REPORT

SAURABH KR VIDYARTHI

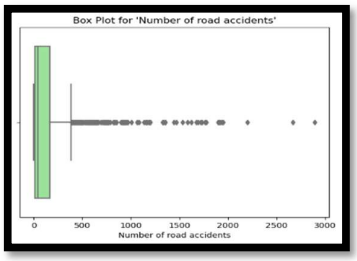
- **INTRODUCTION:**

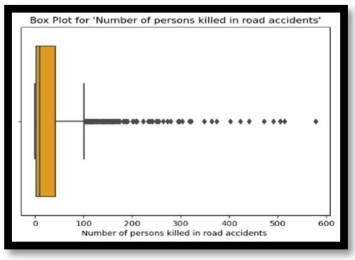
Motivation: This project aims to contribute to the efforts to address the road accident issue in India by utilizing Data Science techniques and methodologies. By analysing the available road accident data, we can uncover hidden trends, identify high-risk areas, and propose data-informed recommendations for improving road safety

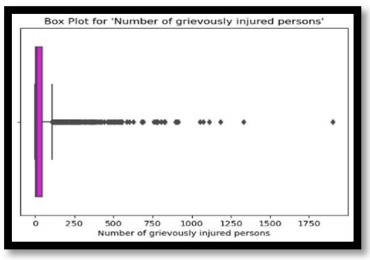
Problem Statement: The objective of this data science project is to analyse and gain insights from road accident data in India, collected across various cities and regions. The primary goals are as follows Identify High-Risk Areas, understand accident patterns, recommend safety improvements

- **DESCRIPTION OF DATASET:** The Dataset used for EDA is "Total Number of Accidents, Number of Persons Killed and Number of Persons Injured in Road Accidents" Dataset.
  This Dataset has been imported from NDAP website, this dataset contains information about road accidents which took place in India during 2019-2021, dataset was uploaded on NDAP on 27th March,23. Till now no work on this dataset has been provided on public platform (Kaggle).
  Dimensions of Dataset: 1350 rows × 9 columns

- **DATA PREPROCESSING:** Data Preprocessing is the Stage where almost 70% of the time is required. It includes cleaning of dataset for further analysis.
  STEP PERFORMED: Missing Values, Outlier detection, Balanced/Imbalanced data, Feature scaling, Feature encoding
  Missing Values: In dataset we had 3 missing values and all the values were missing from one row so we dropped the row containing missing values, there are alternative ways to handle missing values like imputing missing values by mean, median or by applying linear regression but here out of 1340 rows only 1 row had missing values so we dropped that row safely.
  Outlier detection: Box-plot for all above features are showing that almost all points are outlier in dataset, which is not true so box plot failed to capture actual outliers in our use case to find location with high frequency of accidents.







Above is the box plot for all numerical features namely Number of road accidents,Number of person killed in road accidents,Number of person grievously injured persons,Number of minor injured person

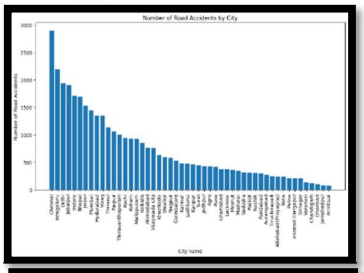
Reason why box plot didn't worked well : Box Plot calculates the Interquartile Range (IQR) which is IQR=Q3−Q1, where Q1 is the median of the lower half of the data. It is the value below which 25% of the data fall, Q3 is the median of the upper half of the data. It is the value below which 75% of the data fall.

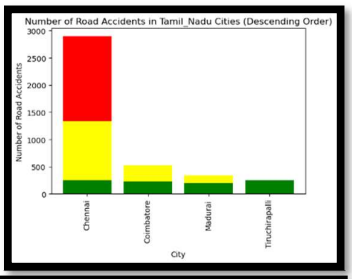
There is high difference in median and mean in features
Mean of Number of road accidents : 154.79
Median of Number of road accidents : 43.0

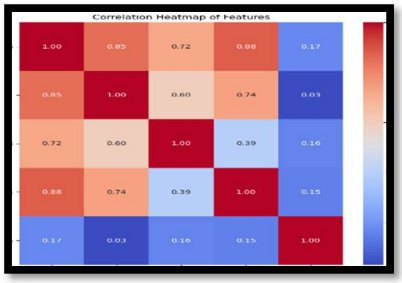
Alternative way for outlier detection So after observing our dataset we realised that declaring the points which has values more than 2x of mean gives location which are highly impacted by road accidents(Red Zone)
In our dataset outliers provide us with very important insight about the areas/cities which are highly impacted by road accidents , so we need to extract more information from outliers so we can't drop them. In further analysis we will continue with outliers

