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 (Dropped columns with too many missing values and removed outliers)
   - Rename and Clean Columns (Made column names more readable)
   - Convert Data types
   - Create Derived Features
   - Filter the Dataset and select appropriate features

   Answer (b) - List of main dependencies:

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

   Answer (b) and (c)

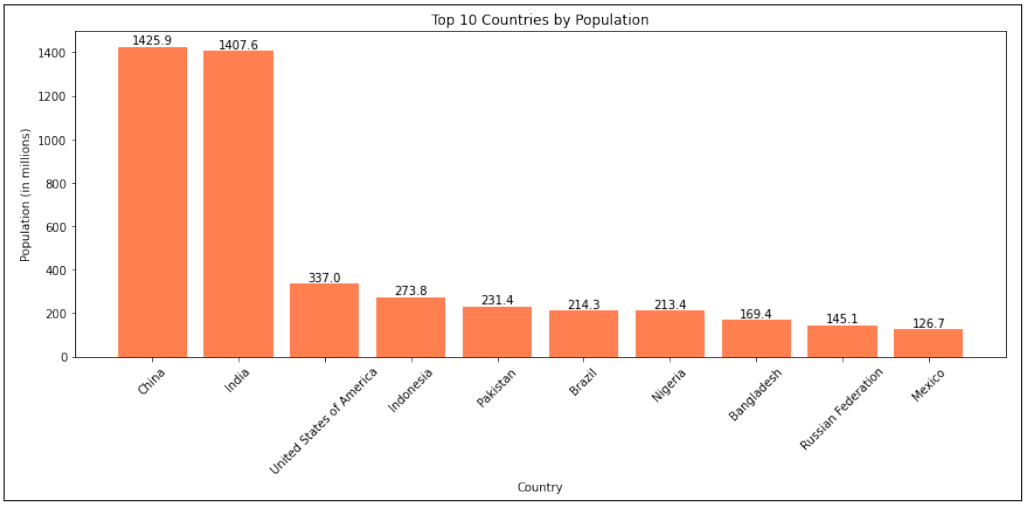


Figure 1: Top 10 countries by population - This bar chart is perfect for comparing population across countries and seeing how each country stacks up in terms of population size, making it easy to spot countries with the highest and lowest populations. The chart shows a dependency between the country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

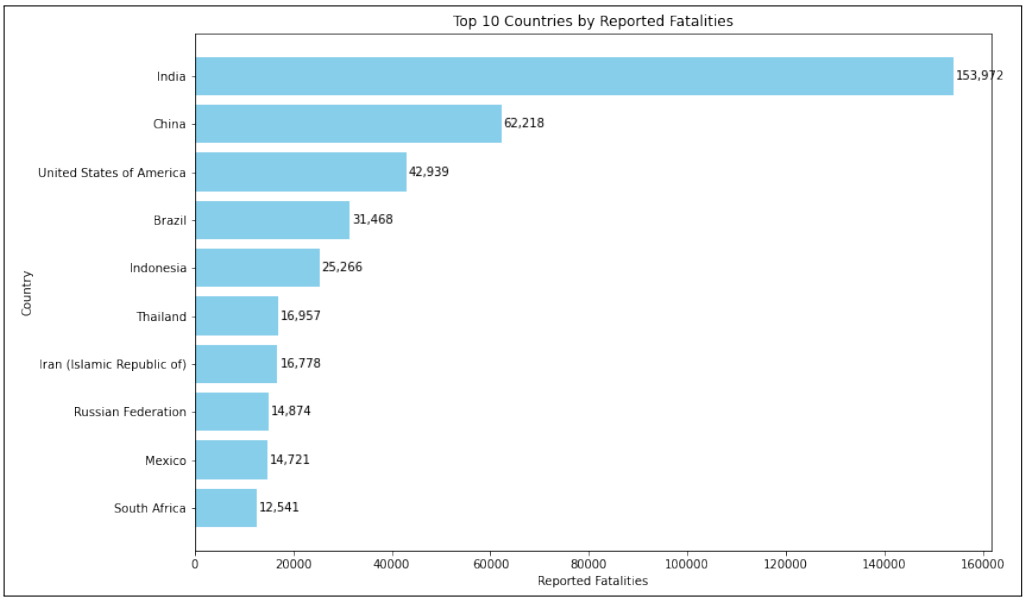


Figure 2: Top 10 countries by reported fatalities.

