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:

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

   Answer (b) - List of main dependencies:

   - The reported fatalities by countries
   - The year by the fatality reduction target by countries
   - Safety laws and fatalities
   - The participation in the GRSSR of countries

   Answer (b) and (c)

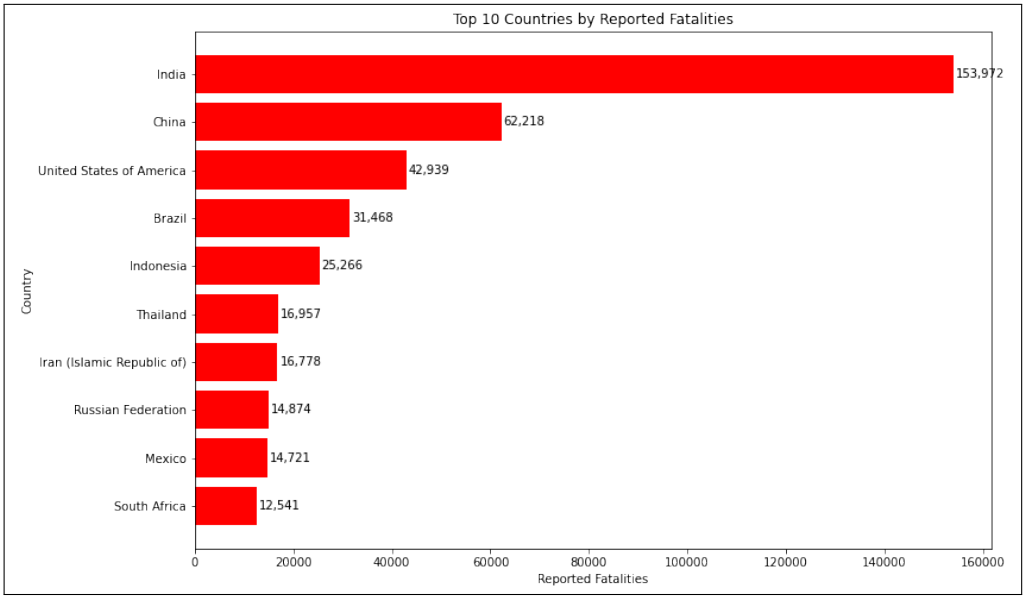


Figure 1: Top 10 countries by reported fatalities -This bar chart is ideal for analyzing the reported fatality rates across different countries, allowing for an easy comparison of how each country ranks. It clearly highlights the variation in reported fatality rates, showing the relationship between each country and its corresponding fatality rate.

