Analysis of the US Census Bureau’s Annual Business Survey

Carol Lopez, Finn McSweeney, Zeeshan Pervaiz, & Rachel Walter

Initial Questions

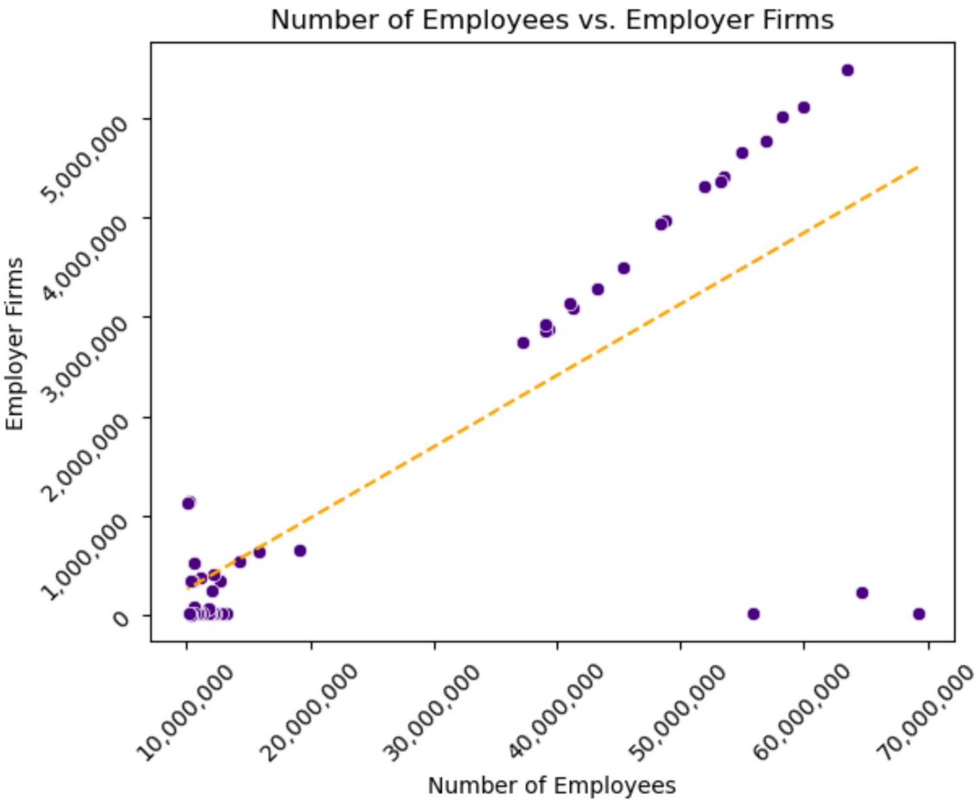
- What is the average salary of employees by state?
- What percentage of workers are male and what percentage are female? How does that compare to the number of male versus female owners?
- What technologies are mostly used by employer firms?
- What is the salary range for jobs that require varying levels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

ETL Steps

- Each dataset was loaded in and the first row was removed (Duplicate header row).
- The index of each row was then reset and previous index was removed.
- Multiple columns in the dataframe were renamed to make the purpose of the columns clearer.
- The data types of the numeric columns were changed from object to either an integer or a float.
- Removed nulls and flagged data.

Company Summary

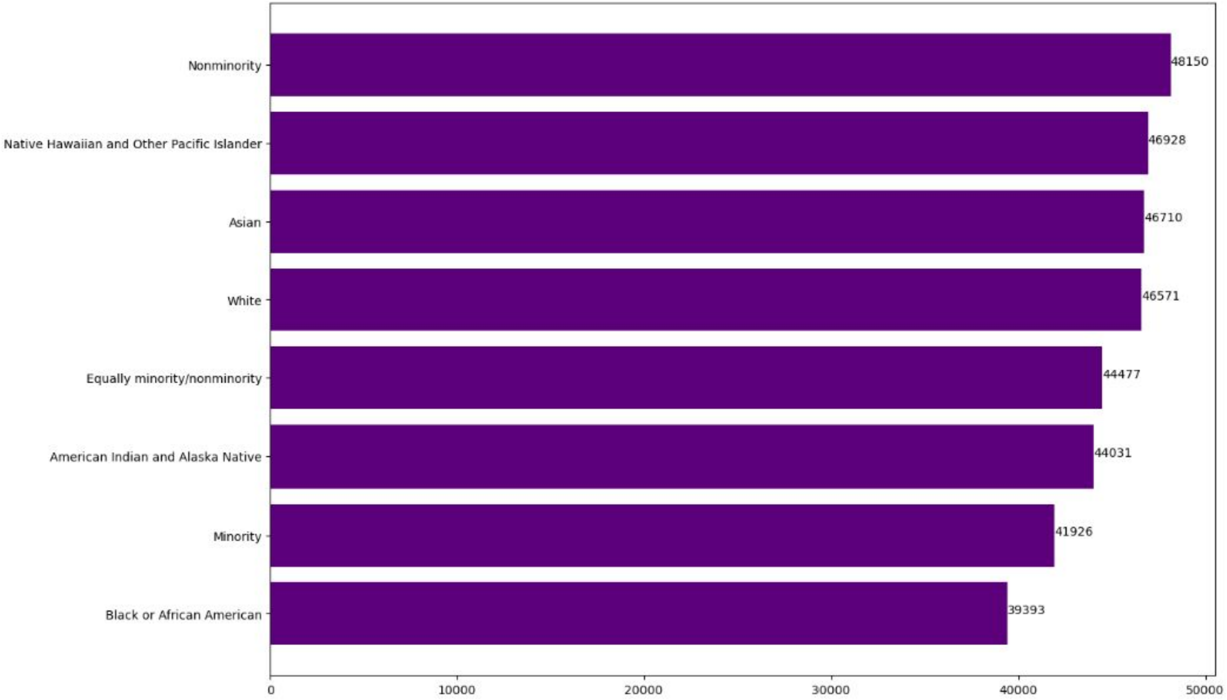
Employee Count and Employer Firms



View the correlation between number of employees and the number of employer firms.

