Retail Industry Analysis

Exploring Top Companies and Trends

Aswani Yaramala

Preethi Reddy Tera

Introduction

We conducted an analysis of prominent retail companies to evaluate their financial stability and provide guidance to managers and investors. The data used for this analysis was collected from Wikipedia and included information on sales, net income, employee count, and the number of stores. Our focus was on the top 100 retail companies globally, with a particular emphasis on the top 5 companies in the United States. Various analytical techniques such as scatter plots, bar plots, line plots, linear regression, and t-tests were employed to extract insights from the data. Additionally, we developed an equation that could help businesses estimate the number of employees to hire based on the number of stores.

GitHub

https://github.com/aswani848/6830project3

Presentation

https://docs.google.com/presentation/d/1ubfFjijN5VCcvGsvYhC-JFh4b3Et_D_cjUI4KAOvO2Q/edit#slide=id.p

Dataset

Our dataset was obtained from Wikipedia, specifically from tables that contained information about the largest retail companies. We narrowed down the list to include only the top retail companies in the United States. We retrieved information from Wikipedia for the top 5 retail companies in USA , namely Walmart, Amazon, Home Depot, Kroger, and Costco. The information we obtained included their revenue, net income, number of employees, and number of stores for different years. We have data for Walmart dating back to 1968, while for Amazon, we have data from 1995. As for the remaining companies - Home Depot, Kroger, and Costco - we only have data from 2006 onwards.

To extract our data from Wikipedia, we utilised two Python modules, namely Beautiful Soup and Requests. This GET request retrieves the HTML code for the entire web page, which includes the table we are interested in. We then use Beautiful Soup to parse the HTML and extract the table data that we want. We created a function that takes a URL and the class name of a specific HTML table as inputs. This function is designed to return a pandas DataFrame that contains the data from the specified HTML table. The class name parameter corresponds to the class attribute of the HTML <table> element.

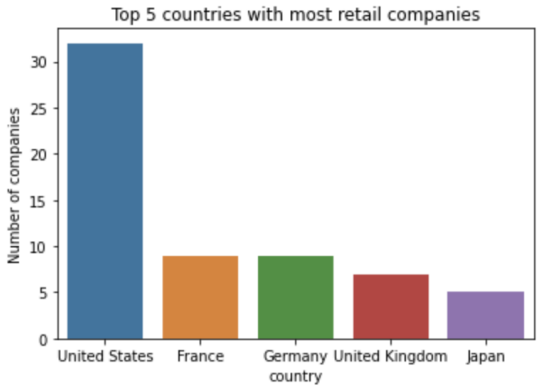
Analysis Technique

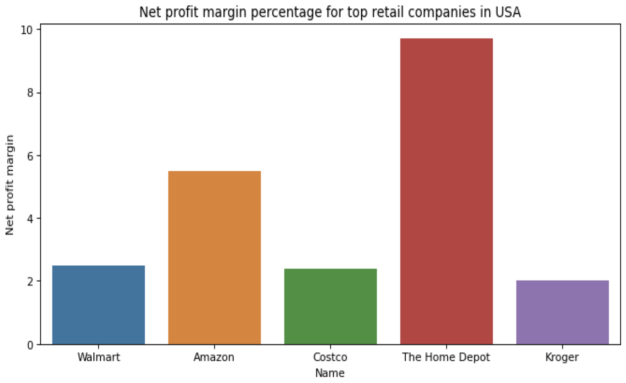
In our analysis, we utilised several types of plots such as bar charts, scatter plots, regression plots, and line charts. We used bar plots to show the top 5 countries with the most retail companies and the net profit margin percentage of the top 5 retail companies in the USA. Additionally, we used bar plots to display the change in the number of employees over time for both Walmart and Amazon. We used a scatter plot to examine the correlation between the number of stores and employees for Walmart, and a regression plot to determine whether there is any correlation between revenues and the number of

employees for Amazon.We performed a t-test to compare the percentage change in net income for Home Depot and Costco. Lastly, we used a line chart to display the revenue changes over time for the top 5 retail companies in the USA.

