**

SIT742 Modern Data Science

Assignment 1

Student name: Yash Ahuja

Student ID: 219608443

Date: 19-04-2021

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** 6](#_Toc67266327)

[**2.1.** **Traffic Analysis** 6](#_Toc67266328)

[**2.2.** **Server Analysis** 7](#_Toc67266329)

[**2.3.** **Geographics Analysis** 9](#_Toc67266330)

[**Part II - School of IT Professor Citation Information** 11](#_Toc67266331)

[**3.** **Professor List Generation** 11](#_Toc67266332)

[**3.1.** **Import Web Crawling Library** 11](#_Toc67266333)

[**3.2.** **Find all professors in School of IT** 12](#_Toc67266334)

[**4.** **Professor Citation Information Generation** 17](#_Toc67266335)

[**4.1.** **Code for generating the Professor Citation Information (include the actual crawling code as well)** 17](#_Toc67266336)

[**4.2.** **Find the Professor with most citations** 24](#_Toc67266337)

[**4.3.** **Find the Associate Professor with most i10-index since 2016** 24](#_Toc67266338)

[**4.4.** **Find all Professors name who have the citations since 2016 > 2500** 24](#_Toc67266339)

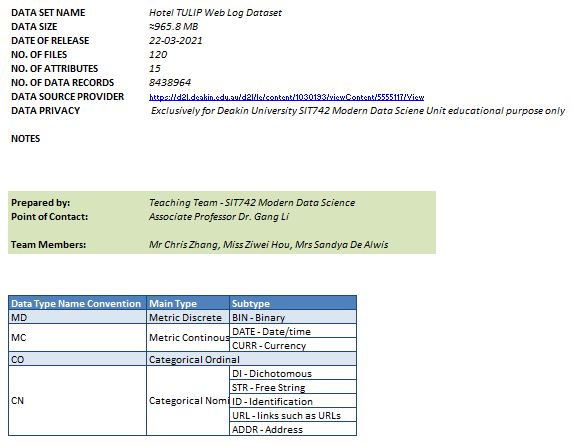
**Part I - Tulip Hotel Web Logs Exploratory Data Analysis**

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



* + 1. Attribute Dictionary
  1. **Data Cleaning**

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

1. *The number NAs for each column*

date 0

time 0

s-sitename 12

s-ip 24

cs-method 24

cs-uri-stem 24

cs-uri-query 7886568

s-port 40

cs-username 8438960

c-ip 36

cs(User-Agent) 36

cs(Referer) 36

sc-status 36

sc-substatus 39

sc-win32-status 80

code-output:

s-sitename has number of NA records: 12

s-ip has number of NA records: 24

cs-method has number of NA records: 24

cs-uri-stem has number of NA records: 24

cs-uri-query has number of NA records: 7886568

s-port has number of NA records: 40

cs-username has number of NA records: 8438960

c-ip has number of NA records: 36

cs(User-Agent) has number of NA records: 36

cs(Referer) has number of NA records: 36

sc-status has number of NA records: 36

sc-substatus has number of NA records: 39

sc-win32-status has number of NA records: 80

1. *The number of rows before removal NAs*

8438964

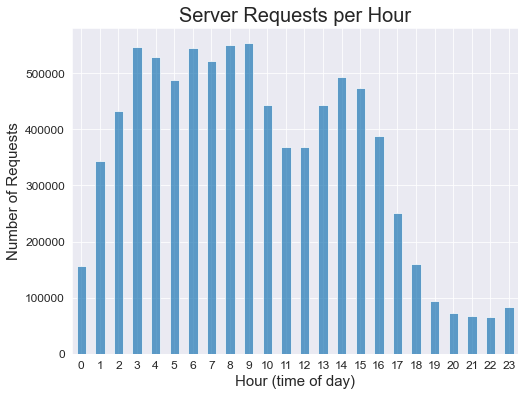
1. *The number of rows after removal NAs*

8438884

1. **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.*

**

*We can see from the above figure that 9th hour of the day (9 – 10 AM) has the highest number of server requests whereas 22nd hour of the day (10 – 11 PM) has the lowest number of server requests. Overall, it can be said that the number of server requests are usually significantly high from 3 AM to 5 AM and 6 AM to 10 AM as compared to the other times of the day.*