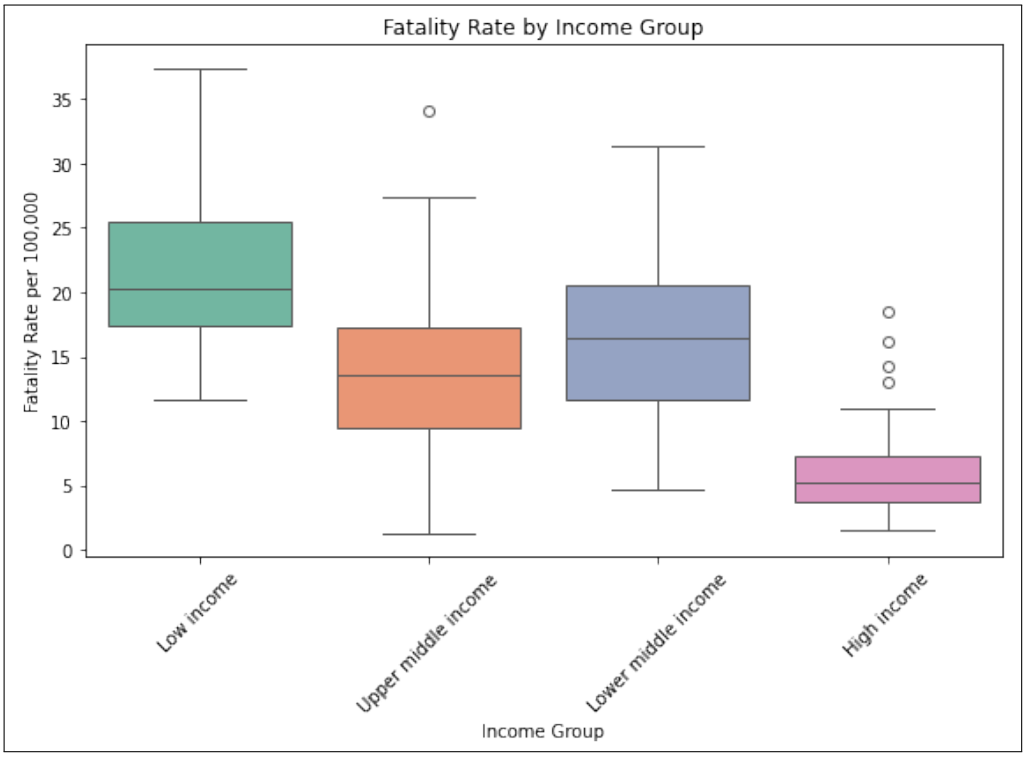


Figure 3: Fatality rate by Income group

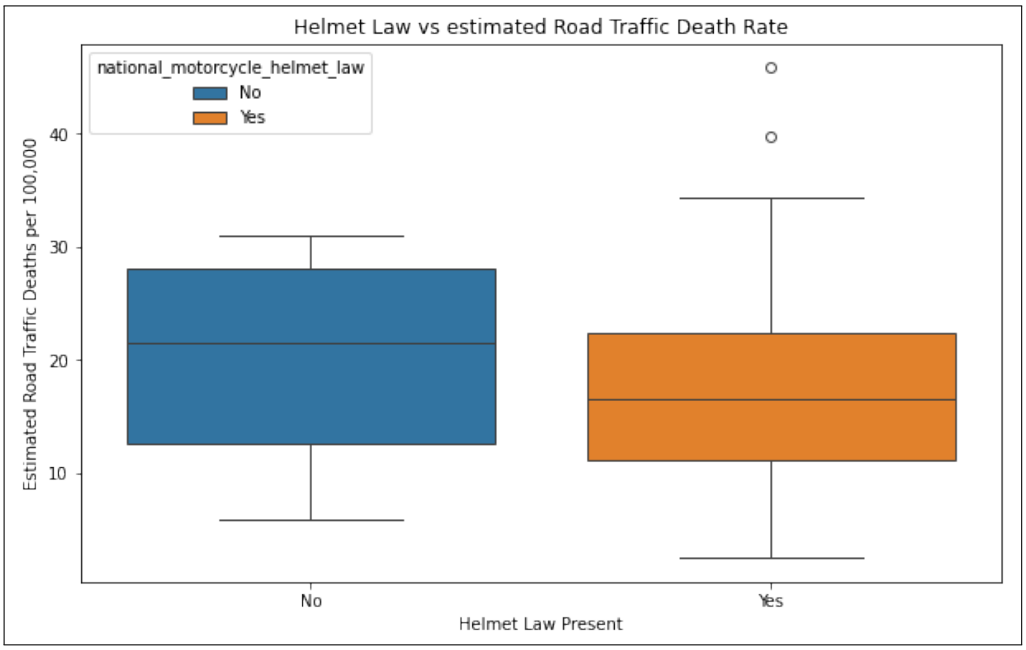


Figure 4: Helmet Law presence vs estimated Road Traffic Death Rate

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

- Finding #1, as illustrated in Figure 1 illustrates the population by country, based on the dataset. China ranks highest with a population of 1.425 billion, followed closely by India at 1.408 billion. The next largest populations are as follows: the United States with 337.7 million, followed by Indonesia, Pakistan, Brazil, Nigeria, Bangladesh, the Russian Federation, and Mexico, which round out the top 10 countries by population.

  This chart provides valuable insights into the distribution of population across the world, highlighting the concentration of large populations in a few key countries. It also illustrates a noticeable decrease in population size as we move down the list, with the gap between China and India compared to the other countries being particularly significant.

- Finding #2, as illustrated in Figure 2, the top 10 countries by reported fatalities are as follows: India leads with 153,972 fatalities, followed by China at 62,218 and the USA at 42,939. The fatalities then decrease in the following order: Brazil, Indonesia, Thailand, Iran, Russian Federation, Mexico, and South Africa.

  India has the highest number of fatalities, which is consistent with its large population. While China also has a high population, the USA, despite a smaller population, has relatively high reported fatalities compared to other countries.

- Finding #3, as illustrated in Figure 3, the boxplot clearly shows that lower income countries tend to have higher and more variable fatality rates, whereas higher income countries generally have lower and more consistent fatality rates. This suggests that income levels might play a significant role in road safety, where wealthier countries may have better road safety measures and healthcare systems, leading to fewer fatalities and less variability in fatality rates.

- Finding #4, as illustrated in Figure 4, the boxplot reveals a significant difference in road safety between countries with and without a national motorcycle helmet law. Countries that have such laws show a lower median fatality rate and less variation in fatalities, suggesting that helmet laws are effective in reducing road traffic fatalities, especially in motorcycle accidents. In contrast, countries without helmet laws experience higher and more variable fatality rates, highlighting the importance of such legislation for improving road safety.

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 - Useful for identifying linear relationships between numerical features and easy to handle high number of variables.

(b) Scatter Plot Method - Help to visualize correlation and indecate the nature of relationships.

(c) Regression Analysis Method - Provides insights into the strength and direction of the relationship.

