LIS705
PROJECT ASSIGNMENT 5

By: Prachi, Rushali & Bhavya

Date: 05/09/2023

Changes from Previous Reports:

We have only changed the research questions since the project assignment 1,2 and 4 Here are the updated research questions which aligns more closely with the requirements of the assignment (as per assignment 4).

Revised Research Questions:

1. What is the relationship between revenue and profit for the top 5 companies in the United States in the top 2 sectors which have the highest profit.Can any patterns or trends be identified in this relationship?
2. How does the presence of a female CEO affect the revenue and profit of companies ? Are there any significant differences in performance between male-led and female-led companies, and if so, what factors may contribute to these differences?
3. How do the top 50 companies in the dataset compare in terms of market capitalization, revenue, and profit?
4. How have the rankings of the top 20 companies changed over time, and what are the current rankings and changes in rankings of these companies?

Data, Cleaning and Preparation:

The dataset for this project was acquired from Kaggle, which provided information about the top 1000 companies based on revenue for the year 2021. It contains information such as the company name, its rank in the list, the revenue and profit generated, the number of employees, the sector it operates in, the state where it is headquartered, the presence of a female CEO, the company's previous rank, its market capitalization, revenue and profit percentage change, and its assets.

The dataset also includes some derived columns such as profit margin, revenue per employee, revenue growth, and profit growth, which can provide further insights into the performance of the companies.

Explanation of each column :

- Company: The name of the company.
- Rank: The rank of the company in the list based on its revenue.
- Rank_change: The change in the rank of the company compared to the previous year.
- Revenue: The revenue generated by the company in millions of dollars.
- Profit: The profit generated by the company in millions of dollars.
- Num. of employees: The number of employees working in the company.
- Sector: The sector in which the company operates.
- State: The state where the company is headquartered.
- CEO_woman: A binary variable indicating whether the CEO of the company is a woman or not.
- Prev_rank: The rank of the company in the previous year's list.
- Market Cap: The market capitalization of the company in millions of dollars.
- Revenue_percent_change: The percentage change in the company's revenue compared to the previous year.
- Profits_percent_change: The percentage change in the company's profit compared to the previous year.
- Assets: The total assets of the company in millions of dollars.
- Profit_margin: The profit margin of the company, which is the ratio of profit to revenue.
- Revenue_per_employee: The revenue generated per employee in thousands of dollars.
- Revenue_growth: The growth rate of revenue for the company.
- Profit_growth: The growth rate of profit for the company.

(last 4 columns are derived columns)

During the data cleaning and preparation phase, several procedures were performed to ensure that the dataset was accurate and consistent. Firstly, missing values were checked and replaced through research to ensure that the values were accurate and consistent with other fields.

Duplicates were also checked and removed to ensure that the dataset contained only unique entries.To make the financial data more readable and comparable, the dollar sign was removed from columns that represented financial information. The number of nulls in the dataset was also checked and addressed accordingly. We also checked the outliers and since it is a financial data, it was expected to have some outliers and hence, we did not remove the outliers.

As a result of these cleaning and preparation steps, the dataset was ready for further analysis. Additionally, some derivative data was prepared, such as change in rank between the current year and the previous year. This data was necessary to understand how each company had fared compared to the previous year and to identify the factors that were driving the growth or decline of a particular industry.

BACKGROUND

The dataset provides information about the top 1000 companies based on revenue for the year 2021, including essential financial information such as revenues, profits, assets, and market value. The number of employees for each company is also included, providing insights into the job market for different industries. Two of the research questions aim to analyze the performance of companies in different industries and understand the factors that contribute to their success or failure. The first question examines the relationship between different financial metrics and the number of employees, while the second question focuses on the impact of gender diversity on company performance.

The other two research questions are related and aim to analyze the performance of companies in the dataset. The third question focuses on the top 50 companies in the dataset and compares their performance in terms of market capitalization, revenue, and profit. This analysis can provide insights into which companies are the most financially successful and which factors contribute to their success. The fourth question examines the changes in rankings of the top 20 companies over time and aims to understand the current rankings and changes in rankings of these companies. This analysis can provide insights into which companies are rising in the rankings and which are falling, and what factors are contributing to these changes.

Overall, the dataset offers a wealth of information that can provide insights into current economic trends and market conditions. By analyzing this data, we can identify the top-performing industries, understand the factors driving their success, and identify areas where improvements can be made.

