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:

- Filter the Dataset and select appropriate features
- Convert data types into the correct data types
- Rename and clean columns (Made column names more readable)
- Handling missing values (Dropped columns with too many missing values and removed outliers)
- Create derived features
- Scaled some features as needed

Answer (b) - List of main dependencies:

- The population by country
- The reported fatalities by countries
- Fatality rate by income group
- Fatality distribution by road user type
- Speed limits and enforcement

Answer (b) and (c)

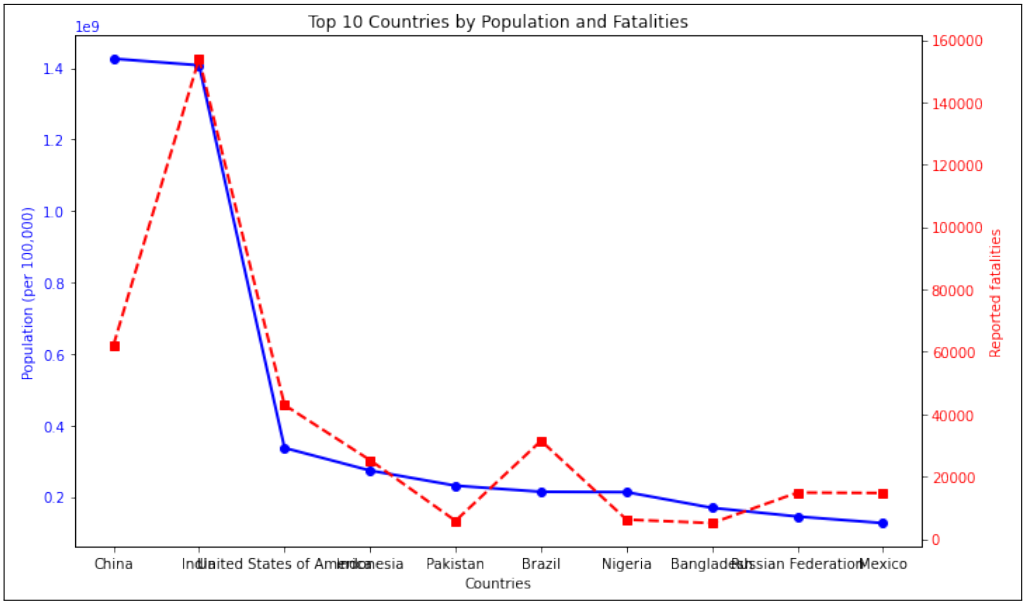


Figure 1: Top 10 countries by population and reported fatalities - This chart illustrates the top 10 countries by population and their reported road fatalities. It provides insight into the relationship between population size and the number of fatalities, allowing for comparisons between countries. By analyzing this data, we can identify which nations are facing a higher burden of road fatalities relative to their population size and where additional measures may be necessary to improve road safety and prevent further loss of life. This comparison highlights the need for targeted interventions in countries with high fatality rates, even as their population sizes vary.

