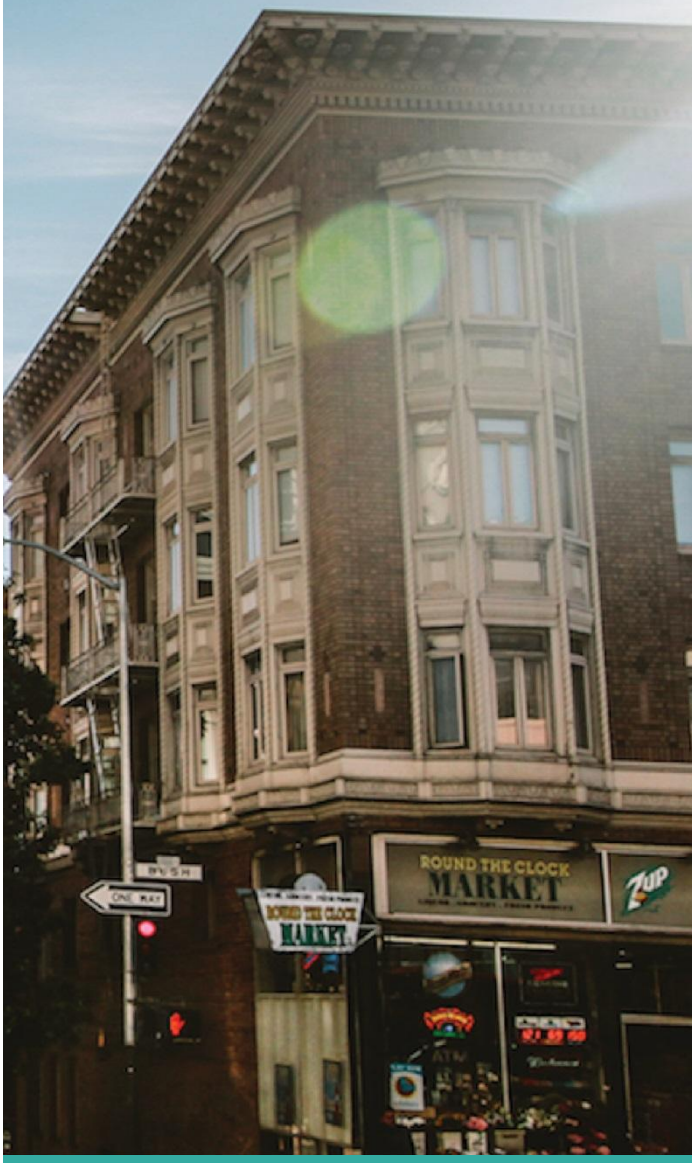
LAB 3: DESCRIPTIVE ANALYTICS



OCTOBER 22

BSAN6070 INTRODUCTION TO MACHINE LEARNING

Authored by: Andres Sebastian Gaibor Heredia

1. Inspecting data and missing values

First, we create a dataframe with the financial data from 39,080 companies during the period 1961-2024. The resulting dataframe has 503,055 records and 981 fields. To narrow down our analysis and make the results more relevant, we take a subset of the data considering only fiscal years 2000-2024. By filtering the data, we reduced the number of records to 235,996. We can also note that there are missing values, so we opted to drop the fields that have a missing value percentage equal to or higher than 50%. This step reduces the number of columns to 317. With the remaining numerical missing values, we decided to impute their value using the field's median. For the object type columns that have missing values, we decided to drop them as we checked that the information is not relevant to our analysis. The resulting dataframe has 235,996 records and 301 fields.

