CSC 3303 – Big Data Analytics | Test 1 | Semester 2 2020/2021

A. TEST QUESTION

Normally, in a data science environment, you will have a set of raw data and you will need to make sense of it. Different data scientists might come to numerous outcomes from the same set. So, it is up to you to figure the out the business questions and transform them to data science questions. In order to achieve this, you need to explore the data thoroughly. Sometimes this is also referred as exploratory data analysis. For example , you will need to inspect data and its properties, use statistics to extract features and test significant variables, use data visualisation to identify significant patterns and trends, etc. Interpretation is the final and crucial step of the process. Simply put, a layman should be able to understand your analytical outcomes. Interpretations should answer your data science questions that you formulated at the beginning and should lead to actionable insights.

An intermediate step between exploration and interpretation can be modelling data. This will help you to create prediction models, understand certain grouping phenomenon's, etc.

This take home test provides you with a datasets pertaining to LinkedIn, highlighting industry employment, talent migration, industry skill needs and skill penetration. The **dataset** folder contains the following files:

1. linkedin_to_isic_rev_4_industry_mapping_0
2. public_use-industry-employment-growth
3. public_use-industry-skills-needs
4. public_use-skill-penetration
5. public_use-talent-migration

This is an **open-ended, exploratory and hands on test**. Your task is to explore the dataset , as per the guidelines below:

- You **must** use **Python** for the analysis.
- Methods are as per your choice & convenience.

You **MUST** prepare and submit your results, analysis & conclusions, etc. using a power point presentation along with the relevant codes or any other material. Documentation within the code is **very important.**

Since this is an open ended test, you can choose to include any material in your presentation that you deem fit. Below are some of the sections that you might want to include:

CSC 3303 – Big Data Analytics | Test 1 | Semester 2 2020/2021

- Overall Summary
- Background
- Data Science Questions
- Objectives
- Methodology
- Results
- Visualizations
- Any other discussion
- References (APA Format)

B. INSTRUCTIONS:

- Name & Matric Number must be included in your submissions

- Submit your files by uploading them as following:

    - **iTaleem**: Only the power point presentation

    - **Google Classroom**: All files

- You must upload **actual** files to iTaleem & Google Classroom. Links to your files will not be accepted, like link to your colab notebook, google slides, etc.

- This is an **open ended**, **exploratory & hands on test**. You are free to refer to any online/offline resources as long as you don't plagiarise and refer to your sources.

- Referring to your sources is of utmost importance and not citing them will be deemed as plagiarism.

- Any form of plagiarism is highly unacceptable.

C. ASSESSMENT:

- **Test 1 bears 20 Marks (20% of your Total Course Evaluation).**

- **If you don't submit your "codes and/or other relevant material", your presentation will not be evaluated.**

- The marks are distributed as follows:

Presentation	**15 Marks**
Codes AND/OR Other Relevant Items	**5 Marks**

D. IMPORTANT DATES/DEADLINES:

STARTS: **21st April 2021, 03:30 PM**

ENDS: **21st April 2021, 06:30 PM**

*****END OF DOCUMENT*****