Project Report: Census Exploration - Annual Business Survey (2019)

# Introduction

The Annual Business Survey (ABS), administered by the U.S. Census Bureau, provides information on various economic and demographic characteristics for businesses and business owners, as well as new information on technology use and production for several technologies. Utilizing [the Census API](https://www.census.gov/data/developers/data-sets/abs.2019.html) allows analysts and developers to explore the datasets and identify possible correlations or areas of growth related to business at the country, state, and metropolitan area. The following research and visualizations were found using the 2019 API datasets; it is important to note that the data represents the reference year 2018.

The four datasets include: **Company Summary, Characteristics of Business, Characteristics of Business Owners, and Technology Characteristics of Business.**

# Initial Questions

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 did the total number of employees compare for Family Owned firms by the owner’s race?
* For firms that were family-owned, how did the total number of employees compare when grouping by owner-race?

**Characteristics of Business Owners**

* In the United States, 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**

* In 2018 how did the total revenue of firms differ when comparing their use of Cloud-Based technology?
* In 2018, how did each industries’ use of Specialized-Software compare with one another?

# Data Sources and Information

Through the [US Census Bureau government website](https://www.census.gov/data/developers/data-sets/abs.2019.html), we accessed the datasets on April 22nd, 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.

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

**Characteristics of Businesses** provides data for respondent\* employer firms by sector, gender, ethnicity, race, veteran status, years in business, employment size, and payroll for the U.S., states, and metro areas.

**Characteristics of Business Owners** provides data for respondent\* owners of employer firms by sector, gender, ethnicity, race, and veteran status, as well as detailed owner characteristics, for the U.S., states, and metro areas.

**Technology Characteristics of Businesses** provides data on technology use and production for Artificial Intelligence, Cloud-Based Computing, Specialized Software, Robotics, and Specialized Equipment technologies for the U.S. and state levels.

\*Respondent refers to the 68.2% of survey recipients who responded. Approximately 300,000 employer businesses were mailed the ABS. ABS Methodology can be reviewed on the [Census Bureau website](https://www.census.gov/programs-surveys/abs/technical-documentation/methodology.2019.html).

Using Jupyter Notebook and imported Python libraries such as pandas, Matplotlib, and seaborn, we created specific API calls related to our initial questions, returned the data in JSON format, created dataframes, completed research on our queries, and created visualizations. The full ETL process, including code, can be located [in this project’s GitHub](https://github.com/Jakob-T13/module-8-assessment) in a general format in the ETL-Report and a step-by-step, specific format in our Jupyter notebooks.

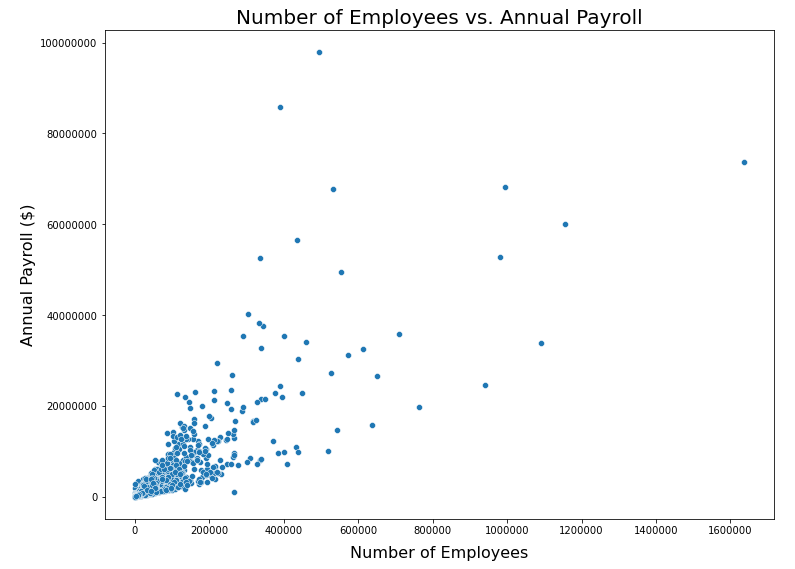
# Research Process, Findings and Visualizations

