PROJECT REPORT

ON

DATA PREPROCESSING AND EDA

BY

Sushant Ovhal

sushantrovhal[@gmail.com](mailto:yogeshbharti888@gmail.com)

30/12/2022

Table of contents

* Aim
* Objective
* Introduction
* Research Questions
* Tools and datasets used
* Methodology
* Exploratory Analysis and Visualization
* Conclusion
* Recommendation and future work
* Reference

Aim:-

This project aims to collect and analyze the data of the US Accidents Dec21 and also analyze the Accidence according to city, country, season wise and find out the Reasons and meaning full information.

Objective:-

The main objective of this project is to data cleaning, remove null values, data pre-processing for machine learning and remove outliers and we need to improve the performance of the data set.

Introduction:-

* The automobile engine picked up the preeminent position as the primary means of transport. From that point forward, no one has challenged the dominance of engine vehicles. Instead, there have been various efforts to improve them, for example: to make the assembly line faster, to make them progressively adapted to the geographical terrain found in individual nations
* At present, automobile transport has become a piece of daily life. Improvement of automobiles is inescapable given the shockingly on-going high rates of terrible accidents and deaths. Unfortunately, vehicle crashes have always been a part of the vehicle driving experience
* This project analyses accident data collected from **49 states**of the United States of America over the period between February 2016 and December 2020. The set is sourced from [Kaggle](https://www.kaggle.com/sobhanmoosavi/us-accidents" \t "_blank). It has a total of **29,06,610 rows and 47 columns**; inclusive of but not limited to variables such as ‘State’, ‘City’, ‘Start\_Time’, ‘Temperature(F)’, ‘Weather\_Condition’ etc.

Research Questions :-

1. Are there more accidents in warmer or colder areas?
2. Which 5 states have the highest number of accidents?
3. Does New York show up in the data? If yes, why is the count lower if this is the most populated city?
4. Among the top 100 cities in number of accidents, which states do they belong to most frequently?
5. What time of the day are accidents most frequent?
6. Which days of the week have the most accidents?
7. Which months have the most accidents?
8. What is the trend of accidents year over year (Decreasing/ Increasing) .

Tools and datasets used:-

1. Kaggel
2. Google colab and Jupyter notebook
3. Google.com

Methodology:-

## Step 1: Data collection

My Project Mentor give me this dataset. I have visited [www.kaggel.com](http://www.kaggel.com/) to find and download the data set on google.

## Step 2: Data collection

I imported the data set with the help of the panda's library in python. The data set is present in the form of a .csv file. I convert the csv file to excel and then import the dataset in python.

# Step 3: Data investigation

In the investigation of data. I have investigated how much data is there like how many rows and columns are there also finalized how many none values are there and how to replace them, what is size, and shape of the data set is.

# Step 4: Data cleaning

In data cleaning I cleaned the data with python because the data-set I found on the website maybe was not have been cleaned and included some missing values, so with the help of the Panda library of python, I cleaned the data and started working on it.

# Step 5: Data interpretation

In this data interpretation step, I found out some important information about the data set.

# Step 7: Visualization

In visualization, I took the help of the ‘Matplotlib’ and ‘seaborn’ library to plot some relevant graphs and visualize.

# Step 7: Pre-processing

1 -Remove non-numerical columns from data-set

2 -Mapping → assign a new value to Segment columns

3 -Preparing data-set for machine learning

4 -Correlation with heat-map :

# Step 7: Highly correlated features

1 -Definition of the predictors and the criterion

2 -Correlations with the output variable

3 -Identification of highly correlated features

4 -Removing highly correlated features selecting numerical variables

5 -Find the correlation among the variables

# Step 8: Pre-processing for Machine Learning

1 -Features variables

2 -Target variables

3 -Training data-set

4 -Reshape data for Scaling and transform

5 -Fitting data-set

Output:-

Total columns: 47

**Columns present in the datasets:**

Index (['ID', 'Severity', 'Start Time', 'End Time', 'Start Lat' , 'Start Lng',

'End Lat', 'End Lng', 'Distance(mi)', 'Description', 'Number', 'Street',

'Side', 'City', 'County', 'State', 'Zip code', 'Country', 'Time zone',

'Airport Code', 'Weather Timestamp', 'Temperature(F)', 'Wind Chill(F)',

'Humidity (%)', 'Pressure(in)', 'Visibility(mi)', 'Wind Direction',

'Wind Speed(mph)', 'Precipitation(in)', 'Weather Condition', 'Amenity',

'Bump', 'Crossing', 'Give Way', 'Junction', 'No Exit', 'Railway',

'Roundabout', 'Station', 'Stop', 'Traffic Calming', 'Traffic Signal',

'Turning Loop', 'Sunrise Sunset', 'Civil Twilight', 'Nautical Twilight',

'Astronomical Twilight'],

dtype='object')

