Web & API Data Integration Report: U.S. Companies by Revenue

1. Introduction

The objective of this project was to combine **web-scraped data** and **API data** to create a unified dataset of major U.S. companies by revenue.
The analysis focuses on companies ranked by total annual revenue (Wikipedia source) and enriched with real-time financial metrics (Financial Modeling Prep API).
The final dataset integrates both business and financial perspectives, allowing comparison between company size, industry, and market performance.

2. Data Sources

2.1 Web Scraping (Wikipedia)

Data was scraped from the *Wikipedia page* titled **"List of largest companies in the United States by revenue."**

2.2 Financial Modeling Prep (FMP) API

To enhance the dataset, additional metrics were fetched via the **Financial Modeling Prep API** using Python's requests library.
Each company name was searched using /search-symbol and /profile endpoints.

3. Data Processing & Integration

3.1 Merging

Both datasets were merged using the normalized Symbol field:

```
df_final = df_merged.merge(df_api_selected, left_on='Symbol', right_on='symbol', how='left')
```

3.2 Cleaning & Normalization

Cleaning the merged data:

```python
# Remove unwanted column (replace 'column_to_remove' with actual column name)
df_final = df_final.drop(columns=['Symbol'])
```

```python
# Remove rows where 'symbol' is empty or null
df_final = df_final[df_final['symbol'].notna() & (df_final['symbol'] != '')]
```

```python
# Reset Rank from 1 to len(dp_final)
df_final = df_final.reset_index(drop=True)  # reset index first
df_final['Rank'] = range(1, len(df_final) + 1)
```

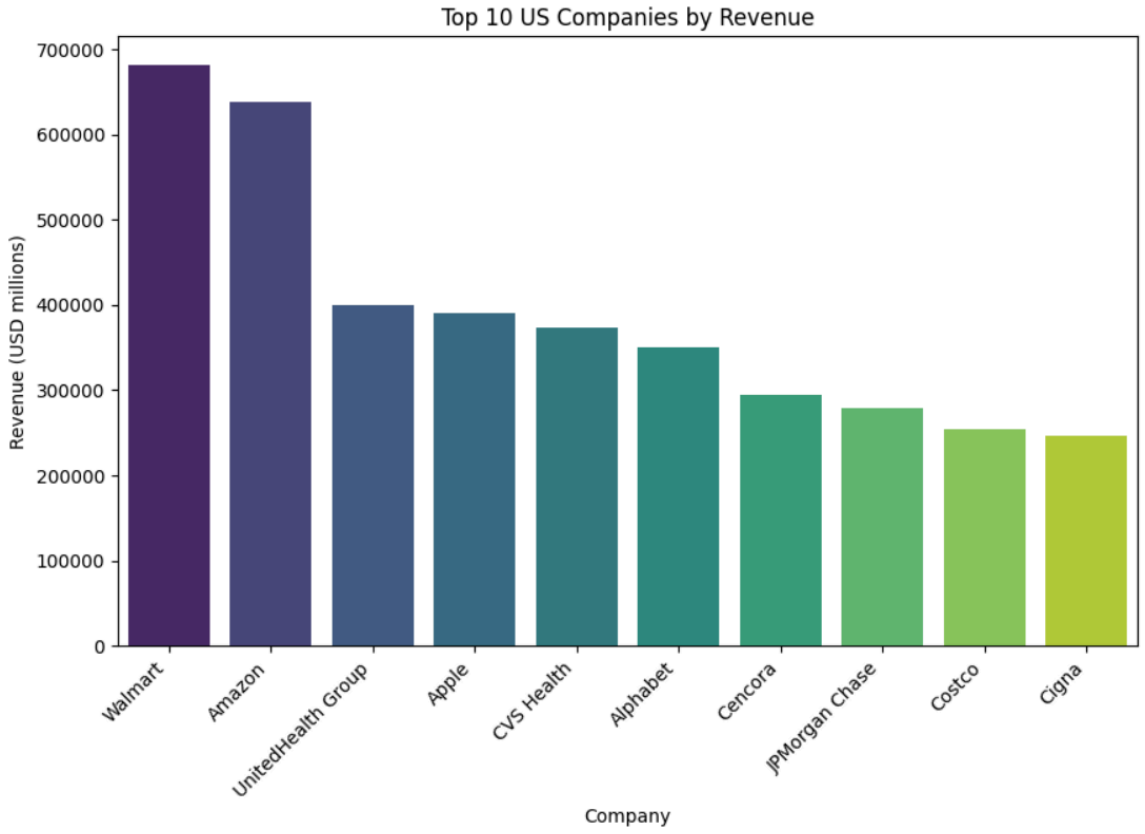
```python
df_final = df_merged.merge(df_api_selected, left_on='Symbol', right_on='symbol', how='left')
```

