THEORY QUESTIONS ASSIGNMENT

Data Science Stream

Maximum score: 100

Maximum score: 100

KEY NOTES

- This assignment to be completed at student’s own pace and submitted before given deadline.
- There are 10 questions in total and each question is marked on a scale 1 to 10. The maximum possible grade for this assignment is 100 points.
- Students are welcome to use any online or written resources to answer these questions.
- The answers need to be explained clearly and illustrated with relevant examples where necessary. Your examples can include code snippets, diagrams or any other evidence-based representation of your answer.

Theory questions	**10 point each**

1. What does “Data Cleansing” mean? What are the best ways to practice this?
2. What is the difference between data profiling and data mining?
3. Define Outlier with an example.
4. What is “Collaborative Filtering”?
5. What is “Time Series Analysis”?
6. Explain the core steps of a Data Analysis project?
7. What are the characteristics of a good data model?
8. Explain and provide examples of univariate, bivariate, and multivariate analysis?
9. What is a Linear Regression?
10. In terms of modelling data, what do we mean by Over-fitting and Under-fitting?

My Answer:

1. What does “Data Cleansing” mean? What are the best ways to practice this?

Data cleansing is the process of removing unwanted elements from datasets.
This can improve the accuracy and quality of succeeding analysis.
The best way to practice Data Cleansing is by:

1. Examining the **Data Quality** with:
    a. **Validity** - Is the data align with defined business rules or restrictions?
    b. **Accuracy** - How close the data to the standard measure?
    c. **Completeness** - How much missing data there is?
    d. **Consistency** - Are the values within the data set match up and are not contradictory?
    e. **Uniforminity** - Does the data have particular same unit of measure or format?
2. **Workflow** process:
    a. **Inspection** - determine the relationship between elements, inspect for unexpected data format and look for erroneous data.
    b. **Cleaning** - clean and fix the irrelevant data, missing values, duplicates, outliers, convert necessary format, standardize and normalize data.
    c. **Verifying** - the results are inspected after cleaning to make sure everything is correct.
    d. **Reporting** - What changes made with the data?

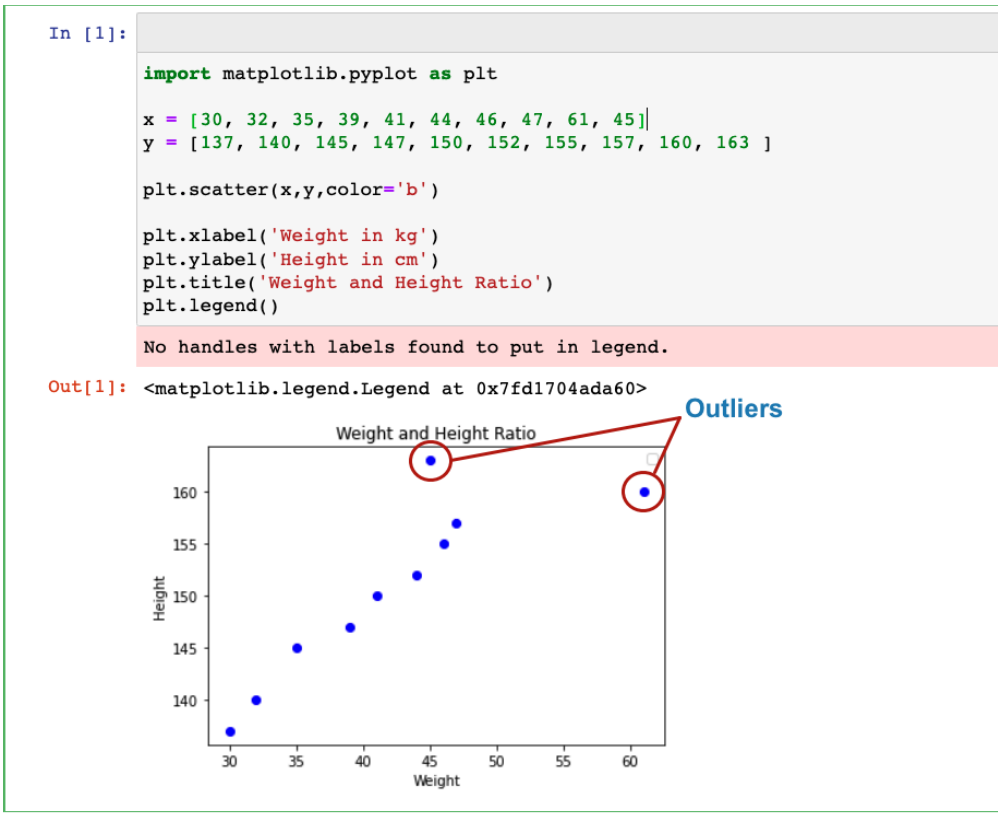
2. What is the difference between data profiling and data mining?

