**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

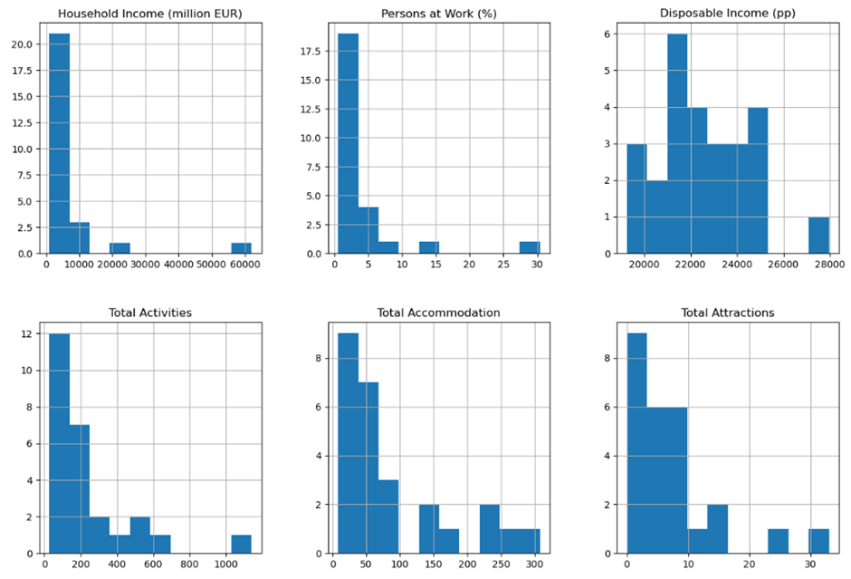
Description automatically generated

**(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

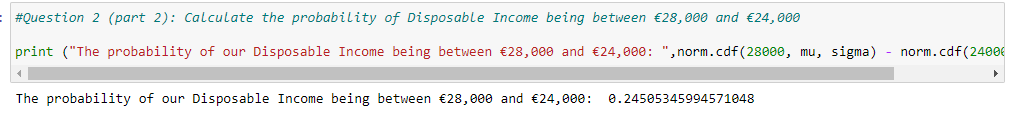
Description automatically generated

**(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.

Characterising the Dataset

When dealing with a new dataset like this, it is important to do some preliminary investigations to gain a better understanding of the type of data that you will be dealing If this is done thoroughly it can give an analyst a better idea of the challenges within the dataset. We can do this by using various methods such as the ones listed below -

shape(): This tells us how many features and entries are contained in our data

head() : This allows us to see a sample of the first rows of the data

info() : This provides a summary of information about the data

describe() : This provides us with the statistical details of the data

nunique(): This helps us to identify if a column is continuous or categorical

**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)**

The process above will give us a good idea of the type of data that we will be handling in this project and there is no substitute for these types of investigations, especially when getting familiar with a new dataset (Müller and Guido, 2017).

**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. Feature creation**

As mentioned above, several data sets have been combined to create the dataset that is being used within this project. From that, we have also created new features from the data sets that have been used. This has been done in the three following ways –

* Feature summation
* Array concatenation
* Data frame Merging

**2. Feature Selection**

The features that we will focus on for the rest of this project will be the ones created and merged in the section above and they are the following –

* Household Income (million EUR)
* Persons at Work (%)
* Disposable Income (pp)
* Total Activities
* Total Accommodation
* Total Attractions

By having an even split of three ‘Economic’ factor features and three ‘Tourism’ factor features we should have enough scope for the investigation outlined in the original problem.

**3. Imputation**

Imputation is an important process in data preparation, it aims to find missing values within a dataset and then a decision needs to be made by the author as to which method will be used to replace or remove these missing or incomplete values.

Looking at our primary datasets used, we must test for missing values. This was tested by looking at the variable that we are targeting (AddressRegion) and making sure that it contains the same range of value entries (no NaN or nulls) as our total dataset.

1. Activities2021

A screenshot of a computer

Description automatically generated

**(Figure 19)**

2, Accomdation2021

A screenshot of a computer

Description automatically generated

**(Figure 19)**

3. Attractions2021

A screenshot of a computer

Description automatically generated

**(Figure 19)**

Within our dataset, we did encounter empty values for two of the Counties in our ‘Attraction’ feature (Offaly and Monaghan) -

A screenshot of a computer

Description automatically generated

**(Figure 19)**

These were automatically dropped in the process of summing all values (due to 0+0 not returning a value) and because of this absence they needed to be manually entered into our dataset with values equalling to zero.

We have also looked at our ‘Economic’ categories from our dataset in relation to finding outliers that may need to be taken into consideration.