Research Questions Analysis:

1. **What is the relationship between revenue and profit for the top 5 companies in the United States in the top 2 sectors which have the highest profit.Can any patterns or trends be identified in this relationship?**

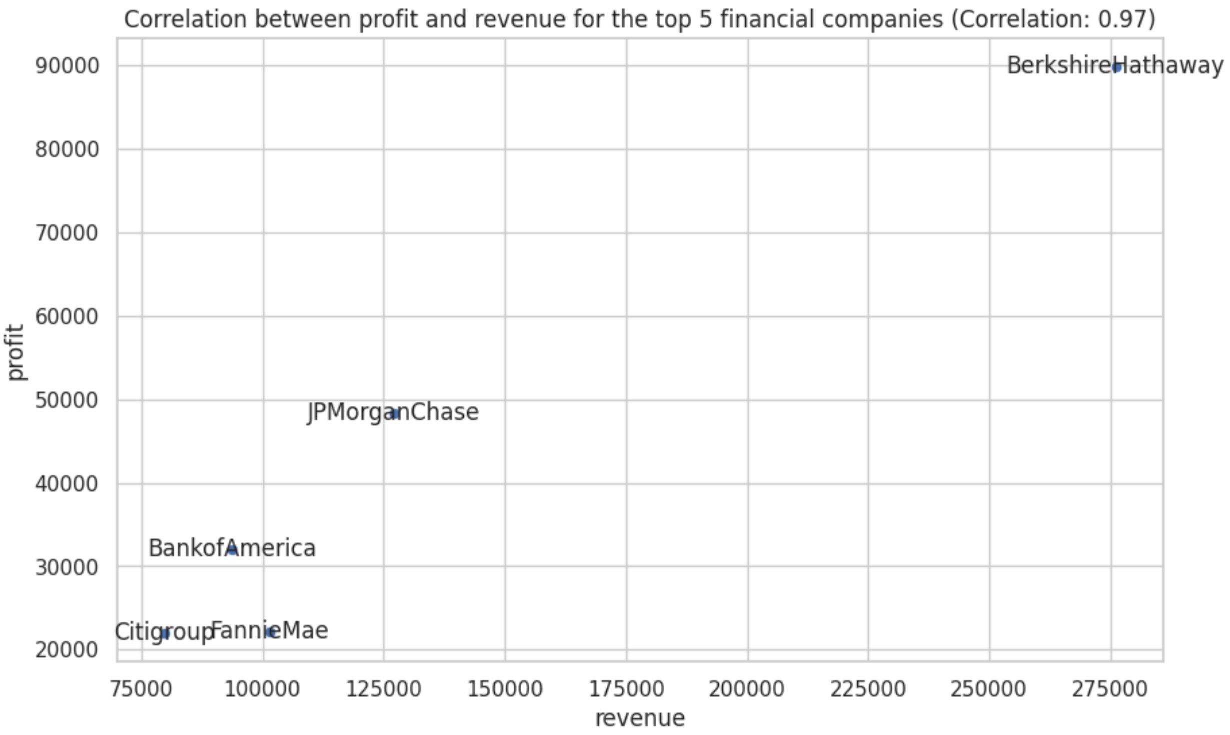
We chose to explore the relationship between revenue and profit for the top 5 companies in the top 2 sectors with the highest profit in the United States. To answer this research question, We first needed to identify the top 2 sectors based on their profit, which turned out to be Financials and Technology.

