

|  |  |  |
| --- | --- | --- |
| **UCCC3073 DATA SCIENCE ASSIGNMENT** | | |
| **Programme(s)** | **:** | Bachelor of Computer Science (Honours) |
| **Trimester** | **:** | June 2022 |
| **Course Leader** | **:** | Dr. Tong Dong Ling ([tongdl@utar.edu.my](mailto:tongdl@utar.edu.my)) |
| **Submission Date** | **:** | **Monday, 22 August 2022, 5 p.m.** |
| **Submission Platform** | **:** | **WBLE** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Group Information** | | | | |
| **Group number** |  | | | |
| **Student Name** | **Student ID** | **Individual Contribution %** | **Signature** | **Final mark** |
| 1) |  |  |  |  |
| 2) |  |  |  |  |
| 3) |  |  |  |  |

Assess and manipulate data sources from various online sources.

CLO2

**Course Learning Outcomes Assessed**

**Assignment mark sheet**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Items** | **Poor** | **Marginal** | **Satisfactory** | **Good** | **Excellent** | **Marks** |
| **Task 1: Data Summary [30 marks]** |  |  |  |  |  |  |
| Data understanding **[10m]** | 0 2 | 4 | 6 | 8 | 10 |
| Data acquisition **[5m]** | 0 1 | 2 | 3 | 4 | 5 |
| Data description **[10m]** | 0 2 | 4 | 6 | 8 | 10 |
| Outlined questions **[5m]** | 0 1 | 2 | 3 | 4 | 5 |
| **Task 2: Data Analysis [30 marks]** |  |  |  |  |  |  |
| Lead sentence**[10m]** | 0 1 | 2 | 3 | 4 | 5 |
| Statistic summary **[10m]** | 0 2 | 4 | 6 | 8 | 10 |
| Clarity of the visualisation **[10m]** | 0 2 | 4 | 6 | 8 | 10 |
| **Task 3: Data Memo [30 marks]** |  |  |  |  |  |  |
| Identification of audience **[5m]** | 0 1 | 2 | 3 | 4 | 5 |
| Insights and recommendations **[10m]** | 0 2 | 4 | 6 | 8 | 10 |
| Clarity of visualisation **[10m]** | 0 2 | 4 | 6 | 8 | 10 |
| Reference sources **[5m]** | 0 1 | 2 | 3 | 4 | 5 |
| **Overall report writing [10m]** | 0 2 | 4 | 6 | 8 | 10 |  |
| **Total mark (100)** | | | | | |  |

|  |  |
| --- | --- |
| **Assessment category** | **Criteria** |
| Poor | Insufficient content, “rush work”, ambiguity, fatal errors |
| Marginal | Insufficient content, fairly written, some major fatal errors |
| Satisfactory | Fair content, reasonably written, minor problems with formatting and coding |
| Good | Good proportion of content, clearly written and coded |
| Excellent | Concise, high density, value creation to organisation |

|  |  |
| --- | --- |
| **Marking item** | **Description** |
| Data understanding | Clear understanding on the chosen data set evident with a relatively detailed explanation on the data set. |
| Data acquisition | The origin of data |
| Data description | Data attributes are clearly explained |
| Outlined questions | questions are meaningful and are answerable from the data |
| Lead sentence | Most interesting thing in the data that would Interest readers |
| Statistic summary | Clear, concise summary to show lead and key findings of the data |
| Clarity of the visualisation to support the lead | The charts clearly illustrated lead sentence, key findings and/or insights of the data. The charts are appropriately sized to be easily read within the report |
| Identification of audience | Appropriateness of audience |
| Insights and recommendations | Clear, compelling points derived from the data and appropriate recommendations given |
| Reference sources | Sources have been well-summarised and referenced in the report. The sources are meaningful and greatly improve the readability and content of the report |
| Overall report writing | Well-structured report with proper use of grammar, punctuation and spelling |