1. *Please add a table of filter result (hourly\_request\_amount >= 400000 & hourly\_request\_amount <= 490000)*

|  |  |
| --- | --- |
| *Hour (time of day)* | *Hourly Request Amount* |
| *2* | *432315* |
| *5* | *487430* |
| *10* | *443621* |
| *13* | *442659* |
| *15* | *473376* |

* 1. **Server Analysis**

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

1. *Please elaborate how many types of reported server status.*

*Types of server status reported: 12*

Server status type Value Counts(Number of records)

200.0 6074352

304.0 2137486

404.0 143646

206.0 72493

302.0 7368

500.0 2020

403.0 703

301.0 436

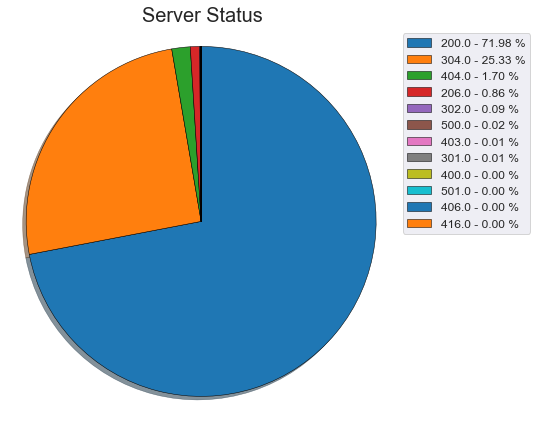
400.0 210

501.0 113

406.0 54

416.0 3

1. *Please add a figure of Server Error Pie Chart from your Notebook and elaborate the findings from the figure.*

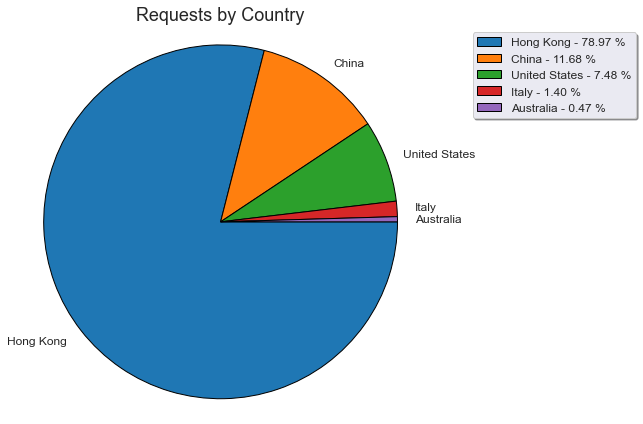
**

*We can see from the above figure we can analyse the protocol status i.e. the status of the server to client action which is represented by a success or error code. From the figure, status code 200 has significantly high i.e. approx. 72% of the total number of server requests followed by status code 304 which has approx. 25% of the total server requests. Whereas, the status codes such as 404, 206, 302 have very less significant number of server requests and status codes such as 403, 301, 400, 501, 406, 416 have insignificant (approx. zero) server requests. Overall, it can be said that the status code 200 has most of the server requests as compared to other status codes.*

* 1. **Geographics Analysis**

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

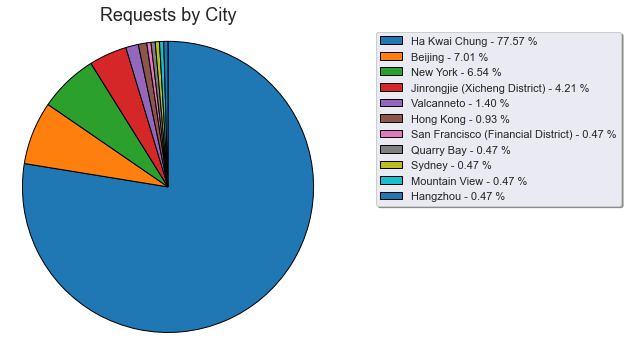
1. *Please add a figure of Country distribution and list top 3 with the number of requests.*
2. The number of requests received in the period: 214
3. The number of countries involved in the period: 5
4. Requests by Country Figure:

