Assessment 8 Project Report

Group 6:
Armin Nouri
Chris Kusha
Jassleen Bhullar
Sara Douglas

January 3, 2022

INTRODUCTION

This project uses data provided by the United State Census Bureau's 2019 Annual Business Survey (ABS). The ABS results are broken into four datasets: Company Summary, Characteristics of Businesses, Characteristics of Business Owners, and Technology Characteristics of Businesses. These datasets supply information about the demographic and financial attributes of the companies that participated in the survey. The Census Bureau made the four datasets available to the public through a Web API. Upon requesting an API key from the site, the group was able to use the API key to make requests to access the data using specific URLS (provided on the site) for each dataset. From the data gleaned from the ABS datasets, each member of the group had specific questions about the businesses that they wanted to explore further. This project explores the following research questions:

1. Does the race of the business owner(s) affect the annual revenue of the company?
2. What factors may influence a company to not disclose its exact revenue per year?
3. Do some states provide better opportunities for minority business owners? Are there differences in revenues generated by companies of different races in different states?
4. Are there industries in which minority business owners are more likely to generate high revenues?
5. How are business owners' ages distributed and how is that impacted by sex or race?

6. How does technology usage differ from state to state?
7. How does technology use differ across industries?
8. What effects has technology had on the worker's skill levels and staffing?
9. How is the data, including nulls, distributed among the customer and worker datasets?
10. How does worker pay vary across industry and worker types?

DISCUSSION

Initial data exploration and analysis led to the formation of several questions and hypotheses regarding employee information, revenue, owner demographics, and technology usage. All the data related to these questions was divided amongst the four datasets, so each member of the group explored the topics they were curious about individually to form a cohesive discussion that summarizes all four datasets.

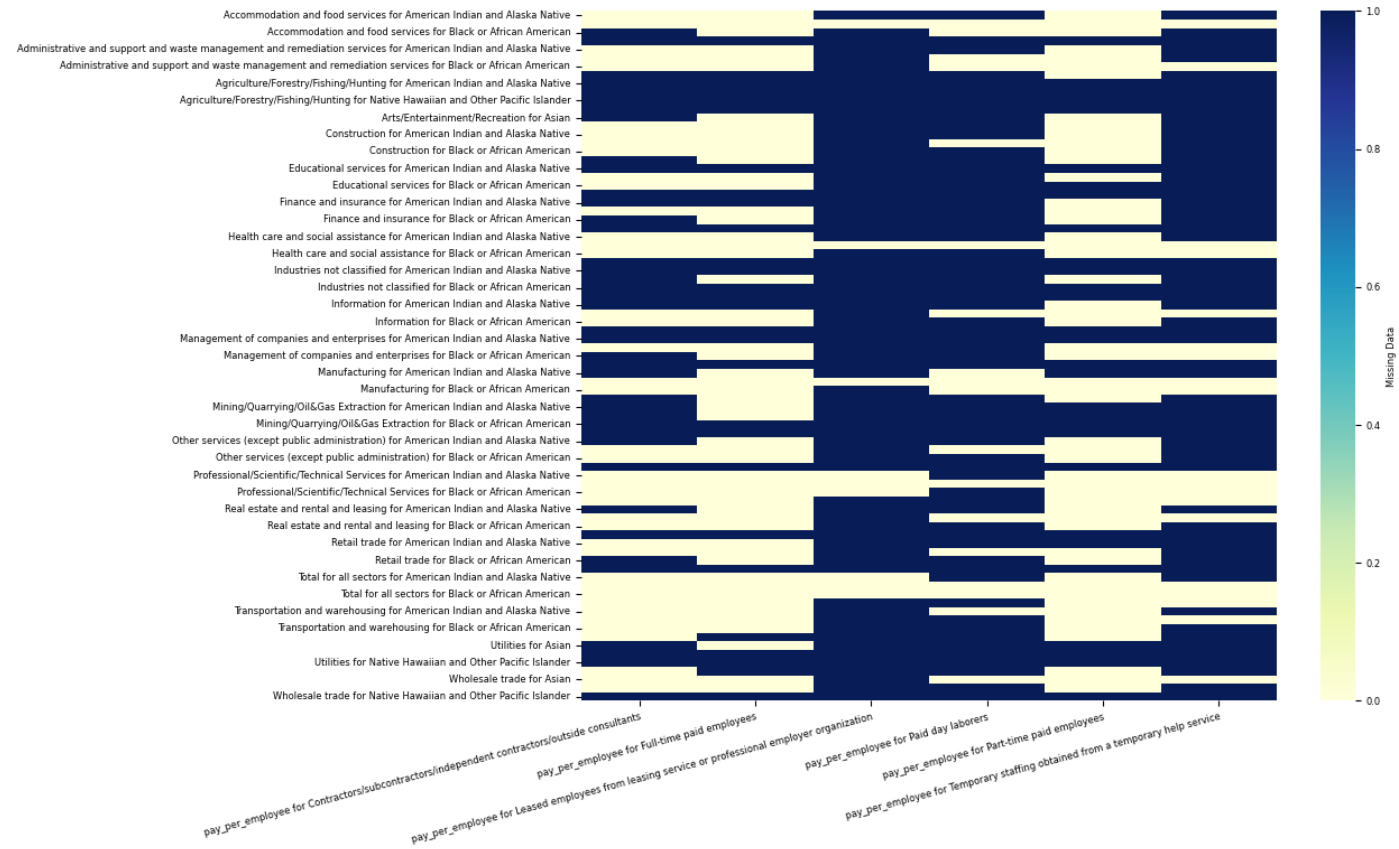
BUSINESS CHARACTERISTICS

Within this dataset, I was interested in looking at the data the Census Department had collected pertaining to workers and customers, which they were so kind as to label for easy access. Seeing as how this data also tracked the industry and race of the reporting party, I thought it might be interesting to look at things aggregated by these two categories.

Let's start with worker data, which was broken up into the following categories (among others indicating totals or missing values):

- Full-time paid employees
- Part-time paid employees
- Contractors/subcontractors/independent contractors/outside consultants
- Temporary staffing obtained from a temporary help service
- Paid day laborers
- Leased employees from leasing service or professional employer organization

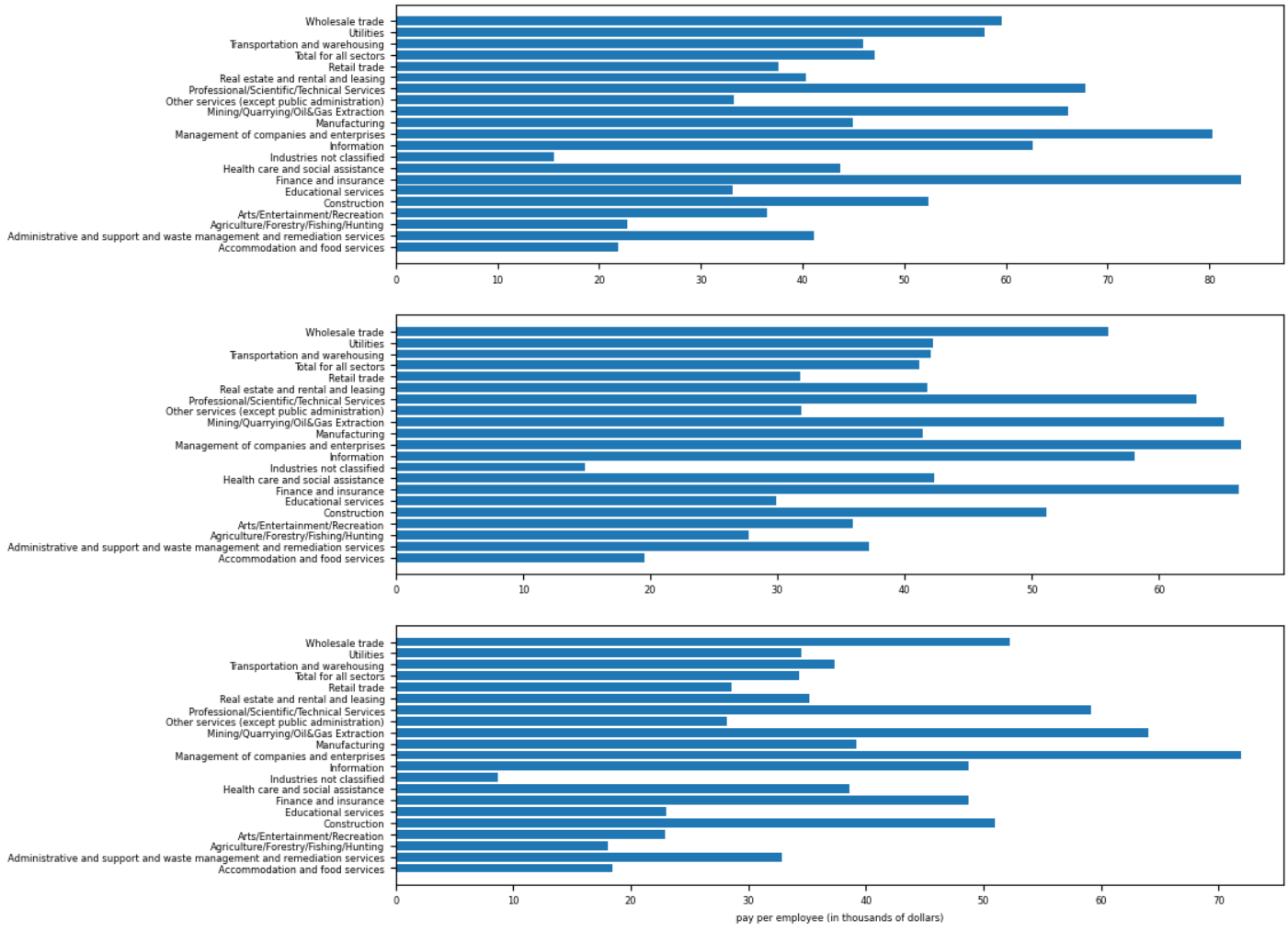
I was immediately curious to see how null values were distributed throughout this dataset, which I visualized using the following seaborn heatmap.