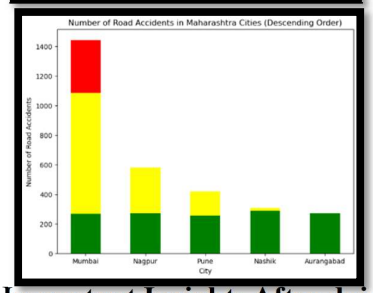
Balanced/Imbalanced data:
Distribution state wise: With respect to state wise distribution the dataset is slightly imbalanced, but from domain knowledge we know that states with larger area and higher population have dense road network therefore some states have more than 10 percent entries
Distribution year wise: In our dataset we have 3 years data i.e., 2019 – 2021, and dataset is equally distributed wrt year feature.
Feature Encoding: We have 3 categorical features in our dataset namely, "State", "City name", "Type of impacting vehicles or objects", We have encoded all above features using label encoder

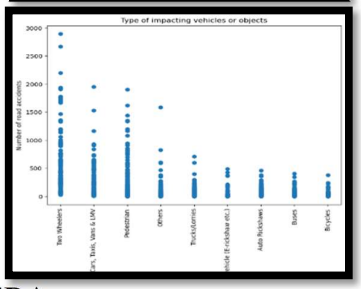
Feature Scaling: We will be using clustering techniques like K-means, K-means++. K-means is sensitive to the scale of features. It treats all features equally and can be influenced by differences in feature scales. Standardization is often recommended. We have used Standard scaler to scale the features
DESCRIPTIVE ANALYSIS: Generate summary statistics, histograms, or box plots to understand the data distributions. You can also Create visualizations (e.g., scatter plots, bar charts, heatmaps) to identify patterns and correlation

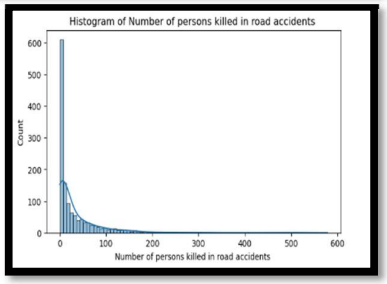












Important Insights After doing EDA:
Feature set A = {"Number of road Accidents" , "Number of person killed in road Accidents", "Number of grievously injured persons" , "Number of minor injured persons"}

Feature set B = {"State", "city name", "type of impacting vehicle"}

1. All feature in set A is highly correlated as we can see in co-relation matrix, reason being that number of road accident is directly proportional to number of persons injured and number of persons killed in the road accident
2. Features in set A follows exponential distribution because there are very few cities with very high number of accident and as we come to less populates and less developed cities where road network is not so strong the number of road accident drops exponentially

Vehicle category which was involves most of the times in cities were two-wheeler and car

Highly impacted vehicle category **set V** = {"two-wheeler", "car", "taxi and cab", "auto-rickshaws"}

RED ZONE CITIES			YELLOW ZONE CITIES		
Cities	Number of road accident	Impacted vehicle set	Cities	Number of road accident	Impacted vehicle set
Chennai	2895	Set V	Hyderabad	1416	Set V
Bengaluru	2197	Set V	Vizag	1403	Set V
Delhi	1935	Set V	Raipur	1034	Set V
Jabalpur	1905	Set V	Ahmedabad	983	Set V

- **CLUSTERING ANALYSIS:**

Objective: Identify cities with a high frequency of road accidents

Using Elbow Method, we found that optimal K i.e., K = 3

K-means++(Fig3.2): Applying K-means++ didn't give satisfactorily results because the shapes of the clusters are not globular. K-means++ assumes that clusters are globular. So, K-Means didn't work here.

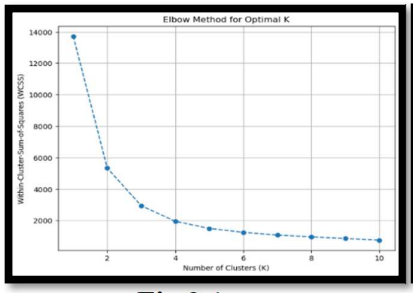


Fig 3.1

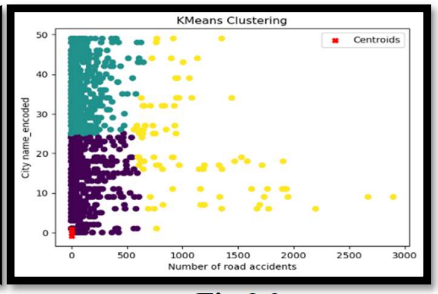


Fig 3.2

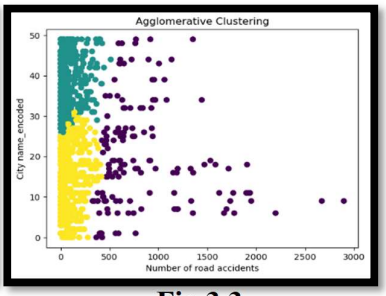


Fig 3.3

Agglomerative Clustering(Fig3.3) : This clustering assumes that each data point is similar enough to the other data points that the data at the starting can be assumed to be clustered in 1 cluster. But here agglomerative clustering was not able to cluster.