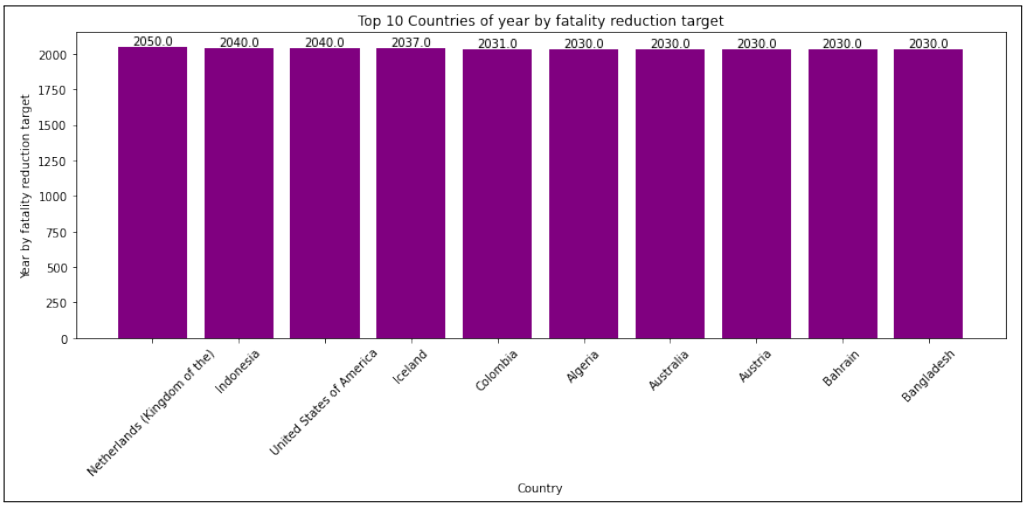


Figure 2: The top 10 countries of the year by fatality reduction target

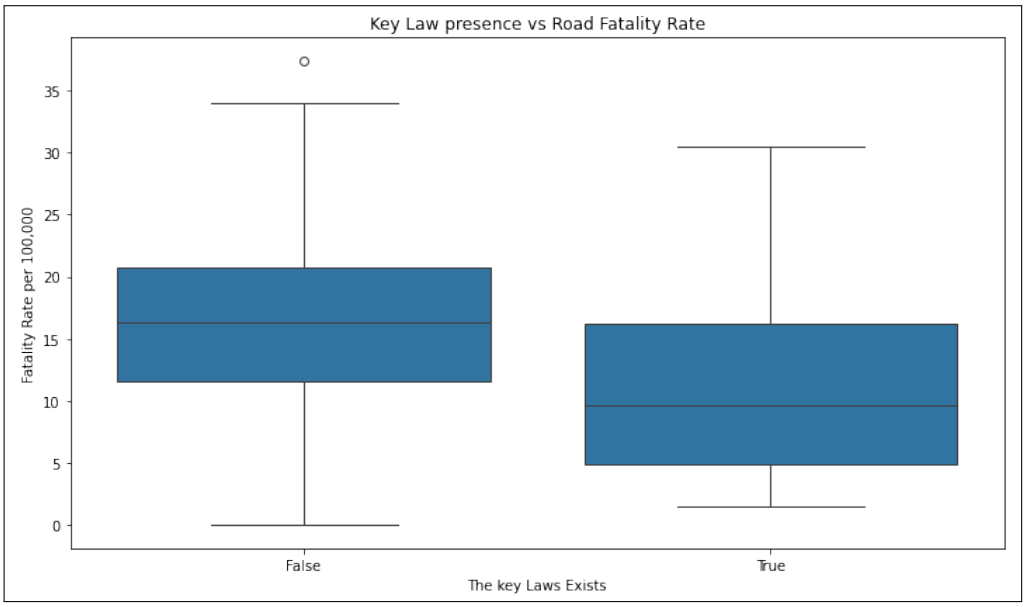


Figure 3: The Key laws presence vs road fatality rate

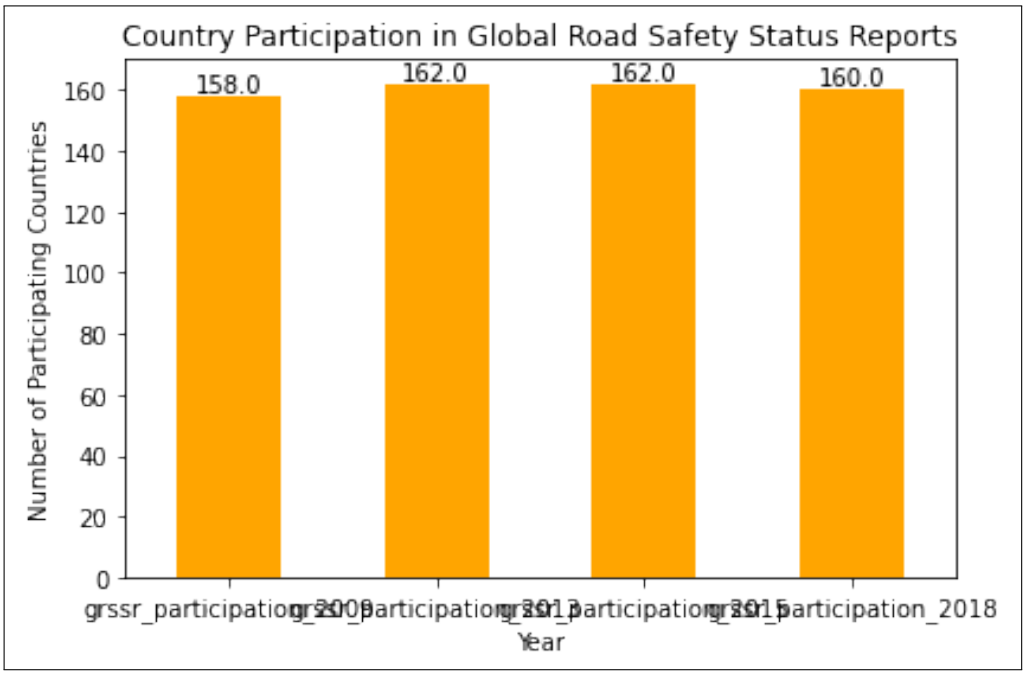


Figure 4: The Participation of countries for Global Road Safety Status Reports

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

- Finding #1, as illustrated in Figure 1 brings into focus the countries with the highest number of reported road traffic fatalities. India leads by a significant margin, with 153,972 deaths, followed by China with 62,218, and the United States with 42,939 fatalities. After these three, the numbers steadily decrease, with Brazil, Indonesia, Thailand, the Islamic Republic of Iran, the Russian Federation, Mexico, and South Africa rounding out the top 10.

  These figures highlight the heavy toll that road traffic accidents continue to take in these countries. They serve as a powerful reminder of the urgent need for stronger road safety policies and interventions to protect lives on our roads.

- Finding #2, as illustrated in Figure 2, presents the year-by-fatality reduction targets for the top 10 countries with the most ambitious road safety goals. The Netherlands leads with the longest-term target set for 2050, reflecting a strong, long-term commitment to eliminating road fatalities. Following closely are Indonesia and the United States, both aiming for 2040. Other countries, including Iceland, Colombia, Algeria, Australia, Austria, Bahrain, and Bangladesh, have also established clear timelines for reducing road deaths. This comparison highlights varying levels of ambition and urgency among nations, emphasizing the global recognition of road safety as a public health priority and the importance of setting measurable, time-bound targets to drive progress.

