Carol Lopez, Rachel Walter, Zeeshan Pervaiz, Finn McSweeney

Group 1’s Project Report on the US Census Bureau's Annual Business Survey (2019)

Introduction

The Annual Business Survey (ABS) is a survey conducted by the U.S. Census Bureau to collect information on the economic and demographic characteristics of businesses and business owners in the United States. The ABS includes data on employer businesses, self-employed entrepreneurs, and microbusinesses, including select data on technology use and production from all sampled businesses. It provides information on business characteristics such as sector, receipts, and employment size, as well as owner characteristics such as sex, ethnicity, race, and veteran status. The ABS is conducted annually and covers the nation, states, metro areas, counties, economic places, and urban and rural areas. Data from the ABS is available for the years 2018, 2019, 2020, and 2021.

Our task for this project was to review 2019 US Census data, access it through the ABS API, and draw conclusions based on our analysis. We used the ABS API to gain a better understanding of the technology adoption and use patterns among small businesses in the US and to identify any trends or disparities in business practices across different industries. We aimed to address several questions through our analysis:

- What is the average salary of employees by state?
- What percentage of workers are male and what percentage are female? How does that compare to the number of male versus female business owners?
- What is the educational background of business owners? Does more education make it more likely for you to own a business?
- What technologies are mostly used by employer firms? What is the salary range for jobs that require varying levels of use of software-based technology? In which industries are individuals who work with Artificial Intelligence likely to earn the highest salaries?

The data used in this analysis is from the 2019 ABS, and the datasets included are:

- **Company Summary:** “Provides data for employer businesses by sector, sex, ethnicity, race, veteran status, years in business, receipts size of firms, and employment size of firms for the U.S., states, and metro areas" (US Census Bureau).
- **Characteristics of Businesses:** “Provides data for respondent employer firms by sector, sex, ethnicity, race, veteran status, years in business, receipts size of firm, and employment size of firm for the U.S., states, and metro areas, including detailed business characteristics" (US Census Bureau).

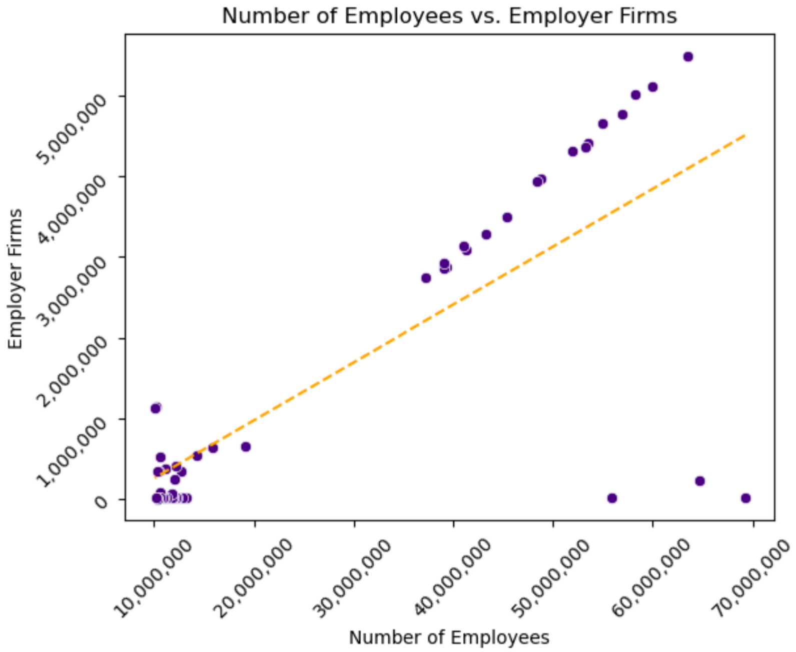
- **Characteristics of Business Owners:** “Provides data for owners of respondent employer firms by sector, sex, ethnicity, race, and veteran status for the U.S., states, and metro areas, including detailed owner characteristics" (US Census Bureau).
- **Technology Characteristics of Businesses:** provides data on technology use and production for Artificial Intelligence, Cloud-Based Computing, Specialized Software, Robotics, and Specialized Equipment technologies data at the U.S. and State level for the reference year 2018.

Visualizations

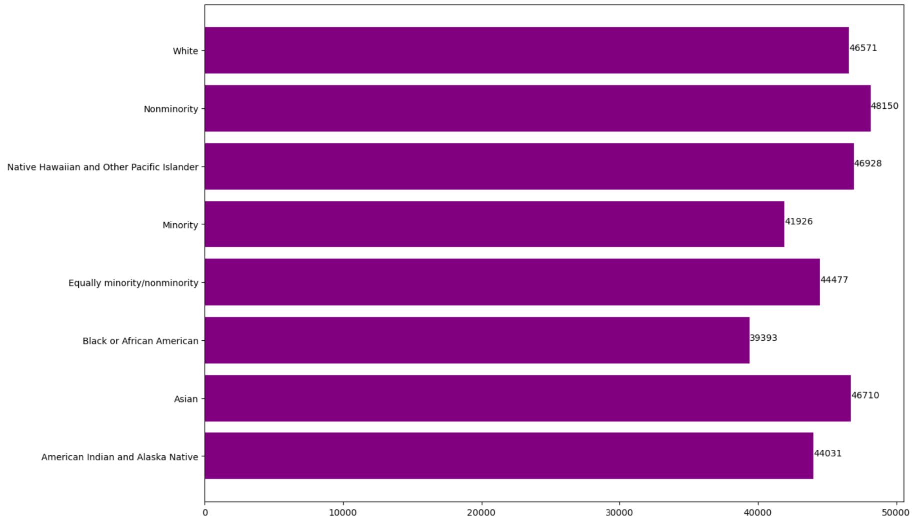
Company Summary

Initial question 1: How strong is the correlation between the number of employees in a company and the amount of employer firms the company has? Initially, it was expected that there would be a very strong correlation and the visualization that was created showed this, however there are some clear outliers.