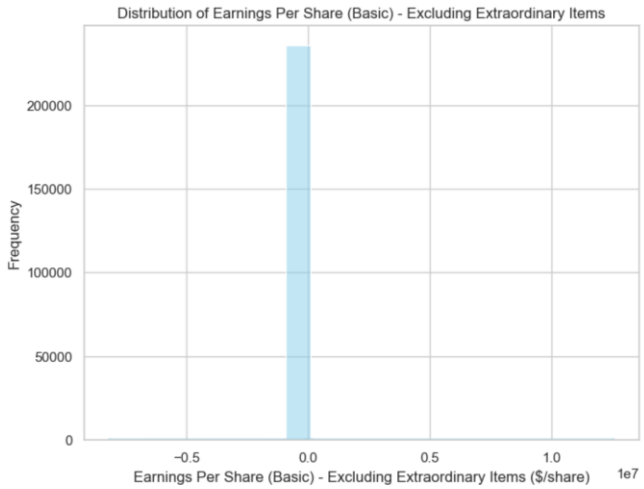
2. Calculating Summary Statistics

We then calculate the mean values for the numerical columns in the data and there is high contrast of the values depending on the field under consideration. By calculating the median of the same columns, we notice there is a significant difference compared to the mean values, which suggests that the data has outliers and is skewed. The mode is also different from both mean and median. We proceed to calculate the standard deviation and range for the numerical columns which gives us an idea of the dispersion of the data.

3. Data Visualization

3.1 Histogram of Earnings Per Share (Basic) - Excluding Extraordinary Items Winsorized

If we do the histogram of Earnings per share without any transformation, we get a plot with significant outliers on the x-axis as seen below. The data is separated in 20 buckets.



In order to address the outliers, we apply a winsorization to the variables we are going to analyze Net Income and Earnings Per Share (Basic). We do this by capping extreme values at the 1st and 99th percentiles.

The limits are:
Net Income (ni) Winsorization Limits: -659.83 to 4582.04
EPS (epspx) Winsorization Limits: -8.32 to 11.36

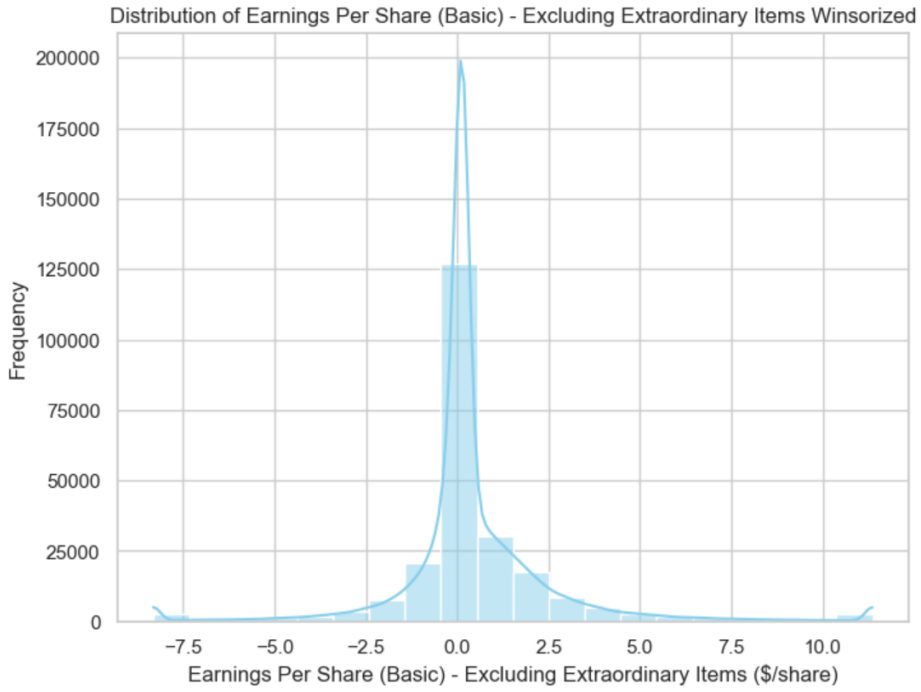
Winsorization keeps most of the data intact while removing extreme distortions from outliers. The resulting summary statistics from the data before and after winsorization are shown on the table:

Metric	Original ni	Winsorized ni_winsorized	Original epspx	Winsorized epspx_winsorized
Min	-99,289	-659.83	-8,182,362	-8.32
Max	104,821	4,582.04	12,580,000	11.36
Mean	175.77	136.31	-6.90	0.45
Std Dev	1,599.35	616.30	33,444.68	2.27

By applying winsorization we:

- Effectively remove extreme outliers while keeping most data points intact.
- Reduce extreme variations in ni and epspx, making the dataset more reliable for financial analysis.
- Prepare data for better visualization and modeling, avoiding the skewing effects of outliers.