- Top 50 companies were used in this visualization.
- One outlier removed from the dataset.
- Correlation of the graph is ≈ 0.7792.
- Correlation not visible with smaller companies, clear trend with the mid-large sized companies.
- Employer firms all > 0, no company in the top 50 has none, some way smaller than others (lowest = 4000).

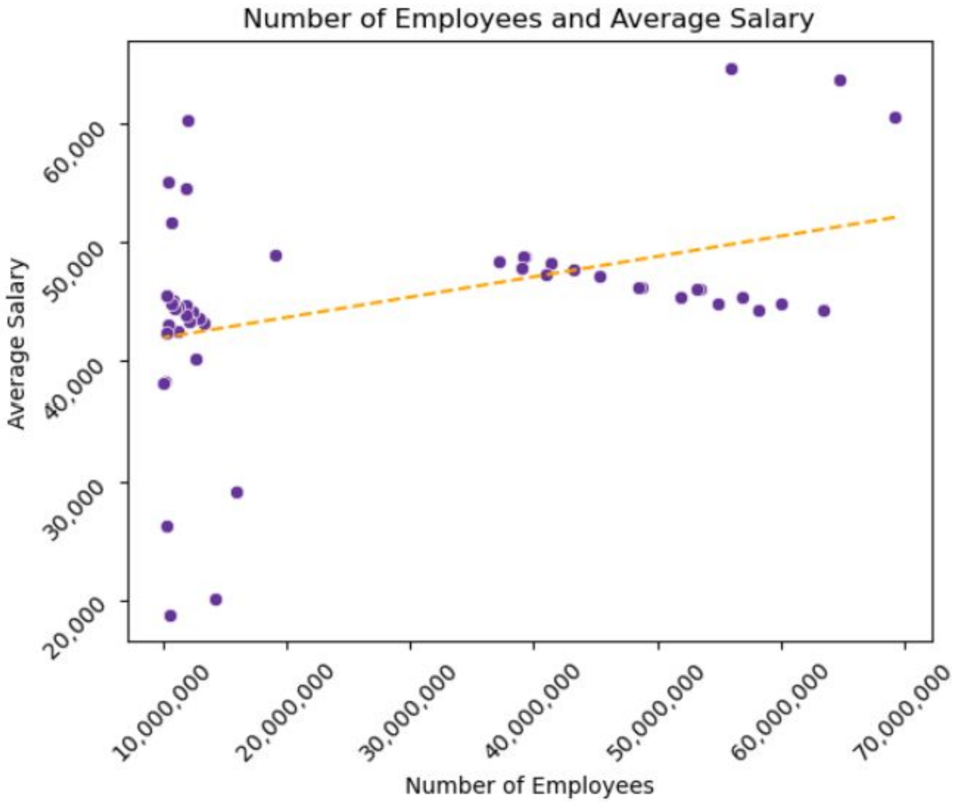
Average Salary Compared to Race



Bar Chart of average employee salary based on race

- Average salary was calculated by the annual payroll / number of employees * 100.
- Average Salary column's mean was taken and grouped by race.
- Total, classifiable, nonclassifiable columns were taken out of the visual.

Average Salary and Number of Employees



Correlation between size of companies and average payroll

- Top 50 companies by size were also used in this visualization.
- Smaller companies had no correlation to the average salary.
- Mid-Large sized companies showed a decline in the average salary as the company size grew.
- Misleading trendline and correlation due to the companies with under 20,000,000 employees.

Code snippets .

1. Creating DataFrame with top 50 companies by employee count
    a. top50 = df.sort_values(by = 'EMP', ascending=**False**)[:50]
2. Checking the correlation of scatter plots
    a. correlation = df['EMP'].corr(df['FIRMPDEMP'])
    b. correlation = df['EMP'].corr(df['Average_Salary'])
3. Creating an Average Salary column
    a. df['Average_Salary'] = (df['PAYANN']/df['EMP'])*1000
4. Rounding the Average Salary column them calculating the mean grouped by race.
    a. df["Average_Salary_Rounded"] = df["Average_Salary"].round(0)
    b. df = df.groupby("RACE_GROUP_LABEL")["Average_Salary"].mean()

Characteristics of Businesses

Question: Percentage of male and female employees.

Steps to generate visual:

- Group the number of workers.
- Determine the count of workers for each gender:

  - Total_gender_count = Total - (Classifiable + Unclassifiable)

  - Male_count = Male + Equally male/female

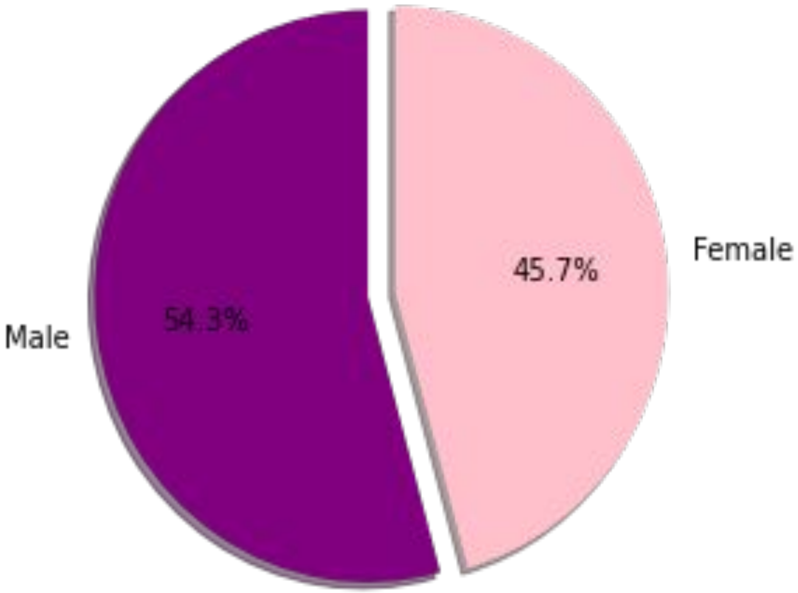
  - Female_count = Female + Equally male/female

- Calculate percentages:

