### KE5106 CA2 programming assignment submission

*Tan Chee Wei (A0179723U)*

*Ramasamy Muthuraman (A0179756H)*

*Laxman Singh (A0178223E)*

*Frank Tran (A0178233A)*

V HireU

# **Abstract**

Graduates and professionals having experience are taking higher education from NUS. This increases their chance to grow professionally. This report based on the professionals who have completed MTECH from NUS ISS and currently working in different positions in different organization. Professionals having MTECH and other graduate certificates are working in different companies. Their earnings and positions in company is the outcome of taking MTECH from NUS ISS.

Positions held after doing certain course reflects the demand of that course. Courses with high impact on the career of people will help upcoming professionals to easily decide the course based on their interest. People will be confused that doing certain course will have ROI in the job market or not. Real world examples have more impact on people. It can be motivational factor for the people-pursuing course.

This document will extract the data for the professionals with their career path who has pursued MTECH course from NUS ISS. It also matches the monetary benefits received for the position as per the market standards. And, then display it as one of the prospects after pursuing the course. This will help to have more and more prospects to look for after completing the course. It also gives you the benchmark of earnings at different stages of your career.

Overall, this report will collect professionals’ position held after doing MTECH and help to visualize the retrospect of doing this course in their career. This will work as a guide to youngsters and professionals to decide their career path. It will surely help to solve many of the questions professionals would like to ask before pursuing courses like MTECH.

# **Introduction**

This report “V HIREU” analyses the working professionals who have done MTECH from NUS ISS. It collects their previous job and company details with aggregate salary. It also takes job taken after completing this course and aggregate salary. This report takes the data for full time and part time students both.

This report is the analyses of career progress of professionals after completing this course. This also helps to give monitory gain received in their career since they completed their course. It helps to provide overview of doing a certain stream of MTECH and benefits received by professional in real.

It collects the job details of an individual and then collects data for same job from various sources like Glassdoor, Indeed and careers future SG. This report compiles the salary data to get minimum, maximum and mean of the salary for the job.

Different people do certain course with different career aspirations. It can be a guiding factory or stimulating factor to come to a conclusion to pursue MTECH. It can help to decide professionals that previous job with the addition of this course in their resume will help them to achieve their targets. It will also help them to give the real figure to make their decisions certain.

This report also compares the salary structures within different stream of MTECH. It also helps to give an idea how the number of years experience correlates to career grown with the acquisition of this course. Course details and future prospects can be searched from online. The benefit of this report is that it gives the conclusion about current demands for the people with this course and certain skills.

# **Objectives**

Hire-U provides the metric and various visualization reports to value the course in the current market situation. The main objective of this report is to study the effectiveness of pursuing MTECH and analyzing the real time data and market information from web portals. This report works as a guide for the professionals who want to pursue this course. Data collected from web is exploited in this report and analyzed using new visualizing tools for expressing market trends and expected salaries.

This report can definitely help professionals to predict their future with role models and career paths. This report certainly helps to give the overview of professionals’ future. Other objectives are as below:

* Guide to professionals
* Value of the Course
* Professionals status before or after the course
* Full time vs Part time jobs and salaries
* KE vs EBAC Trends

# **Definition**

# **Scraping**

Web scraping is the process where program involves downloading the web pages, extracting the useful data and writing it into file for reference in visualization and analyzing. This process is specific for each website.

# **Data Cleaning**

Data cleaning is the process of removing unwanted information and making the data standardized to the requirement. It involves beautifying the data so that meaning can be extracted in next level of analyzing. Different formats used in different websites for same kind of data makes it a challenge to standardize the data.

# **Loading into Mongo DB**

After getting the extracted data from web pages and data cleaning process, data files are loaded into Mongo DB for reference in later stages.

# **Analyzing and Predicting**

Extract the information from DB, aggregate it and visualize it in a manner, which is useful and easy to use by end users. It helps user to understand better in graphical manner.

# **Technical overview**

# **Technology Used**

|  |  |  |
| --- | --- | --- |
| SNO | TECHNOLOGY | Usage |
| 1 | Python | It is used for scraping, cleaning and loading the data. |
| 2 | Mongo DB | It is used for storing data extracted from website. |
| 3 | SQL server | It is used as DWH server for Tableau reports |
| 4 | Tableau | It is used for visualizing reports of the extracted data. |

# **Source Selection Process**

* Investigating Website

This process involves surfing through the different websites and judges the sites for the use of your requirement. It involves estimating the complexity and restrictions on the website. It also involves whether all the information is available or not.