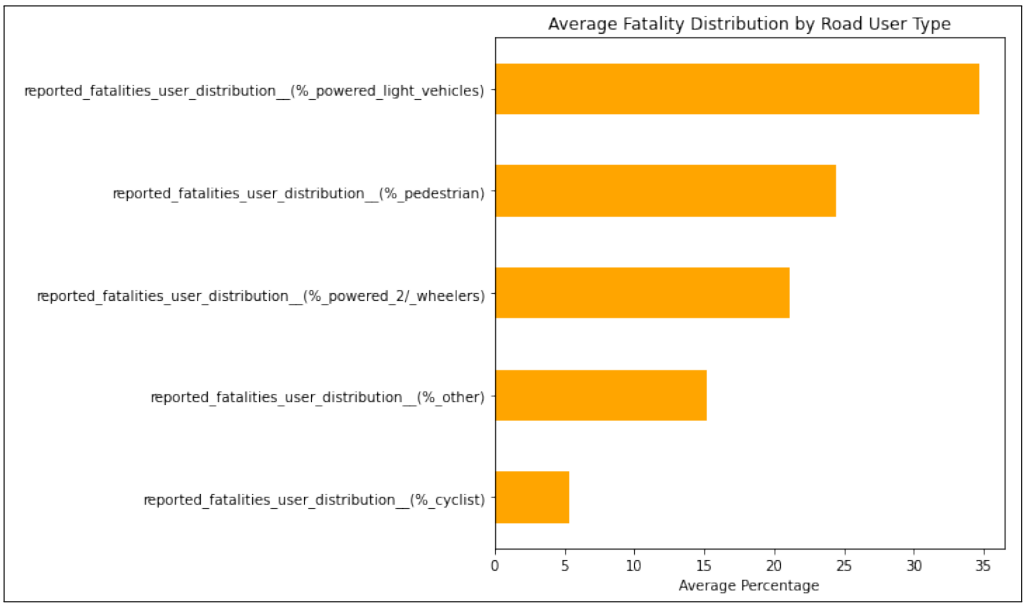


Figure 2: The Average fatality distribution by road user type

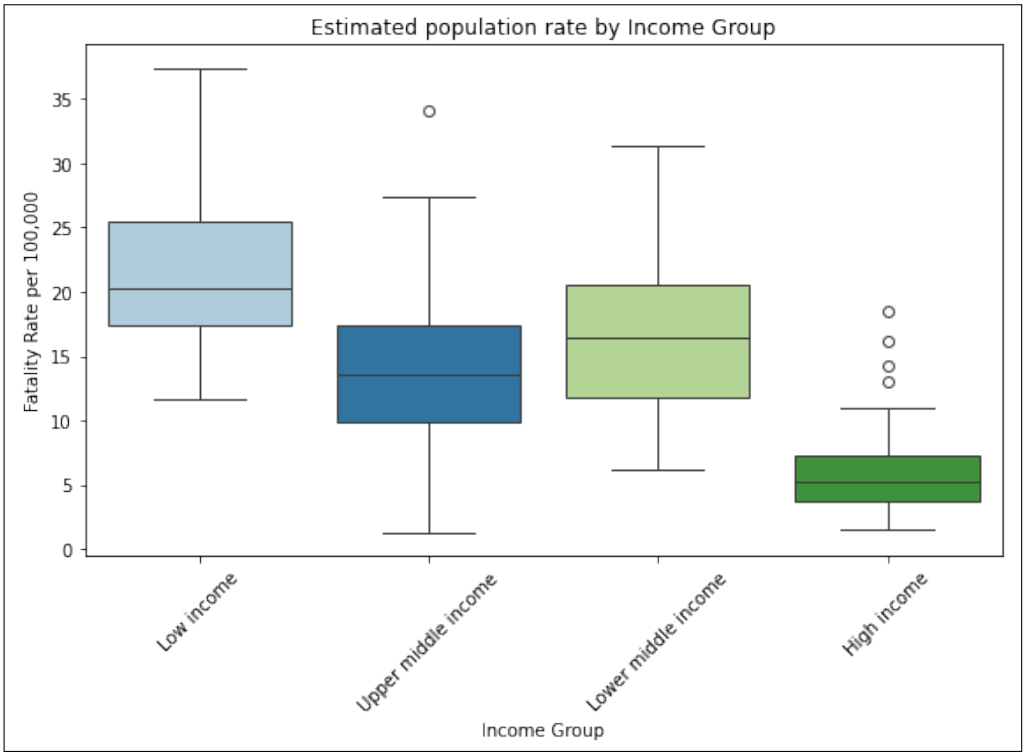


Figure 3: The Estimated rate per 100000 population by Income Group

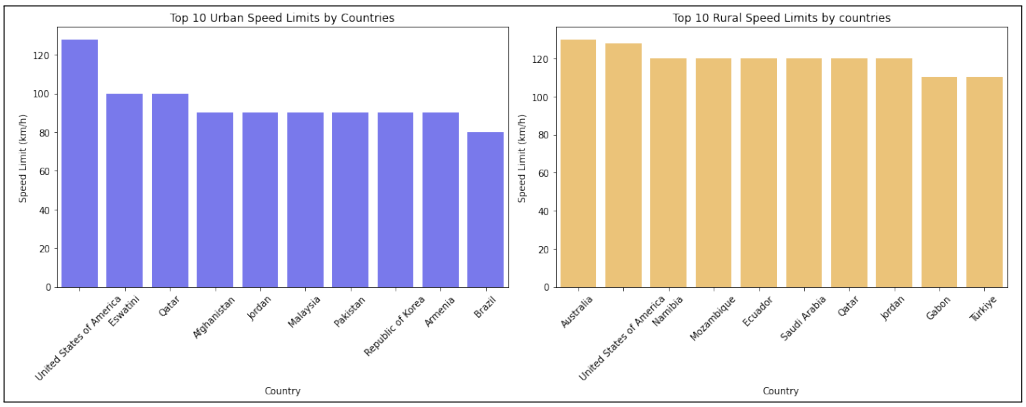


Figure 4: The Maximum speed limits of urban and rural areas by country

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

- Finding #1, as illustrated in Figure 1, the line chart illustrates that the blue line represents the population per 100,000 and the red line represents reported fatalities. China has the largest population, followed by India, the United States, Indonesia, Pakistan, Brazil, Nigeria, and Bangladesh. Although China's population is higher, India reports more fatalities, indicating an urgent need for improved road safety measures in India. Reported fatalities decrease respectively across India, China, the United States, Brazil, and Indonesia, highlighting differences in road safety outcomes among these countries.
- Finding #2, as illustrated in Figure 2, shows the average fatality distribution by road user type: powered light vehicles, pedestrians, powered two-wheelers, others, and cyclists. Fatalities are highest among powered light vehicle users, followed by pedestrians, with numbers decreasing for powered two-wheelers, others, and cyclists. This highlights the need to strengthen road safety measures, particularly for vehicle occupants and pedestrians, while continuing efforts to protect all road users.
- Finding #3, as illustrated in Figure 3, highlights a clear trend: lower-income countries experience higher and more variable fatality rates relative to their estimated populations, while higher-income countries show lower and more stable fatality rates. This pattern indicates that a country's income level may significantly influence road safety outcomes, with wealthier nations likely benefiting from stronger infrastructure, effective safety regulations, and better healthcare systems, resulting in fewer and more consistent fatality rates.
- Finding #4, as illustrated in Figure 4 illustrates the maximum speed limits (in km/h) set by countries for urban and rural areas. In urban areas, the United States records the highest speed limit, followed by Eswatini and Qatar, which share the same limit. Next are Afghanistan, Jordan, Malaysia, Pakistan, Korea, and Armenia, all having slightly lower but identical speed limits compared to the top countries. Brazil also ranks among the top 10 countries with the highest urban speed limits. In rural areas, Australia has the highest maximum speed limit, followed by the United States. Namibia, Mozambique, Ecuador, Saudi Arabia, Qatar, and Jordan have similar speed limits, slightly lower than the top two. Gabon and Türkiye complete the list of the top 10 countries with the highest rural speed limits. This comparison highlights notable regional differences in speed regulations, which may reflect variations in infrastructure quality, road safety policies, and traffic management strategies.