On the y-axis are all the industry-race combinations (e.g. construction for Asian would be businesses labeled as being under the construction industry where the ownership is 50% or more Asian). On the x-axis are all the worker descriptions I mentioned earlier. This heat map shows the spread of null values after grouping by industry, race (data from whites omitted as I was focusing on minority owned businesses), and worker description, aggregated by the average pay per employee (which I calculated very simply by dividing the average annual pay by the number of employees the business listed). The darker the shade of blue, the higher the ratio of nulls for the combination.

As you may have noticed, 3 of the employment categories held the vast majority of nulls as evidenced by the dark blue all throughout the column. These would be the categories involving leased employees, temporary staffers, and paid day laborers. For simplicity then, I chose to remove these categories from my next graphic, and focus on the more common

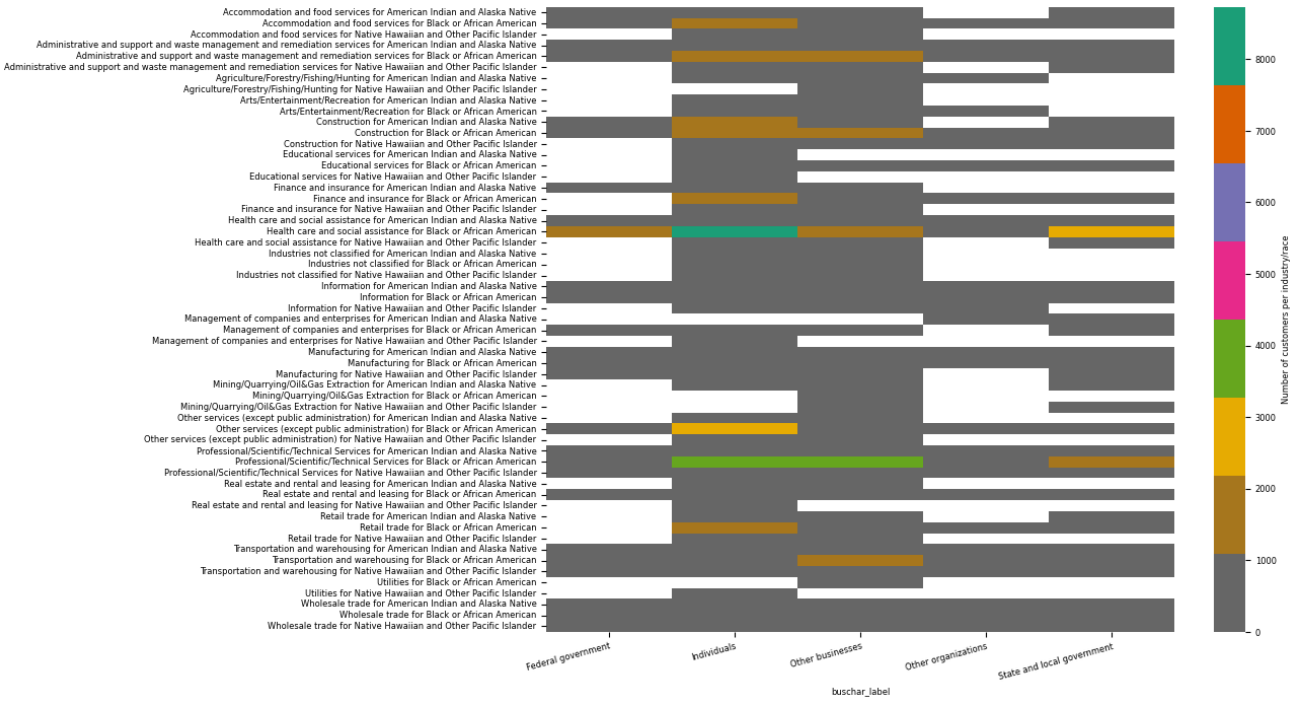
categories: full time, part time, and contractor employees. Since these columns still had some nulls in them, I filled them using the column average (e.g the average pay for a worker of that type, regardless of industry) as I thought that would be a more relevant factor than the row average. I also removed the race column from my initial grouping as I found the data to be too cluttered without offering anything particularly interesting in return. The option to include it was left in the original code, however, for anyone curious to see it.



The axes are similar to the previous graph where the y-axis has the industries of interest and the x-axis has the employee info, this time the annual average pay per employee in thousands of dollars. The graphs are not labeled, but from top to bottom they are for

Contractors/subcontractors/independent contractors/outside consultants, Full-time paid employees, and Part-time paid employees. The graphs look fairly similar at first glance, which is perhaps a little surprising given my first instinct would have been to assume Full-time paid employees made the most on average. This could have to do with the null values being filled in the previous step. Looking back at the previous graph, it seems the Contractors column had the largest number of nulls of the three, so if there were any high paying outliers there, that could've dragged up the amount for all the industries. Given more time, it would be interesting to create another visualization which looks at the variability of pay in the three subplots graphed against the number of nulls in each, which would tell us if the imputation of nulls had any noticeable effect.

I underwent very similar steps for the customer data, this time with all the data on one heatmap:



The null distribution here is visualized as the white spots in the data, with the grays, greens, browns, and blues representing the non-null data. Here I was looking at the number of firms (grouped by industry *and* race this time) which had more than 10% of their clients in a given sector (belonging to various government entities, individuals, other businesses, or miscellaneous organizations).

What I immediately noticed is that, perhaps unsurprisingly, almost every combo had at least some firms where more than 10% of their clientele were individuals or other businesses (i.e the private sector). The ones that didn't, I assume, were combinations that were already very scarce to begin with (e.g Black owned mining/gas extraction companies).

Future groups may be interested at looking into how closely owner race + industry is correlated with clientele category, among other things.

REVENUE

Revenue is one of the most commonly examined factors when analyzing a company, as it informs us how successful, or unsuccessful, the company has been in generating sales. While revenue is different from net profit, it still is a useful metric for the success of a company, especially since the ABS datasets do not provide information about profits. All revenue data was found in the Company Summary dataset.

One of the first things we, as a group, discovered during our initial exploratory analysis is that not all companies disclosed their exact revenues. A large number of companies had disclosed their revenues, but an equally large portion of companies had not. Instead, they provided a range that their revenue fell within. The Census Bureau used the following symbols to denote the ranges:

- B : Less than $1 million
- I : $1 million to less than $5 million

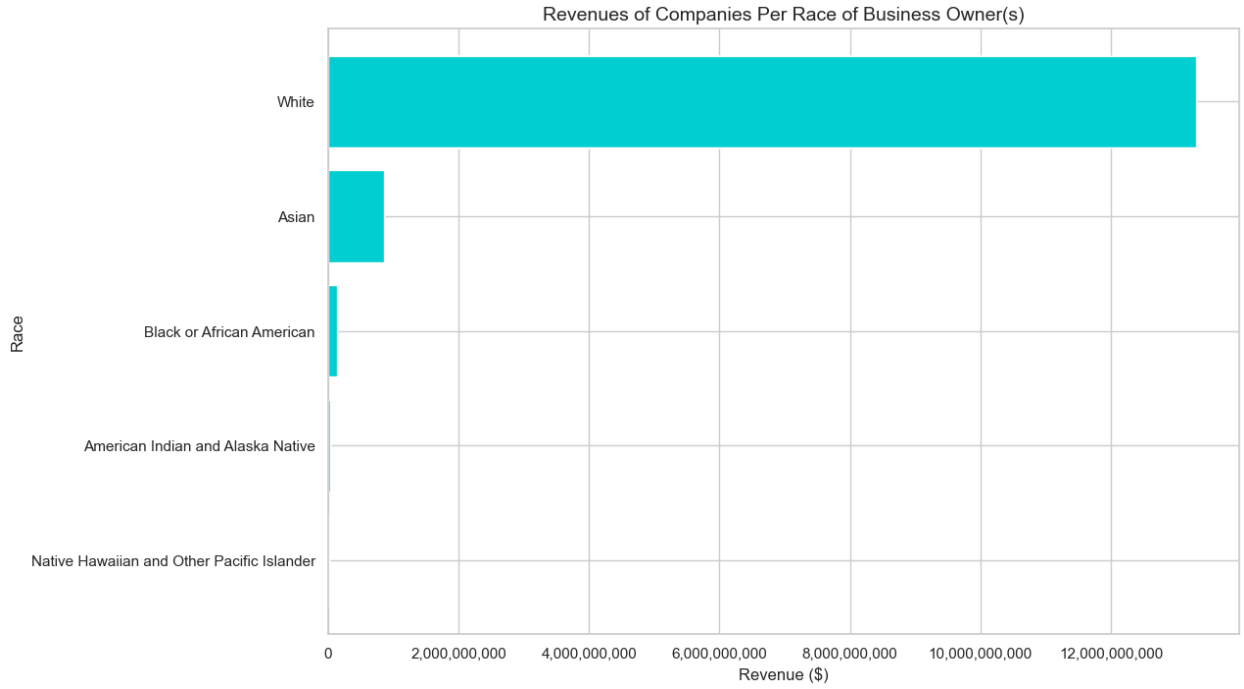
- K : $5 million to less than $15 million
- L : $15 million to less than $50 million
- M : $50 million to less than $75 million
- O : $75 million to less than $150 million
- R : $150 million to less than $500 million
- T : $500 million to less than $1 billion
- U : $1 billion to less than $5 billion
- W : $5 billion or more

These alphabetical letters were later converted to a numeric scale, starting with one and ending with ten, to make data analysis and visualization of the revenue ranges possible. We hypothesized that companies with higher revenues would be less likely to disclose their revenues. We knew that private companies are not required to disclose their annual revenues, and private companies usually generate higher revenues. To test this hypothesis, we filtered for companies that had not disclosed their revenues and created a histogram to see the number of companies in each range of revenue that chose not to disclose.

