Data Warehousing and Business Intelligence

Capstone Project

On

Analysis of the Factors influencing the Selection of Universities

By

Sruthi Boddapati

RU-ID : 191006064

Master's in Information Technology and Analytics (MITA)

Submitted to: Sergei Schreider

Table of Contents

Analysis of the Factors influencing the Selection of Universities

Abstract

The growth of students studying abroad have increased immensely over past few years. So, it is very important to know the basis on which the students choose to opt a country or university. In this project I would like to explore some of the parameters involved in this selection process. We try to understand why certain universities are chosen over other universities in spite of their similarities or dissimilarities.

1. INTRODUCTION

The ratio of students choosing to study in abroad have been increasing rapidly. Times score and rank of the university are the first two criteria that students generally observe. But with the increase in the number of students every year the time score and rank also keep changing. It is also observed in Section-7 that irrespective of the rank and time score of the university there are various other parameters like the location, education quality and the different courses offered by the universities, are a few parameters that help the students in opting for a particular university. American universities have been the center from past few years. The growth in the total number of students opting for American universities has increased by almost double. To understand this a proper research is done in Section-3 where different authors gave their ideas on why students are choosing American countries a lot more in comparison with other countries. In Section-4 a Star Schema is built and the different parameters that help the student in the selection of a particular university is mentioned in Section-5. Further gathering of Raw data is mentioned in Section-2. With the help of R Studio - cleaning, manipulation and automation is done in SSIS, of which in-depth explanation is mentioned Section-6.

2. DATA SOURCES

All the data on American universities is gathered from different sources for building the Data Warehouse.

Source	Type	Brief Summary
Data.world, Website data	Structured	All the necessary columns were found in these data sources. The data is merged together and is used in building the BI queries.
Times Ranking	Structured	Times rank would be helpful for comparing the rank of American universities with other universities.
Twitter	Unstructured	Tweets include reviews, comments etc. Would help people know about the universities better. Sentiment analysis is performed based on these tweets.

2.1 Source 1: Data.world

In this structured type of dataset, I have collected data on World Universities. World and cwur.org has been taken into considerations where all the necessary parameters such as University name, Location, National rank, Employment of Alumni, International students percentage, Student to Staff ratio, Awards given, Patents, Faculty quality, University income, Total number of students and lot of various other factors that would be help build the BI queries.

The Link for the cwur.org for world university rankings is https://cwur.org/2019-2020.php

2.2 Source 2: Times Rank

The second source of structured data that is used for Rank of Universities in America according to Times 2019-2020. It has 25 American universities along with their times rank. The link to the dataset: https://www.timeshighereducation.com/rankings/united-states/2020#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats

2.3 Source 3: Twitter Data

Reviews and ratings play a key role in selection of a University. So, I used Twitter to extract 1 lakh tweets on all the universities mentioned. Sentiment Analysis is done on the tweet data. Further obtained averages of positive, negative and neutral tweets for these universities. This is an unstructured type of data.

3. RELATED WORK

Before knowing what the top universities in the world are, it is very important to know how the universities are rated as top universities, based on several other factors contributing to it. To have an in-depth knowledge about how these ratings are, I have referred to some of the papers on the topic. The author [1] has justified on how the scores are distributed among the universities. It's claimed that the scores are given based on various factors like how competitive the university is in comparison with others and how reliable the scores are. The Times ratings are considered in this and seem to be one of the most reliable ratings given to the universities based on several factors like as Staff, Environment, Financial Cost etc. In my project I have used the Times ratings as one of the influencing parameters in a student's decisions. The scores and ratings are calculated based on the gap between the higher universities and the lower universities. Whereas the author [3] states that the reviews are given by the people which act as an input for these scores and it also states that these inputs can be biased. It's also been stated that the universities compete in order to get their name listed in the world class universities but in this process, they forget their main goal of giving the best to the students. It's also been stated that rather than competing with other universities they should focus on getting the best they have and making it and valuable and more exciting for the students. Which will eventually classify them under the world class universities as the students would be able to achieve their goals and ambitions much more effectively and easily.

How the rankings should be awarded ideally? In this research author [2] has justified and expressed his views based on different inputs. a) distribution score b) correlation coefficients and also two most important things that drew my attention were c) quality versus quantity: In this the author has explained that quality of education is more important rather than quantity which can burden the students and would not be helpful in achieving their goals.

Education versus Research: this is one of the key points that must be considered while rating the universities. The research papers that have been published or the awards obtained by the universities contributes as one of the key points – they should focus on practical skillsets rather than just educating them ,which may or may not be helping them in their real-world scenarios. The rankings are necessary to know the impact and mark a said university has made on a global level. In this author has [5] claimed that every research in an individual stream is equally important in this research a comparison.