- Finding #3, as illustrated in Figure 3, presents the fatality rate per 100,000 population in relation to the presence of key road safety laws, including national motorcycle helmet laws, national seat-belt laws, national child restraint laws, and national laws setting speed limits. The analysis reveals that countries where these laws are in place tend to have a lower range of fatality rates, as well as a lower median fatality rate, compared to countries without such legislation. These findings underscore the critical role that comprehensive road safety laws play in reducing traffic-related fatalities.

- Finding #4, as illustrated in Figure 4, illustrates the participation of countries in the GRSSR (Global Road Safety Status Report) commitments across the years 2009, 2013, 2015, and 2018. The data show that the level of participation remained relatively consistent over these years, with no substantial changes observed. This steady engagement suggests a positive commitment toward ongoing efforts to reduce road traffic 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) Scatter Plot Method - Helps to visualize correlation and indicate the nature of relationships.

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

Answer (b)

Variable 1	Variable 2	Type of dependency	Method	Value
Danglefawn (Cervus dangleus 1067)	Puffpounce (Felis floccus 196)	Linear	Correlation	0.999867
Chirpsnail (Cornu chirpitus 556)	Crunchbeetle (Coleoptera crunchus 1081)	Linear	Correlation	0.992159
Slingshark (Carcharodon slingus 894)	Squeakfluff (Sorex squeakus 1070)	Linear	Correlation	-0.998983
Sparklequail (Coturnix fulgor 1068)	Scruffpaws (Felis scruffus 1061)	Linear	Correlation	0.956447
Vinepaw (Felis liana 1073)	Driftwolf (Canis fluctus 542)	Linear	Correlation	-0.942152

Answer (c)

- "Danglefawn (Cervus dangleus 1067)" causes "Puffpounce (Felis floccus 196)"
  (Strong positive linear relationship - When "Chirpsnail Danglefawn (Cervus dangleus 1067)" increases, "Puffpounce (Felis floccus 196) also increases.)
- "Sparklequail (Coturnix fulgor 1068)" causes "Scruffpaws (Felis scruffus 1061)"
  (Strong positive linear relationship - When "Sparklequail (Coturnix fulgor 1068)" increases, "Scruffpaws (Felis scruffus 1061) also increases.)
- "Vinepaw (Felis liana 1073)" causes "Driftwolf (Canis fluctus 542)"
  (Strong negative linear relationship - When "Vinepaw (Felis liana 1073)" increases, "Driftwolf (Canis fluctus 542) is decreases.)

