ETL PROJECT REPORT

Sam Wimberly | Veronique Singh | Rebeca Hassan
Fiona Nguyen | Justin Stubbs

June 2019

Summary

For this project, we were interested in how the salaries of immigrant workers compared to average salaries in the US. We found transactional data from the Department of Labor and Statistics website detailing individual H-1B visa applications, as well as summary data from the US Census Bureau website showing median household income in each State in the US. After cleaning the datasets in Jupyter Notebook, we loaded the data into two SQL tables using Postgres for later analysis.

1. Extract

We used data from the following sources:

- H-1B Data: https://www.foreignlaborcert.doleta.gov/performancedata.cfm
- Census Data: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml

The H-1B dataset came in the form of two separate Microsoft Excel files, one for the year 2017 and one for 2016. These Excel files initially were about 240 MB with 52 columns and approximately 630,000 entries each. Each entry in this dataset represents a distinct H-1B visa application (denoted by a case number).

The Census dataset came in the form of two separate Comma Separated Value files, one for the year 2017 and one for 2016. These .csv files each contained 4 columns and 53 rows of data. Each entry in this dataset represents the median household income for each State in the US.

2. Transform

To make the H-1B dataset fit our purposes, Jupyter Notebook was used to clean the data. The dataset was trimmed significantly by dropping from 52 columns down to the 20 we needed for our project as well as dropping any entries with no data in any of the columns we care about. The final step was ensuring that the two “Zip Code” and “State” columns were formatted correctly, since this will be the primary means of tying our two datasets together. This was done with the following lines in Jupyter Notebook:

This line splits apart any 9 digit zip codes and drops the part after the ‘-‘

```
df17['WORKSITE_POSTAL_CODE']= df17['WORKSITE_POSTAL_CODE'].str.split('-').str[0]
```

This line ensures leading zeroes are re-added to zip codes (northeast US zip codes start with a 0)

```
df17['WORKSITE_POSTAL_CODE'] = df17['WORKSITE_POSTAL_CODE'].apply(lambda x: '{0:0>5}'.format(x))
```

After cleaning the H-1B data, both years were saved into separate .csv files of about 114 MB, 20 columns, and 620,000 entries each.

For the Census dataset, some minor cleaning was required to ensure the two years downloaded had matching columns as well as removing summary rows for the dataset. The final step was to append the two tables together for easier loading into the final database. All of this was accomplished using Jupyter Notebook.

3. Load

After extracting and transforming the data, the approach chosen was to load it into a relational database. The advantages of SQL were relevant to our project as we identified our database is transactional and there is a need for primary and foreign keys to create joins between the tables with a common identifier. Below is the list of tables created, where data for 2016 and 2017 was loaded by importing the transformed CSV files into “workvisa_stateincome_db” in Postgres.

“H1B”

COLUMN	DATA TYPE
ID (PK)	INTEGER
CASE_NUMBER	VARCHAR
CASE_STATUS	VARCHAR
CASE_SUBMITTED	DATE
DECISION_DATE	DATE
VISA_CLASS	VARCHAR
EMPLOYMENT_START_DATE	DATE
EMPLOYMENT_END_DATE	DATE
EMPLOYER_CITY	VARCHAR
EMPLOYER_STATE	VARCHAR
EMPLOYER_POSTAL_CODE	VARCHAR
JOB_TITLE	VARCHAR
SOC_CODE	VARCHAR
SOC_NAME	VARCHAR
PREVAILING_WAGE	FLOAT
PW_UNIT_OF_PAY	VARCHAR
WILLFUL_VIOLATOR	VARCHAR
WORKSITE_CITY	VARCHAR
WORKSITE_STATE	VARCHAR
WORKSITE_POSTAL_CODE	VARCHAR

“Income”

COLUMN	DATA TYPE
ID (PK)	INTEGER
RANK	INTEGER
STATE	VARCHAR
MEDIAN_HOUSEHOLD_INCOME	FLOAT
MARGIN_OF_ERROR	FLOAT
YEAR	INTEGER