  - Percent_male = (Male_count/ Total_gender_count) * 100

  - Percent_female = (Female_count/ Total_gender_count) * 100

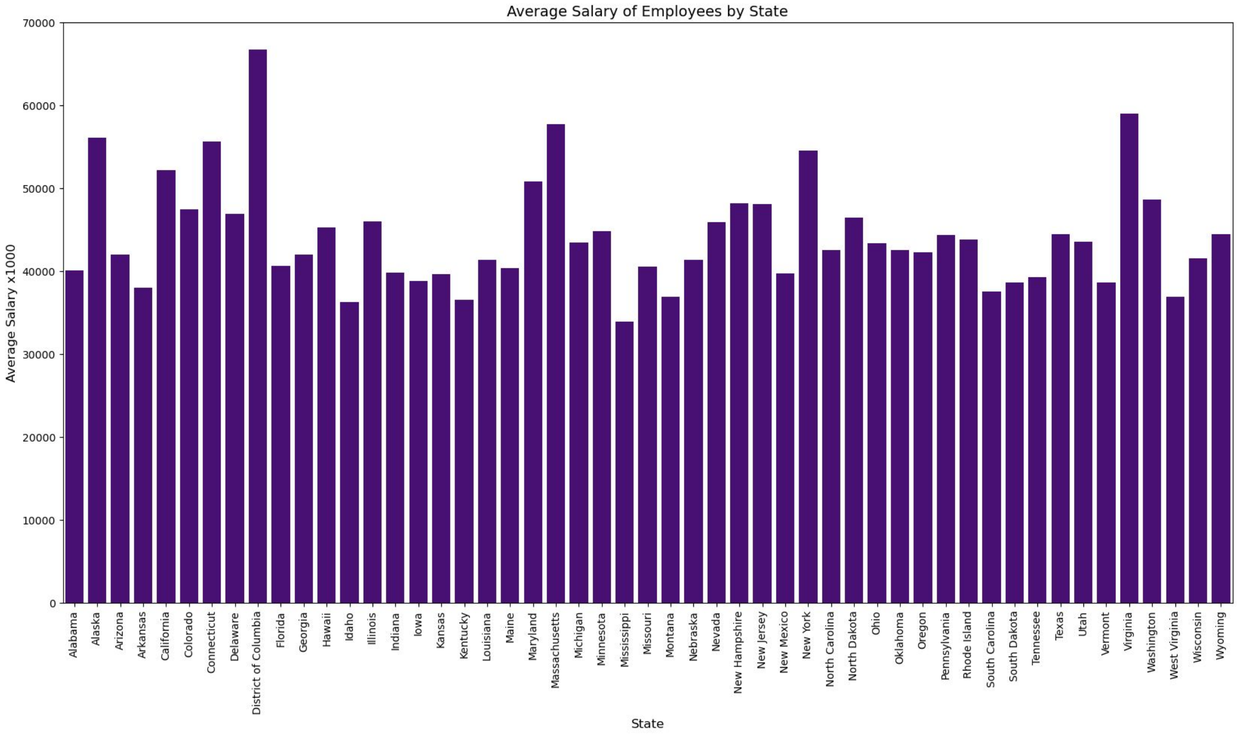
- Plot as Pie chart



Question: Average Salary of Employees per state

Steps to generate visual:

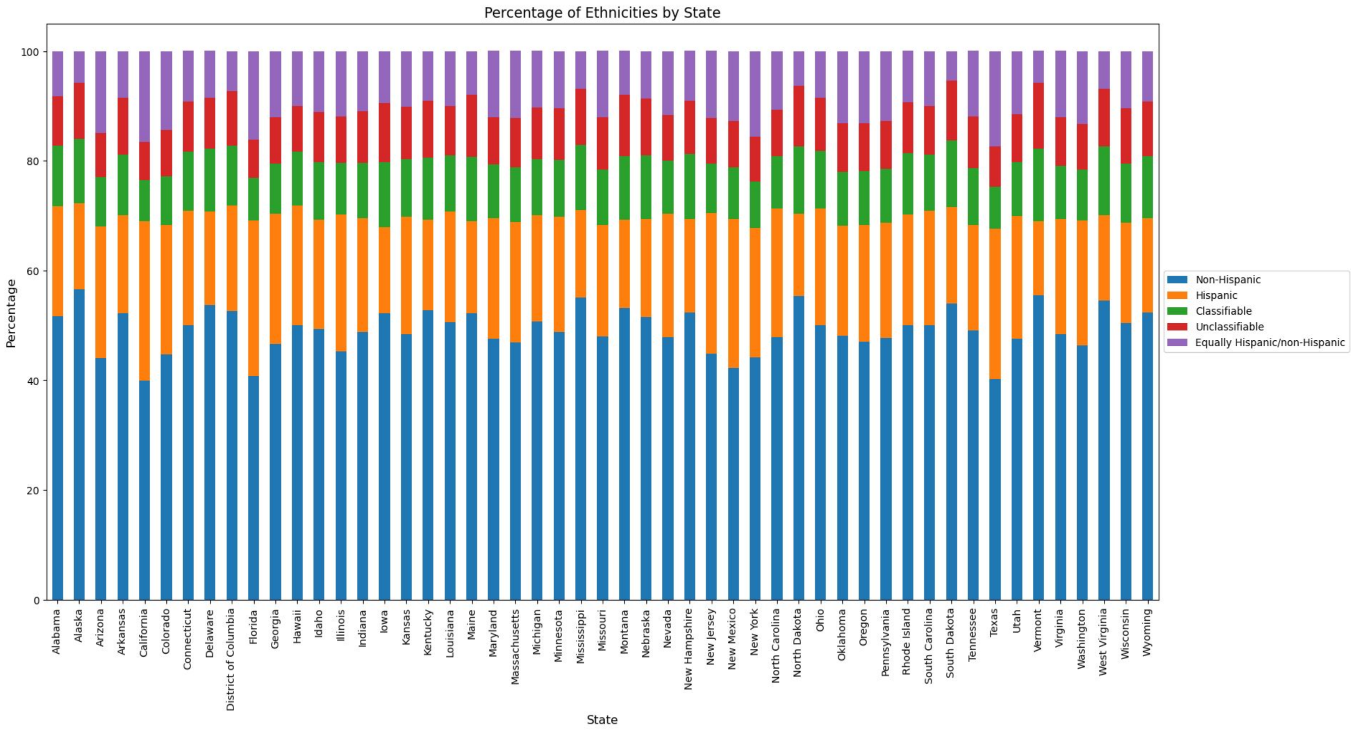
- API call on the business summary data table.
- Group salary values and states.
- Take the average of the salary values for each state.