Answer (b)

Variable 1	Variable 2	Type of dependency	Method	Value
Danglefawn (Cervus dangleus 1067)	Puffpounce (Felis floccus 196)	Linear	Correlation	0.999867
Splashleopard (Panthera splashus 513)	Shivershark (Carcharodon tremorensis 824)	Linear	Correlation	-0.999147
Slingshark (Carcharodon slingus 894)	Squeakfluff (Sorex squeakus 1070)	Linear	Correlation	-0.998983
Shivershark (Carcharodon tremorensis 738)	Squeakfluff (Sorex squeakus 1070)	Linear	Correlation	-0.994845
Fluffernox (Leontodon fluffernus 778)	Driftwolf (Canis fluctus 1065)	Linear	Correlation	0.950261

Answer (c)

- "Danglefawn (Cervus dangleus 1067)" causes "Puffpounce (Felis floccus 196)"
  (Strong positive linear relationship - When "Danglefawn (Cervus dangleus 1067)" increases, "Puffpounce (Felis floccus 196) also increases.)
- "Splashleopard (Panthera splashus 513)" causes "Shivershark
  (Carcharodon tremorensis 824)"
  (Strong negative linear relationship - When "Splashleopard (Panthera splashus 513)" increases, "Shivershark
  (Carcharodon tremorensis 824) 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.)

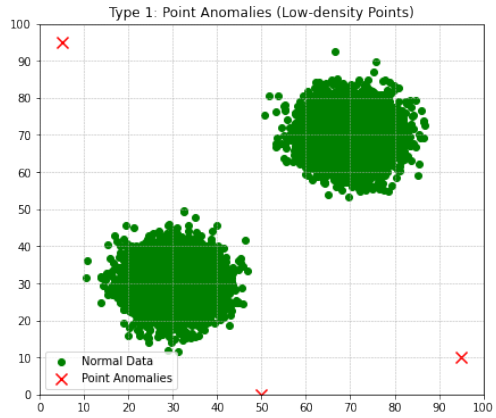
3. *Outliers/Anomalies.* Given two types of anomalies: (1) anomalies are defined to be **datapoints** in low-density regions, (2) anomalies are **regions** of low density.

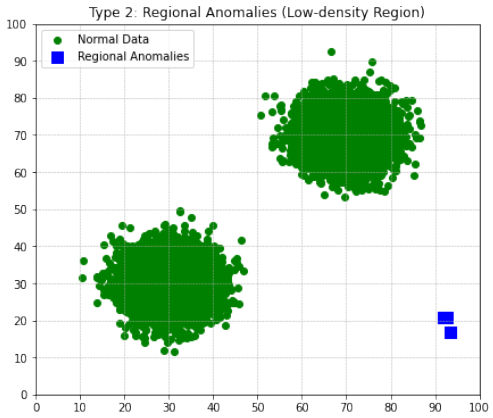
   (a) For both anomalies please create/draw a dataset of 2 features (x and y axis), with 3 anomalies and many normal data points (the normal datapoints should be marked, e.g., green colour)

   (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 Method" was used and 9997 normal points were sampled from a 2D Gaussian cluster, and 3 anomalies were manually placed far from the cluster. This approach was chosen to simulate isolated, rare events, ideal for testing point-based anomaly detection methods like Isolation Forest and Local Outlier Factor.

Dataset 2 "Anomalies as Low-Density Regions Method" was used and 9997 normal points formed a dense circular ring, leaving a low-density center where 3 anomalies were added. This setup models structural gaps in data, suitable for evaluating density-based methods like DBSCAN and cluster-based anomaly detection.

Answer (c) - Describe the main assumptions

Algorithm	**Assumption**
Anomalies as Individual Points Method	Normal data is clustered
Anomalies as Individual Points Method	Anomalies are far from the cluster
Anomalies as Individual Points Method	Gaussian distribution approximates the normal behavior
Anomalies as Low-Density Regions Method	Anomalies are from the low-density region
Anomalies as Low-Density Regions Method	No feature correlation

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	Failed to respond or intentionally left blank
english skills	Missing not at random or preferred not to disclose

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

Variable	**Strategy**
height	Filled with median by grouping gender wise
likes pinapple on pizza	Crated new category as unkown
english skills	Filled with median
semester	Filtered out semester data entries exceeding a value of 40

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