How much value does research play role in the Ranking of a university? To explain this the author [4] has done relative work in which it states that the greater number of citations, the better the knowledge of students and professors, and the more it gives ethical and logical values to the university's rankings. It's also observed that in recent times female Ratio of students is increasing exponentially in American universities. The American universities have higher rate of education and the quality and many universities are listed amongst the top 100 universities globally. The results also stated that among the top 100 they are about 20-27 universities from America. The government and politicians have also discovered that most of the best research papers are written by the students from the American universities. The Times higher education ratings have particularly been used for the prediction modelling; previous scores rated in THE TIMES have been listed as one of the most influential factors for the future analysis of the ratings. The results generated are very accurate and show how the ratings helped in the analysis for the future predictions.

4. DATA MODEL

For building my data model on university rankings, several columns have been taken into consideration that would help me to build and answer my BI queries.

THREE MAJOR ELEMENTS OF DATAWAREHOUSE ARE:

- SQL Server management studio - SSMS
- SQL Server integration services - SSIS

Tools that are used are as follows:

1. SQL Server Management studio: To create a Database and its Dimensions
2. Microsoft Visual studio
3. SSIS: to build the data-warehouse, staging area
4. RStudio: Web scrapping is done to extract data form web
5. Tableau: Data Visualization

In the project I have three structured and one unstructured data source that are used to build the Datawarehouse. Sentiment analysis is performed on the unstructured dataset. R is used for extracting the tweets.

Steps to build a Datawarehouse are:

1. Different data sources are used to gather the data.
2. A Datawarehouse is built using star schema using Kimball's approach.
3. ETL (Extraction, transformation, loading) is performed.
4. Automation of the data using R programming.
5. Finally loading of the data into the data warehouse

In this project Data Warehouse is built using Kimball's approach with star schema. Large amount of data is collected, and cleaning of the data is done using R programming. After cleaning, automation is done. The data from different sources is combined as one CSV file and ETL process is performed on the said CSV file. Kimball's approach is nothing but the creation of dimension table where in each dimension table will have at least one primary key which should be of type int and these primary keys of the dimension table along with some other necessary data from the dimension tables which are numeric together form the fact table. As in fact-table only numeric data can be stored.

KIMBALL'S APPROACH:

The Datawarehouse is built using Kimball's approach which is a bottom-up approach. Kimball's Datawarehouse architecture is also known as Datawarehouse bus. It provides a view in the organization data and can also be merged with a larger Datawarehouse, it doesn't require a normalized data model. In this approach data marts are built first. This approach consists of several dimension tables and one fact table. Each dimension table will have unique identity- that should be numeric and identical for that dimension, this unique identity is called the Primary key.

For every dimension there is a unique Primary key. All these primary keys along with some numeric measures acts as a foreign key in the fact table. Fact table consists of only numbers. With the help of these primary keys the fact table looks into that particular dimension. All the primary keys from the dimension table and measures in fact table must be numeric type of data only.

ADVANTAGES OF KIMBALLS APPROACH:

The importance of using Kimball's approach over Inmon's approach:

1. It is costlier to build Datawarehouse using Inmon's approach in comparison to Kimball's approach.
2. One of the major reasons behind using Kimball's approach over Inmon's approach. Inmon's Approach is not as flexible as Kimball's and it is not easy to make changes in the model.
3. Less work needs be put in ETL phase in Kimball's when compared to Inmon's.
4. In Kimball's approach data marts are built in the database and then with the help of SSAS and SSIS it is moved to the Datawarehouse investigation part of the data warehouse – via SSIS (Now integrated in MS Visual Studio itself).

STAR SCHEMA:

Star Schema is designed using Kimball's approach whereas snowflake schema is designed using Inmon's approach. Since I'm using Kimball's approach, I chose to build a star schema. In the star schema, there will be a fact table at the Centre, surrounded by multiple dimensions tables. Hence the design resembles a Star. It is the simplest type of Datawarehouse schema. It used for querying large datasets. It requires only a small number of joins and the query efficiency is good.

Characteristics of Star schema:

1. It has an easy structure of designing.
2. Every dimension of the star schema has only one-dimension table.
3. Every dimension table consists of number of attributes.
4. These dimension tables are joined to fact table with the help of foreign key.
5. Number of joints used are less as their query efficiency is very good.
6. It is very easy to interpret and uses less space.
7. Dimension tables are not normalized.
8. BI tools widely support this schema.