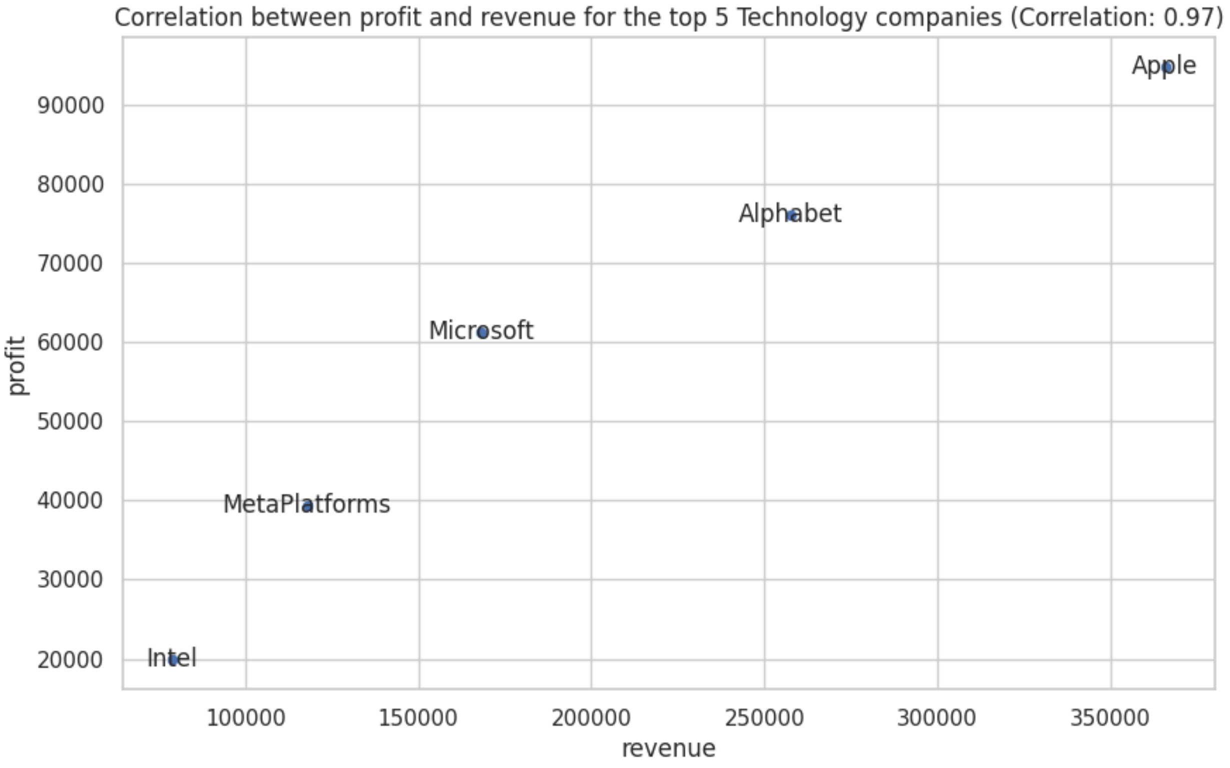
Next, we calculated the correlation between revenue and profit for the top 5 companies in each sector. The results showed that there is a significant positive correlation between revenue and profit for the top 5 companies in both sectors. This means that as revenue increases, profit also increases for these companies.

To further validate our findings, we performed a p-value test with a significance level of 0.01. Therefore, we reject the null hypothesis and conclude that there is a significant correlation between revenue and profit for the top 5 companies in Financials & Technology.

To visualize this relationship, we created a scatter plot of revenue vs profit for each of the top 5 companies in the Financials and Technology sectors. The scatter plot clearly shows the positive correlation between revenue and profit for these companies.

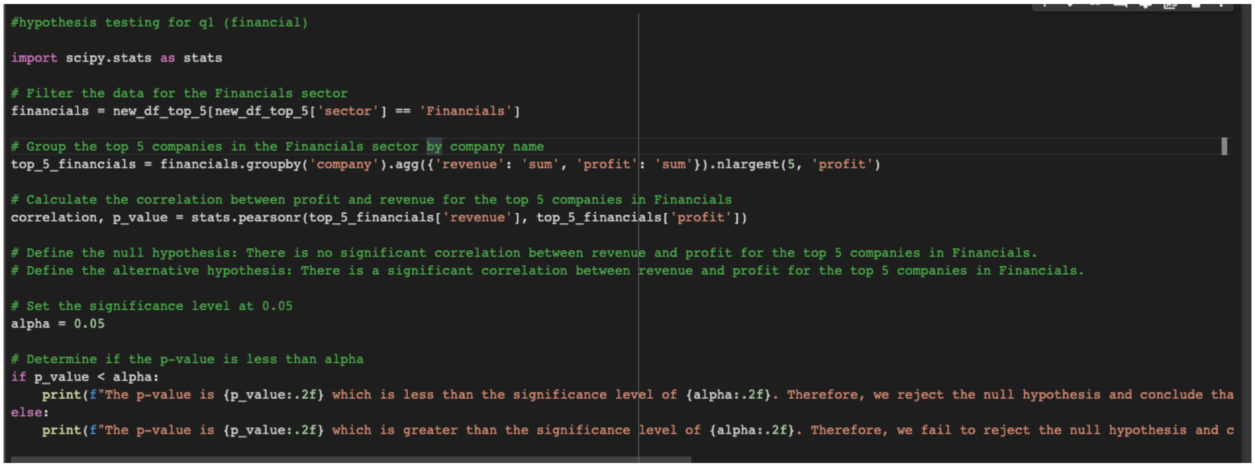
Based on these results, we can conclude that there is a strong positive relationship between revenue and profit for the top 5 companies in the Financials and Technology sectors in the United States. As revenue increases, profit also increases for these companies. This information can be valuable for businesses and investors looking to invest in these sectors or understand the financial performance of these top companies.