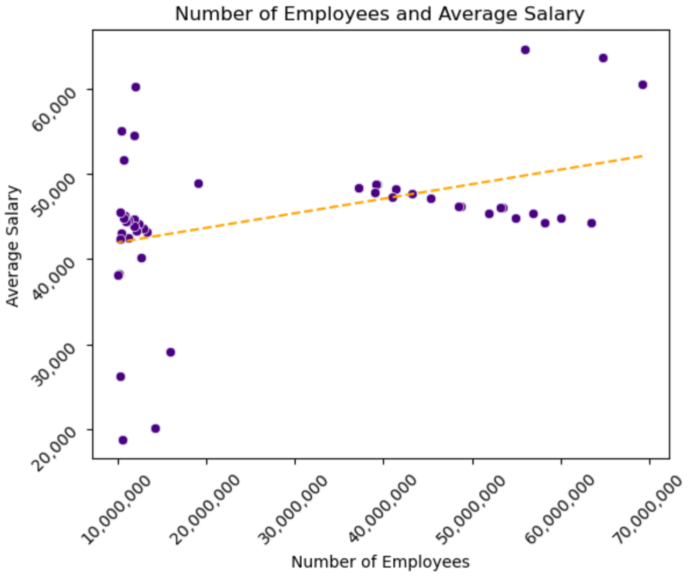
The data that was collected was from the company summary dataset and the top 50 companies in terms of employee count were used for this visualization. The correlation that was calculated from this data was 0.777 however in looking at the graph it is clear to see that the few outliers are lowering this as the majority of the data appears to have a correlation that is very close to one.



The second of the company summary visualizations is intended to answer the question of the average salary by race. A new DataFrame was created to show this and an additional column that calculates the average yearly salary grouped by the race data was used. The entire dataset was not used however, as all rows with null values in ‘PAYANN’ and ‘EMP’ were excluded in order to calculate the averages without any errors.



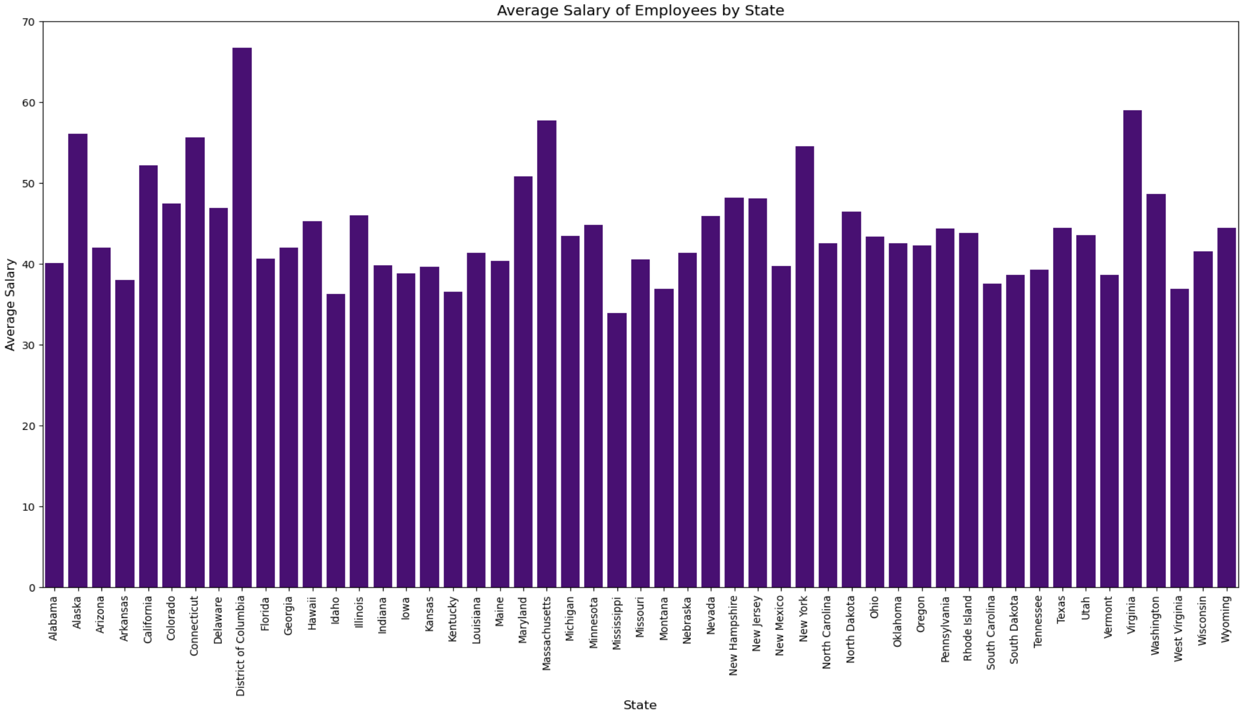
The third visualization for the company summary dataset was created to find a correlation between the number of employees and average salary. The same formula used to get the average salary was used to create this DataFrame and the number of employees in a company was compared to the salary. With only a correlation of .44, there actually was a lot to gain on this graph. The companies with the fewest employees had no correlation between their average salaries; however, as the companies started to get bigger, there was a clear decline in the average salaries.



