**ETL Report Guide: Census Exploration**

Jakob Thunen, Daniel Stein, Christopher Ng, Alysia Halverson  
April 25, 2022

**Introduction**

Utilizing the [Annual Business Survey (ABS) APIs for 2019](https://www.census.gov/data/developers/data-sets/abs.2019.html) from the U.S. Census Bureau, our group wanted to answer a number of questions relating to employer businesses, respondent employer firms and their owners, and technology characteristics of businesses. The four datasets share results from ABS questions on selected economic and demographic characteristics for businesses and business owners by sex, ethnicity, race, and veteran status, as well as data collection regarding innovation, technology, and financing. It is important to note that the 2019 ABS covers data for reference year 2018.

To understand the data from the Census Bureau, it is highly recommended to thoroughly spend time exploring API call-formats, example API calls, geography hierarchy, and variables for each data set. [Technical information guides](https://www.census.gov/programs-surveys/abs/technical-documentation/api.2019.html) are available in PDF format for each of the datasets and can be accessed from the ABS APIs for 2019 site. For many of the API calls utilized, it will be important to access the codes and labels for various ABS API characteristic variables within the technical information guide to ensure both the question asked and the resulted answer are included in the resulting data. In addition, reviewing the [Census Data API User Guide](https://www.census.gov/data/developers/guidance/api-user-guide.Example_API_Queries.html) is highly recommended to ensure the correct dataset name acronyms, variables, and criteria are being passed through the API call.

When parsing through the data, we identified several questions we sought to better understand:

**Company Summary**

* Is there any relation to a company’s number of employees and its annual payroll?
* Which industries have the highest average annual payroll?
* From the highest-ranking industry from the previous question, what firm sizes have the highest average annual payroll?

**Characteristics of Businesses**

* How many Employees classify as only Hispanic and non-Hispanic?
* Does each NAIC have a similar number of employees?
* Is there a relationship between the number of owners and the number of employees?
* How do the number of employees compare for Family-Owned companies when stratified by owner’s race?

**Characteristics of Business Owners**

* In the U.S., which metropolitan areas had the highest spike in new business ownership in 2018?
* What differences are there between total U.S.-born and foreign-born business owners as well as aggregating by race group and gender?
* Are there differences among U.S. states regarding business owners and educational attainment, specifically high school graduates/GED holders versus Bachelor's degree holders?

**Technology Characteristics of Businesses**

1. In 2018 how did the average revenue of firms differ when separating them by use of Cloud-Based technology?
2. In 2018, how did each industry’s use of Specialized-Software compare with one another?

**Data Sources**

Through the [US Census Bureau government website](https://www.census.gov/data/developers/data-sets/abs.2019.html), we accessed the datasets on April 24th, 2022. The four datasets, Company Summary, Characteristics of Businesses, Characteristics of Business Owners, and Technology Characteristics of Businesses can be accessed through different API calls.

The API calls each include the base URL (https://api.census.gov/data/), the year (2018), the dataset name acronym (ABSCS, ABSCB, ABSCBO, ABSTCB) followed by ?get= (see specific dataset variable names), followed by &for= (see specific data set geography hierarchies), followed by one or more values for QDESC (question) code and/or label, OWNCHAR (owner characteristic) code or BUSCHAR (business characteristic) code, and other filters on characteristic data, as applicable.

U.S. Census Bureau (2019). Annual Business Survey (ABS). Retrieved from https://www.census.gov/data/developers/data-sets/abs.2019.html.

**Extraction**

Our data sources were from the four previously mentioned datasets for the 2019 ABS from the U.S. Census Bureau. Using Jupyter Notebook, we imported a number of libraries including pandas, requests, matplotlib.pyplot, and seaborn, created connections with formatted API calls, returned the data in JSON format, and created dataframes. Our Jupyter Notebook files with specific ETL process text blocks and code can be located [in this project’s GitHub](https://github.com/Jakob-T13/module-8-assessment).