SSIS: SQL SERVER INTEGRATION SERVICES:

Before starting the SQL server, a database file is created with the name of the university. For every dimension table there has to be a primary key that should be an integer value. These primary keys of all the dimension tables are used through look up in the fact table - fact table can store only integer values. Once data is been stored in the SSMS SQL server. Dimension and fact table are created and stored in the SSIS SQL Server Integration Services. In the SQL server analysis service, the cube will be deployed.

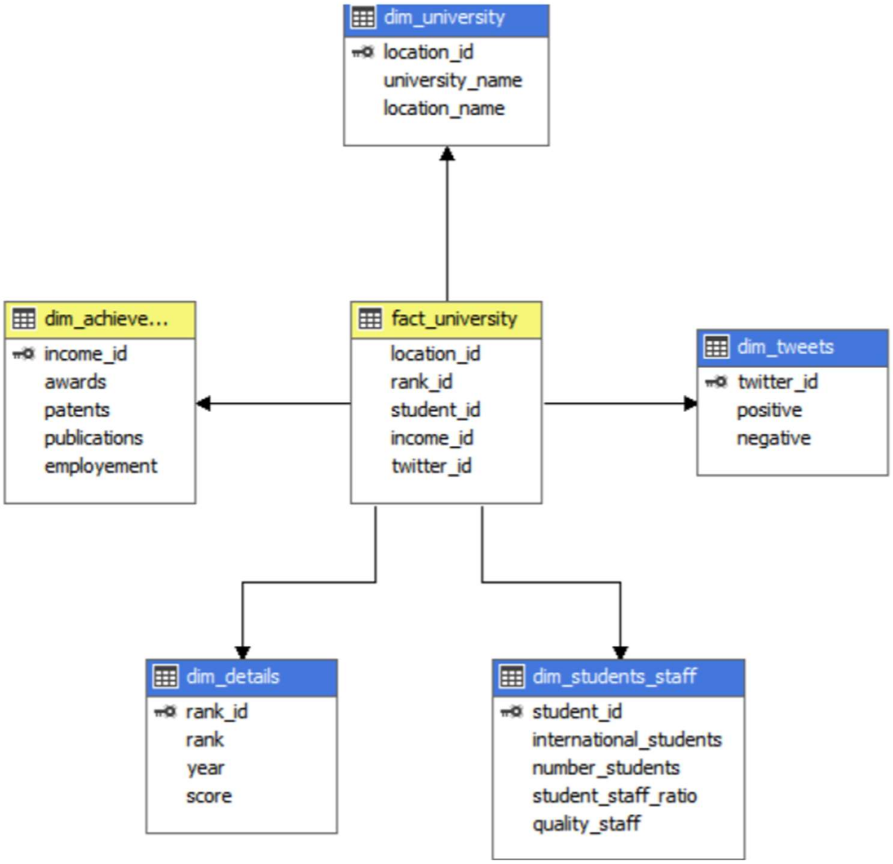


Figure : STAR SCHEMA

5. Logical Data Map

Source	Column	Destination	Column	Type	Transformation
ETL	Stu_id	Dim_faculty-stu	Stu-id	Fact	Primary Key, automatically generated through ETL
1	Total-number-students	Dim_faculty-stu	Stu-num	Dimension	Null, Cleaning performed using R
1	student-staff-ratio	Dim_faculty-stu	Faculty-stu-ratio	Dimension	Null, Cleaning performed using R
1	Faculty-quality-rating	Dim_faculty-stu	Teaching-quality	Dimension	Null, Cleaning performed using R
ETL	Earnings-id	Dim-achive	Earnings-id	Fact	Primary key, automatically generated through ETL
1	University-income	Dim-achive	Earnings	Dimension	Null, Cleaning performed using R
1	Awards-given	Dim-achive	Awards	Dimension	Null, Cleaning performed using R
1	Number-of-Patents	Dim-achive	Patents	Dimension	Null, Cleaning performed using R
1	Number-of-Publications	Dim-achive	Publications	Dimension	Null, Cleaning performed using R
1	Teaching-score	Dim-achive	Employ	Dimension	Null, Cleaning performed using R
1	Ranking-id	Dim-info	Ranking-id	Fact	Primary key, automatically generated through ETL
3	Times Higher Education Score	Dim-info	Times-rank	Dimension	Since in Times Rank only has 25 American university ranks are present. So, the rest of the universities rank are automatically generated through R

