1. *Visual Data Analysis.* Given the dataset "task1-dataset.ods" (available in TeachCenter), which comprises a number of features. Provide a number of meaningful visualisations (4 visualisations) that show key properties of the dataset and dependencies. Based on the visualisations provide your interpretation and insights.

   (a) What pre-processing did your do? (e.g., Did you create new features? Did you normalise the data? Did you filter the dataset? Extended with another dataset?)

   (b) What are the most relevant dependencies between the features (selection of the figures)?

   (c) Provide a series of meaningful plots that show a specific relationship (dependency) or characteristic of the dataset

   (d) Provide a summary of the main insights

   Answer (a) - Preprocessing steps:

   - Handling Missing Values (Removed columns having too many missing values and removed outliers)
   - Filter the dataset and select necessary features
   - Rename and Clean Columns (Made column names more readable)
   - Convert Data types
   - Create new features using existence features

   Answer (b) - List of main dependencies:

   - The Reported Fatalities by Countries
   - Fatality rate by Income group
   - Safety Laws and Fatalities
   - Speed Limits and Enforcement

   Answer (b) and (c)

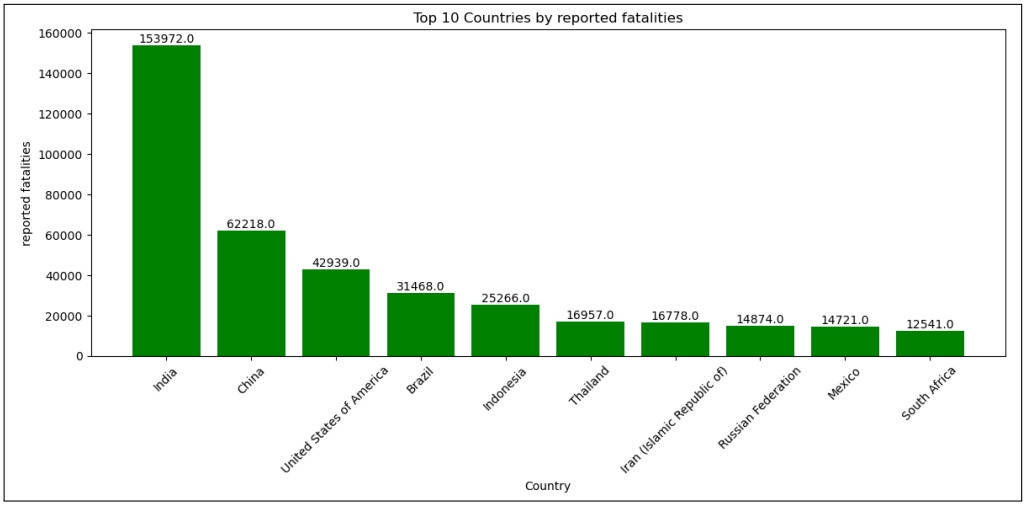


Figure 1: Top 10 countries by reported fatalities - This bar chart is perfect for study reported fatality rate across countries and seeing how each country stacks up in terms of reported fatality rate, making it easy to spot countries with the highest to lowest reported fatality rates. The chart shows a dependency between the country and its reported fatality rates.