Carol Lopez, Rachel Walter, Zeeshan Pervaiz, Finn McSweeney

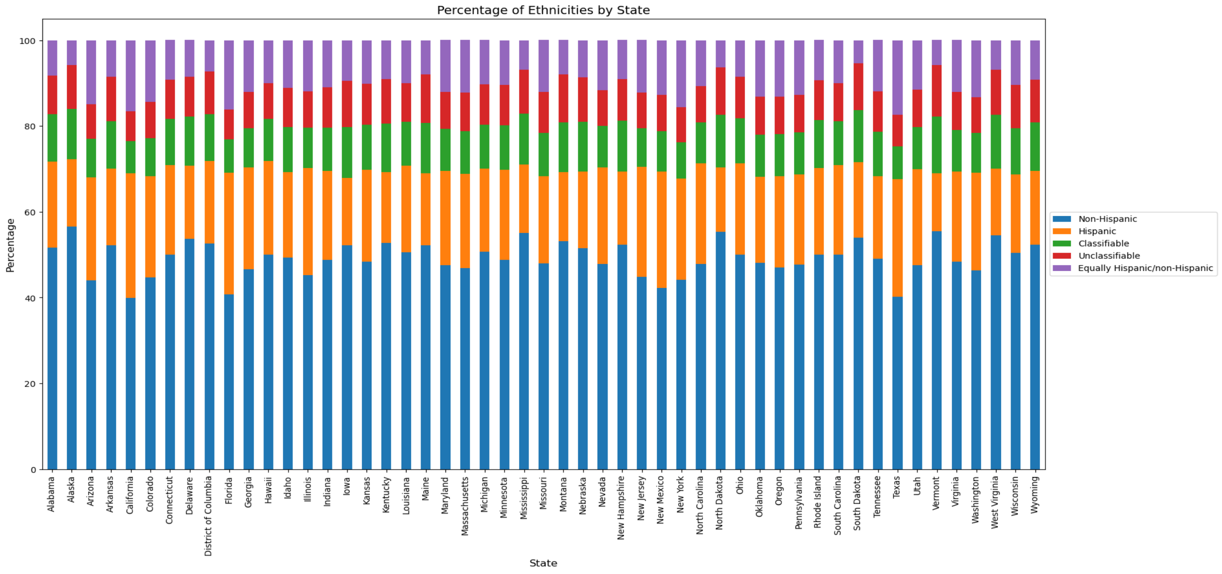
Characteristics of Businesses

What is the average salary of employees by state?
To answer this question we first had to call on the census API to extract the data. Specifically, we used the business summary data table. Once the data was acquired we used the group by function to group each state's salary values together then we took the average of those values leading to a mean salary number for each state. The resulting values are represented in the bar graph below.



What percentage of workers belong to each ethnicity?
To answer this question we again utilized the business summary data table from the census API. We first took the total count of all ethnicities. We then separated each state and divided the count of each ethnicity by the total and multiplied by 100 to find the percentage of each ethnicity per state. The results are shown in the bar graph below.

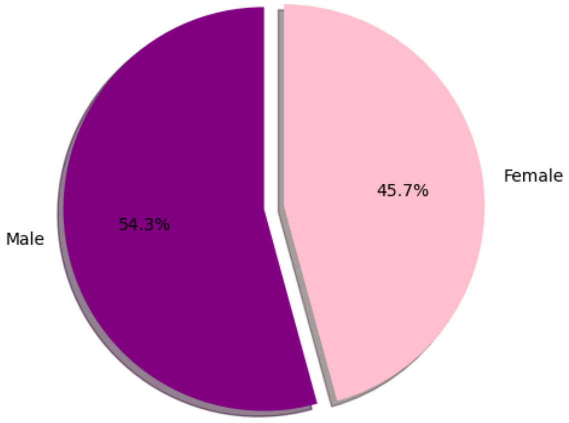
Carol Lopez, Rachel Walter, Zeeshan Pervaiz, Finn McSweeney



What percentage of workers are male/female?

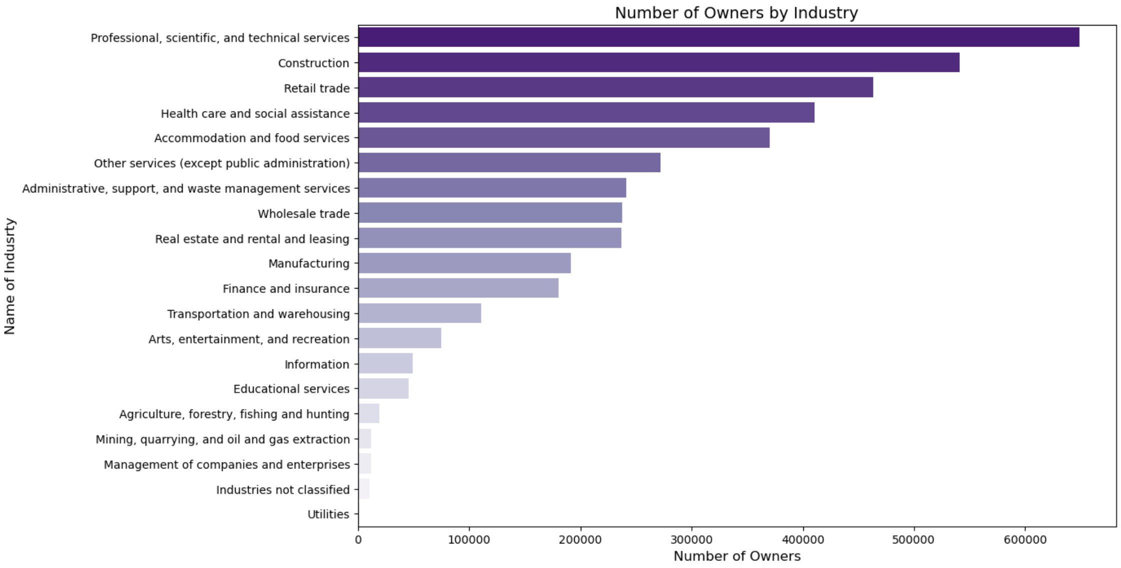
To answer this question we used the business summary data table. First, we grouped the number of male and female workers by state. Then, we had to determine the number of workers for each gender. This required some assumptions to be made since the data wasn’t very clear on how it tallied the values. For example, there were columns for Male, Female, Equally male/female, Classifiable and Unclassifiable. In the end we determined that the values that would determine Male would be the male column and the Equally male/female one and Female would be determined through the sum of Female and Equally male/female. The classifiable and unclassifiable were too ambiguous and as such were excluded from the calculation. Hence our calculation became:

Total_gender_count = Total - Classifiable - Unclassifiable
Male_count = Male + Equally male/female
Female_count = Female + Equally male/female
Percent_male = (Male_count / Total_gender_count) * 100
Percent_female = (Female_count / Total_gender_count) * 100