* Conditions to select source

Availability and reliability of information without violating any rule of secrecy is the criteria for selecting the source.

* Derive attributes

This activity involves identifying the useful fields from web to analyze the data and reach to conclusion.

# **Scraping**

Web scraping is performed on web data to extract linked profiles and salary related information from other relevant sites. These profiles were searched with the keyword MTECH NUS.

This gives us broad view of current job held by professionals who has completed MTECH from NUS. It also helps to provide the details of the company, current job and previous job held by the professionals.

Information extraction from web, data cleaning and analysis shows challenges that need to be addressed. The first problem was to identify and select the source of information and web pages to be used. Sites are selected based on the following attributes:

* Profiles with completed MTECH from NUS ISS
* SG data availability
* International data availability
* Size of data
* Quality of data
* Website limitations

Scraping is simply downloading information which is categorized as “PUBLIC” as it is freely available on the web. Websites like LinkedIn has strong anti-scraping tools. Hence additional care has to be taken to work around the anti scraping mechanism in place such as are you a robot check, and mimic behaviors that is closer to a actual human using the site to prevent the scraping of public profiles from getting blocked. This activity involves downloading, extracting and analyzing the data. Python is used for this activity.

# **Data Reconciliation**

Information collected after scraping the web can be available in different formats. For example, some people mention MTECH and some people mention Masters of Technology. This brings up the challenge to extract the relevant data. It took time to reconcile the data and extract only the useful and relevant information for this report. Job details in different sites are also expressed in different format. It requires effort to standardize the data from various sites and collect it in a same format. Keeping the same format for same type of information helps in analyzing the data better.

# **Duplicate Detection and Removal**

Data extracted after scraping and standardization has lots of duplicates. Jumbled words for same information are quite common in extracted data. Thus, it becomes important to identify duplicates and remove them to make it meaningful data. Around 40-60 percent of the data has to be dropped due to duplication or irrelevancy.

# **Loading into DB**

Output for jobs extracted from different sites is uploaded into CSV format to Mongo DB. CSV files contains the job and related information as:

* Job Name
* Company Name
* Salary Offered

# **Analyzing the Data**

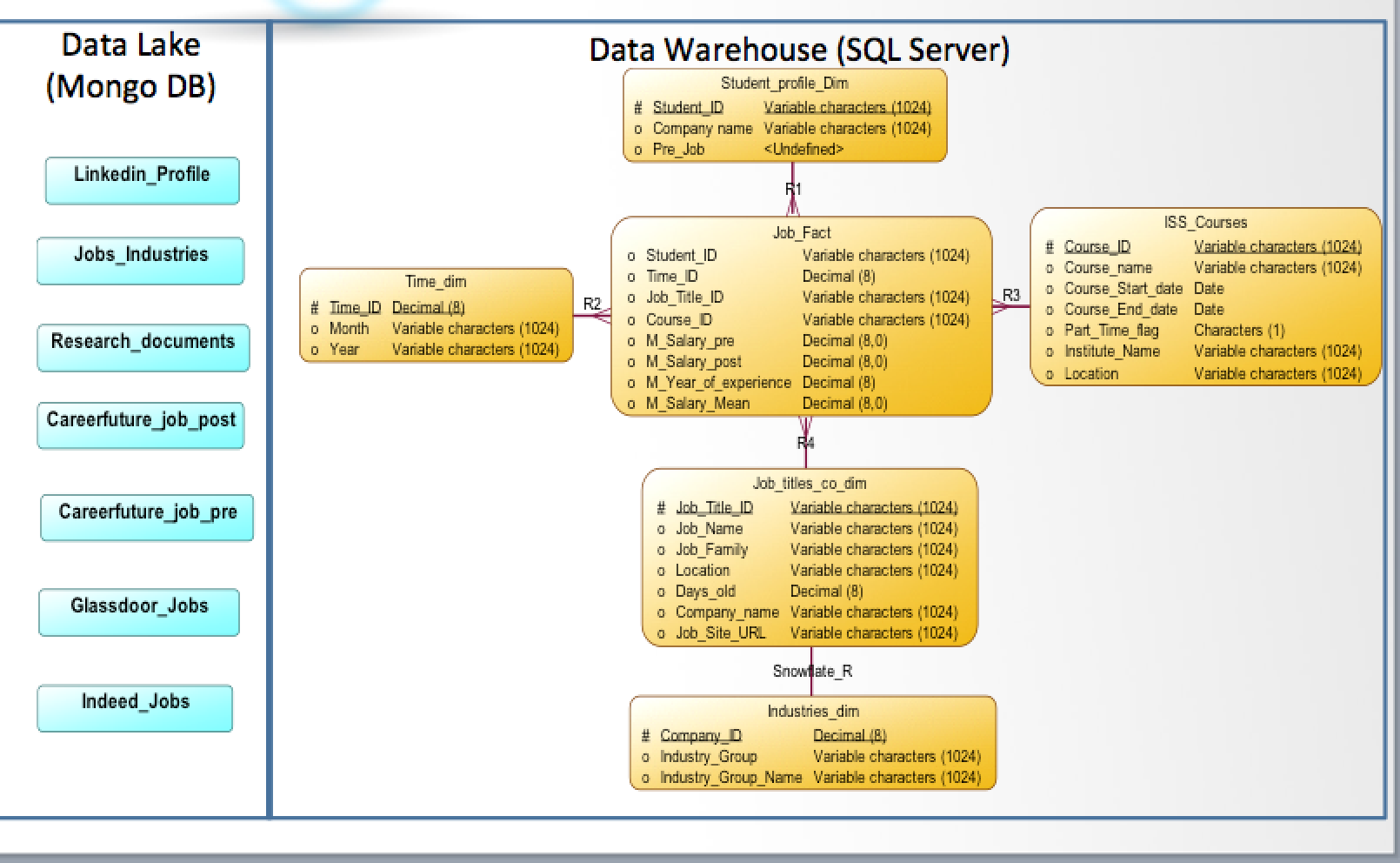
Data is extracted from Mongo DB and aggregated to put in SQL server. Tableau is connected to SQL server refers information and visualizes the aggregated data. Tableau reports are used to visualize the data in different formats and convey the message in easiest and understandable format to end-users.

