Analysis of Road Accidents in UK

PROJECT REPORT

Fall 2021-2022

Submitted by

Sanjana Rajeshwar ***(20BCE0782)***
Arnav Srivastava ***(20BCE0831)***
Sri Shreya Chinamilli ***(20BCE0816)***
Somya Rathi ***(20BCE2323)***

Under the Guidance of
Pravat Kumar Jena
(Assistant Professor Sr.)

<<Computer Science and Engineering >>

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Vellore-632014, Tamil Nadu, India

School of Computer Science and Engineering

November 2021

Index

Project at a Glance

Language Used	Python
Libraries Used	1. Pandas 2. Matplotlib 3. Numpy 4. Seaborn 5. Pywaffle 6. Plotly.express
Number of Plots	20

INTRODUCTION

Road traffic accidents are very common these days and can happen anywhere and at any time. And so, it is the responsibility of each and every one to avoid these traffic accidents and keep everyone safe. And in order to give safe driving instructions, careful analysis of roadway traffic data is critical to find out variables that are closely related to fatal accidents. In this project will be applying data visualization and statistical analysis algorithms on the datasets taken from the official UK government website as an attempt to address this problem. The trend in RTA injuries and death is becoming alarming in countries. The number of fatal and disabling road accidents happening is increasing day by day and is a real public health challenge for all the concerned agencies to prevent it.

OBJECTIVE

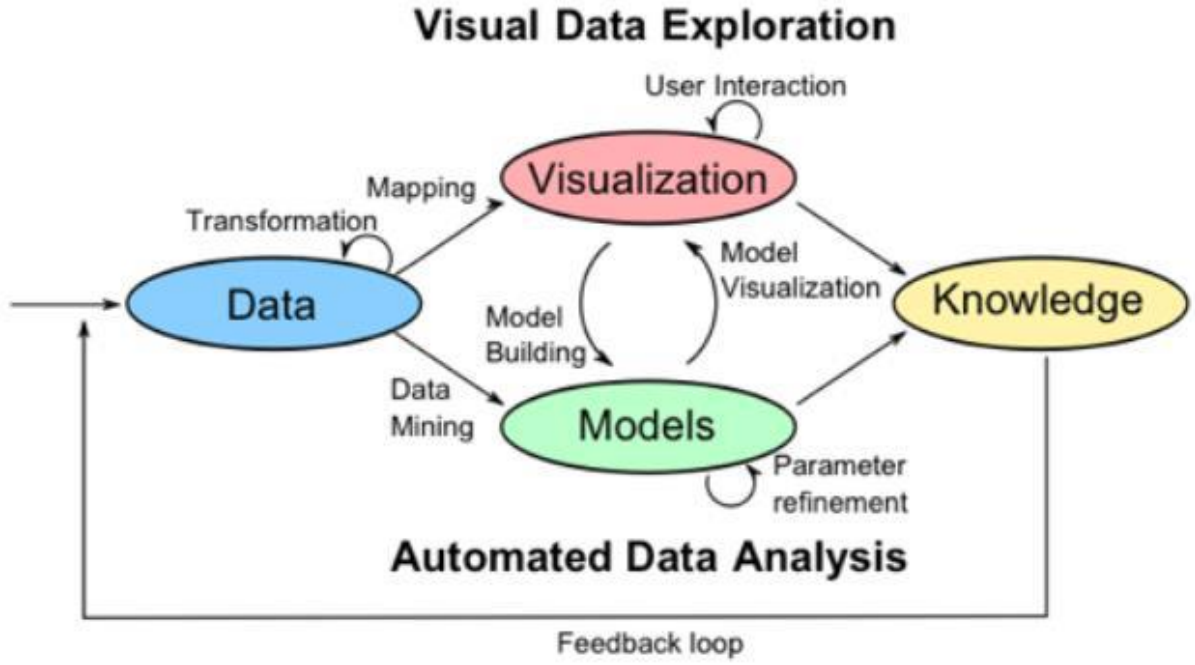
All of us very well know that the road traffic accidents have emerged as a major public health hazard with a number of RTA's taking place every year and so it is our responsibility to derive a suitable relationship between the accidents and the external factors that cause them so we can reduce the number of RTA's and loss of life in the long run. Few factors that cause RTAs include the road and climate conditions, over speeding, distracted driving and jumping the red signal. So, the main objective of this project is to analyse the various road traffic accidents that had taken place in the United Kingdom from the year 2005 to 2014 from the dataset. In this project we will deploy various data visualization techniques and methodologies to find the pattern and insights from the dataset used. Several modelling techniques are used for data analysis in doing so to find out how to drive safely. From this project we will also know the parameters that cause traffic accidents. We can also get a lot of driving suggestions as well and control them in the long run.

PROPOSED METHODOLOGY

The designed proposed methodology is intended to be portable to any visualization challenge; it presents a sequence of important analytical and design tasks and decisions that need to be handled effectively. Visualizing data is a complex process that contains complex problems and efficient methods of interpretation with greater efficiency, effectiveness and elegance. Adopting this methodology is about recognizing the key stages, considerations, and tactics that will help you navigate smoothly through the visualization project

In data analysis the analysis is 100% rigid is a bold statement to make rather a linear process and indeed some of the stages may occasionally switch in sequence and require iteration. That is new factors and branches to data can emerge at any stage and may affect the visualization strategy and solutions to it. so, it is important to keep the visualization technique flexible at all levels.

In data visualization one can rarely, if ever, say that a problem can have a single answer or a single best possible answer. It is much more about heuristic methods to determine the most satisfactory solutions.



INTERPRETING FLOWCHART:

The Visual Analytics Process combines automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data. The figure shows an abstract overview of the flow in between different stages and transitions in the data visualization process. The first step is often to pre-process and transform the data to derive different representations for further exploration. Other typical pre-processing tasks include data cleaning, normalization, grouping, or integration of heterogeneous data sources. After the transformation, the analyst may choose between applying visual or automatic analysis methods.

If an automated analysis is used first; data mining methods are applied to generate models of the original data. Once the pre model is declared and created it is important to evaluate and refine models which best done by interacting with the data with the questions what (what data the user see?), why (why user intent to use visualization tools?), how (how visual encoding and interaction idioms are constructed?). In brief we can conclude that Visual Analytics Process knowledge can be gained from visualization, automatic analysis, as well as the preceding interactions between visualizations, models, and the human analysts. Hence proposed methodology is an important technique for clear data visualization and on point analysis of the data with better understanding and efficient knowledge gained in a time bound manner. Else there is infinite data and infinite ways to interpret it but very few methods are there which efficiently channelize the purpose of studying the particular data in a precise manner.

DATA DESCRIPTION

The dataset used in this project is taken directly from the government of UK's official website. The dataset can be located at

https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

The dataset consists of 3 .csv files, namely

1. Accidents0514.csv
2. Casualties0514.csv
3. Vehicles0514.csv

These files provide us with detailed road safety data across Great Britain from 2005 to 2014. This dataset helps in understanding the circumstances of personal injury road accidents, detailed analysis of all casualties and their types along with the different types of vehicles involved in the consequential casualties.

All the variables in this dataset are coded rather than actual text strings. We can find all these coded references in a separate file named variable_lookup.xlsx.