Correlation between profit and revenue for the top 5 financial companies (Correlation: 0.97)
profit
revenue
90000
80000
70000
60000
50000
40000
30000
20000
75000
100000
125000
150000
175000
200000
225000
250000
275000
BerkshireHathaway
JPMorganChase
BankofAmerica
Citigroup
FannieMae


Correlation between profit and revenue for the top 5 Technology companies (Correlation: 0.97)
profit
revenue
90000
80000
70000
60000
50000
40000
30000
20000
100000
150000
200000
250000
300000
350000
Apple
Alphabet
Microsoft
MetaPlatforms
Intel

Statistical Test:

For Financials:

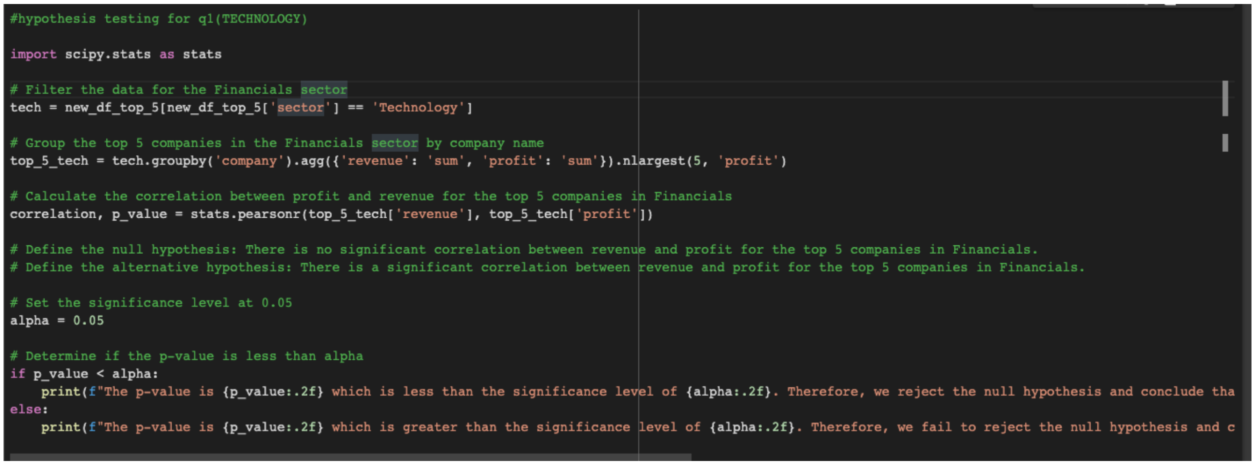


```
#hypothesis testing for q1 (financial)

import scipy.stats as stats

# Filter the data for the Financials sector
financials = new_df_top_5[new_df_top_5['sector'] == 'Financials']

# Group the top 5 companies in the Financials sector by company name
top_5_financials = financials.groupby('company').agg({'revenue': 'sum', 'profit': 'sum'}).nlargest(5, 'profit')

# Calculate the correlation between profit and revenue for the top 5 companies in Financials
correlation, p_value = stats.pearsonr(top_5_financials['revenue'], top_5_financials['profit'])

# Define the null hypothesis: There is no significant correlation between revenue and profit for the top 5 companies in Financials.
# Define the alternative hypothesis: There is a significant correlation between revenue and profit for the top 5 companies in Financials.

# Set the significance level at 0.05
alpha = 0.05

# Determine if the p-value is less than alpha
if p_value < alpha:
    print(f"The p-value is {p_value:.2f} which is less than the significance level of {alpha:.2f}. Therefore, we reject the null hypothesis and conclude tha
else:
    print(f"The p-value is {p_value:.2f} which is greater than the significance level of {alpha:.2f}. Therefore, we fail to reject the null hypothesis and c
```

Output:

The p-value is 0.01 which is less than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is a significant correlation between revenue and profit for the top 5 companies in Financials.

For Technology:



```
#hypothesis testing for q1(TECHNOLOGY)

import scipy.stats as stats

# Filter the data for the Financials sector
tech = new_df_top_5[new_df_top_5['sector'] == 'Technology']

# Group the top 5 companies in the Financials sector by company name
top_5_tech = tech.groupby('company').agg({'revenue': 'sum', 'profit': 'sum'}).nlargest(5, 'profit')

# Calculate the correlation between profit and revenue for the top 5 companies in Financials
correlation, p_value = stats.pearsonr(top_5_tech['revenue'], top_5_tech['profit'])

# Define the null hypothesis: There is no significant correlation between revenue and profit for the top 5 companies in Financials.
# Define the alternative hypothesis: There is a significant correlation between revenue and profit for the top 5 companies in Financials.

# Set the significance level at 0.05
alpha = 0.05

# Determine if the p-value is less than alpha
if p_value < alpha:
    print(f"The p-value is {p_value:.2f} which is less than the significance level of {alpha:.2f}. Therefore, we reject the null hypothesis and conclude tha
else:
    print(f"The p-value is {p_value:.2f} which is greater than the significance level of {alpha:.2f}. Therefore, we fail to reject the null hypothesis and c
```

Output:

The p-value is 0.01 which is less than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is a significant correlation between revenue and profit for the top 5 companies in Financials.