# **Extracting the Information**

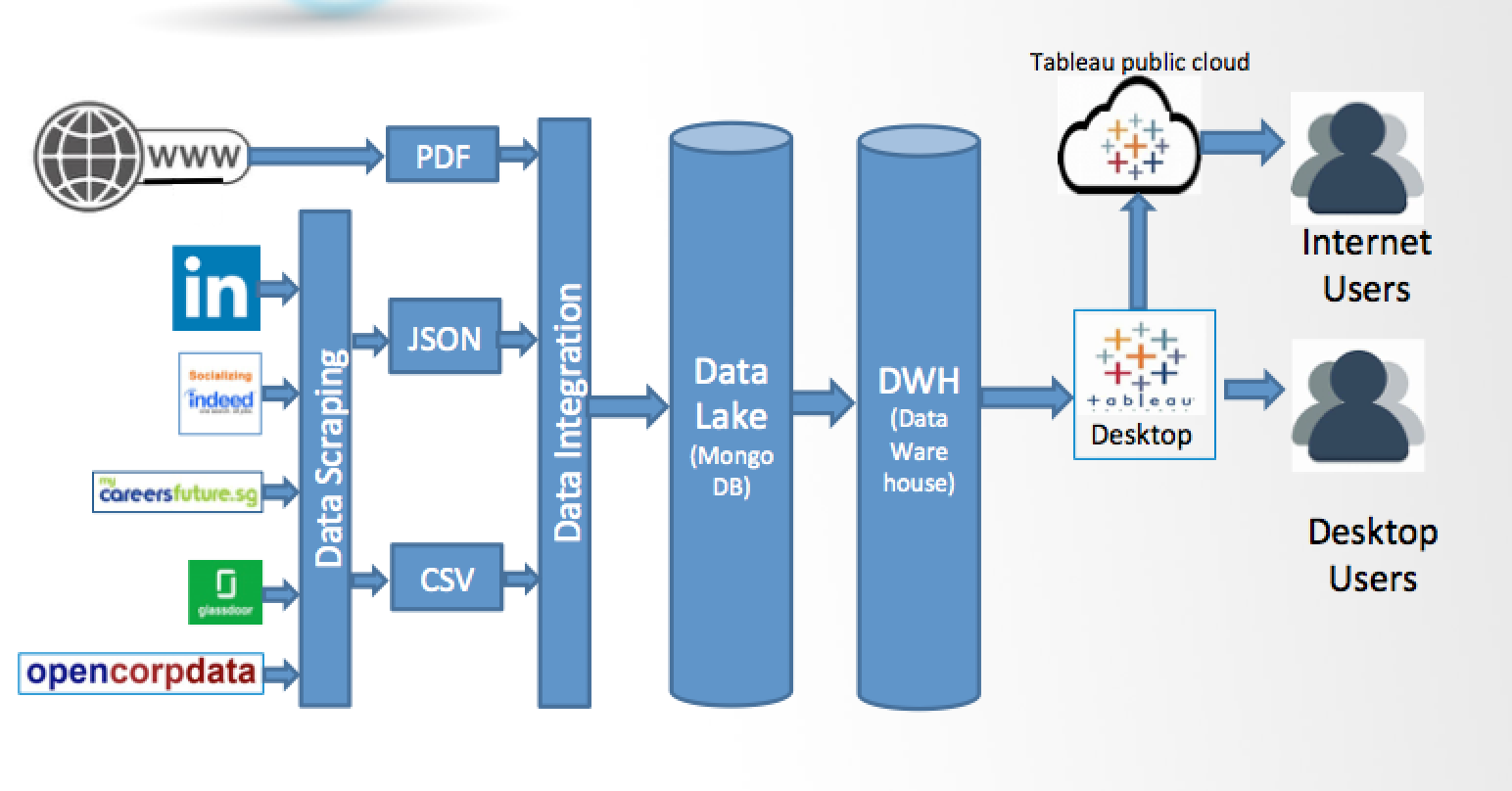
The final step is to get the conclusion drawn from visualizing the data. Usefulness of this report varies based on the requirements of end users.

