By: Tyler Panfil, Jack O'Connor, and Devin Nicholson
Date: 6/25/2021

US Census Hawaii Business Analysis with Python Visualizations

Source: Bureau, U. S. C. (2021, March 8). *Annual Business Survey (ABS) APIs*. The United States Census Bureau. https://www.census.gov/data/developers/data-sets/abs.2019.html.

Introduction

For this analysis of the Annual Business Survey US Census data we extracted information from Company Summary API, Characteristics of Businesses API, Technology Characteristics of Businesses API. The purpose was to explore each of the endpoints and see what the APIs could tell with the data, and see if there was any interconnection that could be made between them. We attempted to analyze the "Urban Honolulu, HI Metro Area" for all APIs, however, the Technology Characteristics of Businesses API did not have that level of granularity. So, we decided to look at "Urban Honolulu, HI Metro Area" for the Company Summary API and Characteristics of Businesses API. We also looked at the Technology Characteristics of Businesses API at the state level- Hawaii. All APIs and documentation can be found at: https://www.census.gov/data/developers/data-sets/abs.2019.html

Company Summary API-

The Company Summary API provided high-level data on businesses in the United States. This included information about the owners of the businesses and their demographics as well as specific numbers about the businesses located in each census block. The following table shows the variables that were requested from the API along with a brief description of each.

GEO_ID	The Geo_ID assigned to a specific area
NAME	Included Metro area and State name
NAICS2017	The NAICS2017 code
NAICS2017_LABEL	The NAICS2017 label that specified industry
SEX	The Sex code of the owners of a business
SEX_LABEL	The Sex Label of the owners of a business
ETH_GROUP	The Ethnicity code of the owners of a business

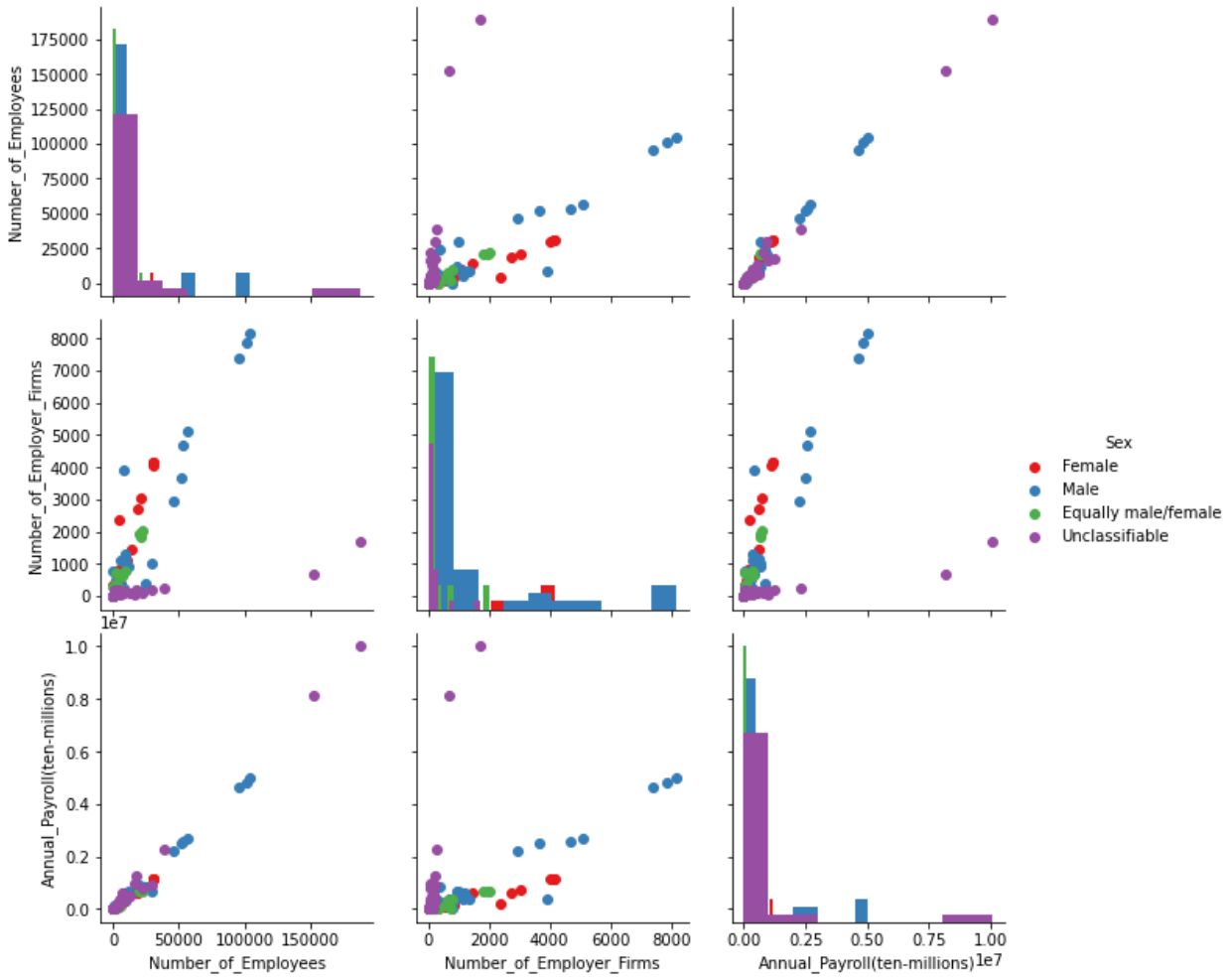
ETH_GROUP_LABEL	The Ethnicity Label of the owners of a business
RACE_GROUP	The Race code of the owners of a business
RACE_GROUP_LABEL	The Race Label of the owners of a business
VET_GROUP	The Veteran status code of the owners of a business
VET_GROUP_LABEL	The Veteran status Label of the owners of a business
FIRMPDEMP	The number of employer firms in the census block
FIRMPDEMP_F	Code used to group observations based on FIRMPDEMP
RCPPDEMP	Sales, value of shipments, or revenue of employer firms
RCPPDEMP_F	Code used to group observations based on RCPPDEMP
EMP	The number of Employees in the business
EMP_F	Code used to group observations based on EMP
PAYANN	The annual payroll for the business
PAYANN_F	Code used to group observations based on PAYANN
FIRMPDEMP_S	The standard error for the FIRMPDEMP variable
FIRMPDEMP_S_F	Code used to group observations based on FIRMPDEMP_S
RCPPDEMP_S	Standard Error of the RCPPDEMP variable
RCPPDEMP_S_F	Code used to group observations based on RCPPDEMP_S

