# Census Business Survey Report

Daniel Rose, Cheyenne Peterson, Zach Bonk, Clara McGrath

1/3/2023

# Data Sources

We used the US Census Bureau 2019 Annual Business Survey datasets to complete our analysis. These datasets included: Company Summary, Characteristics of Business, Characteristics of Business Owners and Technology Characteristics of Business. We obtained this data via calling the Annual Business Survey API, which allows you to aggregate the data over the whole United States or pull data by states. The citation for the API can be seen below.

US Census Bureau. (2022, October 28). Annual Business Survey (ABS) APIs. Census.gov. Retrieved December 26, 2022, from https://www.census.gov/data/developers/data-sets/abs.2019.html

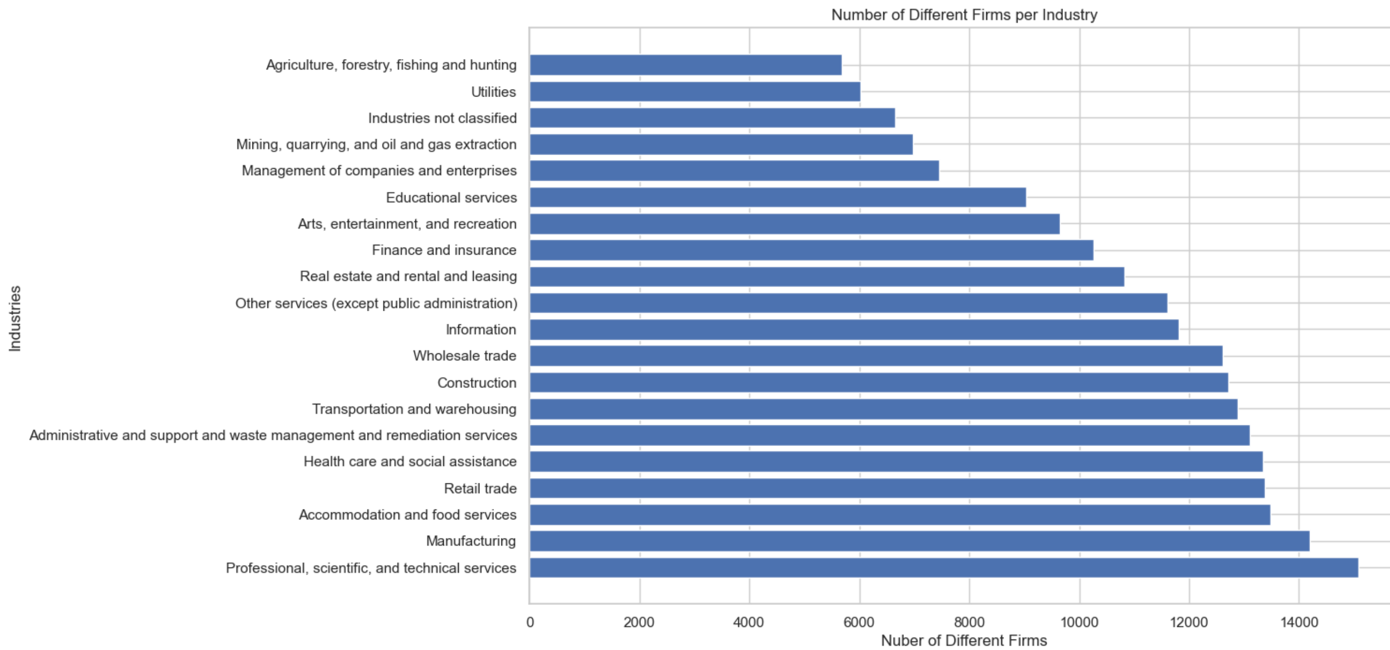
# Initial Questions

- Which states have the most of which industry groups?
- Which industry groups have the most employees?
- Which industry group has the highest payroll?
- Gender by industry group?
    - How much of a difference is there?
    - Pay by gender?
- How does technology use affect salaries?
    - What about revenues?
- How do salaries differ across states depending on technology use?

# Process

## What Industries have the most number of firms and employees?

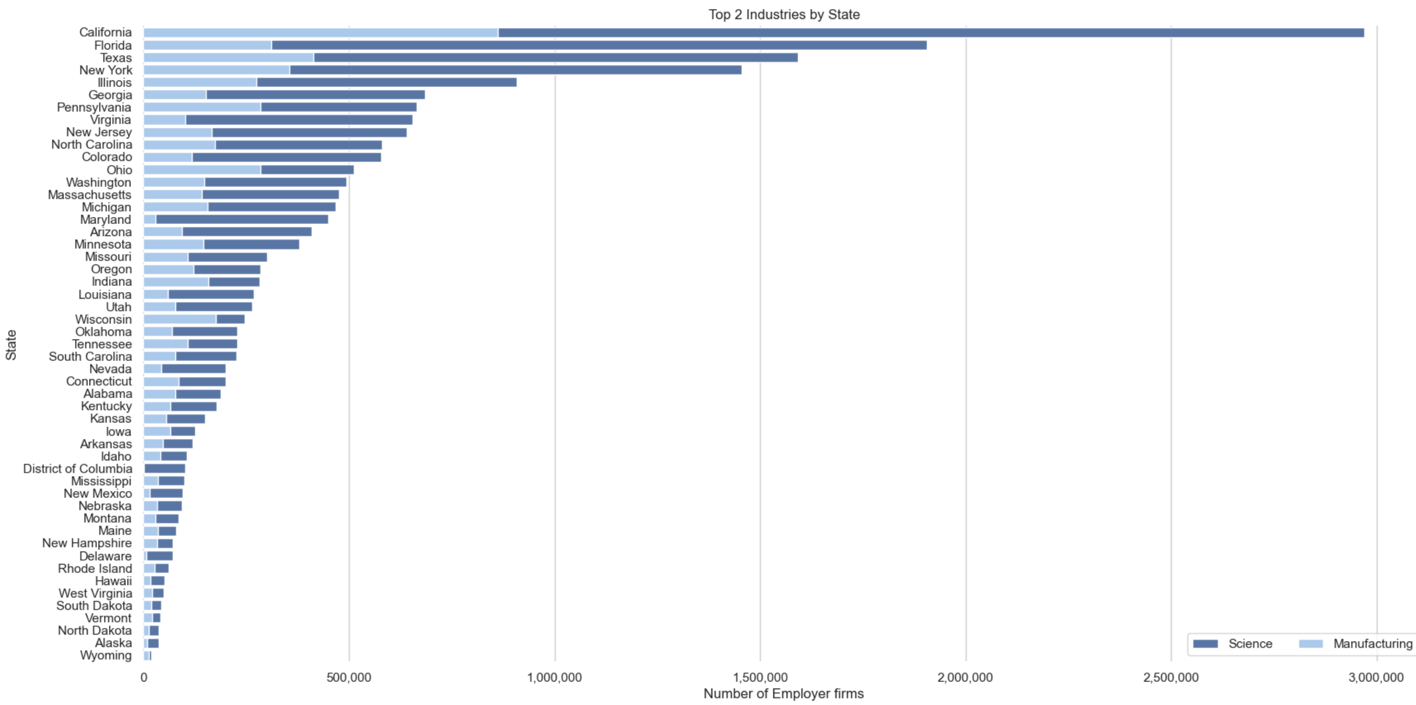
Throughout the United States history the number of firms in different industries has changed over time. Because of this we first started by looking at how many businesses across the United States fell under different Industry categories. This can help us get a baseline of what