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 indicate 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	0.99986
Tangofox (Vulpes tangus 1074)	Chompbeak (Aquila mordus 150)	Linear	Correlation	-0.9888
Shivershark (Carcharodon tremorensis 738)	Slingshark (C0.993804archarodon slingus 894)	Linear	Correlation	0.99380
Shivershark (Carcharodon tremorensis 738)	Squeakfluff (Sorex squeakus 1070)	Linear	Correlation	-0.9948
Scruffpaws (Felis scruffus 1061)	Sparklequail (Coturnix fulgor 1068)	Linear	Correlation	0.95644

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.)
- "Tangofox (Vulpes tangus 1074)" causes "Chompbeak (Aquila mordus 150)"
  (Strong negative linear relationship - When "Tangofox (Vulpes tangus 1074)" increases, "Chompbeak (Aquila mordus 150) is decreases.)
- "Scruffpaws (Felis scruffus 1061)" causes "Sparklequail (Coturnix fulgor 1068)"
  (Strong positive linear relationship - When "Scruffpaws (Felis scruffus 1061" increases, "Sparklequail (Coturnix fulgor 1068).)

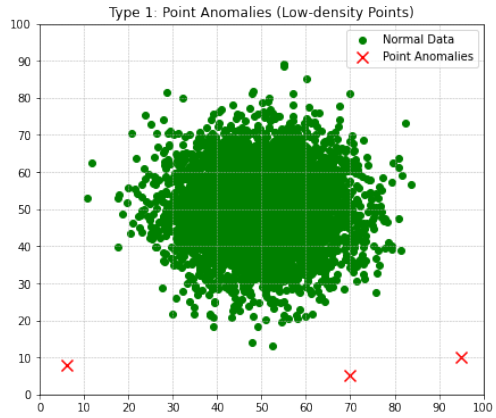
3. *Outliers/Anomalies.* Given two types of anomalies: (1) anomalies are defined to be **datapoints** in low-density regions, (2) anomalies are **regions** of low density.

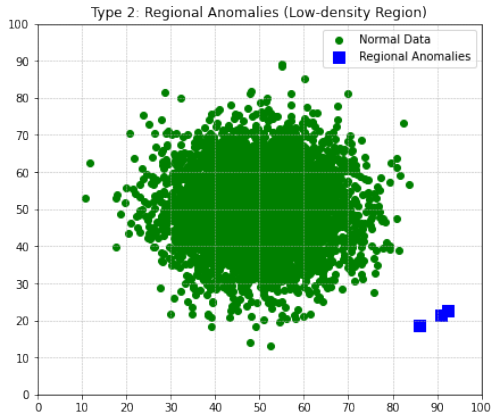
   (a) For both anomalies please create/draw a dataset of 2 features (x and y axis), with 3 anomalies and many normal data points (the normal datapoints should be marked, e.g., green colour)

   (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" generated a cluster of normal data points by sampling from Gaussian (normal) distributions centered at (50, 50), producing 4997 points, and manually placed three isolated points at coordinates far from the cluster, then visualized the data by plotting normal points as green circles, point anomalies as red 'X' markers.

Dataset 2 "Anomalies as Low-Density Regions Method" generated a cluster of normal data points by sampling from Gaussian (normal) distributions centered at (50, 50), producing 4997 points. for regional anomalies, a small cluster of three points is generated using a Gaussian distribution centered around (90, 20) with a small standard deviation (scale=2).these points are close to each other but located in a low-density region

Answer (c) - Describe the main assumptions

Algorithm	**Assumption**
Anomalies as Individual Points Method	Normal data is clustered
Anomalies as Individual Points Method	Gaussian distribution approximates the normal behavior
Anomalies as Individual Points Method	Anomalies are far from the cluster
Anomalies as Individual Points Method	All data points are assumed to fall within the x and y range of 0 to 100
Anomalies as Low-Density Regions Method	Anomalies are from the low-density region
Anomalies as Low-Density Regions Method	No feature correlation between each point

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
gender	likes chocolate	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	Skipped or unintentionally left unanswered
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 unknown
english skills	Filled null values with median
semester	Removed the semester data where the value exceeded 40

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

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