EMP_S	Standard Error of the EMP variable
EMP_S_F	Code used to group observations based on EMP_S
PAYANN_S	Standard Error for the PAYANN variable
PAYANN_S_F	Code used to group observations based on PAYANN_S
metropolitan statistical area/micropolitan statistical area	To which metropolitan area an observation belonged

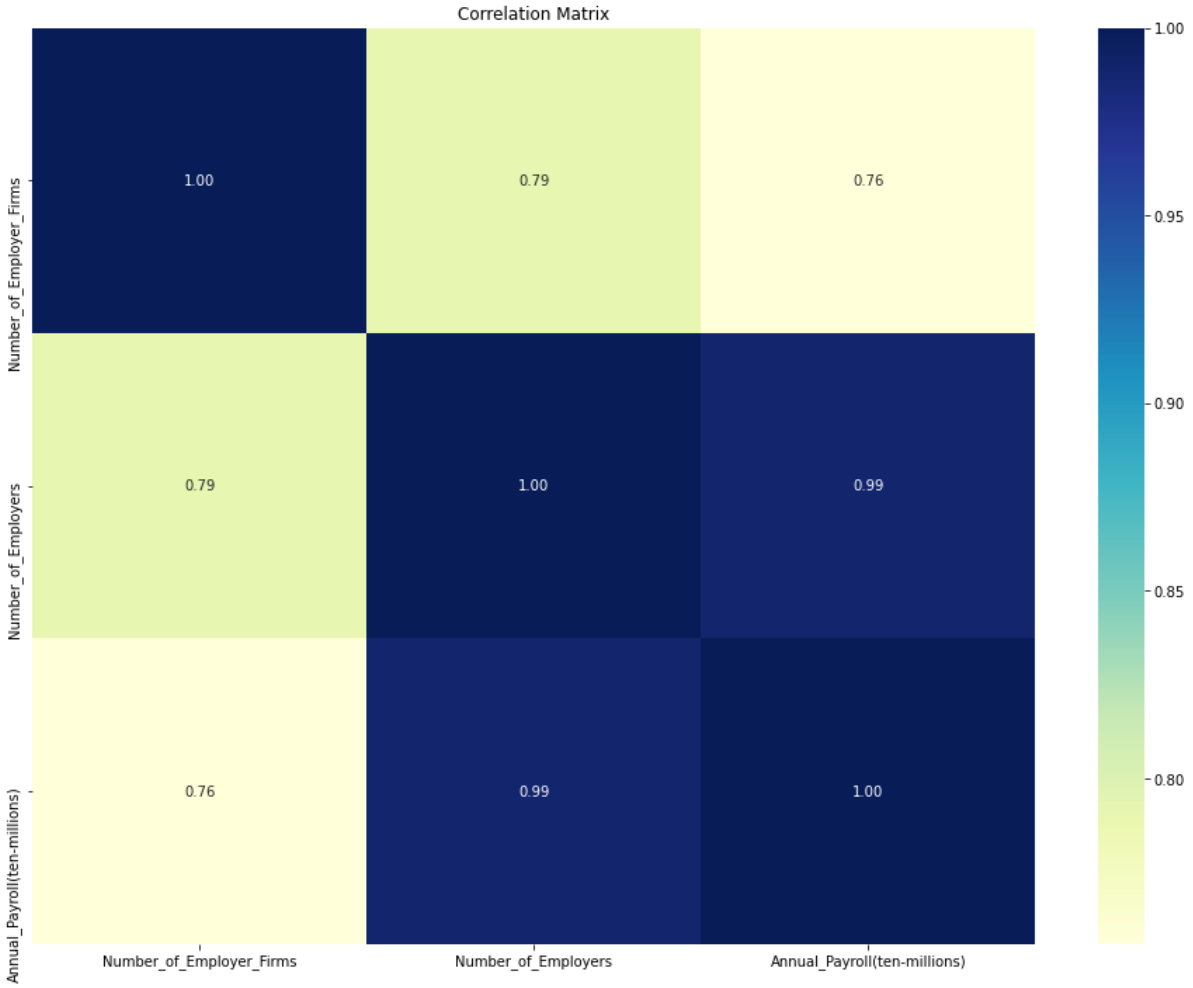
Several of the variables listed above were removed from the dataframe as they were either blank or were represented elsewhere in the data. For example, the variables that were codes (NAICS2017, SEX, ETH_GROUP, RACE_GROUP, VET_GROUP, FIRMPDEMP_F, EMP_F, PAYANN_F, FIRMPDEMP_S_F, RCPPDEMP_S_F, and PAYANN_S_F) were removed because the labels were more intuitive and represented their codified counterparts.

The original,unedited dataset was 227,248 observations long and the decision was made to whittle it down the the Urban Honolulu Metro Area. After the dataset was fileted on the NAME variable to include only those observations in Honolulu and the extra columns were removed, the final dataframe was only 752 observations in total.

