**Written Assessment**

*Instructions: Please answer all questions. Type in your answers directly below each question.*

*Please save your completed word document and upload it to the Student Portal (Lesson 7 -> Capstone Project Submission)*

**SECTION A**

Q1: What is Python, and why is it used for data science?

Python is an open source programming language created by Guido Van Rossum on 1991, it is flexible and easy to learn. Currently Python named as most in-demand coding language for data science, favoured by the data scientists and technology company used for their real-world application such as training the computer with its large data set. With its easy to learn nature, everyone can build the application via Python on top of being able to perform data analysis and data visualization.

Q2: What is a dictionary in Python?

Dictionary is an unordered collection of key-value pair in Python, it used to map information from keys to value. The dictionary key cannot be duplicated and values can be any data structure, 1 key only map to 1 value. Dictionary in Python covered in the brace symbol { }, the separator for key and values is a colon :

**SECTION B**

Q3: How do you access the last element of a list in Python?

Assuming the list is named as list\_numbers, to find the last element of a list, use Print(list\_numbers[-1]), the element will be displayed.

Q4: How do you check if a value is in a list in Python?

Assuming the list is named as list\_numbers, to check if a value is in a list, use print(x in list\_numbers) where x is the element to be checked, it displays True if it is found and False if it is not found, in Boolean form.

**SECTION C**

Q5: What is a NumPy array?

NumPy array is a grid of values, all of the same type, indexed by a tuple of nonnegative integers. It is multidimensional or n-dimensional array of fixed size with homogeneous elements. Rule of thumb for NumPy array is if you create arrays using the array module, all elements of the array must be the same type. NumPy arrays have the benefits of take up less space in size, faster speed than lists in performance, and NumPy has optimized functions such as linear algebra operations built in.

Q6: What is a Pandas DataFrame?

Pandas Dataframe offers powerful, expressive and flexible data structures that make data manipulation and analysis easy. It is a 2-dimensional labelled data structure with columns of potentially different types. Pandas Dataframe simplify the time consuming and repetitive tasks with offering data cleansing, data fill, data normalization, merges and joins, data visualization, statistical analysis, data inspection, loading and saving data functions, and more.

**SECTION D**

Q7: Why do we want to do data visualization?

Data visualization provide quick effective way to communicate information in a universal manner using visual information, it helps business identify which factors affect customer behaviour and pinpoint areas that need to be improved or need more attention.

Q8: Name any 4 common strategies in cleaning data?

Remove missing values, update incorrect values, impute missing or invalid data, deal with Outliers.

**SECTION E**

Q9: What is a correlation?

Correlation measures how strong the statistical relationship between two random variables, indicating how changes in one variable are associated with changes in another. The higher correlation, the stronger the linear relationship.

Q10: What is regression analysis, and how is it performed in Python?

Regression analysis is a set of statistical methods applied to estimate the relationship among variables. Regression is used when the target outputs are continuous with numerical outputs are expected. The regression analysis is used in Python to evaluate the model by import training data with instantiate feature extraction to feature matrix, and then fit to machine learning algorithm to build up the model in training, then the model use as input into predictive model in predicting phase, lastly the result is evaluated by Mean Squared Error (MSE) or R-squared, to compare actual result versus predicted result.

**SECTION F**

Q11: What is classification, and how is it performed in Python?

Classification is used when the target outputs are discrete, a Classification model requires training, validation and test data, input feature X, target features or output Y and measures of improvement. Python use several algorithm such as K-Nearest Neighbours (KNN), Logistic Regression Model and Decision Tree Model to perform the classification.

Q12: What is the KNN algorithm?

KNN algorithm, also named as K-Nearest Neighbour Algorithm is a non-parametric, lazy learning algorithm that predicts outcomes based on the similarity of inputted features to the training set. It doesn’t make assumption about the underlying distribution of the data, the training phase is minimal as KNN use all or nearly all of the training data, and how closely out-of-sample features resemble the training set determines how we classify a given data point.

**SECTION G**

Q13: What is the difference between precision and recall in machine learning?

In machine learning, the precision and recall are the performance metrics used for classification models of predicted classification versus actual classification. Precision is the true positive prediction values, related to the accuracy made by the model. Recall is the sensitivity to measure the model’s ability to correctly identify all relevant instances. The difference between precision and recall is the application, when the cost of False Positives (type I error) are high (eg email spam detection, marking legitimate email as spam) we shall consider precision metric; when the cost of False Negatives (type II Error) are high (eg fail to detect pregnancy), the recall metric shall be considered.

**End of Written Assessment**