**Section 1: Statistics**

**Question 1:** Summarise your dataset clearly, using relevant descriptive statistics and appropriate plots. These should be carefully motivated and justified, and clearly presented. You should critically analyse your findings, in addition to including the necessary Python code, output and plots in the report. You are required to plot at least three graphs. [0-35]

a) Calculate the central tendency

A table with numbers and a number on it

Description automatically generated

**(Figure 1)**

(i) Getting Mean

A screenshot of a computer

Description automatically generated

**(Figure 2)**

(ii) Getting Mode

Results not relevant due to nature of df

(iii) Getting Median

A screenshot of a computer screen

Description automatically generated

**(Figure 3)**

b) Calculate the variance and standard deviation of features

(i) Variance

A screenshot of a computer

Description automatically generated

**(Figure 4)**

(ii) Standard deviation

A white background with black text

Description automatically generated

(**Figure 5)**

c) Compare and give interpretation to these results.

Histograms –

A graph of different types of data

Description automatically generated with medium confidence

**(Figure 6)**

Boxplots -

A graph with a bar graph and numbers

Description automatically generated with medium confidence

**(Figure 7)**

A graph with a blue rectangle and black squares

Description automatically generated

**(Figure 8)**

A graph with a bar and a line

Description automatically generated with medium confidence

**(Figure 9)**

A graph of a number of people

Description automatically generated

**(Figure 10)**

A graph with a bar graph

Description automatically generated with medium confidence

**(Figure 11)**

A graph of disposable income

Description automatically generated

**(Figure 12)**

**Question 2:** Use two discrete distributions (Binomial and/or Poisson) in order to explain/identify some information about your dataset. You must explain your reasoning and the techniques you have used. Visualise your data and explain what happens with the large samples in these cases. You must work with Python and your mathematical reasoning must be documented in your report. [0-30]

Create a new discrete variable class which will categorise our data into three categories which analyses tourist attractions. This will be done by classifying the three of the ‘tourism’ variables that we have included –

* Total Activities
* Total Accommodation
* Total Attractions

By using the quantile values of central tendency measures, we can calculate how each county ranks in relation to each other and then create discrete categories for each of the values above. This is done calculating the values of each on the 33% and 67% quantiles. Our three new discrete variables are –

• Activities Range (High, Middle, Low)

• Accommodation Range (High, Middle, Low)

• Attractions Range (High, Middle, Low)

Now that we have our discrete variables, we can use binomial distribution to analyse the probability of occurrence within our dataset.

The first problem we will look at is, assessing what the probability of occurrence that a country with a high level of household income will also have a high level of amenities across our three ranges.

We will do this by

* Deciding success parameters
* Calculating our number of test cases (n)
* Calculating the number of outcomes we care about (k)
* Calculating the probability of these outcomes (p)

Our question is to –

Calculate the probability of a County having a high Household Income (million EUR) if it has a High 'Accommodation Range' Using Binomial Probability Mass Function (PMF)

**Question 1:** Calculate the probability of a County having an above average Household Income (million EUR) if it has a High 'Accommodation Range' using Binomial Probability Mass Function (PMF)

k = **3**

n = **26**

p = **0.115384**

A screenshot of a computer

Description automatically generated

**(Figure 13)**

Probability of a County having 'High' accommodation range and an above average Household Income: **0.238105**

**Question 2:** Calculate the probability of a County having a below average Disposable Income(pp) if it has a low 'Activity Range' using Binomial Probability Mass Function (PMF)

