Group # 2: Ahad, Derek, Isaac, Jeannine

Date: April 25, 2022

ETL Report

Introduction

The problem that we are trying to solve is to use the Census API data to make observations based on each API. The sources of data include characteristics of business, characteristics of business owners, technology characteristics of business, and company summary datasets. The data first needed to be filtered based on the questions and observations we wished to explore. Afterward, the data from each API needed to be further cleaned as it contained many unnecessary columns and rows.

Data Sources

United States Census Bureau. (October 28, 2021). Annual Business Survey (ABS) APIs. https://www.census.gov/data/developers/data-sets/abs.2019.html (accessed April 21, 2022).

U.S. Census Bureau (2021). *Company Summary.* https://api.census.gov/data/2018/abscs.html (accessed April 24, 2022).

U.S. Census Bureau (2021). *Characteristics of Businesses.* https://api.census.gov/data/2018/abscb.html (accessed April 24, 2022).

U.S. Census Bureau (2021). *Characteristics of Business Owners.* https://api.census.gov/data/2018/abscbo.html (accessed April 22, 2022).

U.S. Census Bureau (2021). *Technology Characteristics of Businesses.* https://www.census.gov/data/developers/data-sets/abs.2019.html (accessed April 24, 2022).

Company Summary Survey:

The Annual Business Company Summary Survey (2018) 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 geographic areas of the entire continental United States, individual US states, and metropolitan/micropolitan areas.

Initial Questions**:**

1. What are the largest industries in the USA based on the number of employees?
2. What states have the largest number of employer firms?
3. What is the distribution of firms based on years of operation?

Extraction:

1. Obtain a Census API key from https://api.census.gov/data/key_signup.html
2. Import pandas as pd
3. You must decompose the API Call for the **Annual Business Survey Company Summary** into its constituent parts and save them as variables:
4. Create a variable called *URL* and save it with the following formatted string:
    a. `URL = f'https://api.census.gov/data/2018/abscs?get={labels}{data_group}{API_key}`
    b. *labels* corresponds to the variables that we want to request
    c. *data_group* corresponds to the geographical area of interest (continental US, individual US States, and metropolitan areas)
    d. The *API_Key* is the API key code that you receive after successfully signing up for a US Census API key
5. Create the variables *labels*, *data_group*, and *API_key* and put them on the lines BEFORE your *URL* variable

6. Top 10 Industries by Number of Employees Horizontal Bar Chart:
    a. `labels = 'EMP,NAICS2017,NAICS2017_LABEL'`
    b. `data_group = '&for=state`
    c. `API_key = '&key=YOUR_API_KEY_CODE'`

7. Top 10 Industries by Number of Employees Horizontal Bar Chart:States by Number of Businesses 2019 Tree Map:

a. labels = 'NAME,FIRMPDEMP'
b. data_group = '&for=state'
c. API_key = '&key=YOUR_API_KEY_CODE'

8. Firms in US by Years in Business 2019 Plot:
    a. labels = 'EMP,NAICS2017,NAICS2017_LABEL'
    b. data_group = '&for=state
    c. API_key = '&key=YOUR_API_KEY_CODE'

9. Read the URL variable into a pandas dataframe:
    a. df = **pd.read_json**(URL)

Transformation*:*

1. You are going to replace the header row, with the names of the first column
2. Using *iloc*, save the first row as a variable named *new_header*
3. Set your dataframe to use the data from row 1 and beyond (df = df[1:])
4. Set df.columns = new_header
5. Top 10 Industries by Number of Employees Horizontal Bar Chart:
    a. Rename *NAICS2017_LABEL* column to *US Industry* and the *EMP* column to *Number of Employees*
    b. Remove any rows containing a *NAICS2017* value of 00
    c. Using a lambda function and the *apply()* function, replace any instances of the word “and” with “&” in the *US Industry* column
    d. Change the *Number of Employees* column to type *int*
    e. Using *groupby*, group the *US industry* column by the sum of the *Number of Employees* column and sort the values by ascending and save to the dataframe
    f. Using a lambda function and the *apply()* function, divide each element in the *Number of Employees* column by 1,000,000, get the last 10 values, and save to the dataframe
    g. Import matplotlib.pyplot as plt
    h. Plot to a horizontal bar chart using matplotlib

6. Top 10 States by Number of Businesses 2019 Tree Map:
    a. Rename *Name* column to *State* and the *FIRMPDEMP* column to *Number of Employer Firms*
    b. Change the *Number of Employer Firms* column to type *int*
    c. Import matplotlib and import squarify
    d. Use a normalization function to get the colors for your Treemap (https://python-graph-gallery.com/202-treemap-with-colors-mapped-on-values)
    e. Use a zip function to get the state and its value for the tree map labels (https://stackoverflow.com/questions/66897045/show-multiple-columns-values-on-labels-with-squarify-plot)
    f. Plot the top 10 values by state using squarify.plot

7. Firms in US by Years in Business 2019:
    a. Rename *YIBSZFI_LABEL* column to *Years in Business* and the *FIRMPDEMP* column to *Number of Employer Firms*
    b. Remove any rows containing a *Years in Business* value of *All firms*
    c. Change the *Number of Employees* column to type *int*
    d. Using *groupby*, group the *Years in Business* column by the sum of the *Number of Employer Firms* column and sort the values by ascending and save to the dataframe
    e. Use df = df.reset_index() to regenerate the data frame from the *groupby* object
    f. Using a lambda function and the *apply()* function, divide each value in *Number of Employer Firms* column by 1,000,000
    g. Import seaborn as sns
    h. Plot the horizontal bar chart using seaborn

Loading:

a. Merge all of the tables into one master table using two Left joins
b. Save master table to csv
c. Using the SQL Import Wizard, import the csv into a SQL database

Characteristics of Business Survey:

The Characteristics of Business Survey provided estimates of different characteristics of businesses. These characteristics included:

- Number of owners in 2018
- Whether the business was family-owned in 2018
- Whether the business was jointly-owned or operated by spouses in 2018
- Types of customers of the business in 2018
- Types of workers used by the business in 2018
- Reasons that the business ceased operations since 2018

Initial Questions:

1. Are there more family owned businesses or non family owned businesses?
2. What is the operating status of the business?
3. What types of employees are there and which are the most popular?

