1. *Visual Data Analysis.* Given the dataset "visual-dataset.csv" (available in TeachCenter), which comprises a number of features taken from Climate Watch. Provide a number of meaningful visualisations (4 visualisations) that show key properties of the dataset and dependencies. Based on the visualisations provide your interpretation. Ideally, one would like to identify some dependency on the yearly values, or the change in yearly values.

   (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) What types of dependency/relationship are there (per figure)?

   (d) Provide a summary of the main dependencies

   Answer (a) - Preprocessing steps:

   - Removed the duplicates;
   - Replaced missing values with mean;
   - Grouped the dataset so that each country only has one row (groupby(['Country']);
   - Grouped the dataset so that each Gas type only has one row(groupby(['Sector']);
   - Added an average column to the dataframe

   Answer (b) and (c)

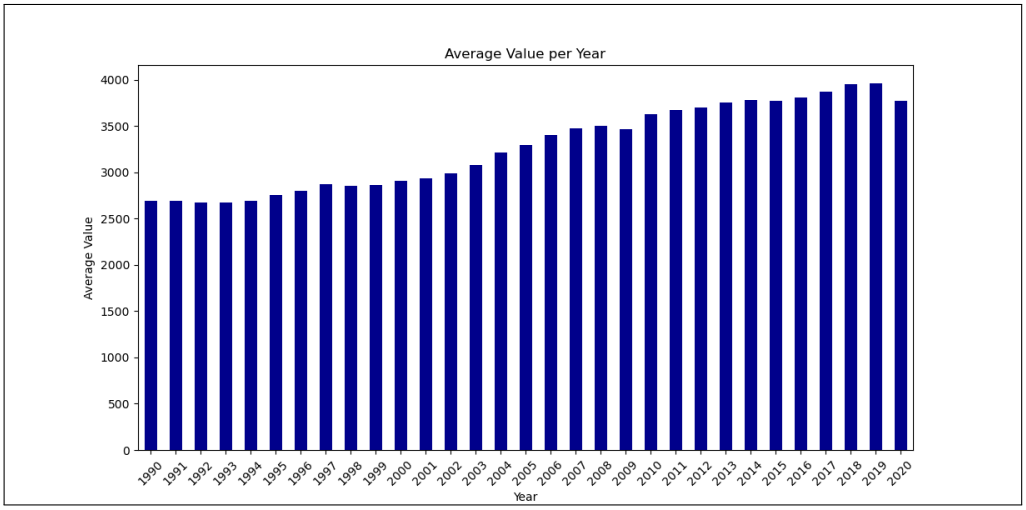

Figure 1: (i) This Visualization allows the viewer to get an idea of what the data is about, and that emissions are increasing over time (ii) Linear dependency showing that emissions rise with the years that go by (iii) The world is increasing it's emissions of CO2

