Group 1 Project Report

**Data Source:**

For this assessment, we used the Annual Business Survey (ABS) APIs from the United States Census Bureau’s website. We used the data for 2019, which broke out into 4 different categories. The different datasets available were for company summary, characteristics of businesses, characteristics of business owners, and technology characteristics of businesses. All of these datasets included variables for different demographics of sex, ethnicity, race, veteran status, years in business, size of the firm, and employment size of the firms/firm owners. There was also data based on different geographical regions available in the datasets.

APA Citation:

Bureau, U. S. C. (2021, September 4). *Annual Business Survey (ABS) apis*. The United States Census Bureau. Retrieved October 8, 2021, from <https://www.census.gov/data/developers/data-sets/abs.2019.html>.

**Questions to answer:**

**Luke:**

1. How does employee count affect the total payroll for each state?
2. Which states have the highest payroll per employee?
3. How does the gender of the owner influence the payroll per employee for firms?

**TJ:**

1. How does ethnicity affect the number of employees a company has?
2. How does owner gender affect the number of employees a company has?
3. How many employees are working at a company that utilizes AI?

**Daniel:**

1. Is there a significant difference between the amount of employed people based on race?
2. Is there a significant difference in the amount of money earned in the US based on race?
3. Are minorities more often employed by larger firms?

**Christian**

1. Do veterans employ more workers? Does this differ between states?
2. Do veteran-owned businesses pay their employees more than non-veteran owned businesses?
3. How many employees does the average employer employ in each state?

**Process for searching through the data:**

**Luke:**

This dataset required us to allow ample time to sift through the data and come up with valuable questions. With a lack of numerous quantitative variables, such as payroll and employee count, this dataset caused us to narrow our focus to how different demographics affected the limited quantitative variables. This narrowing led me to pursue questions based on trends in payroll and employee count given various demographic factors. The questions that I answered only required the use of the first API with a summary of the firms. Other team members pursued questions that required merging of API calls, while my exploration used tactics such as applying functions to create new columns. To answer my initial question of “How does employee count affect the total payroll for each state?”, I had to call in the specific columns of employee count and total payroll from the API for each state and store them in a dataframe. This then allowed for the use of Seaborn to create a regplot showing the trends in the total payroll based on the employee count. After viewing this graph, it looked like there was a positive correlation between employee count and total payroll, as I expected. After viewing this graph, I thought it would be beneficial to take an individual look at each state and see which states had the highest payroll per employee. This analysis required a new column, which I called “pay/emp”. This column simply takes the total payroll divided by the employee count for each state. Then I could select the 10 states with the highest “pay/emp” and display them in a bar chart created by using matplotlib. The final analysis I performed looked at a specific demographic that I thought may affect the “pay/emp”. This demographic was the gender of the owner. Using a grouped bar chart from plotly, I was able to look at the top 10 “pay/emp” values and see how many of them were for women owners.

TJ:

For the first 2 questions, I had, looking through the variables in the technical documents for the API led me to believe that the characteristics of the business dataset had all of the variables that I was looking for, including owner ethnicity group, owner sex, and the number of employees. In python, I called the API with the specific parameters that I needed. I figured out the API will return data for each unique set of parameters that you give it, so when you give it more parameters the dataset gets very large very quickly. To avoid this, I only called the data that I needed. I wanted to answer both of the questions at once with 2 graphs, one for the number of employees by gender of Hispanic-owned businesses and the other for the same axes but of non-Hispanic-owned businesses. To find this out, the first thing I did was create two new data set filtered by Ethnic group. This is easy as the column contained codes, that are readable via the technical documents that come with the datasets. For example, taking only the Ethnic groups that were labeled 029, would only give data for businesses owned by non-Hispanics, this works similarly with the Hispanic-owned businesses. The dataset gave 4 different categories for gender, total, male, female, and male-female equally owned. These were encoded as well so I had to merge the dataset with a description key dataset I made in order to get better labels for the graphs I ended up making.