Extraction:

The data was found in the Annual Business Survey (ABS) APIs provided in the LMS. To get the data, a key must first be obtained. This can be done by going to https://api.census.gov/data/key_signup.html and providing the information asked. Then we create a variable called “labels” which will be the columns that we want to include in the table. We create a variable called “data_group” which lets us use data for the entire United States or separate states. We create another variable called “description” which will be what characteristic we are interested in. Finally, create a variable called “API_key” which will just be the key that we obtained from earlier. We then create a dataframe using all these variables that we created

```
df = pd.read_json(f'https://api.census.gov/data/2018/abscb?get={labels}{data_group}{description}{API_key}')
```

Transformation:

Family Owned vs Not Family Owned

1. Filter the dataset by setting the QDESC_LABEL equal to “FAMOWN” within the description variable.
2. Remove all rows that are not either “Family-owned” or “Not family-owned”.
3. Rename the “FIRMPDEMP” column as “Number of Firms” and “BUSCHAR_LABEL” column as “Business Characteristic”.
4. Remove all columns that are not either “Number of Firms” or “Business Characteristic”.
5. Plot bar chart with “Business Characteristic” on the x axis and “Number of Firms” on the y axis.

Cease Ops:

1. Filter the dataset by setting the QDESC_LABEL equal to “CEASEOPS” within the description variable.
2. Remove all rows that are not either “Business is currently operating”, “Business operations have ceased”, or Operating status not yet reported.
3. Rename the “FIRMPDEMP” column as “Number of Firms” and “BUSCHAR_LABEL” column as “Business Characteristic”.
4. Remove all columns that are not either “Number of Firms” or “Business Characteristic”.
5. Plot horizontal bar chart with “Business Characteristic” on the x axis and “Number of Firms” on the y axis.

Types of Workers:

1. Filter the dataset by setting the QDESC_LABEL equal to “WORKERS” within the description variable.
2. Remove the rows that are “None of the above”, “Item not reported”, “Total reporting”, and “All firms”.
3. Rename the “EMP” column as “Number of Employees” and “BUSCHAR_LABEL” column as “Type of Worker”.
4. Remove all columns that are not either “Number of Employees” or “Type of Worker”.
5. Plot horizontal bar chart with “Number of Employees” on the x axis and “Type of Worker” on the y axis.

Loading:

1. Download database
2. Run SQL express in docker

3. Create a database in Azure Data Studio using DDL
4. Create the necessary tables
5. Right click database and select import wizard
6. Fill out details and select file to be imported
7. Insert the data to populate the tables using Azure Data Studio's import extension
8. If changes need to be made, use update statement to update the rows
9. If anything needs to be removed, use delete statement to delete the rows.

Characteristics of Business Owners Survey

The Characteristics of Business Owners Survey focuses on the demographic breakdown of business owners in the United States. Here, we have chosen to examine the characteristics of race and sex.

Initial Questions:

1. How does the number of male vs. female business owners differ?
2. How does the race of business owners differ?
3. How does the sex of business owners vary across different races?

Extraction:

The data is obtained through the US Census Bureau via an API. A key to the API is available at https://api.census.gov/data/key_signup.html. Then we read in the data via pandas, choosing the relevant labels and data group, and create a data frame from the json file. We use the first row of the data set to create column headers, and sort the data frame by the OWNER_RACE_LABEL and OWNER_SEX_LABEL columns for easier review. See included Jupyter Notebook for further details.

Transformation:

First, we familiarize ourselves with the data by examining the data frame and noting its layout. Then we take a deeper dive into the data by calling df.info, df.OWNER_RACE_LABEL.value_counts(), and df.OWNER_SEX_LABEL.value_counts(). We group the data by sex and race and then use the query method to filter the data in order to

create visualizations of the number of business owners by sex, by race, and by both sex and race, using matplotlib and ggplot . Specifically, for the “Business Owners by Sex” chart, the data are filtered so that the OWNER_RACE_LABEL only contains “All owners of respondent firms.” Similarly, the “Business Owners by Race” chart is filtered so the OWNER_SEX_LABEL only contains “All owners of respondent firms.” The data are not filtered any further for the “Business Owners by Race and Sex” chart. See included Jupyter notebook for further details.

Loading

If we were to load this data into a SQL database, we would take the following steps:

Step 1: Create an ERD to map out the relationships between tables

Step 2: Create the database tables and columns using DDL

Step 3: Add the data (i.e. rows) to the database using DML

Step 4: Join the tables together as outlined in the ERD using JOIN statements

Technology Characteristics of Businesses

This dataset provides data on the use of technology for artificial intelligence, cloud-based computing, specialized software, robotics, and specialized equipment. It also breaks down factors adversely affecting the utilization of specific technologies. This ETL process is recorded in the Jupyter notebook derek-preslar-assessment-8.ipynb file.

Initial Questions:

1. What were the most common factors affecting AI utilization?
2. Does the frequency of adverse factors vary across the business owner’s race?

Extraction:

The data was extracted from an API call. After obtaining a key from the U.S. Census Bureau, the API was called and the data returned in a JSON format that pandas converted into a dataframe. After setting the labels that contain the relevant data, the data group to include the entire US, and the API key, the following code will make the API call and read the JSON response into a dataframe df.

```
df = pd.read_json(f'https://api.census.gov/data/2018/abstcb?get={labels}{data_group}{API_key}')
```

Transformation:

1. First, the columns need to be set to the first row values:

```
df.columns = df.iloc[0]
df = df[1:]
```

2. The data needs to be filtered so that it only includes the totals for sex, veteran status, ethnicity, and sector.
3. The columns used for the dataframe are narrowed down to: 'RACE_GROUP_LABEL','NAICS2017_LABEL','FIRMPDEMP','FACTORS_U_LABEL'
4. The data then needs to be separated into four different dataframes for the business owner's race as White, Black and African American, Asian, and American Indian and Alaska Native.
5. In order to plot the percentage of responses, the total responses from each group needs to be recorded.
6. The responses to Artificial Intelligence then need to be separated off into a separate list for each group and another list for recording the questions corresponding to each point.
    a. This is so that the data can be more easily plotted.
7. Combine all adverse factors to compare them with businesses that reported no issues or that AI was irrelevant for the business.