**

4.Top three countries with highest request numbers:

|  |  |
| --- | --- |
| Country | Number of Requests |
| Hong Kong | 169 |
| China | 25 |
| United States | 16 |

1. *Please add a figure of City distribution and list top 3 with the number of requests.*
2. The number of cities involved in the period: 11
3. Requests by City Figure:

**

1. Top three cities with highest request numbers:

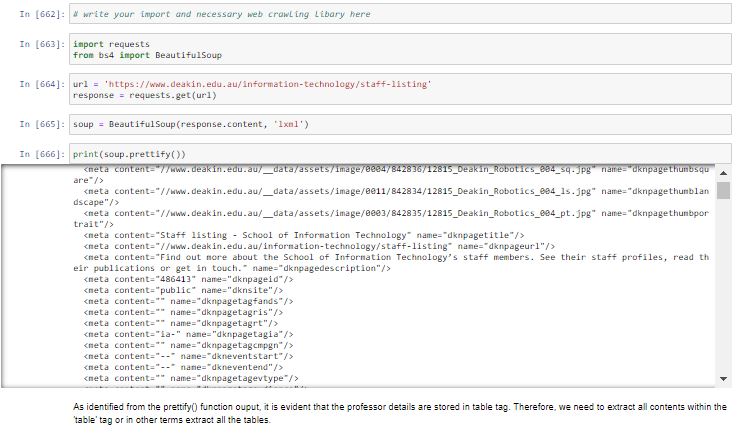
|  |  |
| --- | --- |
| City | Number of Requests |
| Ha Kwai Chung | 166 |
| Beijing | 15 |
| New York | 14 |

**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).

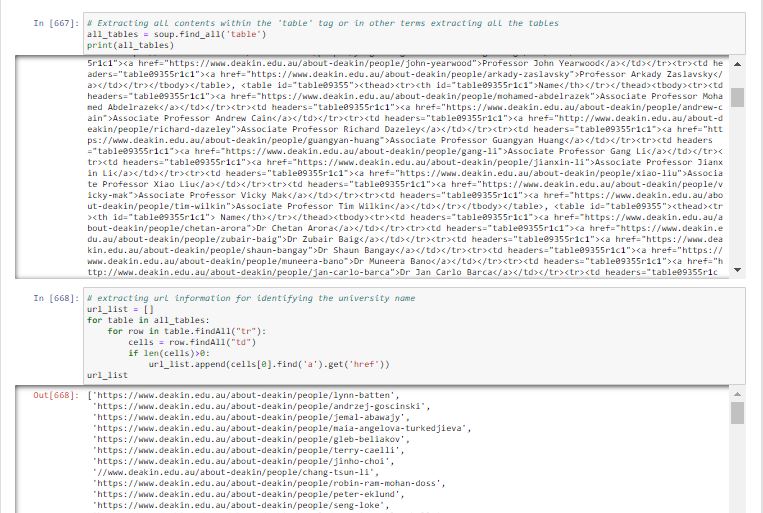
1. **Professor List Generation**
   1. **Import Web Crawling Library**

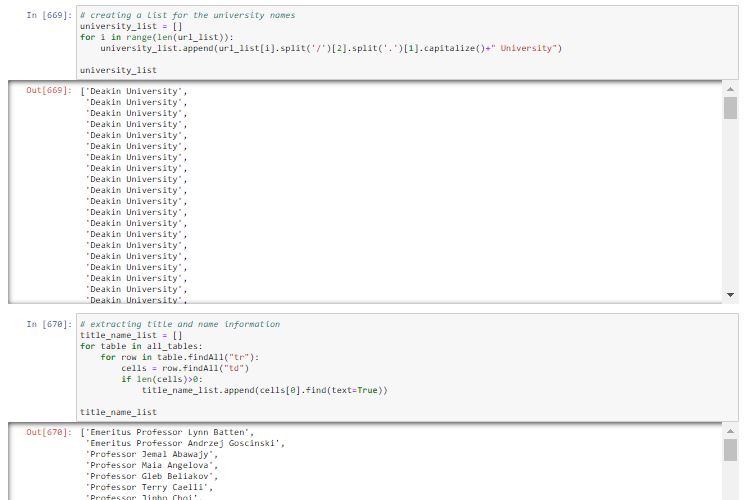
*Please fill this part with the screenshot of your code for import your own web crawling library.*