Following are the attributes of all these files along with their semantics: -

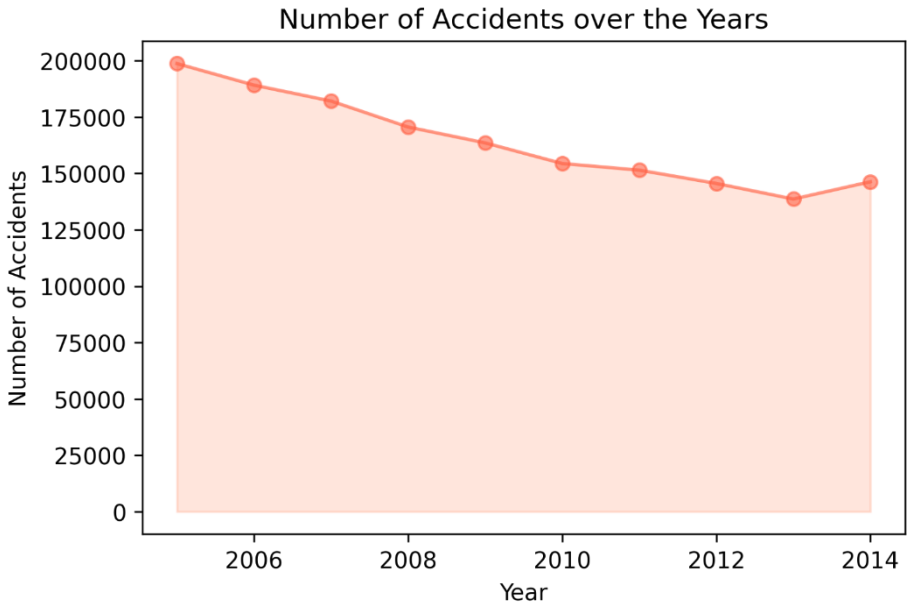
Accidents0514	
Categorical Attributes	
Accident Index	Index of all the accidents according to the year and month.
Accident Severity	Severity of accident: fatal, serious or slight
Road Type	Type of road, what the road is made up of
Light Conditions	Intensity of brightness
Weather Conditions	Description of wind, rain etc.
Road Surface Conditions	Differentiated by dry,wet,oil etc.
Urban or Rural Area	Accident in which area
Did Police officer attend Scene of accident	Attended/ Attended but not reported/ Not Attended
Ordinal Attributes	
Date	DD/MM/YYYY Date of accident
Day of Week	Day of the accident
Time	HH:MM Time of the accident
Quantitative Attributes	
Number of Vehicles	Number of vehicles in accident
Number of Casualties	Number of casualties in accident
Speed Limit	Speed Limit where the accident occured

Casualties0514	
Categorical Attributes	
Accident Index	Index of all the accidents according to the year and month.
Casualty Class	Driver/Passenger/Pedestrian
Sex of Casualty	Male/Female/Others
Casualty Severity	Severity of casualty: fatal, serious or slight
Pedestrian Location	Location of pedestrian when accident happened
Casualty Type	Casualty's Situation
Ordinal Attributes	
Car Passenger	Position of Passenger in the Car
Bus or Coach Passenger	Position of Passenger in the Bus
Quantitative Attributes	
Age of Casualty	Age of the casualty

Vehicles0514	
Categorical Attributes	
Accident Index	Index of all the accidents according to the year and month.
Vehicle Type	Type of vehicle in the accident
Vehicle Manoeuvre	What the vehicle was doing before the accident, reversing/parked/U-turn/turning right etc
Vehicle Location	The lane in which vehicle was moving or parked
Hit Object in Carriageway	Part of the vehicle which hits
Hit Object off Carriageway	Part of the vehicle which gets hit
1st point of Impact	Region of the vehicle hit first
Journey Purpose	Reason why the vehicle was travelling
Vehicle Propulsion	Fuel on which the vehicle runs
Quantitative Attributes	
Age of Vehicle	How old the vehicle was
Age of Driver	Age of the driver in the vehicle

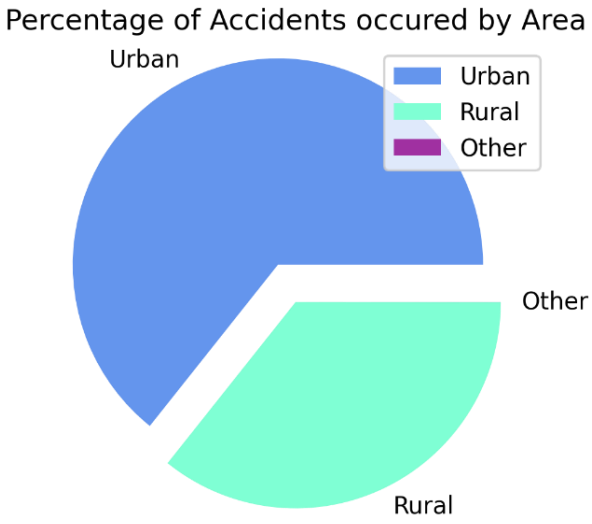
INTERPRETATION OF THE PLOTS

➢ Relationship Between Number of Accidents and Years

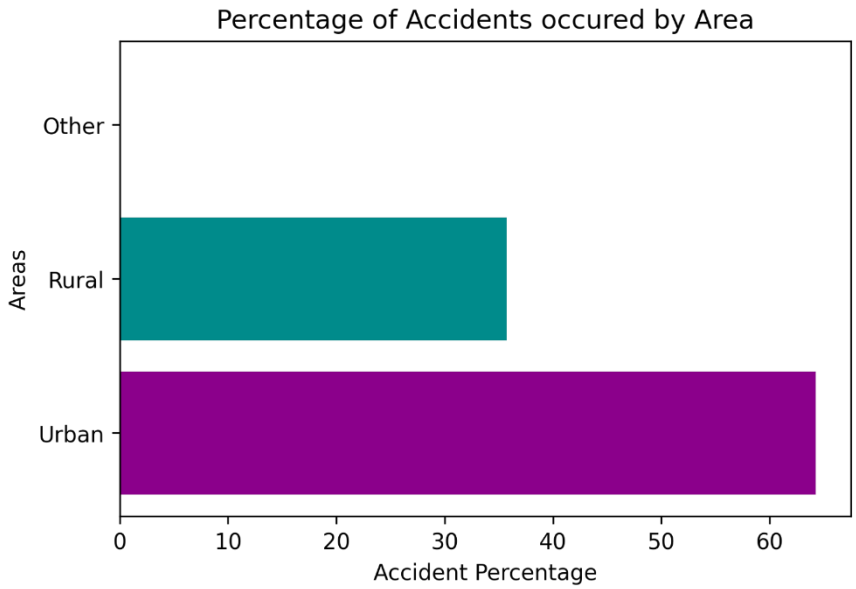


- A **line chart** or **line plot** or **line graph** or **curve chart** is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments.
- It is quite visible from the **line chart** that the number of cases underwent a decrease from 2005 to 2013 and had a slight rise in the year 2014.
- The year **2005** had the **maximum** number of cases whilst **2013** had the **least** number of accidents.

➢ Interconnection Between Accident Percentage and Area

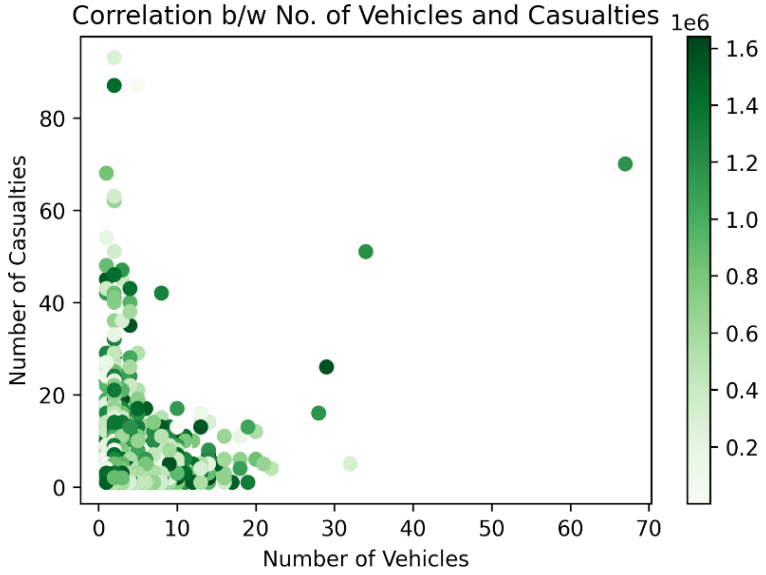


- A **pie chart** is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice is proportional to the quantity it represents.