Loading:

In the loading stage, I would develop a single database containing responses to all technologies (after step 3 in the transformation section). From there, it could be queried to answer questions based on business owner race and response to technology utilization.

Survey Analysis & Diagrams:

Company Summary Survey:

Figure 1: Top 10 Industries by Number of Employees (2019)

Figure 2: Top 10 States by Number of Employer Firms (2019)

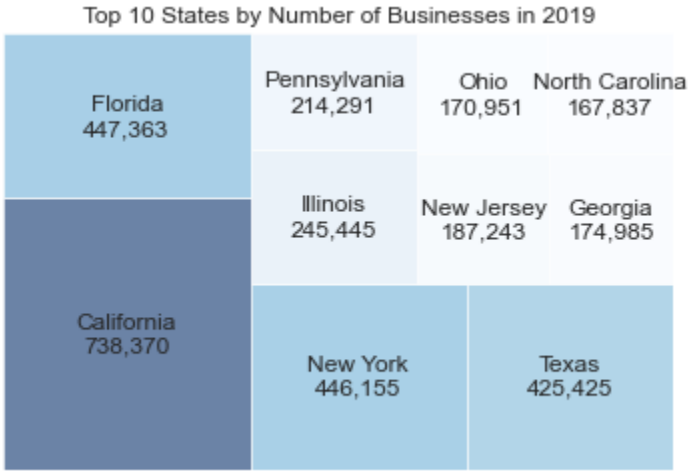
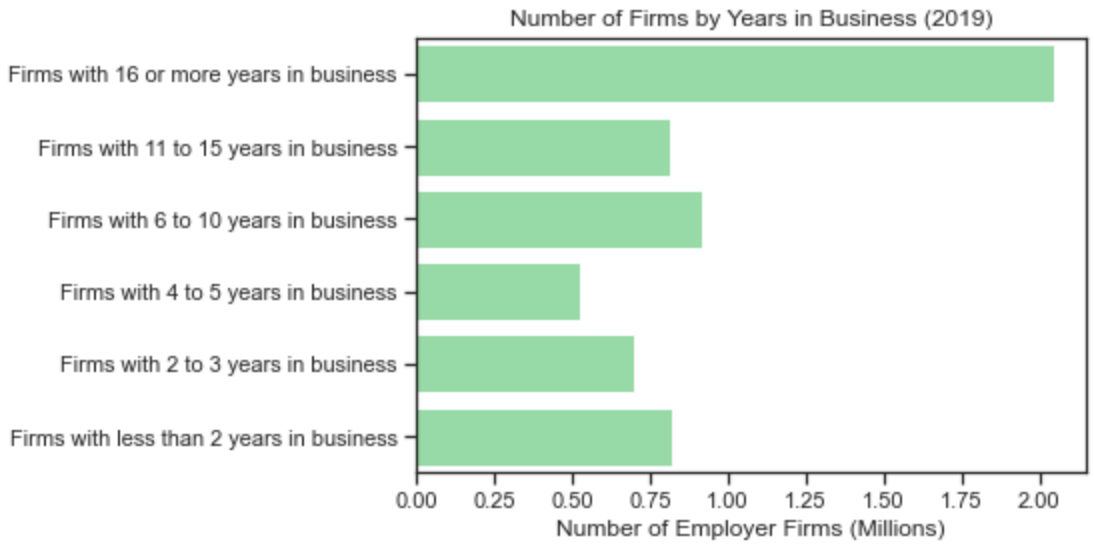


Figure 3: Firms by Years in Business (2019)



Initial Questions & Conclusions:

1. What are the largest industries in the USA based on the number of employees?
    - Based on **Figure 1**, the largest industries in the USA center around healthcare, retail, and hospitality and accommodations
    - This is not surprising as the US has largely shifted away from manufacturing as manufacturing jobs have been outsourced to overseas manufacturers in countries such as China, India, Pakistan, and Thailand
2. What states have the largest number of employer firms?
    - Based on **Figure 2**, California, Florida, New York, and Texas have the largest number of employer firms in the country
    - This should also not be surprising as this reflects demographic trends in population as the these are some of the most populous states in the country, so it figures that these states would also have larger amounts of businesses concentrated there
3. What is the distribution of firms based on years of operation?
    - Based on **Figure 3**, The number of firms with 16 or more years of operation dwarfs the other categories by a wide margin
    - This may result from the greater advantage that an incumbent business has over its competitors as well as the difficulty in maintaining a fledgling business in the

current economy as the category of firms with 4 to 5 years in business is by far the smallest