Out of the companies that did not disclose their revenues, a large number generated revenues of $5 billion or more. The range that had the least number of companies was the lowest range with revenues of less than $1 million. Overwhelmingly, the companies that did not disclose their revenues had revenues of $1 billion or higher (ranges 9 and 10). This histogram suggests that companies with higher revenues are indeed less likely to disclose their exact annual revenue. Information about the companies' public or private status is not provided, but we do know that every company that did not disclose has to be privately owned. Therefore, we can infer privately-owned companies with higher revenues ($1 billion or higher) are less likely to disclose their exact annual revenues.

We were also interested to see if the race of the business owner has an effect on the company's revenue. We had a strong feeling that businesses with white owners would have

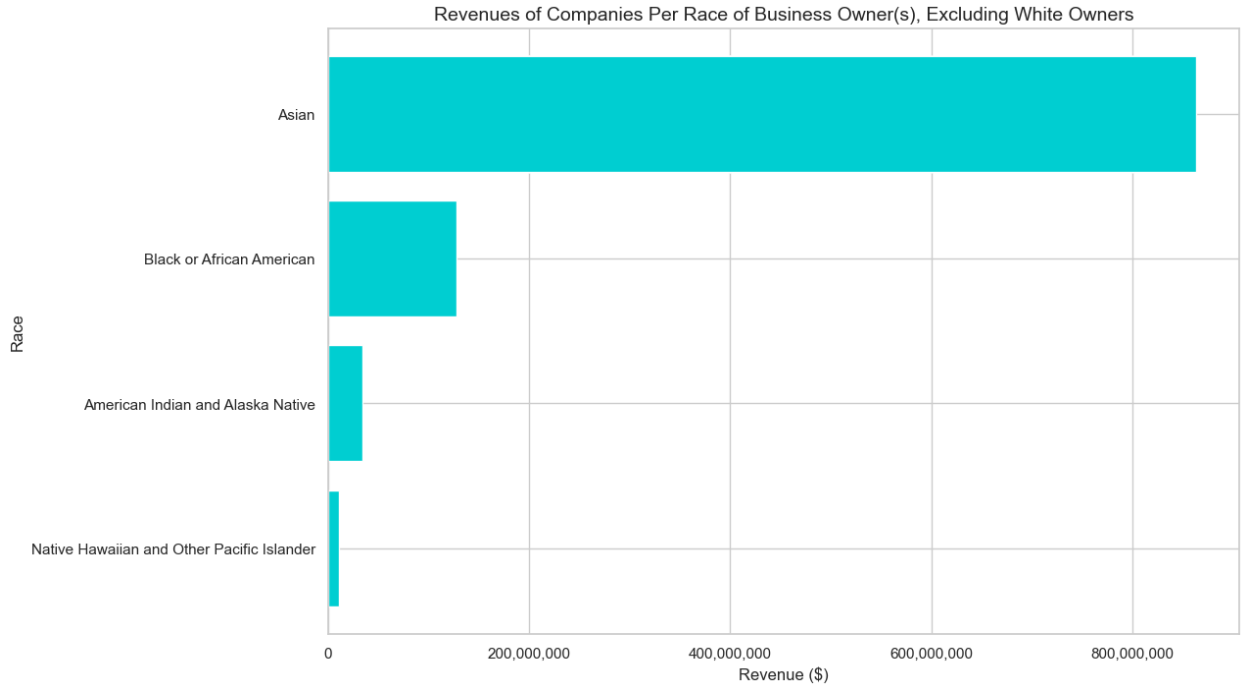
higher revenues than businesses owned by people of color due to racial inequities faced by people of color in the economic sphere.



The data supports the initial hypothesis, but the extremely large difference between the revenues of companies owned by white people and companies owned by people of color was shocking.

Companies owned by white people have much, much larger revenues than companies owned by people of color. No other race comes even close to generating the same amount of revenues. The difference is particularly startling between companies owned by Native Hawaiians/Pacific Islanders and companies with white owners. The most profitable company in this dataset is a white-owned company with over $13 billion reported in revenue. In contrast, the company with the lowest reported revenue is a Native Hawaiian/Pacific Islander-owned company with a revenue of only $1,562. This does make sense considering that people of color face many obstacles when starting businesses, such as lower rates of approvals for loans and less

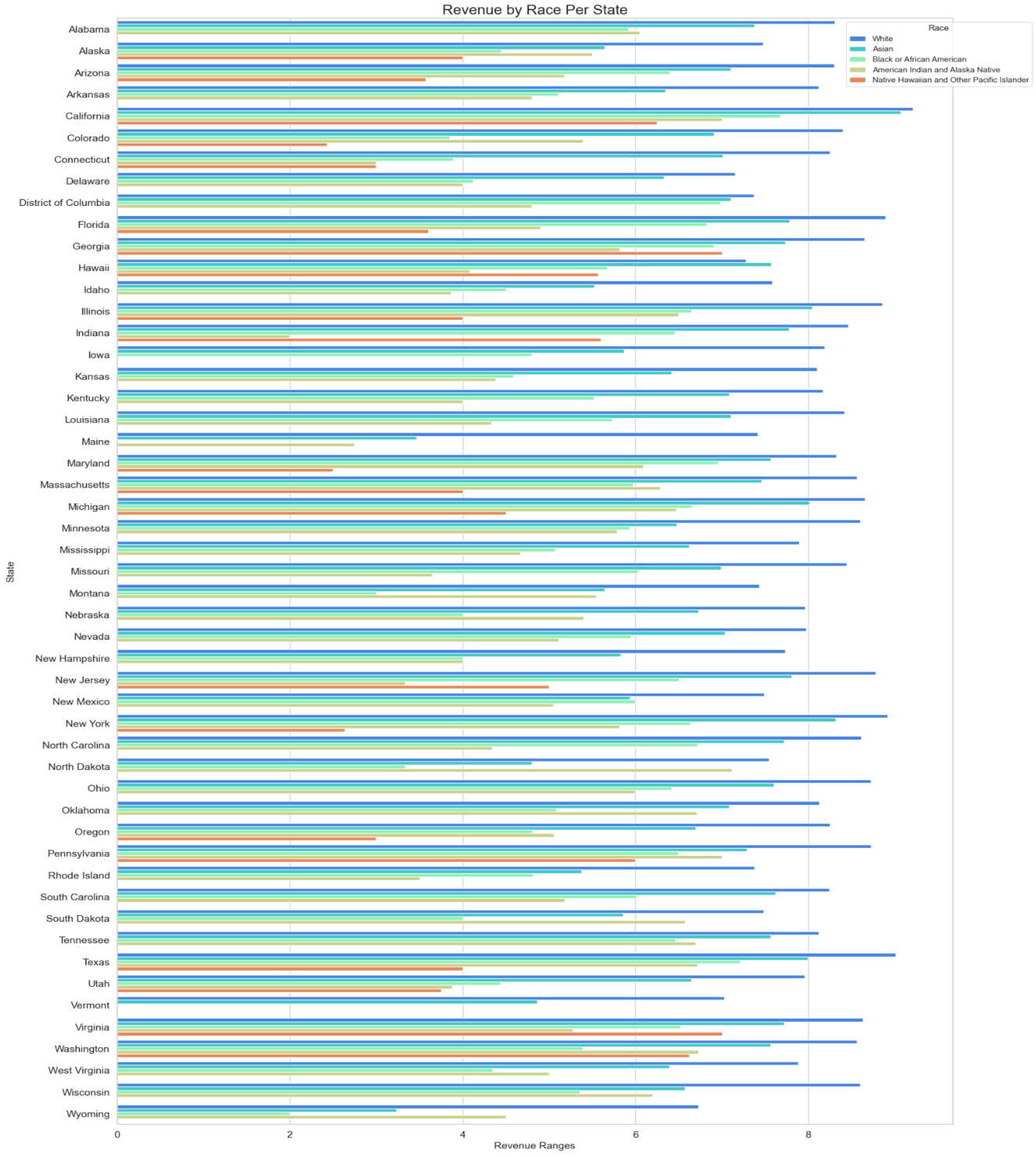
surplus income. Businesses owned by people of color are also more likely to be newer companies, and generally, older companies are able to generate higher revenues.