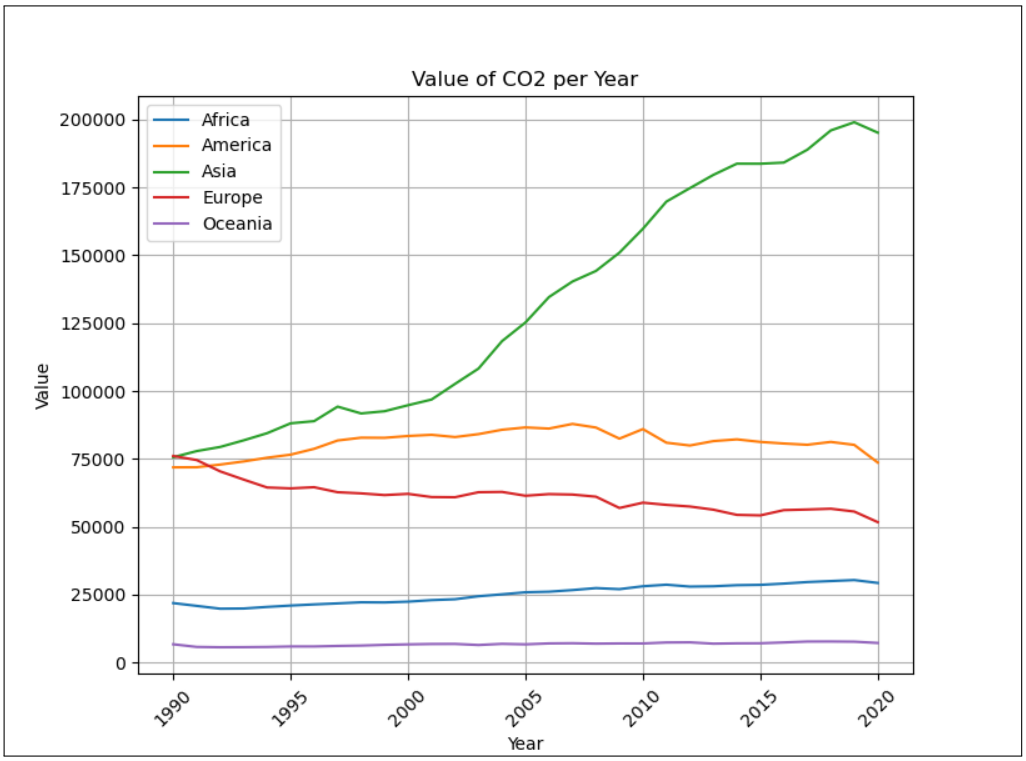

Figure 2: (i) This kind of Visualization allows the viewer to understand that the increase is emissions is manly due to Asia (ii) A exponential dependency where emisions rise dramatically in Asia as the years go by (for most of the time), and decrease slowly but surely in other continents. Note that Africa also rises over time very slightly (iii) Most of the world is decreasing emissions except for Asia