1	Year	Dim-info	Year	Dimension	Null, Cleaning performed using R
1	Score-rating	Dim-info	Score	Dimension	Null, Cleaning performed using R
1	International	Dim-info	International	Dimension	Null, Cleaning performed using R
ETL	Location-id	Dim-uni	Location-id	Dimension	Automatically generated through ETL
2	Name	Dim-uni	Uni-name	Dimension	Null, cleaning and removing all the special characters and blank spaces through R
2	Location	Dim-uni	Location-name	Dimension	Null, cleaning and removing all the special characters and blank spaces through R
4	Tweet-id	Dim-tweeter	Tweet-id	Fact	Primary key, automatically generated through ETL
4	Positive	Dim-tweeter	Positive	Dimension	Average of positive tweets is done separately, and sentiment analysis is been performed
4	Negative	Dim-tweeter	Negative	Dimension	Average of negative tweets is done separately, and sentiment analysis is been performed
4	Neutral	Dim-tweeter	Neutral	Dimension	Average of neutral tweets is done separately, and sentiment analysis is been performed
1	National-rank	Fact	National-rank	Measure	Null, Cleaning performed using R
1	Employment-alumni	Fact	Employ-alumi	Measure	Null, Cleaning performed using R
1	Number-of-citations	Fact	Citation	Measure	Null, Cleaning performed using R
1	Number-of-influences	Fact	Influences	Measure	Null, Cleaning performed using R

6. ETL Process

The most important part of building a Datawarehouse is ETL.

A. Extraction: After proper understanding and knowing what data is required to build the warehouse, extracting the data takes place. In the project the data has been extracted from four different sources has mentioned above in section 2, for university name, rank and location a website is been used, it is a table format hence considered as structured data but extracted using R, that would help in achieving the goal. This is the first step in the ETL process, extraction of data from different sources. For the university name, Rank, location – a website is used to extract - using R programming. The website data is in table format, so it's considered as structured data. The three columns are combined and stored in universities. Twitter dataset is also considered, and 1 lakh tweets are extracted for 1000 universities. Packages like 'rvest' and 'twitter' are installed to extract the tweets. All the 1 lakh tweets are extracted using a single R file. Since twitter allows only 10,000 tweets for every 15 minutes, hence 'testit' is used for it to recall automatically.

B. Transformation:

1. After data is extracted the next step is to perform all the necessary cleaning, manipulations and transformations.
2. The raw data gathered is then cleaned by randomly filling of all the missing values and all the missing values are replaced automatically using R.
3. Special characters and blank spaces are removed through R.
4. Some of the columns and rows are also deleted.
5. Times Data consists of only 25 rows, so data is generated randomly through R for remaining rows.
6. Sentiment analysis is performed using R for the 1 lakh tweets and average of positive, negative, neutral tweets are considered individually.

C. Loading:

First, we use ‘Execute Process’ Task - it automatically calls the CSV file directly through command prompt and loads into the SSIS. After the data is loaded, then all the dimensions are created by using flat file source and destination. After which we must change the directory and delete the output that is generated in the CSV. After that add window script path so that it can be called from command line automatically. Then the final output CSV is generated, in the SSIS, after which the CSV file should be normalized into dimension tables and fact table. All the dimensions and facts are indicated in the above table. Once Schema is generated with the help of dimensions and fact table then finally the cube is deployed.

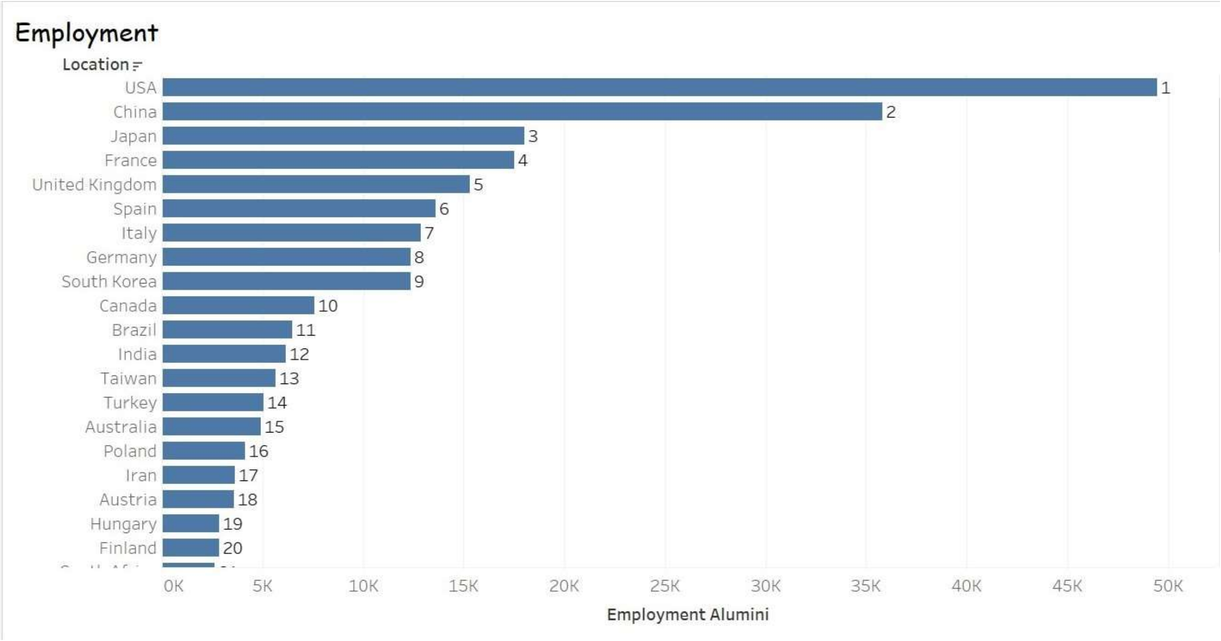
For creating the dimension tables and fact tables MS SQL Server Management Studio is used. Fact table is created with the help of look up that links to the dimension table by using the primary key of the dimension table.