Question: For each state, what percentage of workers belong to each ethnicity?

Steps to generate visual:

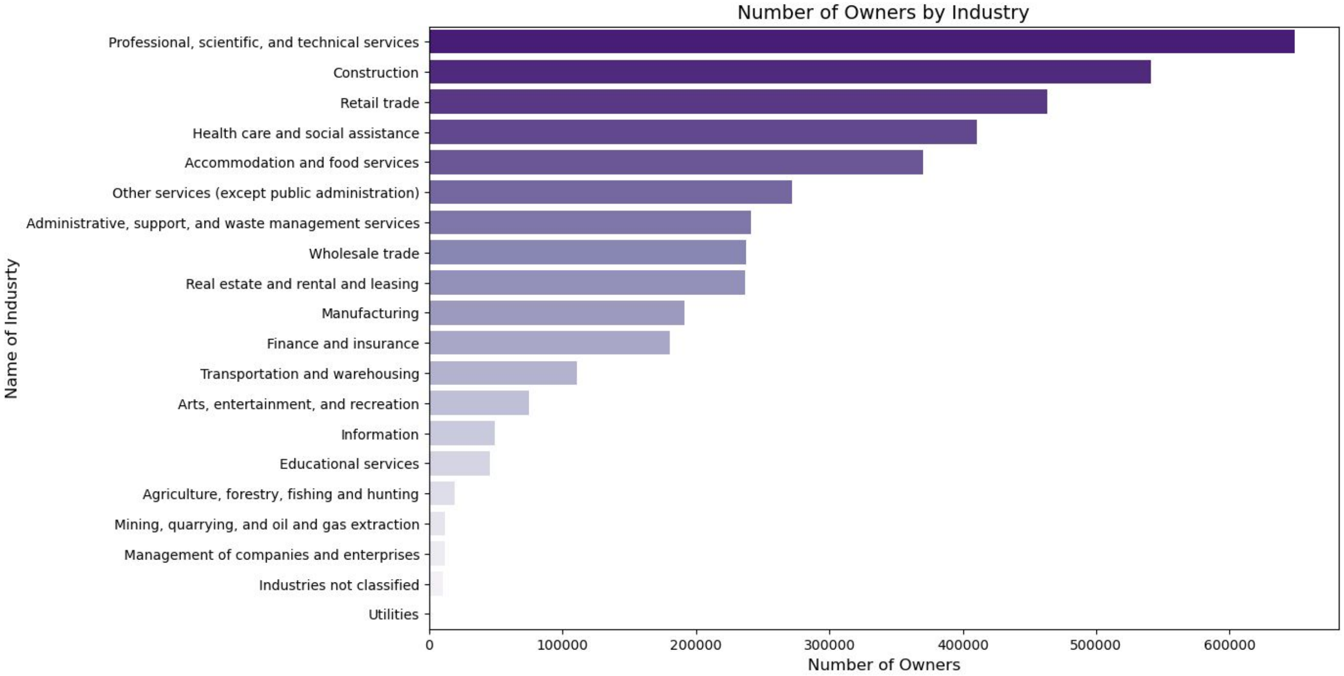
- Drop the native “total” value from data set.
- Group by state name and ethnicity group label.
- Calculate new total ethnicity value for each state.
- Calculate the percentage of each ethnicity by dividing each ethnicities count by the total value for each state.
- Plot the resulting data on stacked bar chart.



Characteristics of Business Owners

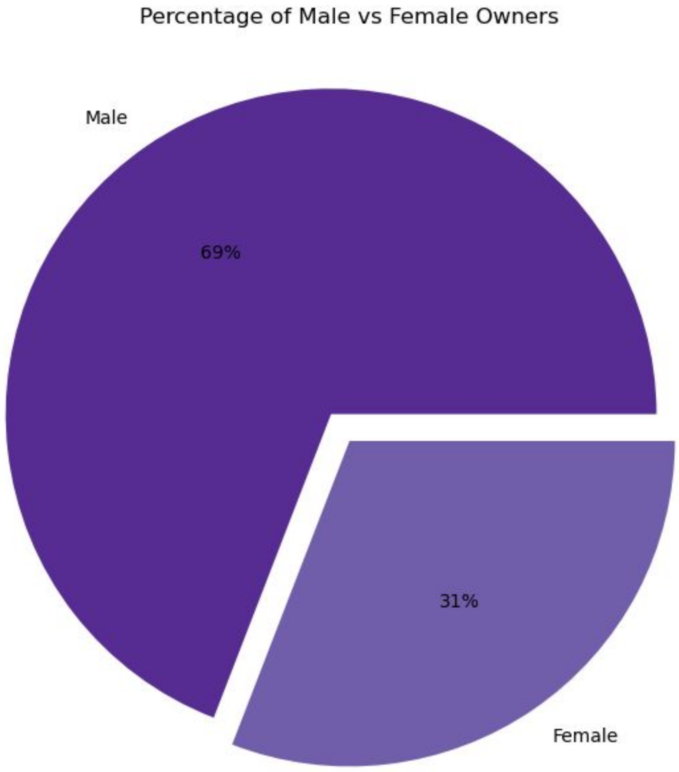
What is the breakdown of number of owners by industry?