kind of jobs people are working and more specifically what industries are the jobs in.

While our graph is fairly simple, it shows us what Industries have the highest and lowest number of firms. From this we are able to gather that professional, scientific, and technical services had the most amount of firms and manufacturing had the second most. We can also see that agriculture, forestry, fishing and hauling had the least amount of firms.

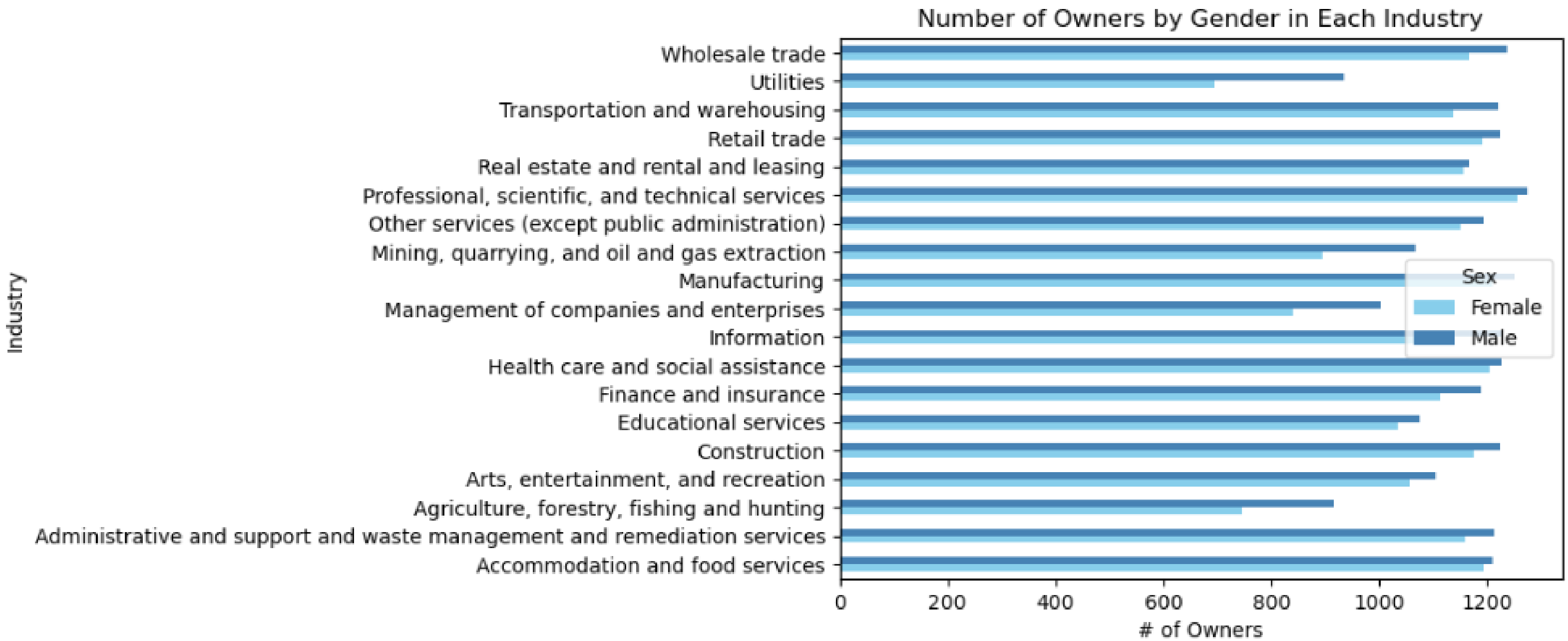
After finding out the amount of firms in the United States by Industries as a whole, we now want to look more specifically at the breakdown by state. Using our data from the previous graph we sum the amount of employees working at the top two industries by amount of firms which are professional, scientific, and technical services, and manufacturing by state This allows us to see the amount of people working in the most popular industries per state.