- A **horizontal bar chart** is a graph in the form of rectangular bars. The bar chart title indicates which data is represented. The vertical axis represents the categories being compared, while the horizontal axis represents a value. This type of chart provides a visual representation of **categorical** data.
- Over **60%** of accidents took place in **urban** areas.
- **Rural** areas are less prone to accidents as compared to accidents in **urban** areas.

➢ Correlation Between Number of Vehicles and Casualties

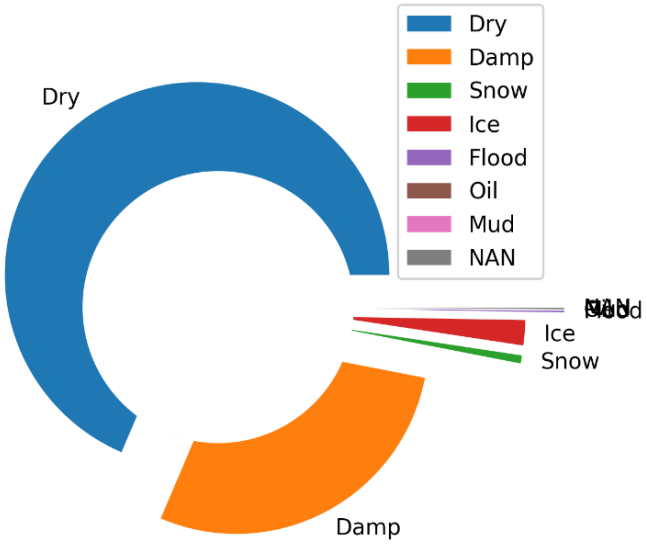


- A **scatter plot** (also called a **scatter graph**, **scatter chart**, or **scatter diagram**) is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two

variables for a set of data. If the points are coded (colour in this case), one additional variable can be displayed.

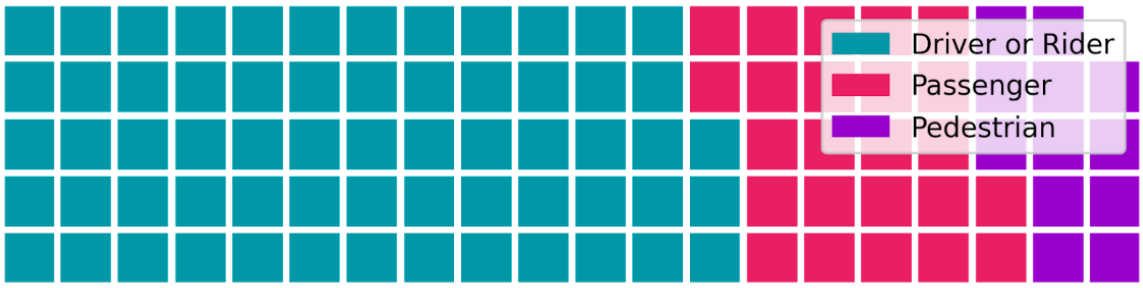
- It is clear from the trend that there is no particular correlation between number of vehicles and casualties.
- It is also evident that there is no possible cluster for any value of colour. The reason being the vast range of values, hence clustering becomes difficult.

➢ Relation Between Number of Accidents and Road Condition



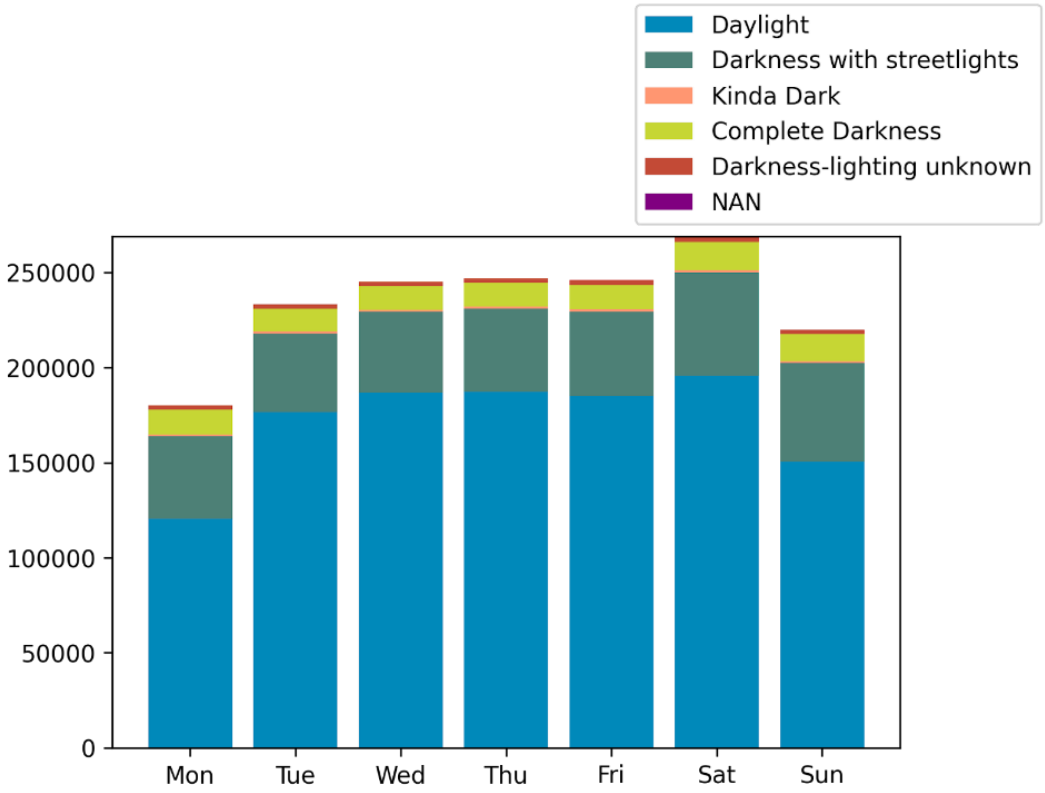
- A **doughnut chart** (also spelled **donut**) is a variant of the pie chart, with a blank centre allowing for additional information about the data as a whole to be included. This type of circular graph can support multiple statistics at once and it provides a better data intensity ratio to standard pie charts.
- A major percentage of accidents took place on roads which were **dry**. This was followed by roads which were slightly wet(**damp**).
- Road surfaces with both **Ice** and **Snow** had a significant number of accidents, the reason being the weather conditions in the **United Kingdom**. It mostly has a winter season and experiences snowfall most of the time.
- The number of accidents on roads with **flood, oil** and **mud** are negligible as compared to the others.

➢ Relation Between Casualty Type and Number



- A **waffle chart** shows progress towards a target or a completion percentage. Waffle Charts are mainly used when composing parts of a whole, or when comparing progress against a goal.
- **Driver or Rider** has the greatest number of squares which implies that they got injured in most of the cases.
- They are followed by **passengers** and then **pedestrian**s.
- It is clear that the number of **driver or rider** casualties is way **more** than **passenger** and **pedestrian** combined.

