**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

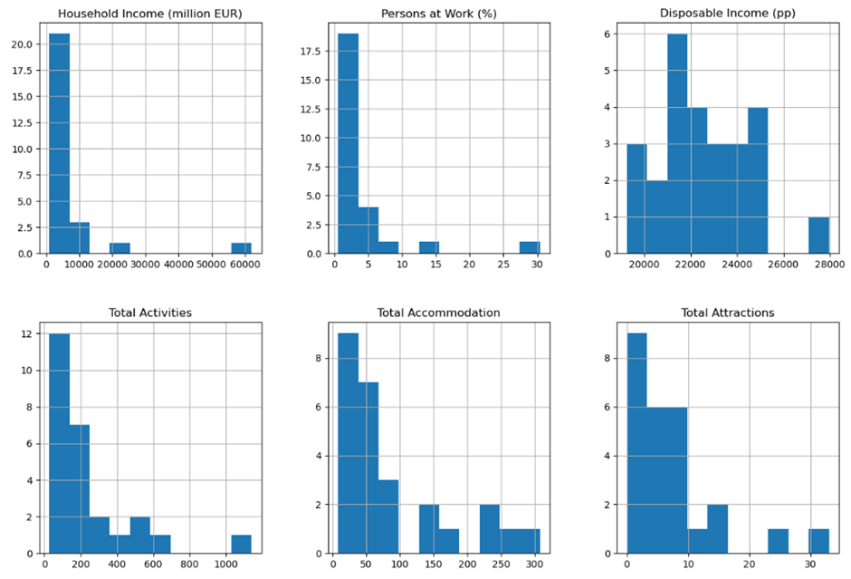
Description automatically generated

**(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

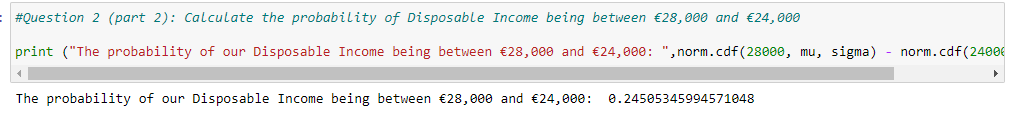
Description automatically generated

**(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]