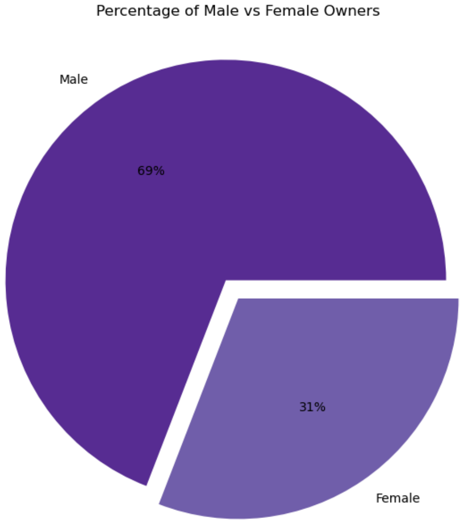


Characteristics of Business Owners

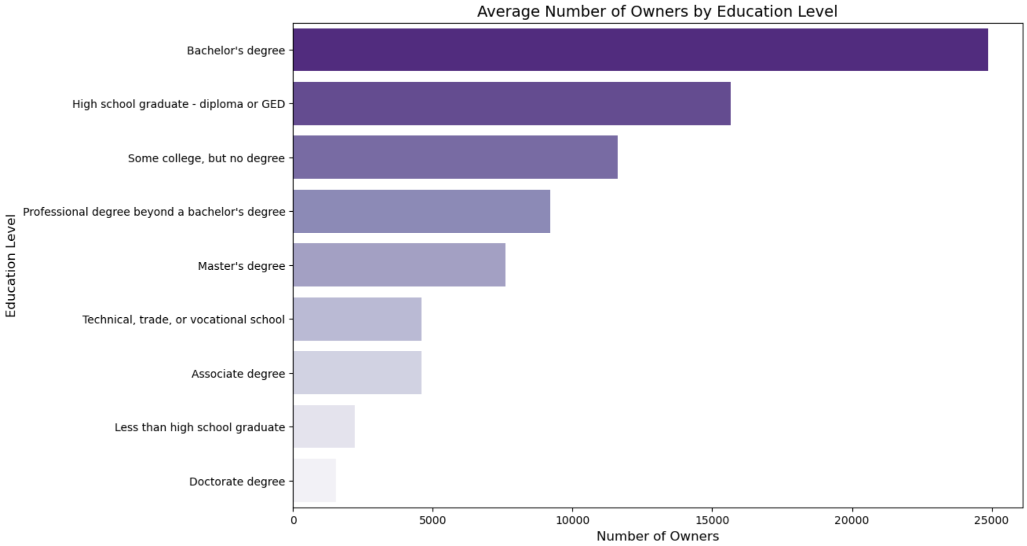
To start off, we used the Characteristics of Business Owners data to look at the breakdown of the number of business owners by industry. To do this, we filtered the data so that we could focus on the individual industries and not the totals for all industries. Then we grouped by industry and summed to get the total for each industry. The industry with the largest number of owners was “Professional, Scientific, and Technical Services.” The US Bureau of Labor Statistics states that this industry includes services like legal advice, payroll, engineering, consulting, and advertising (US Bureau of Labor Statistics). After that the next largest industry was “construction,” then “retail trade,” “healthcare and social assistance,” and “accommodation and food services” rounded out the top five. It makes sense that an industry with such a large array of services would have the largest number of owners, and it is followed by four vital industries. While this question was easily answered, so many different types of businesses fall into the first category so it is a little difficult to get a sense of which types of businesses have the most owners from this data.



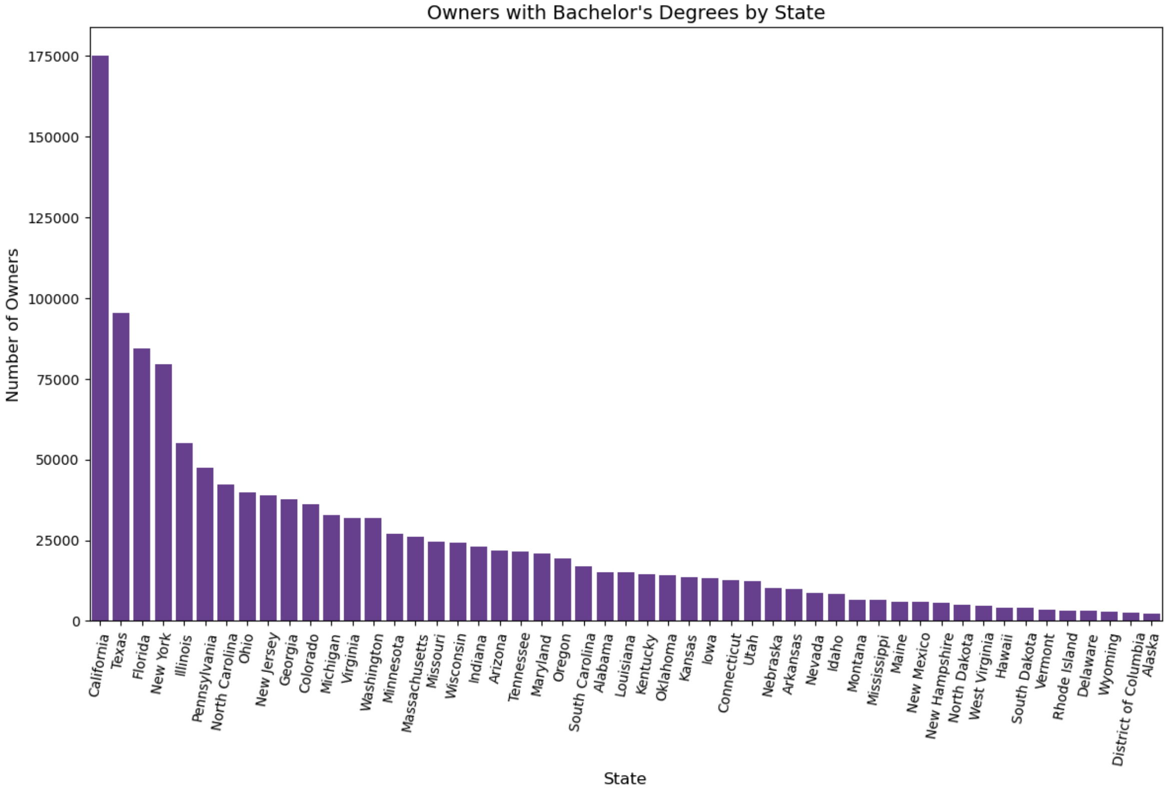
Next we looked at sex, in the previous section we looked at the number of male versus female employees and we wanted to see how those percentages compared to the values for female versus male business owners. To do this, we again filtered the data to focus on the values for the number of male and female business owners and used sum to get the total for each sex. For employees, the data shows that it is not that far from a 50/50 split, male employees made up 54.3% and female employees made up 45.7% of the total number of employees. But with owners, the difference between the two figures is much larger. Male owners make up 69%, as compared to female owners who make up 31% of the total number of owners. While the question was relatively easy to answer, further research would need to be conducted to try and figure out the reason behind it. What is contributing to the smaller number of female business owners and what can be done to improve the situation? The data from subsequent graphs shows us that business owners are more likely to have bachelor’s degrees, but education does not seem to be the problem. Data from the Pew Research Center shows that women are more likely to have a Bachelor's degree than men (Parker 2021). Though, further research could be done on what fields males versus females are more likely to study in college and whether that has an affect on the trajectory of their careers.