**Company Summary**

1. Import required Python libraries: json, Pandas (as pd), MatPlotLib (matplotlib.pyplot as plt), Requests, NumPy (as np), and Seaborn (as sns)
2. Obtain an API key from the Census website
3. Build the URL requesting the needed information from the API:

[https://api.census.gov/data/2018/abscs?get=GEO\_ID,NAME,NAICS2017,NAICS2017\_LABEL,EMPSZFI,EMPSZFI\_LABEL,FIRMPDEMP,RCPPDEMP,EMP,PAYANN&for=state:\*&key=YOUR\_KEY\_HERE](https://api.census.gov/data/2018/abscs?get=GEO_ID,NAME,NAICS2017,NAICS2017_LABEL,EMPSZFI,EMPSZFI_LABEL,FIRMPDEMP,RCPPDEMP,EMP,PAYANN&for=state:*&key=YOUR_KEY_HERE)

1. Use requests.get() to acquire the data from the API
2. Dump the response into a JSON object (interpreted as a list of lists)
3. Convert the JSON object into a Pandas dataframe with pd.DataFrame()
4. Within the dataframe, convert the numeric data to their actual numeric values (they’re strings by default) using astype(int).
5. Trim out the GEO\_ID, NAICS2017, and EMPSZFI columns, as they don’t have any meaningful data and were only included in the API call to ensure all the data would be present
6. Filter out rows containing data values of the sum total per category indicated in the NAICS2017\_LABEL and EMPSZFI\_LABEL columns
7. Create dataframes of subsets of the data, as needed for the visualizations:

Number of Employees vs. Annual Payroll

* 1. Create a dataframe using the EMP and PAYANN columns of the main dataframe
  2. Plot the two columns against each other, using MatPlotLib to create the graphic and Seaborn to format it

NAICS Industry vs. Average Annual Payroll

1. Break out NAICS2017\_LABEL and PAYANN into a dataframe
2. Aggregate the PAYANN values by NAICS2017\_LABEL using groupby().aggregate()
3. Plot the grouped aggregates using MatPlotLib and Seaborn on a bar chart.

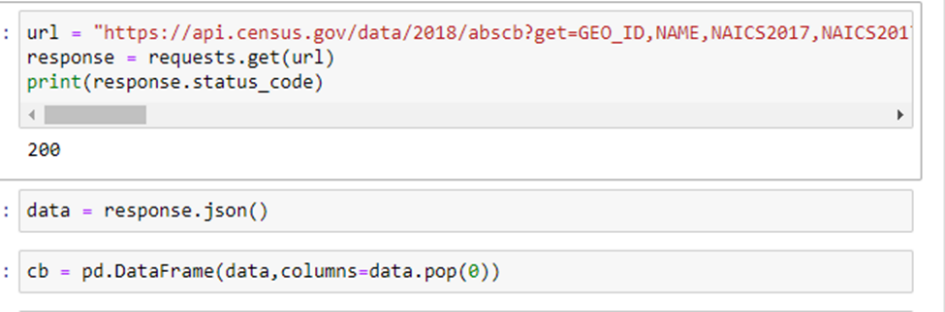
Firm Size vs. Average Annual Payroll for Healthcare

* 1. Break out NAICS2017\_LABEL, EMPSZFI\_LABEL, and PAYANN into a dataframe
  2. Filter the dataframe such that EMPSZFI\_LABEL is only ‘Health care and social assistance’
  3. Remove the NAICS2017\_LABEL column
  4. Aggregate the PAYANN values by EMPSZFI\_LABEL using the same technique as before
  5. Plot the grouped aggregates using MatPlotLib and Seaborn.

