**ETL Report Guide**

Will, Luis, Phil, Sharif  
ETL 4/22/2022

**Introduction**

For this assessment, we were working with Census Annual Business Survey (ABS) dataset. The ABS dataset provides information on selected economic and demographic characteristics for businesses and business owners by sex, ethnicity, race, and veteran status. There are four different tables within ABS dataset, titled as:

**Company Summary** - Provides data for employer businesses by sector, sex, ethnicity, race, veteran status, years in business, receipts size of firm, and employment size of firm for the U.S., states, and metro areas

**Characteristics of Businesses** - Provides data for respondent employer firms by sector, sex, ethnicity, race, veteran status, years in business, receipts size of firm, and employment size of firm for the U.S., states, and metro areas

**Characteristics of business Owners** - Provides data for owners of respondent employer firms by sector, sex, ethnicity, race, and veteran status for the U.S., states, and metro areas

**Technology and Characteristics of the Businesses -**Provides data on technology use and production for Artificial Intelligence, Cloud-Based Computing, Specialized Software, Robotics, and Specialized Equipment technologies data at the U.S and State level for 2018.

Our work is combination of individual, as well as merged data derived from those tables. For each member of our team, we have decided to choose individual hypothesis to check and run analysis upon. Those individual works will be referenced as separate sections of this report.

**Data Sources**

ANNUAL BUSINESS SURVEY (ABS) APIs. Accessed on 4/22/2022

- <https://www.census.gov/data/developers/data-sets/abs.2019.html>

COMPANY SUMMARY.

- Examples: [*api.census.gov/data/2018/abscs/examples.html*](https://api.census.gov/data/2018/abscs/examples.html)

- Variables: [*api.census.gov/data/2018/abscs/variables.html*](https://api.census.gov/data/2018/abscs/variables.html)

CHARACHTERISTICS OF BUSINESSES

- Examples: [*api.census.gov/data/2018/abscb/examples.html*](https://api.census.gov/data/2018/abscb/examples.html)

*-* Variables: [*api.census.gov/data/2018/abscb/variables.html*](https://api.census.gov/data/2018/abscb/variables.html)

CHARACHTERISTICS OF BUSINESS OWNERS

**-** Examples: [*api.census.gov/data/2018/abscbo/examples.html*](https://api.census.gov/data/2018/abscbo/examples.html)

*-* Variables: [*api.census.gov/data/2018/abscbo/variables.html*](https://api.census.gov/data/2018/abscbo/variables.html)

### TECHNOLOGY CHARACTERISTICS OF BUSINESSES

**-** Examples: [*api.census.gov/data/2018/abstcb/examples.html*](https://api.census.gov/data/2018/abstcb/examples.html)

*-* Variables: [*api.census.gov/data/2018/abstcb/variables.html*](https://api.census.gov/data/2018/abstcb/variables.html)

**Extraction**

1. To extract the datasets we first requested an API key for each member of your group in this page (https://www.census.gov/data/developers.html).

2. In Jupytter notebook, we imported pandas, requests, matplotlib.pyplot and seaborn modules.

3. Using instruction provided in the main page of ABS dataset (<https://www.census.gov/data/developers/data-sets/abs.2019.html>), we created a *get\_file(url)*, and using requests.get(url) called the appropriate API for each tables of ABS dataset. Within that function we changed requested API into json() format and returned it. Then called get\_file() function on each API and assigned each result to a separate variable.

**Transformation and cleaning**

We performed following general transformation to datasets, before individual changes:

* Replaced columns row with first row, since 1st row represented actual column named.
* Removed 1st row after above step, since previous column row moved to be the first row, that we didn’t need.

Then each of us did the following transformation and cleaning steps separately:

* **Will**:  
  To begin with, to merge the various tables I created a couple merge point columns which consist of concatenations of other columns which reduce the unrelated overlap. Because the Characteristics of Business Owners didn’t have as many data fields as the other datasets a more specific merge point was made for the other tables and a less specific one to add the CBO data. I chose columns containing GEO\_ID, and details about the owner (sex, race, ethnicity, veteran status) as the first column to merge the CBO data with the company summary data (called [merge\_tag\_1]), and added to that PAYANN, EMP, and FIRMPDEMP which should be the same across all datasets if a row is referring to the same business (called [merge\_tag\_2] in abscs, abstcb, abscb). After dropping columns that would have repeated names and not add new value, I merged the tables together using these merge points as outer joins to keep as much data as possible, and didn’t have overly much overlap, so realized I would have trouble answering many cross-survey questions.  
    