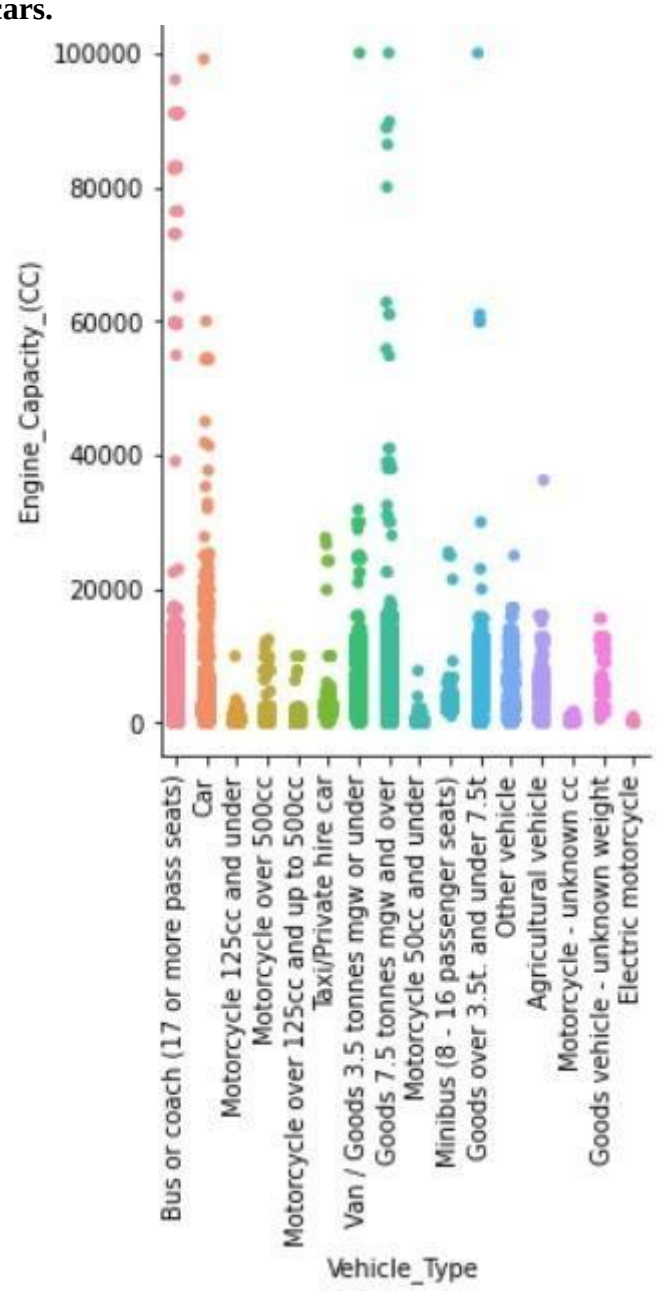
➢ Relationship Between Number of Accidents, Days and Light Conditions



- A **stacked bar chart**, stacks bars on top of each other so that the height of the resulting stack shows the combined result. Stacked bar charts are not suited to data sets having both positive and negative values. Stacked bar charts present the information in the same sequence on each bar.
- Most of the accidents took place in broad **daylight** followed by **darkness with streetlights**.
- **Maximum** number of accidents took place on **Saturday** over the years.
- **Monday** had the **least** number of accidents from 2005 to 2014 in the United Kingdom.
- **Wednesday**, **Thursday** and **Friday** had almost the **same** number of accidents in different light conditions.
- It is evident **most** of the accidents took place in conditions with **proper lighting**. There were a few accidents in areas which had **darkness**.

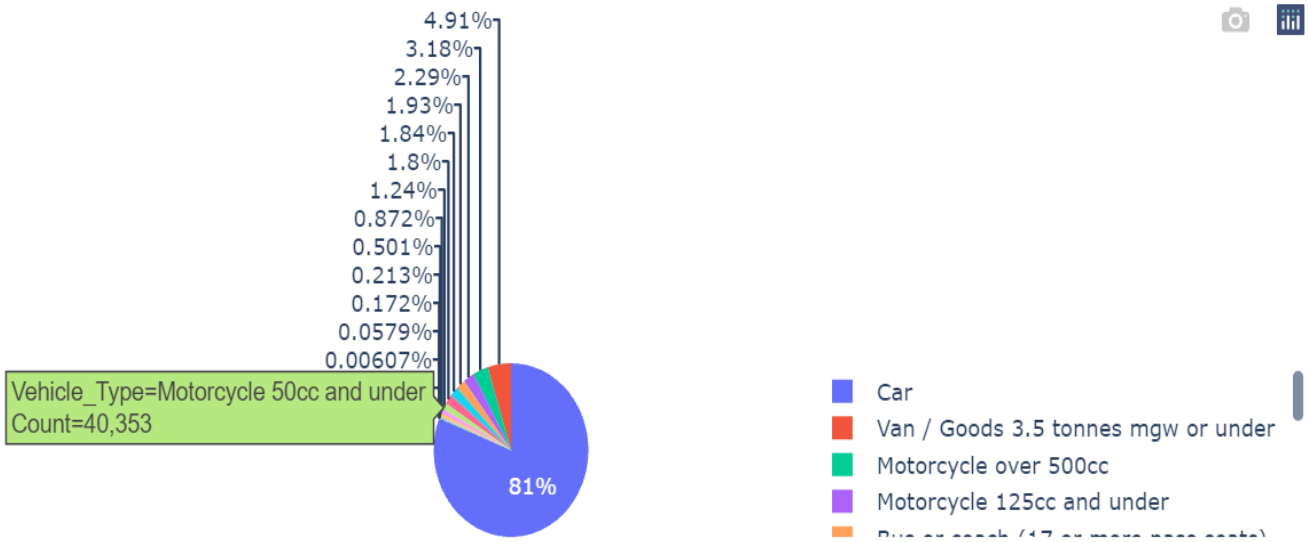
➢ Relationship Between Number of Accidents and Vehicle Type

- The interpretation of the following three plots below is the same, but we used different types of plots for finding the vehicle type which has been involved in most accidents.
- The plot below is a **cat plot.**
- Cat plots are used to show the relationship between **numerical variables and one or more categorical variables**, like boxplot, strip plot and so on.
- From the below cat plot, we can conclude that the vehicle that is **most involved in accidents is cars.**

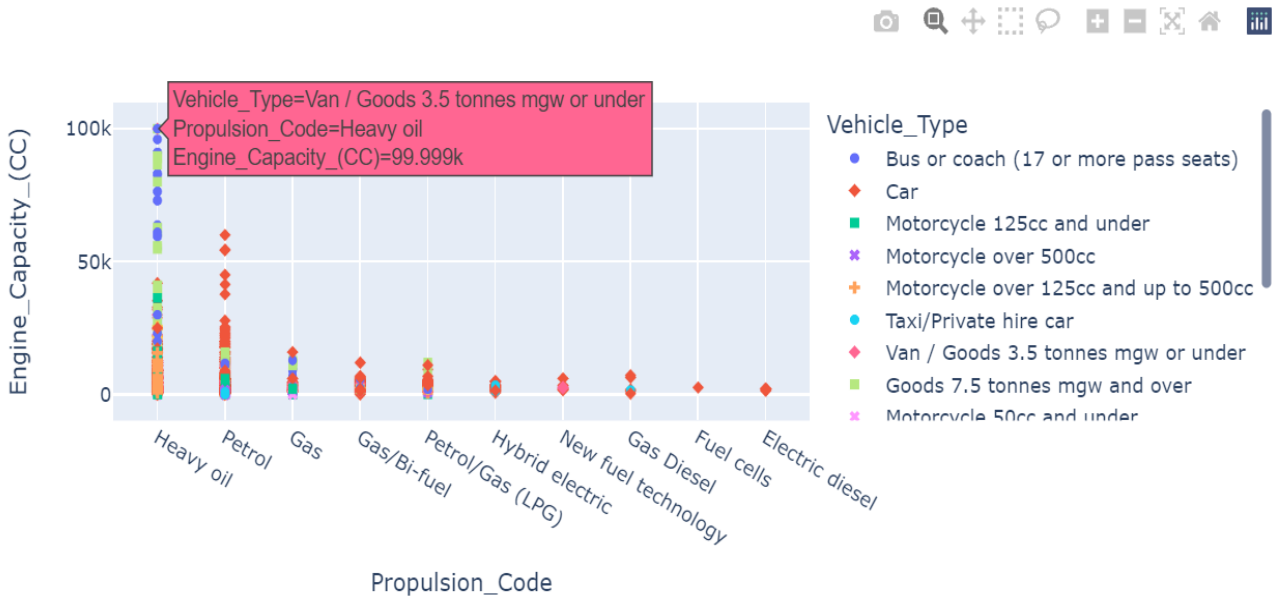


- The plot below is an **interactive pie chart**. These plots are used to show percentages of a whole, and **represent percentages at a set point in time**.