**Characteristics of Businesses Transformation**

1. Import Pandas (import pandas as pd), Import matplotlib (import matplotlib.pyplot as plt), Import Requests (import requests), Import seaborn (import seaborn as sns)
2. Obtain a key from the census website
3. With the key, request the characteristics of business census API and make it into a data frame
   1. "https://api.census.gov/data/2018/abscb?get=GEO\_ID,NAME,NAICS2017,NAICS2017\_LABEL,SEX,SEX\_LABEL,ETH\_GROUP,ETH\_GROUP\_LABEL,RACE\_GROUP,RACE\_GROUP\_LABEL,VET\_GROUP,VET\_GROUP\_LABEL,QDESC,QDESC\_LABEL,BUSCHAR,BUSCHAR\_LABEL,YEAR,FIRMPDEMP,FIRMPDEMP\_F,FIRMPDEMP\_PCT,FIRMPDEMP\_PCT\_F,RCPPDEMP,RCPPDEMP\_F,RCPPDEMP\_PCT,RCPPDEMP\_PCT\_F,EMP,EMP\_F,EMP\_PCT,EMP\_PCT\_F,PAYANN,PAYANN\_F,PAYANN\_PCT,PAYANN\_PCT\_F,FIRMPDEMP\_S,FIRMPDEMP\_S\_F,FIRMPDEMP\_PCT\_S,FIRMPDEMP\_PCT\_S\_F,RCPPDEMP\_S,RCPPDEMP\_S\_F,RCPPDEMP\_PCT\_S,RCPPDEMP\_PCT\_S\_F,EMP\_S,EMP\_S\_F,EMP\_PCT\_S,EMP\_PCT\_S\_F,PAYANN\_S,PAYANN\_S\_F,PAYANN\_PCT\_S,PAYANN\_PCT\_S\_F&for=us:\*&QDESC\_LABEL=SPOUSES

Ex code:



1. For the first question about Business Characteristics, create a new data frame(cbq1) and set it equal to the first data frame (cb in the example above)
2. Change the numemployees column in cbq1 to a float data type
3. Then filter cbq1 to only have ETHNIC\_GROUP\_LABEL and EMP as columns
4. Remove rows that have Equally Hispanic/non-Hispanic, Unclassifiable, and Classifiable for ethnic groups
5. Make the data frame group by ETHNIC\_GROUP\_LABEL count.
6. Create a bar graph using ax=[dataframe\_name].plot(title = 'Number of Employees that are Hispanic and non-Hispanic',kind='bar')
7. Add labels for the x axis label, y axis label, title, column values[ax.bar\_label(ax.containers[0], label\_type='edge', fmt='%d')], and rotate the x axis labels horizontally.
8. For the second question create a new data frame(cbq2) and set it equal to cb
9. Filter cbq2 to have only contain NAICS2017 and EMP columns
10. Remove rows that have ‘Total for all sectors’ and ‘Industries not classified’ as NAICS2017
11. Use a groupby for the count of NAICS2017
12. Create a horizontal bar graph
    1. ex: ax2 = ree2.plot(title = 'NUMBER OF EMPLOYEES BY NAICS',kind='barh')
13. Add x and y axis labels, titles, and make the x axis labels horizontally
14. For the third question using the same key before, make a data frame from the characteristics of business owners census data
    1. [https://api.census.gov/data/2018/abscbo?get=GEO\_ID,NAME,NAICS2017,NAICS2017\_LABEL,OWNER\_SEX,OWNER\_SEX\_LABEL,OWNER\_ETH,OWNER\_ETH\_LABEL,OWNER\_RACE,OWNER\_RACE\_LABEL,OWNER\_VET,OWNER\_VET\_LABEL,QDESC,QDESC\_LABEL,OWNCHAR,OWNCHAR\_LABEL,YEAR,OWNPDEMP,OWNPDEMP\_F,OWNPDEMP\_PCT,OWNPDEMP\_PCT\_F,OWNPDEMP\_S,OWNPDEMP\_S\_F,OWNPDEMP\_PCT\_S,OWNPDEMP\_PCT\_S\_F&for=us:\*&QDESC\_LABEL=YRACQBUS&key=](https://api.census.gov/data/2018/abscbo?get=GEO_ID,NAME,NAICS2017,NAICS2017_LABEL,OWNER_SEX,OWNER_SEX_LABEL,OWNER_ETH,OWNER_ETH_LABEL,OWNER_RACE,OWNER_RACE_LABEL,OWNER_VET,OWNER_VET_LABEL,QDESC,QDESC_LABEL,OWNCHAR,OWNCHAR_LABEL,YEAR,OWNPDEMP,OWNPDEMP_F,OWNPDEMP_PCT,OWNPDEMP_PCT_F,OWNPDEMP_S,OWNPDEMP_S_F,OWNPDEMP_PCT_S,OWNPDEMP_PCT_S_F&for=us:*&QDESC_LABEL=YRACQBUS&key=)
15. Merge the data frames from both websites together and call it ‘merged’
16. Change EMP and OWNPEMP to float types
17. Make a scatter plot with OWNPEMP on the x-axis and EMP on the y-axis
18. To investigate the number of employees working for Family-Owned firms by owner’s race, execute the following.