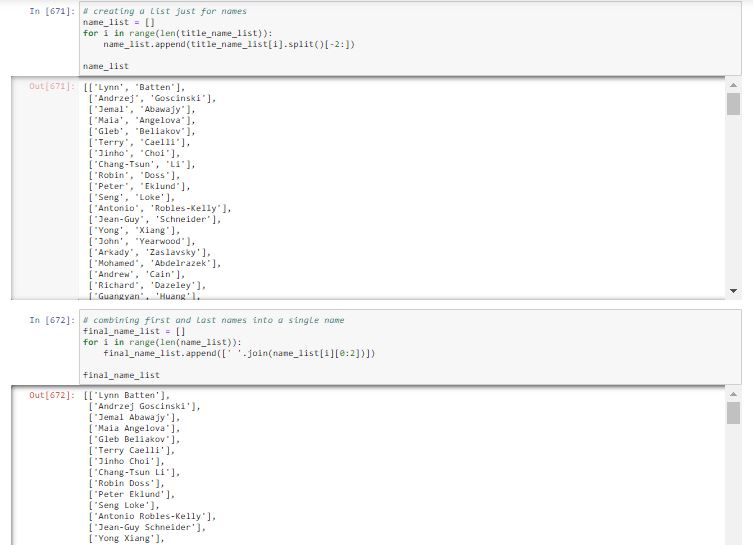


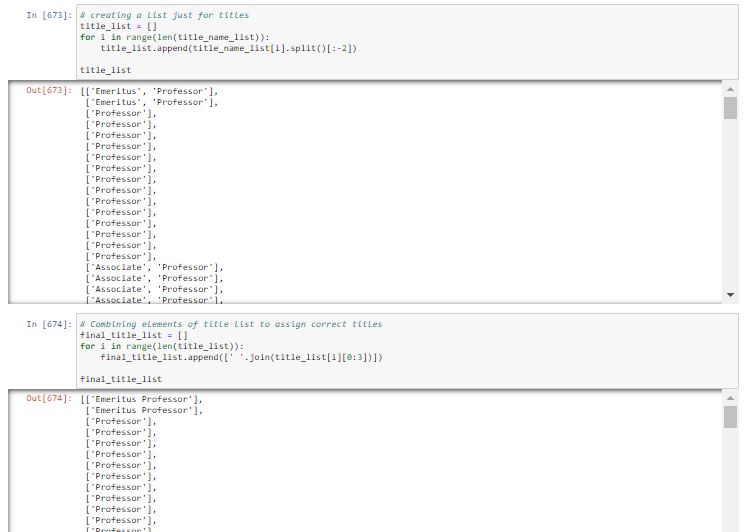
* 1. **Find all professors in School of IT**

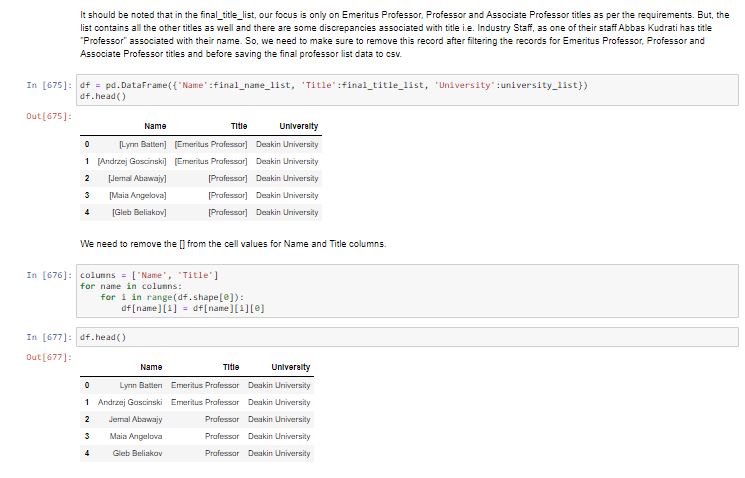
*Please fill this part with the screenshot of your code for generating the professor name list csv. The screen shot will also include the results of the running on the code.*