GMM(Fig 3.6): GMM didn't worked well because GMM assumes that the underlying data distribution is a mixture of Gaussians but in our dataset the feature follows negative exponential distribution.

GMM After feature transformation(Fig 3.8): In this case we have transformed feature using log transformation to get negative exponential distribution to normal distribution to apply GMM. After log transformation we got slight improvement in Silhouette score but still clusters were not so clear because GMM is sensitive to outliers, and the presence of outliers can significantly impact the estimated parameters of the model, and in our case we have significant number of outliers that we kept in data to analyse them.

DBscan(Fig 3.5): Normal DBscan worked as Kmeans but tweaking the values of eps and MinPts resulted in a cluster with a silhoutee score of 0.689. Density based clustering worked well here because this algorithm works for any arbitrary shape. It also works well with Noise and Outliers. It clusters outliers separately.

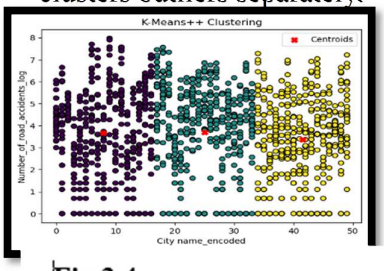


Fig 3.4

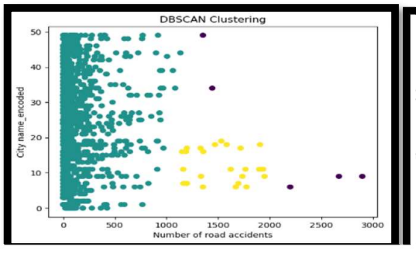


Fig3.5

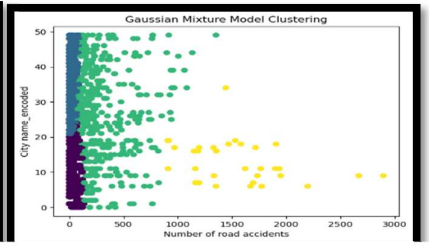


Fig 3.6

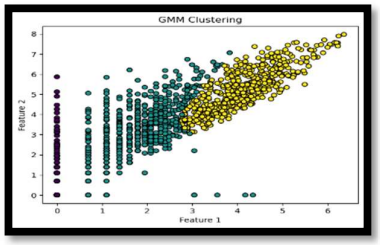


Fig 3.7

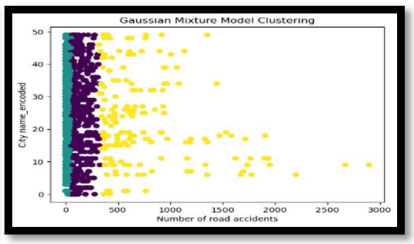


Fig 3.8

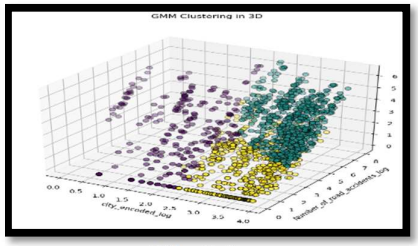


Fig 3.9

Algorithm	K-means++	GMM	GMM with feature transformed	Agglomerative	DBscan
Silhouette score	**0.522**	**0.352**	**0.3739**	**0.500**	**0.689**

Time Complexity based comparison: As or dataset is small so we didn't encounter any problem while running all above clustering algorithms.
Generally Agglomerative clustering can be computationally expensive for large datasets and for small to moderately sized datasets, K-means++ is often fast

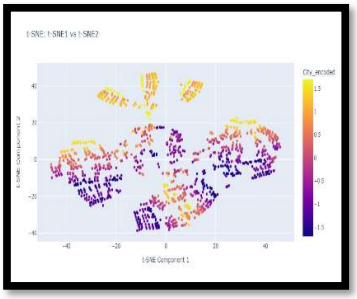
Findings after applying clustering algorithm:
We were able to cluster the cities which are highly impacted by road accidents.