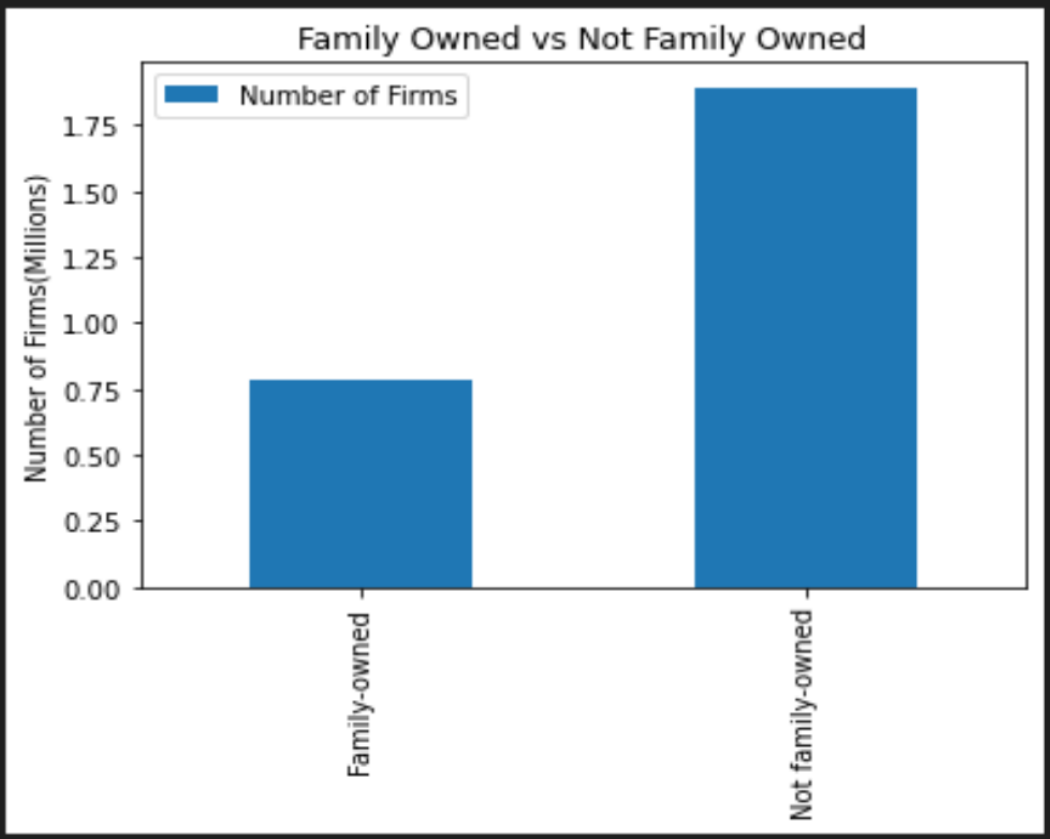
- In other words, plenty of people are starting businesses but are having difficulty maintaining them for long periods of time

Characteristics of Businesses Survey:

Initial question #1: Are there more family owned businesses or non family owned businesses?

Figure 1: Number of Family Owned Firms vs. Non-Family Owned Firms (2019)

	Business Characteristic	Number of Firms
2	Family-owned	0.787775
3	Not family-owned	1.893857

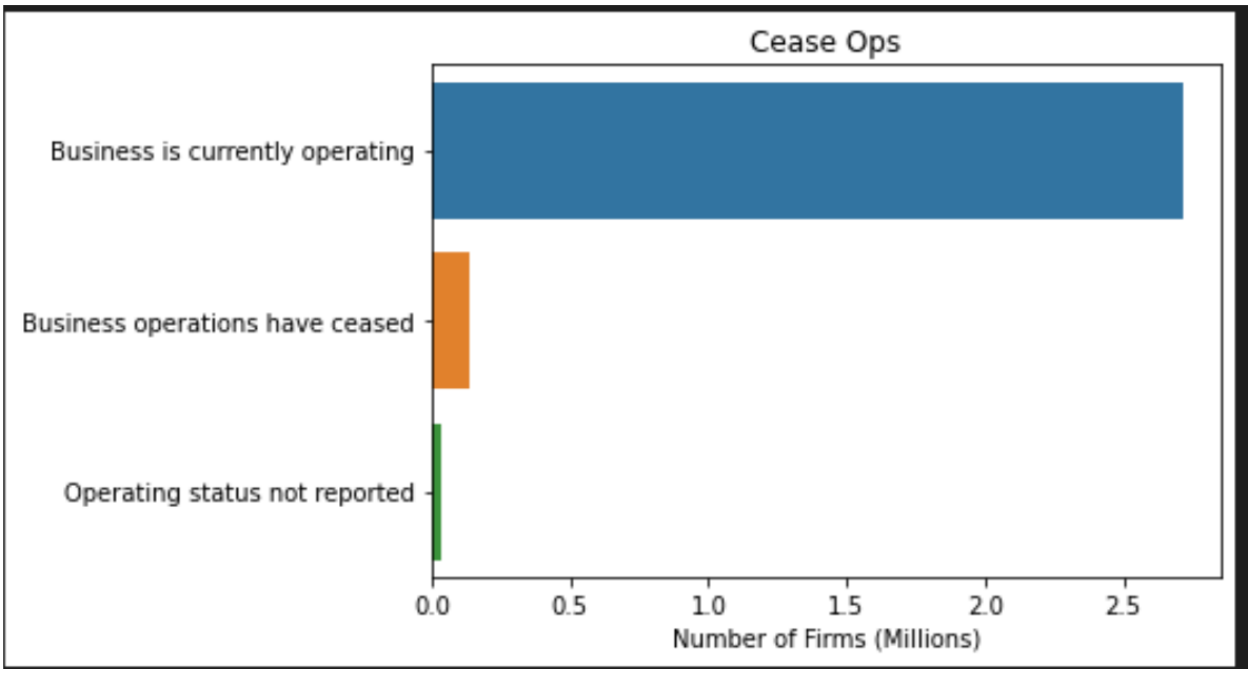


As one can see, there are over twice as many non-family owned businesses than family owned businesses. This data was pulled by filtering the QDESC_LABEL to “FAMOWN”. This filters the data to only how the businesses are owned. Since there are many challenges that come with

family owned businesses, it makes sense that there are a lot more non-family owned businesses.

Initial question #2: What is the operating status of the business?