*Explaining the process to answer those questions*

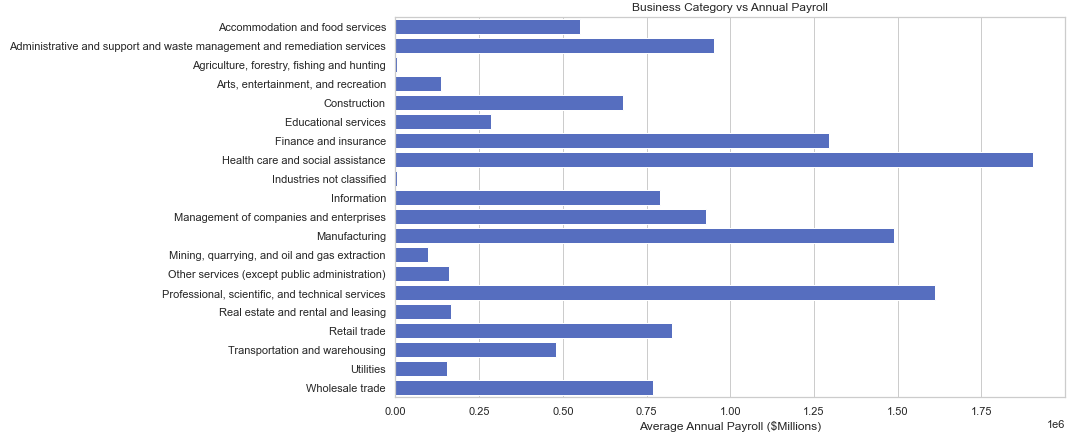
**Company Summary**

The Company Summary (CS) dataset includes data on the surveyed businesses, such as the number of firms within the company, the NAICS industry that the business operates in, and the company’s annual payroll. The structure of this data allows to make broad, general observations without getting too bogged down in details.

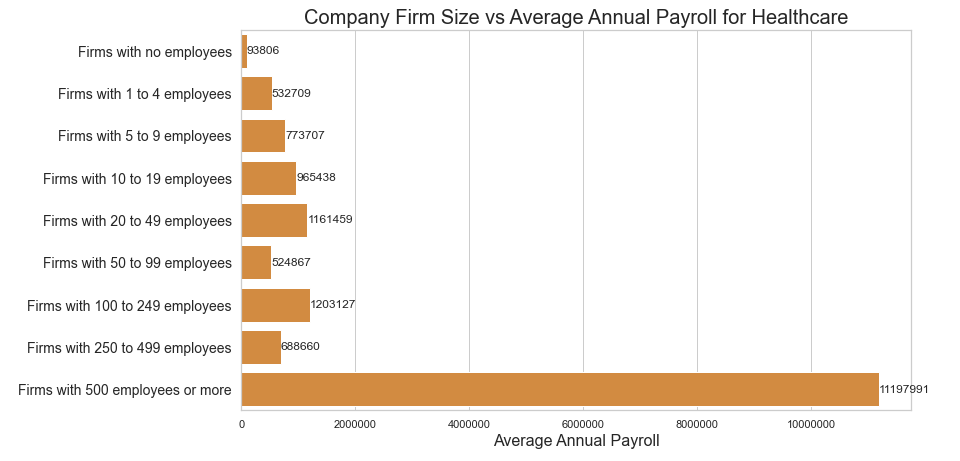
The first question from this dataset was whether there was a correlation between a company’s total number of employees and its total payroll. In the below graphic, it would seem there may be a loose correlation, but most of the data is clustered in the lower ranges, with several outliers. Thus, it cannot be conclusively determined that there is a direct correlation between a company’s number of employees and annual payroll without some other explanation.



To follow up on this, the next question was which industries have the highest annual payroll, to see which ones are causing the outliers seen in the above scatter plot. As seen in the next graphic, healthcare has a clear lead over the other industries in terms of annual payroll, followed by professional and scientific services and manufacturing. However, this chart also shows there may be a degree of selection bias or skewed data, as the average payrolls for Agriculture and Mining seem to be extremely low for the relative size of those industries.