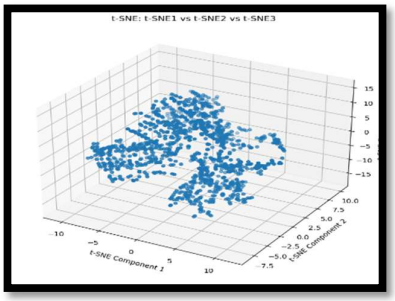
- **DIMENSIONALTIY REDUCTION:**

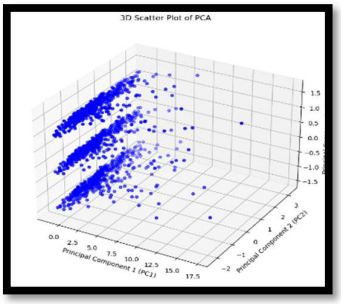
PCA (z-transformation): We standardized the data using z transformation $[z = (x - \mu)/\sigma]$. vis ualization from PCA was not helpful, as PC1 and PC2 captures 57.46 % variance only and in 3-D P C1, PC2, PC3 captures 69.97% variance only.
Reason why PCA didn't performed well: PCA is sensitive to outliers, and a few extreme data poi nts can disproportionately influence the principal components
MDS (Multidimensional Scaling): MDS didn't helped in visualizing data, data points were highly overlapped.
Reason why MDS didn't performed well: MDS is a linear technique. If the underlying structure of data is non-linear, MDS may not be able to capture this complexity effectively, and also MDS is sensitive to outliers, as it aims to preserve pairwise distances. Outliers can disproportionately affect the distances, leading to distortions in the resulting configuration.

t-SNE (t-Distributed Stochastic Neighbour Embedding): t-SNE works for our dataset and we were able to clusters of cities where high, moderate and less number of road accidents takes place and we divided the cities into 3 groups according to risk factor i.e. red zone(cities with number of road accident), yellow zone(cities with moderate number of road accidents) and green zone(cities with least number of road accidents)

Reason why t-SNE performed well: It is a non-linear dimensionality reduction technique and t-SNE may be more robust to global outliers compared to PCA due to its emphasis on local relationships.







Dimensionality Reduction analysis:

- ➔ We tried 2 linear techniques i.e., PCA, MDS (Classical MDS is a linear technique) and one non-linear technique t-SNE, we were able to visualize some important cluster after applying t-SNE.
- ➔ We were able to separate out cities with high number of road accidents but cities moderate and less number of road accident were still overlapping
- ➔ While applying these techniques we didn't had any complexity as our dataset is small but MDS time complexity is O(n^3) which is very high and it might create some problem in large datasets

- **ISSUES/CHALLENGES FACED:**

1. We had to keep outliers because they were the actual cities where accidents occurs very frequently and proper safety measures are required there, most of the algorithms are sensitive to outliers.
2. Most of the numerical features were following negative exponential distribution, due to which it was challenging to form clusters because in small range we have lot of data points.

- **Methods used to outcome challenge:**

1) We performed feature transformation to get normal distribution among numerical features.
2) Used Algorithms which has less affected by outliers like K-means++, this algorithm will get affected by outliers but it gets less affected as compare to K-means.
3) It may not be feasible method in real life, but when Dimensionality Reduction didn't help to visualize and cluster formation, then we came up with a method where to visualize the data, we take 2 relevant columns from the data and see the distribution of those columns whether it can give us some insights. Similarly, we did by selecting 3 relevant columns and visualizing it.

- **RESULT OF ANALYSIS USEFUL FOR INDIAN GOVERNMENT:**

We have categorized the cities in 3 categories i.e., red zone cities, yellow zone cities and green zone cities
We request government to deploy more strict traffic rules in red zone cities as these are cities with high risk for road accidents and casualties, and most often vehicles which are involved in accidents are car, two-wheeler.
We request government to monitor violation of traffic rules in yellow zone cities to prevent accidents
Red Zone cities = {Chennai, Bengaluru, Delhi, Jabalpur, Indore, Bhopal, Jaipur, Mumbai}
Yellow zone cities = {Hyderabad, Vizag, Raipur, Trivandrum, Ahmedabad, Nagpur, Kochi}
Green zone cities are safe

- **CONCLUSION/INSIGHTS GAIN:** The categorized cities, along with their associated risk zone, provide a practical and actionable framework for the Indian government to prioritize and implement targeted road safety interventions. The insights generated from this analysis aim to contribute to more effective accident prevention strategies and enhanced road safety nationwide.

- **REFERENCES:**

[1] https://ndap.niti.gov.in/dataset/6606?tab=data

[2] https://cs.wmich.edu/alfuqaha/summer14/cs6530/lectures/ClusteringAnalysis.pdf

[3] https://towardsdatascience.com/use-this-clustering-method-if-you-have-many-outliers-5c99b4cd380d

[4] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html