- Filtered down to focus on each individual industry
- Used group by and sum to get the totals
- Sorted values so the bars would be in descending order
- Difficult to draw conclusions from the graph since the largest industry includes a wide array of businesses



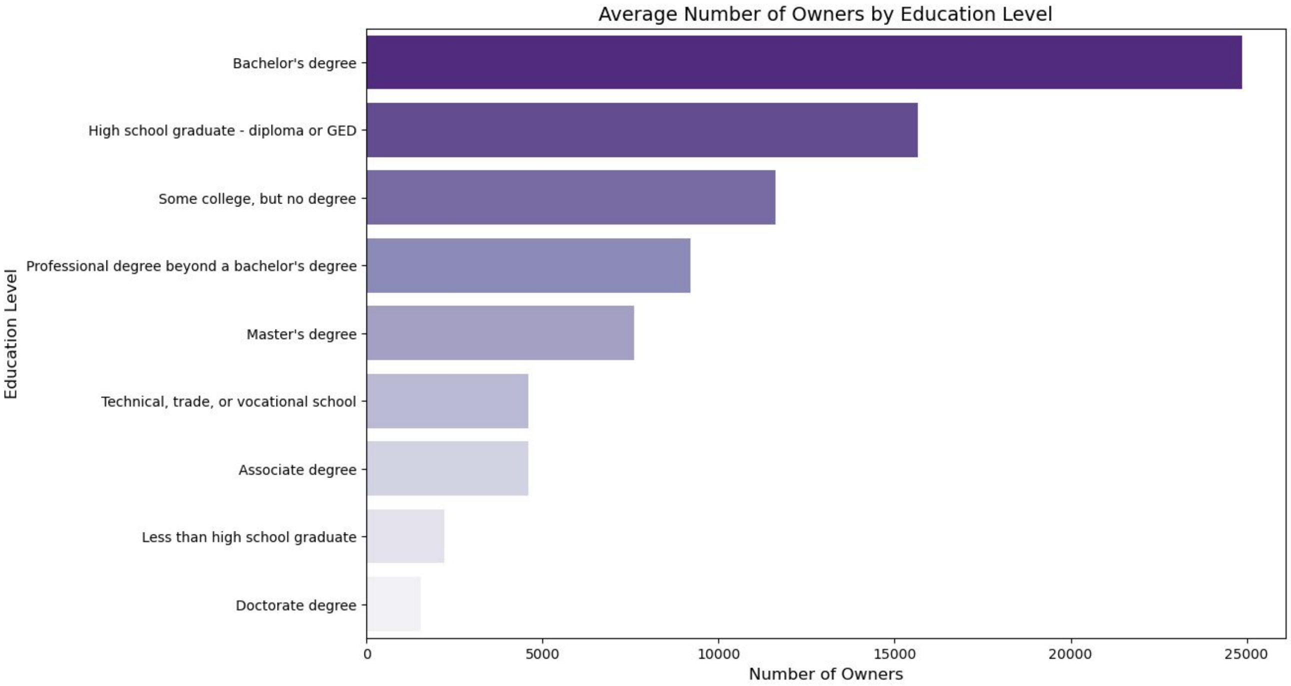
Is there are a large difference between the percentage of male versus female business owners?

- Filtered again to focus on the data for male vs female business owners
- You can immediately see that the difference is much larger than the one between the percentage of male vs female employees
- Used explode so that the pieces of the pie would pop out



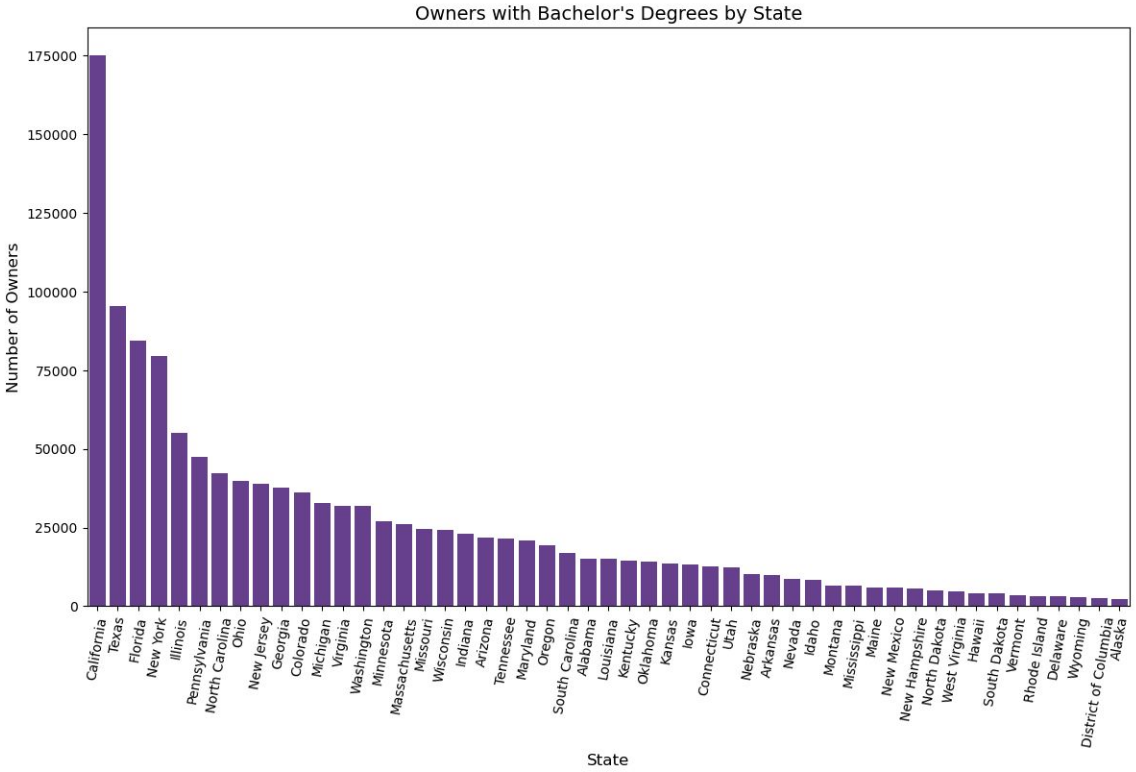
What is the educational background of business owners?