This created a unified DataFrame containing **Wikipedia metrics** (revenue, employees, headquarters) alongside **API-based financial data** (price, market cap, CEO, website).

---

4. Visualizations and Insights

4.1 Top 10 U.S. Companies by Revenue



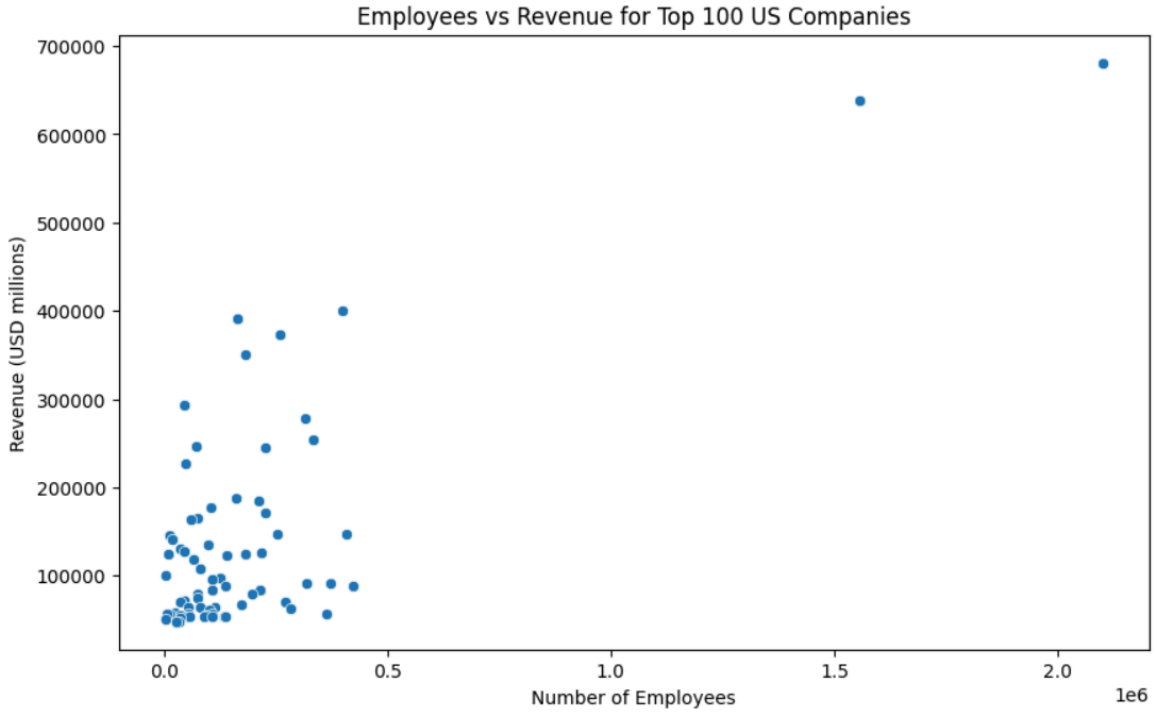
The bar chart visualizes the top 10 U.S. companies by total annual revenue.

- **Walmart** and **Amazon** dominate, earning nearly double the revenue of the next companies (UnitedHealth Group, Apple, CVS Health).

- While retail and cloud-based companies lead, healthcare and technology firms also hold strong positions.

Insight:
Retail remains the top revenue-generating sector, but its dominance is shared with tech and healthcare, reflecting the diversification of the U.S. economy.