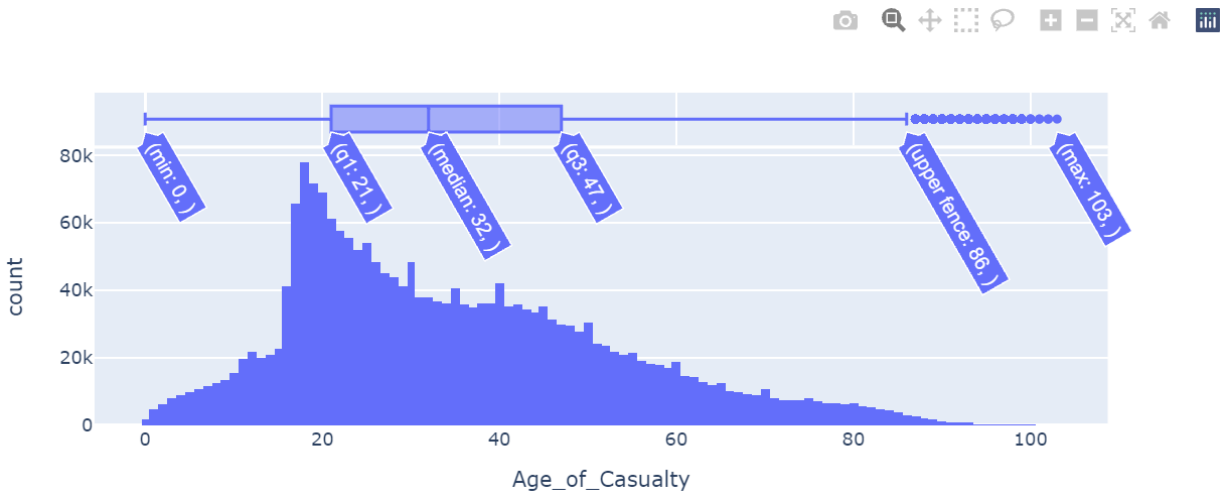
- This plot shows the **percentage of different vehicle types** that are involved in accidents.
- From the below interactive pie plot we can conclude that the vehicle type **car is involved in the greatest number of accidents (81%) followed by vans (4.91%).**



- This graph below is an **interactive scatter plot.** This type of graph uses **dots to represent values** for two different numeric variables.
- This graph gives the relationship between the **engine capacity** and **propulsion of various vehicle types** from which we can know what vehicle type is involved in **a greater number of accidents.**
- From this graph we can interrupt that the **car** which was the **most involved vehicle** in accidents works on all propulsions, followed by **bus/coach** works on **petrol, heavy oil and gas** and has the highest engine capacity.

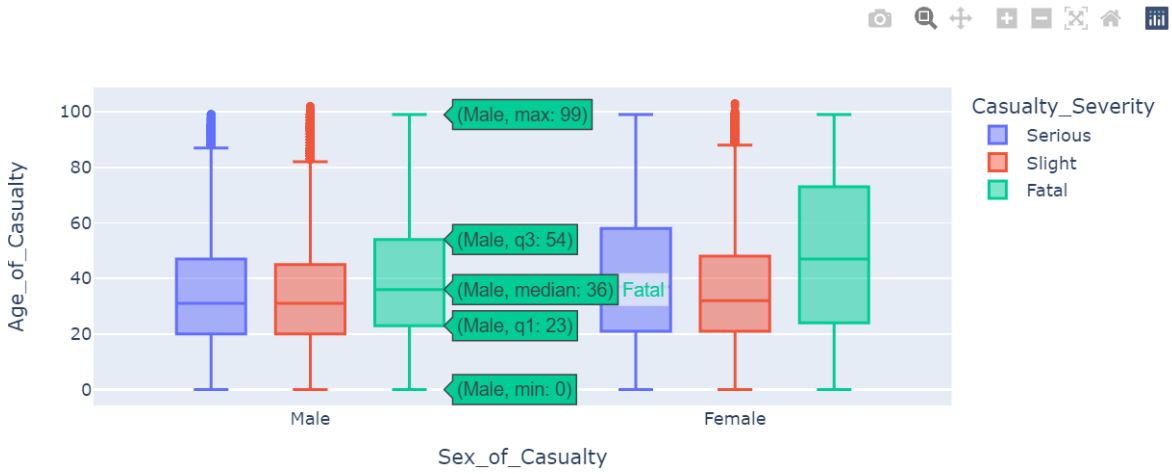


➢ Relationship Between Number of Accidents and Casualty Age



- This plot is an **interactive histogram plot with boxplots**. A histogram represents the **frequency distribution by means of rectangles** whose widths represent the class intervals and whose areas are proportional to the corresponding frequencies and box plot is a standardized way of representing statistical **five number summary** of a dataset which includes **minimum,1st quartile, median (2nd quartile), third quartile and maximum.**
- This graph shows the relation between the **age of the person** involved in casualty and the **count of the accidents**.
- From the above given graph, we can conclude that people aged from **18 to 20 years** have been involved in **more accidents** when compared to the other age groups.

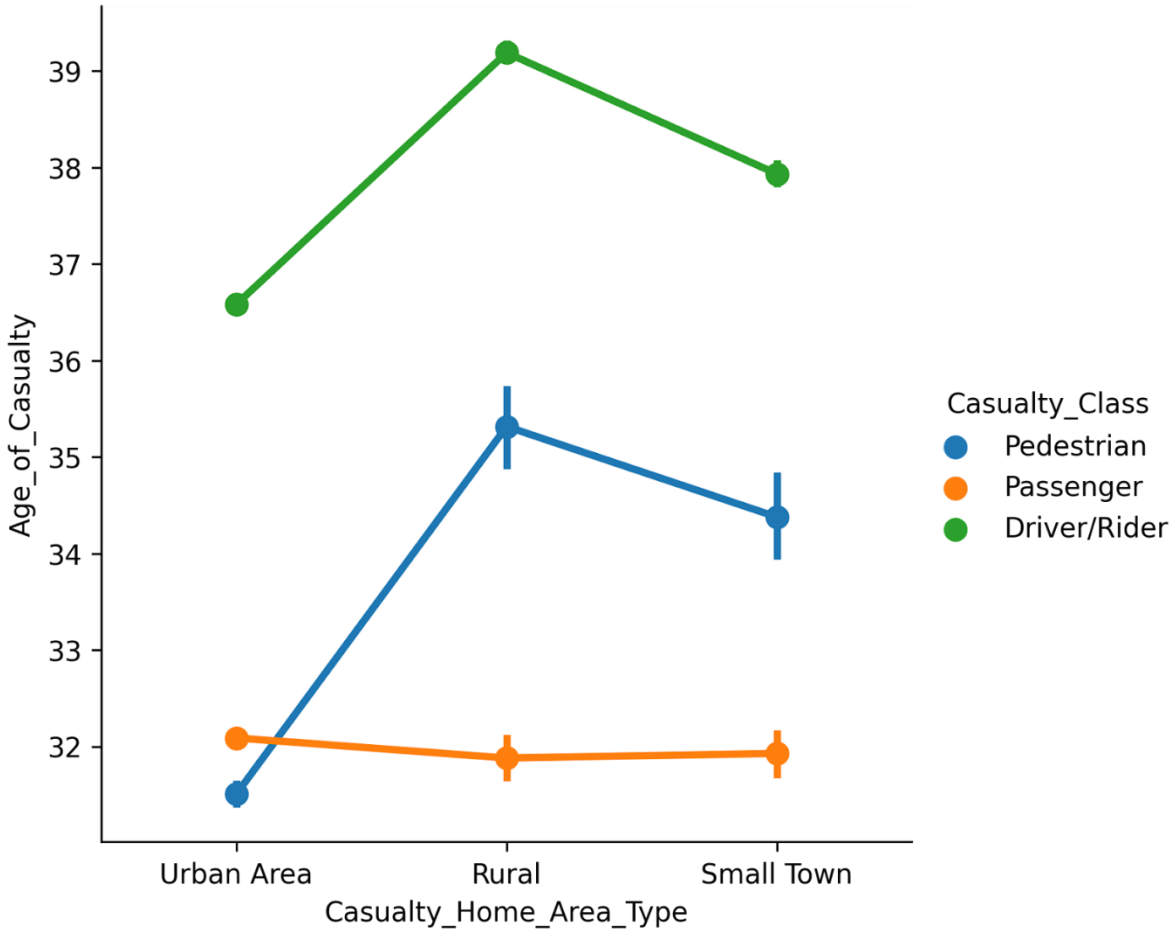
➢ Relationship Between Casualty Sex and Casualty Age



