CODE FIRST GIRLS

CFGDEGREE

DATA ASSESSMENT MATERIAL RELEASE

THEORY QUESTIONS

SECTION	MARK
1. Theory Questions	25
2. Pandas Questions	25
3. Matplotlib Challenge	25
4. Numpy Questions	25
TOTAL	100

Important notes:

- This document shares the first section of the Data Assessment which is composed of 5 Data Theory Questions
- The answers do not have to be long, but they have to answer each of the mention points for each question
- It is worth a quarter of your assessment mark
- You have 24 hours before the assessment to prepare.
- If any plagiarism is found in how you choose to answer a question you will receivea 0 and the instance will be recorded.
- Consequences will occur if this is a repeated offence. You can remind yourself of the plagiarism policy here.
- You are allowed to use any online images to support your answers.

Section 1: Theory Questions [25 points]

1.1 In your own words, what does the role of a data scientist involve?	**2 points**
Answer: A data scientist analyses and interprets complex data to extract meaningful insights that can be used in a multitude of ways. This also includes data collection, pre-processing, visualisation and applying statistical model techniques to solve specific business problems, advance scientific and medical understanding or identify social/political trends.	

1.2 What is an outlier? Here we expect to see the following: a. Definition b. Examples c. Should outliers always be removed? Why? d. What are other possible issues that you can find in a dataset?	**4 points**
Answer: **a. An outlier is a unit of data that is significantly different from the others in a particular dataset. This can indicate an error in the data or a fringe case.** **b. This could be a substantially higher value for alcohol percentage in a wine dataset.** **c. Whether to remove or not depends on the context; it is up to the analyst to investigate. If it is a statistical anomaly- like a country with particularly high carbon emissions in comparison with their neighbours, there may be value in understanding why they are an outlier. Removing that country might decrease the quality of the dataset.** **d. Other dataset issues include missing values, duplicate entries, and irrelevant data.**	

1.3 Describe the concepts of data cleaning and data quality. Here we expect to see the following: a. What is data cleaning? b. Why is data cleaning important? c. What type of mistakes do we expect to commonly see in datasets?	**4 points**

Answer: **a. Data cleaning is the process of correcting/altering/normalising/changing the type of data or removing incorrect, incomplete, or irrelevant data from datasets.** **b. It's essential for accuracy in analysis, as the program can only group things that share parameters. If the input is not clean, the output will be inaccurate. If it's not inn a recognisable state, it won't be counted by the system. Consistency ensures accuracy.** **c. Some common mistakes are: mismanaging missing values, inconsistent formatting, and not removing problematic outliers.**	

1.4 Discuss what is Unsupervised Learning - Clustering in Machine Learning using an example. Here we expect to see the following: a. Definition. b. When is it used? c. What is a possible real-world application of unsupervised learning? d. What are its main limitations?	**7.5 points**
Answer: **a. Unsupervised learning is a type of machine learning where algorithms identify patterns in a dataset without needing pre-existing labels. Clustering groups similar items with similar features.** **b. It is used when the data you have is unlabelled and you need to discover the patterns in the data.** **c. It's often used for market segmentation and customer profile analysis- where similar customer behaviours are grouped together. It can also be used for anomaly detection in network security.** **d. Limitations include incorrectly determining the number of clusters and interpreting results without external validation, therefore producing incorrect results.**	

1.5 Discuss what is Supervised Learning - Classification in Machine Learning using an example. Here we expect to see the following: a. Definition. b. When is it used? c. What is a possible real-world application of supervised learning? d. What data do we need for it? Is there any processing that needs to be done?	**7.5 points**
Answer: **a. Supervised learning involves training a model on labelled data to classify new data.** **b. It is used when classifying data into file types. E.g. Image, doc, text** **c. Used in email spam detection (classifying emails as spam or not).** **d. It requires accurate, representative labelled data. Pre-processing may likely need to be done. Data pre-processing includes normalization and dealing with missing or imbalanced data.**	