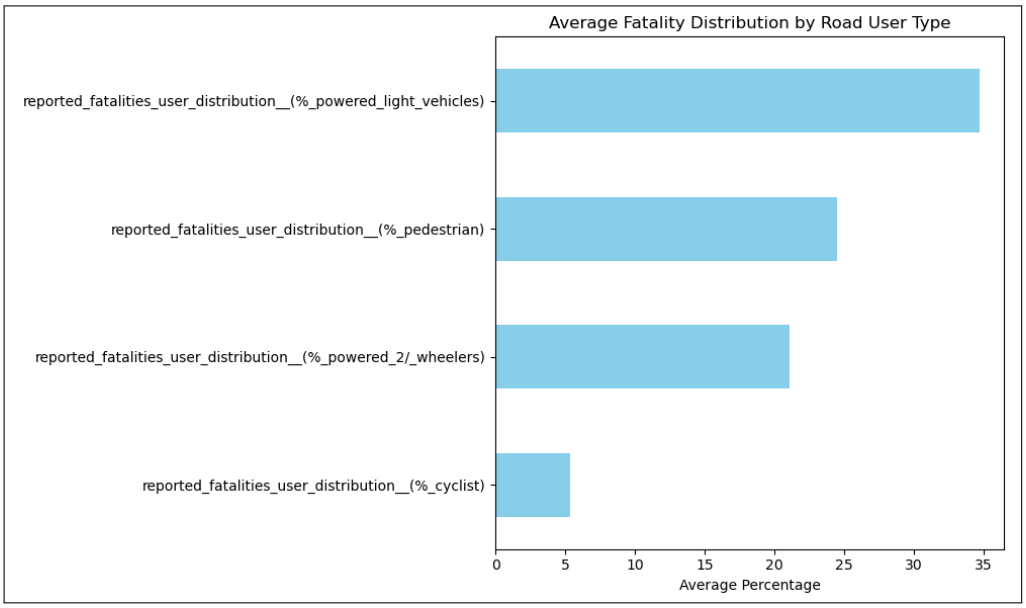


Figure 2: The average fatality rate by User Types

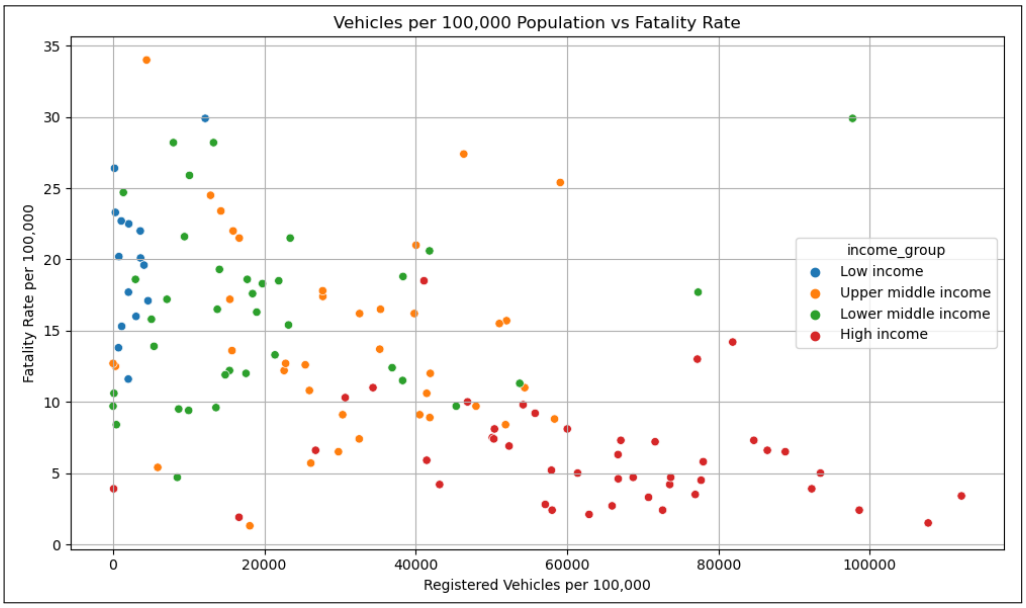


Figure 3: The Vehicles per 100,000 Population vs Fatality Rate by Income group

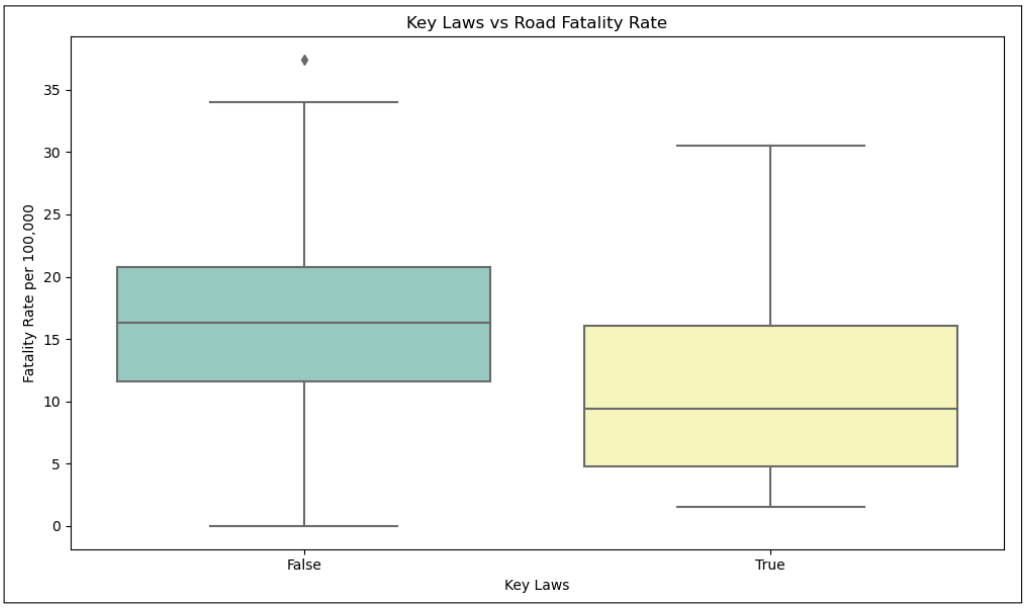


Figure 4: Key Law presence vs estimated fatality Rate

Answer (d) - Short summary of the main insights (with references to the corresponding image)

- Finding #1, as illustrated in Figure 1 highlights the countries with the highest number of reported fatalities. India stands out with the highest fatality count, reporting 153,972 deaths. China follows with 62,218 fatalities, and the United States is third with 42,939 fatalities. Beyond these three, the numbers gradually decline, with Brazil, Indonesia, Thailand, the Islamic Republic of Iran, the Russian Federation, Mexico, and South Africa completing the top 10 list.

