Coursework 1 pdf

1. Data preparation & understanding

1.1 Python code to prepare and understand the data

See data.py file

1.2 Original data set

See dataset.csv file

1.3 Prepared dataset

See dataset_prepared.csv

1.4 Explanation of code for preparation & understanding

The first action carried out to prepare the dataset was creating a pandas DataFrame. This was done using the read_csv function. Because the dataset.csv file contains some lines given general information about the data such as the title, data source, license and more, it is necessary for the dataframe to be created without these lines. Therefore, the argument skiprows= was added to the read_csv function to ignore these lines.

Once the dataframe was created, some general functions for information on the dataframe were used. This includes the functions shape, columns, head, tail, dtypes. These functions allowed me to get a general understanding of the dataset including its size, the columns involves and an idea of the lists within each column. The dtypes function was also especially useful to start to give an idea of what issues there may be with the dataframe. The dtypes function showed that the value column was categorised as a string whereas I expected its type to be an integer based on the first and last five rows. This suggested there were null values in the value columns which would present a problem.

To get an idea of how many data points may have issues, I first ran some code to see how many rows contain null values. This showed me that there were 43,712 rows with nulls so it will not be possible for me to analyse and rectify them individually to clean the dataset. With the ten rows will null values shown in the terminal, the null values were all in the 'Value' column. To confirm this, I used the isnull.sum function to know how many null values are in each column. This function confirmed that all the null values are in the 'Values' column. I have inferred that these null values are present where the Higher Education Institution did not collect or provide data for that metric. Because all the values are independent from each other, they can be categorised as Missing Completely At Random (MCAR). Because my dataset is so large with over 245,000 rows and 18% of the rows have missing values, I decided to delete the rows with missing data. Deleting these rows won't have an impact on any of the other rows. This can be seen in the prepare_df function.

Continuing the data preparation of the dataset, I decided to improve the table column. From the HESA website where the dataset was downloaded from, the data is split into five tables based on the 'Class' of the data as shown in Table 1 below When the dataset is downloaded, there is no column for 'Class', there is only a column for the 'Table' value. When it's time to create the dashboard, I felt as though it would be more useful and intuitive for me to have a 'Class' column rather than a 'Table' one. Therefore, I decided to rename the 'Table' column to 'Class' and replace all the values for Table for their corresponding 'Class'.

Table 1 'Table' Column in raw dataset and corresponding 'Class' column in prepared dataset

Table	**Class**
Table-1	Building and spaces
Table-2	Energy
Table-3	Emissions and waste
Table-4	Transport and environment
Table-5	Finances and people

2. Product definition

2.1 Product overview

For staff and students of Higher Education Institutions (HEIs) as well as researcher who want to compare the environmental impact of HEIs. The HEI Environmental Dashboard is a data visualisation dashboard that allows user to easily compare and rank a wide range of environmental metrics. Unlike impact ranking tables which have a methodology to determine environmental impact of HEIs, our product will allow easy visualisation of raw UK HEI data and allows for exploration of correlations and patterns.

2.2. Persona

3. Tools & Techniques

3.1 Source code control

Github repository: https://github.com/ucl-comp0035/comp0035-cwi-4jjnaomi

3.2 Use of AI