The histogram below represents the distribution of Earnings Per Share (Basic) - Excluding Extraordinary Items after applying Winsorization to cap extreme values at the 1st and 99th percentiles.



Key Observations:

Highly Skewed Distribution

The histogram is sharply peaked at 0, meaning that most companies report EPS close to zero. This is expected in financial data, as many firms barely break even or operate near zero profit over the long term.

Extreme Values Have Been Trimmed

Compared to a raw EPS distribution, the winsorized data has removed extreme outliers, ensuring that the bulk of the data is more interpretable.

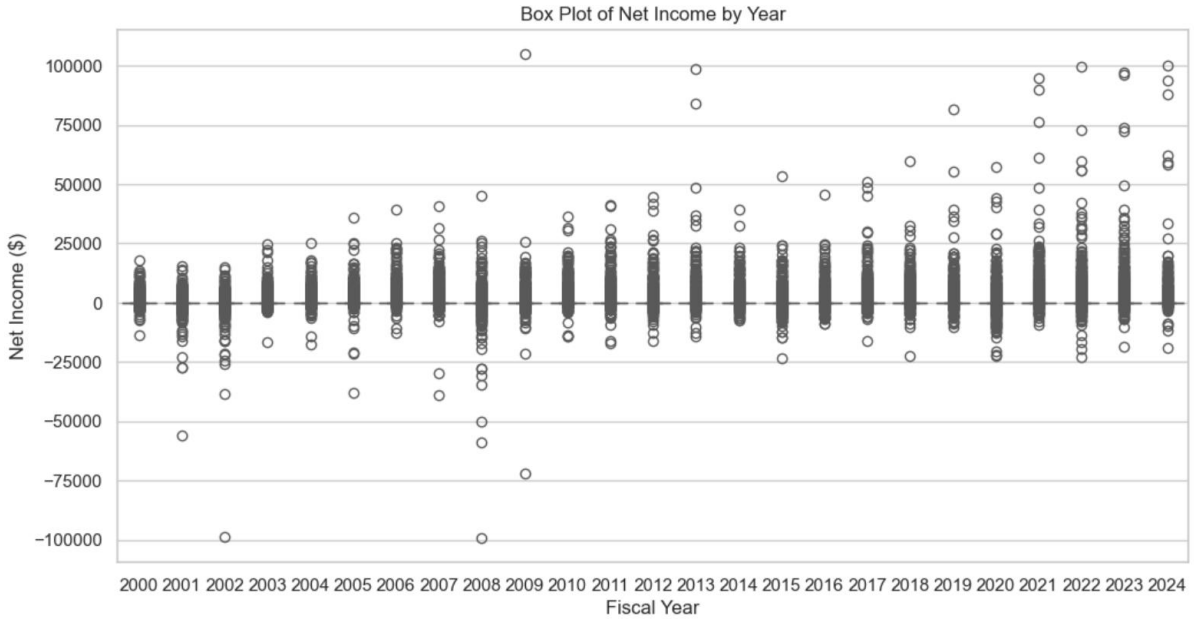
The range is now approximately between -8 and 11, which aligns with the Winsorization thresholds applied earlier (-8.32 to 11.36).