k = **7**

n = **26**

p = **0.269230**

A screenshot of a computer

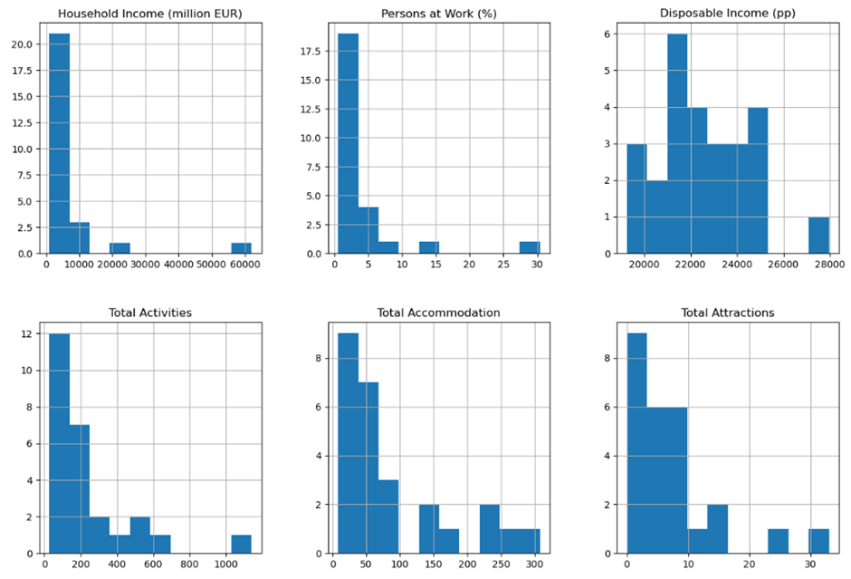
Description automatically generated

**(Figure 14)**

Probability of a County having 'Low' activity range and a below average Household Income: **0.174097**

**Question 3:** Use Normal distribution to explain or identify some information about your dataset. [0-20]

First we will need to look at the distribution of our data and see which of our variables have normal or skewed distribution levels.



**(Figure 15)**

From the above we can see that the distribution of all of our 6 variables is skewed, except for one – ‘Disposable Income (pp)’ . Let us take a closer look at this variable and it’s distribution –

A graph of disposable income

Description automatically generated

**(Figure 16)**

We can test this theory on our dataset which looks like it has a standard distribution by formulating questions around this theory and our dataset, we could look at something like the two questions below -

**Question 1:** What is the probability that our Disposable income (pp) is less than €24,000?

First, we can calculate the mean and standard deviation of our desired variable to get our mu and sigma values and then use the scipy norm cdf (cumulative distribution function) to find our answer -

mu = **22626.134615**

sigma = **2028.213264**

A screenshot of a computer code

Description automatically generated

**(Figure 17)**

Answer = **0.750916 %**

**Question 2:** What is the probability that our Disposable income (pp) is between €24,000 and €28,000 (the right-hand side of our graph)?

First we need to calculate the values less than or equal to €28,000 (using the same method as above)

mu= **22626.134615**

sigma= **2028.213264**

A screenshot of a computer

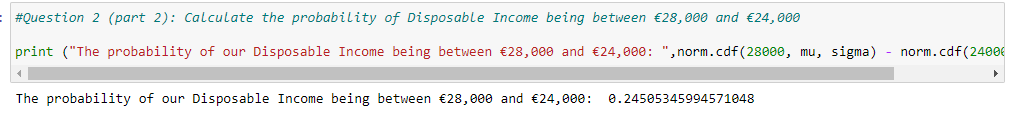
Description automatically generated

**(Figure 18)**

Answer **= 0.995970 %**

Next we need to calculate the difference between our two results by using the following –

norm.cdf (amount 1, mu, sigma) - norm.cdf (amount 2, mu, sigma)



Answer **= 0.245053 %**

**Question 4:** Explain the importance of the distributions used in point 3 and 4 in your analysis. Justify the choice of the variables and explain if the variables used for the discrete distributions could be used as normal distribution in this case. [0-15]

**Section 2: Data Preparation and Visualization**

**1. Exploratory Data Analysis:**

**Question 1**. You must perform appropriate EDA on your dataset, rationalizing and detailing why you chose the specific methods and what insight you gained. **[0-20]**

This project aims to infer the correlations between common tourist amenities and economic factors specific to local regions (Counties) of Ireland. At the outset of this task, it was clear that several datasets would need to be combined to create data that might offer insights into this problem. This being the case, EDA was utilised across all of these datasets in different ways and for different reasons. Below is a sample of these investigations, the methods that were used and the insights gained through this process.

**Head:**

A screenshot of a computer

Description automatically generated

**(Figure 19)**

**Shape:**

A screenshot of a computer

Description automatically generated

**(Figure 19)**

**Describe:** By getting the central tendencies and other information from our data we can better understand what other steps may need to be taken for efficient data preparation.