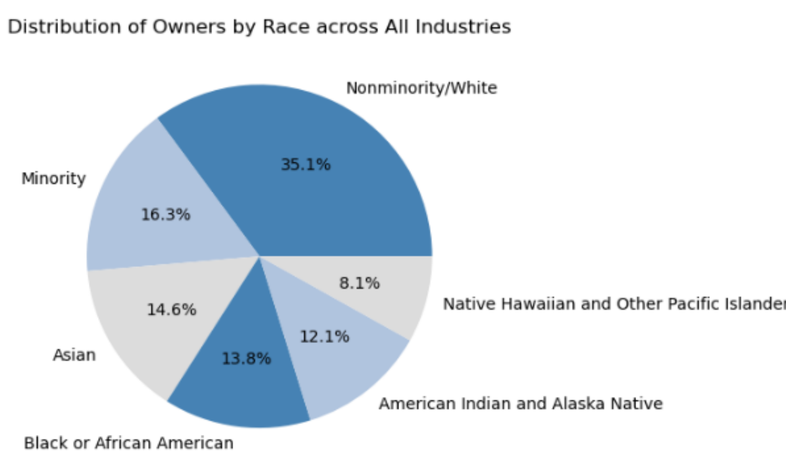

We can see from our graph that the most people work in California, Florida, Texas, and New York. This aligns with those states having the highest populations in the country. We can also see that like our previous graph, professional, scientific, and technical services have more employees for every state than manufacturing. We can also notice specific areas like D.C. where it is almost entirely Professional, scientific, and technical services and very little Manufacturing.

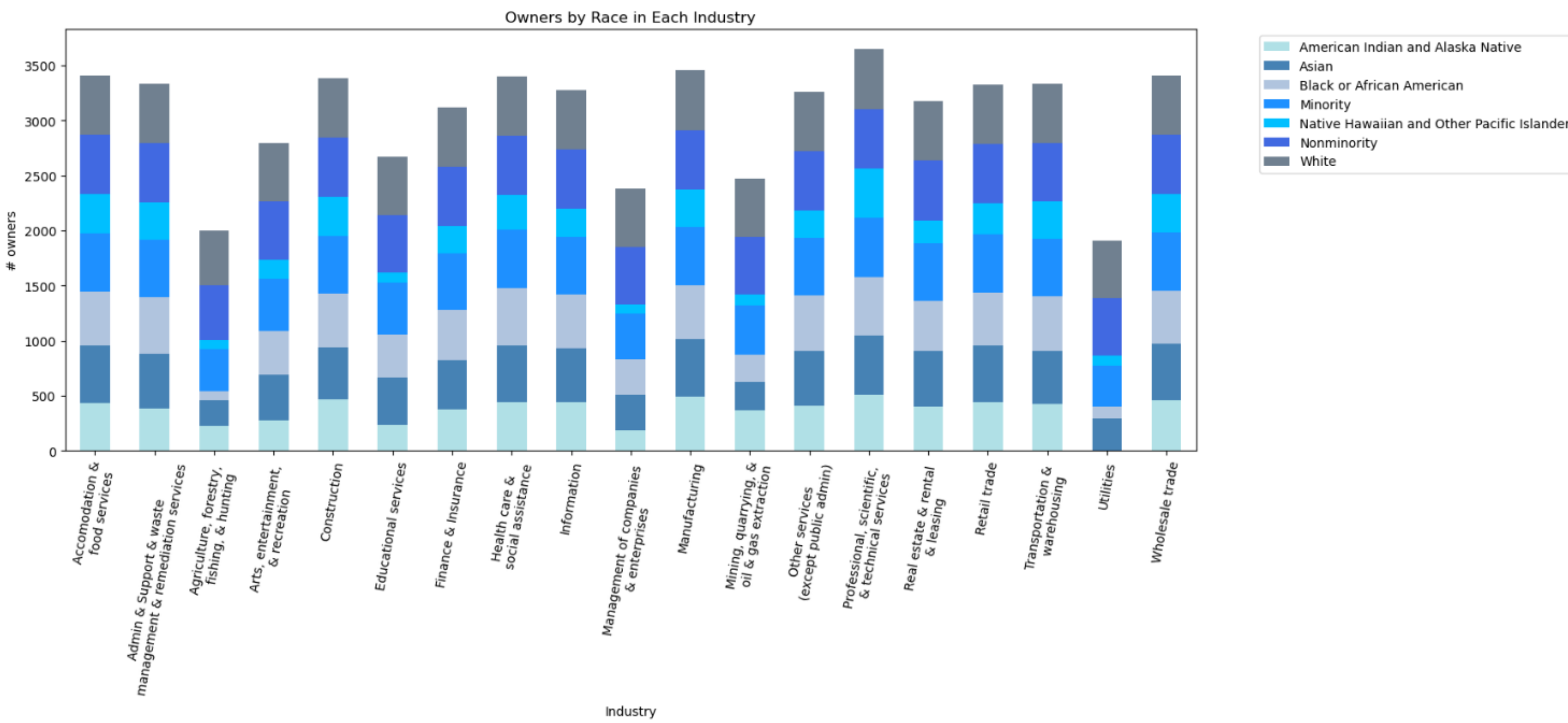
## What is the demographics of different business owners?

One of the characteristics within the business owners' information that we decided to focus on was the gender of business owners. Over many years, the amount of women business owners has grown significantly, so we decided to look at the number of female versus male owners by industry. We decided to use a horizontal bar chart to visualize the male versus female comparison by industry. The industry with the highest number of employer firms is the Professional, scientific, and technical services, which is consistent with the firms and employees data seen earlier. There were a few industries that had notably larger disparities between the number of female owners versus male, including Utilities, Management of companies and enterprises, and Agriculture, forestry, fishing and hunting. However, it is important to keep in mind the context of these results, since we know that the Agriculture, forestry, fishing, and hunting group, for example, had the lowest number of firms.

