**

SIT742 Modern Data Science

Assignment 1

Student name:

Student ID:

Date:

Table of Contents

[**Part I - Tulip Hotel Web Logs Exploratory Data Analysis** 3](#_Toc67266321)

[**1.** **Data ETL** 3](#_Toc67266322)

[**1.1.** **Data Loading** 3](#_Toc67266323)

[1.1.1. Dataset Description 3](#_Toc67266324)

[1.1.2. Attribute Dictionary 4](#_Toc67266325)

[**1.2.** **Data Cleaning** 4](#_Toc67266326)

[**2.** **Data Statistics Description** 4](#_Toc67266327)

[**2.1.** **Traffic Analysis** 4](#_Toc67266328)

[**2.2.** **Server Analysis** 5](#_Toc67266329)

[**2.3.** **Geographics Analysis** 5](#_Toc67266330)

[**Part II - School of IT Professor Citation Information** 6](#_Toc67266331)

[**3.** **Professor List Generation** 6](#_Toc67266332)

[**3.1.** **Import Web Crawling Library** 6](#_Toc67266333)

[**3.2.** **Find all professors in School of IT** 6](#_Toc67266334)

[**4.** **Professor Citation Information Generation** 7](#_Toc67266335)

[**4.1.** **Code for generating the Professor Citation Information (include the actual crawling code as well)** 7](#_Toc67266336)

[**4.2.** **Find the Professor with most citations** 8](#_Toc67266337)

[**4.3.** **Find the Associate Professor with most i10-index since 2016** 8](#_Toc67266338)

[**4.4.** **Find all Professors name who have the citations since 2016 > 2500** 8](#_Toc67266339)

**Part I - Tulip Hotel Web Logs Exploratory Data Analysis**

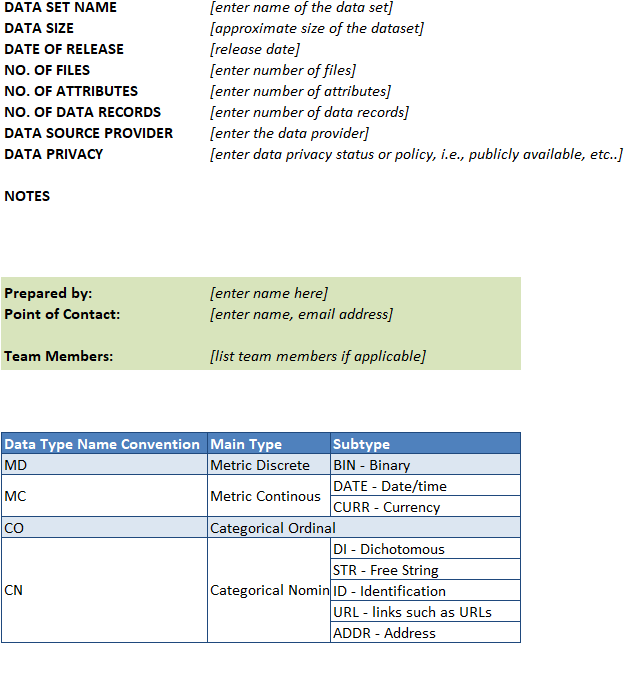
Hotel TULIP a five-star hotel located at Deakin University, and its CIO Dr Bear Guts has asked the Team-SIT742 team to analyse the weblogs files. As an employee for Hotel Tulip, working in the Information Technology Division, it is required to prepare a set of documentation for Team-SIT742 to allow them to understand the data being dealt with. Throughout this report, some source codes are to explore the weblog, which afterwards the information is presented to Dr Bear Guts in the format of a report.

1. **Data ETL**
   1. **Data Loading**

*Fill the DataDictionary.xlsx with discovery from the result of 1.1 Data Loading from your notebook.*

* + 1. Dataset Description

*Please add a screenshot of Dataset Description from your DataDictionary.xlsx.*



* + 1. Attribute Dictionary

*Please add a screenshot of Attribute Dictionary from your DataDictionary.xlsx.*



* 1. **Data Cleaning**

*Please add description of the following contents by yourself.*

1. *The number NAs for each column*
2. *The number of rows before removal NAs*
3. *The number of rows after removal NAs*
4. **Data Statistics Description**
   1. **Traffic Analysis**