7. APPLICATION

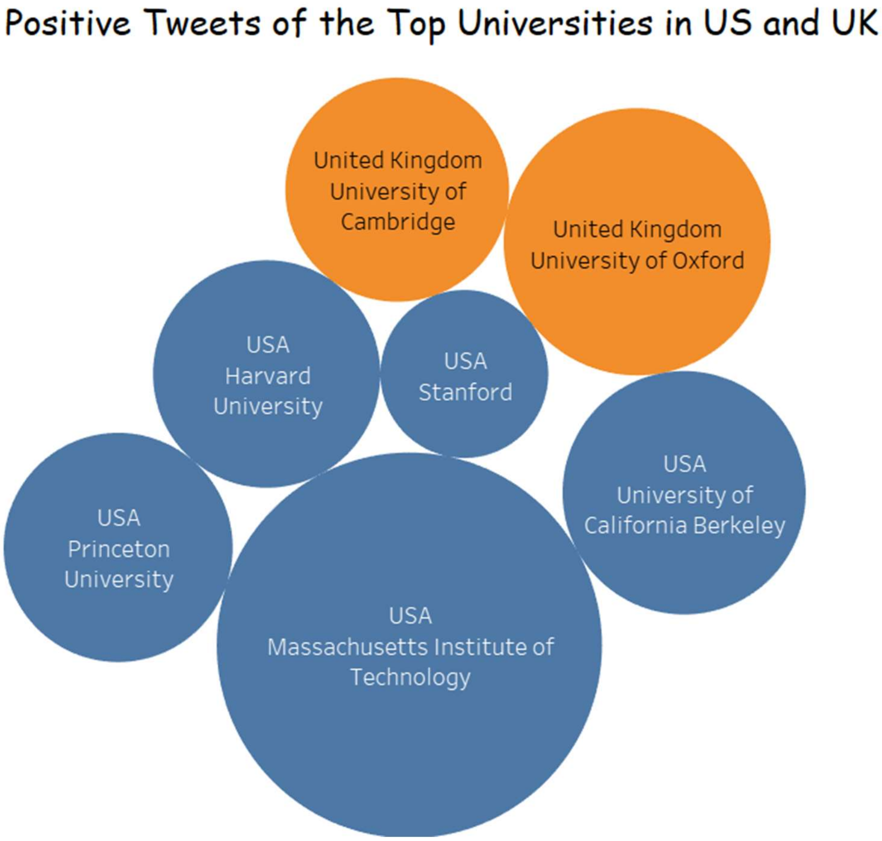
The below BI queries would not only help the students to decide in their choose of universities and course work, but it will also help the universities know their competition and what they can do to help the students and make the overall university experience better in comparison.

Query 1: Employment Rate World-wide

The below visualization is about the employment rate and it is shown by comparing various countries. It can be inferred that USA has the highest employment rate and employment status compared to other countries, followed by the other states like China and Japan. This shows and helps many other states to find in depth knowledge on why USA has a good employment rate and the other state universities can work and organize themselves in a similar way. As it is a well-known fact that USA is considered as a place where people come to fulfill their dreams, it seems only appropriate that the status of employment is very good compared to many other countries all over the world.

