**Census API Data – Annual Business Survey 2019: ETL Report**

Darrell Gerber, Grace Seiler, Pemba Sherpa, Amy Yucus  
1/17/2021

**Introduction**

Using the Annual Business Survey (2019) Technology Characteristics of Businesses, our team developed the real-world example of an Artificial Intelligence Company that is interested in knowing information regarding business that report the highest use of AI technology (for marketing and sales purposes). To further explore this scenario, we pose these questions:

* What firms already have high acceptance of AI (use is high)?
  + Where are high level adoptees located?
  + What types of industry are high level adoptees?
  + What size of businesses are high level adoptees?
* Does AI technology use affect the number of workers and worker skill level?
  + Does location change the impact of AI usage?
  + Does the size of the business change the impact of AI usage?
* What is the motivation behind firms using AI (specifically firms who report high use)?
  + What is the motivation behind business using AI that report high use?
  + Does motivation depend on the location?
  + Does motivation depend on industry type?
  + Does motivation depend on business size?
* What is the main factor that adversely affects the adoption or utilization of AI (besides not applicable and no factors affected)?
  + What is the proportion of Number of Firms by business that indicated this factor?
  + What is the proportion of Number of Firms by State that indicated this factor?

The data from the Annual Business Survey (2019) Technology Characteristics of Businesses API is broad and needs to be narrowed heavily to answer the questions our team has posed.

**Data Sources**

API data was provided by the U.S. Census Bureau. Our team used the Technology Characteristics of Businesses (2019) from the Annual Business Survey APIs (<https://www.census.gov/data/developers/data-sets/abs.2019.html>). API calls were made during the period of 1/14/2022-1/16/2022.

Bureau, US Census. “Annual Business Survey (ABS) APIs.” *Census.gov*, US Census Bureau, 14 Oct. 2021, https://www.census.gov/data/developers/data-sets/abs.2019.html.

**Extraction**

Data was extracted from the Annual Business Survey Technology Characteristics of Businesses (2019) API. The base URL for this API is: <http://api.census.gov/data/2018/abstcb>.

Question 1 -

1. Retrieve the data comparing the level of technology use by sector in US
   1. Query the US Census API with the call: https://api.census.gov/data/2018/abstcb?get=NAME,TECHUSE,TECHUSE\_LABEL,NAICS2017,NAICS2017\_LABEL,FIRMPDEMP&for=us&key=YOUR\_KEY
   2. Load the JSON response into a pandas dataframe TechUse\_Sector\_us
2. Retrieve the data comparing the level of technology use by business size in US
   1. Query the US Census API with the call: https://api.census.gov/data/2018/abstcb?get=NAME,TECHUSE,TECHUSE\_LABEL,NSFSZFI,NSFSZFI\_LABEL,FIRMPDEMP&for=us&key=YOUR\_KEY
   2. Load the JSON response into a pandas dataframe TechUse\_Size\_us
3. Retrieve the data comparing the level of technology use by state
   1. Query the US Census API with the call: https://api.census.gov/data/2018/abstcb?get=NAME,TECHUSE,TECHUSE\_LABEL,FIRMPDEMP&for=state&key=YOUR\_KEY
   2. Load the JSON response into a pandas dataframe TechUse\_state
4. Retrieve the dataset with the US Census regions and divisions for the states
   1. Read the raw csv file from the Github repository at https://github.com/cphalpert/census-regions/blob/master/us%20census%20bureau%20regions%20and%20divisions.csv

Question 2 –

1. Retrieve data comparing impact of technology use by firm size in US
   1. Query the US Census API with the call: http://api.census.gov/data/2018/abstcb?get=IMPACTWF\_U,IMPACTWF\_U\_LABEL,NSFSZFI,NSFSZFI\_LABEL,FIRMPDEMP,EMP &for=us:\* &key=YOUR\_KEY
2. Retrieve data comparing impact of technology use by firm size in US
   1. Query the US Census API with the call: http://api.census.gov/data/2018/abstcb?get=IMPACTWF\_U,IMPACTWF\_U\_LABEL,NSFSZFI,NSFSZFI\_LABEL,FIRMPDEMP,EMP &for=us:\* &key=YOUR\_KEY
3. Retrieve data comparing impact of technology use by US state
   1. Query the US Census API with the call: http://api.census.gov/data/2018/abstcb?get= NAME,IMPACTWF\_U,IMPACTWF\_U\_LABEL,NSFSZFI,NSFSZFI\_LABEL,FIRMPDEMP,EMP&for=state:\* &key=YOUR\_KEY

Question 3 -

1. Retrieve the data comparing the motivation of technology use by sector in US
   1. Query the US Census API with the call: https://api.census.gov/data/2018/abstcb?get= MOTUSETECH, MOTUSETECH\_LABEL, NAICS2017, NAICS2017\_LABEL, FIRMPDEMP &for=us&key=YOUR\_KEY
   2. Load the JSON response into a pandas dataframe motivation\_by\_industry
2. Retrieve the data comparing the motivation of technology use by business size in US
   1. Query the US Census API with the call: https://api.census.gov/data/2018/abstcb?get= MOTUSETECH, MOTUSETECH\_LABEL, NSFSZFI, NSFSZFI\_LABEL, FIRMPDEMP&for=us&key=YOUR\_KEY
   2. Load the JSON response into a pandas dataframe motivation\_by\_size
3. Retrieve the data comparing the motivation of technology use by state
   1. Query the US Census API with the call: https://api.census.gov/data/2018/abstcb?get= NAME,MOTUSETECH,MOTUSETECH\_LABEL,FIRMPDEMP&for=state&key=YOUR\_KEY
   2. Load the JSON response into a pandas dataframe motivation\_by\_state
4. Retrieve the dataset with the US Census regions and divisions for the states
   1. Read the raw csv file from the Github repository at https://github.com/cphalpert/census-regions/blob/master/us%20census%20bureau%20regions%20and%20divisions.csv