Data Profiling	**Data mining**
Gathers details about data quality to find irregularities in the dataset	Extracts useful information using complicated mathematical algorithms
Kinds:	Techniques:
● Structure discovery ● Content discovery ● Relationship discovery	● Association learning ● Classification technique ● Clustering technique ● Prediction technique ● Sequential patterns
Data mining techniques are applicable to data profiling. Both might have similar concepts in data cleaning and data preparation, but the ultimate goal of each process makes them distinct from one another	

3. Define Outlier with an example.

An **outlier** is a data point that does not conform to the general pattern of the rest of the data set.
It is an abnormality or deviation from what is expected.

For example below, I have a dataset where I have values on my X-axis with "Weight" and on my Y-axis with "Height" which showing the Weight and Height ratio. In normal scenarios, when the value of weight increases, the height increases. But when there's a value that is far from the corresponding value, in which that case is a very rare condition and you don't have that kind of value normally in a dataset, then it is an outlier.



In [1]:

```
import matplotlib.pyplot as plt

x = [30, 32, 35, 39, 41, 44, 46, 47, 61, 45]
y = [137, 140, 145, 147, 150, 152, 155, 157, 160, 163 ]

plt.scatter(x,y,color='b')

plt.xlabel('Weight in kg')
plt.ylabel('Height in cm')
plt.title('Weight and Height Ratio')
plt.legend()
```

```
No handles with labels found to put in legend.
```

Out[1]:

```
<matplotlib.legend.Legend at 0x7fd1704ada60>
```

4. What is “Collaborative Filtering”?

The **Collaborative Filtering** concept is that past similar preferences can help predict future preferences. It works by looking at a large group of people and finding a smaller set of users whose taste is similar to the particular user in question. By examining the items they like, we can then create a ranked list of suggestions.

5. What is “Time Series Analysis”?

Time series analysis is a particular method of examining data points that have been collected over time which take measurements at regular intervals instead of just recording sporadic data points. It can be very useful in spotting trends and patterns, which can then be used to make informed business decisions or forecasting.

6. Explain the core steps of a Data Analysis project?

Core steps of a Data Analysis project:

1. **Defining the problem**: coming up with a clear and concise question that you want to answer with your data. This defines the objective of what we want to achieve in a Data Analysis project.
2. **Collecting the data**: gathering all relevant information that can be used to answer your question.
3. **Cleaning the data**: making sure that your dataset is free of any errors or inaccuracies which could lead to incorrect results later on.
4. **Analyzing the data**: using various statistical methods in order to find trends, patterns, and relationships within your dataset which will help you answer your original question.
5. **Sharing your results**: presenting what you've found in a comprehensible way so others can understand it as well.
   *__Embracing failure__: Not every experiment turns out perfectly – sometimes analyzing the data can reveal unforeseen complexities or roadblocks . Accepting this fact is crucial for continued learning and progress.

Core Steps of a Data Analysis Project

STEP 1 Defining the problem

STEP 2 Collecting the data

STEP 3 Cleaning the data

STEP 4 Analyzing the data

STEP 5 Sharing your results

7. What are the characteristics of a good data model?

Characteristics of a good date model:

1. A good data model is easy to understand and use.
2. Changes in big data do not negatively impact the performance of the model, making it reliable.
3. It provides foreseeable results so you know what to expect from them.
4. They can also adapt to changing requirements without compromising on quality or performance.

8. Explain and provide examples of univariate, bivariate, and multivariate analysis?

Type of Analysis	Description	Statistical Technique to conduct analysis	Example
Univariate analysis	● Refers to a single variable. ● The simplest form of analysis involves describing patterns within the data, without considering any possible causes or relationships.	● Frequency Distribution Tables ● Histograms ● Frequency Polygons ● Pie Charts ● Bar Charts	An example of univariate analysis would be if we looked at the weight of the cabin crew. In this case, the data would just reflect a single variable (weight) and its quantity. The main goal of this kind of analysis is to describe the data in order to find patterns within it.
Bivariate analysis	● Involves two different variables. ● This type of analysis is concerned with discovering potential causes and relationships between the two variables being studied.	● Correlation coefficients ● Regression analysis	An example of bivariate analysis is examining the correlation between two variables, such as gender and weight of cabin crew. In this case, we would have two variables - gender (the independent variable) and results (the dependent variable). This would allow us to measure the relationship between the two variables.
Multivariate analysis	● Refers to data that involves three or more variables.	● Factor Analysis ● Cluster Analysis ● Variance Analysis ● Discriminant Analysis ● Multidimensional Scaling ● Principal Component Analysis ● Redundancy Analysis	When looking at multivariate data, one example would be the relationship between weight, gender and race of cabin crew members. By analyzing multiple variables simultaneously, patterns and trends can be more easily uncovered than if only examining a single variable at a time.

9. What is a Linear Regression?

Linear regression is a statistical method that is commonly used to estimate the relationships between two variables.
In particular, it can be used to quantify the relationship between one dependent variable and multiple independent variables.

10. In terms of modelling data, what do we mean by Over-fitting and Under-fitting?

Over-fitting	is when a model performs well on training data but poorly on test data.	This happens when the model doesn't generalize well to new data.
Under-fitting	is when a model doesn't perform well on either training or test data.	This usually happens when the model is too simple and is unable to capture the complex patterns in the data.