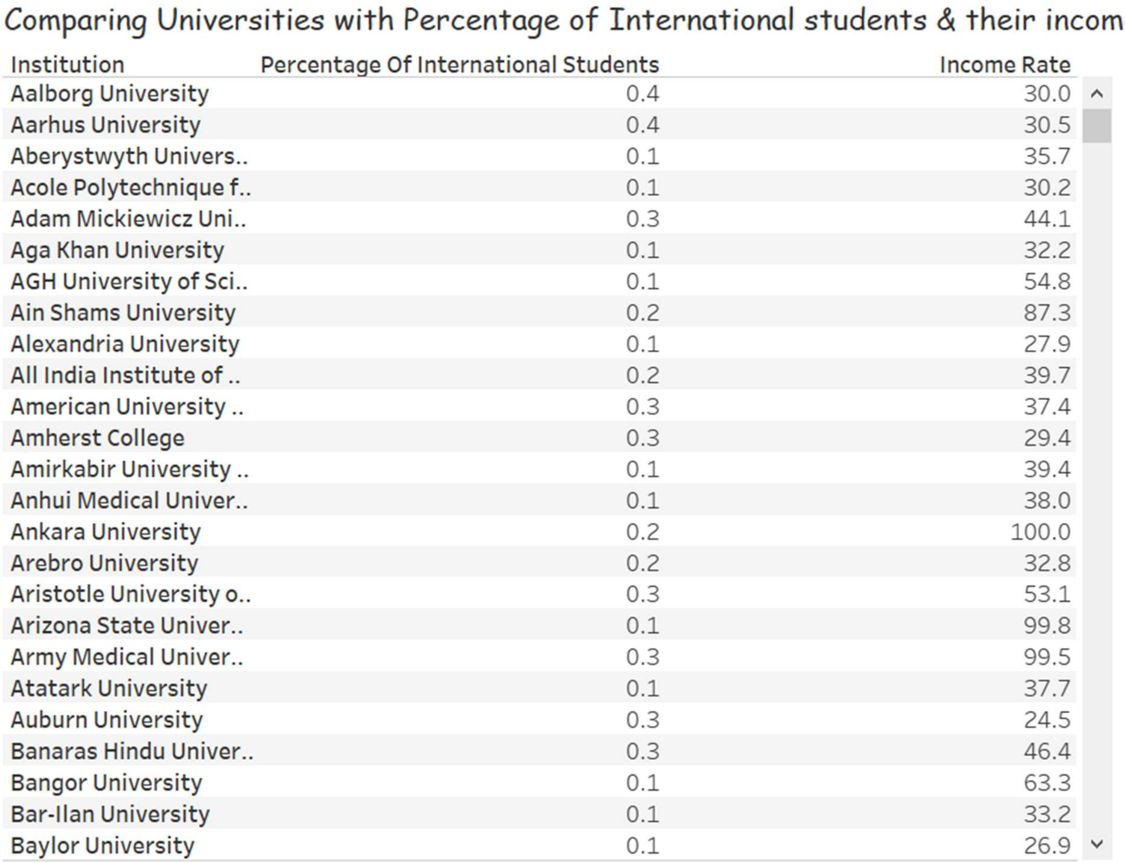


Query 2: Average Positive Tweets in USA and UK



In the above visualization, the Positive tweets are taken into consideration along with the comparison of two different Countries. I have considered universities from UK like University of Oxford, University of Cambridge and three other universities and compared them with different Universities in the USA like Harvard, Massachusetts and three other universities. And it is clearly shown that Massachusetts Institute of technology has highest number of Positive tweets compared to UK's Universities like Oxford and Cambridge.

Query 3: Comparing Universities with Percentage of International Students and Income Rate