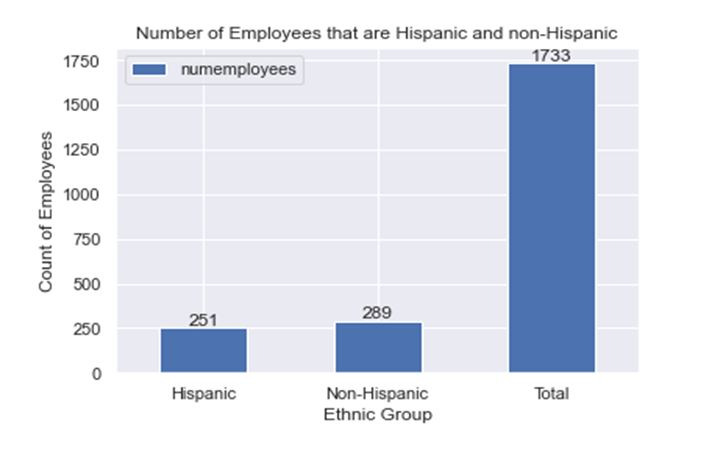


Disregarding the possible issues with the data posed, a third and final question about this dataset was how different-sized firms within the healthcare industry compare to each other in terms of payroll. As shown in the following bar chart, large firms have *significantly* higher payroll. This makes sense, as having more people working at a firm obviously requires more money to pay them. The inverse is true as well, for the most part; small firms like private practices only need to pay the doctor/owner, maybe a secretary as well, meaning much lower total payroll. The only odd figure here is the category of 50-99 employees having lower average payroll than the 1-4 employee category, but this may be another case of selection bias.

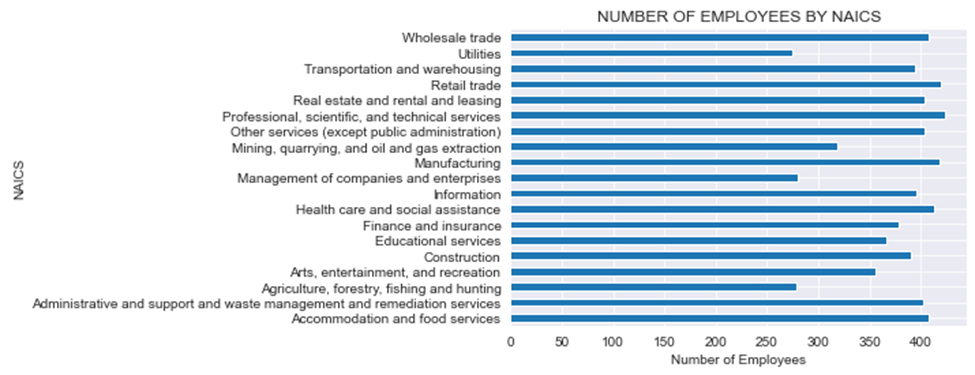


**Characteristics of Businesses**

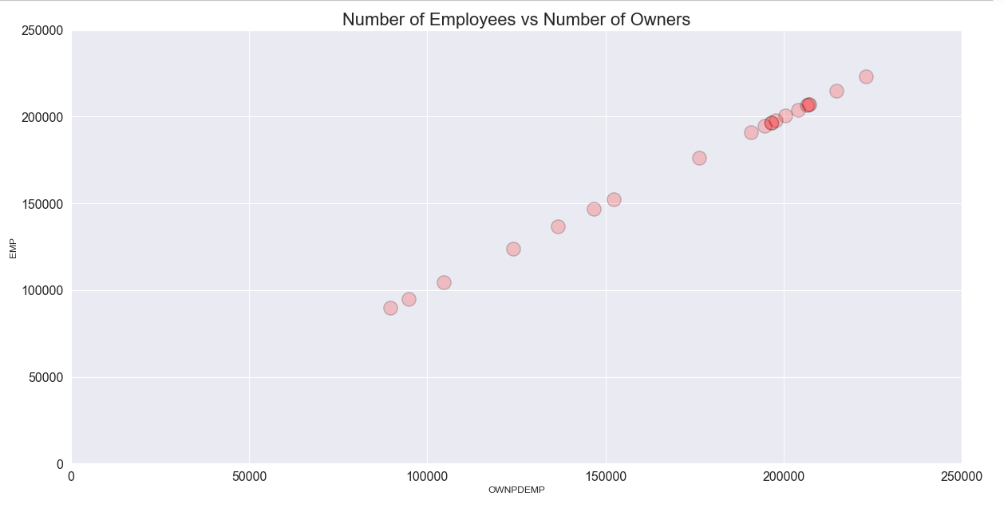
When looking at the census data, one of the questions that we wanted to figure out was, how many employees identify as only Hispanic or only non-Hispanic and compare them to the total sample size. We utilized the characteristics of business data; from the data we can see that about 15% of the employees in the firms are Hispanic and about 17% of the employees identify as non-Hispanic. We are issuing a caution when using these numbers because there are people who identify their ethnicity as ‘Equally Hispanic/non-Hispanic’, which are not included in calculating the 15% and 17%. This can be seen in the bar chart below.