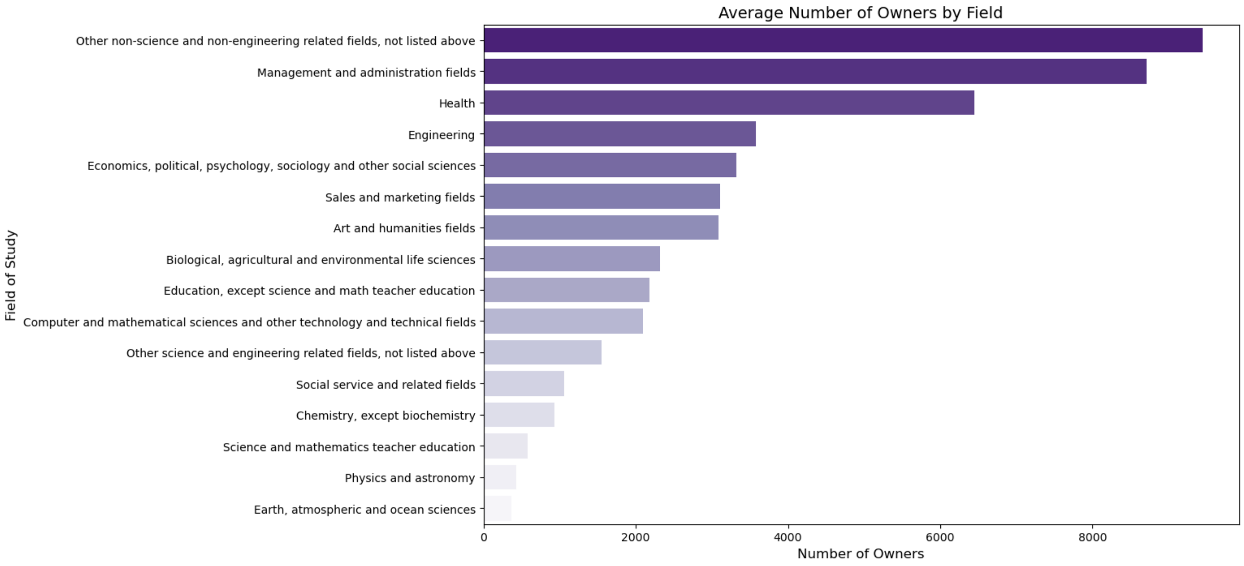
Next we decided to organize the data by state instead of industry. To do this, we made another request but modified it by changing "us" to "state." Of the data available, we were most interested in education so we filtered the data to focus on the number of owners by education level. We grouped by education level and then used the mean function to get the average number of owners for each education level. The results were not too surprising, the education level with the most owners was Bachelor's. Though, next was high school diploma or GED and some college but no degree demonstrating that a degree is not necessary to run your own business.



Since business owners were more likely to have a Bachelor's degree, we next took a look at the number of owners with Bachelor's degrees by state to see if there was a difference at the state level. While there was definitely a big difference between states, California had the highest number with 175,196 and Wyoming had the lowest with 56,1662, the difference seems to be largely caused by the difference in population size between the states. When these values were compared with the population values from the US Population by Zip Code dataset there was a .99 correlation between the two sets of values (US Census Bureau 2017).