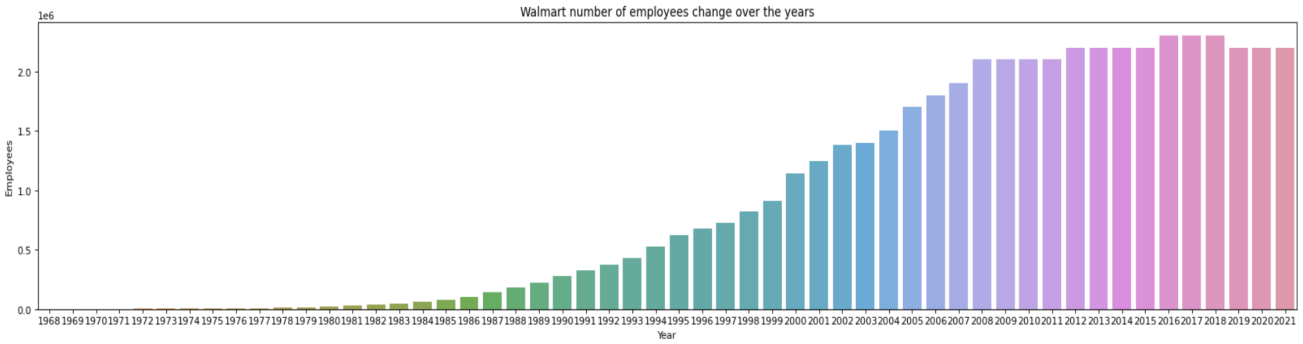
Results

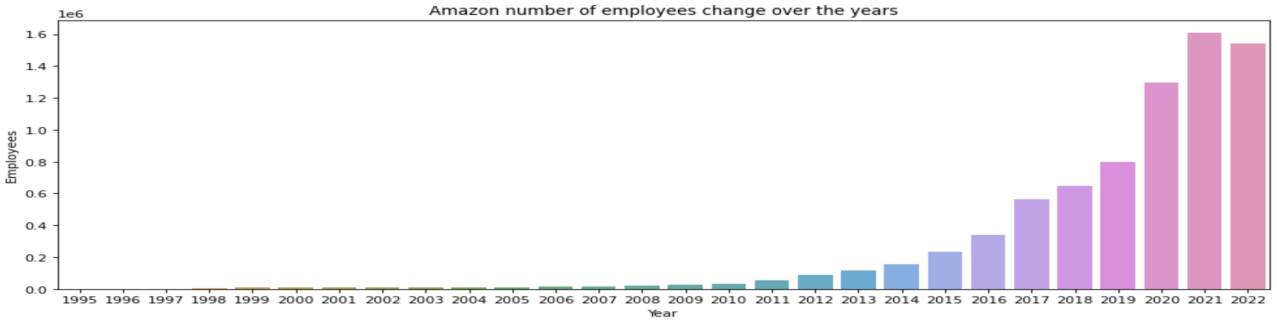
In our first analysis, we created a bar graph to display the top 5 countries with the most retail companies. The United States had the most retail companies with a total of 32, followed by France and Germany with 9 retail companies each. Given the dominance of the United States in this list, we then focused our analysis on the top 5 retail companies in the US. We plotted a bar graph to display the net profit margins for these companies, and found that Home Depot had the highest net profit margin among the top 5, despite Walmart being the top company in terms of revenue.