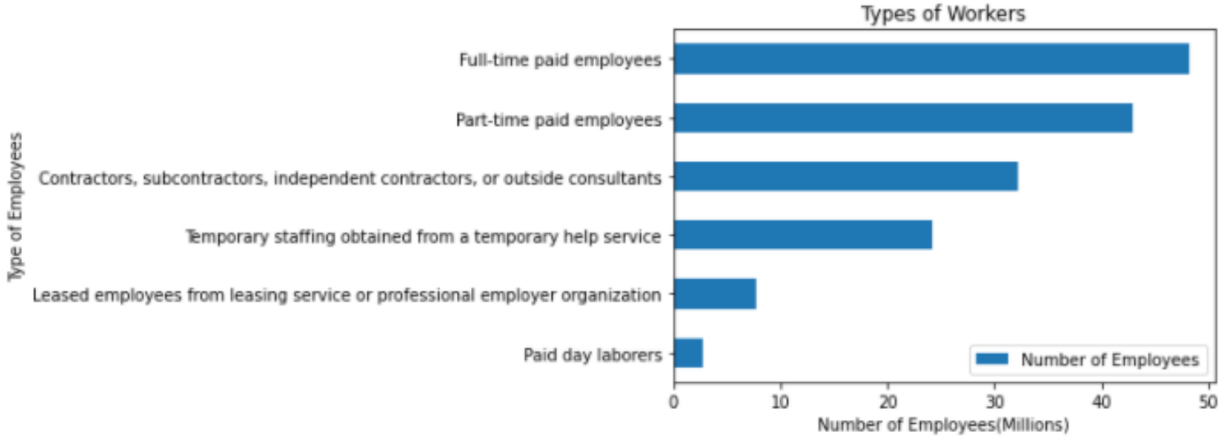
Figure 2: Operating Status of Businesses (2019)



The overwhelming number of businesses are currently operating. This data was pulled by filtering the QDESC_LABEL to “CEASEOPS”. This will show how many businesses are currently operating, how many have ceased, and how many have not been reported. This plot was made using seaborn.

Initial question #3: What types of employees are there and which are the most popular? Least popular?

Figure 3: Number of Employees by Employee Type



Based on this graph, full time paid employees are the most common type of worker with just under 50 million. The least popular type of worker are paid day laborers. This data was pulled by filtering the QDESC_LABEL to “WORKERS”. This will show the type of workers and how many are in each type.

Characteristics of Business Owners Survey:

Figure 1: Business Owners by Sex (2019)

4: All Owners

5: Female Owners

6: Male Owners

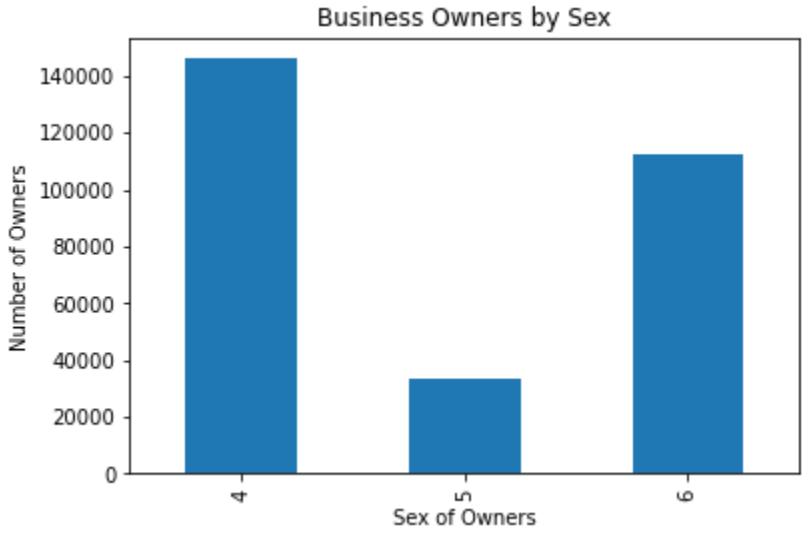


Figure 2: Business Owners by Race (2019)

```
3:              American Indian and Alaska Native
4:                 All owners of respondent firms
7:                                           White
10:                    Black or African American
13:                                          Asian
16:  Native Hawaiian and Other Pacific Islander
19:                                       Minority
22:                                    Nonminority
```