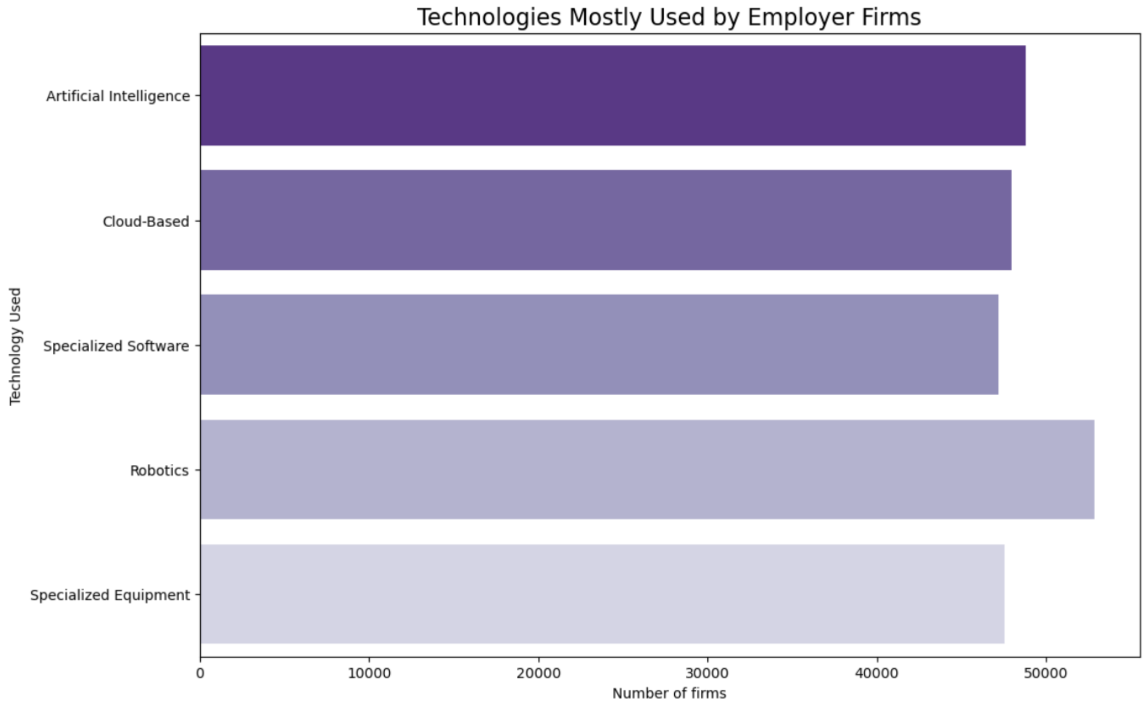
Because population sizes largely accounted for the difference between each state's number of owners with Bachelor's degrees, we wanted to take a look at a different aspect of education. We decided to pivot our attention to what field those with degrees studied. To do this, we again filtered our data organized by state but looked at a different owner characteristic. We then grouped by field and used the mean function to get the average number of owners for each field. The result for the field with the largest number of owners was not very clear since “Other non-science and non-engineering related fields, not listed above” had the highest value. This result suggests that the categories for this question need to be restructured since it is hard to get insights from the data when the “other” group is the largest portion. “Management and administration fields,” “health,” “engineering,” and “economics, political, psychology, sociology, and other social sciences” rounded out the top five fields.



Further research would need to be done to figure out why these fields are most popular with business owners. Is the field they studied directly related to the business they own and the industry they are in? Are these fields more popular because there are more businesses in these fields? Or possibly because there is more growth potential in that field? The idea that it is directly related to the number of businesses in certain fields is likely but more research is needed. To return to the issue of there being less female business owners, since education does not seem to be the problem, it would be interesting to look more closely at the number of women and men in each industry and whether the fields women choose to go into are hurting their chances at owning their own business. For example, teaching is a women dominated field, but most teachers do not own their own business since they most likely teach in a public school. Of course, it might be a totally different issue that is causing the difference between the number of male and female business owners, but more research is definitely required.

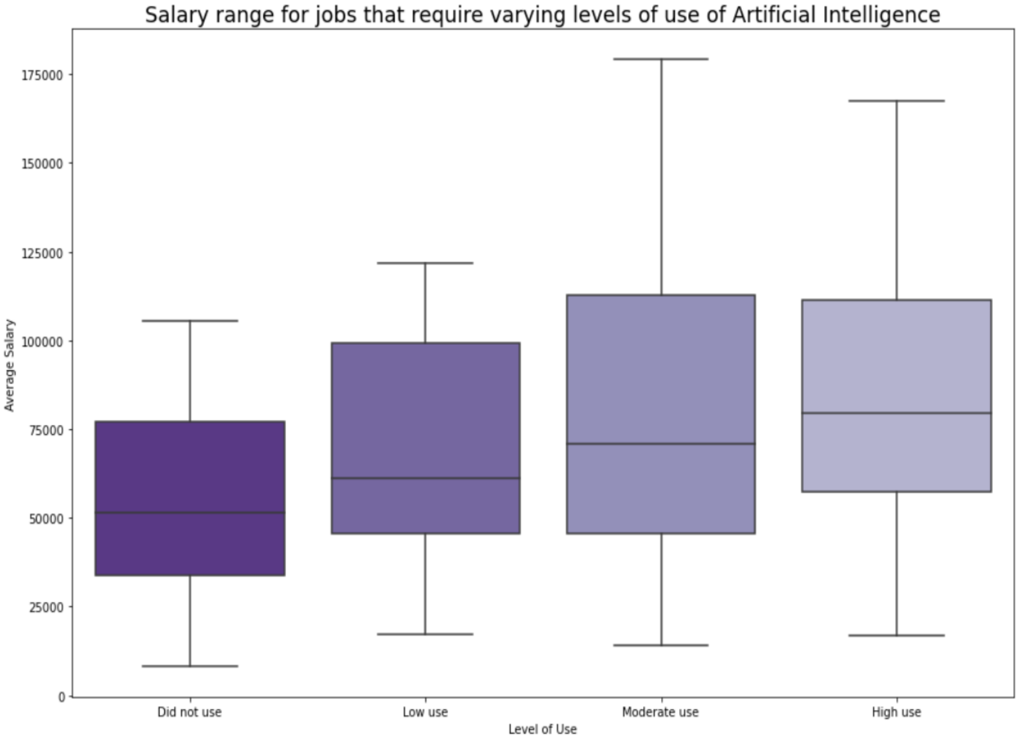
Technology Characteristics of Businesses