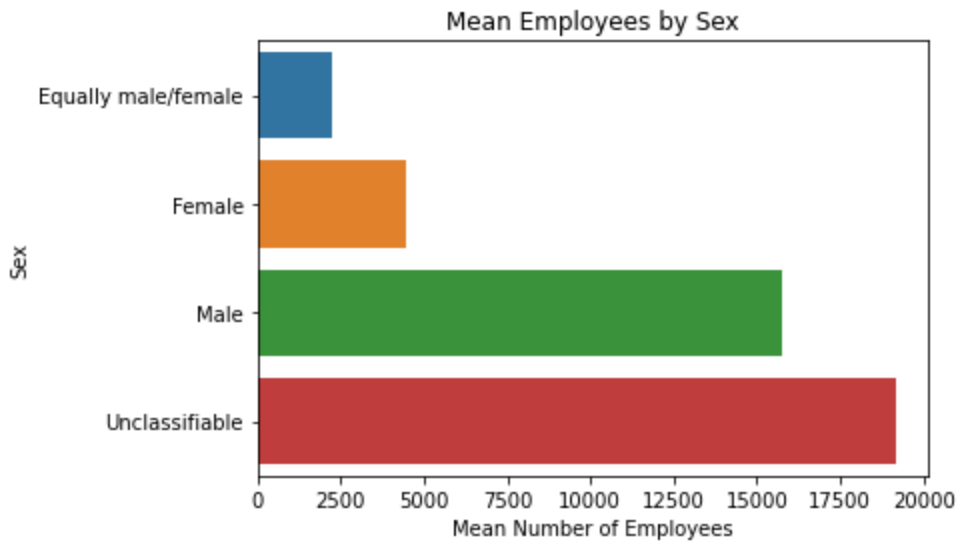
The points of interest from this API were the number of employees and annual payroll broken down by the gender status of the owners of the businesses in Honolulu. The following chart shows the number of employees, the number of employer firms and the annual payroll plotted against one another as well as the distribution of each variable along the diagonal.



There are obvious positive relationships between these variables and a few outliers in the distributions. The next chart shows the quantified relationships in a correlation matrix.



The final chart created for this API was a bar chart that broke down the average number of employees in companies owned by each gender status. The Unclassifiable category had the largest average number of employees followed by male-owned businesses.



Characteristics of Businesses API-

The first part of the process was to request an API call to get all the metropolitan areas in the dataset. The decision was made to pull in all the columns of the data to make sure nothing was filtered out before looking at all the data. After the API was called, it was formatted into JSON to be placed into a Pandas data frame. When putting the JSON data into the data frame, the first row was utilized for the header names. After the first row was removed for the headers, the rest was iterated through to create the rows of data. After the data was placed in a data frame, a glance at the data revealed multiple columns that were irrelevant. It also showed that there were 16,996 rows and 50 columns. The irrelevant columns were removed; once that was complete, the data was narrowed down to the Urban Honolulu metro area. This made a more manageable dataset at 45 rows and 43 columns. Then to see if the data could be broken down even more, the unique values were gathered from the NAICS2017_LABEL, which broke down the survey by sectors. Unfortunately, breaking the data set down to the metropolitan level provided only the Total for all the sectors in the characteristics of business API. The first thing visualized with the data was to see if there was any correlation between the numerical data in this API endpoint. Before this could be completed, the data type had to be identified and made sure it was either an int or float to perform these checks. The data type for each of the columns came up as objects. Before changing the types using astype, the columns were checked for which ones would be helpful for a correlation matrix. After reviewing all the fields, it was concluded that the only areas worth further investigation were the number of employees (EMP), number of employer firms (FIRMPDEMP), and the annual payroll (PAYANN) columns. These were all cast to float types using .astype(*float*). The data type was rechecked to make sure the change went through. The next step was to change these column names to something more meaningful before the visualizations were made. (FIRMPDEMP) was changed to the number of employer firms; (EMP) was changed to the number of employees, and (PAYANN) was changed to annual payroll ($10-Millions).

The second step in learning more about this endpoint was to correlate the three fields that were cast to floats. This showed that all three areas had a positive correlation with each other,

whether it was the number of employees and annual payroll or number of employees and number of employer firms.

	Number of employer firms	Number of Employees	Annual payroll ($10-Millions)
Number of employer firms	1.000000	0.947510	0.934994
Number of Employees	0.947510	1.000000	0.998126
Annual payroll ($10-Millions)	0.934994	0.998126	1.000000

After seeing the correlation between the three fields, the data needed to be represented better than just a set of numbers. An annotated heatmap was used to communicate the positive correlation values better. Using this visual, it can be seen that the number of employees and annual payroll has the highest positive correlation at 1.00. The first version of the annotated heatmap did not have enough contrast between the fields to make any discerning difference. The manual entry of having a minimum number of -1 on the gradient was removed in the second version (See Figure 1). This provided a higher contrast between the fields that were being compared. At first glance, this can be misleading when seeing the colors, even though all three have a positive linear correlation. To help get past the color difference, it was decided to create a grid of graphs for the three fields (See Figure 2). This consisted of scatterplots on the three comparisons happening in the correlation to show the positive linear correlation. This also provided a quick way to visualize if there were any outliers. Then going diagonally through the set of graphs, there are histograms. Having the histograms provided a quick way to see where most of the businesses that took part in the annual business survey fell. Looking at each of the histograms the most of the companies fell in the first bin. This could partially be due to the first bin, including zero. The zero in these columns for these data fields meant either the company did not report or chose not to submit a value.