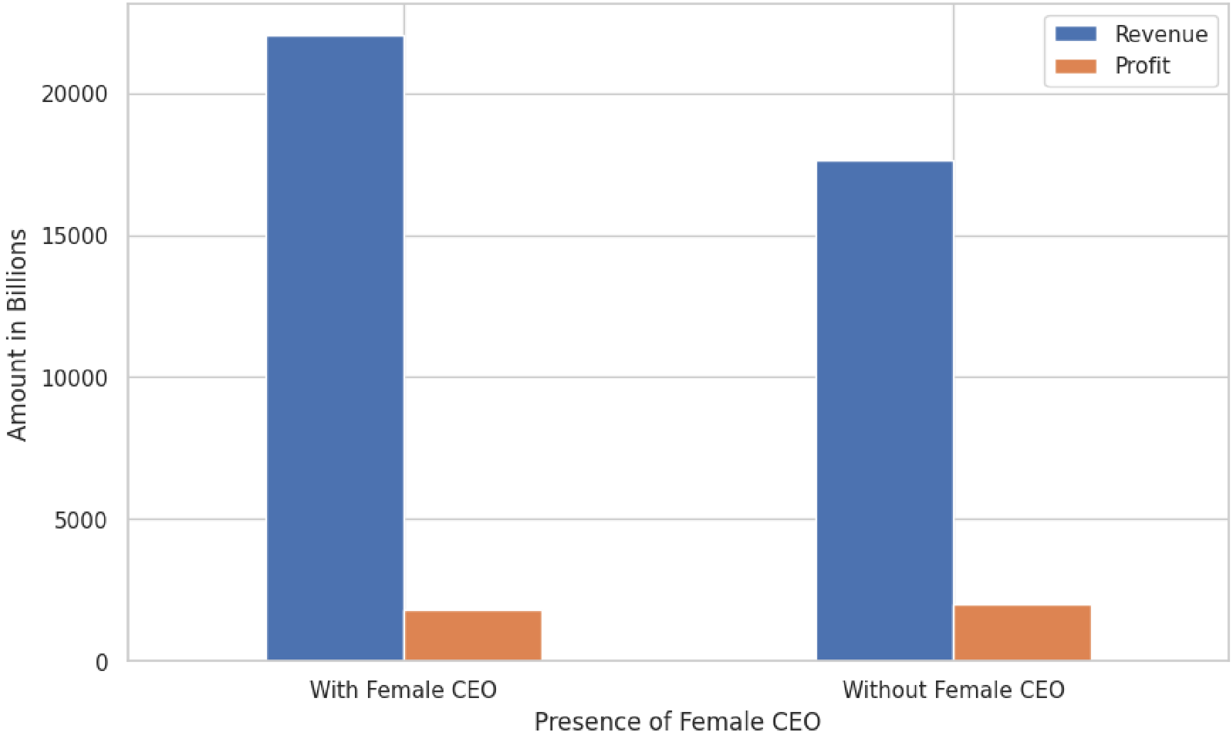
2. **How does the presence of a female CEO affect the revenue and profit of companies ? Are there any significant differences in performance between**

male-led and female-led companies, and if so, what factors may contribute to these differences?

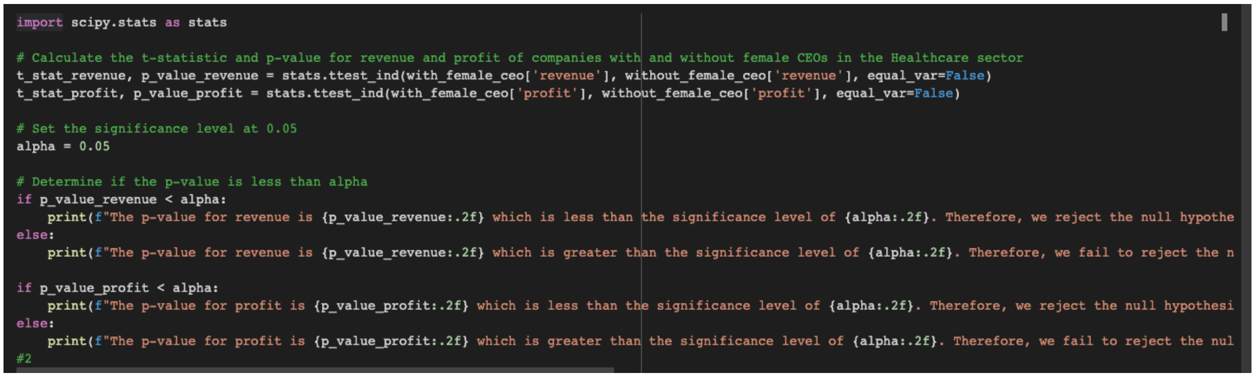
From the analysis of the dataset, we calculated the average revenue and profit for companies with and without female CEOs. We plot the bar chart which displays the average revenue and profit for companies with and without female CEOs. We also conducted a hypothesis test to determine if there was a significant difference in revenue and profit between companies with and without female CEOs .