- Did another request to get data at the state level
- Used group by and the mean function to get the combined average for each education level



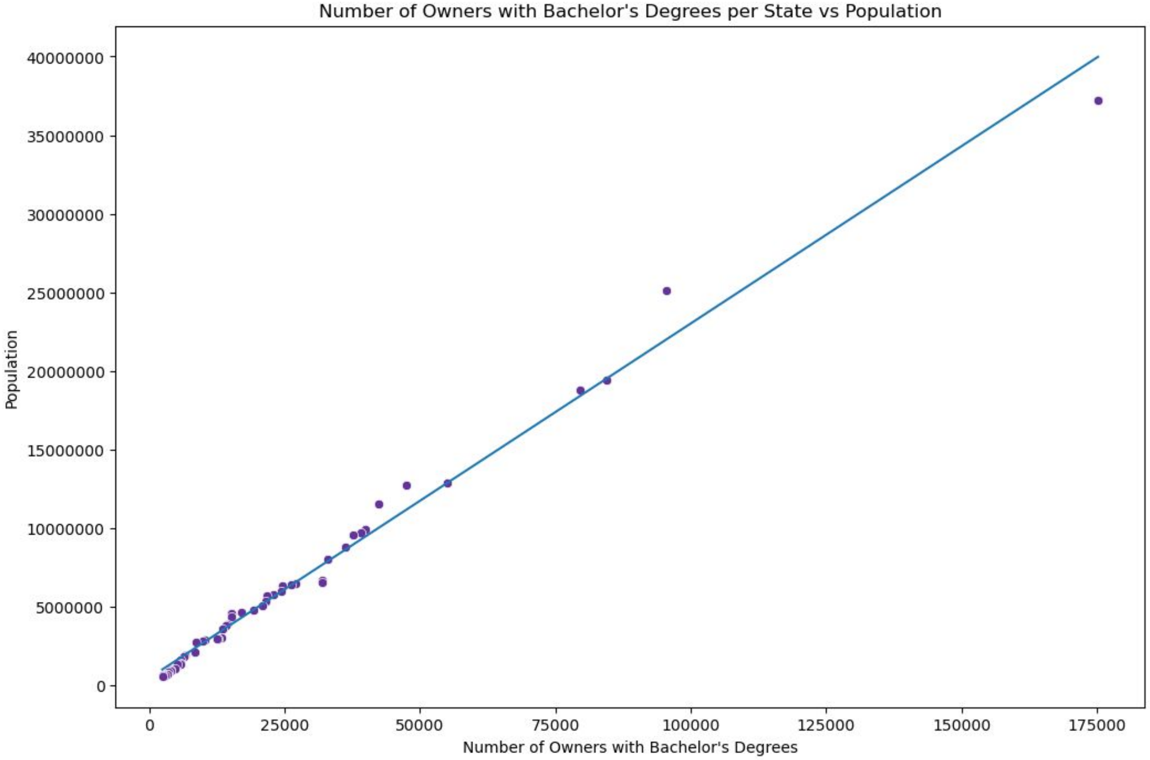
Is there a difference in number of owners with Bachelor's degrees at the state level?

- Wanted to see if there was a difference in the number of owners with Bachelor's degrees per state
- At first glance it looks like there is a large difference, but…



Comparing data from the previous graph to US Census population data

- When the data from the last slide is compared to state population, there is an r value of .99
- Tip: To get graph values out of scientific notation use “ticklabel_format (style = ‘plain’)”



Technology Characteristics of Businesses

1. Dropped rows where *NAICS2017_LABEL* equals *Total for all sectors*

```
indexNA = techuses[(techuses['NAICS2017_LABEL'] == 'Total for all sectors')].index
techuses.drop(indexNA , inplace=True)
```

2. To easily obtain the different levels of use for each particular technology, split column *Tech Use* into two columns named *Technology Used* and *Level of Use*

```
techuses[['Technology Used','Level of Use']] = techuses['Tech Use'].str.split(': ', expand=True)
techuses= techuses.drop(['Tech Use'], axis=1)
```

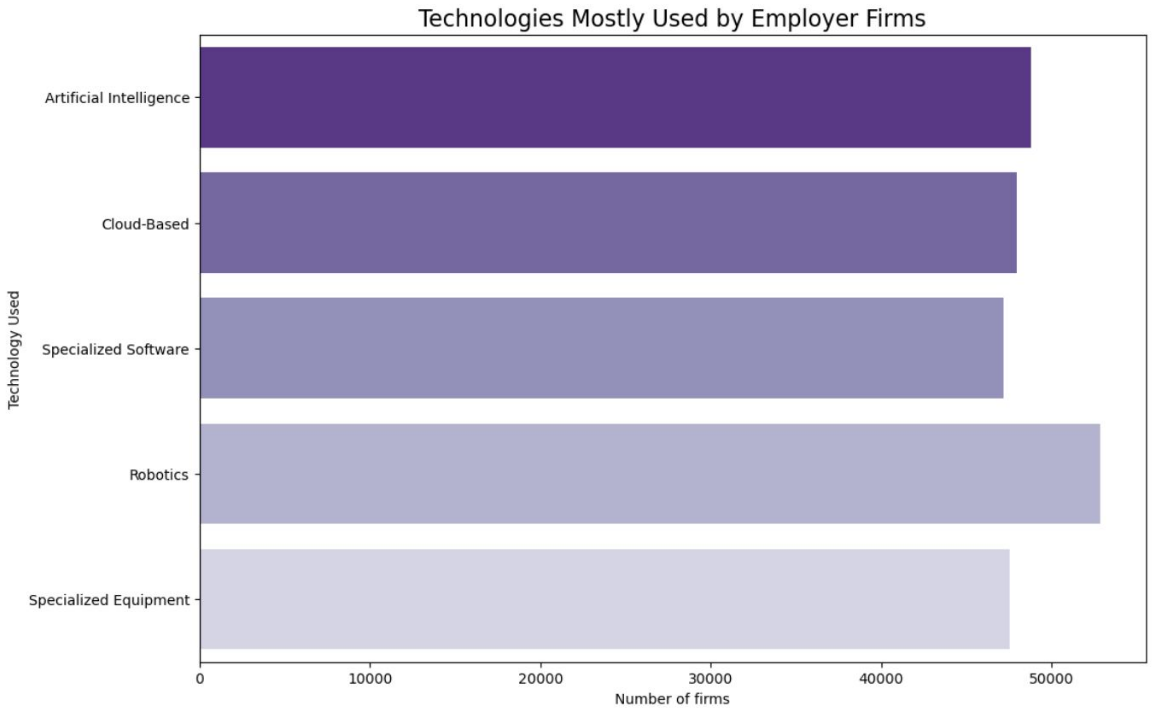
3. Exclude rows using the ~ isin( ) method.

```
techuses= techuses.loc[~techuses['Level of Use'].isin(['Total use','Total Reporting',"Don't know"])]
```

4. Calculated the Average Salary

```
techuses['Average Salary']= (techuses['Annual Payroll']/techuses['Number of Employees'])*1000
```

What technologies are mostly used by employer firms?