For the final question that I wanted to answer in relation to the AI use of companies and how many employees they have. I wanted to do this to see the number of employees that are affected by and work with AI directly at least at the firm level. From my looking through the data, I saw that limited the API call is the most efficient way to get exactly what you are looking for. So, in order to do this, I only need to include the number of employees and the tech use variables for the API call. Another thing that I needed to filter for was the specific rows in the tech use column that was associated with AI use. To do this I made another pandas data frame that included the codes and the descriptions for the AI tech use rows, so that when I inner joined the data frames I would only get those specific rows. From there I was able to directly make the bar graphs that I wanted to make, the number of employees working at firms and the level of AI use the company reported. Looking at the initial graph, having the row for no AI use would heavily skew the y axis because most firms do not use AI. In order to better see the data for companies that use AI, I dropped the row for no AI use and then made another graph as a drill down.

Daniel:

After going through each of the four predetermined APIs and examining the different types of available data, I decided to use this data to mathematically determine if there were significant differences in employment and earnings based on race in the United States. I first determined which APIs I would need to use in order to answer my questions. Since the Technology Characteristics of Businesses API mostly focused on data regarding the motivations, uses, and effects of enhanced technological methods in the workplace, I decided against using that API. The Characteristics of Business Owners API was more focused on data for business owners as opposed to the overall business, so that API was also disregarded for my analysis. Both remaining APIs had data that I could examine, but I decided to focus on the Characteristics of Businesses API for my analysis.

While it is commonly stated that people of certain racial demographics tend to have an easier time getting employed by companies and holding jobs, I wanted to use this data to attempt to see if this was true. For the first question I proposed, I decided to examine the number of employees in the United States for each race.

For my second question, I decided to take the next logical step from my first question and determine if there was a significant difference in earnings between different racial demographics. To examine this, I observed the summation of the payroll amounts of each racial demographic and compared them to the number of employees to see if any disparity in pay truly existed.

For my last question, I was simply curious to see if minorities were more often employed by larger firms than smaller ones. To help visualize this, I looked at the data for firms that earned specific amounts of revenue and determined how many employees those firms hired happened to be minorities.

Christian:

Understanding the API data was very difficult for me at first, and required reading the API documentation multiple times to determine what exactly was stored and how it could be aggregated. I had to determine exactly which variables I needed to gather, what each code in the data stood for, and how the tables could be joined.

Answering the first question required learning exactly how to make an API call to the Characteristics of Businesses API, and how to link the codes in the data with the description linked to the code in the documentation. I determined that the best way to link the descriptions was to simply manually create a DataFrame of the codes in question, seeing as there were only four codes to list out. I also filtered the data down to the data for Minnesota and Texas, states which I thought might have significant enough cultural differences to be reflected in the data. Finally, I had to convert the counts of employees to proportions of the total, as the difference in population between Texas and Minnesota might make the data hard to visualize. The data showed that Minnesota and Texas have virtually identical proportions of veteran-employed to non-veteran-employed employees.

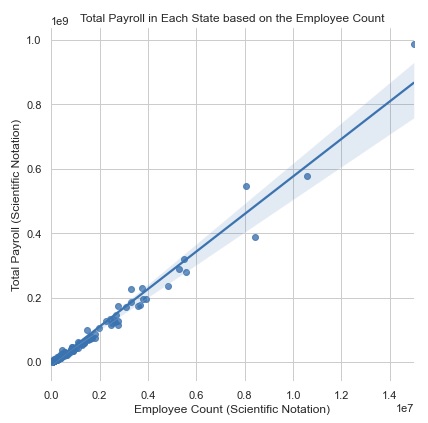
The second question required no filtering, as the data to be compared was nationwide. I pulled the data on employee counts and annual payroll from the Characteristics of Businesses API, grouped by state and veteran status. I converted both the employee count and annual payroll columns to integers for the purposes of graphing the data, and found an extremely strong correlation between company size and annual payroll, with very little deviation based on veteran status. However, as the previous question described, veteran-owned companies employed far fewer employees than non-veteran-owned companies on average.