With your key, request the Characteristics of Business API with the following modified url, and put the results into a data-frame.

Use: '[https://api.census.gov/data/2018/abscb?get=GEO\_ID,NAME,NAICS2017,NAICS2017\_LABEL,SEX,SEX\_LABEL,ETH\_GROUP,YIBSZFI,YIBSZFI\_LABEL,ETH\_GROUP\_LABEL,RACE\_GROUP,RACE\_GROUP\_LABEL,VET\_GROUP,VET\_GROUP\_LABEL,QDESC,QDESC\_LABEL,BUSCHAR,BUSCHAR\_LABEL,YEAR,FIRMPDEMP,FIRMPDEMP\_F,FIRMPDEMP\_PCT,FIRMPDEMP\_PCT\_F,RCPPDEMP\_PCT\_F,EMP,EMP\_F,EMP\_PCT,EMP\_PCT\_F,PAYANN,PAYANN\_F,PAYANN\_PCT,PAYANN\_PCT\_F,FIRMPDEMP\_S,FIRMPDEMP\_S\_F,FIRMPDEMP\_PCT\_S,FIRMPDEMP\_PCT\_S\_F,RCPPDEMP\_S,RCPPDEMP\_S\_F,RCPPDEMP\_PCT\_S,RCPPDEMP\_PCT\_S\_F,EMP\_S,EMP\_S\_F,EMP\_PCT\_S,EMP\_PCT\_S\_F,EMPSZFI,EMPSZFI\_LABEL,PAYANN\_S,PAYANN\_S\_F,PAYANN\_PCT\_S,PAYANN\_PCT\_S\_F&for=us:\*&QDESC\_LABEL=FAMOWN&key=](https://api.census.gov/data/2018/abscb?get=GEO_ID,NAME,NAICS2017,NAICS2017_LABEL,SEX,SEX_LABEL,ETH_GROUP,YIBSZFI,YIBSZFI_LABEL,ETH_GROUP_LABEL,RACE_GROUP,RACE_GROUP_LABEL,VET_GROUP,VET_GROUP_LABEL,QDESC,QDESC_LABEL,BUSCHAR,BUSCHAR_LABEL,YEAR,FIRMPDEMP,FIRMPDEMP_F,FIRMPDEMP_PCT,FIRMPDEMP_PCT_F,RCPPDEMP_PCT_F,EMP,EMP_F,EMP_PCT,EMP_PCT_F,PAYANN,PAYANN_F,PAYANN_PCT,PAYANN_PCT_F,FIRMPDEMP_S,FIRMPDEMP_S_F,FIRMPDEMP_PCT_S,FIRMPDEMP_PCT_S_F,RCPPDEMP_S,RCPPDEMP_S_F,RCPPDEMP_PCT_S,RCPPDEMP_PCT_S_F,EMP_S,EMP_S_F,EMP_PCT_S,EMP_PCT_S_F,EMPSZFI,EMPSZFI_LABEL,PAYANN_S,PAYANN_S_F,PAYANN_PCT_S,PAYANN_PCT_S_F&for=us:*&QDESC_LABEL=FAMOWN&key=) + your key.