  These figures reflect the significant burden of road traffic fatalities in these countries, underscoring the urgent need for stronger road safety measures and interventions to save lives.

- Finding #2, as illustrated in Figure 2, presents the average distribution of fatalities across different vehicle user types. The highest proportion of fatalities is seen among occupants of light powered vehicles. This is followed by pedestrians, who represent the second most affected group. Two-wheel powered vehicle users, such as motorcyclists, come next, while cyclists account for the smallest share of fatalities.

- Finding #3, as illustrated in Figure 3, illustrates the relationship between the number of registered vehicles per 100,000 population and the fatality rate per 100,000 population, colored by income group.

  The data reveals that high-income countries have the highest number of registered vehicles, yet they experience relatively low fatality rates. In contrast, low-income countries have far fewer registered vehicles but suffer from significantly higher fatality rates.

  Among the middle-income groups, upper-middle-income countries show a higher number of registered vehicles compared to lower-middle-income countries. However, lower-middle-income countries exhibit a higher fatality rate than their upper-middle-income counterparts.

  These findings highlight the disparities in road safety outcomes across different income groups and emphasize the need for improved road safety infrastructure and interventions, particularly in lower-income settings.

- Finding #4, as illustrated in Figure 4, the boxplot reveals the fatality rate per 100,000 population is analyzed in relation to the presence of key road safety laws, such as national motorcycle helmet laws, national seat-belt laws, national child restraint laws, and national speed limit laws. The results show that countries with these laws in place generally report a narrower range of fatality rates and a lower median fatality rate compared to those without such regulations. These findings highlight the significant impact that comprehensive road safety laws have in mitigating traffic-related fatalities.

2. *Correlation.* Given a dataset, which consists of 1,000 variables (hint: most of them are just random), the goal is to find the relationships between variables, i.e., which and how do the variables relate to each other; what are the dependencies. The dataset "task2-dataset.csv" can be downloaded from TeachCenter.

   (a) Which methods did you apply to find the relationships, and why?

   (b) Which relationships did you find and how do you characterise the relationships (e.g., variable "Lurkowl (Strix umbra #1068)" to "Frosthawk (Accipiter glacies #1064)" is linear with correlation found via method X of 0.9)?

   (c) Which causal relationships between the variables can you find (e.g., variable "Rattlepuff (Lynx rattleus #1067)" causes "Slingshark (Carcharodon slingus #1068)")?

Answer (a) - Method and motivation:

(a) Pearson Correlation Coefficient Method - Used because it provides a quantifiable way to assess how strongly two variables are related and in what direction (positive or negative).

(b) Regression Analysis Method - Used to examine the relationship between a dependent (or response) variable and one or more independent (or predictor) variables.

(c) Scatter Plot Method - Helps to visualize correlation and indicate the nature of relationships.

Answer (b)

Variable 1	Variable 2	Type of dependency	Method	Value
Lurkowl (Strix umbra 1068)	Trunkasaurus (Trunkasaurus ancientus 304)	Linear	Correlation	-0.932852
Chirpsnail (Cornu chirpitus 556)	Crunchbeetle (Coleoptera crunchus 1081)	Linear	Correlation	0.992159
Slingshark (Carcharodon slingus 894)	Squeakfluff (Sorex squeakus 1070)	Linear	Correlation	-0.998983
Fluffernox (Leontodon fluffernus 778)	Driftwolf (Canis fluctus 1065)	Linear	Correlation	0.950261
Skydash Falcon (Falco aeris 490)	Shimmercat (Felis lucere 323)	Linear	Correlation	0.934820

Answer (c)

- "Chirpsnail (Cornu chirpitus 556)" causes "Crunchbeetle (Coleoptera crunchus 1081)"
  (Strong positive linear relationship - When "Chirpsnail (Cornu chirpitus 556)" increases, "Crunchbeetle (Coleoptera crunchus 1081) also increases.)
- "Lurkowl (Strix umbra 1068)" causes "Trunkasaurus (Trunkasaurus ancientus 304)"
  (Strong negative linear relationship - When "Lurkowl (Strix umbra 1068)" increases, "Trunkasaurus (Trunkasaurus ancientus 304) is decreases.)
- "Fluffernox (Leontodon fluffernus 778)" causes "Driftwolf (Canis fluctus 1065)"
  (Strong positive linear relationship - When "Fluffernox (Leontodon fluffernus 778)" increases, "Driftwolf (Canis fluctus 1065) also increases.)