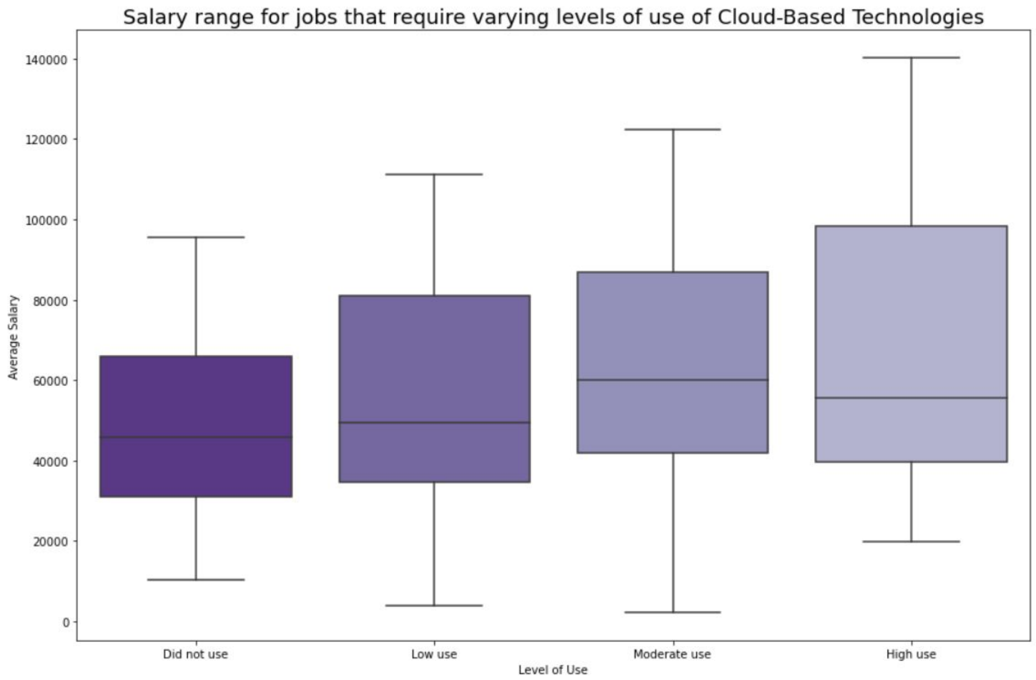


- We created a new DataFrame called firmtech that includes relevant columns from the techuse dataframe.
- Used firmtech as the data for our barplot.
- *Set confidence interval to None.*

```
sns.barplot(y= firmtech['Technology Used'], x=firmtech['Number of firms'], data=firmtech,
ci=None);
```

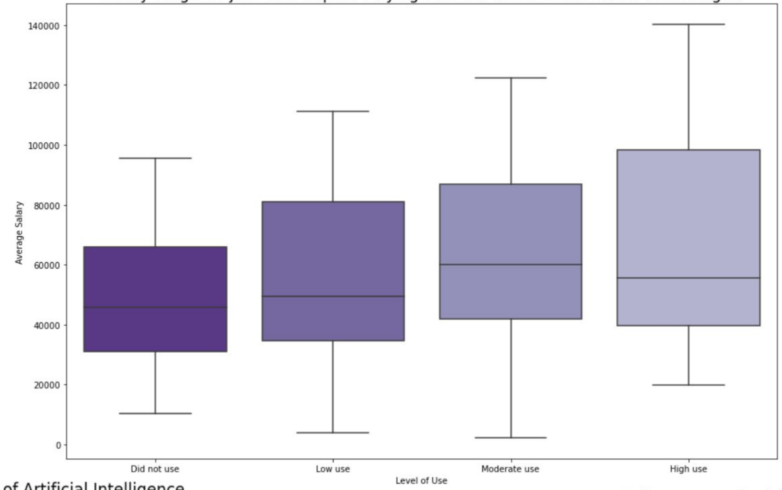
What is the salary range for jobs that require varying levels of use of software-based technology?

- Created new data frames for each technology, called ai, cloud, and software, respectively.
- Removed rows that are flagged as S or D in the PAYANN_F column.
- *Manually set the order according to the level of use with 'order='*

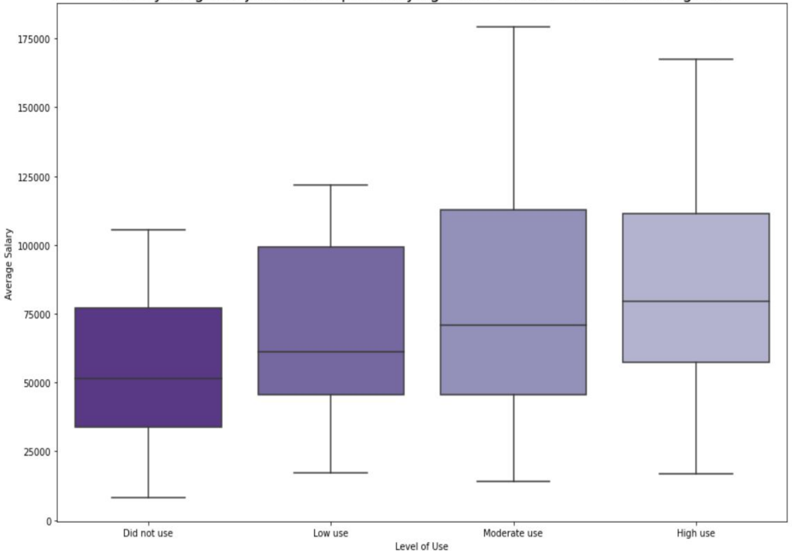



```
software_plot= sns.boxplot(x='Level of Use', y='Average Salary', data=software, order=['Did not use','Low use','Moderate use','High use'])
```