Number of Employer Firms vs Number of Employees vs Annual Payroll

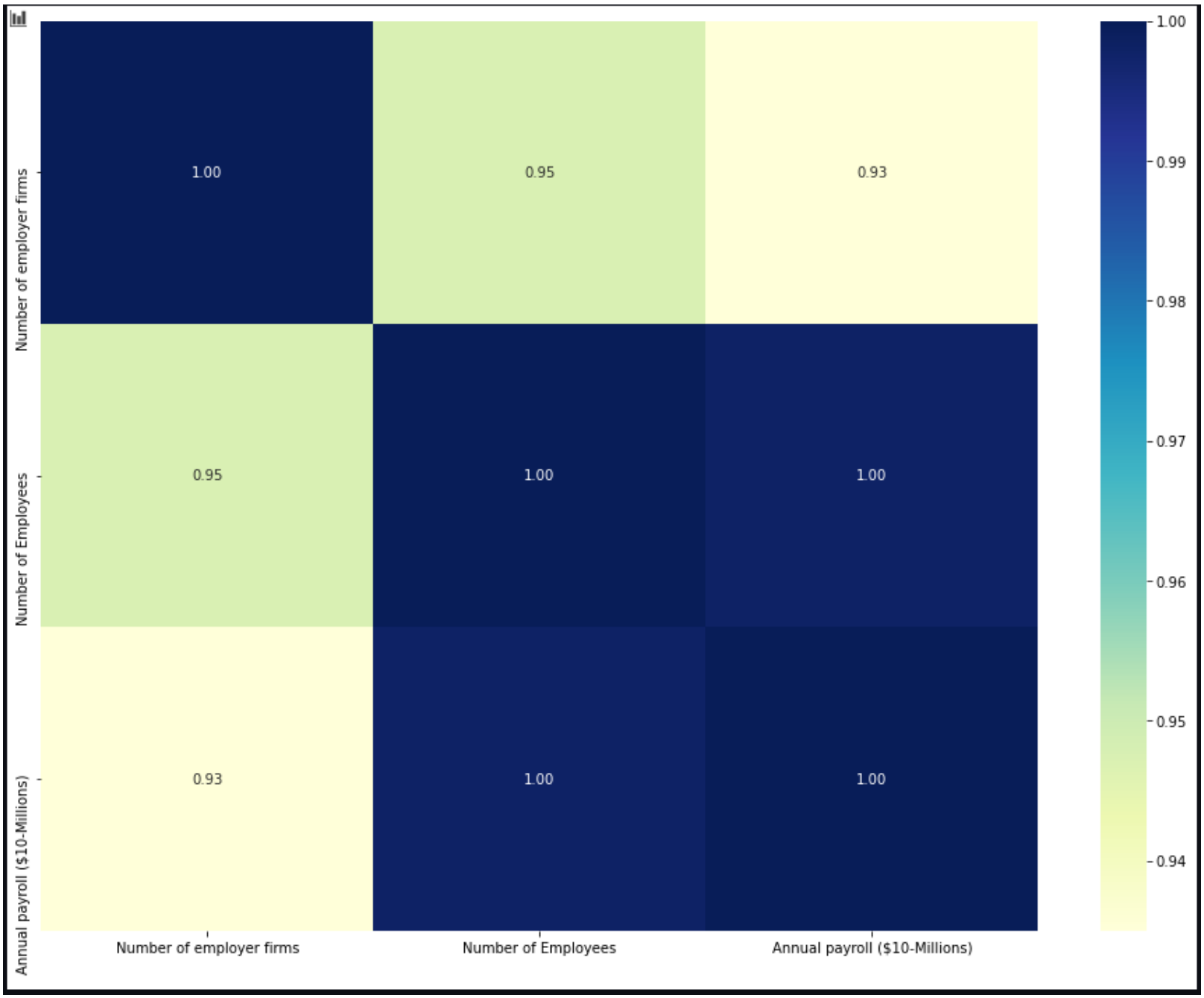


Figure 1

Number of Employer Firms vs Number of Employees vs Annual Payroll
Histograms on the Diagonal of each of the three fields
Scatterplots for each of the three comparisons