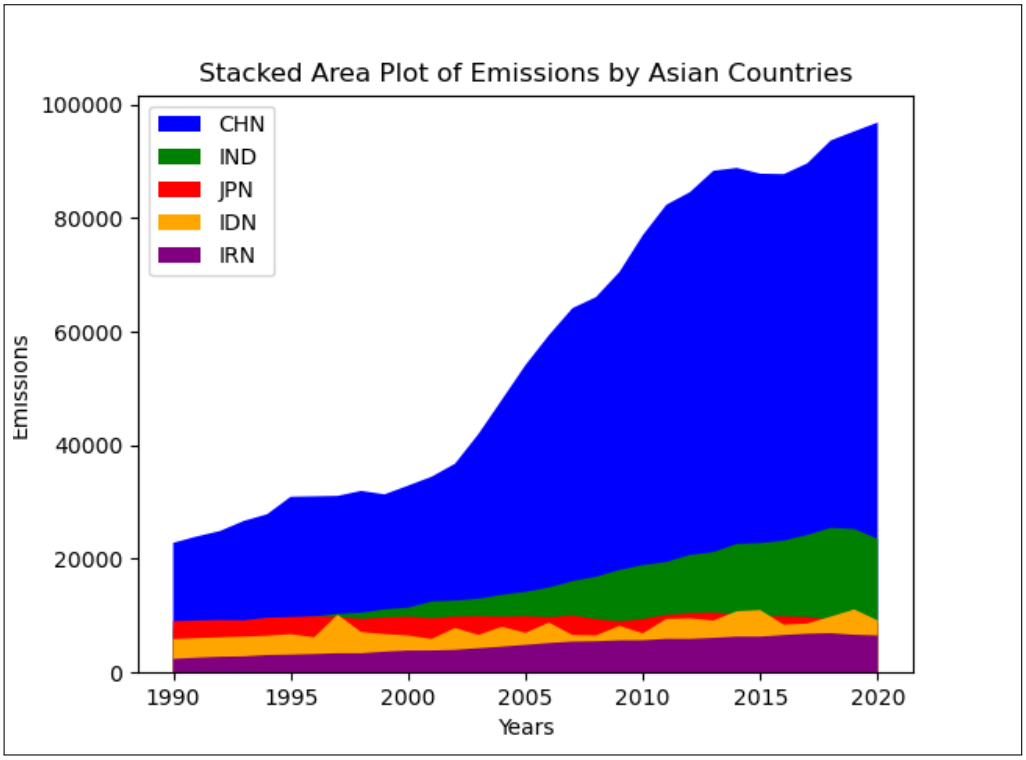

Figure 3: (i) This Visualization allows the user to see whithin Asian the biggest culprits of CO2 emission (ii) An exponential emission of CO2 by China and linearly by India and Iran. Japan also presents a negative linear correlation in which it decreases it's emissions over time (iii) Given the plots from above, Asia is is the biggest responsible for world emissions of CO2 rising, of which, China is the biggest responsible country within.

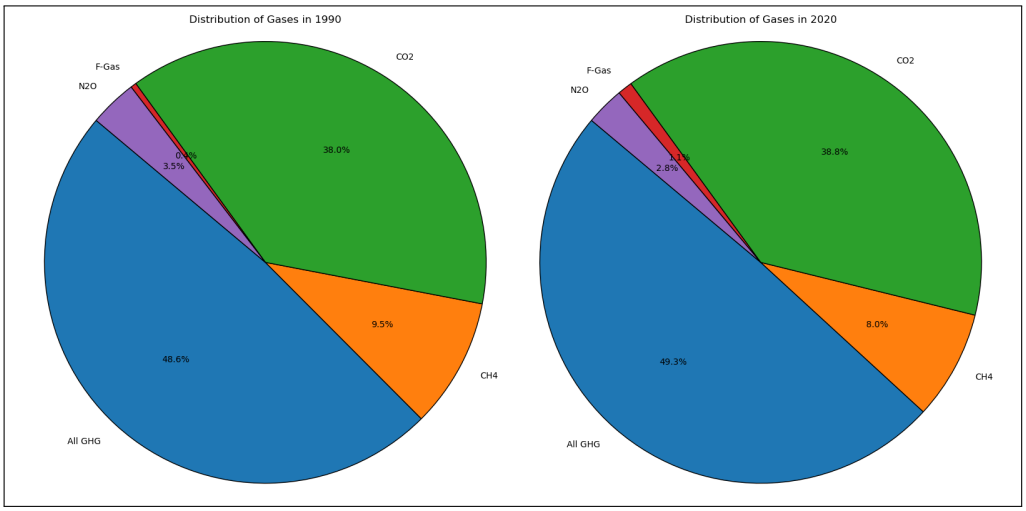

Figure 4: (i) Pie charts allows the user to view information grouped by categories that add up to a total of 100 (ii) This plots displays two pie charts that have identical percentages (iii) The percentages of emitted gases has not changed much since 1990.

Answer (d) - Short summary of the main findings

- The world as a whole is increasing it's CO2 Emissions;
- Most of continents and countries are in fact decreasing it's emission;
- The main entity responsible for emitting CO2 is the continent of Asia, and within Asia, China is biggest emitter;
- During the recorded period of time (1990-2020) the way our emitted gases are distributed has not changed significately;

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 "correlation-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 "Munchen" to "Delmenhorst" is linear with correlation found via method X of 0.9)?

   (c) Which causal relationships between the variables can you find (e.g., variable "Hamburg" causes "Bielefeld")?

Answer (a) - Method and motivation:

(a) Pearson's Correlation was chosen to detect linear correlations between the different variables;

(b) Spearman's Correlation was chosen to detect non-linear monotonic relationships between variables

Answer (b)