As a group we also wanted to know which NAICS (North American Industry Classification System) had the most employees. For this question we used the characteristics of business data and took out entries that were labeled “Industries not classified”. From the horizontal bar chart below, we can find out the top 3 industries with the highest number of employees which are “Professional, scientific, and technological services”, “Retail trade”, and “Manufacturing”. Also, we can see the industries with the lowest number of people which are “Utilities”, “Agriculture, forestry, fishing, and hunting”, and “Management of companies and enterprises”.

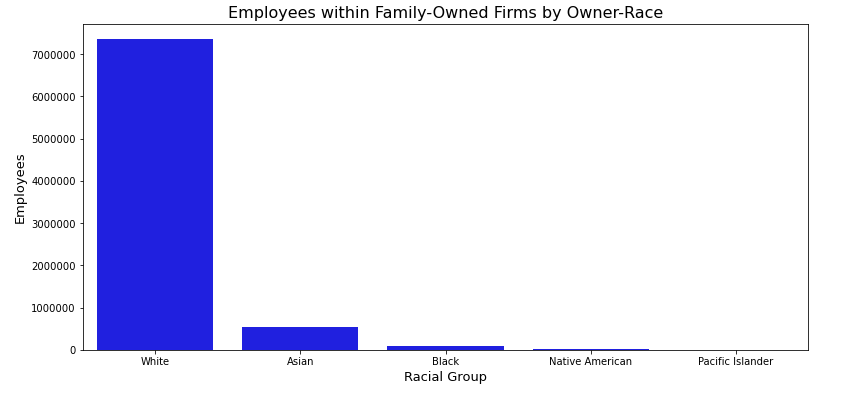


Another finding we wanted to know was whether there was a correlation between the number of employees and the number of owners. What we found is that there is a positive correlation between the two attributes, so as the number of employees increases the number of owners increases. The data points look like it could be modeled by a first order linear regression model. The scatter plot also shows that most firms have around 200000 owners and 200000 employees. These findings can be shown in the scatterplot below.



Another question we wished to answer was “For firms that were family-owned, how did the total number of employees compare when grouping by the owner’s race?” We wished to see how the total employee size for firms differed when compared between firm-owners of different racial groups.

We filtered the data so that all columns that were not of interest were set to their ‘total’ values. We then filtered the data to examine only those firms which were family-owned. Finally, we grouped the employee values by owners' race-group and built a bar plot.



As is evident from the plot, when firms are family owned, most employees work for White-owned firms. This result hints at two possibilities, 1; most family-owned firms are likely White-owned, and 2; these firms are most likely larger, on average, when compared to other family-owned firms.

While, due to that fact that Whites have a majority in the U.S population, we would have expected a sizeable proportion of employees to fall within White-owned-family-firms, the differences shown by this plot are, likely, too vast to be due to population difference, alone. Might there be some factor, or roadblock, preventing families of color from advancing their company? Might the longevity of large family-owned companies suggest that it has been too short of a time since periods of greater systemic discrimination in the U.S? These are important questions, and there is potential value in studies that would seek answers to them.

**Characteristics of Business Owners**

The Characteristics of Business Owners (CBO) dataset and API allows developers to complete research related to business owner characteristics including year of acquired ownership, owner’s highest degree or level of school completed prior to acquiring the business, and whether owners were born in or outside of the United States. To investigate possible disparities, owner classification can be utilized to investigate responses based on gender, race, and ethnicity and find possible correlations or areas of growth.