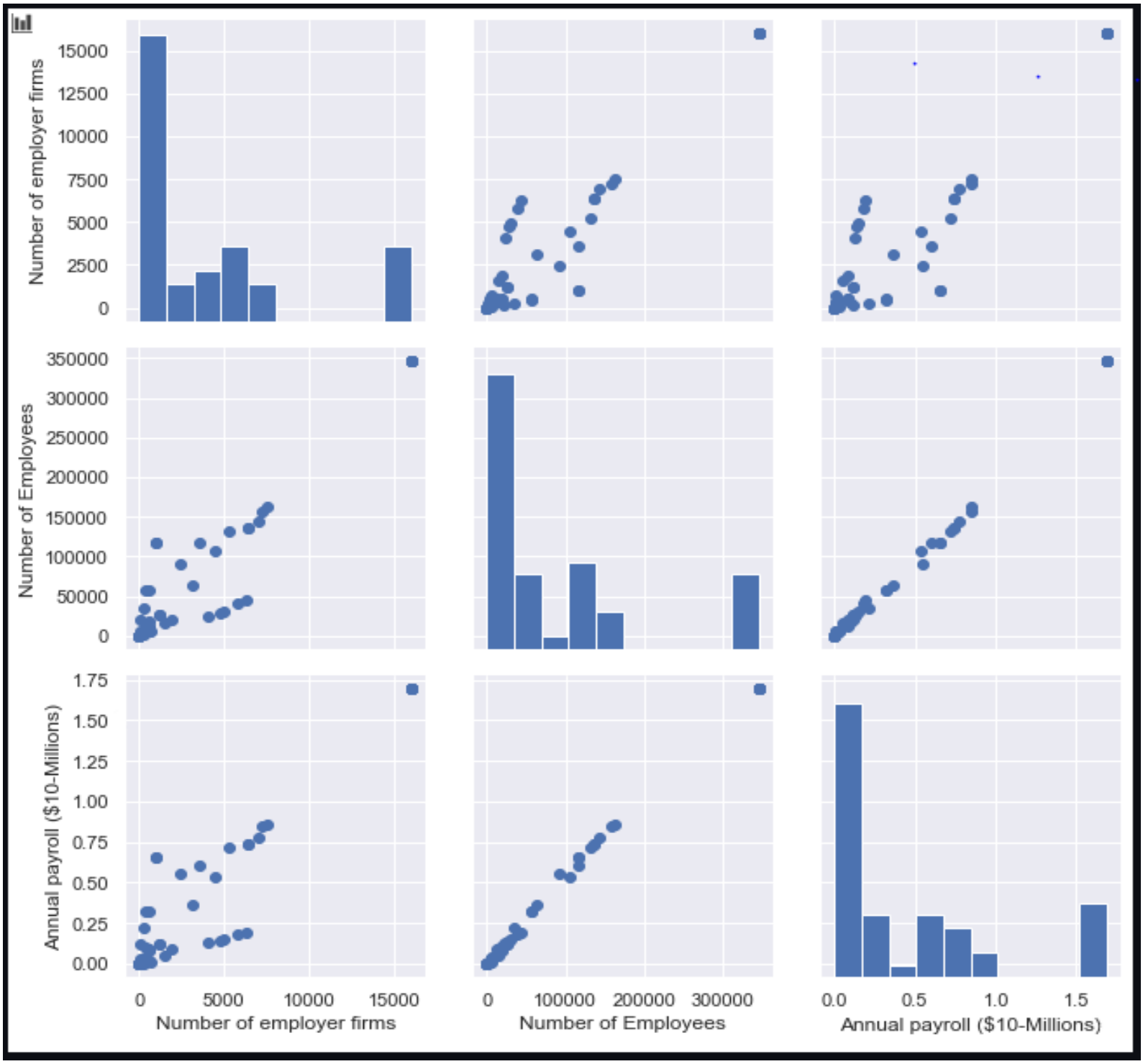
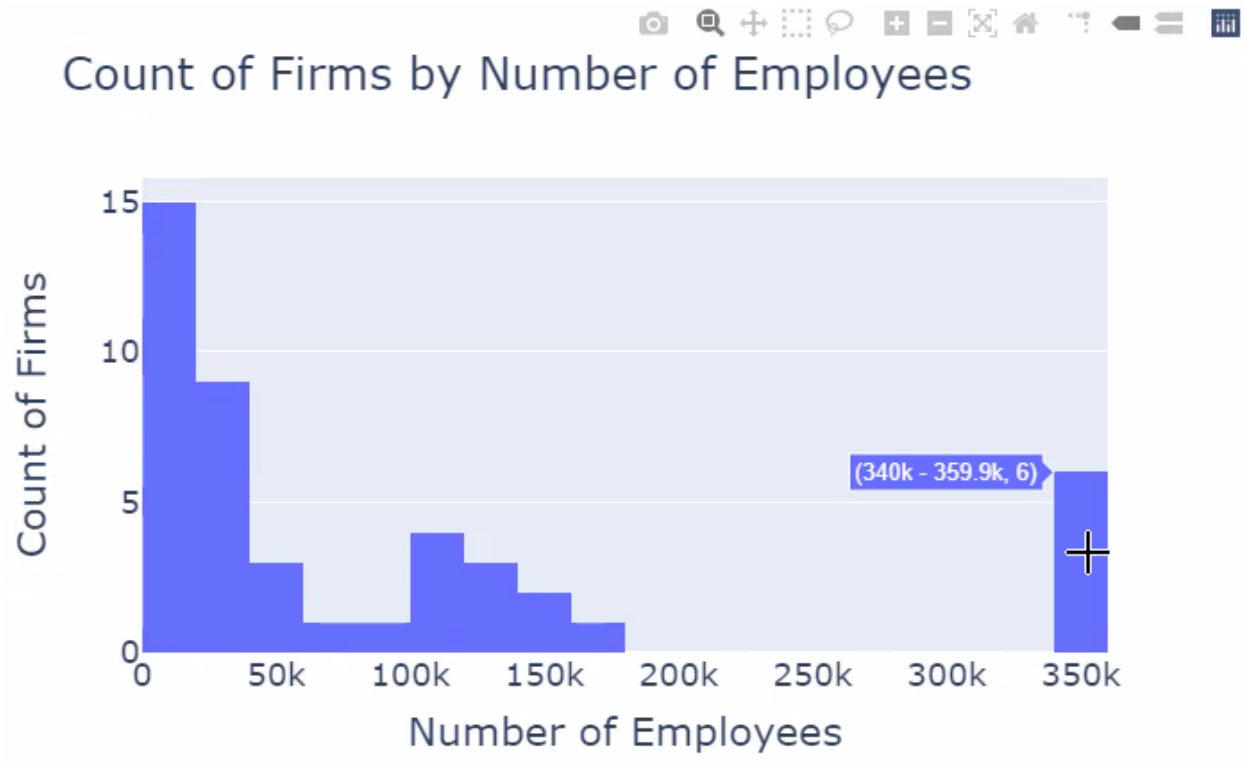