- These graphs shown above are interactive **box plots**. These plots are the visual representations of a **statistical five number** summary of a dataset which includes **minimum,1st quartile, median (2nd quartile), third quartile and maximum**.

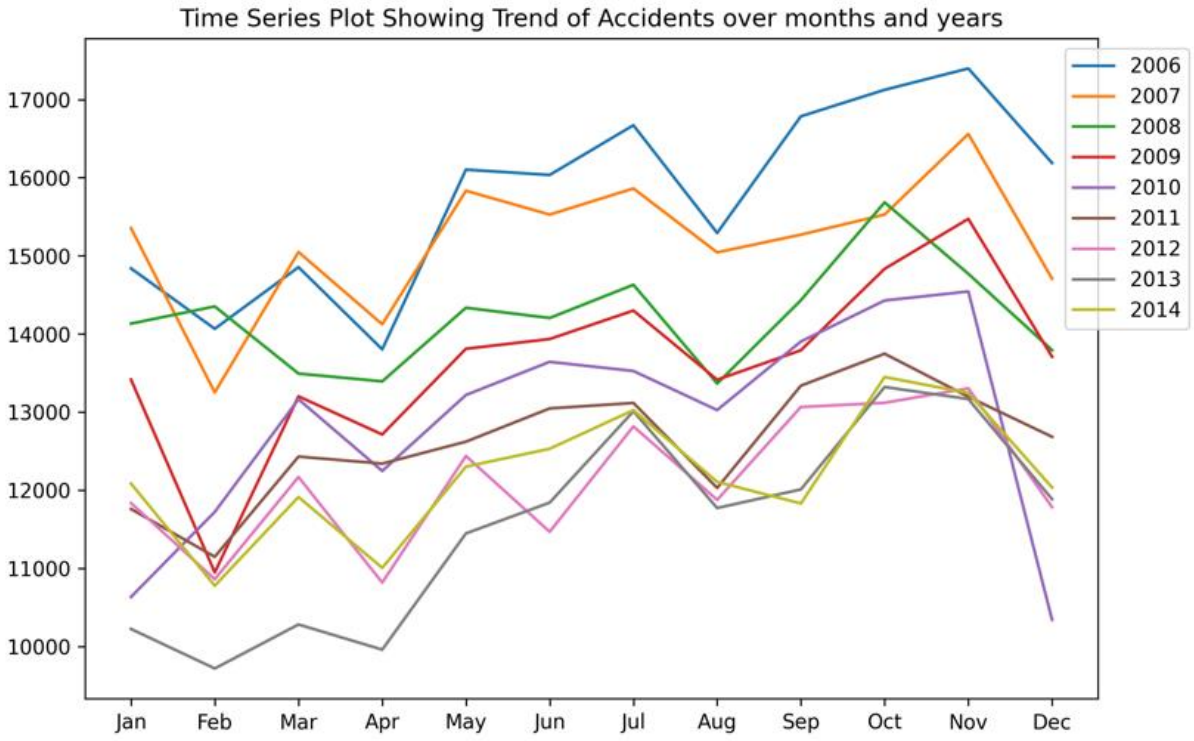
- In this graph x axis represents the **sex of the person**, y axis represents the **age of the person** who were involved in the casualty and hue represents the **severity of the casualty.**
- From the above plot we can interrupt that the **females** who were aged between **24 and 73** have been involved in **fatal accidents**
- **Males** who were aged between **21 and 57** have been involved in **fatal accidents** compared to females and males of other age groups.

➢ Relationship Between Casualty Home Area and Casualty Age



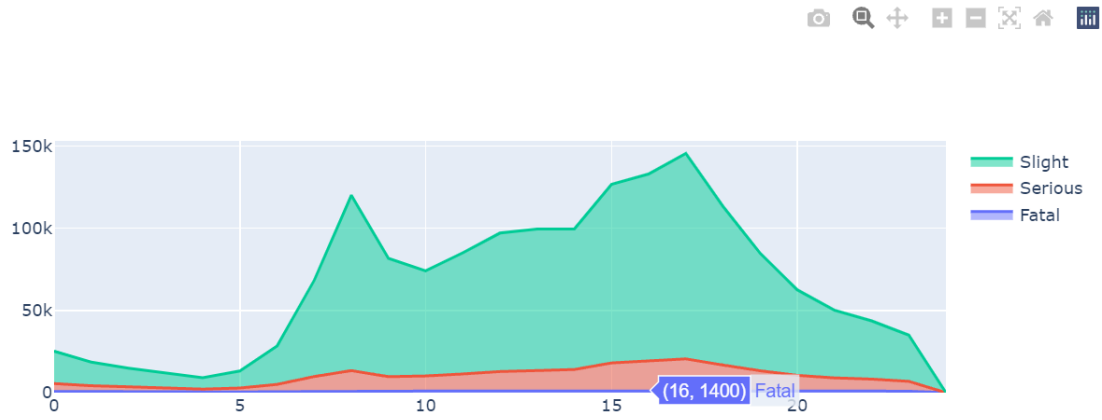
- The plot above is a **point plot**. A point plot represents an estimate of central tendency for a numeric variable by the position of scatter plot points and provides some **indication of the uncertainty** around that estimate using error bars.
- From the above point plot, we can conclude that the **age of drivers/riders** who are involved in the accident casualties are **older compared to the passengers and the pedestrians** involved.
- **Mean age** of pedestrians who were involved in casualties was between **35 and 36 years.**

