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**
The role of a data scientist involves machine learning and other statistical techniques and programming skills in order to provide insights from data. When compared against data analysts, data scientists use more complex and unstructured data and produce models to create predictions based on data.	

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**
An outlier is a result that is significantly different than all other results within a dataset and therefore deviates from the overal distribution and pattern of data, and this can distort the results so it may seem like there is an effect when there may not be. An example of an outlier would be if you have a set of heights for a group of people within a classroom of year 4 children and they have also included the height of the teacher, this is an outlier that would bring the average height up significantly. Outliers should not always necessarily be removed, as they can be important to understand how a dataset was collected, or what other factors may be at play that could effect the overall dataset. Therefore, sometimes it can be better to highlight the outliers but not necessarily remove them unless it is affecting your analysis and skewing the measures needed. Other than outliers, another possible issue you could have is null/missing data as this may lead to biased or incomplete conclusions. Also, a problem faced in data is duplicates, as you may come to conclusions in your analysis based on incorrect data and may overestimate an effect that may not be as important as expected as there may only be half the cases.	

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**

Data cleaning is the process of removing bad data from a dataset that may be inaccurate or incorrect. This includes but is not limited to removing null data, checking for and handling outliers, normalising data, removing duplicates, and much more. Data cleaning is important as it improves the integrity and reliability of our data and ensures we are measuring the true effect and not the effect of any errors.	

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**
Unsupervised learning is a type of machine learning that clusters and analyses data without any human intervention, this means it does not need set categories in order to analyse it and will create the categories for you. Clustering is an unsupervised technique that will group unlabeled data based on how similar or different it is, ie. it may look to group together movies on netlfix with similar themes. A limitation of unsupervised learning is the sensitivity to outliers. Some models may be sensitive to outliers and may over or underestimate the effect of a relationship whereas human intervention may be easily able to identify this. Therefore a way to correct this would be before entering data into a model, to do some preliminary manual checks to see if there are any outliers. Another limitation is the difficulty in validating the success of the model. For example, with supervised learning you can see clearly if something was sorted in correctly by seeing if the label fits, it may be harder to see in unsupervised learning whether the model successfully associated data without digging further into it. A way this could be addressed would be to split the data into smaller subsets and then see if the results are consistent across all models.	

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**
Supervised learning is a type of machine learning in which datasets are labelled and the model is then trained on recognising the input and output variables. Classification is a type of supervised learning in which the model will assign items into distinct categories. A real world application of supervised learning (classification) is the classification of email as either spam or not spam and filtering your email based on this. Before the model can be used, preprocessing is required in which the data is cleaned, normalised, the categories are encoded and then the data must be split into three sets of data. The data needed for supervised learning classification is a set of training data, validation data, and testing data. The input training data is labelled with a desired output and this is then used to create the model.	