1. Within the data-frame, filter ‘ETH\_GROUP\_LABEL’, ‘NAICS2017\_LABEL’, ‘SEX\_LABEL’, ‘VET\_GROUP\_LABEL’, ‘EMPSZFI\_LABEL’, and ‘YIBSZFI\_LABEL’ so the only instances remaining are those that represent the totals for those columns.
2. Filter the data-frame so that the only columns remaining are ‘BUSCHAR’,’RACE\_GROUP\_LABEL’, and ‘EMP’.
3. Filter ‘BUSCHAR’ to instances of the response ‘BM’.
4. Change the data-type of ‘EMP’ from string to float.
5. Filter ‘RACE\_GROUP\_LABEL’ so that the only instances remaining are ‘Black or African American’, ’Native Hawaiian and Other Pacific Islander’, ‘American Indian and Alaska Native’, ‘White’, and ‘Asian’.
6. Create a group-by. Group ‘EMP’ by ‘RACE\_GROUP\_LABEL’ and return the sum.



1. Create a data-frame pairing a list of the RACE\_GROUP\_LABEL instances from the group-by and a list of the numeric values from the group-by.



1. Plot this data-frame in a bar-chart.

**Characteristics of Business Owners**

\*See [Jupyter notebook](https://github.com/Jakob-T13/module-8-assessment.git) as ETL & formats differ per API call\*

**Extraction**

1. Import libraries: pandas as pd, requests, matplotlib.pyplot as plt, and seaborn as sns
2. Obtain key from [Census Bureau developers website](https://www.census.gov/data/developers/guidance/api-user-guide.Help_&_Contact_Us.html)
3. Check response code of API
4. Save API response in JSON format into a dataframe and use .pop(0) to create column headers with first row values

Graphical user interface, text, application

Description automatically generated

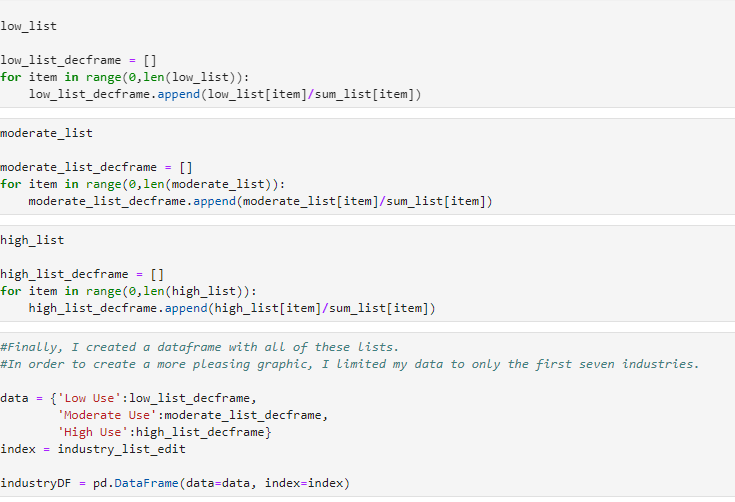
**Transformation**

1. Change data type for OWNPDEMP from object to float using .astype()
2. Drop ‘0’ values in column OWNPDEMP
3. Filter out columns with redundant information
4. Split NAME string to separate city from other location info, create new dataframe
5. Merge dataframes on NAME, rename columns for readability, filter out unnecessary columns
6. Sort values with .sort\_values() function
7. Use sns and plt to create visualization