Finally, the third question required merging two separate API columns into one table. Ensuring that both queries gathered state-level data ensured that the tables could be merged neatly, as each would have 51 rows (50 states and D.C.). The tables were simply merged on the state column. I then generated a new column representing the number of employees per employer, which was simply created by dividing the employee column by the employer column. Creating a histogram of the data let me find that the mean number of employees per owner was around 810, with a right skew due to the states of Nevada and Arizona.

**Results(visualizations):**

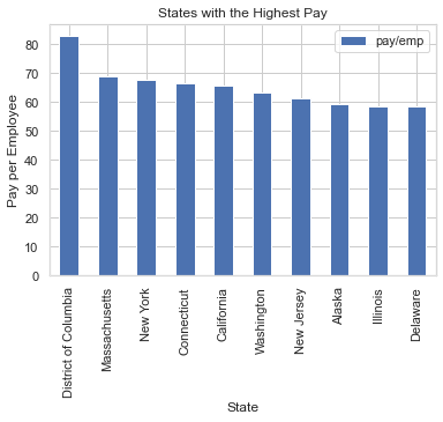
**Luke:**

Question 1:



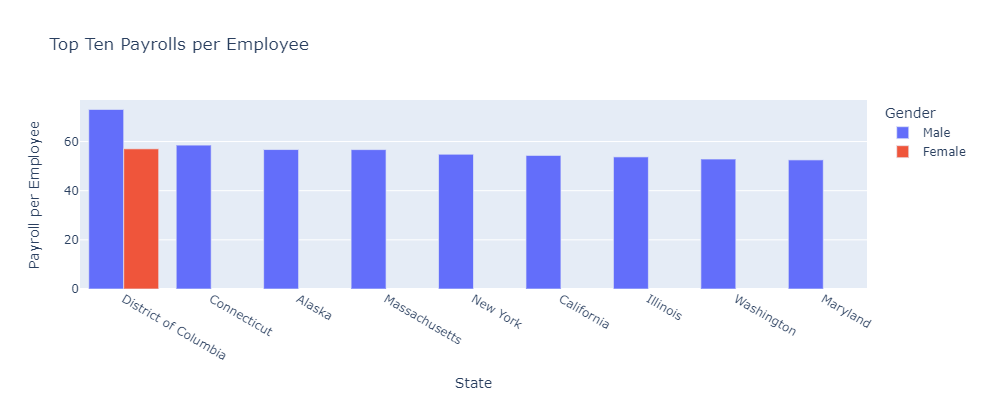
**Fig 1.** scatter plot with trendline created using Seaborn showing the rise in total payroll for states with larger employee counts. Axes in scientific notation for functionality and readability. Shading to show the 95% confidence interval.

Figure 1 shows the trend in payroll based on employee count. The results of this graph are as hypothesized, because each employee is paid some amount. This means that total payroll will increase each time an employee is added to a firm. The trendline represents the average payroll at a certain employee count based on all states. Many states with smaller economies are clustered at the lower left portion of the chart, whereas states with larger economies are at the top right portion of the chart. There are much fewer states with large economies. Any point on figure 1 that is above the trendline will have higher payroll per employee than average and any point below the trendline will have lower payroll per employee than average.



**Fig 2.** bar graph created using matplotlib to show the top ten states with the highest payroll per employee.

Figure 2 normalizes the data for the size of the economy and looks at the states with the highest payroll per employee. The top 3 states include the District of Colombia, Massachusetts, and New York. The top ten contains a range of large economies and small economies. The District of Columbia outpaces the rest of the field in pay/emp by more than ten dollars.

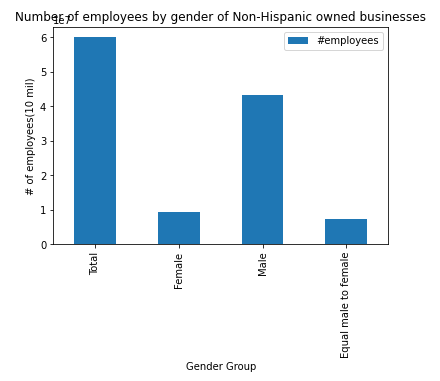


**Fig 3.** top ten payroll per employee values broken up by gender of the owner. Created using plotly with genders represented by color.