A graph of a number of household income

Description automatically generated

**(Figure 19)**

A graph with a blue rectangular bar

Description automatically generated

**(Figure 19)**

A graph of disposable income

Description automatically generated

**(Figure 19)**

In the first two boxplots above (figure 25 and figure 26) we can see that there is a lot of rightward skew which indicates to us that the skew is positive and that the majority of extreme (higher data points) taper off more slowly. The third boxplot (figure 27) is the most normal distribution of the three with the mean appearing close to the centre of the box and just to the left of the middle of the plot. While we can infer that there are extreme values within the first two plots which are causing this skew, it is not appropriate to our problem to remove or impute these outlying values. As per the testing done previously within the preliminary EDA and the above, no imputation was needed for this project.

**4. Encoding**

As we are preparing our dataset for the application to machine learning models, we will need to convert all the categorical variables that we want to look at into numerical variables. We will apply two types of encoding to our data, ordinal/label encoding and one hot encoding.

**(i) Ordinal Encoding:**

We have three features that we created by finding the statistical ranges of the entries and grouping our entries based on their relative values. These three features are –

* Activity Range
* Accommodation Range
* Attraction Range

Within these we have three values (based on percentiles) of High, Middle and Low. These are ordinal values as they denote if an entry has a specific number of a Tourist amenity, so for this encoding we can use ordinal encoding with High =3, Middle =2 and Low =1.

We can do this by using the Ordinal Encoder from sk learn preprocessing library and fitting it to our target variables (Activity Range, Accommodation Range , Attraction Range) and mapping our new values to the range and replacing the existing values with the mapped ones. This gives us the result below -

A white background with black text

Description automatically generated

**(Figure 19)**

**(ii) One Hot Encoding:**

For our remaining categorical data feature (County) a decision was made to use One-hot encoding to encode because it will allow us to encode data in a way that is non-ordinal, preserve the descriptive parameters of our data and let us handle nominal categories that are fit for machine learning purposes ( ref \* Page 78, Feature Engineering for Machine Learning, 2018.) One of the major setbacks of OHE is the increase in the dimensionality of the dataset but as we have used a targeted approach to create our data frame from combined datasets, we are starting with a dataset that isn’t very large so dimensionality shouldn’t be an issue.

We apply our one-hot encoding by first using the Pandas get-dummies function which will create a new feature to our data frame for each entry contained within our target feature of County and assign a value of True or False where this is present –

A table of text on a white background

Description automatically generated

**(Figure 19)**

Once this has been done, we will then have to convert these True/False values into their binary representatives 1 and 0. We can do that by using the Pandas ‘astype(int)’ function. That will give us the following result –

A screenshot of a number table

Description automatically generated

**(Figure 19)**

Now that our data has been transformed with numerical encoding, we can move on to scaling and get closer to apply machine learning models to our problem.

**3. Scaling**

As we have used OHE to encode our data, we can scale our data afterwards without any negative side affects ( \*ref)

Now we need to assess our data to identify if any scaling might be needed, we can do this by looking at the central tendencies of our features.

A table with numbers and text

Description automatically generated

**(Figure 19)**

From the above (figure 31) we can see that the six features that we are looking at, only ‘Disposable Income (pp)’ has a mean that is close to the 50% percentile mark which indicates that it is likely normally distributed. The other five features all have some degree of variance and as such, are likely skewed and not normally distributed. We can remind ourselves of the distribution to prove this –

A group of blue and white graphs

Description automatically generated

**(Figure 19)**

**Standard Scaler:**

In the first case, we will take our feature that displays normal distribution ‘Disposable Income (pp)’, and apply the SkLearn Preprocessing Standard Scaler to it to create a new feature in our dataframe called ‘Dis\_Income\_Stand’

The standard scaler will centre the new mean of the data at 0 and apply a standard deviation of 1.

**MinMax Scaler:**

Then for our remaining five features (Household Income (million EUR), Persons at Work (%), Disposable Income (pp), Total Activities, Total Accommodation, Total Attractions) that have high levels of rightward skew, we will apply the MinMax scaler to create new corresponding features for each of these.

This will map a new range of values to our features based on 0 being the lowest value in the range and 1 being the highest.

When our data has been transformed it will look like this –

A screenshot of a computer

Description automatically generated

**(Figure 19)**

Although the visual distribution of values will remain the same, we have reduced the dimension of range between our values so that they are in a scale that has a closer proportionality to each other. Our data is now in a form that it can be used for Machine Learning.

**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]**

Through the process of exploratory data analysis and data preprocessing, several key insights that have been gained. These have been documented along the way, but below are the ones that have particular significance or illustrate meaningful insights.