# **Architecture / Data Model**

Architecture



Data model: Files received from three sources: PDF, JSON and CSV. CSV and JSON files are extracted via web scraping



# **Assumptions**

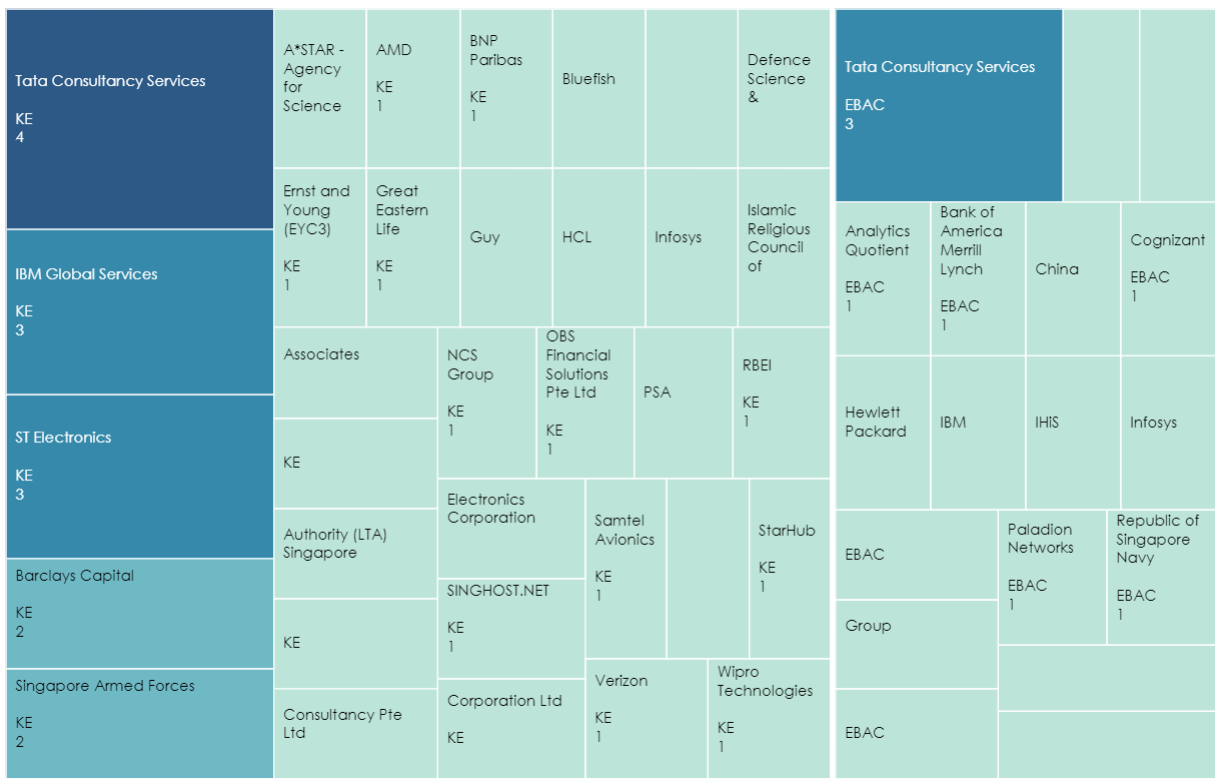
|  |  |
| --- | --- |
| Sno | Assumptions |
| 1 | Data scraping is executed on the MTECH profiles on NUS. Only MTECH students with NUS completion status are considered. It is assumed that change in their career is the result of MTECH degree. Other factors are assumed to be constant. |
| 2 | Salary is taken from different sources based on Job titles extracted. It is assumed that matched Job title and salary can be attached to the current salary of an individual. |
| 3 | Industry difference with the Job titles is ignored. It is assumed that Job title and salary are constant in various industries. |