**Number of null values in each column**:

ID 0

Severity 0

Start\_Time 0

End\_Time 0

Start\_Lat 0

Start\_Lng 0

End\_Lat 0

End\_Lng 0

Distance(mi) 0

Description 0

Number 615177

Street 1

Side 0

City 55

County 0

State 0

Zipcode 435

Country 0

Timezone 1026

Airport\_Code 3647

Weather\_Timestamp 15516

Temperature(F) 21746

Wind\_Chill(F) 214122

Humidity(%) 22728

Pressure(in) 18585

Visibility(mi) 22534

Wind\_Direction 22626

Wind\_Speed(mph) 57687

Precipitation(in) 231683

Weather\_Condition 22892

Amenity 0

Bump 0

Crossing 0

Give\_Way 0

Junction 0

No\_Exit 0

Railway 0

Roundabout 0

Station 0

Stop 0

Traffic\_Calming 0

Traffic\_Signal 0

Turning\_Loop 0

Sunrise\_Sunset 1662

Civil\_Twilight 1662

Nautical\_Twilight 1662

Astronomical\_Twilight 1662

dtype: int64

**Information of datasets**:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1048575 entries, 0 to 1048574

Data columns (total 47 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 ID 1048575 non-null object

1 Severity 1048575 non-null int64

2 Start\_Time 1048575 non-null object

3 End\_Time 1048575 non-null object

4 Start\_Lat 1048575 non-null float64

5 Start\_Lng 1048575 non-null float64

6 End\_Lat 1048575 non-null float64

7 End\_Lng 1048575 non-null float64

8 Distance(mi) 1048575 non-null float64

9 Description 1048575 non-null object

10 Number 433398 non-null float64

11 Street 1048574 non-null object

12 Side 1048575 non-null object

13 City 1048520 non-null object

14 County 1048575 non-null object

15 State 1048575 non-null object