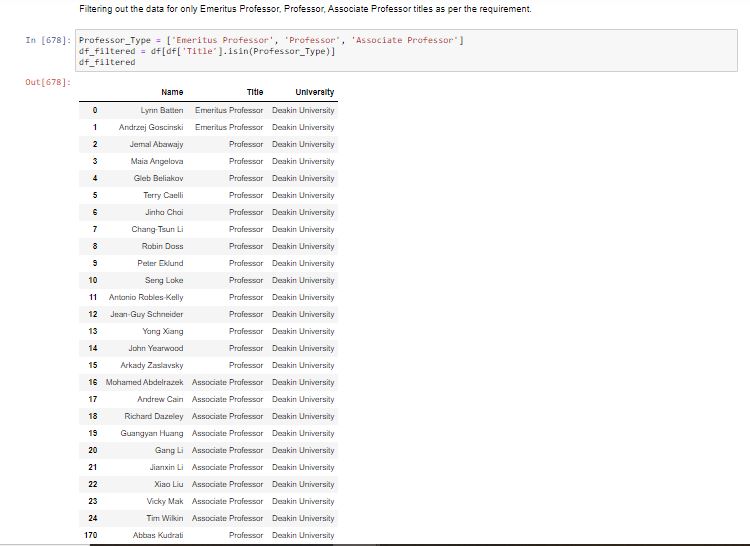














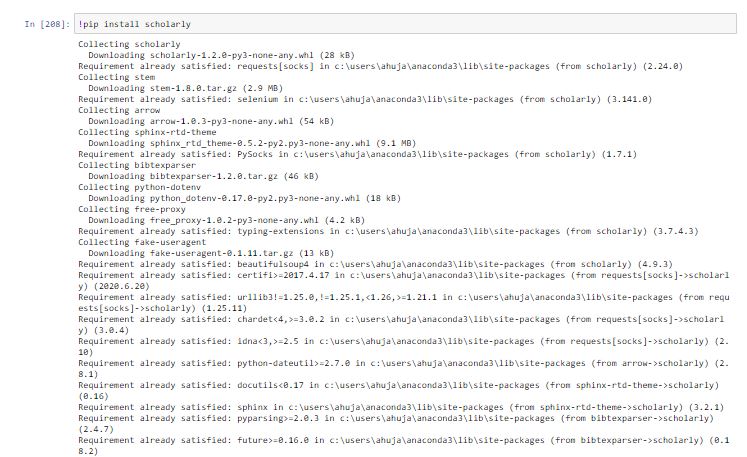
* + 1. **Professor Name List CSV**

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

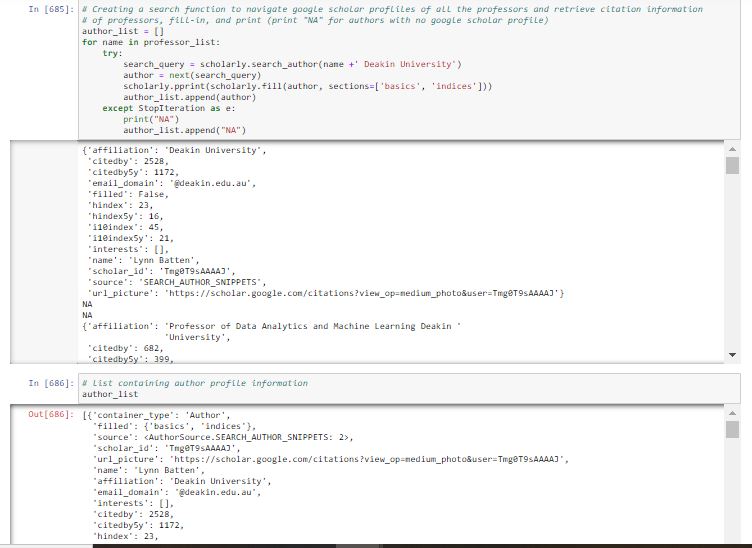


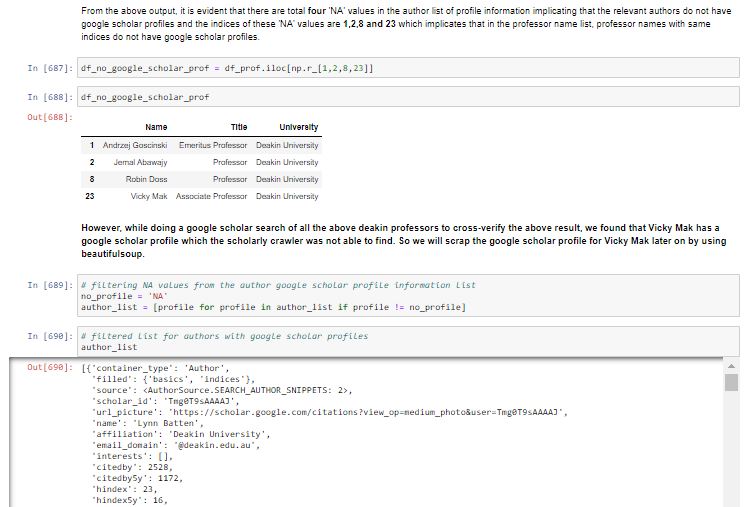