Long Tails Are Still Visible

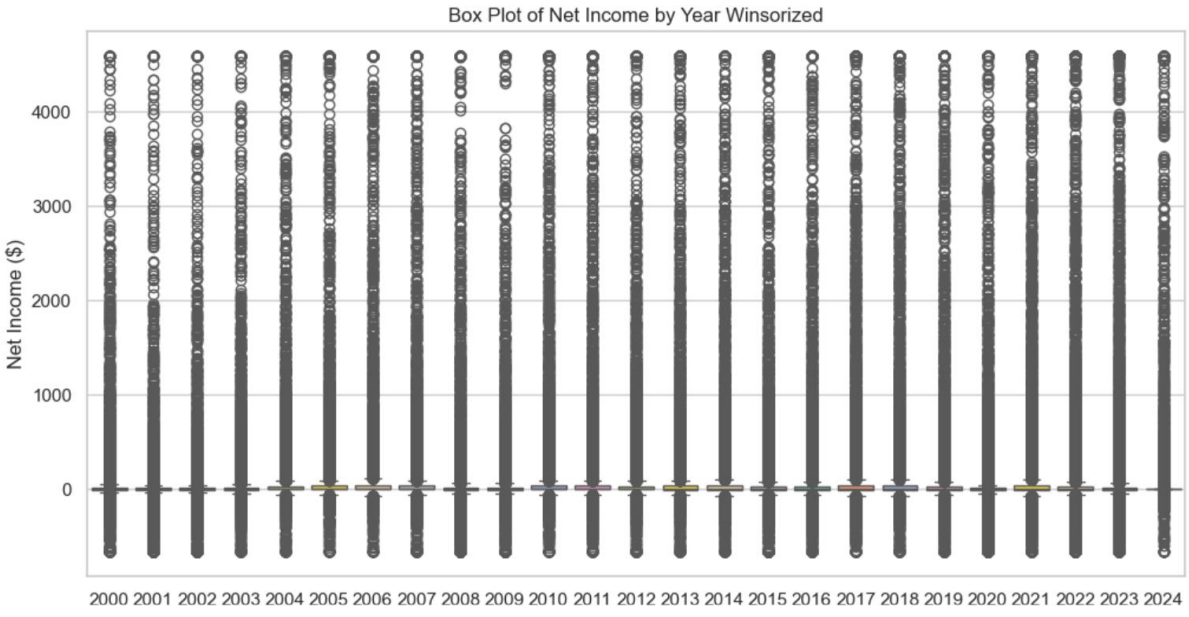
Despite winsorization, the distribution retains a slight long-tail effect, meaning some companies still have relatively high or low EPS, but extreme distortions have been mitigated.

3.2 Analysis of the Winsorized Box Plot for Net Income by Year

The box plot visualizes Net Income across different years (2000-2024), displaying its distribution, median, and presence of outliers.



Given that we previously winsorized the data at the 1st and 99th percentiles, the resulting plot using winsorization we can expect extreme values to be limited, but key distribution patterns to remain. The plot is shown below.



Key Observations

High Concentration Around Zero Net Income

The interquartile range box (IQR) is very small for all the years, indicating that a large portion of firms have net income near zero. This suggests that many firms are barely profitable or operating at break-even levels.

Significant Outliers Despite Winsorization

Even after Winsorization, numerous outliers remain above the upper whisker. This suggests that while extreme outliers (1% most extreme values) were removed, a wide distribution of profitability still exists.