Figure 3 displays the top ten payroll per employee values when broken down by gender of the owner. As shown in figure 3, there is only a single value in the top ten for firms owned by women in a state. The results in figure 3 are similar to figure 2 in which states are listed but differ through displaying a demographic influence on the payroll per employee.

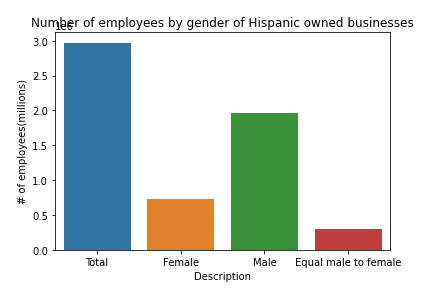
TJ:

Questions 1 & 2:



The graph above was made with matplotlib in python from the 2018 business census data. The graph shows the distribution of employees between male and female owned businesses that are of Non-Hispanic Ethnicity

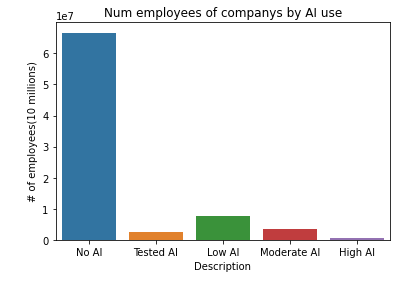
This graph above starts to answer the second question that I had in that how does owner gender affect the number of employees a firm has. As we can see from the graph, the majority of the employees are at firms that are male owned by a good margin. This tells us that there are most likely more male owned businesses than female owned businesses.



The graph above was made using seaborn and shows the distribution of employees based on owner gender, of owners who are of Hispanic

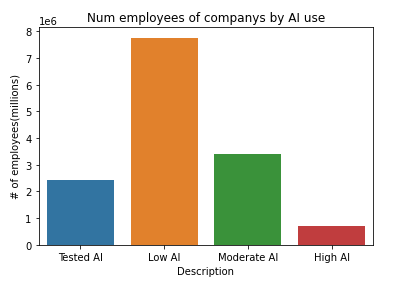
For this graph, we see proof for the answer to my second question as well, that again within Hispanic owned businesses, there are more employees at male-owned firms than female-owned firms. So, the overall answer to the second question is that owner gender does affect the number of employees that the firm has on a high level. In order to answer my first question, we need to look at both of the previous graphs, specifically, the Y-axis is the most important part here. The difference between the two graphs is one is only for non-Hispanic-owned firms, and the second is for Hispanic-owned firms. Looking at the number of employees, the total for Hispanic-owned firms is about 3 million, while the total for non-Hispanic-owned firms is just under 60 million employees. From this data we can say that the ethnicity of an owner does affect the number of employees. We can say for both of these questions that there are more male owned businesses and also there are less owners that are of Hispanic Ethnicity.

Question 3:



This graph above shows the number of employees in firms and their level of AI use reported at the company. The graph was made using seaborn.

Looking at the graph above we can easily conclude that the majority of employees are working at a firm that uses no AI in their business. This would make sense due to AI being a relatively new concept, especially in 2018 when this survey was conducted.



The graph above is a drill-down of the previous graph to get a better view of how many employees are working at firms that use AI and the level at which they use AI.

I wanted to see the distribution of employees that worked at firms that did use some level of AI, since the in the previous graph the majority of employees were working at firms that did not use AI at all, which makes the other descriptions hard to see. We can see that in firms that use AI the majority of employees work at firms that use AI at a low level. This makes sense as well as again, AI is a newer concept and unless you work at a firm specializing in AI, it makes sense that there would not be heavy use of AI.