4.2 Relationship Between Employees and Revenue



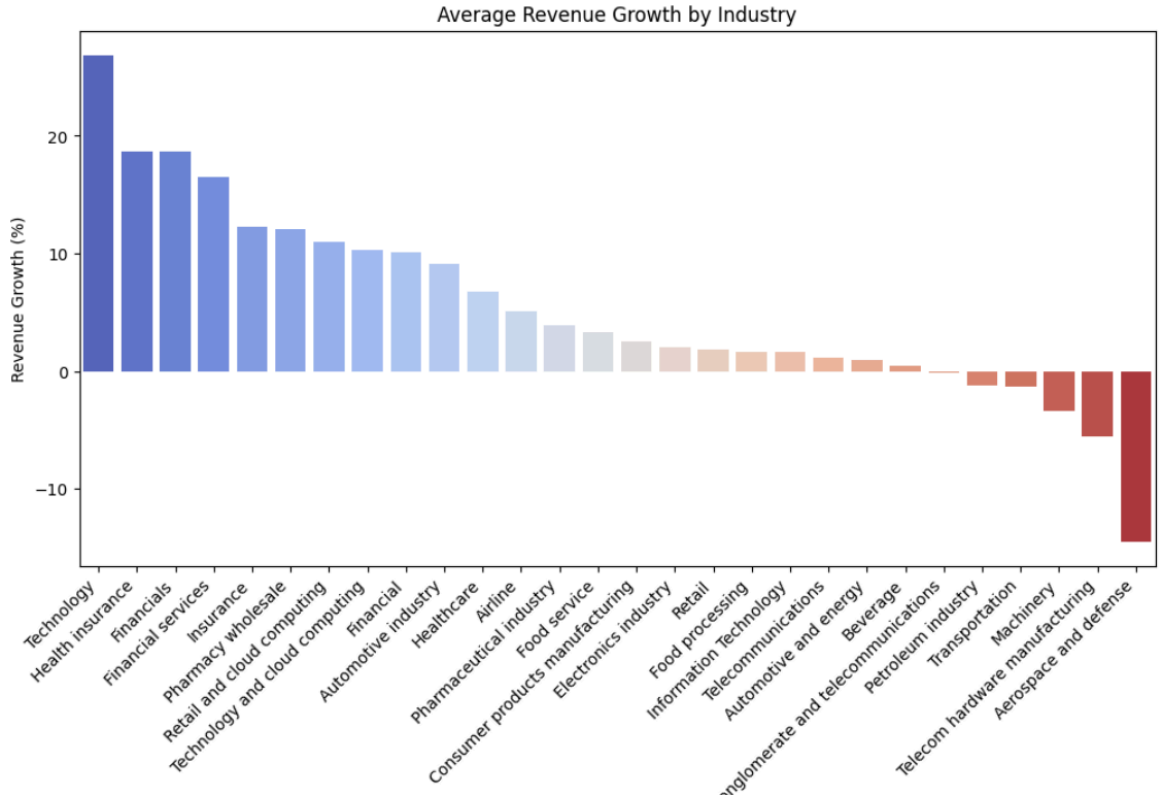
This scatter plot explores the correlation between company workforce size and total revenue.

- Companies like **Walmart** and **Amazon** appear as outliers, with both high employee counts and massive revenues.

- However, firms like **Apple** and **Microsoft** achieve high revenues with relatively fewer employees, indicating **higher productivity and automation**.

Insight:
There is no strict linear relationship between employee count and revenue. Technology companies demonstrate **higher efficiency per employee**, whereas retail and service sectors rely more heavily on workforce scale.

4.3 Average Revenue Growth by Industry



This visualization shows the average revenue growth across industries.

- **Technology**, **Health Insurance**, and **Financials** lead with the highest average growth rates (above 15–20%).

- Traditional sectors such as **Aerospace and Defense**, **Machinery**, and **Transportation** show negative or minimal growth.

Insight:
Industries tied to innovation and digitalization (e.g., tech, finance, and healthcare) continue expanding, while capital-intensive and manufacturing-based sectors are experiencing slower or even negative growth trends.

Conclusion

By merging publicly scraped Wikipedia data with API-based financial data, a richer analytical dataset was created.

This integration approach demonstrates the power of combining **static web content** with **dynamic API information**, providing both operational scale (revenue, employees) and market sentiment (price, market cap)