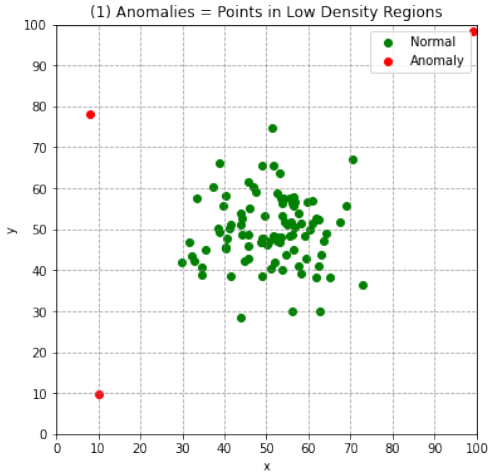
3. *Outliers/Anomalies.* Given two types of anomalies: (1) anomalies are defined to be **datapoints** in low density regions, (2) anomalies are **regions** of low density.

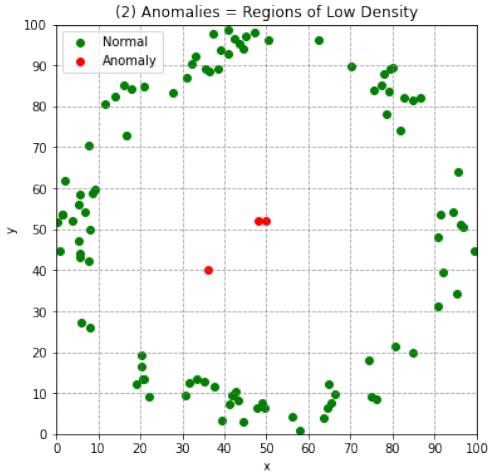
   (a) For both anomalies please create/draw a dataset of 2 features (x and y axis), with 3 anomalies and many normal data points (the normal datapoints should be marked, e.g., green colour)

   (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 Data Points in Low-Density Regions" : A fixed random seed is set for reproducibility. Ninety-seven normal data points are generated from a multivariate normal distribution centered at [50, 50] with a small covariance. Anomalies are added at sparse locations: [10, 9.9], [99, 98.5], and [8, 78]. The normal data and anomalies are combined into a single dataset, with labels assigned as 0 for normal points and 1 for anomalies.

Dataset 2 "Anomalies as Regions of Low Density (Hole in the Center)": A random seed is set, and ninety-seven normal data points are generated in a ring pattern around [50, 50] with random radii between 40 and 50. Anomalies are placed near the center of the ring at [48, 52], [36, 40], and [50, 52]. The normal and anomaly points are merged into a dataset, with labels assigned as 0 for normal points and 1 for anomalies.

Answer (c) - Describe the main assumptions

Algorithm	**Assumption**
Anomalies as Data Points in Low-Density Regions	Normal data are clustered
Anomalies as Data Points in Low-Density Regions	These outliers are positioned far from areas where normal data is concentrated.
Anomalies as Data Points in Low-Density Regions	The labels are assumed to be binary, with 0 for normal points and 1 for anomaly points
Anomalies as Regions of Low Density (Hole in the Center)	The data follows this ring-shaped distribution and the radius of the ring is randomly determined
Anomalies as Regions of Low Density (Hole in the Center)	Anomalies are assumed to occur in a small, distinct area within the center of the ring
Anomalies as Regions of Low Density (Hole in the Center)	There is no overlap between the normal data points and the anomalies

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
age	semester	Positive correlation
books per year	english skills	Weak positive correlation
gender	height	Association
gender	likes chocolate	Association

Answer (b) - Describe the reason for missingness

Variable	Reason
height	Measurements not taken or Data entry error
likes pinapple on pizza	Omitted or missed answering
english skills	Data missing randomly or Chose 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
semester	Eliminated data points from the semester where values surpassed 40
likes pinapple on pizza	New category as unknown was created
english skills	Filled with median

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
english skills	87.009721	87.133983
semester	26.1560	10.343484