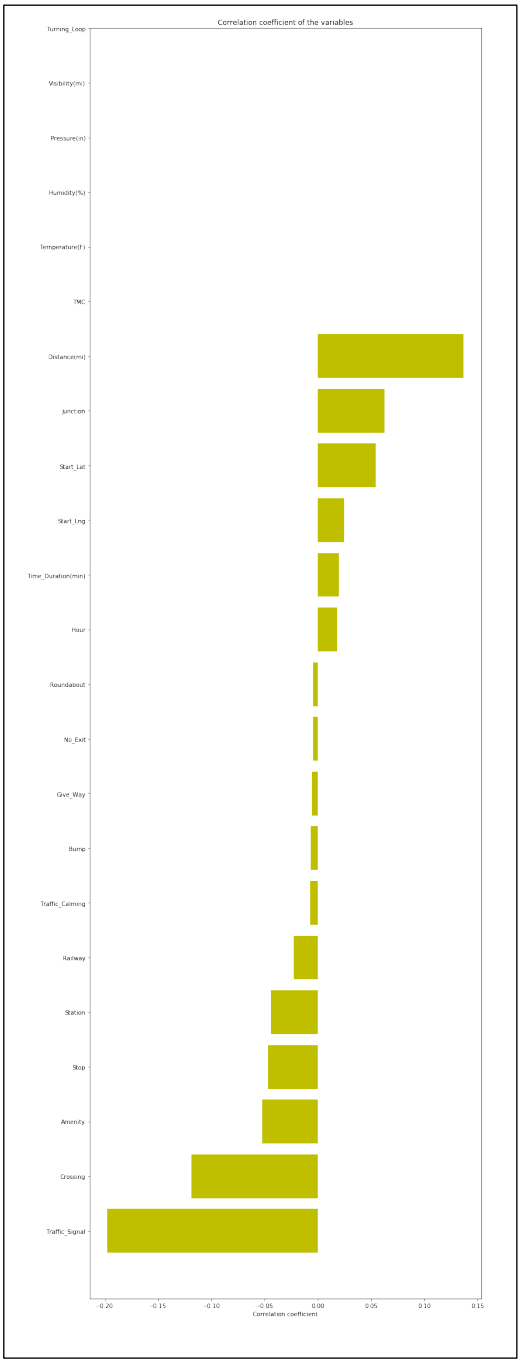
Σx = Total of the First Variable Value

Σy = Total of the Second Variable Value

Σxy = Sum of the Product of first & Second Value

Σx^2 = Sum of the Squares of the First Value

Σy^2 = Sum of the Squares of the Second Value



- We notice that the presence of traffic signals reduce The severity of the accident also crossing, amenity and Stops point reduce the number of the severity of Accidents and the number of them.

- We advice the government to increase the number of Traffic signals, crossing, stops and stations in the Country as the help in reducing the severity of Accidents and the number of them

Linear Regression:

Linear regression is a statistical technique that aims to model the relationship between a dependent variable and one or more independent variables. It is commonly used to predict or estimate the value of the dependent variable based on the given independent variables. In our project, we employed linear regression as an analytical tool to investigate the factors that contribute to car accidents. The main objective was to understand the position of the car in the beginning of the accident relate to the severity of accidents. The linear regression model assumes a linear relationship between the dependent variable (severity of accidents) and the independent variables (Start latitude). It estimates the coefficients for each independent variable to determine their impact on the dependent variable. To conduct the analysis, we first divided our dataset into a training set and a testing set. The training set was used to fit the linear regression model, while the testing set was used to evaluate its performance. After fitting the model, we examined the coefficients of the independent variables to understand their influence on the severity of accidents. A positive coefficient indicated a positive relationship, meaning that an increase in the variable would lead to an increase in accident severity. Conversely, a negative coefficient indicated a negative relationship, implying that an increase in the variable would result in a decrease in accident severity. Additionally, we evaluated the goodness of fit of the model using metrics such as mean squared error (MSE) and R-squared. The MSE provided a measure of the average squared difference between the predicted and actual severity values, while the R-squared indicated the proportion of variance in the dependent variable explained by the independent variables.

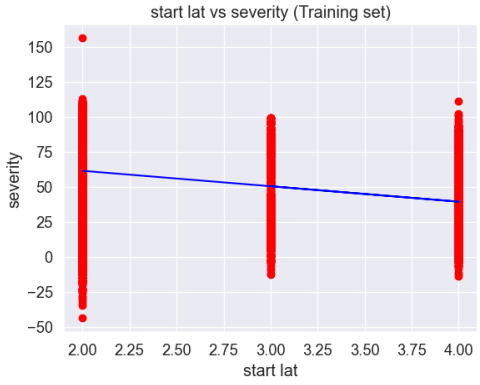
the equation for a line: y = mx + b

- Y = the vertical value.
- M = slope (rise/run).
- X = the horizontal value.
- B = the value of Y when X = 0 (i.e., y-intercept).

Coefficients : -11.02228183
Intercept : 83.33171653

$$Y = -11.022\,X + 83.33$$

So if we want to know the severity of any accidents we could know by substitute the value of the starting point of the accident in (X)



Conclusion :

Traffic accidents affect society in many ways (socially and economically), Human losses are estimated to be 6000 citizens annually in Egypt and it costs the country a lot of money, the loss in money is estimated to be 1-3% of the total local product which is 5 billion Egyptian bounds annually. So, it was essential to reduce accidents' harmful effects.

How to reduce traffic accidents ?

1- Spreading working hours all over the day so that there is no traffic jam at the same time.
2- Allocating centers of trade and industry in the country to various places
3- Warning citizen not to drive cars when the humidity is high.
4- The country should increase the number of traffic lights, stations, stop places and Pedestrian road.