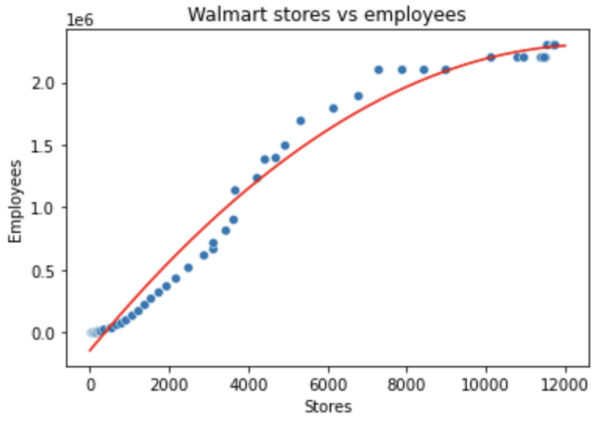
In our next analysis, we focused on Walmart and Amazon, and examined the trends in the number of employees for both companies over time. For Amazon, we observed a significant increase in the number of employees from 2019 to 2021, which could be due to the COVID-19 pandemic and increased online shopping. We found a positive correlation between revenue and the number of employees in Amazon, as evidenced by the low p-value of the Pearson correlation coefficient.

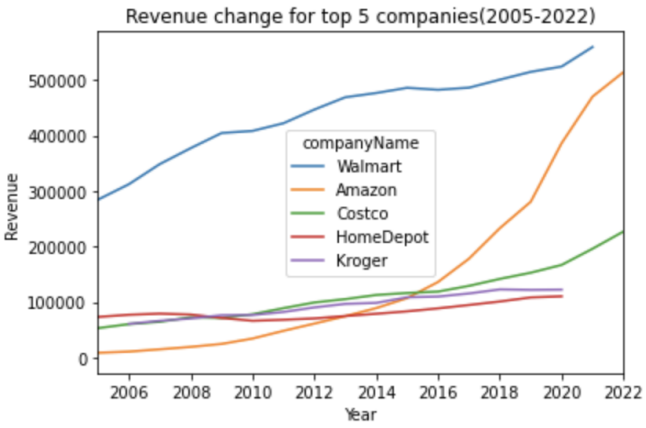


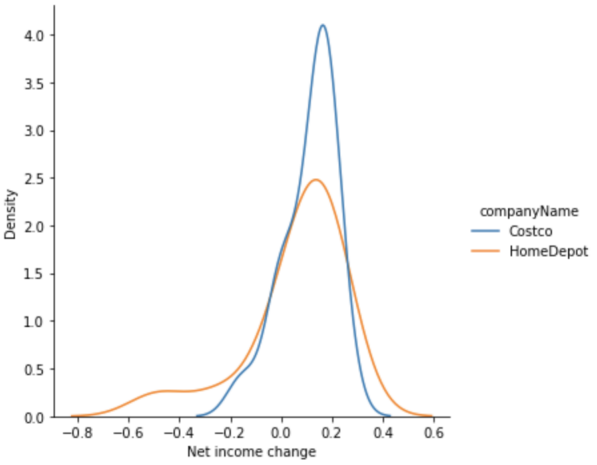


For Walmart, we observed a steady increase in the number of employees over time. We investigated the correlation between the number of stores and the number of employees in Walmart and found a positive correlation, which we validated with a Pearson r-test. We also fitted a curve to the data

points and obtained a polynomial equation, which we can use to estimate the number of employees Walmart needs to hire given the number of stores.
pred_employees=-0.015*num_stores2+383.71*num_stores-141545.86**







The revenue of the top 5 retail companies in the USA from 2005 to 2022 was compared in our analysis. Walmart had the highest revenue, while Amazon's revenue increased significantly from 2019 to 2021, possibly due to the impact of the Covid-19 pandemic. The revenues of the other three companies, Kroger, Costco, and Home Depot, remained relatively stable. In addition, we compared the percentage of net income change over the years for Home Depot and Costco, and found that Costco had a higher mean net income percentage change than Home Depot. We got the p value >0.05, which is not statistically significant.

Technical

To extract the data for analysis, we utilised the 'beautifulsoup' and 'requests' packages to extract the necessary data from Wikipedia, as detailed in the dataset section. The data preprocessing steps we followed included dropping null values using dropna(), renaming several column names, and formatting the data by converting string values to float and int values, replacing non-numeric characters such as %, and splitting some of the columns as needed. To create visualisations of our analysis, we imported the matplotlib and seaborn Python packages. To validate our results, we used the stats package from scipy to perform Pearson correlation coefficients and t-tests.