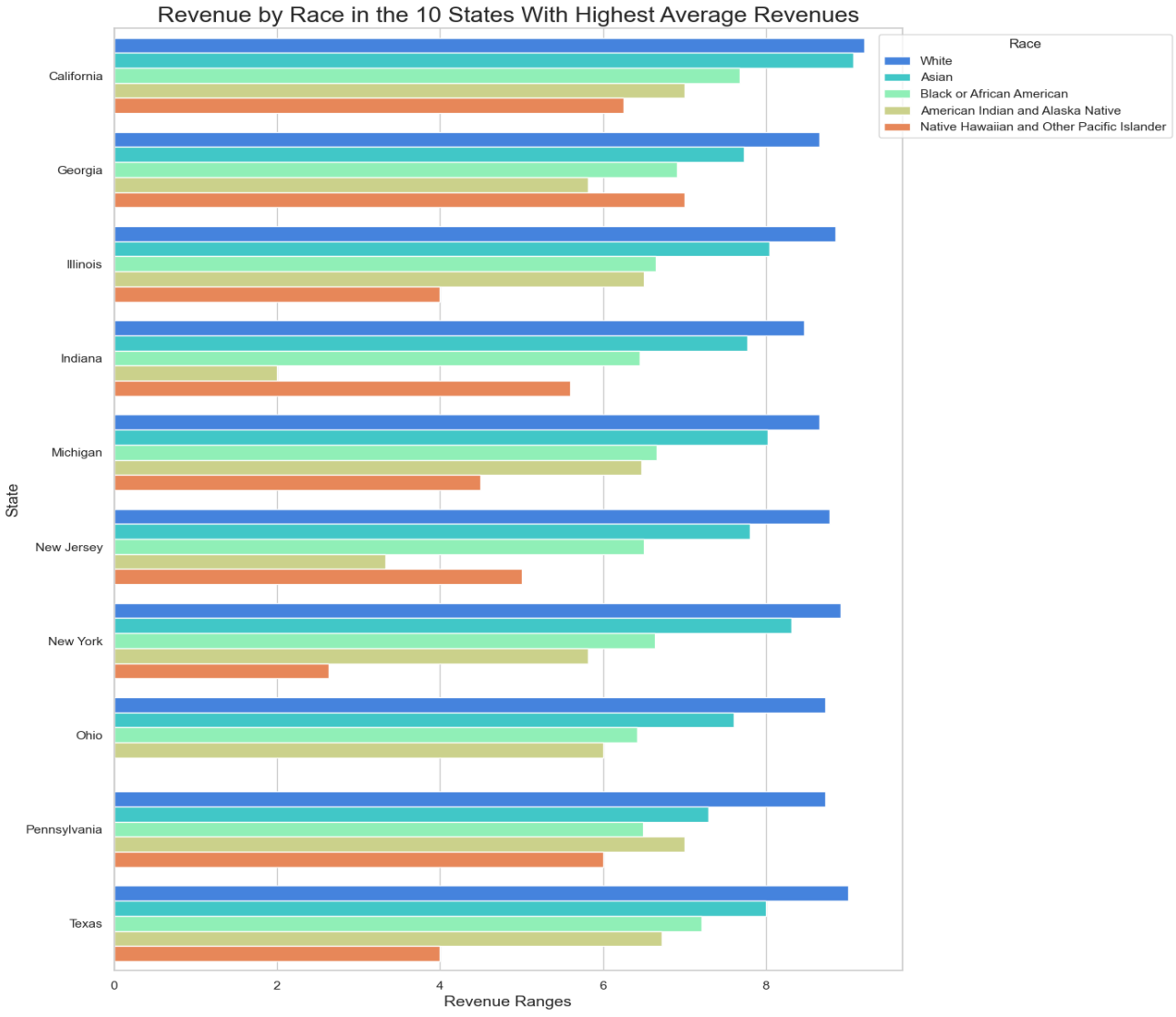
For Native Hawaiian/Pacific Islander owned companies to show on the graph, white-owned companies had to be removed from the dataset since they are such a large outlier. Even after doing so, it is clear that Native Hawaiian/Pacific Islander owned companies fare worse, in terms of revenue, than companies owned by other races. Asian-owned companies have the highest revenues out of minority-owned businesses. Several factors could influence such results. Perhaps, a large portion of Native Hawaiian/Pacific Islander business owners did not participate in the Annual Business Survey. It could also be that Native Hawaiian/Pacific Islanders are a very small subset of the U.S. population. More data would be required to examine why Native Hawaiian/Pacific Islanders owned companies generate far less revenue.

The previous two figures are based on companies that did disclose their revenues, but the ABS Company Summary dataset provided state information for companies that did not disclose

their revenues. We wondered if some states provide better opportunities for minority-owned businesses than other states, so we conducted some data analysis to find the revenue ranges of companies by race in each state.

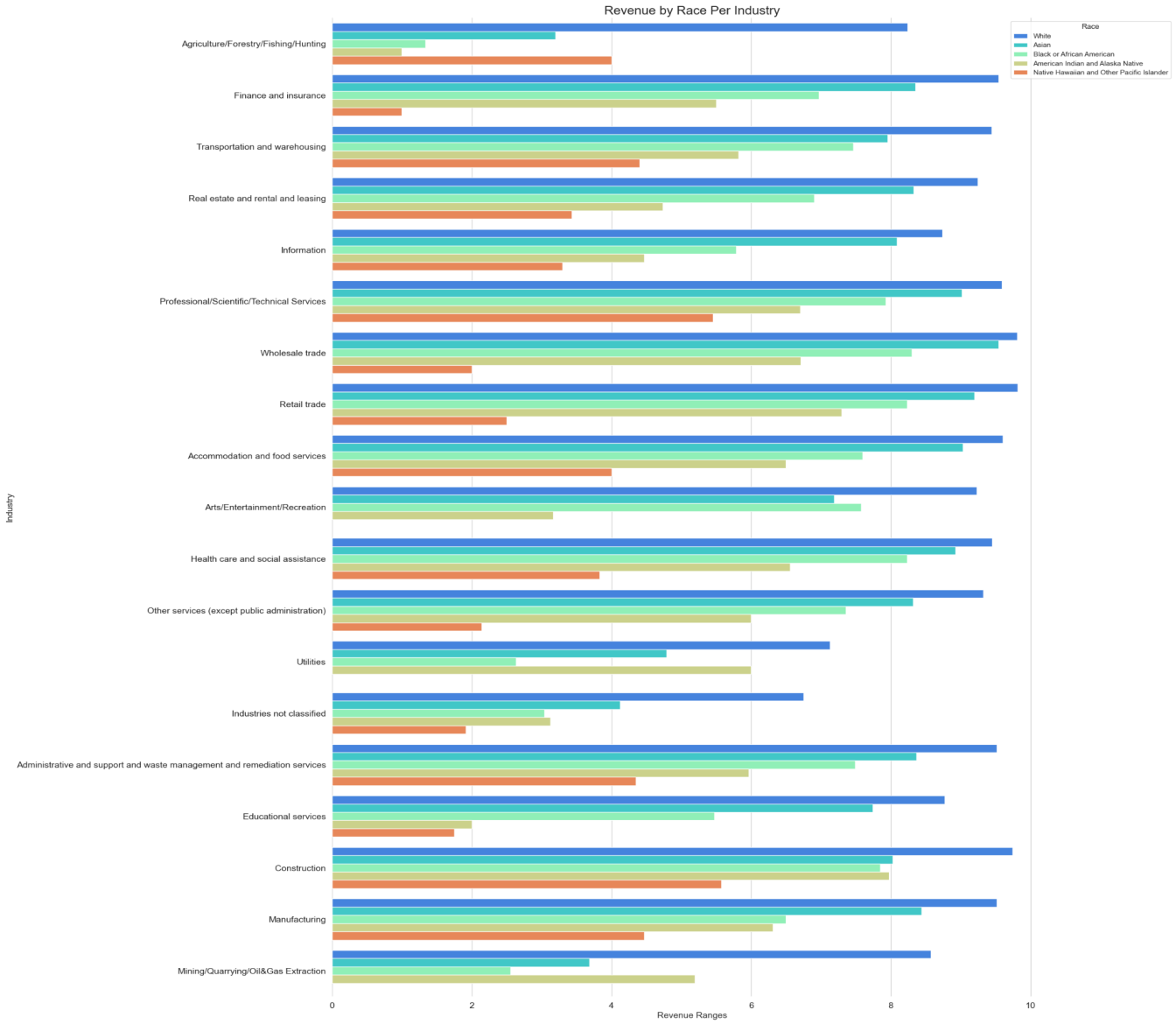


This visualization is quite large and overwhelming, but we could see that the same pattern of white-owned businesses generating the highest revenues, with Asian-owned businesses following. Several states like Maine, Wyoming, and Vermont did not have enough data for races other than white and Asian. In some states like California, New York, and Massachusetts, the disparities in revenues appeared to be less striking. Many states like Iowa, Alabama, and Wisconsin still had vast gaps in terms of revenue based on the race of the owner(s).We drilled down into the ten states that had the highest average revenues to create a more comprehensible image.



Amongst the ten states with the highest average revenues, white-owned companies are still generating the highest revenues, but the gaps between Asian-owned and Black-owned companies are noticeably smaller. Native American and Native Hawaiian/Pacific Islander owners trail far, far behind, except in Georgia and Indiana. Curiously, Native Hawaiian/Pacific Islander owners, in particular, seem to do well in those two states. This is a very interesting finding we would love to explore further, but at the moment, we do not have any data that corresponds to this finding.

The dataset also provided industry information for companies that did not disclose their revenues. We wondered if there were greater or smaller disparities between revenues of white-owned companies and minority-owned companies in specific industries.



Once again, white-owned companies generate much higher revenues than businesses owned by people of color in every industry. There are some industries where the differences are much less overwhelming and point to progress. Businesses owned by people of color are succeeding and actually competing against white-owned businesses in Finance and insurance, and Professional/Scientific/Technical Services. Trade (Wholesale and Retail) and Healthcare appear to be the industries in which minority-owned businesses generate the highest revenues in. Native American owned businesses seem to be thriving in the Utilities and Mining/Quarrying/Oil

& Gas Extraction industries. Other industries like Agriculture/Forestry/Fishing/Hunting still have much catching up to do, with white-owned businesses generating significantly more revenues than businesses owned by other races. Unfortunately, Native Hawaiian/Pacific Islander owned companies still trail behind other races in terms of revenue in most industries.