Stable Median Net Income Across Years

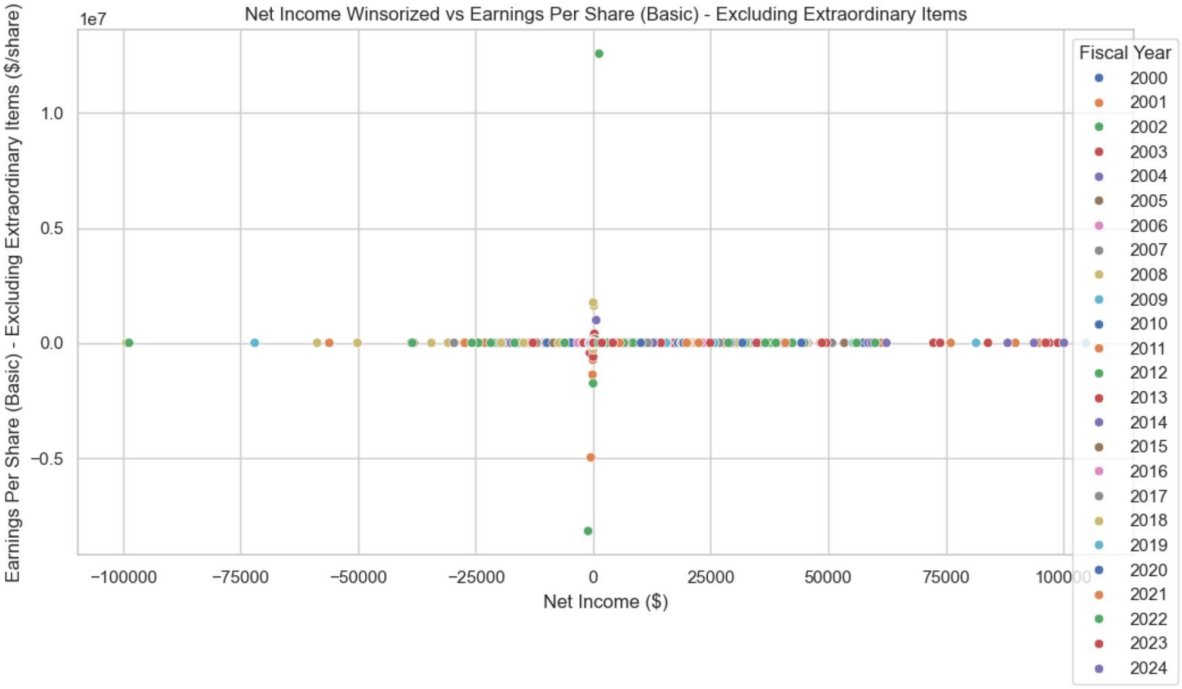
The median (central line in each box) appears relatively stable across years, implying that the central tendency of net income hasn't drastically changed over time.

High Dispersion of Net Income

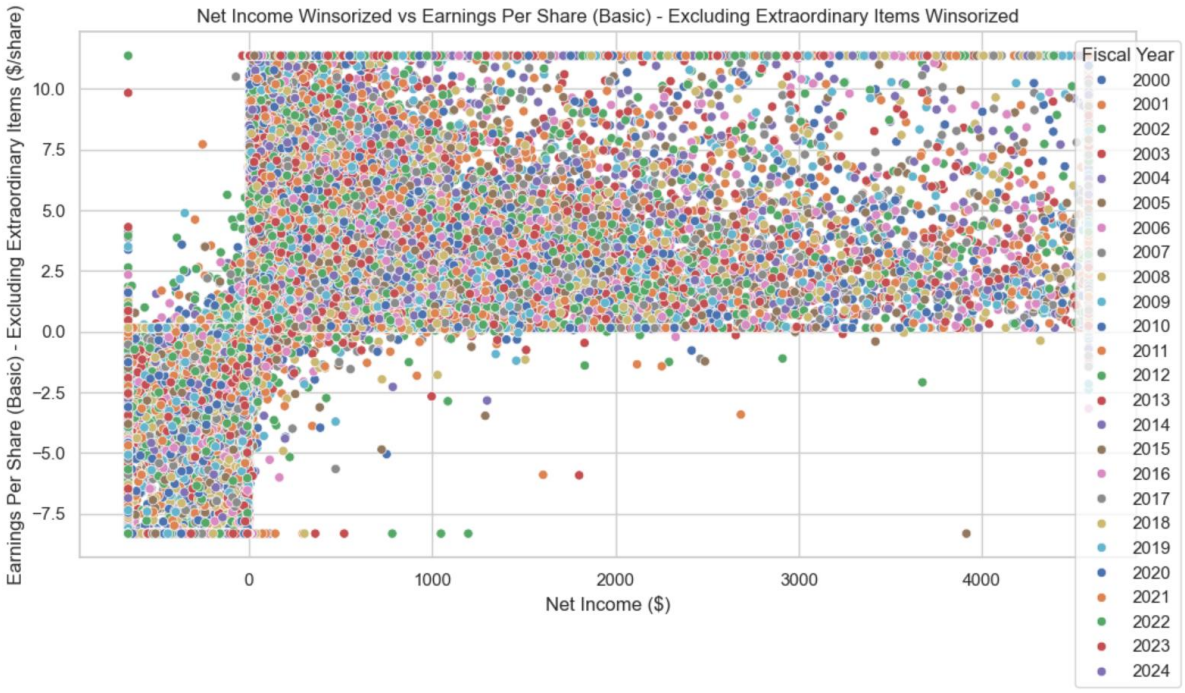
The whiskers extend quite far, reinforcing that companies each year have a broad range of profitability levels.