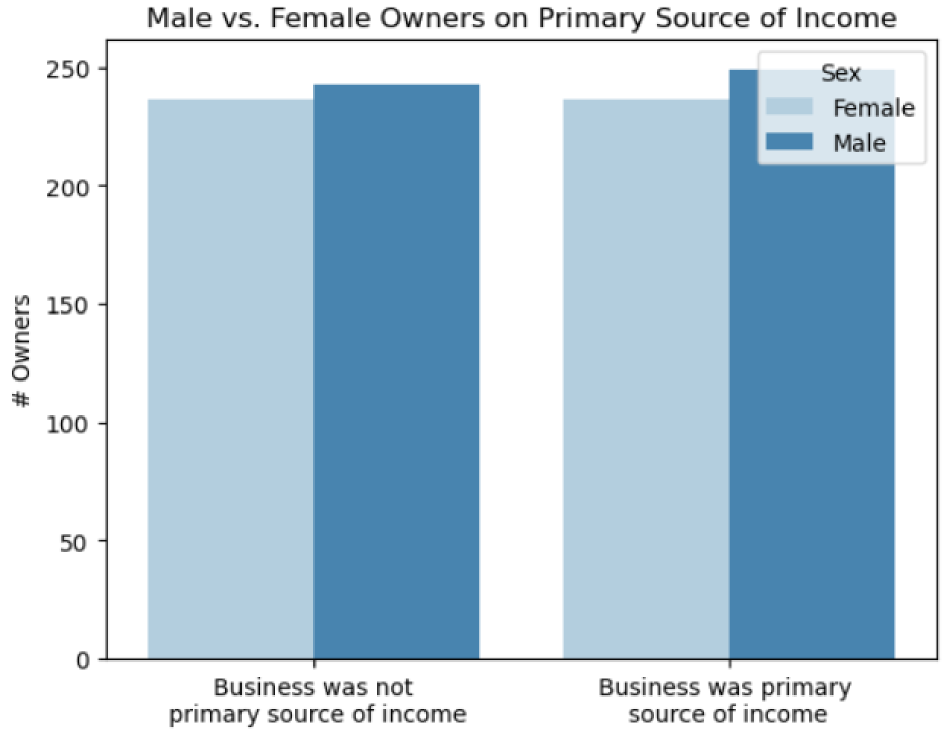



After seeing the breakdown of business owners by sex, we were curious to see what the breakdown might be by racial demographics. We chose to visualize this with a stacked column chart to highlight the ratio of each racial group to the whole industry. The industries that had less equal representation of each race among their owners included Utilities, Agriculture, forestry, fishing and hunting, and Management of companies and enterprises. It is important to note that according to the Census Technical Documentation information, owners had the option of selecting more than one race and is included in each selected race, and the survey requests information from no more than four individuals that own the largest percentages of each business. Because the majority of each industry was nonminority and white, we wanted to see the breakdown of racial groups across all industries. We grouped together the nonminority and white groups so that the pie chart could act as a contextual supplement to the stacked column chart.



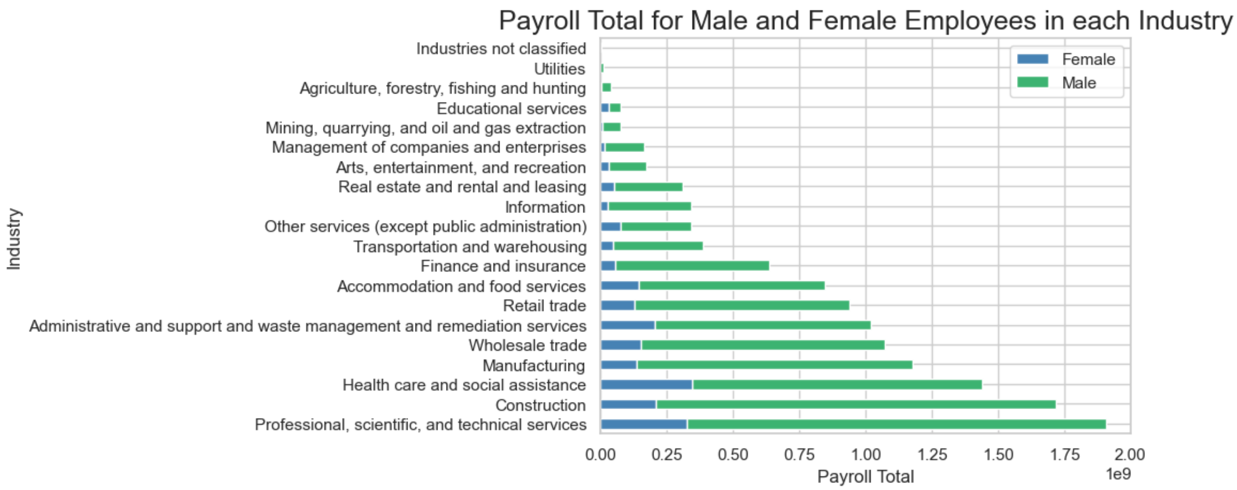



The third owner-related question we chose to investigate involved the survey question that asks whether the business was the owner's primary source of income upon completion of the survey. We chose to use clustered columns to show the comparison between sexes, and the clusters are grouped by answer (yes or no). The number of male owners that reported the business as their primary source of income was slightly higher than the number of female owners. That being said, the first visualization from the owner characteristics data showed that each industry had a higher number of male owners than female, so if that were weighed in this visualization, the results may be slightly different. There was a slim difference, if any, between the number of owners that reported the business as their primary source of income versus those who did not. The results may differ if examined on the industry level, but we did not choose to use industry as a variable because this specific survey question only represents a small fraction of the survey respondents.