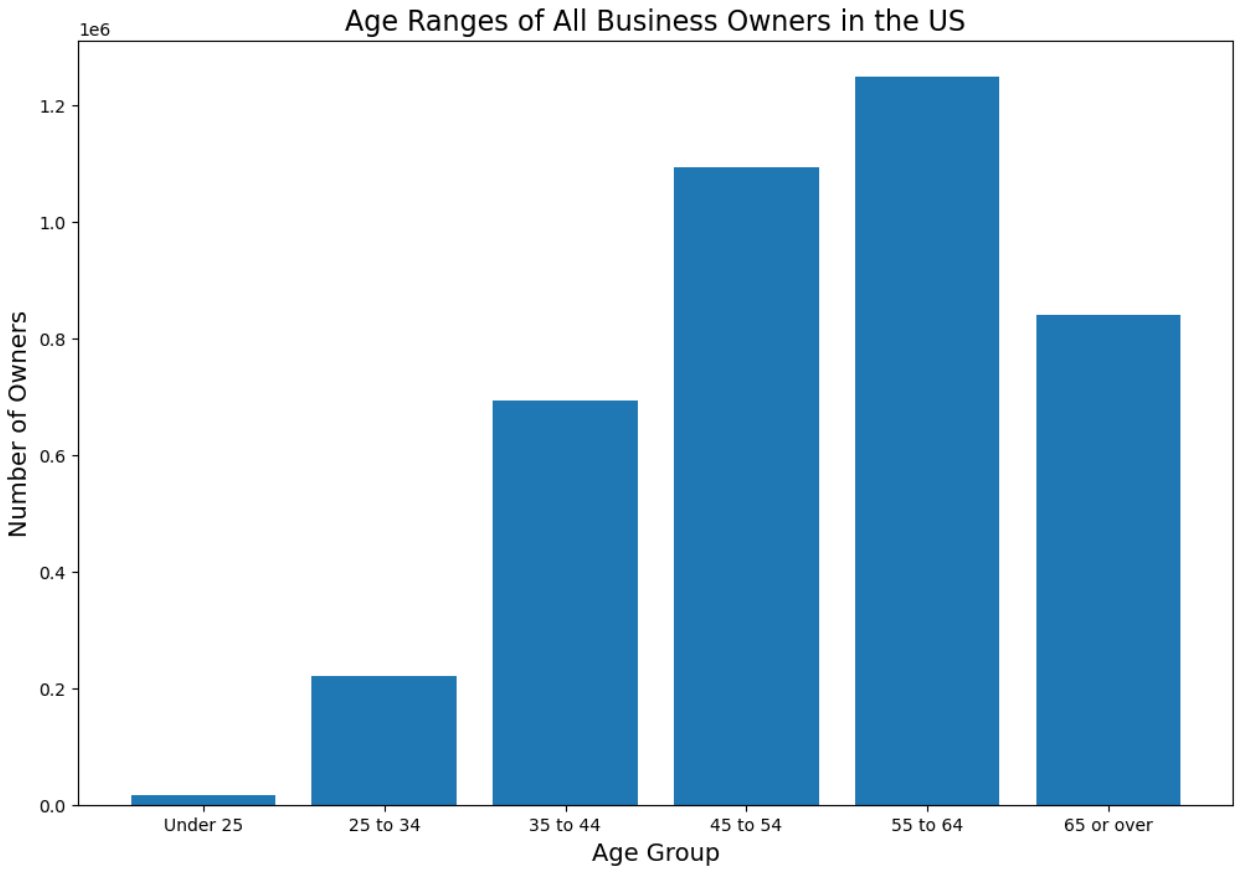
AGE RANGES OF BUSINESS OWNERS

To examine the ages of business owners the Characteristics of Business Owners dataset was used. Within the dataset multiple age ranges are provided each with their own unique code. With this analysis we were interested to explore how a business owner's age was distributed, as well as how sex and race affected the distribution. For this analysis we decided to exclude the "total reporting" and "item not reported" categories so we can focus on declared ages. Additionally, we filtered the owner ethnicity to include only "total reporting" so as to not break down by ethnicity. Since we wanted to focus on the sex and race of business owners it was important to leave sex and race unfiltered as filtering can cause loss of data down the line.

Our initial expectations were that business owners tend to be in older age groups. We expect older people to have had more years to grow professionally and tend to have more resources and connections, allowing for the success of business ownership. Examining the chart below we determined that our assumption was correct. Business ownership increases with age groups until the age group of "65 or older". Interestingly 65 is considered the age when many retire, and this appears to be reflected in our data. However, the number of business owners over 65 is still larger than some of the younger age groups which suggests that many do not retire and still own their business. Another interesting thought is the "65 or older" and the "under 25" age groups span a larger age range than the other groups. All other age groups span a range of 9 years, while "under 25" spans 25 years and "65 or older" is open and has no limit. There is a

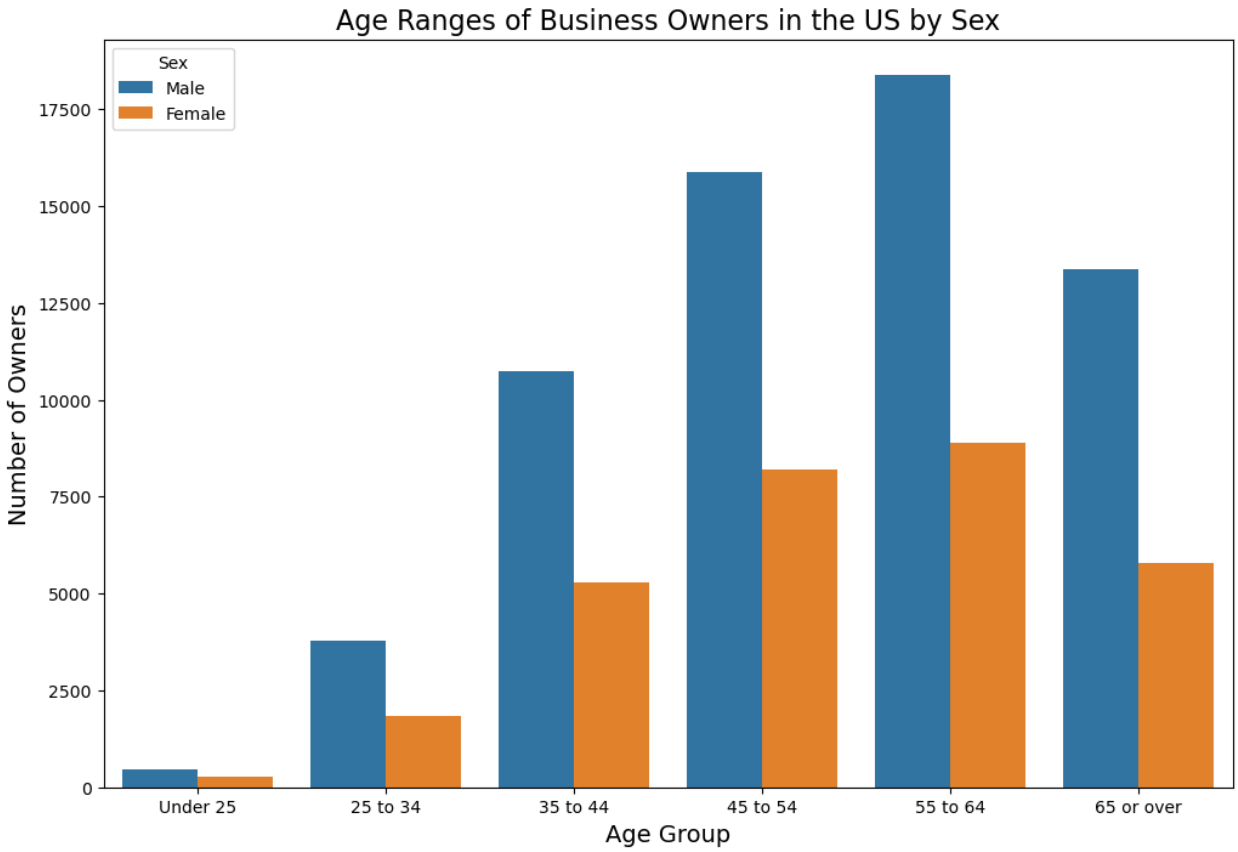
limit to human age however having a “65 or older” age group can account for the high number of business owners as they are all grouped together after the age of 65.

On the other end of the spectrum the “under 25” age range has the least business owners. Which makes sense as until age 18 people are still considered children, and thus may not be able to legally own businesses. So realistically this age group only includes business owners from the ages of 18 to 25, which only spans 7 years. This can partially account for the lack of business owners reported for this age group; however, our initial assumptions can also explain this. Younger adults have had less experience purely due to having lived for less time. Additionally, business ownership requires a certain amount of financial security that younger adults may not have achieved yet.



When filtering our data by sex the number of business owners significantly decreases. This can be explained by the methodology behind the collection of the data, since business owners may not provide this information. Our initial expectations were that males were more likely to be business owners as they have had legal rights for longer than females. Since the prestige of business ownership can be inherited by the following generation, those in power will often pass that power to their children. Thus, the idea that business owners are more likely to be male aligns with the idea of inheritance and passing down family wealth.