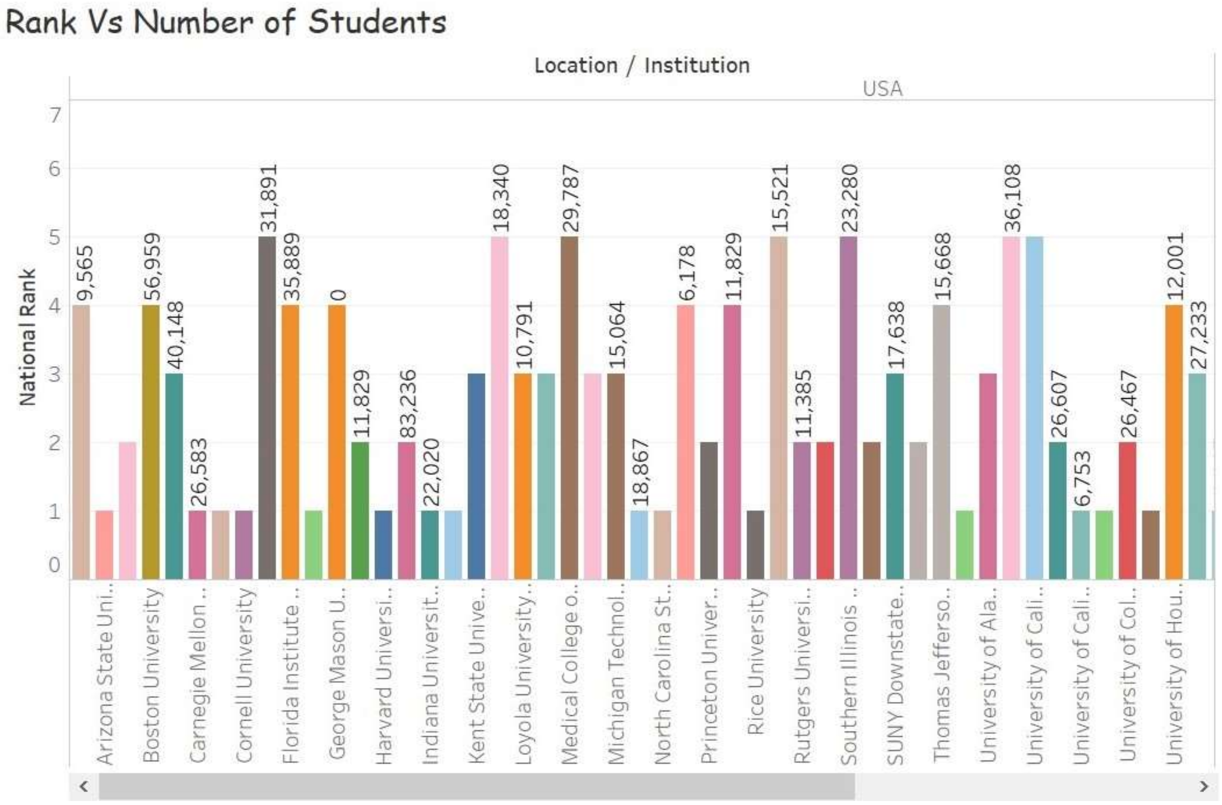


Comparing Universities with Percentage of International students & their income

Institution	Percentage Of International Students	Income Rate
Aalborg University	0.4	30.0
Aarhus University	0.4	30.5
Aberystwyth Univers..	0.1	35.7
Acole Polytechnique f..	0.1	30.2
Adam Mickiewicz Uni..	0.3	44.1
Aga Khan University	0.1	32.2
AGH University of Sci..	0.1	54.8
Ain Shams University	0.2	87.3
Alexandria University	0.1	27.9
All India Institute of ..	0.2	39.7
American University ..	0.3	37.4
Amherst College	0.3	29.4
Amirkabir University ..	0.1	39.4
Anhui Medical Univer..	0.1	38.0
Ankara University	0.2	100.0
Arebro University	0.2	32.8
Aristotle University o..	0.3	53.1
Arizona State Univer..	0.1	99.8
Army Medical Univer..	0.3	99.5
Atatark University	0.1	37.7
Auburn University	0.3	24.5
Banaras Hindu Univer..	0.3	46.4
Bangor University	0.1	63.3
Bar-Ilan University	0.1	33.2
Baylor University	0.1	26.9

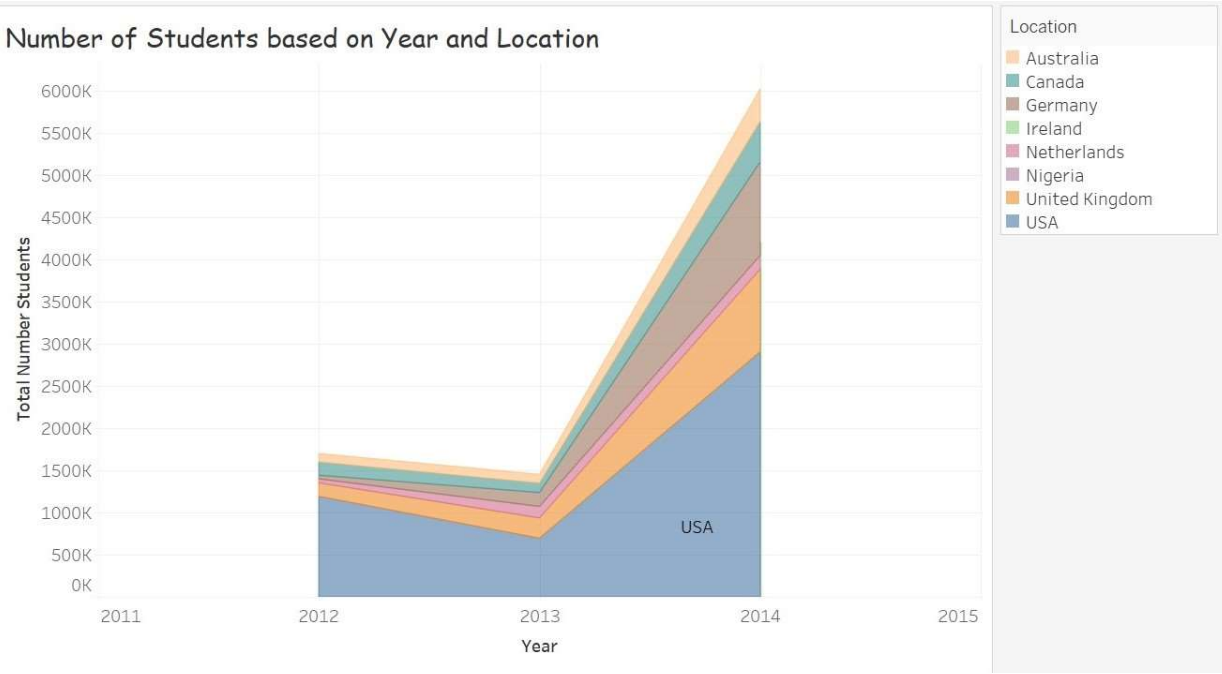
In the above visualization, we can see that the international students' admissions are less even if the university has high income standards. So, it is more likely that the university has many high requirement standards that many of the international students could not meet, or they are more likely to have good benefits for local students compared to international students. Admission or Tuition fee could also be one of the criteria that either it might have created this gap for international students. So, this would also help for the students in future.