The results of the hypothesis test showed that the p-value for revenue was 0.37, which is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis and conclude that there is no significant difference in revenue between companies with and without female CEOs. Similarly, the p-value for profit was 0.60, which is also greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis and conclude that there is no significant difference in profit between companies with and without female CEOs.

In conclusion, the presence of a female CEO does not seem to have a significant impact on the revenue and profit of companies. However, it is important to note that other factors may also influence the performance of companies, and further research may be required to gain a more comprehensive understanding of the relationship between gender diversity at the top and organizational outcomes.



Statistical Test:



```
import scipy.stats as stats

# Calculate the t-statistic and p-value for revenue and profit of companies with and without female CEOs in the Healthcare sector
t_stat_revenue, p_value_revenue = stats.ttest_ind(with_female_ceo['revenue'], without_female_ceo['revenue'], equal_var=False)
t_stat_profit, p_value_profit = stats.ttest_ind(with_female_ceo['profit'], without_female_ceo['profit'], equal_var=False)

# Set the significance level at 0.05
alpha = 0.05

# Determine if the p-value is less than alpha
if p_value_revenue < alpha:
    print(f"The p-value for revenue is {p_value_revenue:.2f} which is less than the significance level of {alpha:.2f}. Therefore, we reject the null hypothe
else:
    print(f"The p-value for revenue is {p_value_revenue:.2f} which is greater than the significance level of {alpha:.2f}. Therefore, we fail to reject the n

if p_value_profit < alpha:
    print(f"The p-value for profit is {p_value_profit:.2f} which is less than the significance level of {alpha:.2f}. Therefore, we reject the null hypothesi
else:
    print(f"The p-value for profit is {p_value_profit:.2f} which is greater than the significance level of {alpha:.2f}. Therefore, we fail to reject the nul
#2
```

Output:

The p-value for revenue is 0.37 which is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis and conclude that there is no significant difference in revenue between companies with and without female CEOs in the Healthcare sector.