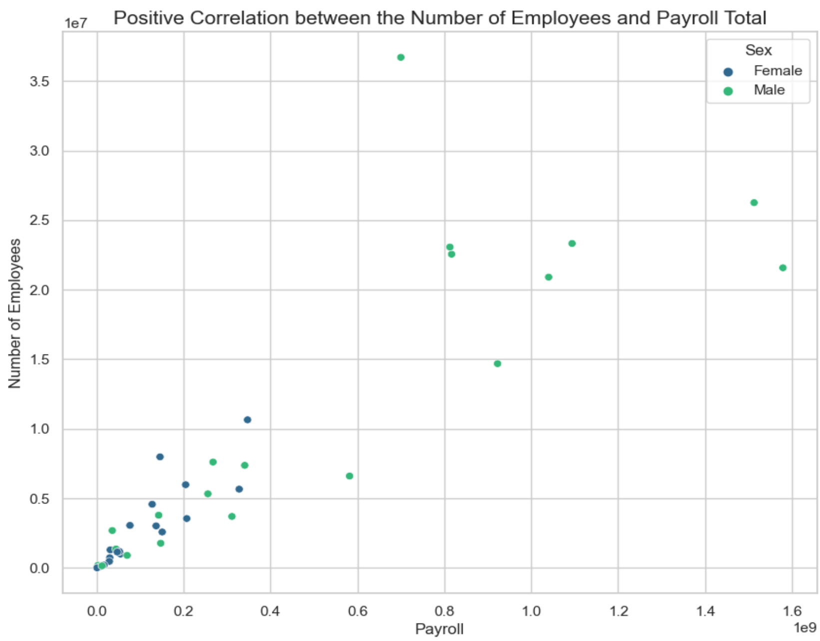



## Gender and payroll by industry group?

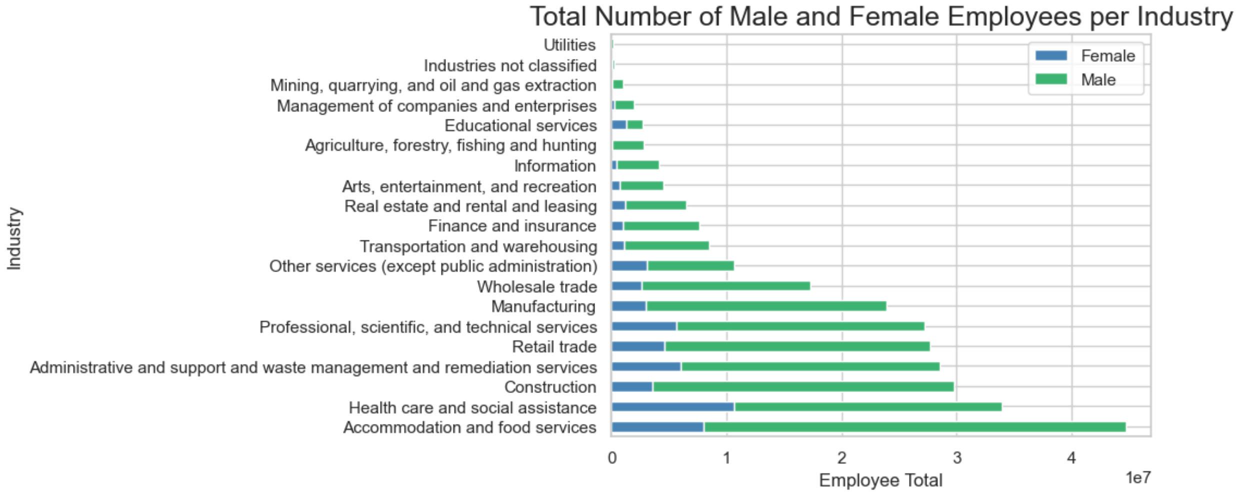
Another focus for our group was to look at the differences of the genders when it came to the different industries. For one of the graphs, I decided to focus on the payroll for the amount of employees based on sex per industry. From the first stacked bar graph, we can see that the industry with the highest payroll amount is the professional, scientific, and technical services sector. The next highest payrolls would be construction and healthcare/social assistance sectors. From the other stacked bar graph, we see that the food service/accomodation industries have the most amount of employees but are not very high when it comes to payroll total. Professional, scientific, and technical services do not have as many employees as quite a few other sectors but it has the highest payroll. With this, we can infer that people in the latter industry are typically paid more when it comes to their salaries than those in other industries. We can use the information from both graphs to compare both the amount of employees and their payrolls. Nonetheless, a scatterplot was created to compare the two amounts directly.




As you can see, there is a strong positive correlation between the amount of employees and the payroll total. This makes sense as the total of the employees payroll increases when the amount of employees increases. With knowledge from the previous graph, we can infer that the values that have a lower amount of employees and higher payroll totals are ones where employees typically have higher salaries. What is surprising in this graph is the drastic difference between the total payroll for males and females. From the graph, none of the females cross the 4 billion threshold while a majority of males are more than double that. You can also see that the number of female employees is also very low when compared to males.




To find out why there would be such a difference, I calculated the amount of males and females in each industry to better understand why males would have a pattern of such higher payrolls and employee rates. What I found in the stacked bar graph was that there tended to be a lot more males in almost all of the industries. With the highest number of females being in the healthcare and social assistance industry. The only industry where the amount of males and females seemed to be equal was the education industry.