*Please add description of the following contents by yourself.*

1. *Please add a figure of Hourly Requests Bar Chart from your Notebook and elaborate the findings from the figure.*
2. *Please add a table of filter result (hourly\_request\_amount >= 400000 & hourly\_request\_amount <= 490000)*
   1. **Server Analysis**

*Please add description of the following contents by yourself.*

1. *Please elaborate how many types of reported server status.*
2. *Please add a figure of Server Error Pie Chart from your Notebook, and elaborate the findings from the figure.*
   1. **Geographics Analysis**

*Please add description of the following contents by yourself.*

1. *Please add a figure of Country distribution and list top 3 with the number of requests.*
2. *Please add a figure of City distribution and list top 3 with the number of requests.*

**Part II - School of IT Professor Citation Information**

To better introduce all the professors including the emeritus professor, the professor and also associate professor in Deakin University School of IT, faculty will need to know all the citation information on all professors. Google Scholar is a web search engine that freely indexes the metadata of articles on many authors. Majority of the professors choose to use google scholar to track their publications and research works. Therefore, the web crawling on google scholar will be able to have the citation information obtained across all the professors (who have the google scholar profile).

1. **Professor List Generation**
   1. **Import Web Crawling Library**

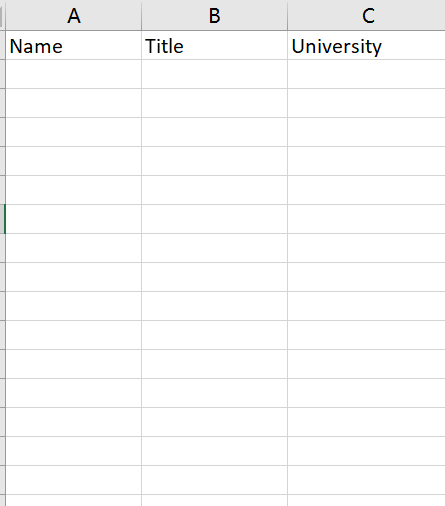
*Please fill this part with the screenshot of your code for import your own web crawling library.*

* 1. **Find all professors in School of IT**

*Please fill this part with the screenshot of your code for generating the professor name list csv. The screen shot will also include the results of the running on the code.*

* + 1. **Professor Name List CSV**

*Please fill this part with the screenshot of your csv.*

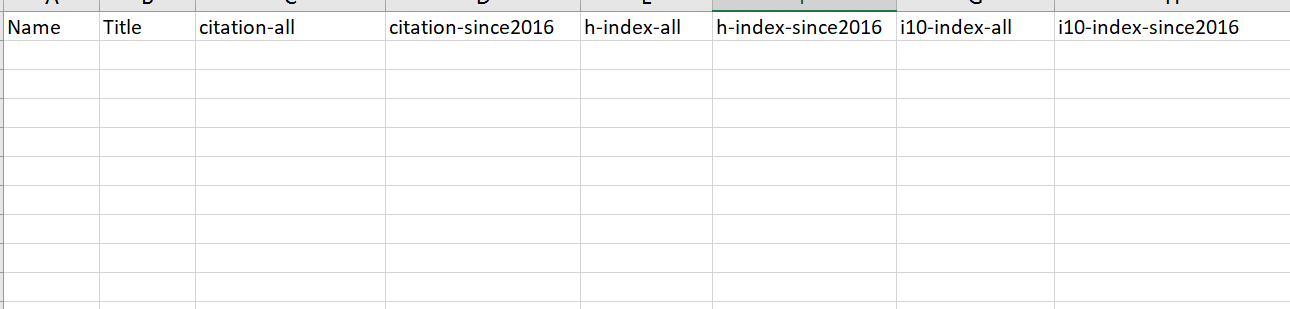


1. **Professor Citation Information Generation**
   1. **Search the google scholar for all**

*Please fill this part with the screenshot of the code for generating the professor citation information (include the actual crawling steps).*

* + 1. **Professor Citation Information CSV**

*Please fill this part with the screenshot of the professor citation information CSV.*



* 1. **Find the Professor with most citations**

*Please fill this part with the screenshot of the code (include the results of the code running).*

* 1. **Find the Associate Professor with most i10-index since 2016**

*Please fill this part with the screenshot of the code (include the results of the code running).*

* 1. **Find all Professors name who have the citations since 2016 > 2500**

*Please fill this part with the screenshot of the code (include the results of the code running).*