16 Zipcode 1048140 non-null object

17 Country 1048575 non-null object

18 Timezone 1047549 non-null object

19 Airport\_Code 1044928 non-null object

20 Weather\_Timestamp 1033059 non-null datetime64[ns]

21 Temperature(F) 1026829 non-null float64

22 Wind\_Chill(F) 834453 non-null float64

23 Humidity(%) 1025847 non-null float64

24 Pressure(in) 1029990 non-null float64

25 Visibility(mi) 1026041 non-null float64

26 Wind\_Direction 1025949 non-null object

27 Wind\_Speed(mph) 990888 non-null float64

28 Precipitation(in) 816892 non-null float64

29 Weather\_Condition 1025683 non-null object

30 Amenity 1048575 non-null bool

31 Bump 1048575 non-null bool

32 Crossing 1048575 non-null bool

33 Give\_Way 1048575 non-null bool

34 Junction 1048575 non-null bool

35 No\_Exit 1048575 non-null bool

36 Railway 1048575 non-null bool

37 Roundabout 1048575 non-null bool

38 Station 1048575 non-null bool

39 Stop 1048575 non-null bool

40 Traffic\_Calming 1048575 non-null bool

41 Traffic\_Signal 1048575 non-null bool

42 Turning\_Loop 1048575 non-null bool

43 Sunrise\_Sunset 1046913 non-null object

44 Civil\_Twilight 1046913 non-null object

45 Nautical\_Twilight 1046913 non-null object

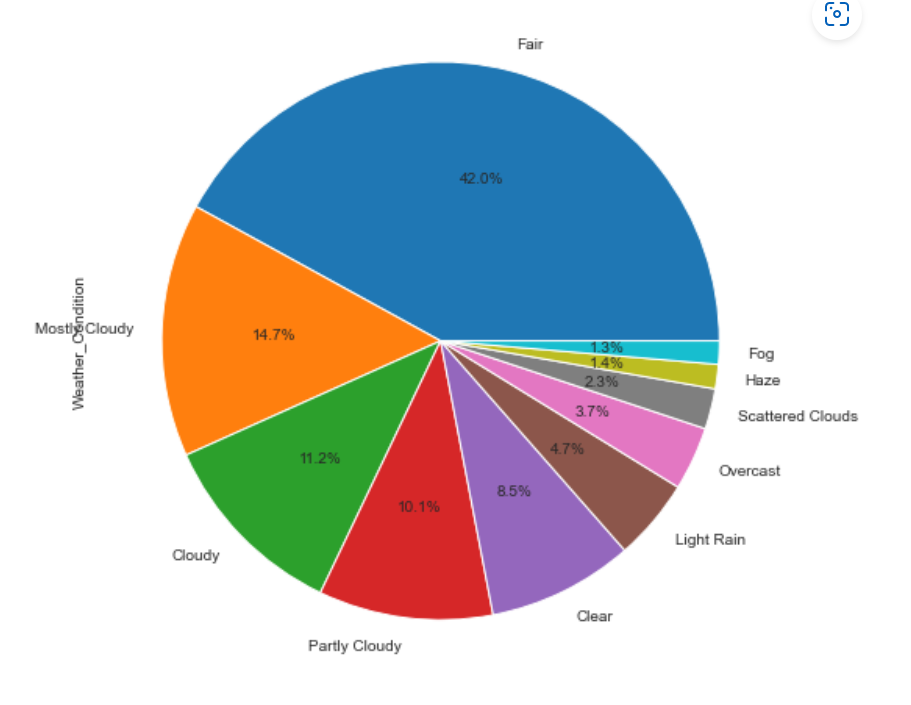
46 Astronomical\_Twilight 1046913 non-null object

dtypes: bool(13), datetime64[ns](1), float64(13), int64(1), object(19)

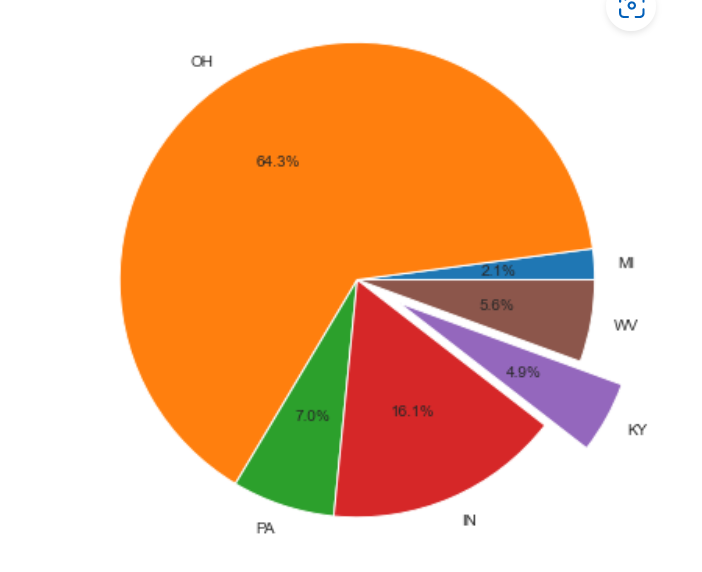
memory usage: 285.0+ MB

Exploratory Analysis and Visualization :-

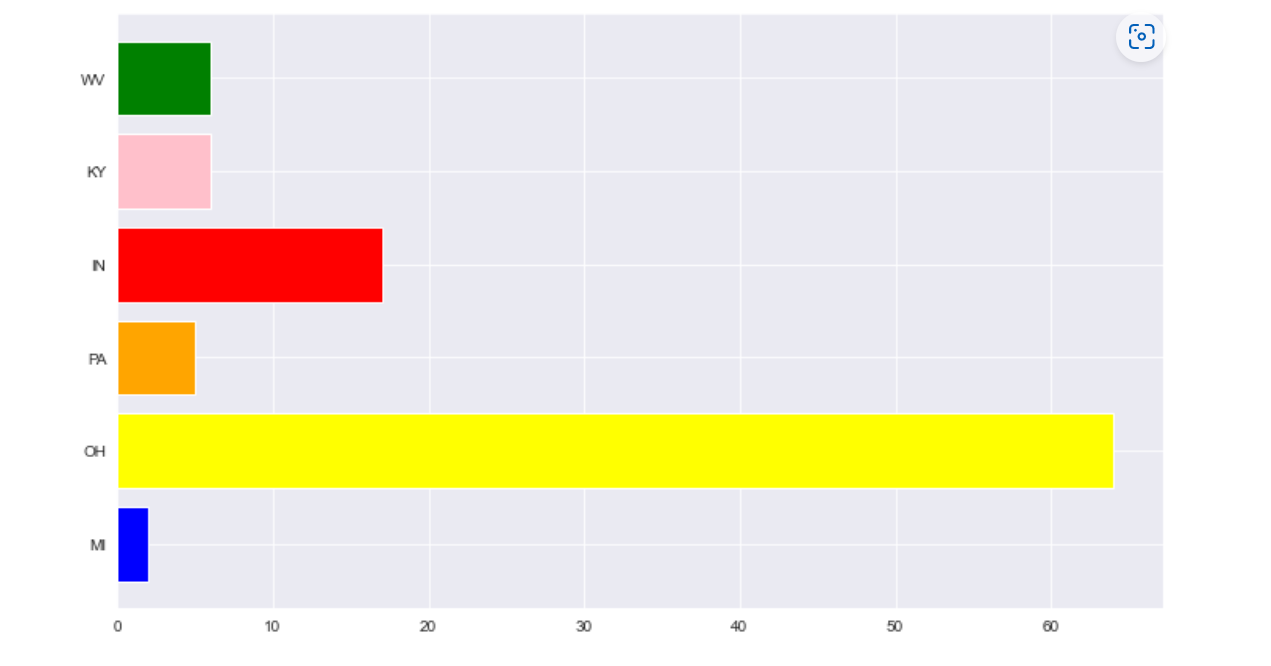
1. Plot the pia chart of the Accident occure because of the weather condition.