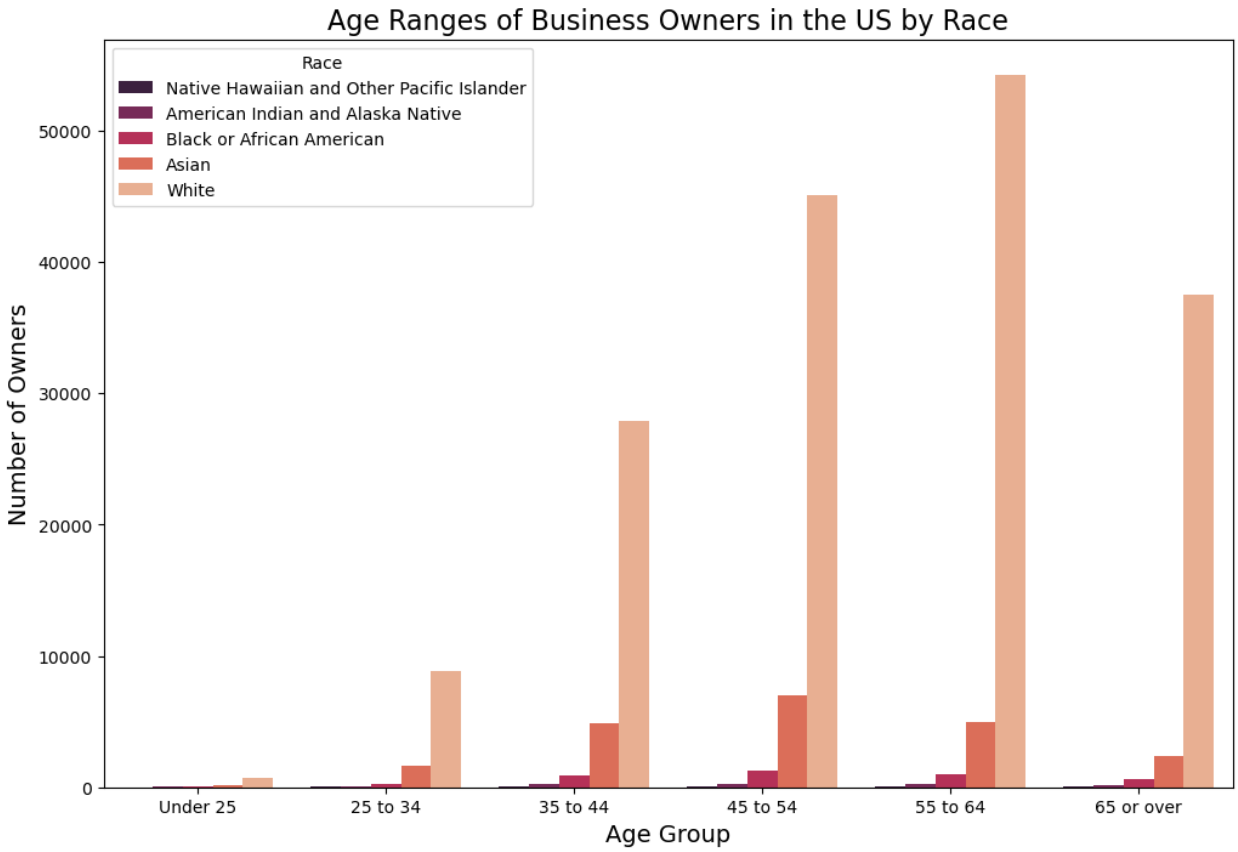
According to the chart below, our assumptions appear to be correct. In all age groups males own approximately twice the number of businesses when compared to females. This difference is a lot larger than initially expected, although the trend appears to be consistent within each age group. Comparing this chart to the chart of all business owners both male and female ownership increase with age until the “65 or older” age group. This confirms that our filtered data is consistent with our primary data.



When filtering our data by race the number of business owners decreases just as it had when filtered by sex. This is also due to the methodology of the data collection as business owners may not have disclosed their race. Our initial expectations were that white business owners would have the highest count, but we didn't have any assumptions regarding other race groups. However, we did assume that the "Native Hawaiian and Other Pacific Islander" and "American Indian and Alaska Native" groups would likely be a lot smaller in population, which would affect the number of business owners when compared to larger groups. Since this data was collected in the US and the white race has had the longest time with legal rights our assumption made sense.

According to the chart below our assumptions were more correct than we had assumed. White business owners significantly outnumber business owners of any other race in every age

group. The difference is so drastic it can almost appear as an anomaly. This can make us question the methodology of the data collection, however more data from other sources would be needed for further analysis. The distribution of business owners across most races also appears to follow the initial trend analyzed in the chart of all business owners. However, it is noted that most of the data appears to be from white business owners, so the age of all business owners' chart is highly dependent on the age distribution of white business owners. Ignoring the number of white business owners, all the other race groups do appear to follow the same general trend of increased business ownership with increased age. Unfortunately, the drastic difference between white business owners and those of other races distracts from being able to deeply analyze the age distributions within other races. This is an area of further study we would like to investigate with additional data focusing on business owners of nonwhite race.



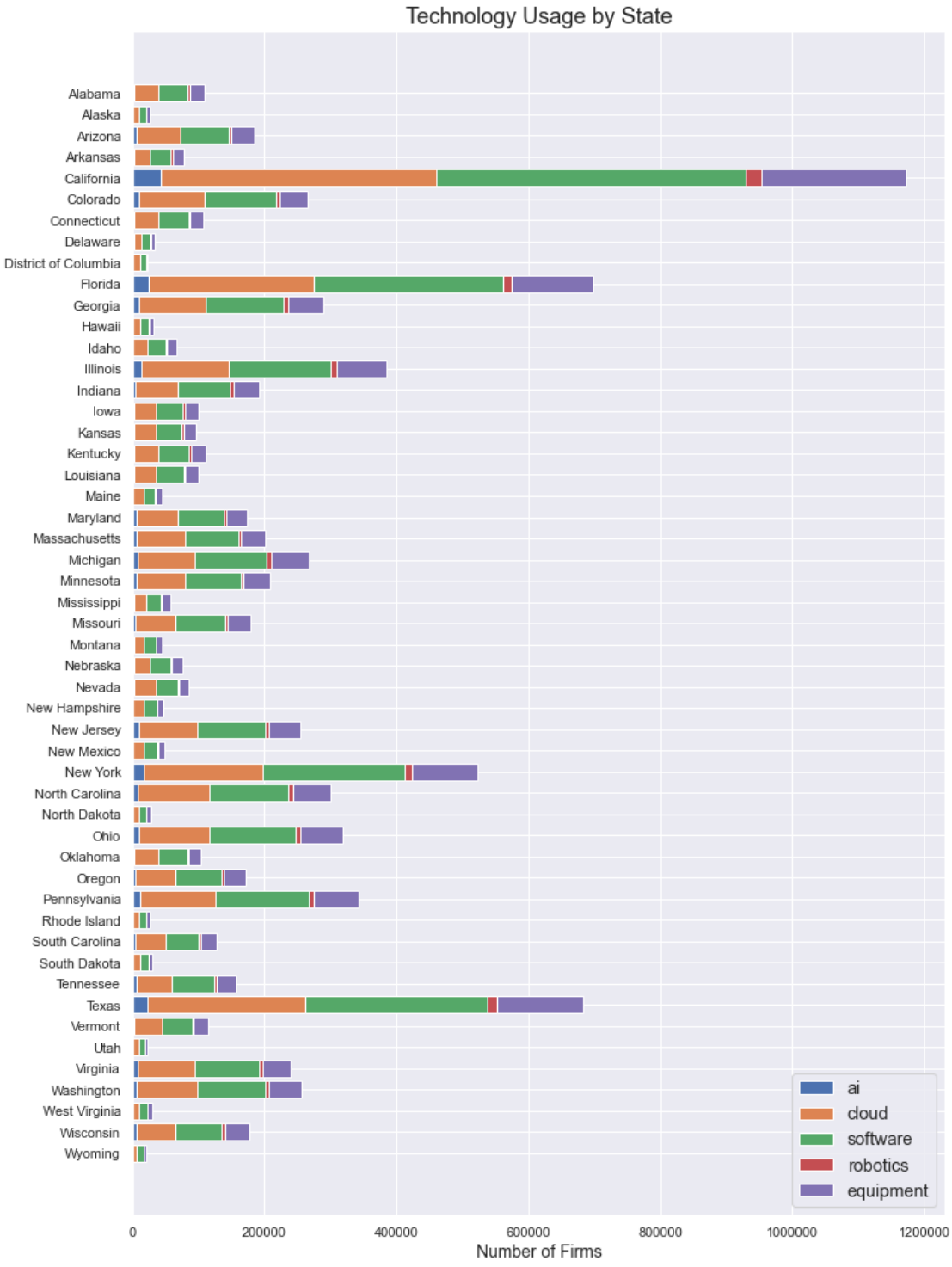
TECHNOLOGY

Everyone would agree technology has had monumental effects throughout most facets of our lives. While not always in a beneficial way, it's impossible to ignore. Without advances in data collection, processing and programming, you wouldn't be reading this analysis! The business world is no different. In our dataset, technology use was tracked by five different categories:

- Artificial Intelligence
- Cloud-based Computing Systems and Applications
- Specialized Software
- Robotics