Variable 1	**Variable 2**	**Type of dependency**	**Method**	**Value**
Hollabrunn	Lienz	Linear	Pearson's	0.598
Feldbach	Hard	Linear	Pearson's	0.925
Vocklabruck	Enns	Linear	Pearson's	0.699
Rankweil	Liez	Linear	Pearson's	0.824
Knittelfeld	Neunkirchen	Linear	Pearson's	0.850
Knittelfeld	Korneuburg	Linear	Pearson's	0.885
Korneuburg	Neunkirchen	Linear	Pearson's	0.956
Liez	Hollabrunn	Non-linear	Spearman's	0.596
Hard	Feldbach	Non-linear	Spearman's	0.932
Vocklabruck	Enns	Non-linear	Spearman's	0.756
Knittelfeld	Neunkirchen	Non-linear	Spearman's	0.688
Knittelfeld	Korneuburg	Non-linear	Spearman's	0.825
Korneuburg	Neunkirchen	Non-linear	Spearman's	0.822

Note: Only correlations with a higher score than 0.55 were presented. The first three rows in each table represent the same variables with very close values, meaning, that the variables in questions have a strong association wheter we evaluate them from a linear or monotonic perpective. To simplify, only the first three from the first method could have shown, however showing all six reinforces the previous explored idea.

Answer (c) No causal relationships between the variables can be established with complete certainty because correlation does not mean causation. Proving causation would require further analysis, like the correlation being consistent and strong over different studies and years, changing one variable and accurately predicting the other one, experimental studies (RCT's), etc... However, given a strong enough correlation, like what is the case with some variables presented above (for example Korneuburg and Neunkirchen or Knittelfeld and Korneuburg) we can say with confidence, that there exists a dependency between them.

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

   (a) For both anomalies please create/draw a dataset, 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

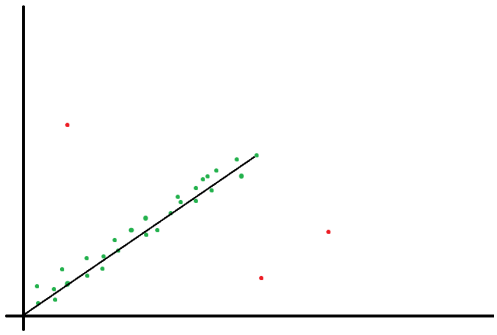

Figure 5: Scatter plot with 3 outliers. Green represents normal points in the dataset while red represents outliers/anomalies.

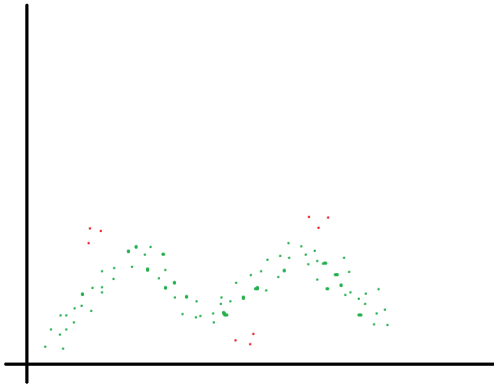

Figure 6: Dataset with 3 groups of 3 outliers each. Green represents normal points in the dataset while red represents outliers/anomalies.

Answer (b) - Describe the algorithms

Dataset 1

Distance Algorithm

Isolation Forest

K Nearest Neighbour (K-NN)

Dataset 2

Local Outlier Factor (LOF)

K Nearest Neighbour (K-NN)

Angle Based Outlier Degree (ABOD)

Answer (c) - Describe the main assumptions

Algorithm	**Assumption**
K-NN	Euclidean space/distance, low dimensionality (2D) and outliers expected to have a low density
ABOD	Angles are linear, not too much noise
LOF	Outliers have different local densities than their neighbours, optimal choice of k
Distance	Euclidean space/geomtry, optimal threshold selection
Isolation Forest	Outliers are isolated and in low density regions

4. *Missing Values.* The dataset "missing-values-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