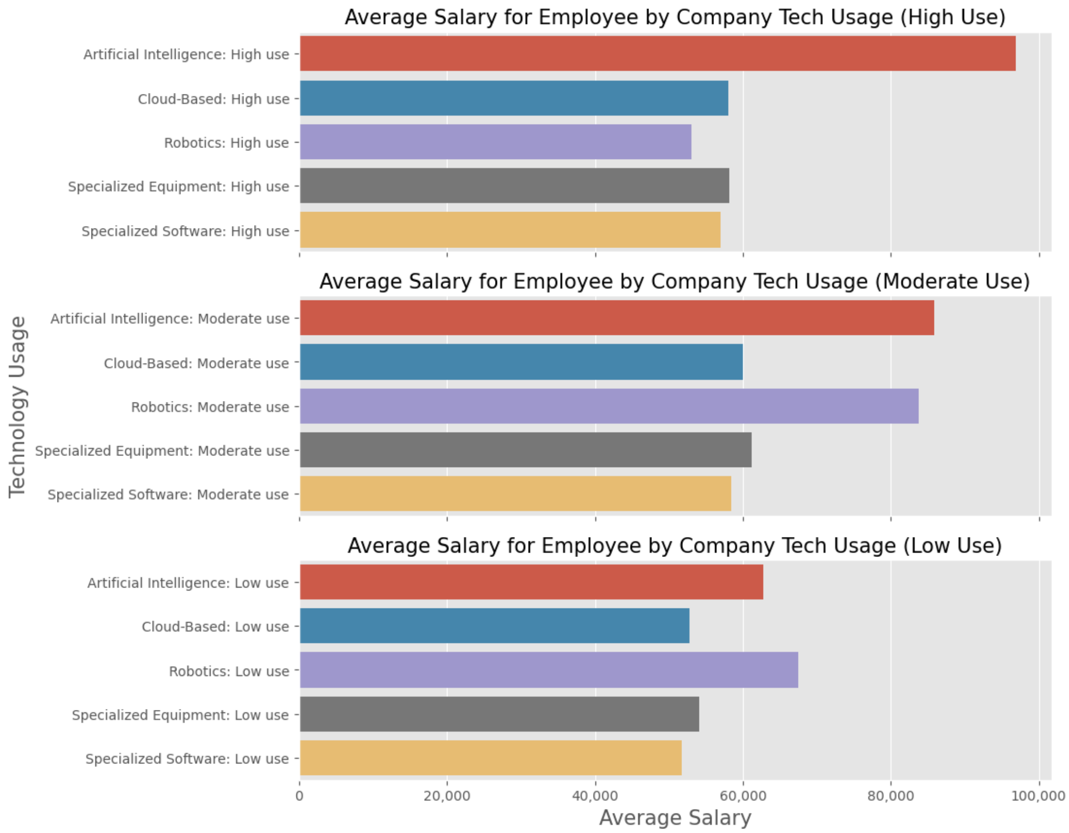



With this knowledge, it can be easy to deduce that the vast difference between payrolls could be due to the amount of employees for the sexes. However, there could be another underlying reason that was not reported in the APIs, like females being paid less than males for instance. But that would be an inference based on guessed reasoning. Another reason that these graphs could be skewed is that in the modern world, there are people who do not identify as male or female. There is no way for us to understand how much this would skew the data without taking other reported genders into consideration. However, by reading from the graphs illustrating the differences between reported genders in US industries, people or companies can take steps to minimize those differences. With notable and great differences like the ones displayed, it would be in the best interest for companies to take note of their employees' demographics.

## How does technology use affect salaries and revenues?

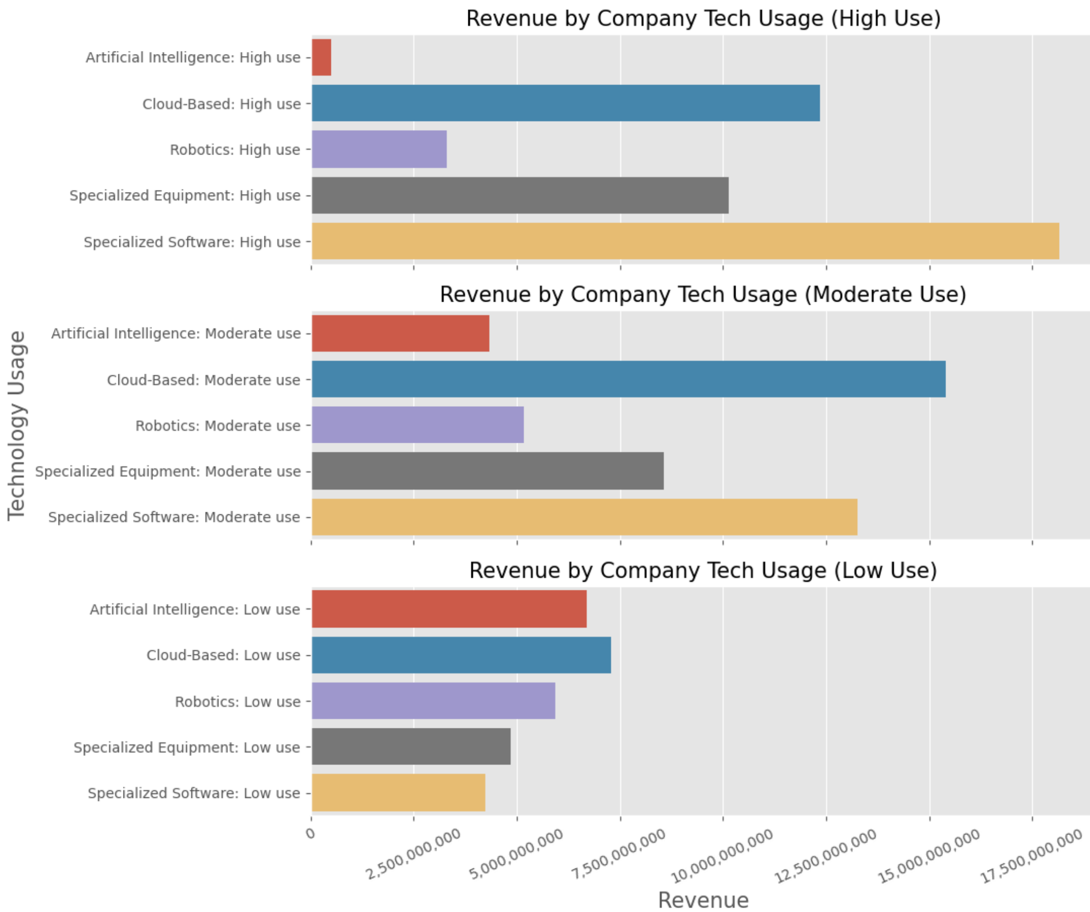
Once we had analyzed the differing industry group, we turned our attention towards questions regarding technology use. Our first question regarding technology use related to its effect on salaries. Before getting into that question though, we needed to decide what technology use factors we wanted to include. There are 5 different technology categories (Artificial Intelligence, Cloud-Based, Specialized Equipment, Specialized Software, and Robotics) with 8 different use-level identifiers. These use-levels are as follows: don't know, did not use, tested but not implemented, low, moderate, high, total use, total reporting. For this particular visual, we decided to only look at the low, moderate, and high use levels of this variable. We decided this since we are more interested in how implementing these technologies at different levels affects the salaries. Low, Moderate, and High represent all of the companies that have implemented those technologies while the other levels of tech use do not.

While we have information on the number of employees and the overall annual payroll, we lack any direct information on salaries. In order to get this measure, we must divide annual payroll by number of employees. Once that column was created, we were ready to make our visualization which can be seen below.