  My merge resulted in very little overlap between the tables so I decided to focus on questions that could be answered using only one data set, namely “What impact does having Female ownership have on payroll, sales, and employees?” For my questions I wanted to compare the amounts of a company’s sales with their payroll, unfortunately all the state data we were looking at didn’t have non-zero values in the RCPPDEMP (Sales) categories, so I additionally loaded overall national data using the same process as the state data. In attempts to get as much usable data as possible, many unnecessary columns were included in the original url request, and then reduced down to the desired information afterwards. The columns retained were PAYANN, EMP, RCPPDEMP, SEX\_LABEL, as well as the column upon which to merge with the existing data, the created [merge\_tag\_2].  
    
  Once all the data was merged together I reduced the table to only those rows having non-zero data for PAYANN, EMP, and RCPPDEMP, as well as removing a few of the less relevant SEX\_LABEL values, keeping only those for Male, Female, and Equally male/female. With these restrictions in place, various visualizations were made examining my question of interest.
* **Luis**:

1. For my transformation and cleaning, I decided to do left join merges for the four data frames we had. For my base dataframe I used the c\_summary dataset.

2. After that I dropped all the columns that I would not be using as well as the duplicate columns using “.drop().”

3. After that I checked to see if there are any missing values.

4. Then, I checked to see if there were any duplicates, which there weren’t any.

5. Afterwards, I renamed my column names to something cleaner.

6. Finally, I changed my “NUMBER\_OF\_EMP” column and my “NUM\_OF\_OWNERS\_FOR\_FIRM” column to integer types.

* **Phil** :

1. After Importing the data from the API, we had 4 tables to pull from:

c\_summary

char\_business

char\_business\_owners

tech\_char

I merged the tables with left joins on the ’NAME’ column, which was the name of the states involved

1. Though I was then able to use information from all 4 tables at this point, the data I selected for exploration was in the Summary table, ‘c\_summary’.
2. For my first visualization, I selected only the ‘NAME’, ‘EMP’, and ‘SEX\_LABEL’ columns. I changed the type of the ‘EMP’ (number of employees) from an object to an integer. I made sure to select counts of employees that were not overlapping by selecting only the labels of demographics that were labeled ‘Total’
3. For my second and third visualizations, I selected much of the same data, however for the ‘SEX\_LABEL’, I selected ‘Male’ and ‘Female’ respectively to get the breakdowns of numbers of employees by gender.
4. My fourth visualization involved looking at the ‘NAME’, ‘YIBSZFI\_LABEL’ (years in business label), and the ‘FIRMDEMP’ (number of firms) columns. I changed the value type of ‘FIRMDEMP’ from an object to an integer.

* **Sharif** :

1. Filtered columns in each table. For *c\_summary* table only kept 'NAME','RCPSZFI\_LABEL','EMP','PAYANN','YIBSZFI\_LABEL','EMPSZFI\_LABEL','SEX\_LABEL','ETH\_GROUP\_LABEL','RACE\_GROUP\_LABEL','VET\_GROUP\_LABEL','FIRMPDEMP' columns. For *char\_business* table only kept 'NAME','EMP','PAYANN','FIRMPDEMP' columns, for *char\_business\_owners* table only kept 'NAME','OWNPDEMP' columns and *for tech\_char* table only kept 'NAME','EMP','PAYANN','FIRMPDEMP' columns.

2. Changed all columns data types from object type to either string or intiger based on what the values represented.

3. Renamed all columns to names of what they represent based on variable’s list

4. Merged c\_summary table with char\_business table based on State column

5. Merged above merged table with char\_business\_owners tabe based on State column

**Conclusion**

Below are list of findings and conclusions we made, based on each individual work:

* **Will**  
  I merged the tables by constructing arbitrary keys and then needed to add additional data to answer the questions I had, so merged that in to and reduced the data down to just what I needed to begin to answer the questions I had.
* **Luis**

The first question I was trying to answer was: Comparing firms with least amount of years in business to firms with most amount of years in business, do the number of employees grow as years in business grows? (first 5 states). For this question I saw that yes, as years in business grows, so do the number of employees within that business. My second dealt with finding whether there is a relationship between the number of owners for all firms and the number of employees for all states. The answer to that was no. As the number of employees is decreasing, you can see that the number of owners goes up and down, so there was no relationship there. For my last question I was trying to figure out what the total number of minority employees in all firms for the first 10 states was. Although my visualization came in handy to give you an idea of the answer to my question, my visualization is strengthened by the table I created for it.

* **Phil**

Though it was possible to use the data from all of the endpoints, I found that the only useful portions for my visualizations came form the main ‘Summary’ table. The values in the data seemed to be fairly consistent, however the need to change value types of numbers from ‘string’ to ‘int’ was necessary.

* **Sharif**

Data within ABS dataset came mostly clean but disorganized. I needed to do transformation and datatype conversion. Other minor work included removing columns and rows, changing column names and later manipulating data to get needed visuals working.