Assessment ETL Report

Group 5:
Jack Lynn
Stanley Perez
Tavianne Kemp
Sam Wainright

Process Date: 17 July 2022

## Introduction

The data supplied by the Census Bureau provided the initial raw data. The ETL process aimed to keep as much relevant data so that filtering could be conducted through Python.

The ETL process was completed after forming hypothesis and exploratory data analysis.

The 12 questions: align with final goal which is to investigate what demographics are disproportionally affected in the US workforce:

1. Which states contribute the most to the overall data?
2. Do males and females differ in the number of business firms across the US?
3. Of the races being studies, how much do they own and in what industries?
4. Do businesses differ in number of employees by sex? By ethnicity?
5. Does the owner ethnicity or sex affect the average amount of income for their employees or business?
6. On average, how much yearly income do the different races receive?
7. Which race contributes the most to number of businesses owned?
8. What is the degree of race representation in business ownership?
9. What is the degree of race representation throughout the industries?
10. How frequently do businesses record joint ownership?
11. What is the trend for amount of businesses owned by females?
12. To what degree do males and females own businesses and in what industries?

## Data Sources

Links to the data were supplied in the Pandas and Visualizations module for the Dev10 Data Fundamentals course on BrightSpace (Genesis10, 2021). The Section “Part 1: Get into Groups and Review the API” provided three links: “Census Bureau” (United States Census Bureau, 2022), “Annual Business Survey (ABS) APIs for 2019” (United States Census Bureau, 2021), and “API User Guide” (United States Census Bureau, 2021a). All three to be used in the ETL process. The Census Bureau (2022) page links to all API datasets, ABS 2019

(2021) navigates to the specific dataset used in this report, and the API User Guide (2021a) is a guidance page for developers. Additionally, one extra dataset comes from the Census Bureau's QuickFacts web interface system and is used to get data regarding racial and ethnic breakdowns of the United States (2022a).

**Extraction**

*Where did you get the data from? How did you get the data? What format is the extracted data? What steps were taken to extract the data? Be sure to number steps when the order matters.*

Initial extraction

config.py file holds the API key and the URL's

The Census Bureau allows for a custom data extract process:

1. Requires that the researcher navigate to the API User Guide (2021a) and request a key. The API key is returned to the email address entered.
2. The ABS dataset page (2021) allows for custom extraction. Each dataset, except the demographic dataset which is extracted from a public web interface, used has a unique API call and variables that can be associated with it. To create a custom data extraction begin the URL with the API call (the host name, year, and acronym) followed by ?get=.
3. Add any variables and variable predicates if applicable, separated by commas. Both predicates and geographies are separated with an ampersand to separate it from the get clause, i.e. &for=state:*.
4. The last aspect of the custom extraction URL is the emailed API key formatted as &key={api key}.
5. Variables chosen for the company summaries, states summaries, and characteristics of business reflects the goal of the project; all variables that correspond to demographic info were included as well as unique variable to each dataset. The demographic dataset does not allow for the selective inquiry of variables, since it is pulling from an HTML webpage rather than an API.
6. All three 2019 ABS and the one 2020 QuickFacts datasets were extracted and uploaded via Jupyter Notebooks. Group 5 found that looking at the data as a whole prior to transformation allowed for more efficient planning and better understanding of the data. The custom extraction URLs are stored directly in the ETL Jupyter notebook, with the API key stored in a separate config.py file that is not saved on the public directory. The demographic dataset does not require the API key, as it is extracting from a public web interface.
7. In Jupyter Notebooks, import requests and Pandas. Check the connection of each URL, convert it into JSON format, and append the data into a list to create a data frame for each dataset.

**Transformation**

*Did you use all of the data you extracted as-is? Did you remove columns? Did you change columns' names? Did you change your column formats? What steps were taken to get the data in a form that you could use it? Be sure to number steps when the order matters.*

Each dataset, except the demographic dataset, underwent the same initial cleaning process by being passed through a defined function (def commonClean). This function removes census code groups (any column with _S and/or _F), assigns float to data types that correspond to percent columns, and int to other numeric columns, assigns None to any columns that are equal to zero, creates a dictionary of variables that are labeled as meaningless for .drop(), and creates a dictionary for renaming the data headers into titles more descriptive.

```
def commonClean(data, code_groups=[], meaningless_groups=[],
rename_dict=dict(), number_cols=[])
```

Each dataset undergoes its own unique cleaning that involves removing columns that are not relevant to the hypotheses posed by passing in the variables into the lists established by the commonClean function.

Summary dataset holds census variables that involve the sex, ethnicity, and veteran status of the owner, the degree of ownership for a business, employee count, revenue, and payroll data. States Owner dataset includes the same, but also includes state name. Owner dataset addresses the owner's business establishment year and the business industry. Data included in the Tech dataset is anything that addresses the business' extent of technology use, production, and avoidance level.

The demographic dataset goes through a unique cleaning process, as it is the product of a web interface instead of an API. After the requests library is used to get the HTML, this data is passed into BeautifulSoup to interpret the HTML. From there, the table, defined by "tbody" tag and "Race and Hispanic Origin" "data-topic" attribute, is pulled from the HTML using BeautifulSoup's find function. Next, each row of the table is iterated through, and both BeautifulSoup's find function and string interpolation are used to extract racial and ethnic population percentages. These values are saved as dictionary instances in a list, which are eventually converted to a CSV file using the CSV library's DictWriter function.

**Load**

Each dataset is given a meaningful name--summary_dataset, states_dataset, owner_dataset, and race_ethnicity_dataset--and saved as CSVs in the data directory. All datasets, and the ETL Jupyter Notebook can be found in the GitHub repository.

**Conclusion**

Following ETL process create visualizations. Exploratory data analysis and visualizations are to be completed with Python using Pandas, Matplotlib, and Seaborne.

**References**


Genesis10. (2021, August 12). Assessment: Census API Data - Annual Business Survey 2019. Retrieved July 15, 2022, from Brightspace website: https://stage3talent.brightspace.com/d2l/le/content/6899/viewContent/17439/View

United States Census Bureau. (2021a, October 8). Census Data API User Guide. Retrieved July 15, 2022, from census.gov website: https://www.census.gov/data/developers/guidance/api-user-guide.Overview.html

United States Census Bureau. (2021b, October 28). Annual Business Survey (ABS) APIs. Retrieved July 15, 2022, from census.gov website: https://www.census.gov/data/developers/data-sets/abs.2019.html

United States Census Bureau. (2022, July 13). Census API: Datasets in /data and its descendants. Retrieved July 15, 2022, from census.gov website: https://api.census.gov/data.html

United States Census Bureau. (2022a). QuickFacts. Retrieved July 15, 2022, from census.gov website: https://www.census.gov/quickfacts/fact/table/US/PST045221