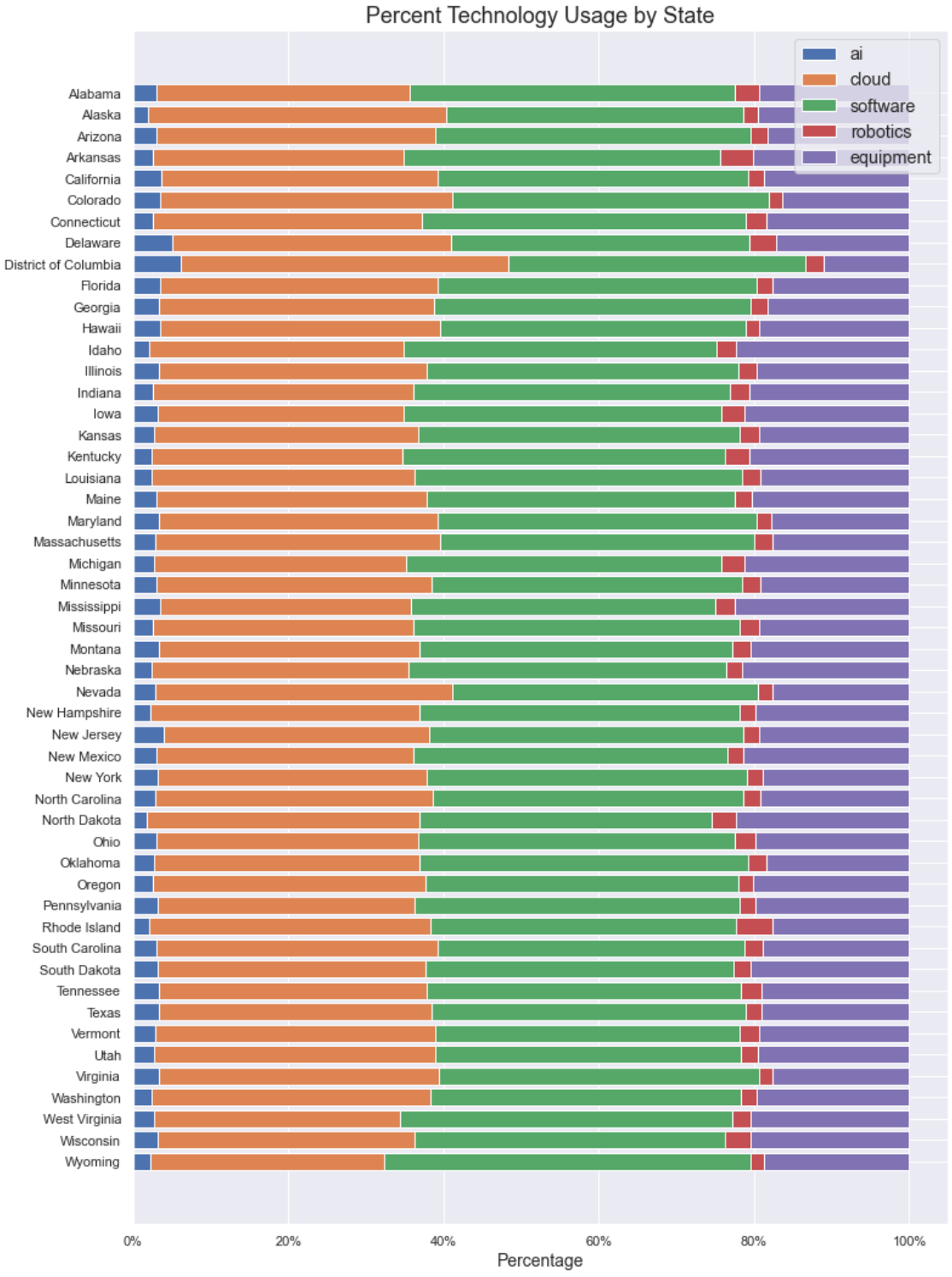
- Specialized Equipment

There were many questions asked in the dataset, such as demographic information, questions about what factors adversely impacted the production and/or utilization of various technologies, reasons behind why businesses had chosen to implement the various technologies, and whether or not the businesses sold products that included the various technologies. The questions we decided to explore were to what extent the businesses used the various technologies, as well as their effects on the workforce, both in terms of number of workers, and skill levels during the years 2016 to 2018. First we have technology usage by state.



Here we can see the density of firms in the various states and can see that cloud services and specialized software are very much the dominant forms of technology over the last few years, with specialized equipment being more of a middleground, with robotics and artificial

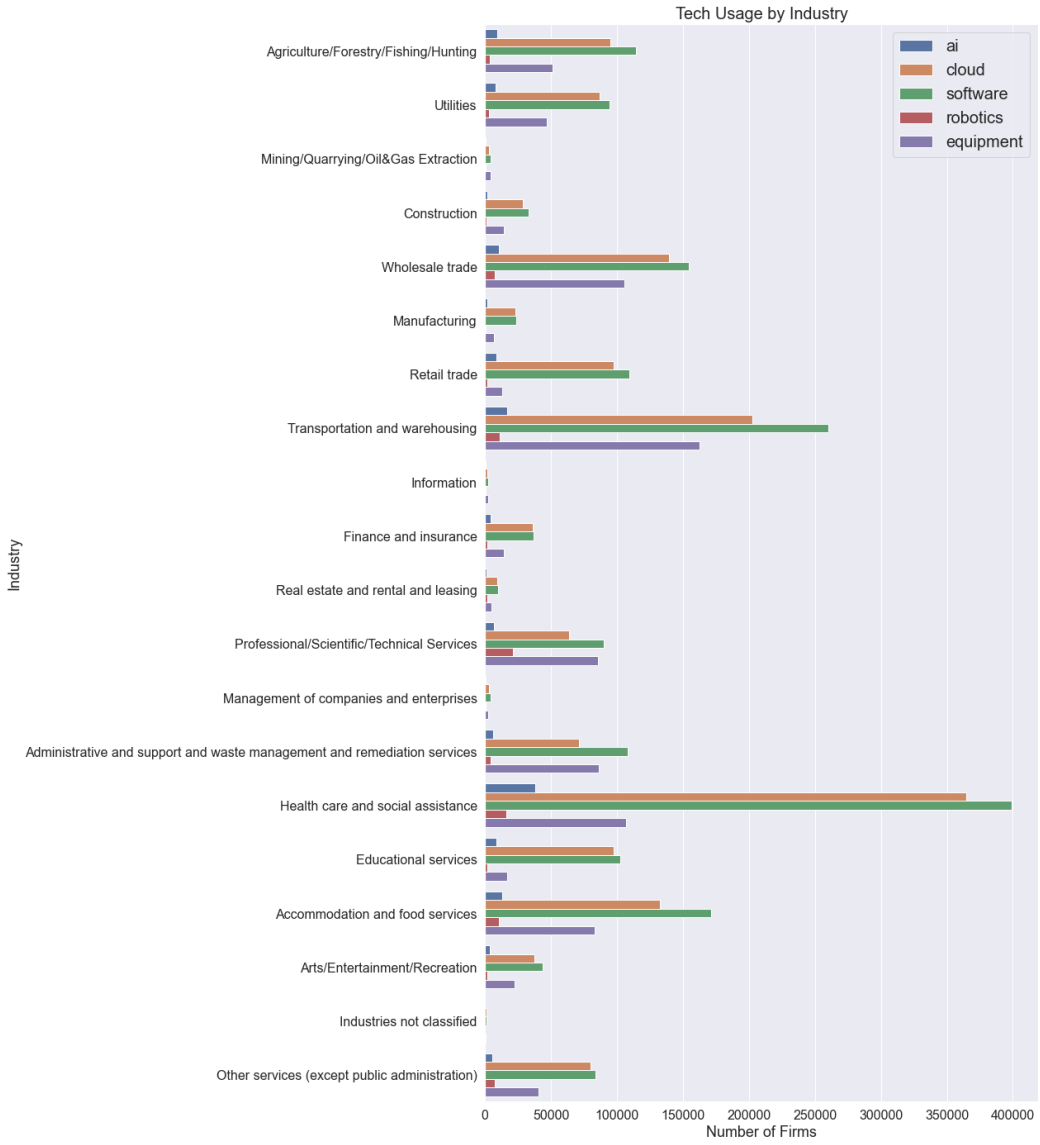
intelligence being much less used. However, while this chart shows for example California has the most firms using artificial intelligence, California also has way more firms than any other state, so a better metric would be to see the technology use in comparison to the total firms in the state. This will give a better idea of what percent of firms used the various technologies.