3.3 Analysis of the Winsorized Scatter Plot: Net Income vs. Earnings Per Share (EPS)

First, we plot the scatterplot without any previous winsorization, and the results are very hard to interpret. We can note the presence of outliers in both variables and any pattern is not easily detected.



After applying the winsorization on both variables the relationship between the variables is clearer. The following scatter plot visualizes the relationship between Net Income (Winsorized) and Earnings Per Share (EPS) - Excluding Extraordinary Items, color-coded by Fiscal Year (fyear). Extreme values have been capped at the 1st and 99th percentiles.



Key Observations:

Clear Positive Correlation Between Net Income and EPS

The general upward trend indicates that higher net income is associated with higher EPS. This aligns with fundamental financial principles: EPS is derived from net income, so higher earnings logically lead to higher EPS.

Winsorization Has Trimmed Extreme Values

The x-axis (Net Income) is capped at around $4,582, and the y-axis (EPS) is capped at around $11.36 (as expected from the previous Winsorization). Despite Winsorization, dispersion remains high, suggesting firms still exhibit high variability in profitability.

Large Cluster of Firms with Low Net Income and EPS

There is a dense region around Net Income = $0 and EPS = 0. Many firms operate at break-even levels, showing low profitability. Many companies barely make a profit or reinvest heavily, leading to near-zero net income. Some firms may have temporary losses or accounting adjustments.

Two Clusters in the Data

First Cluster: Net Income < 500 & EPS between -2.5 and 2.5

Most firms belong here, showing small profits or minor losses. Likely mature, stable companies or firms with fluctuating earnings.

Second Cluster: Net Income > 1000 & EPS above 5

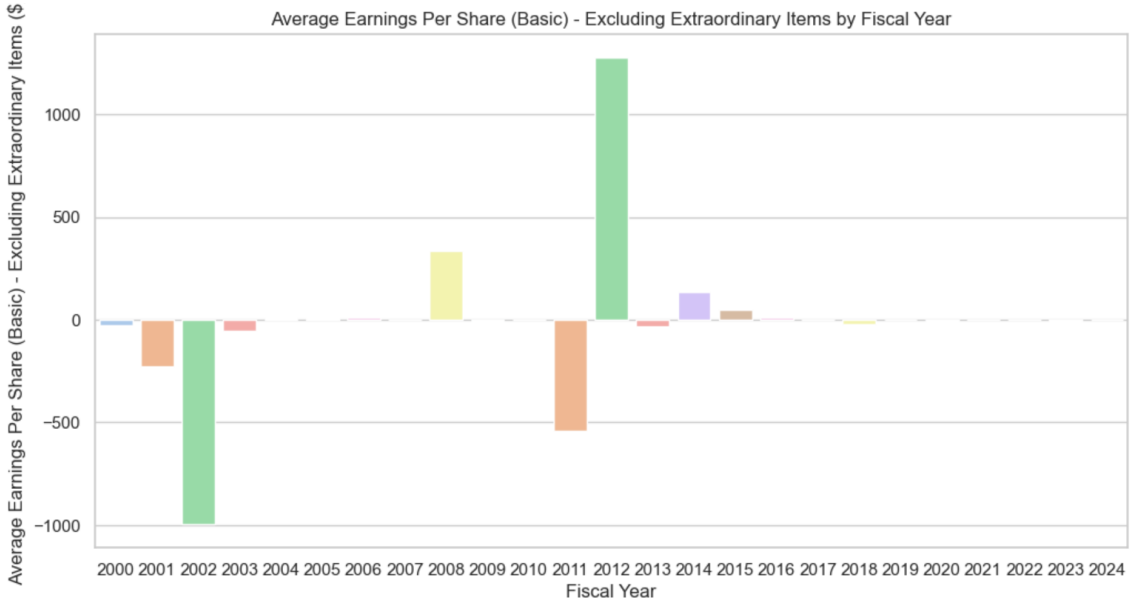
Represents firms with significant profitability and high shareholder value. Likely younger firms here, as maintaining consistently high earnings is difficult.