1. **Professor Citation Information Generation**
   1. **Search the google scholar for all**

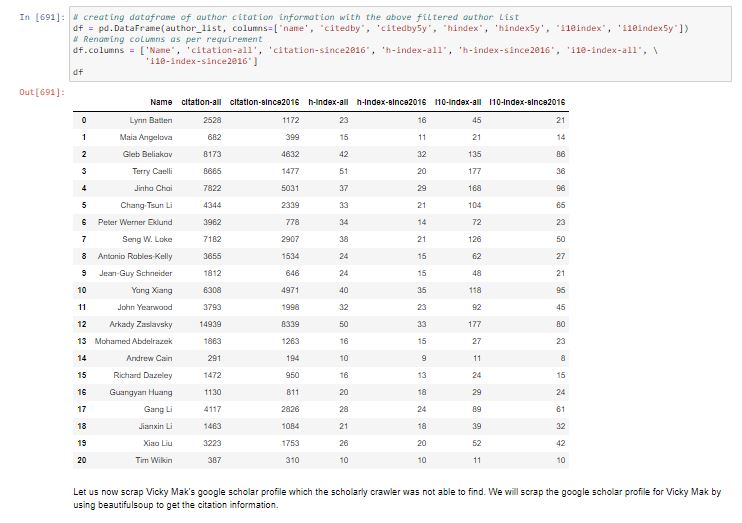
*Please fill this part with the screenshot of the code for generating the professor citation information (include the actual crawling steps).*