**Technology Characteristics of Businesses**

1. With the key, request the Technology Characteristics of Business API and put the results into a data-frame.
2. Use: response = [https://api.census.gov/data/2018/abstcb?get=GEO\_ID,NAME,NAICS2017,NAICS2017\_LABEL,TECHUSE,TECHUSE\_LABEL,MOTPRODTECH,MOTPRODTECH\_LABEL,MOTUSETECH,MOTUSETECH\_LABEL,SEX,SEX\_LABEL,ETH\_GROUP,ETH\_GROUP\_LABEL,RACE\_GROUP,RACE\_GROUP\_LABEL,VET\_GROUP,VET\_GROUP\_LABEL,NSFSZFI,NSFSZFI\_LABEL,FACTORS\_P,FACTORS\_P\_LABEL,YEAR,FIRMPDEMP,FIRMPDEMP\_F,FIRMPDEMP\_PCT,FIRMPDEMP\_PCT\_F,RCPPDEMP,RCPPDEMP\_F,RCPPDEMP\_PCT,RCPPDEMP\_PCT\_F,EMP,EMP\_F,EMP\_PCT,EMP\_PCT\_F,RCPPDEMP\_S,RCPPDEMP\_S\_F,IMPACTWF\_P,IMPACTWF\_P\_LABEL,IMPACTWF\_U,IMPACTWF\_U\_LABEL&for=us:\*&key=](https://api.census.gov/data/2018/abstcb?get=GEO_ID,NAME,NAICS2017,NAICS2017_LABEL,TECHUSE,TECHUSE_LABEL,MOTPRODTECH,MOTPRODTECH_LABEL,MOTUSETECH,MOTUSETECH_LABEL,SEX,SEX_LABEL,ETH_GROUP,ETH_GROUP_LABEL,RACE_GROUP,RACE_GROUP_LABEL,VET_GROUP,VET_GROUP_LABEL,NSFSZFI,NSFSZFI_LABEL,FACTORS_P,FACTORS_P_LABEL,YEAR,FIRMPDEMP,FIRMPDEMP_F,FIRMPDEMP_PCT,FIRMPDEMP_PCT_F,RCPPDEMP,RCPPDEMP_F,RCPPDEMP_PCT,RCPPDEMP_PCT_F,EMP,EMP_F,EMP_PCT,EMP_PCT_F,RCPPDEMP_S,RCPPDEMP_S_F,IMPACTWF_P,IMPACTWF_P_LABEL,IMPACTWF_U,IMPACTWF_U_LABEL&for=us:*&key=) + your key.
3. Clean the IMPACTWF\_P\_LABEL. Set the response of
   1. ‘Cloud-Based: Increased skill level of workers employed by this business’ to 'Cloud-Based: Increased STEM skills of workers employed by this business'
   2. 'Cloud-Based: Decreased skill level of workers employed by this business' to 'Cloud-Based: Decreased STEM skills of workers employed by this business'
   3. 'Cloud-Based: Did not change skill level of workers employed by this business' to 'Cloud-Based: Did not change STEM skills of workers employed by this business'
   4. 'Specialized Software: Increased skill level of workers employed by this business' to 'Specialized Software: Increased STEM skills of workers employed by this business'
   5. 'Specialized Software: Decreased skill level of workers employed by this business' to 'Specialized Software: Decreased STEM skills of workers employed by this business'
   6. 'Specialized Software: Did not change skill level of workers employed by this business' to 'Specialized Software: Did not change STEM skills of workers employed by this business'
   7. 'Robotics: Increased skill level of workers employed by this business' to 'Robotics: Increased STEM skills of workers employed by this business'
   8. 'Robotics: Decreased skill level of workers employed by this business' to 'Robotics: Decreased STEM skills of workers employed by this business'
   9. 'Robotics: Did not change skill level of workers employed by this business' to 'Robotics: Did not change STEM skills of workers employed by this business'
   10. 'Specialized Equipment: Increased skill level of workers employed by this business' to 'Specialized Equipment: Increased STEM skills of workers employed by this business'
   11. 'Specialized Equipment: Decreased skill level of workers employed by this business' to 'Specialized Equipment: Decreased STEM skills of workers employed by this business'
   12. 'Specialized Equipment: Did not change skill level of workers employed by this business' to 'Specialized Equipment: Did not change STEM skills of workers employed by this business'
   13. 'Artificial Intelligence: Increased skill level of workers employed by this business' to 'Artificial Intelligence: Increased STEM skills of workers employed by this business'