Query 4: How does Rank of the universities effect the number of students taking admissions



In the above visualization, we can see the Correlation between rank and number of students amongst American universities. It can be inferred that even if the rank of university is more the number of students studying in that university could vary (less). Hence there is no direct correlation between the Rank given to the University and Number of students taking admission in that university. This shows that some university may not have better rank but have other good facilities and quality of education (Or any other influential factors).

Query 5: Is it necessary that the country with highest number of students in the past is more likely to have the same in future.



In the above visualization, we are predicting the number of students going to different countries for future studies based on the Location and Year. Based on the visualization in 2012 USA had the highest number of students and it has doubled in the year 2014. Whereas for Germany, the number of students in the year 2012 was comparatively low but has drastically increased more than double in the year 2014. Despite the number of students being more in USA, it can be predicted that students are more likely to choose Germany over USA in the future based on the trend.

8. CONCLUSION

The objective of this project is to determine the influence of parameters like World Rank, Location, Employment Rate, Publications, Awards, Tweets- whether Positive or Negative have over a students' choice of a university.

The data and the BI queries have solved many problems on how score rank and other factors play a major role in the universities rankings and ratings. Twitter is also seen as one of the important influences as the tweets are generally come from either the students or staff in the university or Alumni of the university. It can also be seen that the number of students going to America has increased immensely. Hence, we can predict that majority of International students would prefer American universities in the future for their career growth and advancement. Irrespective of the rank, positive tweets are higher in number with respect to American universities in comparison to universities in other countries.

9. FUTURE WORK

The limitation that I found in my Datawarehouse would be that, I could not find a proper, consistent and accurate data and could not completely solve the BI problems. If in the future, various other parameters based on students and University details are available and are taken into consideration then it would be easy to understand and predict more accurately as to what basis and circumstances influence a student's choice to opt for a university or country. This would help the business expand in the good and standard way.

10. References

https://www.4icu.org/us/

https://cwur.org/2019-2020.php

https://www.timeshighereducation.com/world-university-rankings/2020/world-ranking

https://www.timeshighereducation.com/rankings/united-states/2020#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats

https://tdan.com/data-warehouse-design-inmon-versus-kimball/20300#

https://www.mssqltips.com/sqlservertip/2976/comparing-data-warehouse-design-methodologies-for-microsoft-sql-server/

https://www.geeksforgeeks.org/star-schema-in-data-warehouse-modeling/

Balatsky, E. V. & Ekimova, N. A. (2018), `World class universities: Experience of identification', MIROVAYA EKONOMIKA I MEZHDUNARODNYE OTNOSHENIYA 62(1), 104{113

Barra, C., Maietta, O. W. & Zotti, R. (2018), `Academic excellence, local knowledge spillovers and innovation in europe', Regional Studies pp. 1{12.

Dowling-Hetherington, L. (0), `Transnational higher education and the factors influencing student decision-making: The experience of an irish university', Journal of Studies in International Education 0(0), 1028315319826320.

Gentile, T. A. R., De Nito, E. & Vesperi, W. (2016), A survey on knowledge management in european universities through e-learning, in `European Conference on Knowledge Management', Academic Conferences International Limited, p. 282.