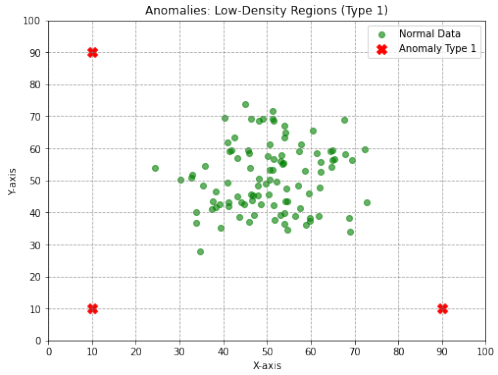
3. *Outliers/Anomalies.* Given two types of anomalies: (1) anomalies are defined to be **datapoints** in low density regions, (2) anomalies are **regions** of low density.

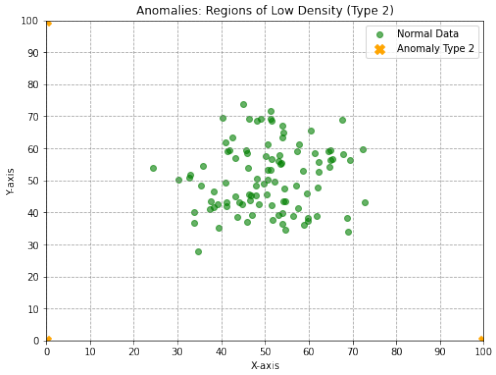
   (a) For both anomalies please create/draw a dataset of 2 features (x and y axis), with 3 anomalies and many normal data points (the normal datapoints should be marked, e.g., green colour)

   (b) Name the algorithms or describe the algorithmic way of how to identify this anomalous behaviour (you may also describe any necessary preprocessing)

   (c) Name the assumptions made by your algorithms

Answer (a) - Draw two datasets





Answer (b) - Describe the algorithms

Dataset 1 "Anomalies as Individual Points Algorithm" starts by creating normal data points clustered around the center of the grid at (50, 50), using a Gaussian distribution. Then, anomalies are introduced in low-density regions at the grid's corners, such as (10, 10), (90, 10), and (10, 90). After generating the data, the normal points are plotted in green, and the anomalies are marked in red "X" on a scatter plot. A grid with tick marks at intervals of 10 is added for clarity, with axis limits from 0 to 100.

Dataset 2 "Anomalies in Isolated Regions Algorithm" was used normal data points are clustered around (50, 50), while anomalies are placed in isolated regions far from the main cluster, such as (0, 0), (100, 0), and (0, 100). The normal data is plotted in green, and the anomalies are marked with orange "X". Similar to Dataset 1, the plot includes a grid with tick marks every 10 units, and the axis limits are set from 0 to 100.

Answer (c) - Describe the main assumptions

Algorithm	**Assumption**
Anomalies as Individual Points	Normal data are clustered
Anomalies as Individual Points	Anomalies are located at a significant distance from the normal data clusters.
Anomalies as Individual Points	Gaussian distribution approximates the normal behavior
Anomalies in Isolated Regions	Anomalies are from the low-density region
Anomalies in Isolated Regions	Both x and y dimensions are considered equally important

4. *Missing Values.* The dataset "task4-dataset.csv" (available on TeachCenter) contains a number of missing values. Try to reconstruct why the missing values are missing? What could be an explanation?

   (a) What are the dependencies in the dataset?

   (b) What could be reasons for the missingness?

   (c) What strategies are applicable for the features to deal with the missing values?

   (d) For each feature provide an estimate of the arithmetic mean (before and after applying the strategies to deal with missing values)?

Answer (a) - Describe the dependencies in the dataset

X	Y	Type of dependency
gender	height	Association
age	semester	Positive correlation
books per year	english skills	Weak positive correlation

Answer (b) - Describe the reason for missingness

Variable	Reason
height	Measurement not taken or Data entry error
likes pinapple on pizza	Forgot to answer, or skipped
english skills	Missing not at random or preferred not to disclose
semester	Missing at random

Answer (c) - Describe the strategies for dealing with missing values

Variable	Strategy
height	Fill missing values with the median by grouping the data based on gender
likes pinapple on pizza	Created new category as unknown nulls
english skills	Null values were filled with median
semester	Removed semester data greater than 40 as outliers

Answer (d) - Arithmetic mean of original dataset (with the missing values), and the one after applying the strategies

Variable	Before Strategy	After Strategy
height	172.161446	172.764921
semester	26.1560	10.343484
english skills	87.009721	87.133983