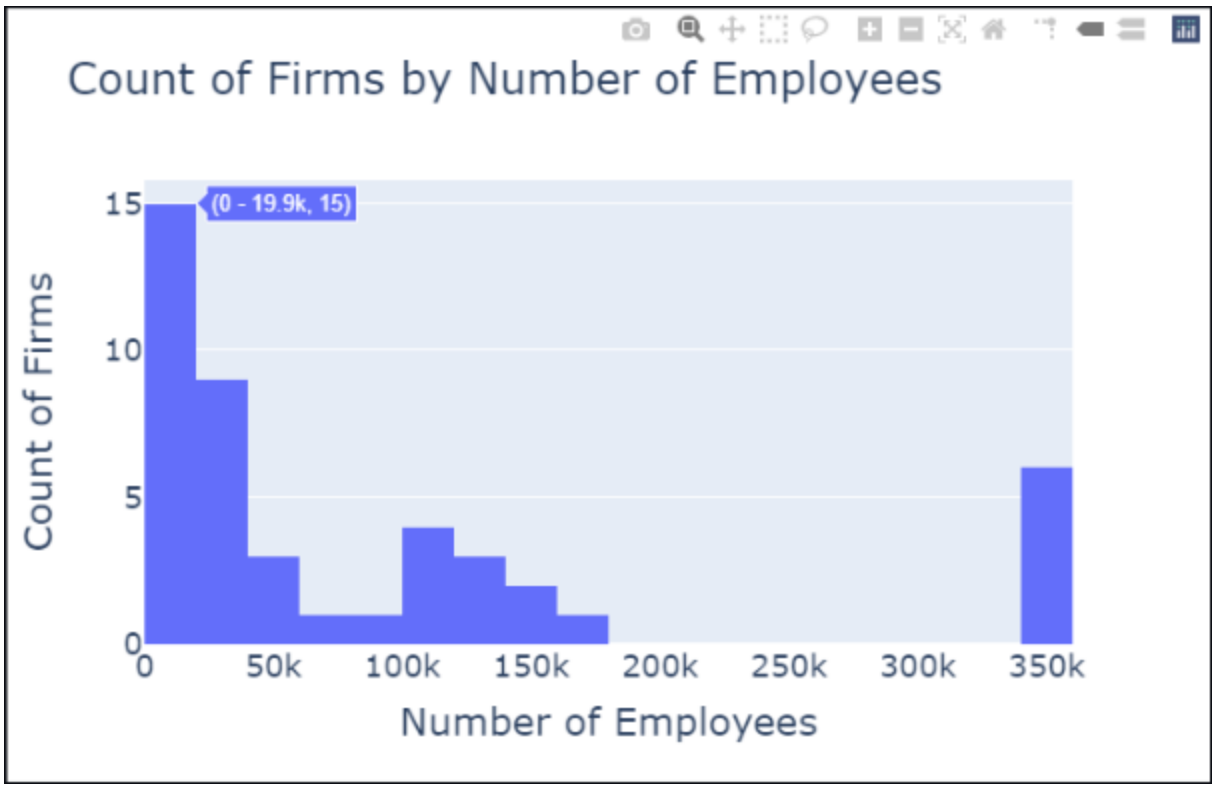