Daniel:

Question 1:

|  |  |
| --- | --- |
| **Race** | **# of Employees** |
| American Indian and Alaska Native | 400200 |
| Asian | 10180130 |
| Black or African American | 2377636 |
| Native Hawaiian and Other Pacific Islander | 108571 |
| White | 113754864 |

Table 1: Table showing data for Average US Employee Count by Race in 2018

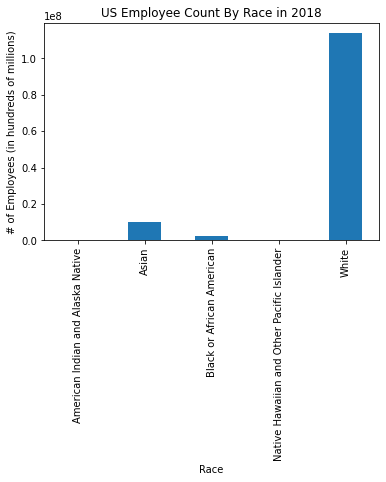


Figure 1: Bar graph created using matplot.lib showing the number of employees in the US for different racial demographics

In the graph shown above, there is a clear disparity in the number of US workers. People who are white are in the vast majority with 113,754,864 employees (89.70% of the total) with the next highest group being the Asian group with 10,180,130 employees (8.03% of the total). This visualization makes it explicitly clear that there is a rather large racial disparity in the workplace, at least in 2018.

Question 2:

|  |  |
| --- | --- |
| **Race** | **Employee Earnings** |
| American Indian and Alaska Native | 16475163000 |
| Asian | 351600528000 |
| Black or African American | 73878456000 |
| Native Hawaiian and Other Pacific Islander | 4247283000 |
| White | 5160208241000 |

Table 2: Table showing data for US Summation of Employee Earnings by Race in 2018



Figure 2: Horizontal bar graph made using matplot.lib showing the amount of money earned in 2018 by different racial groups.

At first glance the bar graph above makes it seem like there is another clear disparity between racial groups, however there is more to consider than the bar graph itself. The clear advantage of white employees in the earnings graph could easily be explained by figure 1 and table 1, the data from the previous question. Since white people make up the vast majority of US employees, it would be rational that their total earnings would also be much higher. In order to determine if a true disparity exists, I divided the earnings of each racial group in 2018 by their respective number of employees in 2018. The results are shown in the following table:

|  |  |  |  |
| --- | --- | --- | --- |
| **Race** | **Employee Earnings** | **# of Employees** | **Earnings/# of Employees** |
| American Indian and Alaska Native | $16,475,163,000 | 400200 | $41,167.32 |
| Asian | $351,600,528,000 | 10180130 | $34,537.92 |
| Black or African American | $73,878,456,000 | 2377636 | $31,072.23 |
| Native Hawaiian and Other Pacific Islander | $4,247,283,000 | 108571 | $39,119.87 |
| White | $5,160,208,241,000 | 113754864 | $45,362.53 |

Table 3: Table showing the results of dividing employee earnings by the number of employees

Table 3 shows that the white employees in 2018 made approximately $4195 more than the next highest racial group and $14290 more than the lowest earning racial group. The high variance of the data does display a disparity between different racial groups, however there are many other potential factors besides race that could explain the variance. As such, it is my opinion that more research needs to be conducted to determine if a difference in race truly leads to a difference in pay.

Question 3:

|  |  |
| --- | --- |
| **Revenue\_Desc** | **# of Minority Employees** |
| Firms with sales/receipts of less than $5,000 | 11537 |
| Firms with sales/receipts of $5,000 to $9,999 | 6823 |
| Firms with sales/receipts of $10,000 to $24,999 | 22234 |
| Firms with sales/receipts of $25,000 to $49,999 | 40714 |
| Firms with sales/receipts of $50,000 to $99,999 | 125490 |
| Firms with sales/receipts of $100,000 to $249,999 | 490432 |
| Firms with sales/receipts of $250,000 to $499,999 | 865638 |
| Firms with sales/receipts of $500,000 to $999,999 | 1269106 |
| Firms with sales/receipts of $1,000,000 or more | 6600106 |

Table 4: Table showing the number of US minority employees by firm revenue in 2018