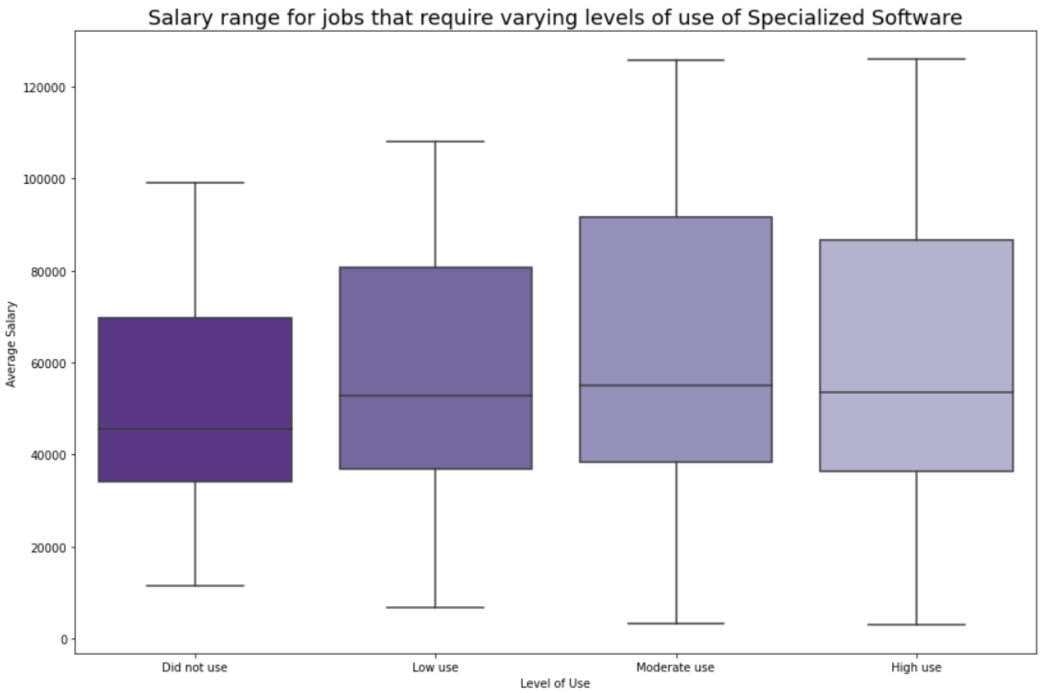
The main ETL steps taken for our principal DataFrame, named techuse, included promoting the second row to column headers, changing data type to numeric when it applied, changing column names, splitting TECHUSE column by the semicolon delimiter to obtain the Technology Used and Level of Use, and filtering out the rows where Level of Use equals ‘Total use’, ‘Total reporting’, and ‘Don’t know’. To understand which technologies are most commonly used by employer firms, we created a new DataFrame called *firmtech* that includes relevant columns from the techuse DataFrame. This includes the columns Industry, Technology used, Level of Use, and Number of firms. Based on the data from the dataset Technology Characteristics of Business, which included surveys from 2018, we can see that robotics is the technology most commonly used by firms. Among software-based technologies, artificial intelligence (AI) is the most popular, followed by cloud-based and specialized software. It would be interesting to see how this distribution may have changed with the increased availability of cloud-based software and services with embedded AI, which make it easier for non-experts to use AI. (Cloud-Based Services*)*

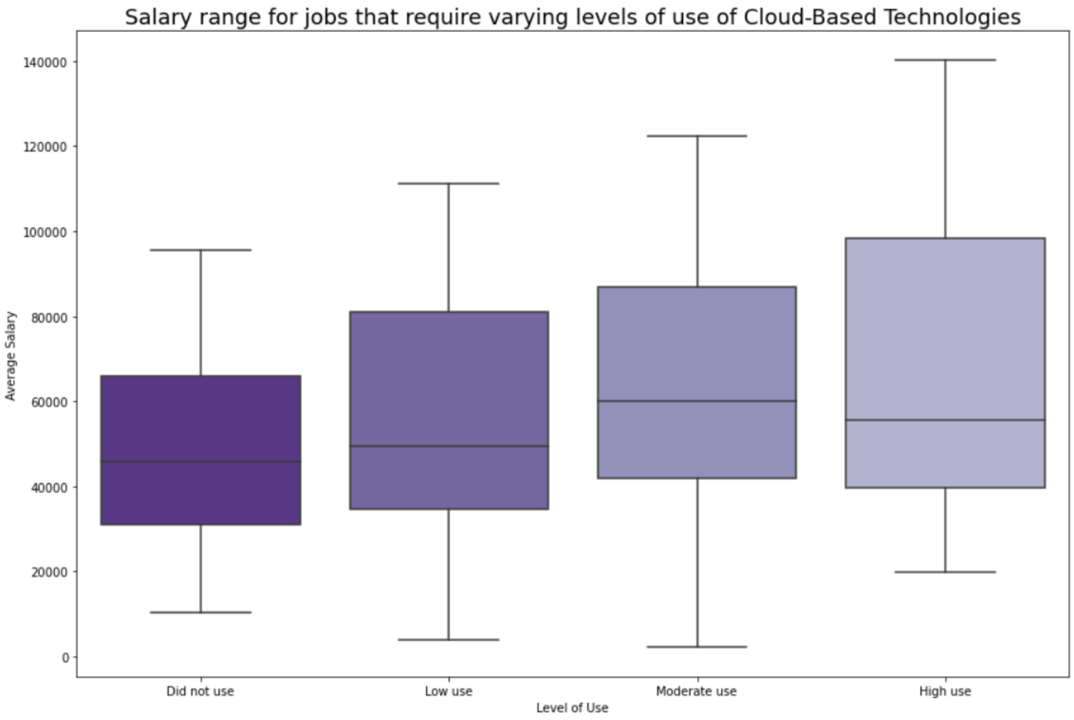


To determine the salary range for jobs that require varying levels of use of software-based technology (Artificial Intelligence, Cloud-Based, and Specialized Software), we first calculated the average salary. This was done by dividing the annual payroll by the number of employees, and multiplying the result by 1000. The data for the annual payroll is provided in thousands of dollars, and includes all forms of compensation paid to employees during the reporting year. Payroll does not include employer costs for benefits such as payroll taxes, insurance premiums, pension plans, and other employer-paid benefits. (U.S. Census Bureau) We then created new DataFrames for each technology, called ai, cloud, and software, respectively. These DataFrames only include rows for the specified technology in the Technology Used column. In the PAYANN_F column, rows with a value of 0 for annual payroll are flagged. This is because data was not collected for these rows in order to protect the confidentiality of individual companies, or because the estimate does not meet publication standards. Therefore, we removed rows that are flagged as S or D in the PAYANN_F column. (U.S. Census Bureau)