# General Instruction

1. The total mark for the assignment is 100 and contributes 20% to the total final grade.
2. This is a **group assignment**. Each group can have a **maximum of 3 students**. Group members can come from different practical groups.
3. Each group has to submit **One report (\*.pdf)** before the deadline. Only the group leader is required to upload the report.
4. Late submission may be subjected to a penalty. For example,
   * **Late submission** = -10 marks
   * **Copying/plagiarism** = 0 marks
   * **Do not include cover page** = - 5 marks
   * **Do not include marking rubric** = - 5 marks per rubric
5. Your submission should consist of:
   * The cover page
   * Assignment mark sheet
   * Report
6. Each member has to evaluate their team member(s) effort in completing this assignment. Total contributing score should be 100%. The contribution score should be acknowledged by the respective member. **Marks will be allocated based on a team members’ contribution**.

**ASSIGNMENT SPECIFICATION**

In this assignment, you are required to evaluate data of your choice using exploratory data

analysis (EDA) tools and to prepare a report reflecting on what you have discovered from the data. This assignment assesses your ability to derive useful insights from the data. Therefore, you are allowed to use any EDA tools to analyse your data.

This assignment containing 3 tasks. Task 1 required you to summarise the data of your choice. You must understand your data fully. Task 2 required you to explore the data and identify interesting findings in the data. Task 3 required you to wrap out all your findings and organise these findings in the report format.

You need to find a large data set that will serve as the basis for this assignment. The data set can be on any topic you choose — ideally, pick one that is something you are interested in and/or familiar with. This data set MUST consisting of at least 2,000 observations/records from one of the following, but not limited to, sources:

* UCI repository (<https://archive.ics.uci.edu/ml/datasets.html>)
* The U.S. Centers for Disease Control ([https://data.cdc.gov](https://data.cdc.gov/))
* The World Bank (<https://data.worldbank.org/indicator>)
* Malaysia Open Data Portal (<https://www.dosm.gov.my/>)
* EU Open Data Portal (<https://data.europa.eu/euodp/en/home>)
* Kaggle (<https://www.kaggle.com/datasets>)

## Task 1: Data summary (maximum 500 words) – 30 marks

Write a short summary about your data. The summary should describe the following:

* The problem(s) or challenge(s) that you want to solve.
* The nature of the chosen data set, data types and reference its origin.

If the chosen data set has 15 or less attributes, table them with attribute name, description, and attribute type (e.g., interval, ratio, nominal, etc.). If the chosen data set has more than 15 attributes, group them into themes (e.g., customer, orders, pollution, etc.) and describe the type of information and attribute types in each theme. You may want to highlight significant/interesting attributes in the themes.

* Outline at least 3 questions that you would like to find out from your data set and these questions should lead you to solve your problem(s) or challenge(s) above.

## Task 2: Initial analysis – 35 marks

Write an analysis report containing a lead story and descriptive analysis for your data set. The lead story and analysis should relate to your questions outlined in Task 1.

This report should include:

1. A single sentence (e.g., “headline” of a news or story) identifying the most interesting finding you have discovered in your data set. This “lead” sentence should intrigue the reader and make them want to learn more.
2. A short write-up summary on your lead sentence and other interesting findings that you have discovered in your data set. This write-up summary should involve statistics/descriptive findings.
3. At least two visualizations that illustrate your findings. At least one of the visualizations should support the lead sentence. Visualizations can include of tables, charts, graphs, pictures, plots, and figures.

## Task 3: Data Memo (1500 – 2000 words) – 25 marks

Write a short data memo report outlining all the key findings of your analysis as well as suggestions for your intended audience. This task will build on the findings in your previous tasks, exploring them in greater depth.

This report should include:

* + A brief description of the intended audience of your report.
  + Your findings and your suggestions. You are advised to use different heading sections for your findings and your suggestions to make your report clear (i.e., write “Recommendations” above the section where you provide your recommendations). Your recommendations should be clear for the audience of what steps they should take in response to your analysis.
  + Reference at least 6 external sources (studies, news articles, reports) that put your data into context. You need to include the source details after you have cited it in your text (you can list the source after the stated paragraph before you start summarizing another source).
  + At least six visualizations to support your findings and these visualizations must be your own works. You should include these visualizations when you explain your findings.