Karatina University

BIT 316: Big Data Analytics (Assignment 3)

DATE: 22/10/2025 **TIME: 0800 hrs – 1100 hrs**

Requirements

- You need a computer installed with the following packages: *pandas, numpy, seaborn, matplotlib, wordcloud* and *nltk.*

Instructions: Answer Question 1 and Any Other Two Questions

Question 1: (30 Marks)

a) Consider the *Stores.csv* data-set.
   i. Read this data into Python Pandas as a DataFrame.
   (2 marks)
   ii. Using `apply`, write a function to determine the totals for the values in each column and save them in a new row called "Column_Totals".
   (3 marks)
   iii. Using `lambda` with `apply`, write a function to determine the totals for the values in each row and save them in a new column called "Row_Totals".
   (2 marks)

b) Consider the *Airbnb_Open_Data.csv* data-set.
   i. Scrutinize the country column for NaN values
   (2 marks)
   ii. If the country column has NaN values, delete from the data-set all records with these values.
   (3 marks)
   iii. Examine this data-set for records that are a replica of each other by selecting duplicate rows except the first one.
   (3 marks)
   iv. If duplicate values exist, prove that indeed at-least one record has been duplicated.
   (2 marks)
   v. If duplicate values exist, remove them. Justify your answer.
   (3 marks)
   vi. Generate a data-frame showing the number of listings in each neighborhood group.
   (3 marks)
   vii. Using `seaborn`, visualize the number of listings in the neighborhoods.
   (7 marks)

Question 2: (15 Marks)

Consider the *Politics_Tweets.csv* data-set.

The media column of the this data-set has data that has the following format for photos:

```
[{"url":"https://pbs.twimg.com/media/
FUBWWSOWAAIzEJ6.jpg","type":"photo","media_key":"3_1531318814155079682"},
```

{"url":"https://pbs.twimg.com/media/
FUBWWd1XoAMVYmd.jpg","type":"photo","media_key":"3_1531318817271554051"}]

The most important part of this string that is required for the current project is the url of the image i.e., **`https://pbs.twimg.com/media/FUBWWd1XoAMVYmd.jpg`**.

Automate the extraction of the urls for the media from these strings of data for all records. It is expected that the column of media should only retain the urls.

(15 marks)