# **Results**

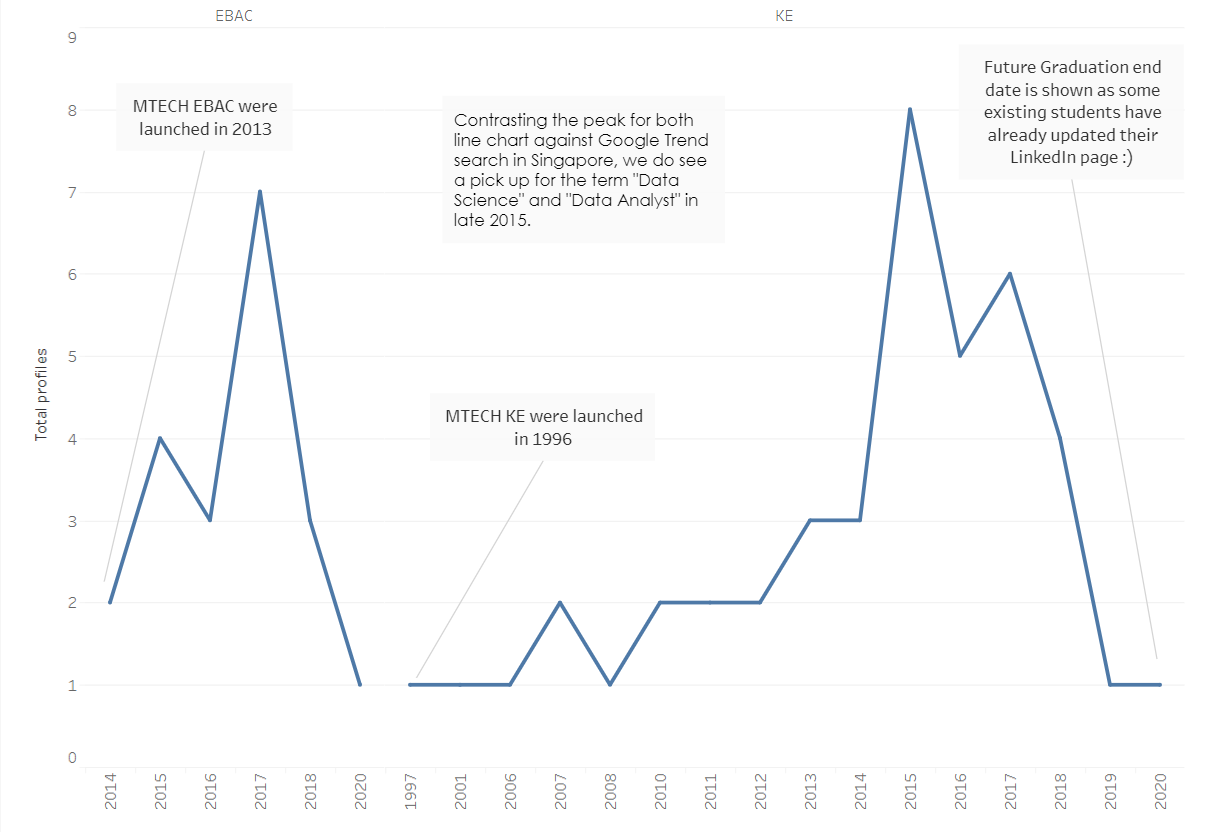
Some of the results from visualization tool are as below. Whole report can be accessed at the below URL:

