Project Report ETL

TEAM AGILE

Faisal Malik ***Miki Minhui Cai*** ***Praveena MS***

Table of Contents

Abstract

Our client WA University Sales and Marketing Department are planning to update their course structure based on "in-demand" qualifications resulting to high level of employment rate and income level.

Purpose

This report is to provide the information with relation to employability and EBITDA.

Scope/ Methodology/ Approach

Scope of this report is limited to the Australian industries.

ETL is the methodology used for this project i.e. extraction, transformation and loading of data.

The approach used is to provide the latest data from authentic government sources for further decision making to university management.

Tools/Programing

Jupiter Notebook

Python

PostgresSQL

Originality/Values

The learning outcome of this project is understanding of ETL project.

Introduction

During the last decade higher education has been evolved as an industry. Most of the universities around the world are designing their courses as per industry/market demands. Also, with the advancement of technologies the overall business dynamics are changing resulting in changing the business revenues. In future some business may become more profitable than others. This can be seen from the changing revenues movements over subsequent years. This revenue movement has effect on the employability of the different industry sectors. The sectors which have potential to be more profitable in future attracts more people, so academia focus more on courses related to these courses.

Universities receive feedback from the industries about the future requirements in different sectors, so they design the course contents and capacity as per upcoming requirements. Recent reports from some authentic sources like government also help to universities in future planning as well. The combination of revenue and employability movement will facilitate the decision process.

Data Sources

- ***Data Sources***

  Main source of our data is ABS Australian Bureau of Statistics. We chose this source as it is authentic and extract from different sources done by government. We can download various datasets related to the research topic. The datasets include estimates derived using a combination of data from the Economic Activity Survey and business tax data sourced from the Australian Tax Office.

- ***Data Period***

  2017-19 financial year data is used to keep the scope to research limited to recent past years.

- ***Limitations***

  Data for FY2020 is not available because this year the world went through the covid-19 pandemic and data might not be reliable.

  We also planned to get some data by web scrapping but dropped due to limitation of time.

ETL Process

The steps involved in ETL process have been discussed below.

- ***Data Extraction***

  Data was extracted from ABS web site (abs.gov.au). Data was available for downloading in different formats. We choose to download in CSV format as it is easy to use/load in Python. Two files downloaded were:

  1) Annual industry EBITDA movements
  2) Annual industry employment movements

- ***Data Transformation***

  In the transformation process we have done some operations on the extracted data. It includes data cleaning, data formatting and data sorting. In detail we have done renaming the columns header, cleaning the date to drop any null values and duplicates, set the index as per industry. We didn't need to drop any column as we have most relevant information.

  We faced the problems related to format of data as it numbers was in string format, so we have to convert string into integer format. Also, we had to extract the comma (,) as it was causing problem in loading into SQL database.

These transformations were necessary to load data in SQL specific format.

Table 1: Raw Data from CSV File

	Annual industry EBITDA movements, 2016-17 to 2018-19	Unnamed: 1	Unnamed: 2	Unnamed: 3
0	NaN	2016-17 ($m)	2017-18 ($m)	2018-19 ($m)
1	Agriculture, forestry and fishing	3,694	-2,258	-998
2	Mining	23,313	21,399	33,746
3	Manufacturing	-1,064	4,201	3,346
4	Electricity, gas, water and waste services	3,254	481	670

Table 2: Clean Data for use in SQL

industry	y2016_17m	y2017_18m	y2018_19m
Agriculture, forestry and fishing	3694	-2258	-998
Mining	23313	21399	33746
Manufacturing	-1064	4201	3346
Electricity, gas, water and waste services	3254	481	670
Construction	435	3404	2504

- ***Data Loading***

  The final step was to load the data in SQL database. We created simple schema for Postgres because it was simple two tables, we didn't need any data modeling took like http://www.quickdatabasediagrams.com.

  We used pandas in Jupyter notebook to load data into Postgres SQL database. Postgres SQL is used instead of MongoDB because we don't have not massive data base and our data source is structured CSV based data.

  The two tables were merged on Industry bases combining the earnings and employability in one query table. Further data analysis by queries was possible but it dropped due to limited time and scope of this project.

Table 3: Combined employability and earning data query in SQL

	industry text	emp_2017 text	emp_2018 text	emp_2019 text	earnings_2017 integer	earnings_2018 integer	earnings_2019 integer
2	Mining	-7	12	10	23313	21399	33746
3	Manufacturi...	-7	10	14	-1064	4201	3346
4	Electricity, g...	1	3	5	3254	481	670
5	Construction	14	47	4	435	3404	2504
6	Wholesale tr...	15	-3	17	-310	3197	1099
7	Retail trade	-16	27	21	873	516	-325
8	Accommod...	34	20	2	-972	23	132

Conclusion

As the two tables related to employability and earnings have been combined in SQL, it can be further analyzed by universities to see which professions are in demand as well as profitable. Although there are multiple factors involved but this data can be one of the factors in planning the future courses. By plotting the graphs and data visualization can also help in decisions.

Further research based on this report can be done in the future as a more advanced and detailed project.

Abbreviations

- ABS: Australian Bureau of Statistics
- EBITDA: Earning before Interests, Taxes, Depreciations and Amortization
- SQL: Structured Query Language

Reference List

- https://www.abs.gov.au/statistics/industry/industry-overview/australian-industry/latest-release