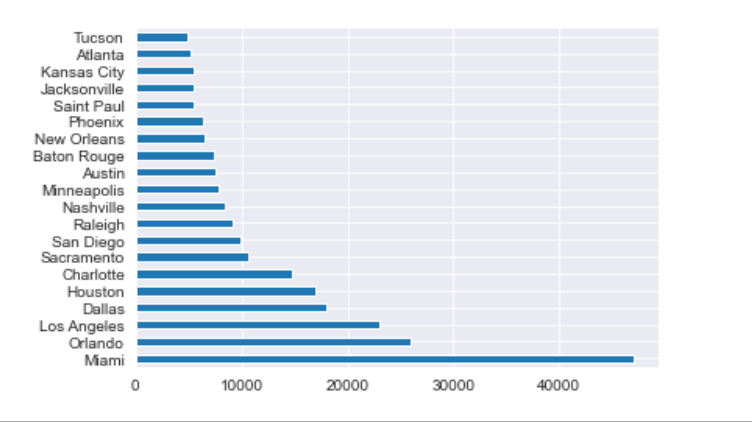


1. Plot the Pia chart of states have the highest number of accidents



1. Among the top 100 cities in number of accidents, which states do they belong to most frequently.





Conclusion:-

Traffic accidents are a main public safety issue, with much research devoted to the analysis and prediction of these rare events. The study helped us to derive factors that are responsible for accidents. From this dataset, a variety of insights concerning the location, time, weather, and points-of-interest of an accident are found. The analysis helps us understand the best month, day, and hour of the day to travel. Also, it can help us to predict what are the accident prone areas in each state.

1. Are there more accidents in warmer or colder areas?

Ans :- Warmer - 0

colder - 0

1. Which 5 states have the highest number of accidents?

Ans :- California, Florida, Texas, Oregon, Virginia.

1. Does New York show up in the data? If yes, why is the count lower if this is the most populated city?

Ans :- The New York is not showing in the data.

1. Among the top 100 cities in number of accidents, which states do they belong to most frequently?

Ans:- Miami, Los Angeles, Orlando, Dallas, Houston.

1. What time of the day are accidents most frequent?

Ans :- 11am to 7pm

1. Which days of the week have the most accidents?

Ans :- Monday

1. Which months have the most accidents?

Ans :- December has the most number of accidents.

1. What is the trend of accidents year over year (Decreasing/ Increasing) .

Ans:- It is exponentially increasing

RecommendationandFuturework:-

The available Dataset will further data set can be collected and after merging with this data set we can perform this same analysis on the complete data set.

* EDA
* Pre-processing for machine learning
* We can find confusion matrix, Lenear Regression, Random forest tree etc.
* Per-Capita accident figures can be analyzed if the population data is sourced from authentic sources.
* An exponential rise in accidents was seen in the year 2020-2021, further analysis using external sources can be done to figure out the reason.
* Predictive analysis can be done which will help in devising ways to prevent further accidents.
* Various parameters like Humidity, Pressure, Severity, etc. which haven't been considered for this particular analysis, can be analysed to gain more insights.

Reference:-

Github.com

kaggel.com

Google.com