Figure 2

Looking at each of the scatter plots and histograms, it can be seen that there is an outlier. So to focus on this outlier, it was decided to look into the number of employees more closely. A histogram of the count on the number of employees was created with Plotly Graph Objects. This was made with 20 sets of bins. Once created, it can be seen that six businesses reported the number of employees in the firm between 340,000 to 359,999. After seeing this, the max value was checked in the number of employees column, 346,144 employees for the largest business in Urban Honolulu. Due to six firms reporting in this bin, it was decided to leave these rows in the dataset. This is because it makes up about 13 percent of the remaining dataset that the visualizations were made from. Looking at the histogram, it can be seen that the majority of our firms fall in the first two bins, with 15 companies in the first bin of 0 to 19,999 employees and nine companies in the second bin of 20,000 to 39,999 employees.

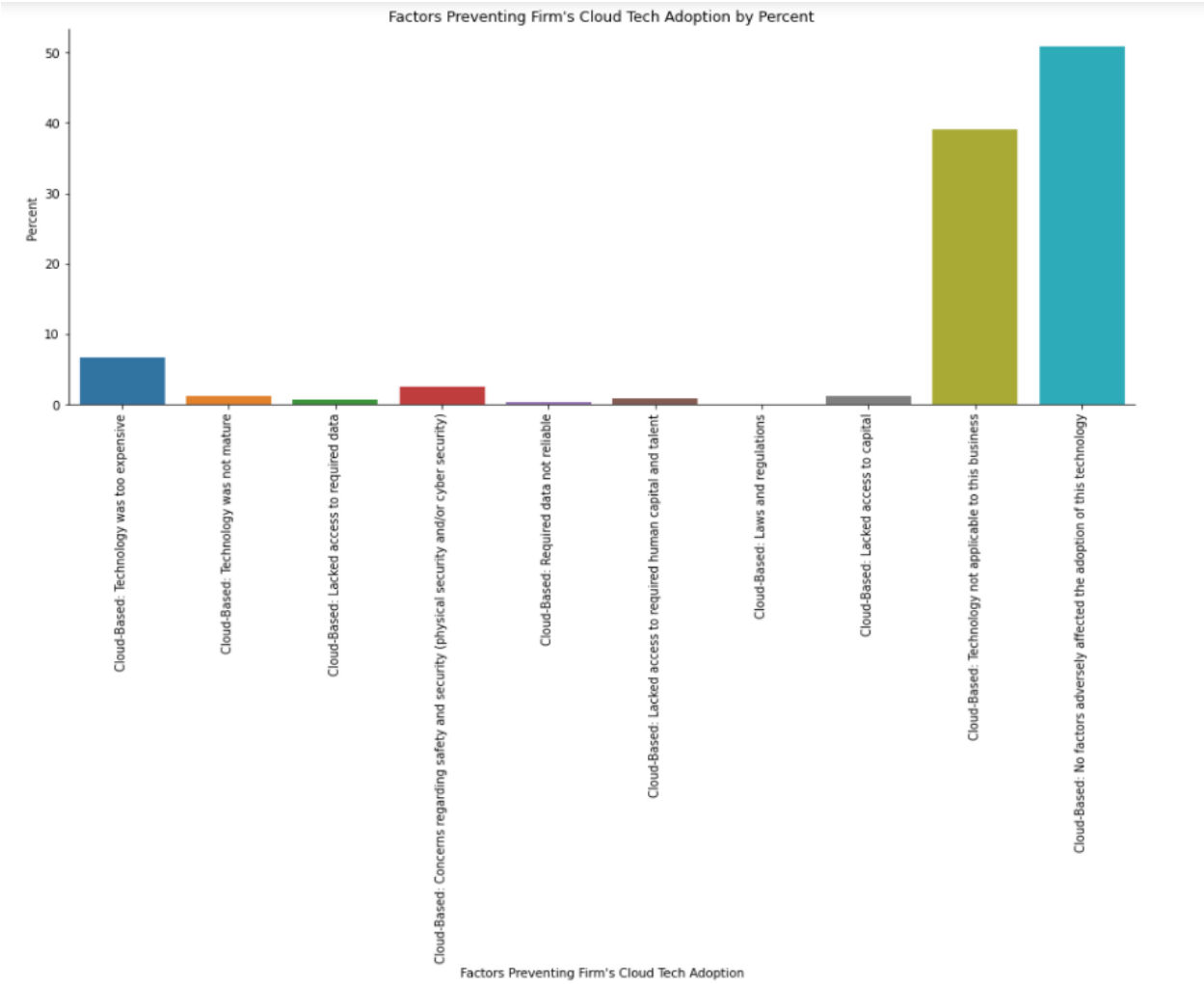




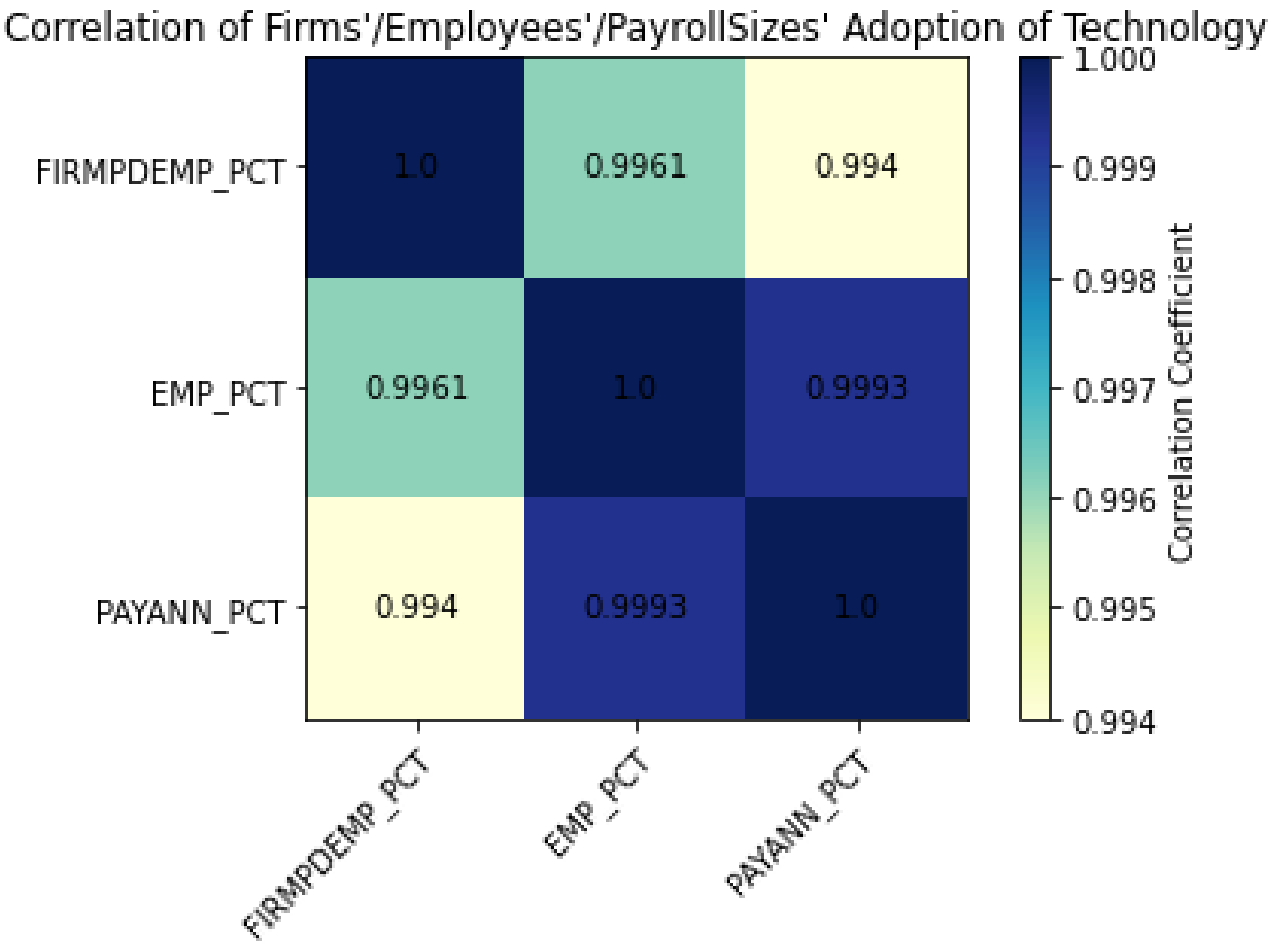
Technology Characteristics of Businesses API