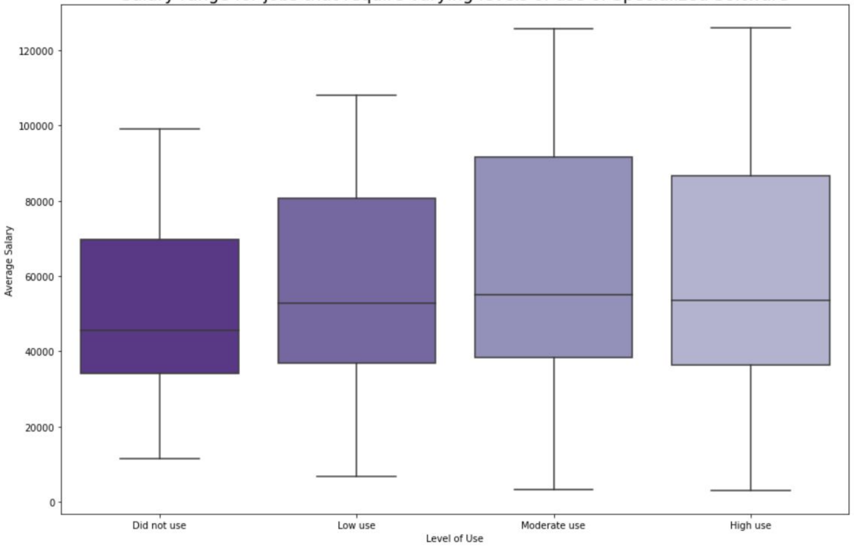
Salary range for jobs that require varying levels of use of Cloud-Based Technologies



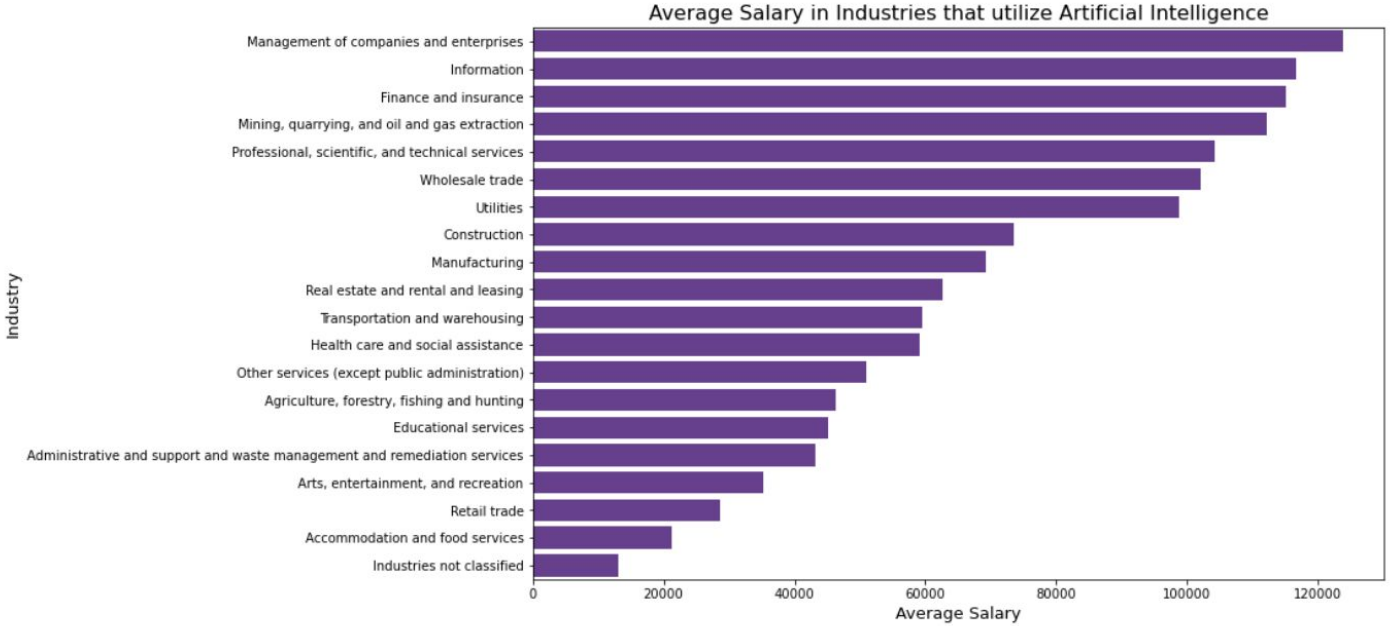
Salary range for jobs that require varying levels of use of Artificial Intelligence



Salary range for jobs that require varying levels of use of Specialized Software



In which industries are individuals who work with Artificial Intelligence likely to earn the highest salaries?



- We grouped the Technology Used, Industry, and aggregated Average Salary into a new dataframe called group_tech.
- Created a numpy array called ai_industry that contains the unique values for Industry from the ai DataFrame and sorted it.
- Created a new DataFrame called Payroll_ai by transposing a list of the group_tech and ai_industry data.

Code snippets

```
software= techuses[techuses['Technology Used'] == 'Specialized Software']
ai= techuses[techuses['Technology Used']== 'Artificial Intelligence']
cloud= techuses[techuses['Technology Used'] == 'Cloud-Based']
```

```
group_tech= ai.groupby(['Technology Used','Industry'])['Average Salary'].mean()
ai_industry= ai['Industry'].unique()
ai_industry.sort()
Payroll_ai= (pd.DataFrame([list(group_tech),ai_industry])).T
```

THANK YOU

ANY QUESTIONS