<https://public.tableau.com/profile/steve7an#!/vizhome/investing_nus_iss/NUSISS>

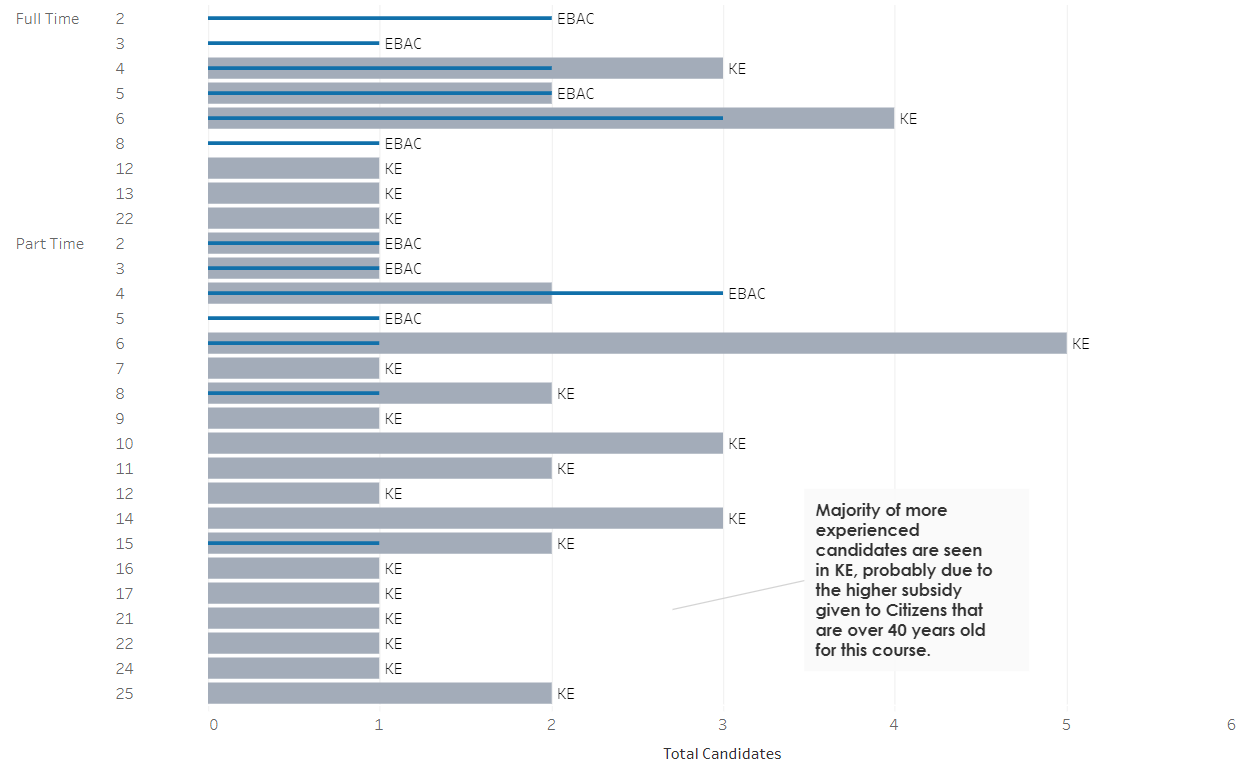
* Visualizing Candidates’ Profile



* Visualizing Graduation year and number of students



* Visualizing Candidates’ work experience with their stream (EBAC/KE)



# **Conclusions**

The result of the report analyzes MTECH student career progression and benefits achieved after completing MTECH from NUS ISS. This reports shows that scraper is able to find many profiles with NUS ISS education background. Out of which relevant to this report were below 100. The results of this report are competitive to be a starting point. Improvement to this can be incorporated in future research to make it a better product in guiding upcoming professionals. The analysis of the data extracted from web shows that there is visible growth in the career of professionals after completing the MTECH. Overall, this report provides the good overview of market trends, salary structure in different job bands and value addition of this course to students.

# **References**

1. (Chopras, 2017)
2. (Steen, n.d.)
3. (Legatt, 2017)
4. (Rundell, 2016)
5. (Huang, 2016)
6. (Media, 2017)
7. (Sandhu, 2018)
8. (Media O. , 2017)
9. (Institute, 2017)
10. (Page, 2018)
11. (Services, 2017)
12. (Adecco, 2018)