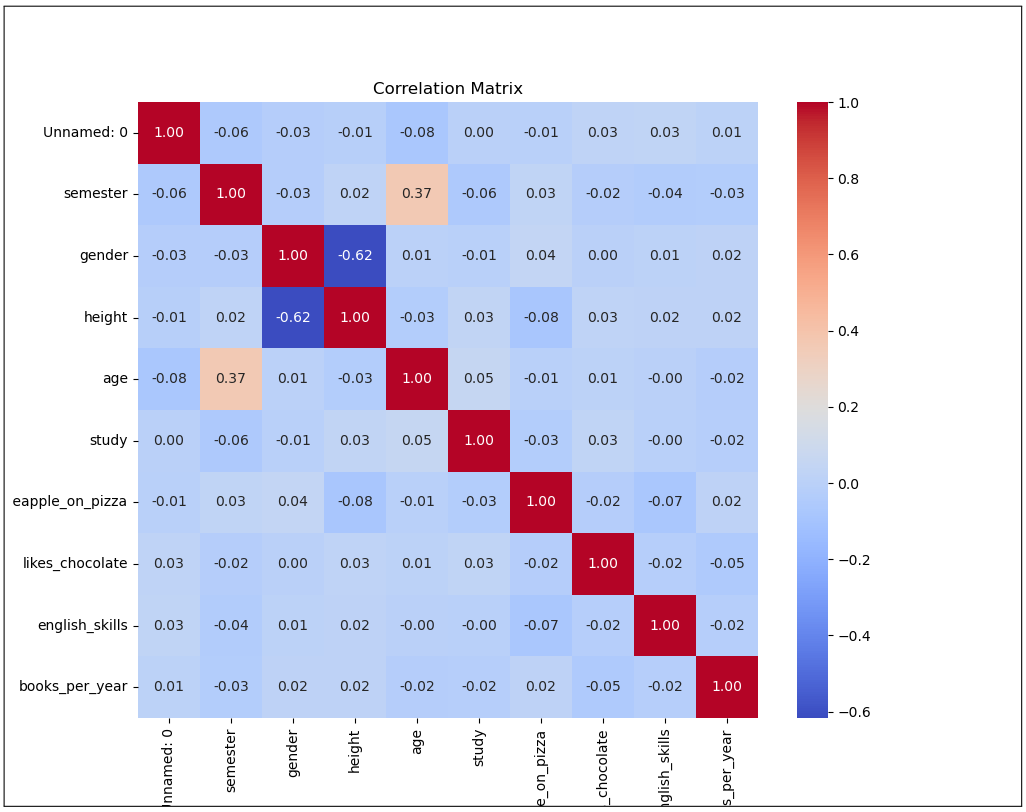

Figure 7: Heatmap for the Pearson's correlation between all the variables in the dataset. The upper or the lower part of the triangle can be ignored since they are mirrored by the diagonal line. A moderate correlation between gender and height can be observed, as well as, a weaker but still significant correlation between semester and age.

X	**Y**	**Type of dependency**
Gender	Height	Categorical
Semester	Age	Linear (Pearson's)

Semester and Age: Students who are have more semesters in the university tend to be older
Gender and Height: Man tend to be taller than Women

Answer (b) - Describe the reason for missingness

Variable	**Reason**
Semester	Bad user input (negative values and 99's). MAR
Height	Missing values from 184 onwards (N/A). NMAR
likes-pineapple-on-pizza	Not beeing socially accepted to like pineapple on pizza thus missing values. MCAR
likes-chocolate	Seems random/no pattern found. MCAR
english-skills	Seems random/no pattern found. MCAR
books-per-year	User forgot to input (blanks). MCAR

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

All the variables were imputed as a way to deal with the misisng values.

Variable	**Strategy**
Semester	Logistic regression to predict number of semester having the student age
Height	replace with mean
likes-pineapple-on-pizza	replace 0 and 1 with 0.0 and 1.0 respectively. Replace N/A with median
likes-chocolate	replace with median
english-skills	replace with average
books-per-year	replace blanks with 0's

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

Variable	**Before Strategy**	**After Strategy**
Semester	27.751	11.57
Height	172.171	172.171
likes-pineapple-on-pizza	0.318	0.24
likes-chocolate	0.814	0.879
english-skills	86.772	86.772
books-per-year	3.775	3.775