The first, initial question was to identify which metropolitan areas experienced the highest level of growth in terms of new acquired ownership of businesses. This may provide information regarding metropolitan economies, accessibility to owning a new business, and overall vibrancy of urban areas. As shown in the below visualization, the top ten metro areas by new business ownership as of 2018 includes New York-Newark-Jersey City as the clear leader with over 3,500 new business owners followed by Los Angeles-Long Beach-Anaheim with over 2,500 new business owners and various other metropolitan areas having between 1,000 and 1,500 new business owners including Miami-Fort Lauderdale-West Palm Beach, Chicago-Naperville-Elgin, and Boston-Cambridge-Newton. Metropolitan areas that ranked in the bottom half of new business owners in the year 2018 include Las Vegas-Henderson-Paradise, Cleveland-Elyria, and Charlotte-Concord-Gastonia. While this may suggest booming and less-than-booming metropolitan areas in terms of business success, it would be important to compare these findings with previous years’ survey data and overall population data.

Chart, bar chart

Description automatically generated

Better understanding disparities between U.S.-born and foreign-born business owners across gender and race groupings could lead to improvement in policy-making and programming to ensure that there is equal opportunity for those interested in owning a business. As shown in the below visualizations, U.S.-born business owners by race show an extremely large disparity between white business owners and other race groups including Asian, Black or African American, American Indian and Alaska Native, and Native Hawaiian and Other Pacific Islander. When looking based on gender, it is also clear that business owners are predominantly male. Interestingly, when evaluating foreign-born business owners by race and gender, we see the disparity shrink between White and Asian business owners while seeing a significant disparity remain looking at Black or African American and Native Hawaiian and Other Pacific Islander business owners\*. While business owners are still predominantly male, there are more Asian foreign-born female business owners than White foreign-born female business owners. These findings suggest that there is a need for an improvement in policies and programming to encourage non-White and female business ownership interest, accessibility, and opportunity.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

\*Important to note that while included in this second visualization, it is not surprising to see that American Indian and Alaska Native numbers are low considering they would be overwhelming considered U.S.-born; however, for consistency, they are still included.

Lastly, when researching at the state level, comparing business owners’ educational attainment may shed light on the benefits of attaining various college degrees. For simplicity, comparisons in the below visualizations focus on business owners with a high school diploma or equivalent and business owners with a Bachelor’s degree in any major; however, future research including other levels of educational attainment could provide stronger findings. When sorting by states with the highest numbers of business owners in each educational attainment category and comparing between the two, it was interesting to note states that have close to a 50/50 split: Wyoming, South Dakota, West Virginia, Maine, Mississippi, Iowa, Indiana, Wisconsin, and Michigan, and states that have higher numbers of business owners with a Bachelor’s degree: California, Texas, Florida, New York, and Nebraska, in the below visualizations. This comparison may provide prospective business owners with information relating to their own educational attainment and the possible likelihood of acquiring business ownership based on state patterns. Of course, it would be important to better understand population data as well as omit outliers like California, Texas, Florida, and New York to provide a more clear visual of state patterns.

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

**Technology Characteristics of Business**

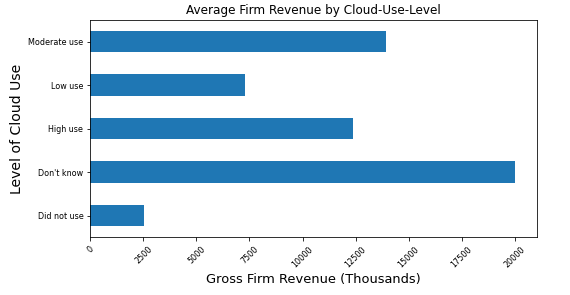
The Technology Characteristics of Business API presents survey data from numerous respondents describing several aspects of technology, and its relationship to business. Responding firms described their use of various technologies, which included: Cloud-Based, Specialized Equipment, Specialized Software, Artificial Intelligence, and Robotics. Some of the questions answered were: “Why was a certain technology used?” “How did it affect the technical skill of employees within the firm?” and “How did that technology influence the firm’s capacity to engage with markets?”

As with the other APIs presented in the Annual Business Survey, there are columns which describe the demographic background of the firm’s owner. There are also several numeric columns that contain information ranging from, the number of employed persons, to the total revenue, to the total payroll of firms throughout the United States.