**Scatter Plots:**

**Scatter/Bubble Plot:**

**Heat Map:**

**Density Plot/Histogram:**

**Choropleths:**

A final interesting visualisation that we can look at for our data is the geospatial choropleth plot. This allows us to conceptualise our data for all 26 Counties of the Republic of Ireland at once in a way relative to each other. Like the heatmap, this plot gives the viewer of the weight of the data value based on the intensity of colour. While it is impossible to determine the exact figures we are looking at, it is a great visual aid for broad data conceptualisation.

A map of ireland with a green area

Description automatically generatedA green map of ireland

Description automatically generated

A map of ireland with different shades of green

Description automatically generatedA map of ireland with different colored areas

Description automatically generated **(Figure 19) (Figure 19)**

**( Figure 19) (Figure 19)**

A map of ireland with different shades of green

Description automatically generatedA map of ireland with different colored areas

Description automatically generated

**(Figure 19) (Figure 19)**

**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]**

**Tufts Six principles of graphical integrity that have been considered and utilised in this project are –**

**1. The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities measured.**

**2. Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data**

**3. Show data variation, not design variation**

**4. In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units.**

**5. The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data**

**6. Graphics must not quote data out of context (page 74).**

These have been utilised in the following ways

**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

CRISP-DM ( Cross-Industry Standard Process for Data Mining) has been selected as the project management framework for the implementation of this project for the following reasons –

* Structured approach to data collection
* Experimentation and refinement at core of approach
* Iterativie approach to deploying machine learning models

CRISP-DM operates along the distinct phase-based iteration of projects, which is very suitable for this project as the author has been required to satisfy individual component elements from statistics, data preparation and visualisation, machine learning and programming. The data being used for each of these individual parts of the project needs to be constantly updated and re-assessed so this iterative approach has been ideal to manage these considerations. This has allowed the author to produce a clean and structured data set that can now be taken forward for applying to machine learning algorithms to address the problem outlined at the beginning of this project.

The data modelling and feature engineering of this project has been a focused process which has required the author to constantly experiment and re-evaluate at all stages. CRISP-DM has allowed this to be conducted in an efficient manner while allowing the flexibility needed for ongoing testing.

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

The question that we are going to look at for this project is ‘Can we try and predict the 'Total Activities' of a County by the additional features that we have selected using machine learning models?’

As this problem will use continuous data that we already have labelled, we will use supervised machine learning approaches. This problem is a regression task due to the nature of the target variable 'Total Activities', which is continuous. Our main goal will be to develop two machine learning models which can understand the relationships between our selected (independent) features and the target variable. These models aim to produce precise predictions regarding the 'Total Activities' based on the chosen relationships.

**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]**

The three machine learning models that we will use to address our problem are:

* Random Forest
* Linear Regression
* Support Vector Regression

We will develop and fit our models with test and train data and select hyperparameters to obtain preliminary results. Once this has been done, we will utilise GridSearchCV to hyper-tune these parameters and then get new results for our models and compare and examine them in a comparative results table.

The metrics that we will use to assess our machine learning models are -

MSE

MAE

R2

These have been chosen based on the research available.

**Part 2:** Select Hyperparameters and test

**Random Forest:**

First we will look at the random forest model.

**Linear Regression:**

**Support Vector Regression:**

**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

**Comparative Results Tables**

**Evaluation Metric: R2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Default**  **Tune** | **Manual**  **Tune** | **GridSearchCV**  **Tune** | **Scaled and Optimised** |
| Linear Regression | 0.2057 | 0.1466 | n/a | 0.2057 |
| Random  Forest | **0.3953** | **0.2915** | **0.46147** |  |
| Support Vector Regression | -1.4181 | -1.4137 |  |  |

**Evaluation Metric: MAE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Default**  **Tune** | **Manual**  **Tune** | **GridSearchCV**  **Tune** | **Scaled and Optimised** |
| Linear Regression | 38.7011 | 39.1208 | n/a | 0.0347 |
| Random  Forest | **31.6866** | **31.3138** | **26.4331** |  |
| Support Vector Regression | 61.9804 | 61.9609 |  |  |

**Evaluation Metric: MSE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Default**  **Tune** | **Manual**  **Tune** | **GridSearchCV**  **Tune** | **Scaled and Optimised** |
| Linear Regression | 2020.9854 | **2171.3883** | n/a | 0.0016 |
| Random  Forest | **1538.5821** | 2195.8576 | **1370.3093** |  |
| Support Vector Regression | 6153.2259 | 6141.9810 |  |  |

**(Table 1)**

**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]