Question 4 -

1. Retrieve the data comparing the factors that adversely affected the adoption or utilization of the technologies to produce goods or services in US.
   1. Query the US Census APi with the call: [https://api.census.gov/data/2018/abstcb?get=NAME,FACTORS\_U,FACTORS\_U\_LABEL,FIRMPDEMP&for=us:\*](https://api.census.gov/data/2018/abstcb?get=NAME,FACTORS_U,FACTORS_U_LABEL,FIRMPDEMP&for=us:*)&key=YOUR\_KEY
   2. Load the JSON response into a pandas dataframe factors\_data
2. Retrieve the data comparing the factors that adversely affected the adoption or utilization of the technologies to produce goods or services and size of the firms in US.
   1. Query the US Census APi with the call: [https://api.census.gov/data/2018/abstcb?get=NAME,FACTORS\_U,FACTORS\_U\_LABEL,FIRMPDEMP,NSFSZFI,NSFSZFI\_LABEL&for=us:\*](https://api.census.gov/data/2018/abstcb?get=NAME,FACTORS_U,FACTORS_U_LABEL,FIRMPDEMP,NSFSZFI,NSFSZFI_LABEL&for=us:*) &key=YOUR\_KEY
   2. Load the JSON response into a pandas dataframe firm\_size\_data
3. Retrieve the data comparing the factors that adversely affected the adoption or utilization of the technologies to produce goods or services by state.
   1. Query the US Census APi with the call: [https://api.census.gov/data/2018/abstcb?get=NAME,FACTORS\_U,FACTORS\_U\_LABEL,FIRMPDEMP&for=state:\*](https://api.census.gov/data/2018/abstcb?get=NAME,FACTORS_U,FACTORS_U_LABEL,FIRMPDEMP&for=state:*%0d)
   2. Load the JSON response into a pandas dataframe state\_data
4. Retrieve US map .shp file by state from US census bureau.
   1. Download the .zip file from <https://www2.census.gov/geo/tiger/TIGER2019/STATE/> and extract it.
   2. Read .shp file using geopandas.read\_file(‘filepath’)

**Transformation**

Question 1 -

1. Remove the unneeded rows from the dataframes
   1. Remove all rows except TECHUSE==T3E03B05, TECHUSE==T3E03B06, and TECHUSE==T3E03BA9. These are the high use, Don't know, total responses for Artificial Information technology use.
   2. Reset indices to sequential
2. Remove the unneeded columns from the dataframes
   1. Remove the NAME and 'us' column from the us-based dataframes.
   2. Remove the 'state' column from the state-based dataframe.
3. Merge state regions with the state-based dataset
   1. Change column names in the datasets for the state abbreviations to match.
   2. Do an inner join to add the region and division labels to the state-based dataframe.

Question 2 –

1. Drop first row to make first row the column headers
2. Remove all rows except where IMPACTWF\_U is T1E05D01, T1E05D02, T1E05D03, T1E05D99, and T1E05D01
3. Remove ‘us’ and ‘state’ columns from dataframes

Question 3 -

1. Remove the unneeded rows from the dataframes
   1. Remove all rows except MOTUSETECH==T1E04C01, MOTUSETECH==T1E04C02, MOTUSETECH==T1E04C03, MOTUSETECH==T1E04C04, MOTUSETECH==T1E04C05, MOTUSETECH==T1E04C06, and MOTUSETECH==T1E04C99. These are all motivations related to AI.
   2. Reset indices to sequential.
2. Remove the unneeded columns from the dataframes
   1. Remove the 'us' column from the us-based dataframes.
   2. Remove the 'state' column from the state-based dataframe.
3. Merge state regions with the state-based dataset
   1. Change ‘NAME’ column in the state-based dataframes to ‘State’.
   2. Do an inner join to add the region and division labels to the state-based dataframe.

Question 4 -

1. Make 1st row as column header and remove that row from dataframe.
2. Make a list named AI that contains FACTORS\_U codes for Artificial Intelligence.
3. Filter factors\_data by AI list.
4. Filter firm\_size\_data by FACTORS\_U\_LABEL = 'Artificial Intelligence: Technology was too expensive'.
5. Filter state\_data by FACTORS\_U\_LABEL = 'Artificial Intelligence: Technology was too expensive'.

**Load**

1. Save files for later use
   1. Save each dataframe as a csv files. Example:
      1. TechUseHigh\_Sector\_us.csv,
      2. TechUseHigh\_Size\_us.csv, and
      3. TechUseHigh\_state.csv
2. Loading into a SQL database
   1. Use DDL to create tables for each of these dataframes
   2. Load the data into the SQL database using DML

**Conclusion**

Using the US Census Bureau’s Annual Business Survey (2019) Technology Characteristics of Businesses, our team is answering 4 major questions (each with several sub-questions) to explore the scenario of helping an Artificial Intelligence company sell and market to companies who might be interested. To answer these questions, we had to narrow the fields provided by the API to those that would contain the pertinent information needed. Once we retrieved the information from the API, we further reduced rows and columns to refine our dataframes. Finally, these dataframes were saved to CSVs, where they can be loaded into an SQL database, or (in our case) read back into Python for analysis.