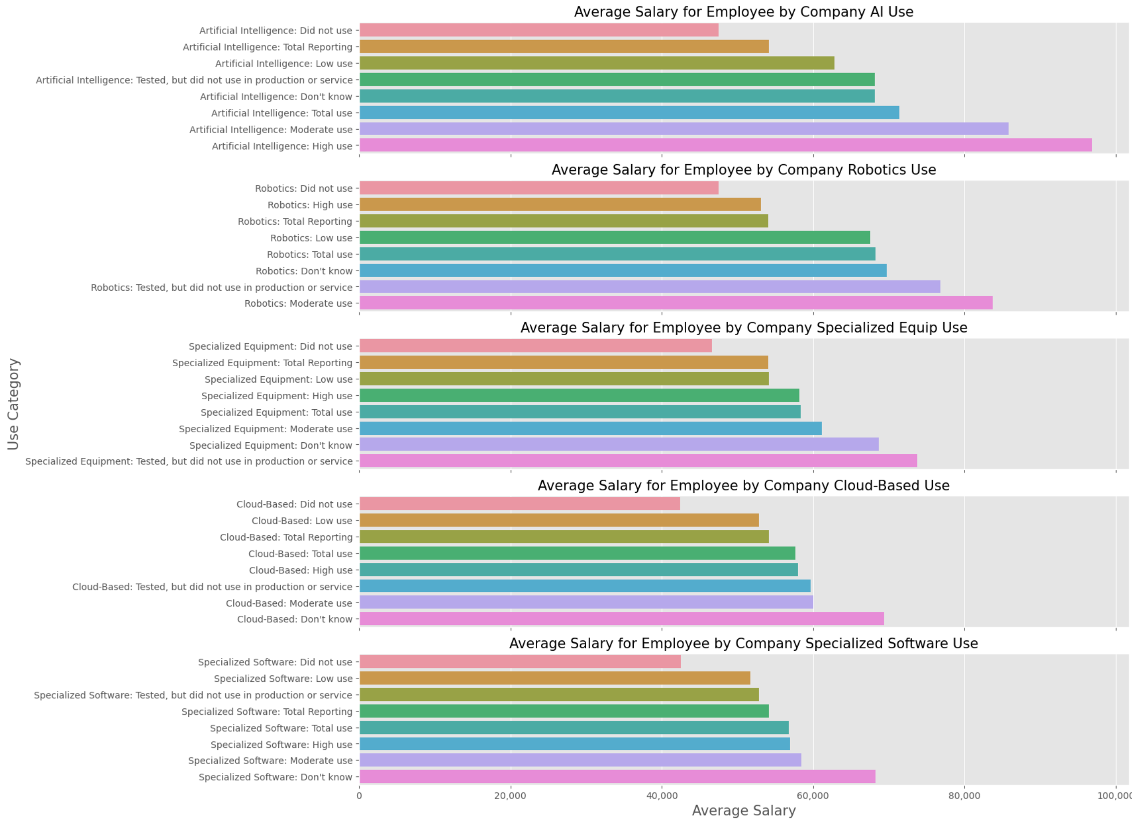

This visual shows the average salary for employees that work at differing usage rates with given technologies. When we discussed this visual, we assumed that as the usage got higher, so would salary. This is true when looking exclusively at AI. However, this is not true for any of the other technologies. In fact, for all four other technologies, moderate usage leads to a higher average salary. Some categories of technology display even crazier relationships. For example, employees that have a low use rate of robotics make on average roughly $15,000 more than employees that have a high use rate. Generally speaking though, it appears that there is one definitive trend, moving from low to moderate usage projects an increase in average salary. Overall it appears that there is a salary gap between low and moderate use, but there is not one between moderate and high use.

Following that visual, we found ourselves with a very similar but slightly different question. How does differing levels of technology use affect revenues for companies? In order to complete this analysis and visual, we used the same process described for the visual above. We again wanted to analyze each of the technologies from low to high use. Additionally, revenue was already found within the data so we did not have to create a new column for this visualization.




This visual shows the effect that differing usage rates of technologies has on the revenue of companies. We expected a similar correlation between tech use and revenues as we did for salaries. Again, this visual had some technologies follow the expected pattern and others that did not. For example, both specialized equipment and software revenues increase as usage increases. However, the exact opposite relationship is true for robotics and AI. Additionally, the high use AI revenue is noticeably smaller than any other revenue metric in the visual. One possible explanation for this is that AI is relatively new and not as well established at a high level yet. Therefore, using this technology at a high level is yet to be profitable, but this idea cannot be confirmed through this data.

We were able to distinguish some patterns that stood out among the data; however, the data wasn't as straightforward as we expected. From an overview perspective the graphs seen when analyzing revenues make sense. As the technology use gets higher, we generally see more revenue. This is however not true for all of the technologies. There are examples such as AI where low use makes the most revenue and high use makes the least. The data seems even less logical when looking at it from a salary perspective. Again, we would expect salaries to increase with usage, since these employees would be doing more intensive work and should be paid accordingly. This as shown by the visual though, is not always true. Outside of AI, the other technology categories are paid slightly higher at the moderate use level. This was not the expected result and in order to analyze this we created the following graphic.