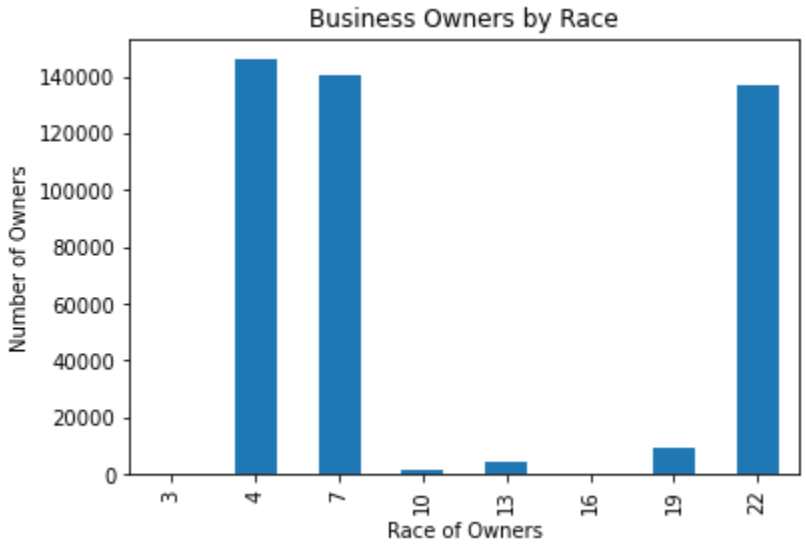
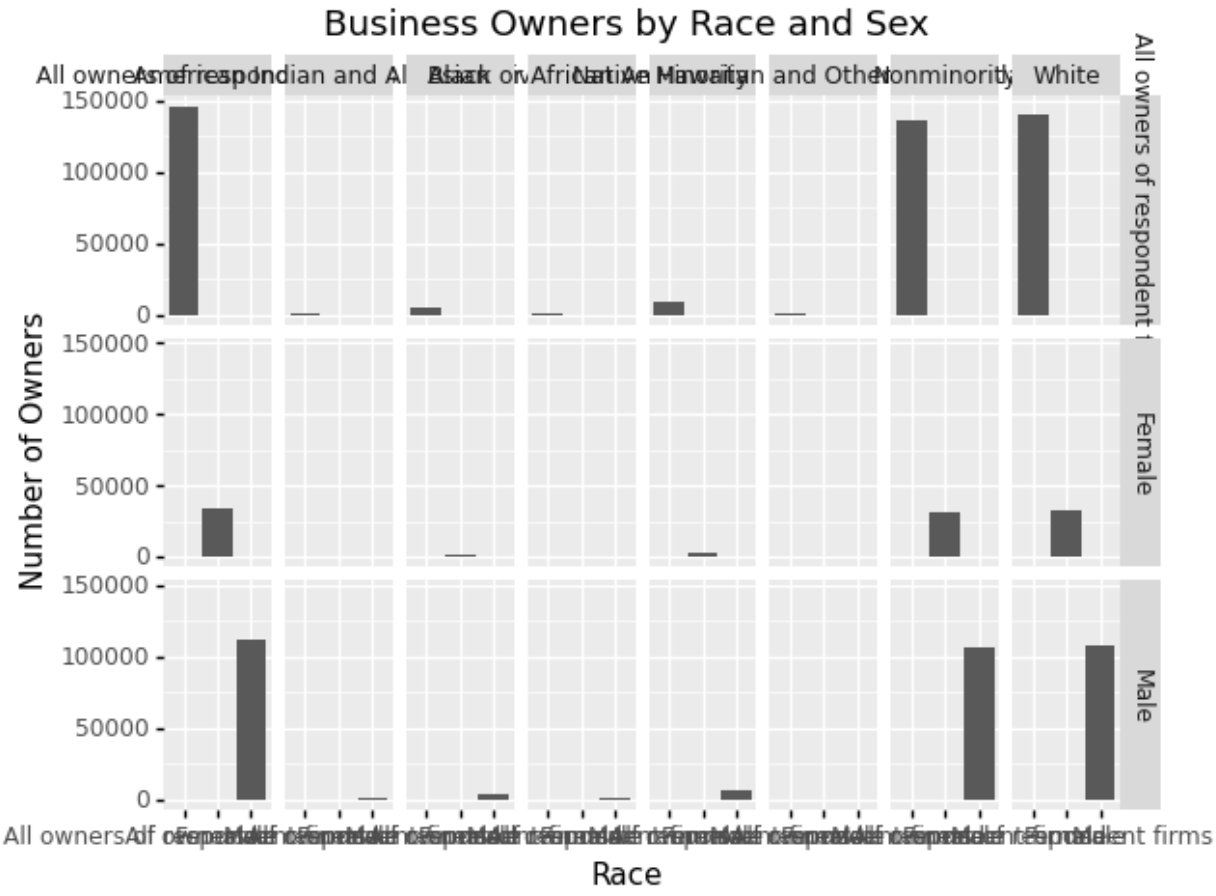


Figure 3: Business Owners by Race & Sex (2019)



Conclusions

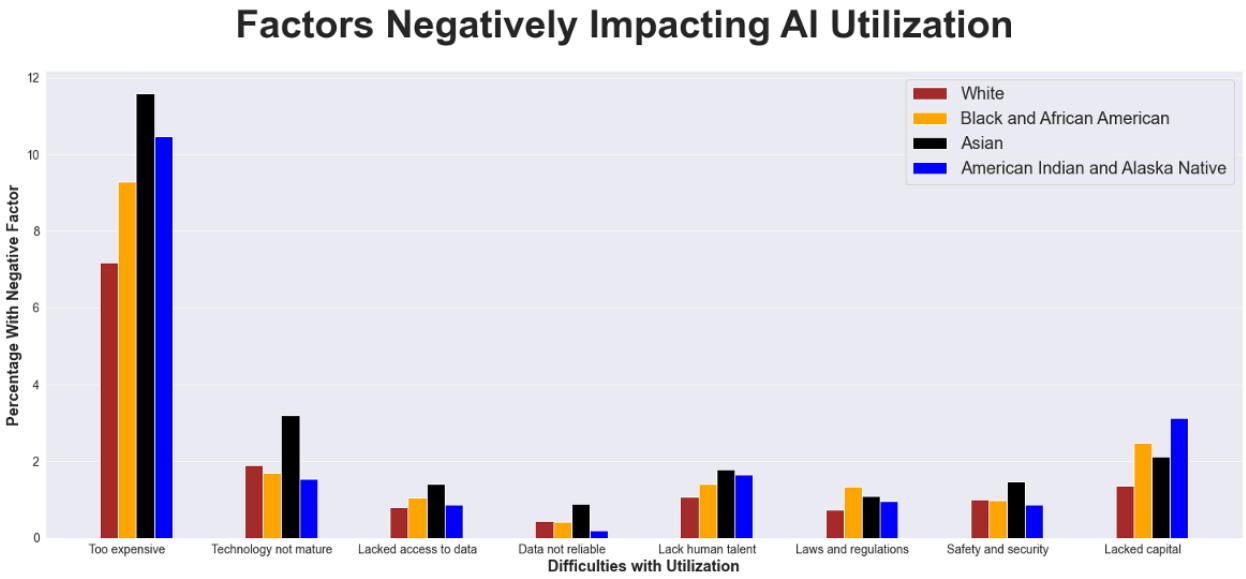
Most businesses in the United States are overwhelmingly owned by white males. In the interest of diversifying business ownership in the US, more study to ascertain the causes for this disparity is warranted. Additionally, study and funding for programs aimed at preparing minorities and women for business ownership is needed.

Technology Characteristics of Businesses:

Initial Questions:

1. Investigate the factors that adversely affect the utilization of artificial intelligence by businesses. What are the most common factors preventing the utilization of artificial intelligence?
2. Investigate any relations between those factors and the race of a business's owner. Are there any racial disparities in the utilization of artificial intelligence?

Figure 1: Factors Adversely affecting the Utilization of AI



As Figure 1 shows, the most common negative factor in the utilization of AI was the cost; this holds true regardless of a business owner's race. However, it does show that white business owners were less likely to experience cost or capital as a prohibiting factor. The following figures (Figures 2 to 6) will explore the relations between an owner's race and AI utilization.