The amount, and breadth, of information contained in the API allows for numerous approaches of study.

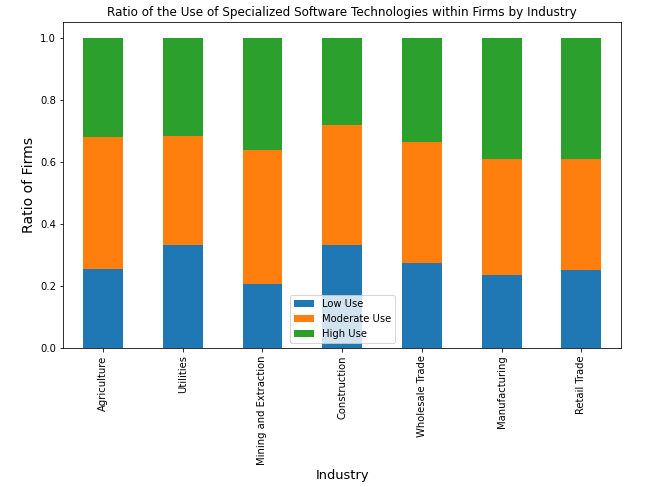
The first question to arise from this API was, “In 2018 how did the average revenue of firms differ when comparing their use of Cloud-Based technology?” We wished to know if firms that identified as using a high amount of Cloud-Technology were larger (revenue-wise) than those that did not, or that used less.

We began by filtering the data to rows that solely contained data which pertained to Cloud Technology. We then transformed the revenue columns, dividing their fields by the total number of firms listed for each row. We then executed a group-by, averaging these values across the levels of Cloud Technology use. Finally, we built a bar plot using this transformed data.



There is evidence that firms which do not use Cloud Technology are smaller, in comparison, to those that do. Furthermore, firms that identify as having a “Low” use of Cloud Technology seem to be worth less than those who use a “Moderate", or “High” amount. When analyzing the difference between firms who identify as using a “Moderate” and “High” amount of Cloud Technology, the answer is less conclusive. As with all data, it is important to interpret these results with a degree of skepticism. Because this data is self-reported, there is a high amount of subjectivity present. What constitutes “Moderate” vs “High”? And are the respondents guaranteed to be honest when describing a firm to which they belong? These questions indicate that this data, while having value, cannot be considered as having complete statistical accuracy.

The second question to arise from this API was, “In 2018, how did each industries’ use of Specialized-Software compare with one another?” To investigate this question, we began by filtering the data to just the ‘TECHUSE\_LABEL’ column’s responses of “Specialized Software: Low use,” “Specialized Software: Moderate use,” and “Specialized Software: High use.” Then, following transformations, we created a series of stacked bar-charts containing the ratio of all three instances, each chart’s values adding to one.



The above plot represents a sample of the plots for all industries. Of these, the manufacturing category has the greatest proportion of firms to make “High” use of Specialized Software. The construction and utilities category has the highest proportion of firms to make minimal use of specialized software.

The reasons for these differences are not immediately evident. They could, however, serve as basis for a new study into the matter. Such a study may involve listing some of the most common types of specialized software and charting their sales across various companies and industries.

There is another way in which the use of software could be quantified, and it would be mostly straight-forward. Suppose a researcher gathered data measuring the total number of hours employees spent working with Specialized-Software in various companies. After a distribution of these measurements has been created, a company could be classified as having “High” use, “Moderate” use, or “Low” use based on where the percentage of their total hours fell within the it: if below 33%, a company would be classified as “Low”, below 66% as “Moderate”, and so on.

# Discussion

As previously stated, it is important to note that the findings and visualizations presented are from largely respondent-only survey responses from business firm owners. While various steps were taken by the Census Bureau to review, correct, and verify operations, processes, methods, and response data were carried out as specified in the ABS methodology, considerations should be made. Specific to questions related to race classification, respondents were instructed to select all that apply; this means that an owner and/or firm could be duplicated within the race classification data if they selected more than one race. Throughout the research, we noted that future research should cross-reference important data like population, previous ABS survey years, and other business-related metrics to strength the presented findings.