A screenshot of a computer screen

Description automatically generated

**(Figure 19)**

**Info:**

A screenshot of a computer

Description automatically generated

**(Figure 19)**

**Null / NaN values:**

A screenshot of a computer

Description automatically generated

**(Figure 19)**

**Unique Values:**

A screenshot of a computer program

Description automatically generated

**(Figure 19)**

**Cleaning:**

The first thing that we must do when dealing with a dataset of this size is to clean it. This should be done at the beginning, once you have a general idea of the issues that you may be faced with in the dataset. The data used here was cleaned by

**Dropping unneeded columns:**

A white background with red text

Description automatically generated

**(Figure 19)**

**Renaming features for ease of use:**

A screenshot of a computer code

Description automatically generated

**(Figure 19)**

**Data cross validation:**

A screenshot of a computer code

Description automatically generated

**(Figure 19)**

**Merging Data:**

Due to the nature of the problem we are trying to solve, it is necessary to merge several data sets. For this project, elements from the four data sets below will be used –

1. Activities2021
2. Accomdation2021
3. Attractions2021
4. Household\_Income2021

After the target variables were identified, the following methods were used to combine data from these sources –

**Feature summation:**

A screen shot of a computer code

Description automatically generated

**(Figure 19)**

**Array concatenation:**

A computer screen shot of text

Description automatically generated

**(Figure 19)**

**Dataframe Merging:**

A screenshot of a computer code

Description automatically generated

**(Figure 19)**

**2. Preparing data for Machine Learning**

**Question 2.** You must also rationalise, justify, and detail all the methods used to prepare the data for ML (Scaling, Encoding, imputation etc…). **[0-40]**

**Five Elements of Pre-processing and Feature Engineering for ML:**

1. Imputation

2. Encoding

3. Scaling

4. Feature creation

5. Feature Selection

**3. Insight through Visualisation**

**Question 3.** Appropriate visualizations must be used to engender insight into the dataset and to illustrate your final insights gained in your analysis. **[0-20]**

**4. Visualization Design and Tufts Principles**

**Question 4.** All design and implementation of your visualizations must be justified and detailed in full., referring to Tufts Principles **[0-20]**

**Section 3: Machine Learning**

**1. Project Management Framework**

**Question 1:** Explain which project management framework (CRISP-DM, KDD or SEMMA) is required for a data science project. Discuss and justify with real-life scenarios. Provide an explanation of why you chose a supervised, unsupervised, or semi-supervised machine learning technique for the dataset you used for ML modeling. **[0 - 20]**

**Part 1:** PM Framework: CRISP-DM

**Part 2:** ML Technique used (supervised, unsupervised, semi-supervised)

**2. Machine Learning Models**

**Question 2:** Machine learning models have a wide range of uses, including prediction, classification, and clustering. It is advised that you assess several approaches (at least two), choose appropriate hyperparameters for the optimal outcomes of Machine Learning models using an approach of hyperparameter tunning, such as GridSearchCV or RandomizedSearchCV. **[0 - 30]**

**Part 1:** Select ML Models

**Part 2:** Select Hyperparameters and test

**Part 3:** Use Gridsearch Cross Validation to hypertune parameters

**3. ML Model Results**

**Question 3:** Show the results of two or more ML modeling comparisons in a table or graph format. Review and critically examine the machine learning models' performance based on the selected metric for supervised, unsupervised, and semi-supervised approaches. **[0 - 30]**

**Part 1:** Create table of results

**Part 2:** Examine and discuss results

**4. Result Interpretation and Findings**

**Question 4:** Demonstrate the similarities and differences between your Machine Learning modelling results using the tables or visualizations. Provide a report along with an explanation and interpretation of the relevance and effectiveness of your findings. **[0 - 20]**

**Part 1:** Explain results and interpret

**Part 2:** Conclude report on ML

**Section 4: Programming**

**1. Code Justification and Explanation**

**Question 1:** The project must be explored programmatically; this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook. Your codebook should be properly annotated. The project documentation must include sound justifications and explanation of your code choices (code quality standards should also be applied). **[0-50]**

**2. Programming Paradigms**

**Question 2:** In a dedicated section in your report, discuss your use of aspects of various programming paradigms in the development of your project. For example, this may include (but is not limited to) how they influenced your design decisions or how they helped you solve problems. Note that marks may not be awarded if the discussion does not involve your specific project. [0-50]