Key Takeaways

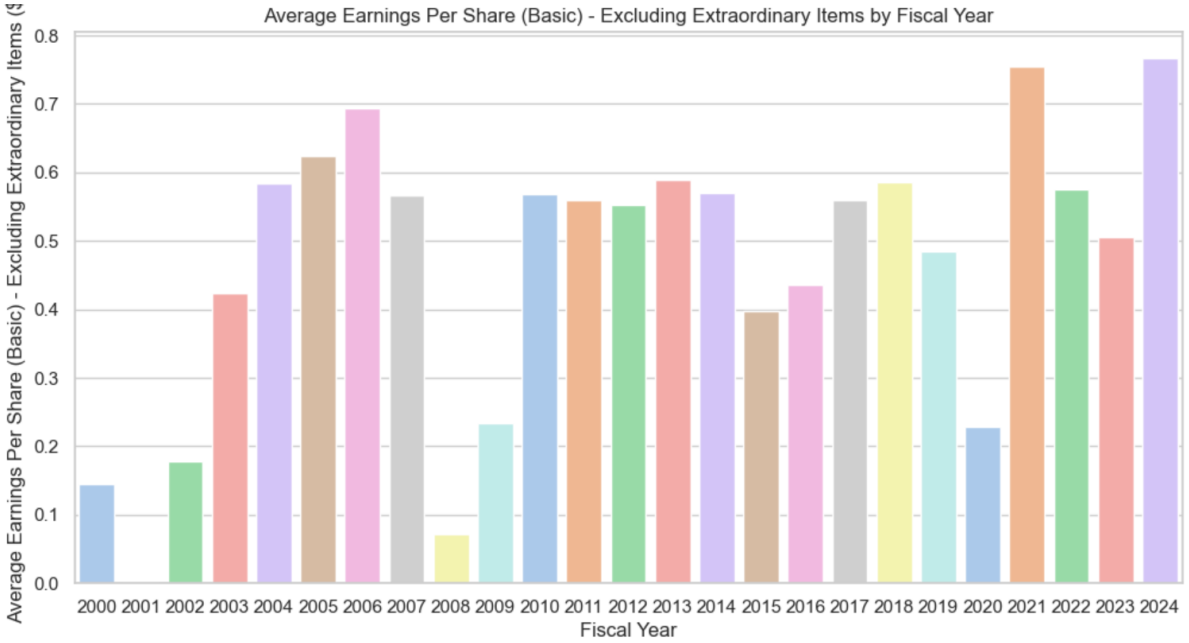
- A strong positive correlation is evident, reaffirming that higher earnings lead to higher EPS.
- Most companies operate near zero profit, as seen in the dense cluster around (0, 0).
- Companies with very high profitability are fewer but well-distributed, reflecting real-world financial patterns.

3.4 Analysis of the Winsorized Bar Chart: Average Earnings Per Share (EPS) by Fiscal Year

If we first run the plot without previously applying winsorization to EPS, we can note that outliers affect heavily the average for the years. This is expected, as the average is a measure that is highly influenced by extreme values. The bar chart is shown below.



After applying winsorization to EPS the results drastically change. The following bar chart displays the average Earnings Per Share (Basic) - Excluding Extraordinary Items, Winsorized, across different Fiscal Years (2000-2024). Given that winsorization was applied at the 1st and 99th percentiles, extreme values have been capped, making the data more reliable for trend analysis.



Key Observations

Winsorization Has Smoothed Out Extreme Variations

The EPS values are within a reasonable range (~0 to 0.8) due to the Winsorization process.

Without Winsorization, years have shown extreme spikes or deep drops, distorting the overall trend.

Clear Yearly Variability in EPS

Some years, such as 2004, 2011, 2021, and 2024, exhibit higher average EPS, indicating stronger overall profitability. Other years, such as 2000, 2008, and 2020, show lower average EPS, potentially reflecting the dotcom bubble, the Global Financial Crisis and the COVID pandemic economic slowdowns.

