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 is to collect, analyse and interpret the specified data. A Data Scientist solves a problem by using various techniques to spot patterns that may occur. Data Scientist can make educated predictions based off of what patterns may have been found.

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**

A:

An outlier is a data point which lies far from the rest of the dataset, this means that is lies far from the expected results, which lie around the median and mean. Outliers can be either significantly large or small in value.

B:

E.g. The Height of people that work at a certain company
If a person's height was significantly higher or lower than the average/expected height of 169/170cm then they can be considered an outliersince hey are drastically different in height, that of the general population.

E.g. The income of a Small Town
If the income of a few households don't align with the overall trend then this could be considered an outlier (if a household income is considerably much higher or lowe of than area) An example of such events could be: A widow who has lost their partner therefore the most logical explanation would be a n expected instant decrease of income of a half. Another situation could be that if within the area somebody won the lottery and the obvious result would be a massive increase in income.

E.g. School Run Competition performance
If a well known team that normally performs well, and let's say that this particular day the team drastically decreases in performance not only is it an outlier but we can also trace back and consider their previous plays/wins and therefore suggestv and explore other

factors of influence rather than pure skills which may be though upon first glance based on just the results obtaibned.

C:

Outlier should not always be removed this is because the decision should be based on the context of the problem at hand and the aims of the solution we need to produce. What is found to be the most important time to ensure that outliers are removed is within predictive analysis. An example of this would be when training ML (Machine Learning) models as these need accurate data in order to reproduce higher accuracy it is preferred to remove these outliers. If the data is not classified as an erroneous data type. Then it can be assumed that the is not a result of an error in when entering the data and therefore there is no need for data removal.

D:
Other issues include dat bias which is a good indication of the influence from where the orgination of that data collection. Therefore Data Scientists should make an effort to try to mitigate/reduce data bias where possible.

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**

A:

Data cleaning is the process where we identify and attempt to fix errors which may in our dataset. This involves managing missing values as well as the removal of supliacted data, this may also be alter formats to increase clarity and deal with outliers.

B:

Data cleaning is essential as these could produce inaccurate results resulting which negative impacts important decision making processes. By providing better accuracy data we can provide a correct insight to what the data may show.

C:

Most commonly in any data: outliers, inconsistency in the may data is formatted, errors occurred during the netery of data,

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**

A:
Unsupervised learning is where the dataset has no pre-determined labels/categories.
A type of unsupervised learning is clustering. This is where similar data points are grouped together by an algorithm which helps identify patterns found within the dataset as well as discover any characteristics.

B:
Unsupervised learning is where the dataset has no pre-determined labels/categories. This means that it is particularly useful in anomaly detection e.g. in fraud which is becoming ever importantly as our modern world develops.

C:

An example of real-world application would be when companies use unsupervised learning to better the retail side of the company . This is particularly useful as the algorithm produces clusters when customers are grouped based on their similarities. This especially useful for companies as they have bette rinight into who their top spenders may be. They could then provide further bundles and discount which will be more enticing. These is very useful knowledge to companies in order to strategically plan wher they can tailor tatics for marketing and sales.

D:

The main limitation is where the algorithm produce clusters the results may differ every time it is run.

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 unsupervised learning? d. What data do we need for it? Is there any processing that needs to be done?	**7.5 points**

A:
Supervised learning is where the dataset pre-determined labels/categories.
This is where the algorithm learns from existing labels and categories.
B:
Supervised learning is where the dataset has pre-determined labels/categories. This means that it is particularly useful in spam filtering where certain characteristics can be picked up on within mail, further determining that it is a spam email. Furthermore, supervised learning can also be used in image recognition e,g, within devices we use biometrics.

C:

An example of real-world application would be when companies use supervised learning to reduce the amount of unnecessary emails arriving to their inbox by using spam filtering which uses supervised learning.

D:

The main drawback is that the dataset provided must have clear/distinct labels in order to process the data. Again managing missing values is very important.