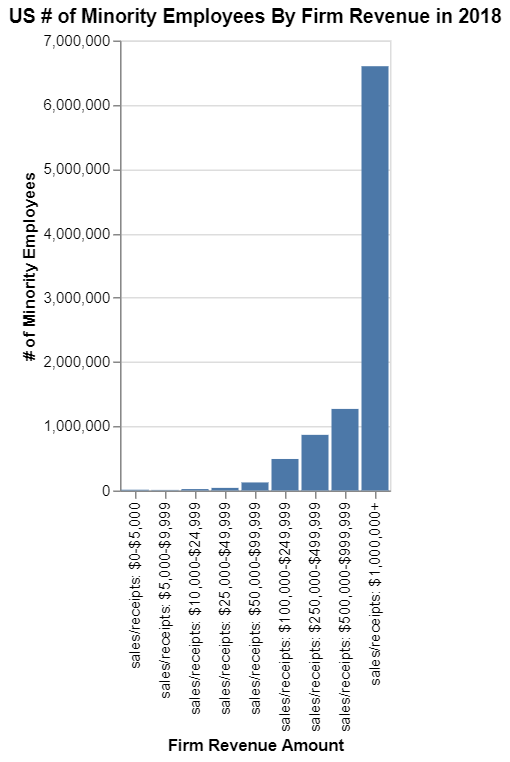
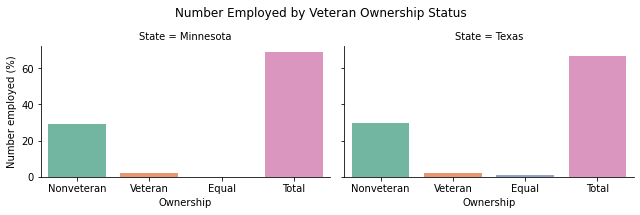


Figure 3: Bar graph made using altair that shows the number of minority employees employed at firms who earned specified amounts of revenue in 2018.

Table 4 and figure 3 deal with data involving the population of minorities rather than observing each race individually. The data is divided into firms that earn certain ranges of revenue to differentiate between bigger firms and smaller firms, since bigger firms likely earn more revenue on average than smaller ones. The graph clearly shows that, for the most part, firms have a tendency to hire more minority workers the more they grow. This is a natural conclusion since bigger firms require more people to keep them going, but there is a substantial increase in minority workers for firms that earn at least one million dollars in sales. More research needs to be conducted in order to determine why there is a sudden spike in minority employment.

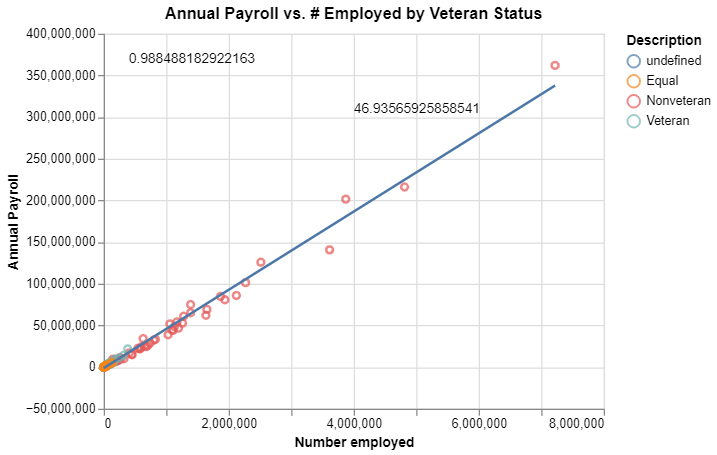
Christian:

1. Do veterans employ more workers? Does this differ between states?



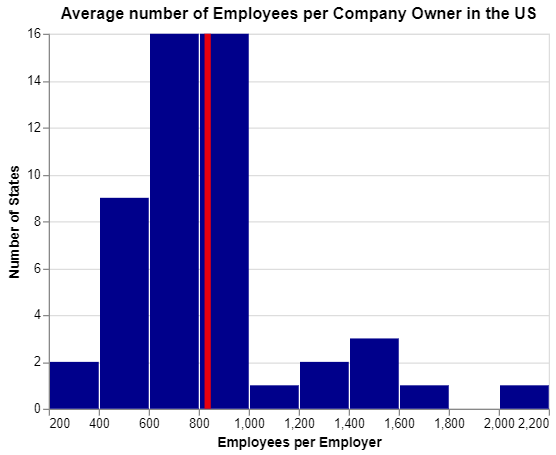
This graphic shows that there is very little difference between Texas and Minnesota in terms of the proportion of employees working at veteran- and nonveteran-owned companies. The only noticeable difference is in the lack of equally veteran- and nonveteran-owned companies in Minnesota, while Texas seems to have a small percentage.

1. Do veteran-owned businesses pay their employees more than non-veteran owned businesses?



This graph shows a very strong correlation between the number employed and total payroll, which in retrospect should have been obvious, as more employees necessarily means more employees on payroll. We can also see that veteran-owned companies are much smaller on average than nonveteran-owned, which is another way of stating the previous graph’s findings. The r-squared value of .988 and the slope of 46.93, however, give us some good mathematical evidence for our correlation.

1. How many employees does the average employer employ in each state?



Most owners in the data have between 600 and 1,000 employees under their supervision, indicating many large companies in the ACS’ participants. However, we have a heavy rightward skew, which includes the state of Nevada and its economic hub of Las Vegas. If this data is an accurate estimation of all American companies, then the average American firm is quite large.

**Conclusions:**

**Luke:**

My analysis provides a brief look at the trends in payroll numbers based on factors such as geographic location, employee count, and gender of the owners. Some results were expected, such as an increase in total payroll as employee count increases. While others, such as which states had the highest payroll per employee and the disparities in gender of the owner for the highest payroll per employee, were not as apparent before our analysis. Higher payroll per employee in certain states could have a lot to do with the industries that are common in those states. Further research into the reasoning for the higher payroll per employee in states such as the District of Columbia, Massachusetts, and New York could be beneficial. Further research on how certain demographics affect payroll per employee could also be beneficial. The race and ethnic groups that the owners of firms are a part of could influence the payroll per employee just as gender did. Research into why gender had such a profound effect could also be valuable. A disparity in profits for male versus female owned businesses could be the reasoning behind seeing a lack of female payroll per employee values in the top ten.

This exploration into the *Annual Business Survey* provided a preliminary analysis of trends for companies in the United States. Our analysis only scratched the surface of information within this survey and opened numerous doors to further research.

TJ:

Overall, I found that there are less employees working at firms that have owners of Hispanic ethnicity. This would make sense as ethnicity is split up by Hispanic versus non-Hispanic, so the Hispanics are a minority in this sample. I found that the majority of employees are working at firms that are male owned as well. Something to note is that the proportion of male to female owned businesses is actually lower within the Hispanic ethnicity, meaning that if an owner is Hispanic, they are slightly more likely to be female than if the owner is non-Hispanic. Finally, we see that the majority of employees work at firms that use no AI at all, but of those that use AI, the majority of firms use AI at a low level.

Daniel:

The cultures people are born into and the way they are raised greatly make up who they are both in life in general and in the workplace. Differences in race are very obvious when looking at the professional world, but the reason why is much less opaque. Looking at the US populations of employees by race, it is clear that white workers make up much of the professional scope. It is also clear that people of different races earn different amounts overall, but it is still unclear why. Similarly, minorities are highly employed by bigger and more successful companies, but there is not enough evidence in the current data to generate a probable cause as to why that may be. There may never be a clear answer to questions such as these that have a seemingly endless list of potential factors, but with more work and more research, we may come closer to finding the true answer.

Christian:

I feel that my second question could have used some more thorough exploration – I’m not confident that the visualization I produced quite answers the question. If I were to re-do my second question, I would perform further data transformation and find a way to account for a clearly vast difference in the size of veteran-owned and nonveteran-owned firms. However, I feel pleased with my work in questions one and three. Question one proves that there’s quite a strong economic similarity between states that may seem different at face value, and question three indicates a robust American economy.