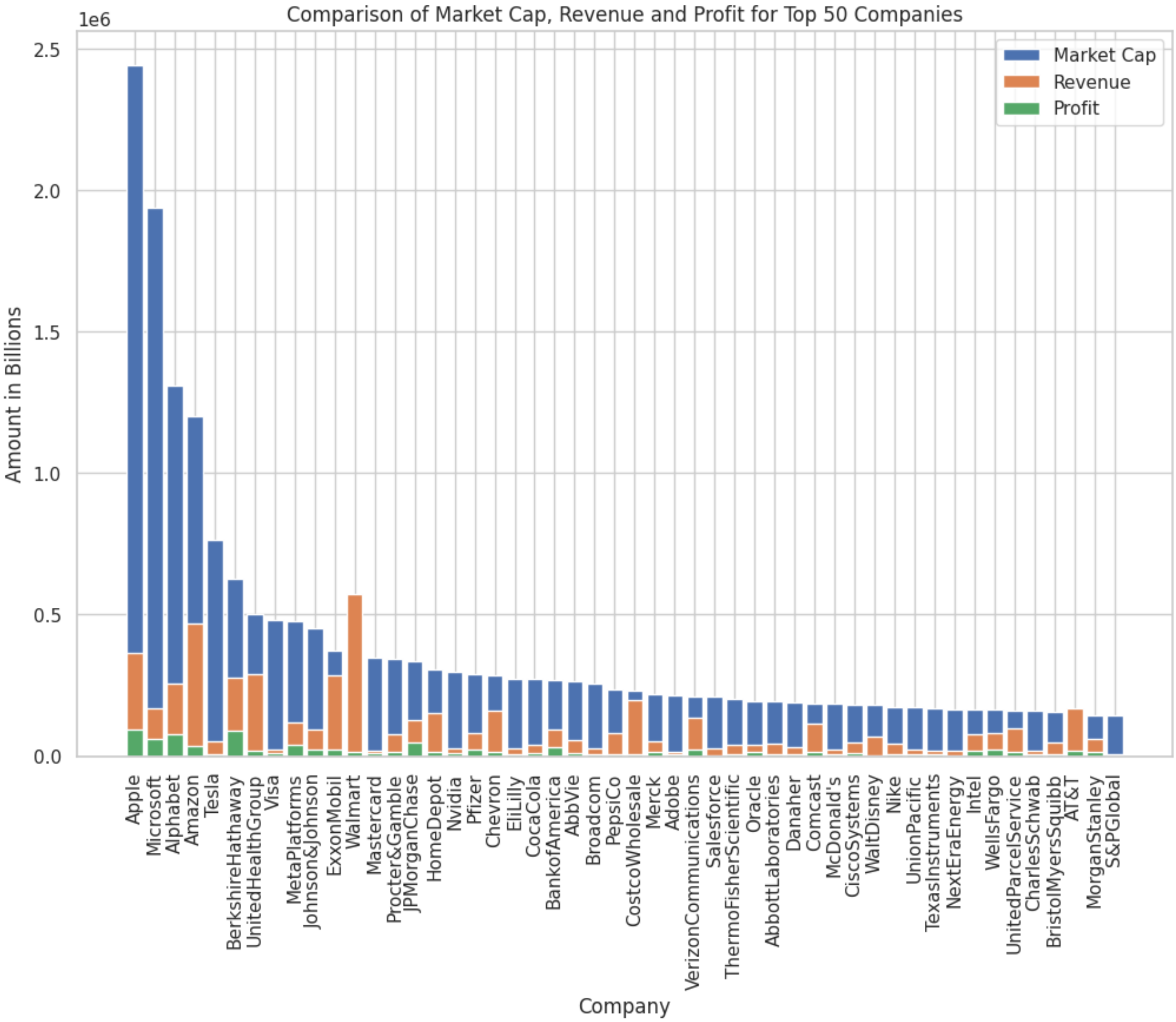
The p-value for profit is 0.60 which is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis and conclude that there is no significant difference in profit between companies with and without female CEOs in the Healthcare sector.

3. **How do the top 50 companies in the dataset compare in terms of market capitalization, revenue, and profit?**

To answer this research question, we used a stacked bar chart that compares the market capitalization, revenue, and profit of the top 50 companies in the dataset. We first extracted the top 50 companies based on their market capitalization using the nlargest() method in pandas. We then plotted a stacked bar chart using the matplotlib library, with each bar representing a company and the three stacked components representing its market capitalization, revenue, and profit.

The visualization shows that the market capitalization of the top 50 companies is generally much higher than their revenue and profit, indicating that investors value these companies based on their potential for future growth rather than their current financial performance. There is also a wide variation in the market capitalization, revenue, and profit of the top 50 companies, with some companies performing much better than others.

No statistical test was performed for this research question as it is a descriptive analysis of the data. Based on the visualization, we can conclude that the market capitalization of the top 50 companies is generally higher than their revenue and profit margins, indicating that investors are willing to pay a premium for these companies based on their future growth potential and market position. The stacked bar chart also allows us to compare the relative performance of individual companies within the top 50 based on their revenue and profit margins.



The graph also allows us to compare the relative performance of individual companies within the top 50 based on their revenue and profit margins. Overall, the graph provides a useful summary of the financial performance and valuation of the largest companies in the dataset.