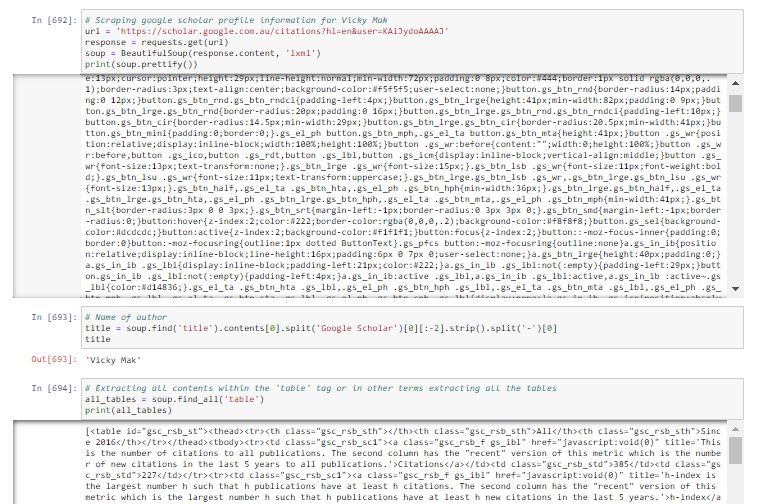


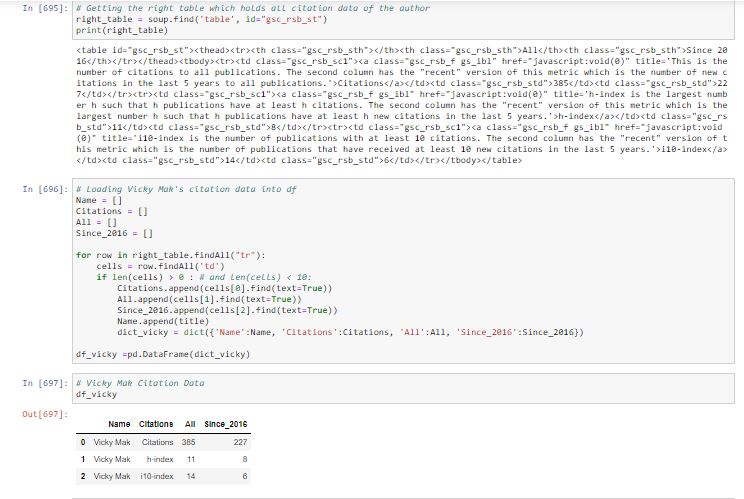


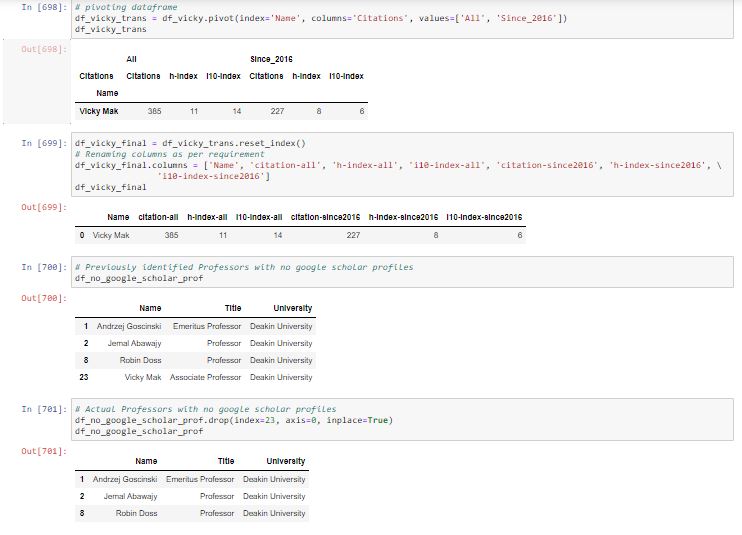


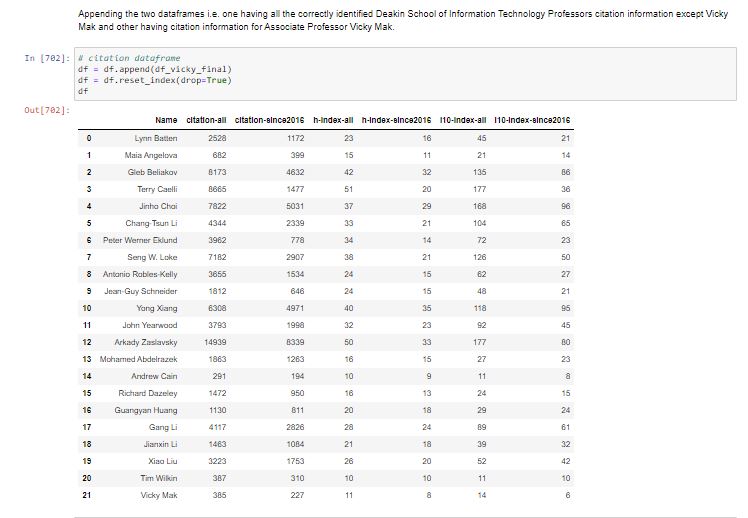




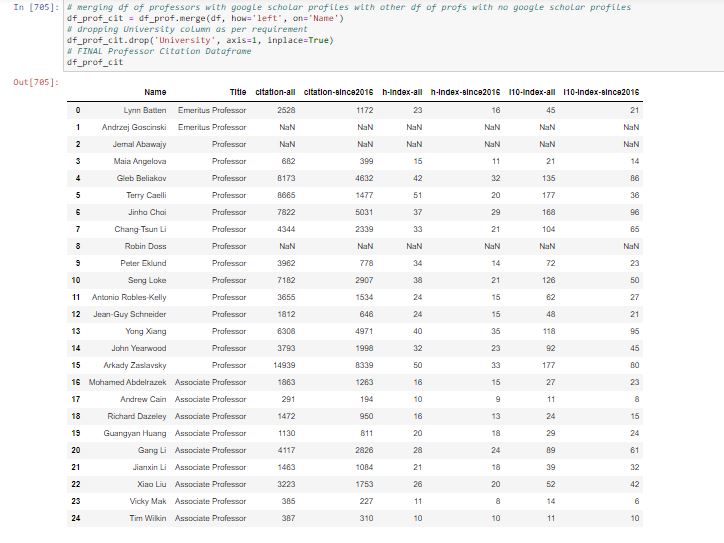








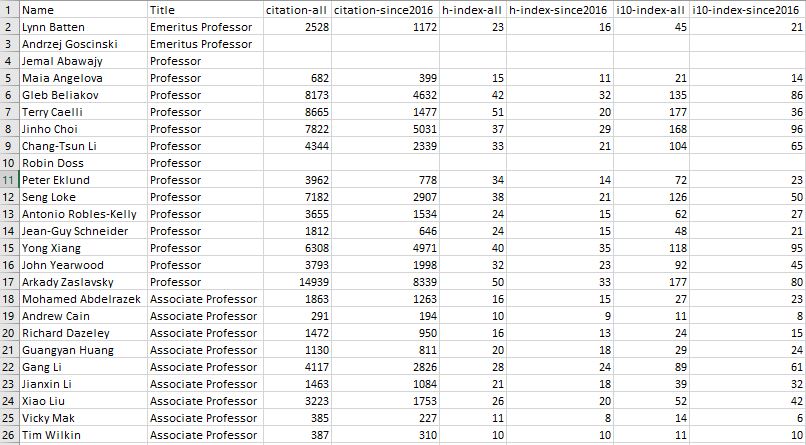