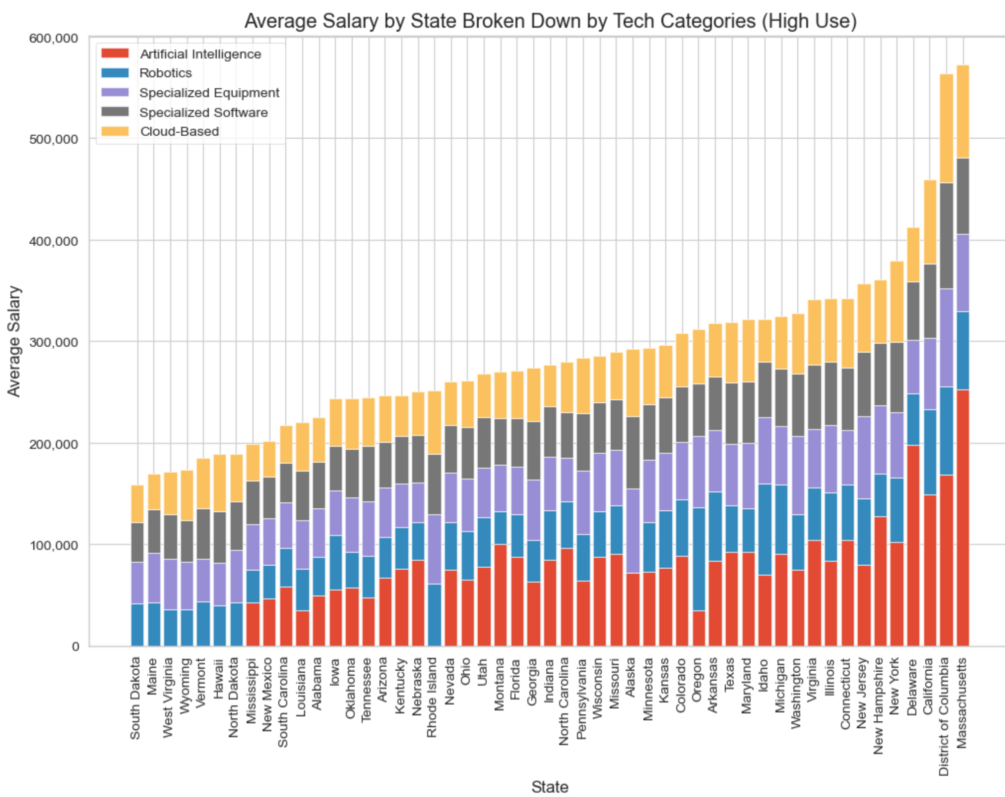



This graphic looks to generalize salaries across the differing usage rates of each technology. This further displays some of the trends that were discussed previously. AI is the only technology where high usage equates to the highest average salary. For each of the other technologies, you actually expect a higher average salary at moderate use rates. This graphic also displays potential issues within the data itself. For example, multiple technologies have the level “don't know” appear as their highest average salary. The level “don’t know” could realistically fall into any of the other levels, but whoever took the survey is not knowledgeable about the company using these technologies. This means that companies that should be in other levels may all be grouped together in a don’t know level, which could potentially skew the distributions.

Overall, I think more analysis would need to be done on this section to specifically answer these questions. Differing levels of technology usage has an impact on both salaries and revenue, but this impact cannot be generalized across the other technology categories. Patterns may exist individually for example with AI, but these patterns aren’t necessarily the same across technology and even appear to be different. In the next steps, we would proceed by zooming out and analyzing salaries and revenues based on using a given technology or not instead of level of technology use. This will allow us to establish a baseline for each technology and at least determine if any usage is better than none.

# Do salaries differ across states due to technology?

The final focus for our group related to technology, was how technology use affects salaries by state. Particularly, we were curious to see the average salary by technology at the high use level for every state. In addition to this, the visual can then provide a quick overview of which states are the best and worst to work with certain technologies or technology in general. To make this visualization, we needed to collect information on tech use, number of employees, and annual payroll for each state within the US. Using the number of employees and the annual payroll, we can create an average salary column as discussed previously in this report.




While the visual makes it look that salaries do differ across states based on the technology, we can't prove that this is due to the technology directly. While it could very well have an effect, there are already established relationships like population, cost of living, type of work, etc that also affect salary levels.

Nevertheless, there are some things that can definitively be stated about technology use and salary by state. The field of AI appears to be the most lucrative from a salary standpoint, but it is also the only technology that isn't commonly found within every state. Additionally, if you are looking to potentially get into any of these fields it is recommended to target places like New York, Delaware, California, DC, and Massachusetts. Conversely, we recommend avoiding states like South Dakota, Maine, West Virginia, Mississippi, and New Mexico.

# Conclusion

By focusing on industries on both the state and national level for the United States, we were able to answer a number of questions. We found that the states with the higher numbers of industries were the ones with the largest populations across the US like California, Florida, and Texas. Furthermore, these states typically had more businesses within the professional, scientific, and technical services, and manufacturing industries. However, it is typically service industries or ones that help people like food services/accommodation, healthcare, and retail that seem to have the higher amounts of employees. By looking more in depth into the businesses within these industries, we discovered significant differences in the demographics. For example, in almost every industry, there are significantly more male than female employees. This causes a huge skew when it comes to the payroll amounts as it seems that males are consistently being paid more than females. However, it's hard to infer whether that's because of the difference in the amount of employees or if females are generally paid less. Yet, when it comes to the business owners, the visualizations do not seem to demonstrate such harsh inequality when it comes to sex and race. Removing or combining the White and Nonminority racial groups gives a better visual breakdown. Utilities, management of companies and enterprises, and agriculture, forestry, fishing and hunting were the industries with the most differences when it came to owner demographics.

Lastly, our group wanted to figure out the answer to technology use and its role within salaries and revenues. To begin, we looked at how the average salary of an employee changed by the rate of usage for each of the five main technologies. Doing this we were able to distinguish a generalized trend of moving from low use to moderate use resulting in higher salaries. However, the same was not true moving from moderate to high use, where we actually see decreases for many of the technologies. These decreases are usually very small and indicate potentially that there is no significance difference among salaries of moderate use to high use technology employees. We then did the same thing again, but instead analyzed revenues. When analyzing this relationship we found no general trend. Some technologies generated more revenue at low usage, while some were better at high use. Overall, this relationship did not appear statistically meaningful. Our final technology analysis, then involved looking at salaries across US states for all high use technologies. While this visual did show definitive differences in salaries across states, this cannot be attributed fully to the tech factors due to underlying relationships between states and salaries that already exist. Instead, this visual provides a good insight to someone who may be looking at the best or worst states to work in related to the technology fields.