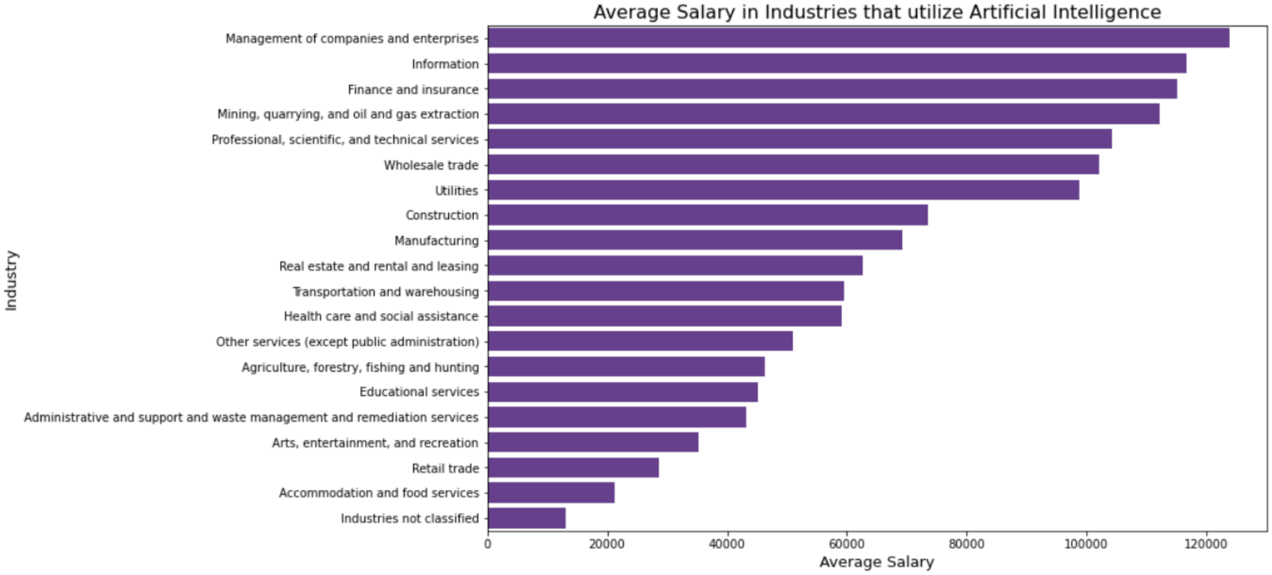
We used boxplots to visualize our data and found that AI and specialized software technologies had the widest and highest spreads when used moderately. AI had the highest maximum salary at moderate use, while specialized software had maximum salaries tied at moderate use. AI also had a higher median for high usage. Cloud-based technologies had the highest and widest spreads on high usage, but the median was lower than when used moderately. There appears to be a linear increase in higher-end salaries for firms using cloud-based technologies. Interestingly, the minimum salaries for AI, cloud-based, and specialized software were very low when used moderately and sometimes high. Only high use of cloud-based technologies had the lowest minimum end of the average salaries. It is not surprising to see higher average salaries for firms that do incorporate some level of a software-based technology.







To gather the breakdown of salaries for industries that work with AI, we grouped the Technology Used, Industry, and aggregated Average Salary from the ai DataFrame into a new dataframe called group_tech. We created a numpy array called ai_industry that contains the unique values for Industry from the ai DataFrame and sorted it. We then created a new DataFrame called Payroll_ai by transposing a list of the group_tech and ai_industry data. We renamed columns 0 and 1 to Average Salary and Industry, respectively, and sorted the values in Payroll_ai by Average Salary. The top five industries with the highest average salaries, in order, are Management of companies and enterprises, Information, Finance and insurance, Mining, and Professional, scientific, and technical services.



Conclusion

While we were able to draw significant conclusions from the above visualizations, it is important to recognize that this is only a brief glance at the patterns and tendencies of businesses in the United States. It would be interesting to examine how population differences across the US may affect the distribution of our data. Additionally, as we only analyzed datasets from the year 2019, it is not possible to predict new trends from our analysis. However, this report provides a valuable introductory look at the demographics of US businesses, even without a complete understanding.

Works Cited

Cloud-Based Services Are Making It Easier for Companies to Use AI - SPONSOR CONTENT FROM DELOITTE. (2019, June 21). Harvard Business Review. https://hbr.org/sponsored/2019/03/cloud-based-services-are-making-it-easier-for-companies-to-use-ai

Industries at a Glance: Professional, Scientific, and Technical Services: NAICS 54. (n.d.). US Bureau of Labor Statistics. https://www.bls.gov/iag/tgs/iag54.htm.

Parker, K. (2021, November 8). *What's behind the growing gap between men and women in college completion?* Pew Research Center. https://www.pewresearch.org/fact-tank/2021/ 11/08/whats-behind-the-growing-gap-between-men-and-women-in-college-completion/.

U.S. Census Bureau. (2022, October 28). *Annual Business Survey (ABS) APIs*. Census.gov. https://www.census.gov/data/developers/data-sets/abs.2019.html Documentation for the US Census Bureau's Annual Business Survey API. The data can be assessed by calling the API. An API Key is required.

U.S. Census Bureau, ASD, WSCS. (n.d.-b). *U.S. Census Bureau: Page not found*. https://www2.census.gov/programs-surveys/economic-census/about/fields_variables_glossary/PAYANN+FINAL_021919.docx

U.S. Census Bureau (2021, November 30). *Technical documentation*. Census.gov. Retrieved January 3, 2023 https://www.census.gov/programs-surveys/abs/technical-documentation.htm l

U.S. Census Bureau, ASD, WSCS. (n.d.). *U.S. Census Bureau: Page not found*. https://www.census.gov/data/developers/guidance/api-user-guide.Help_ API keys can be obtained by following the link and clicking on 'Request a Key' on the left-hand side of the page.

Valcic, M. (2017). *2010* [Dataset].https://www.kaggle.com/code/mvalcic/add-city-state-longitude-and-latitude-data/data.