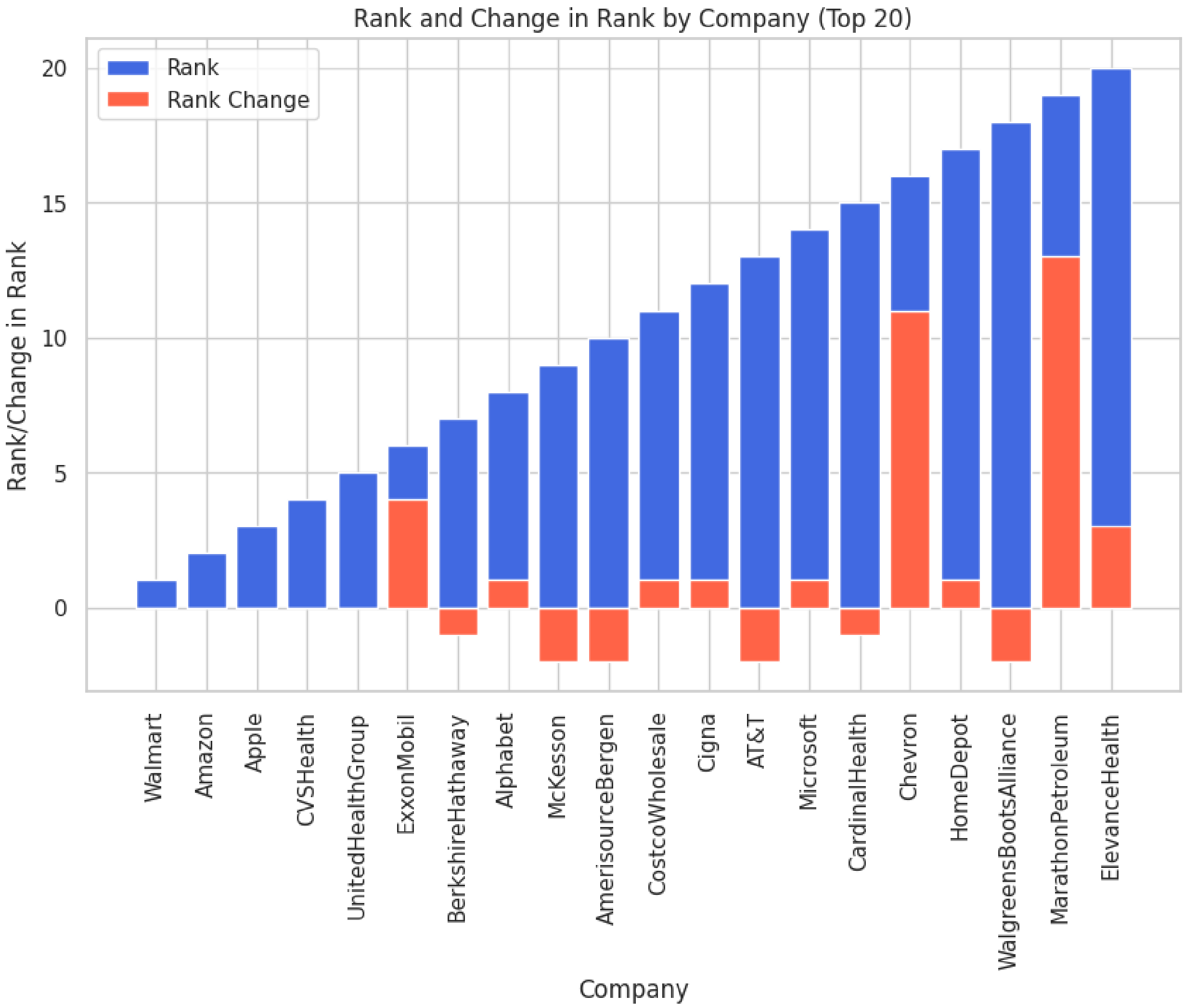
4. **How have the rankings of the top 20 companies changed over time, and what are the current rankings and changes in rankings of these companies?**

We can learn the answer from the data by first sorting the dataset by rank in ascending order and selecting the top 20 companies. Then, we can create a bar plot to visualize the rank and rank change for each of these companies.

The resulting graph shows us the current rankings and changes in rankings of the top 20 companies. We can use this information to analyze the market position of these companies and identify any trends or patterns in their performance over time. By examining the data, companies can determine their strengths and weaknesses, and make strategic decisions to improve their position in the market.

No statistical test was used for this research question, as it is primarily exploratory in nature. The graph provides valuable insights into the rankings and changes in rankings of the top 20 companies, which can inform strategic decision-making for businesses.

In conclusion, the graph created for this research question provides a clear visualization of the current rankings and changes in rankings of the top 20 companies, which can help businesses to analyze their market position and make informed decisions about their business strategy and operations.



The above graph shows the rank and rank change for the top 20 companies. This information can help businesses to understand their position in the market and their progress over time. By analyzing this data, companies can identify their strengths and weaknesses, and can make strategic decisions to improve their market position. For example, a company that has seen a significant drop in rank may need to reassess its business strategy or make changes to its operations to improve its performance. On the other hand, a company that has seen a positive rank change may use this data to highlight its success and make investments in growth areas. Overall, this graph provides valuable insights for companies to make informed decisions about their business strategy and operations.

Conclusion:

In conclusion, while this project provides valuable insights into the performance of top companies in the United States and the impact of certain factors on their success, further research is needed to validate the findings and ensure their accuracy. For instance, a larger sample size or more comprehensive dataset could provide a more robust analysis of gender diversity and its impact on company performance. Additionally, analyzing additional factors such as innovation, customer satisfaction, and employee engagement could provide a more holistic understanding of a company's success. Therefore, it is important to continue exploring and analyzing data to gain a more complete understanding of the factors that contribute to a company's performance.