➢ Relation showing trends of accident over months and years



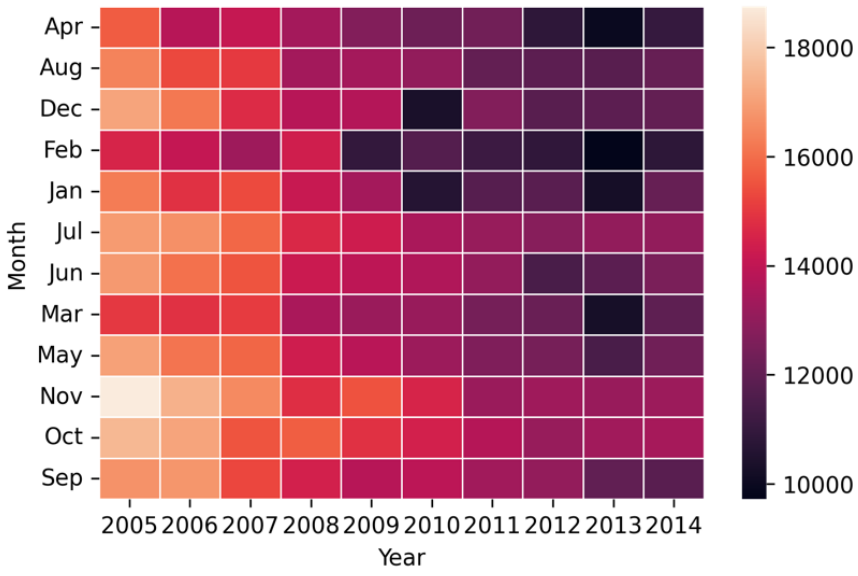
- The above graph is **a time series graph**, it's a data visualization visual analytics tool that illustrates data points at successive intervals of time. Each point corresponds to both a time and quantity that is been compared with base considering a bigger time span
- The above graph has comparing time span as months from **January to December** (X axis) and quantity that is been compared as **accidents** (Y axis) and **base as years** which are depicted by different colors line sketched.
- Here the relation between months and years are shown for the years 2006-2014 and each year is depicted in **different color** to give a clearer and wide idea to person interpreting the graph.
- For example: in 2006(blue) there is a rise in number of accidents in March, May, July, September, October, November and there is fall in accidents for February, April, June, august, and December.
- Since the number of accidents have a **greater scale** so minor changes of accidents in months cannot be detected properly that can give certain **flaws** in analyzing then too this graph is good for analyzing a dataset and get idea of data.

➢ Relationship Between Number of Accidents and Time of Accident



- The graph above is an **interactive stacked area chart.** The interactive stacked area chart displays the **evolution of several groups on the same graphic.**
- In the above graph - x axis represents the **time** at which the accidents had taken place, y axis represents the **number of accidents** and hue represents the **severity of the accidents.**
- From the above graph we can interpret that the **majority of the severe accidents** have taken place from the timings **2pm to 6pm and 5am to 9am** as we can observe that the graph is rising during those timings.

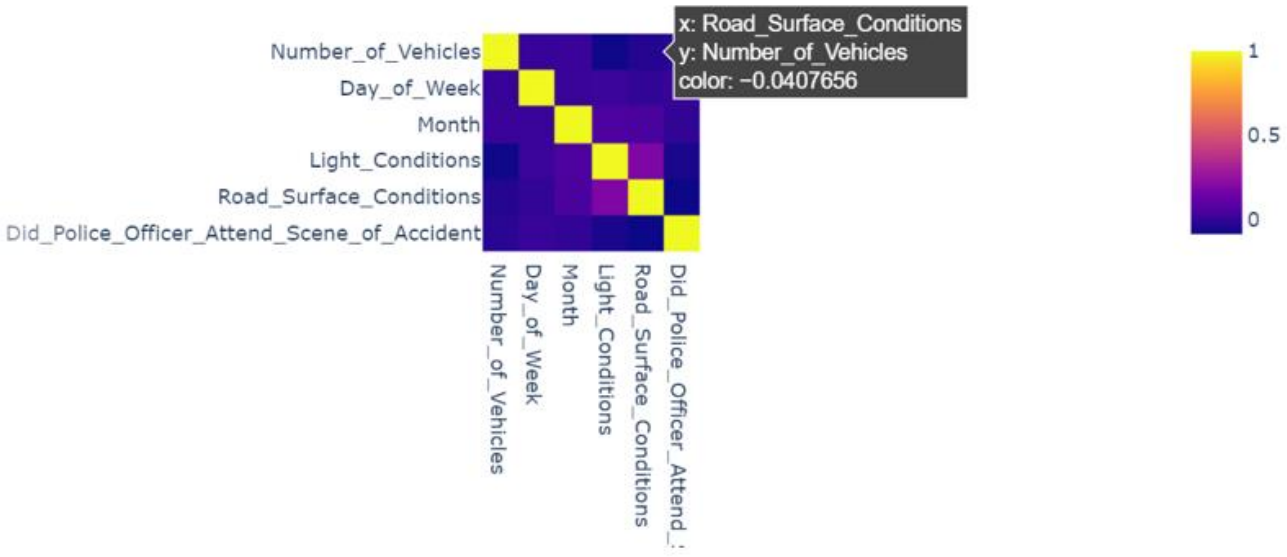
➢ Distribution of Number of Accidents from 2005 to 2014 with respect to months



- The above graph is a **calendar heat map** which is variation of a traditional heat map where data is laid-out on grids.
- The dataset interpretation is provided in a **12*10 grid** which specifically have **12 rows** depicting months and **10 columns** depicting years (2005-2014).
- This graph has a color key that depicts color scheme for number of accidents, such that the accidents reduce from color **peach to dark magenta** as peach color depicts a greater number of accidents (18000) and dark magenta represents a smaller number of accidents (10000) these are the 2 extreme colors of color key in between peach and dark magenta there are many shades that depicts progressive decrease in number of accidents with respect to colors

  (18000→16000→14000→12000→10000 reduction of 2000 accidents every time, color changes as a matter of fact)
- So, the color affiliated in the rectangular portions depicts number of accidents in particular year for example- there are minimum accidents in February 2013 since the small grid color is dark magenta so accidents are near around 10000
- Similarly, the accidents are high in December 2005 as the small grid color is peach so color key line suggests that numbers of accidents are closer to 18000, similarly others. We can also say that this graphic analytics is an alternate visualization to analyze time series data.

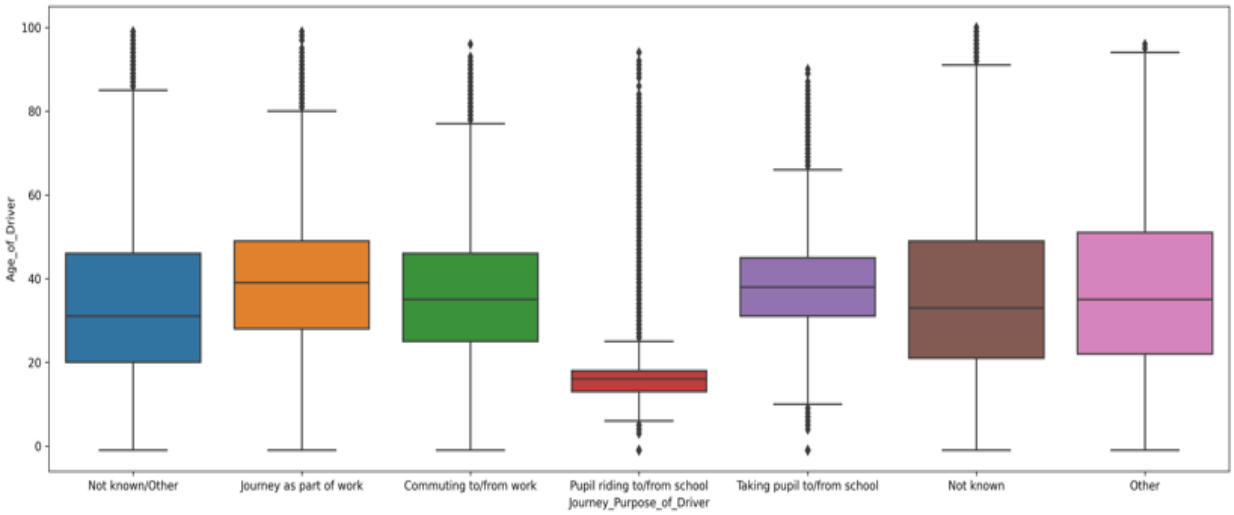
➢ Correlation graph between various attributes of Accidents Dataset