The first step was to get the data. The data was pulled into a Jupyter notebook with the requests library, turned into JSON, and put into a pandas data frame with the first JSON piece set as the metadata. It was decided that this API would not be analyzing the breakdowns unless more visualizations were needed. So a concise data frame was created that contained all firms, all sexes, all ethnicities, all races, and all sectors. This brought the data frame down from 1155 to 60. When looking at this data frame, total reporting was a factor in the data, which due to using percents for this analysis and the fact that it was obvious that 100% firms who responded...responded, we removed these rows by identifying them with a lambda function. This function found rows that ended with “Total Reporting” and created a True/False column accordingly. Then, all True rows were removed from the dataframe. Now the categorical differences between all the rows were the 11 potential factors(mutually exclusive) for not adopting every 5 different business technologies. There was numeric data that corresponded with each of these 55 combinations as well. However, the numeric data was registering as an object. This was fixed with a simple expression using astype(float). The first visualization narrowed in on the factors preventing the adoption of Cloud-Based Technology. Using a lambda

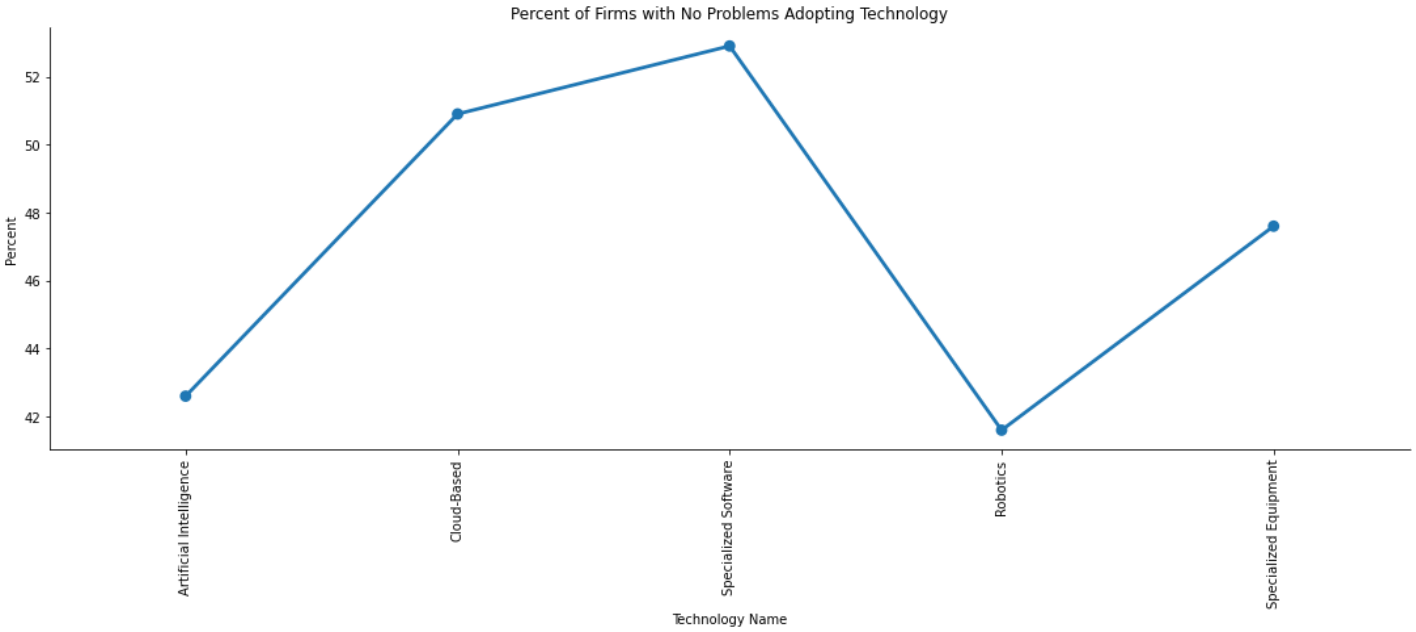
function, the factors were individually tested for whether or not they started with “Cloud-Based”. This test returned True/False booleans in a new test column. Then, all rows with False were removed. Now that the data frame was down to just 11 rows of Cloud-Based Factors, a bar graph was created to show firms’ main factor that prevented them from adopting Cloud-Based technology.



Two responses were much more popular than the others. 1. No factors and 2. Cloud-Based Technology wasn’t relevant to the business. Everything else was under 10%. Next, a correlation matrix was created to see if firms, employees, and payroll needed to be analyzed separately for factors preventing the adoption of technologies. If they were very very highly correlated with each other, then they would be practically interchangeable in our study of factors. If they are practically interchangeable, only one of them should be analyzed and the other 2 will have results so similar they are not worth exploring separately. To do this correlation matrix, the 55-row data frame with multiple technologies was used. Specifically the firms, employees, and payroll responses to the technology prevention factors(in percent). Here is the result:



In this graph, the weakest correlation is .994, which is almost a perfect correlation. Therefore, we concluded that these columns are so highly correlated that only one needs to be analyzed. Due to the Firms already being used for analysis above, Firms would be the 1 column of the 3 to be analyzed. Lastly, this study wanted to determine which technologies were being adopted the most. Starting with the 55-row Data Frame with all the technologies and questions, a lambda function was used to create a “test” column that returned True if a row had a “No factors adversely affected the adoption of this technology” at the end of the Factors column” and a False otherwise. Then, all the rows with “False” in the test column were removed. This resulted in a 5 row data frame where rows had responded with “No factors adversely affected the adoption of this technology” for each technology. Here is the graph created from this data:



This graph makes it clear that all technologies were adopted at 41-53%, with Robotics being the least adopted technology and Specialized Software being the most adopted technology.