Figure 2: AI Utilization Responses: Difficulties and Applicability

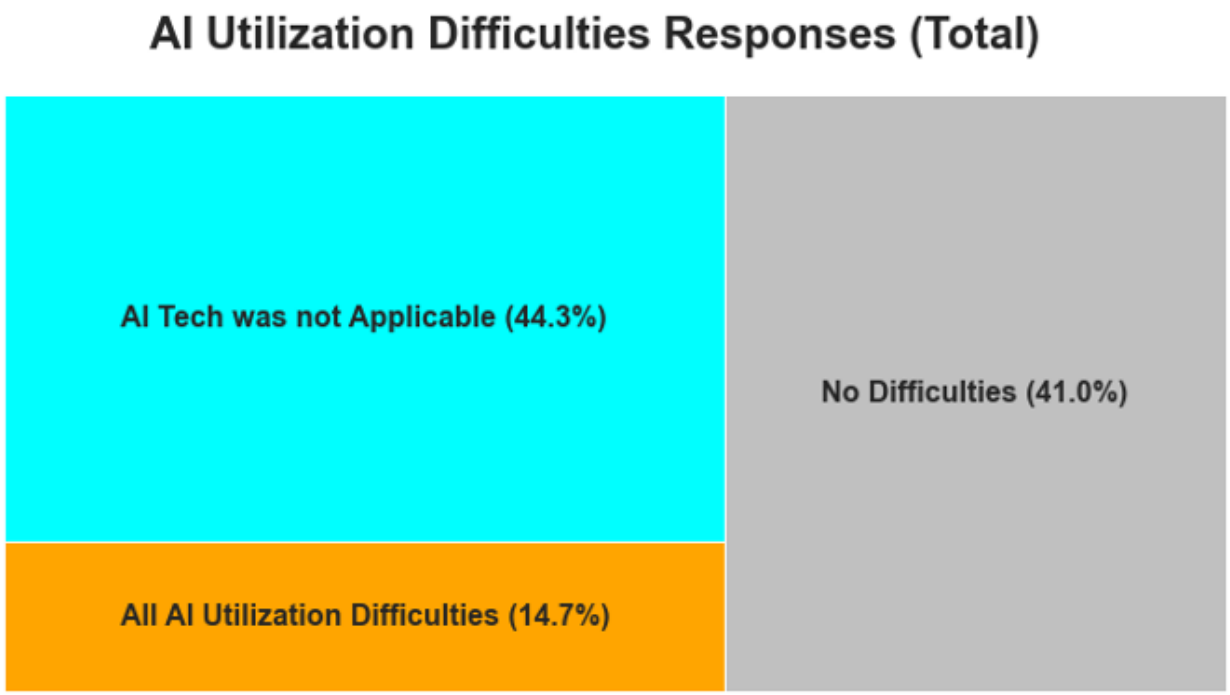


Figure 3: AI Utilization Responses (Asian Business Owner): Difficulties and Applicability

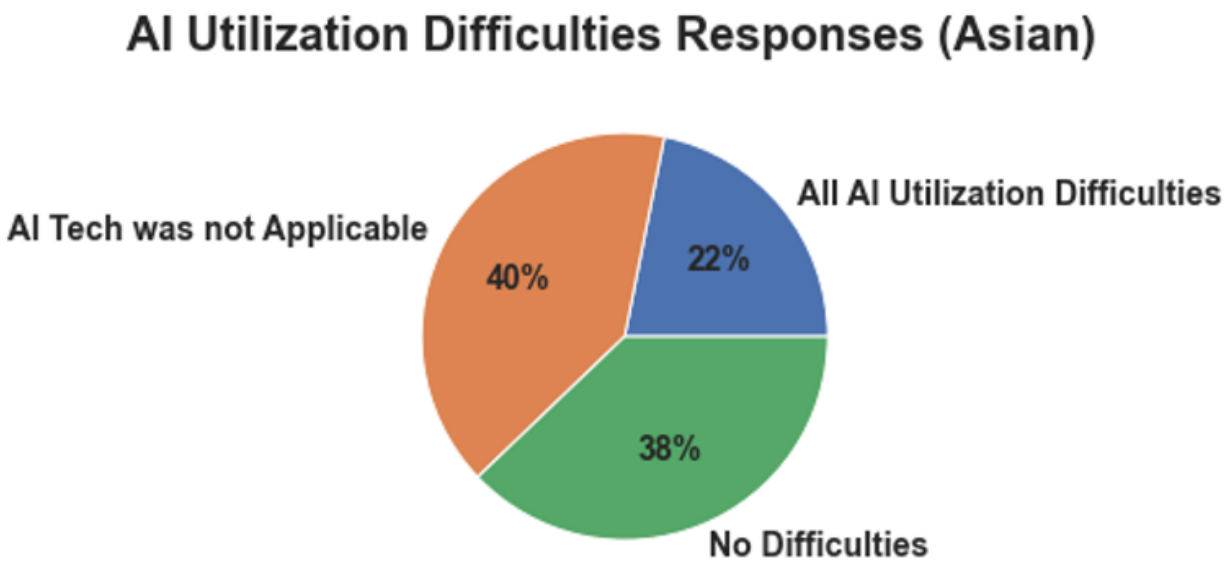


Figure 4: AI Utilization Responses (Black and African American Business Owner): Difficulties and Applicability