Here we have a better picture of the technology usage, as now we can see that Washington DC leads the artificial intelligence usage, not California. This falls in line with the

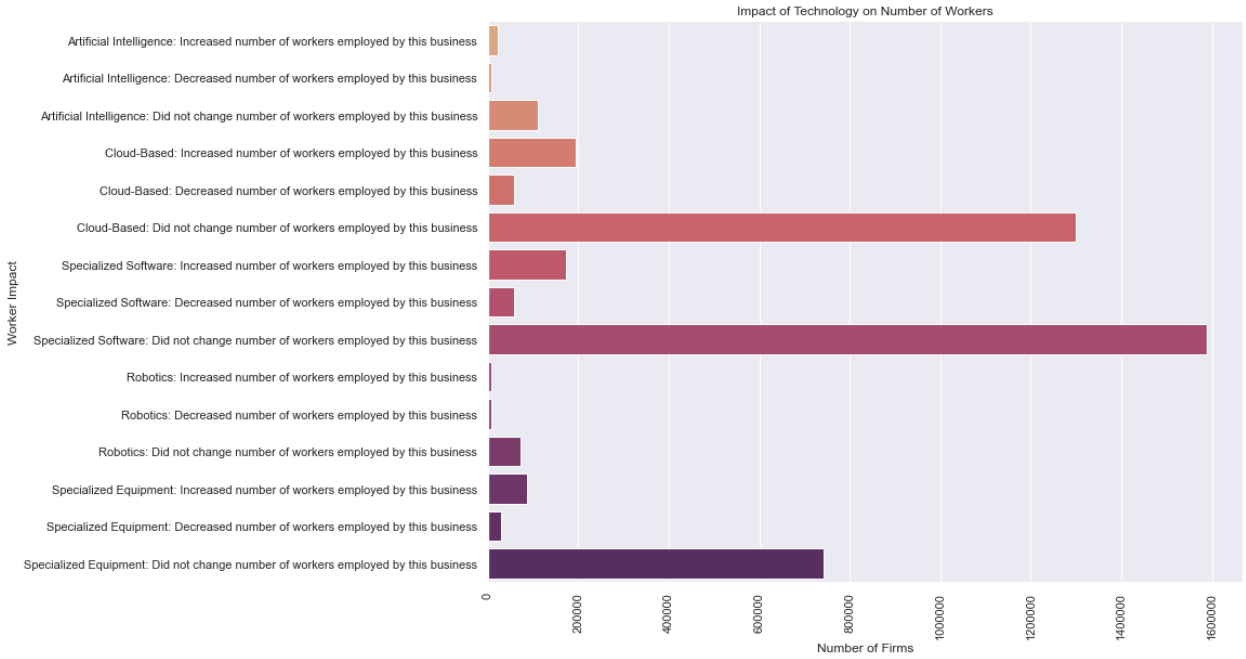
thought that the nation's capital would be on the cutting edge of technology within the central government. Some key takeaways are that robotics seem to be a very niche technology, with the highest density being in Rhode Island, and even then only 4.58% of firms employ it. This could be a function of how useful the technology is in certain industries, but one thing is clear, with the high uses of cloud computing and software, data culture is very integral to many businesses already. It would be interesting to see how these numbers change over time as the internet of things continually expands and more advances in artificial intelligence and automation develop.

We decided to also explore technology usage on the industry level across the country as this is where we would expect to see much more disparity.

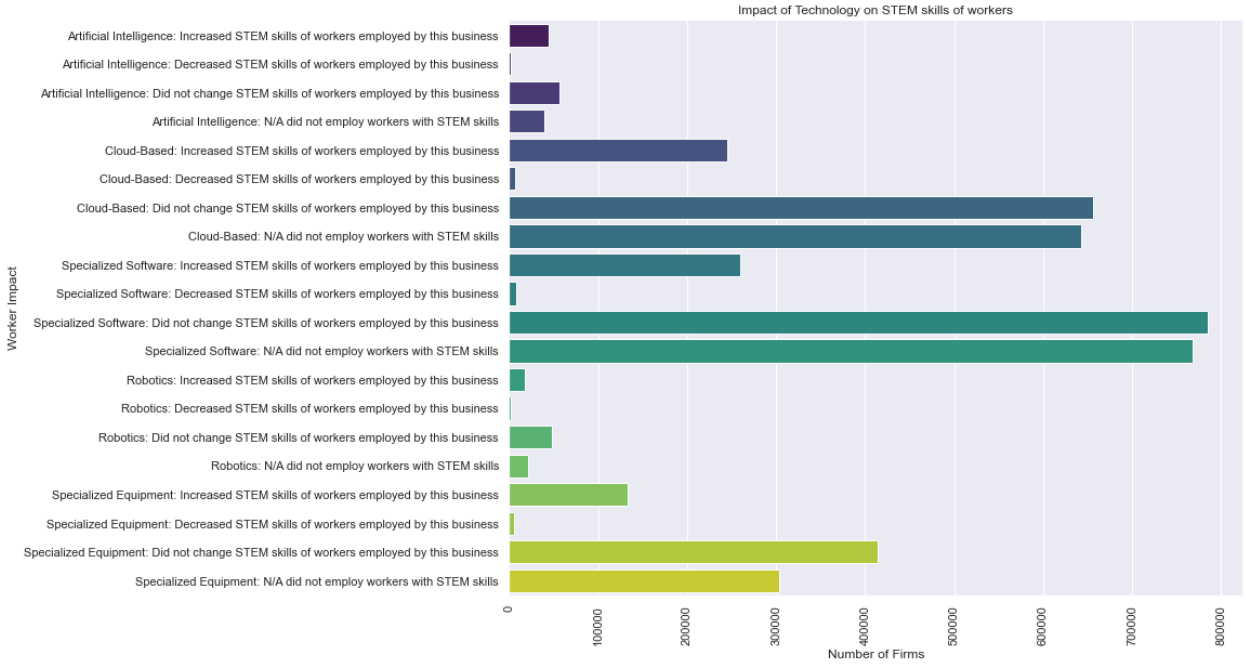


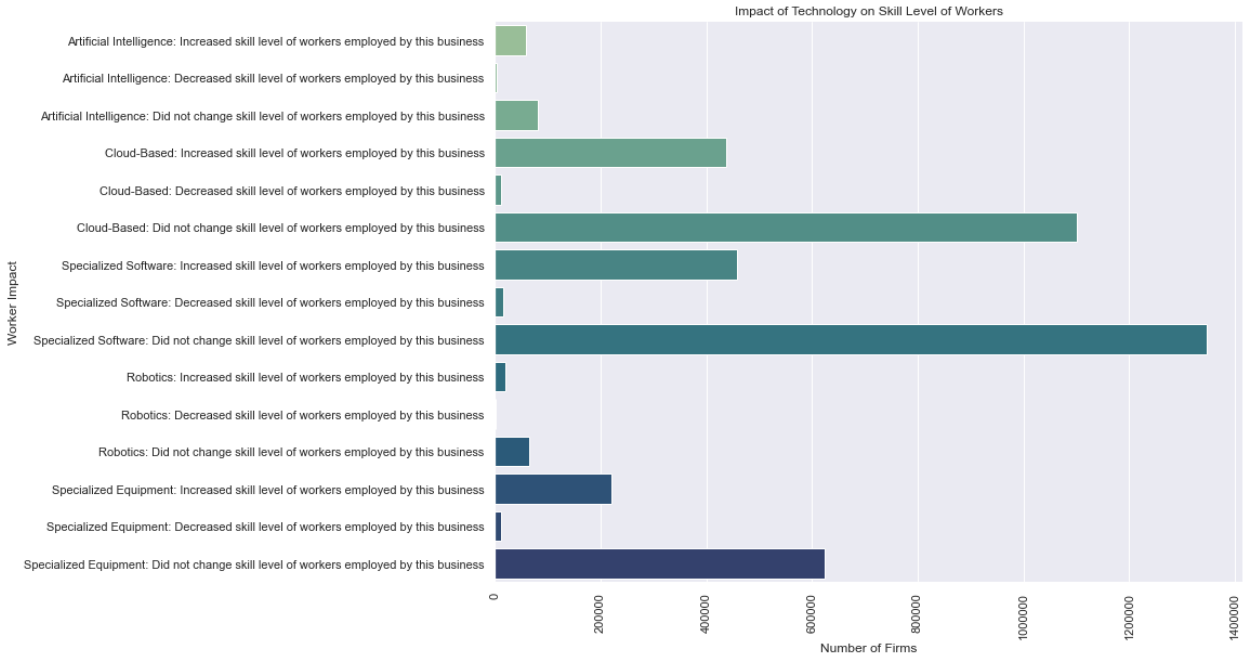
Cloud services and software are being used by just about every industry, with very few being outdone by equipment. This makes perfect sense as mining, quarrying, oil & gas extraction is an industry where equipment should be more dominant, at least until there are advances for automatons to perform these tasks. The healthcare industry appears to be the biggest adoptor of both artificial intelligence and robotics, with transportation and food services not falling far behind. While advanced technology in medicine is seen as an overwhelming positive, there have been concerns over automation in the transportation and food industries potentially hurting job

markets which is another point of interest worth looking into.



At least on an overall level, we can see that even though it does happen on an occasional basis, technology replacing the jobs of workers is not at a level to cause alarm. This is also a good metric to track year over year and ensure that too many jobs aren't being lost. At least in 2019, the effects have either been positive, or had no effect on the workforce count. Next we will look at the skill levels of workers.





Here we see a similar trend that we would expect, technology in almost all cases either didn't affect the skill levels of the workers, or has a positive impact. One thing that's interesting is how technology is increasingly accessible to non technical users. We can see this in software

and applications that allow users to handle tasks like building web sites and apps without needing to know how to write code.

CONCLUSION

Our findings were numerous and varied as we all pursued a different angle while analyzing this dataset.

Looking at the technological data, we found that cloud services and specialized software are widely popular, while artificial intelligence and robotics are not yet. Technology in business tends to be an overall boost to jobs and skills of workers.

Looking at the ownership data, broadly speaking, we found that there are large inequities in the distribution of owner race, and how that affects revenue generation. Through our analysis it was demonstrated that business owners are predominantly white males in older age groups. Asian-owned companies are the only minority-owned companies that can remotely come close to the revenues of white-owned companies, but there is still a big gap. While this is somewhat expected, it also warrants more analysis into why that is the case. These large gaps seen in the data could also be a result of data sampling and more white business owners responding to the ABS survey than minority business owners. There, an effort to incentivize minority business owners to respond and make the sample more representative should be made. Advocating for programs to support minority-business owners, especially black, Native American, Native Hawaiian, and Pacific Islanders, should also be a priority. Identifying the disparity between white business owners and those of other races is the first step in creating systems where other races are able to achieve comparable success in business ownership.