- The above graph is a **color heat map** which is showing correlation between various dataset attributes with the correlation range we can interpret that attribute is correlated to another in some way weather

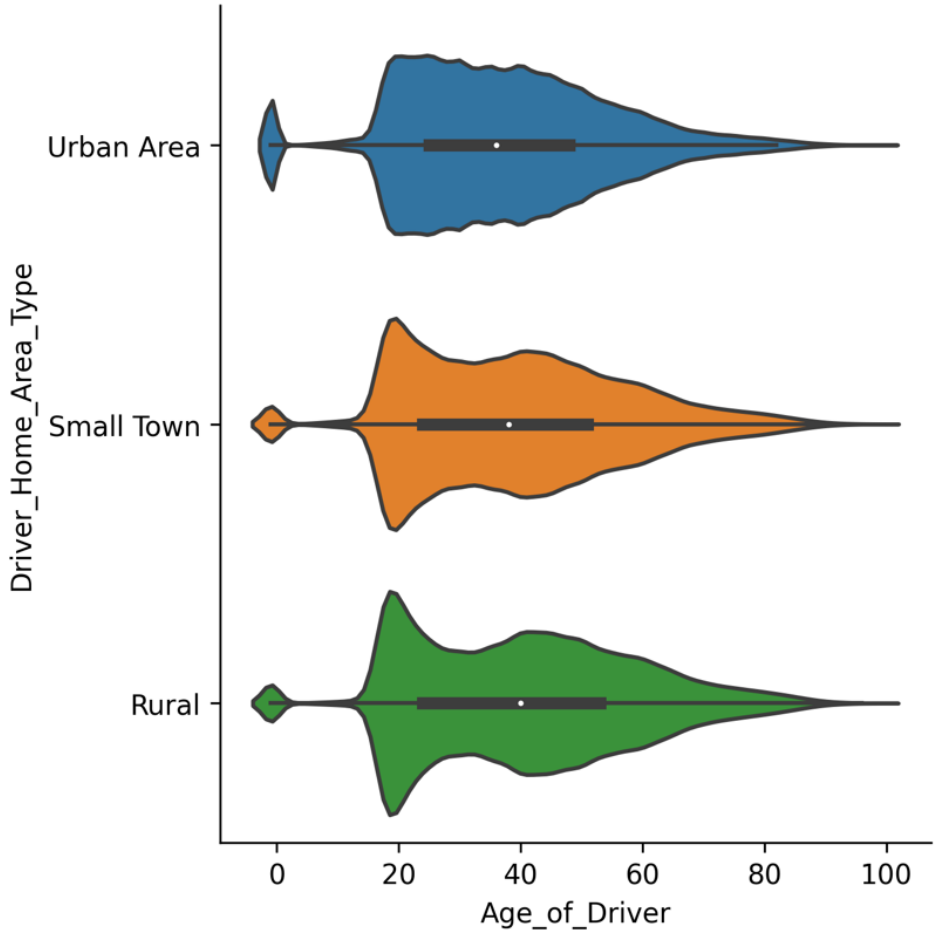
- The correlation is weak or strong (0-1) that means there is **only positive correlation** between the attributes, or there is 0 correlation between the attributes but there is no negative correlation.
- Color **yellow** showing **strongest positive correlation** (1) and **dark blue** showing **weakest or no correlation** between the attributes (0).
- For example-number_of_vehicles are self-correlated to itself hence the correlation is strong correlation of 1 which depicts perfect correlation but number_of_vehicles does not have any relation with day_of_week so the correlation is 0. Correlation between road surface conditions and light conditions is about 0.2 which is depicted by light magenta. Similarly, that others also show correlation between 0 to 1 depending upon their interdependency.

➢ Relation Between Age of the Driver and Journey Purpose of Driver



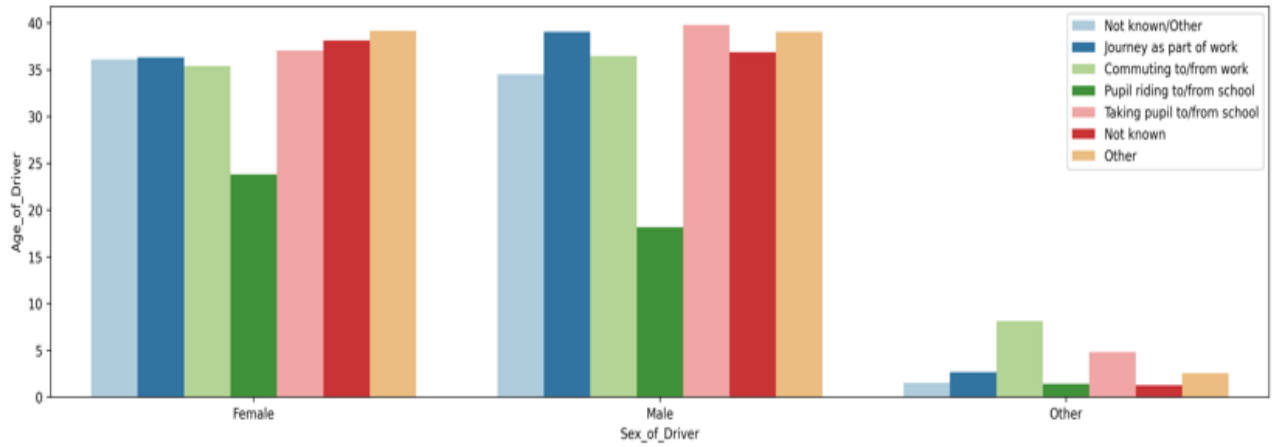
- Above shown is a **box plot** which is a method for graphically depicting groups of numerical data through their quartile. It may have lines extending from boxes indicating variability outside the upper and lower quartiles. It also shows the skewness through displaying the data quartiles (percentiles) and averages.
- The line inside the box depicts median values which suggest that 50% of data is above the median and 50% is below but within the box itself we have 25% data above the median and 25% below that inform us that a total of 50% data is in the box plot also have outliers beyond the whiskers.
- Example – if we consider age of drivers' vs purpose of travelling, we can conclude that **most of the drivers aged between 18-20 were travelling from/to school** with a median of around 15

➢ Relation Between Driver Home Area and Age of Driver



- Above graphic representation is called a **violin plot**, it's a method of plotting numeric data and analyzing it its similar to box plot but the only difference is it have rotated kernel density plotted to each side and shows probability density of different data at different values
- The **white dot** represents the **median** value of drivers age which is 40 left to it is first quartile with the tip as lower adjacent value and to the right it is third quartile further followed by upper adjacent values and outside points
- The shape of distribution indicates that the **ages of driver of small town and rural are highly concentrated in quartile 1**, i.e., the probability of accidents caused by drivers aged from 20-40 is high, and the probability of those aged 40- 70 and above is low (40 as median value).

➢ Relation Between the Driver Age and Driver Sex by Several Purposes



- The above graphic representation is the **group bar chart** that analyze the dataset comparatively, here we are comparing male and female drivers on various parameters indicated by different colors which are: -

  (Light blue) Not known/others

  (Blue) Journey as part of work

  (Dark green) Pupil riding to/from work

  (Pink) Taking pupil to/from school

  (Red) Not known

  (Yellow) others

- For example: **with respect to age in male journey "as a part of work" is more than in females, but "pupil riding to/from school" has more female population.**

Link for Code File and Datasets

Note: - Please download the code file with .html extension to view the code along with the outputs.

Google Drive Link: -
https://drive.google.com/drive/folders/1RjZNwa_6BbLqKnc5LoNVEVFkYGluFQjW?usp=sharing