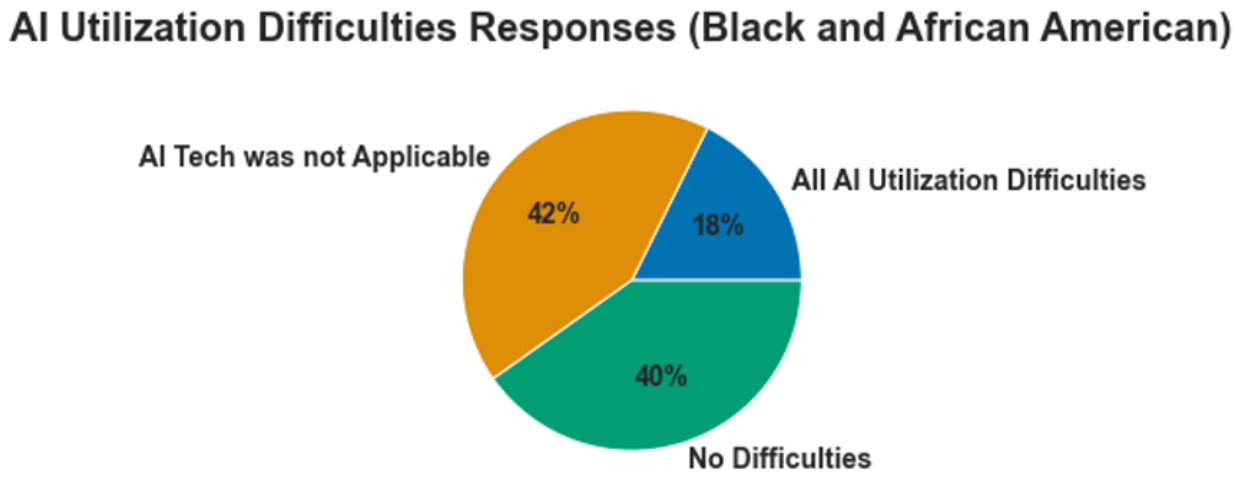


Figure 5: AI Utilization Responses (White Business Owner): Difficulties and Applicability

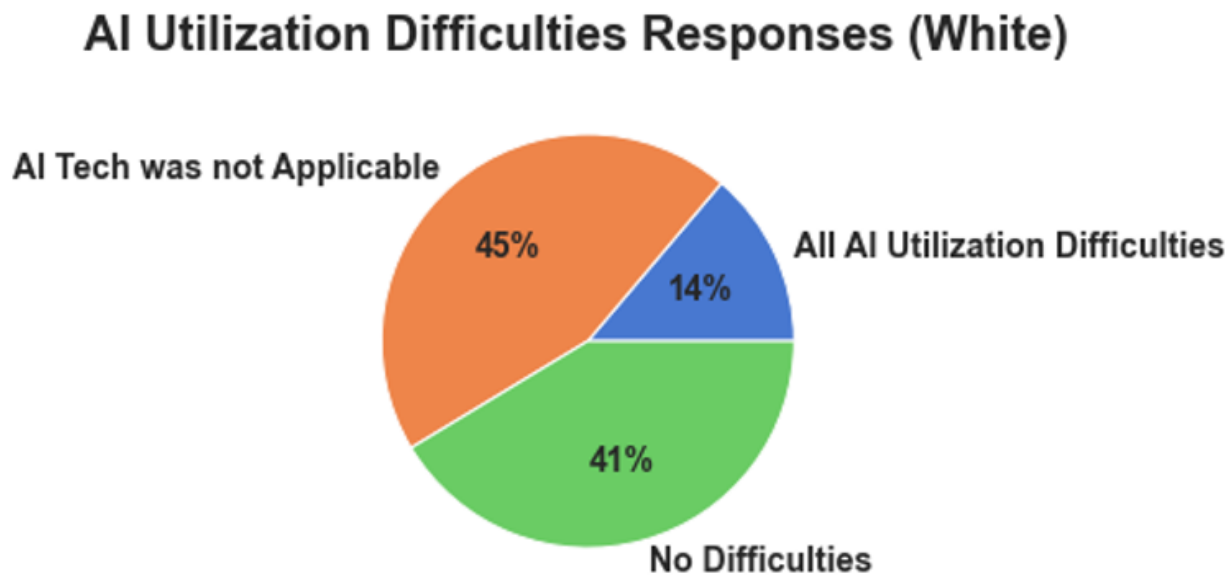
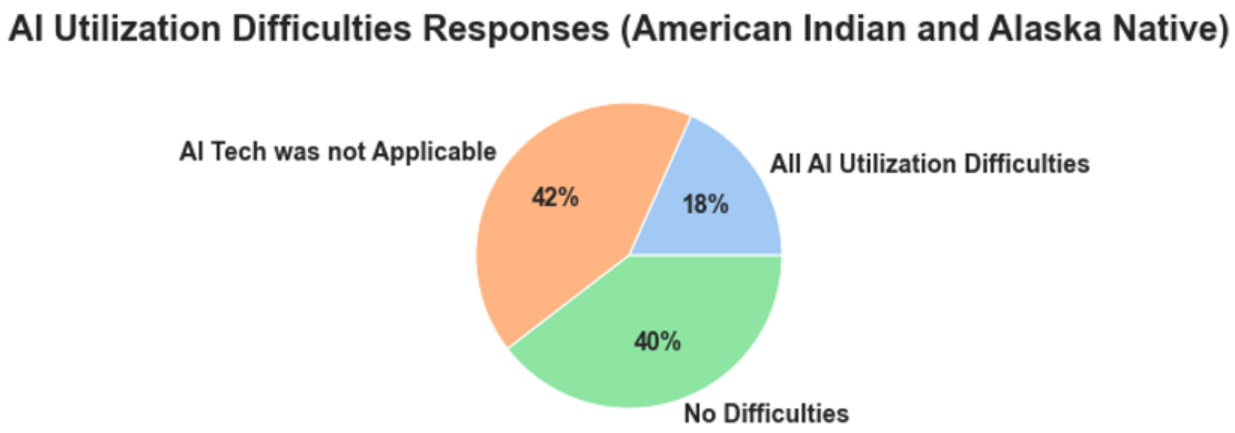


Figure 6: AI Utilization Responses (American Indian and Alaska Native Business Owner): Difficulties and Applicability



Conclusions:

These plots show that overall, businesses whose owners were white had the lowest levels of adverse factors affecting the utilization of artificial intelligence technologies at 14%. As Figure 1 shows, most of these differences appear to be related to the costs of using artificial intelligence, but more research is needed. Another area of further research is to investigate any overlaps with the job sector, as most businesses, regardless of the owner's race, either did not consider artificial intelligence to be applicable to the business or did not have difficulties in utilizing artificial intelligence.