**ETL Report Guide**

Parth, Nick, Alex, and Han

04/22/2022

**Introduction**

The Annual Business Survey provides information on selected economic and demographic characteristics for businesses and business owners by sex, ethnicity, race and veteran status. Using the data provided by the Census Bureau, we looked to understand how a person’s background affects their business ownership and employer status. We were also interested in the average number of employees per state. Before we can answer any questions, the first step is to transform the data. Census data is very messy and the shear size of the dataset can make it confusing to understand. Cleaning the data is very important to help us truly understand what the data can reveal.

**Data Sources**

Bureau, U. S. C. (2021, October 14). *Annual Business Survey (ABS) apis*. Census.gov. Retrieved April 22, 2022, from https://www.census.gov/data/developers/data-sets/abs.2019.html

**Extraction**

The data we used was from the Census Bureau website. To retrieve the information we used API calls. We all had special keys in order to retrieve the data to our devices. Using jupyter notebook/labs, we used html requests to retrieve the data via json format. We created a variable to save our url and make changes to the url for the columns needed. After declaring our url into a variable, we used that variable to request a ‘GET’ method to retrieve the data in a json format into a different variable with all the data. Using this new variable, a function was used to format the data into a dataframe where we did our transformation in order to merge all the tables necessary into one table. There were multiple data frames due to requesting different tables from the Census Bureau website. Below is a code snippet of the process taken to retrieve the data.

**Transformation**A screenshot of a computer

Description automatically generated with medium confidence

Once the data was extracted from the Census API, we were left with 4 different data frames that consisted of company summaries, business characteristics, business owner information and technology characteristics of a business. The following steps were taken to merge the company summary, business characteristics and business owners dataframes:

1. Merge the business owners and business characteristics data frames on GEO\_ID.
2. Rename the intersecting columns to useful names.
3. Drop columns that are doubled from the merge.
4. Using the new data frame, merge the company summary data frame on GEO\_ID.
5. Again rename the intersecting columns to more useful names.
6. Drop the duplicate columns due to the merge.
7. Change the EMP (number of employees) column to integer format

**Load**

To best load in the api’s into a database, a download to csv files would be the most efficient. Once the select api’s have been converted into csv files, the database is ready to be created.

1. Create a database
   1. use master;   
      go   
      drop database if exists (*name of database*); *------for ex. Census Dataset*  
      go   
      create database (*name of database*);   
      go *------this function will create the database and if it were to exist already it would drop it first before making to avoid conflicts*
2. Create tables (repeat for according table) *------ insert names of tables \*\*\*from ERD if created\*\*\* (company summary, characteristics of business, characteristics of business owners, technology characteristics of businesses)*
3. For tables that have foreign keys
   1. constraint fk\_(own table name)\_(borrowed table name)
   2. foreign key (ID of borrowed table)
   3. references [borrowed table name](ID of borrowed table)
   4. repeat for all FK keys
4. Once table schema is complete, load in the data
5. Navigate to *import wizard*
6. Input connection details
7. Select file to be imported
8. Verify data, data types, and fix any errors
9. Once cleared, data will be able to be transferred in.
10. To transfer from denormalized state, use the *insert* function
    1. Repeat until all data is transferred in the respected tables.

**Conclusion**

In conclusion, from the Annual Business Survey we were able to extract the information needed. The data we needed were called using the HTML ‘GET’ method from multiple different sources. After extracting all the information we had to transform our data into data frames in order to merge our data into one table. Some columns were not necessary and dropped them to use for our visualizations and data gathering. Once this was completed, we could load our files into any multiple different formats such as json, csv, excel, and many more. We were able to understand more about the business owner’s characteristics, company summary, characteristics of the company, and technology characteristics.