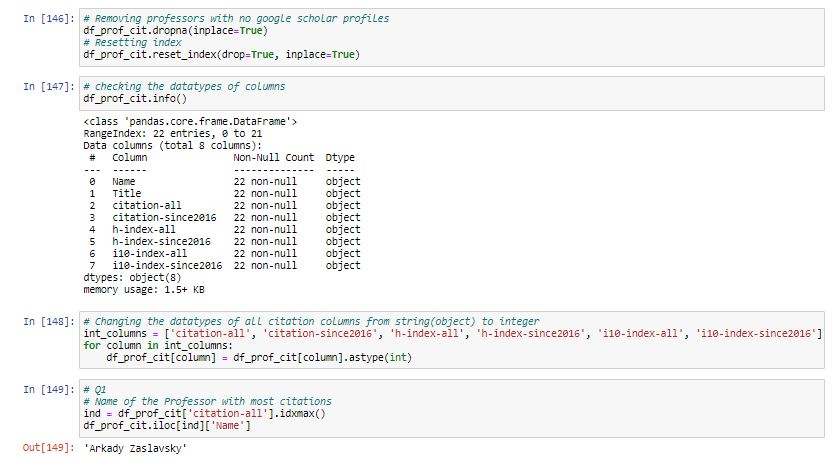
* + 1. **Professor Citation Information CSV**

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



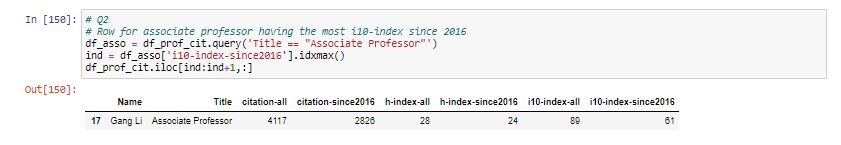
* 1. **Find the Professor with most citations**

*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).*