   14. 'Artificial Intelligence: Decreased skill level of workers employed by this business' to 'Artificial Intelligence: Decreased STEM skills of workers employed by this business'
   15. 'Artificial Intelligence: Did not change skill level of workers employed by this business' to 'Artificial Intelligence: Did not change STEM skills of workers employed by this business'
   16. 'Artificial Intelligence: Total use' to 'Artificial Intelligence: Total Reporting'
   17. 'Robotics: Total use' to 'Robotics: Total Reporting'
   18. 'Cloud-Based: Total use' to 'Cloud-Based: Total Reporting'
   19. 'Specialized Equipment: Total use' to 'Specialized Equipment: Total Reporting'
   20. 'Specialized Software: Total use' to 'Specialized Software: Total Reporting'
   21. Keeping both sets of categories is redundant and would likely cause issues of overlap.
4. Change the datatype of RCCDEMP to ‘float’.
5. Change the datatype of FIRMPDEMP to ‘float’.
6. In order to investigate Cloud-Technologies without overlap, create a subset of the data were all ‘LABEL’ columns that aren’t ‘TECHUSE\_LABEL’ are set to the categorical instance that is their ‘total’ values. These columns are ETH\_GROUP\_LABEL, IMPACTWF\_P\_LABEL, IMPACTWF\_U\_LABEL, MOTPRODTECH\_LABEL, MOTUSETECH\_LABEL, NAICS2017\_LABEL, NSFSZI\_LABEL, RACE\_GROUP\_LABEL, SEX\_LABEL, and VET\_GROUP\_LABEL
7. Within the subset, filter to the columns ‘TECHUSE\_LABEL’, ‘RCPPDEMP’, and ‘TECHUSE’.
8. Filter the TECHUSE\_LABEL columns to be only ‘Cloud-Based: Did not use’, ‘Cloud-Based: Low use’, ‘Cloud-Based: Moderate Use’, ‘Cloud-Based: High use’ and ‘Cloud-Based: Don’t know’.
9. Rename these values to ‘Did not use’,’Low use’,’Moderate use’,’High use’, and ‘Don’t know’.
10. Create a group-by for this subset. Group ‘RCPPDEMP’ by ‘TECHUSE\_LABEL’ and get the mean.
11. Create a bar-plot using this group-by.
12. To investigate the ratio of Specialized-Software use in firms by industry, do the following.
13. Create a subset of the data where all ‘LABEL’ variables are filtered to their totals aside from ‘NAICS2017\_LABEL’ and ‘TECHUSE\_LABEL’.
14. Filter the subset to the columns of ‘NAICS2017\_LABEL’, ‘TECHUSE\_LABEL’ and ‘FIRMPDEMP’.
15. Within the ‘TECHUSE\_LABEL’ column replace all instances of 'Specialized Software: Tested, but did not use in production or service' with 'Specialized Software: Low use'.
16. Filter the ‘TECHUSE\_LABEL to 'Specialized Software: Low use', 'Specialized Software: Moderate use' and 'Specialized Software: High use'.
17. Filter out all instances of 'Total for all sectors' and 'Industries not classified' from the ‘NAICS\_2017’ column.
18. Create four new subsets by filtering on ‘NAICS2017\_LABEL’ to 'Specialized Software: Low use', 'Specialized Software: Moderate use', 'Specialized Software: High use', and one that contains all three instances.
19. Create four lists that group the FIRMPDEMP values by the NAICS2017\_LABEL column using each subset. This should result in four lists of eighteen numbers.
20. For the three lists that contain the counts of firms by Specialized-Software level, create three new lists that are those values divided by those from the list that is the summation of all three instances.
21. Load these lists into a data-frame along with a list of their respective industry names (NAICS2017\_LABEL values).





1. Create a stacked bar plot using this data-frame.