Post-2010 Shows a General Recovery

After 2010, the EPS values remain positive and relatively stable, with moderate fluctuations. This suggests economic stabilization and corporate recovery from the 2008 financial crisis. The high EPS in 2021 & 2024 may indicate strong post-pandemic recovery or shifts in corporate profitability.

Key Takeaways

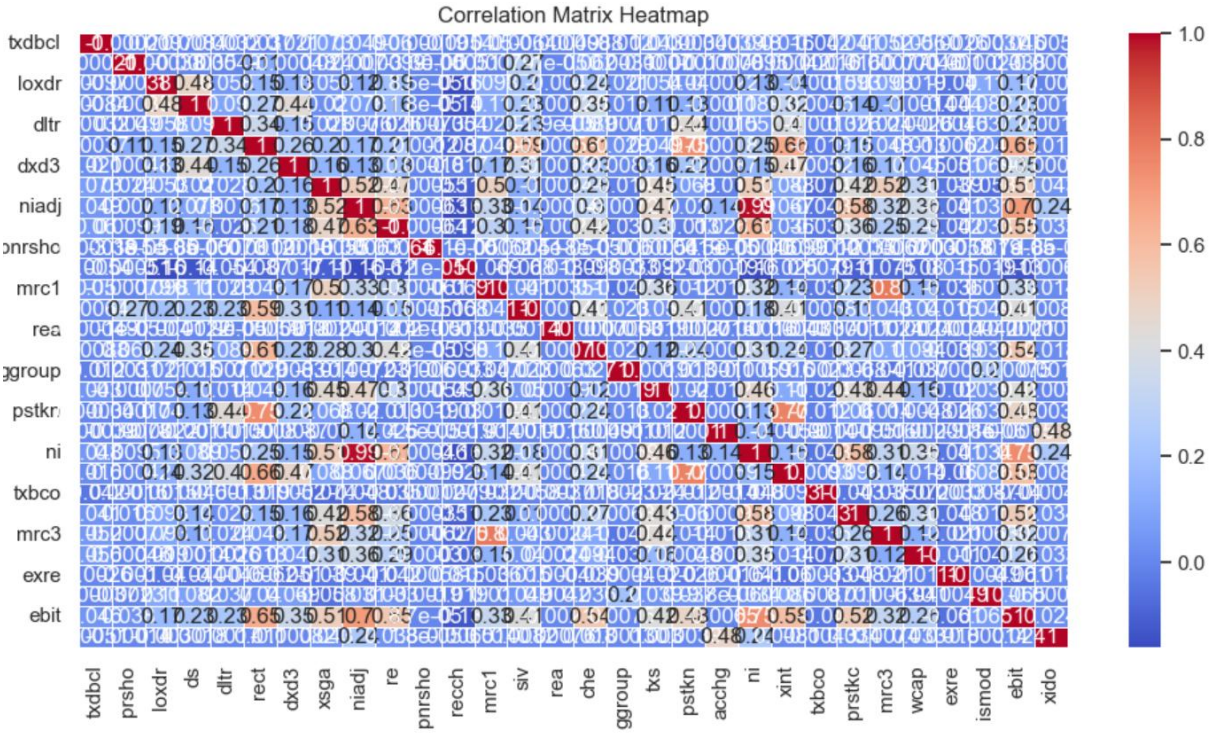
- Economic downturns (2008, 2020) correspond with lower average EPS, highlighting external financial shocks.
- Certain years (2004, 2011, 2021, 2024) exhibit higher EPS, possibly due to strong corporate earnings or favorable market conditions.
- Post-2010 data appears more stable, suggesting a maturing financial environment.

3.5 Heatmap of Correlation Matrix

This correlation matrix heatmap visualizes the pairwise correlation coefficients between 30 randomly selected financial variables. Considering the number of fields in the dataset (301) we had to sample a smaller group. The color gradient represents the strength and direction of correlations:

Red (close to +1) → Strong positive correlation.

Blue (close to -1) → Weak or no correlation.



Key Observations

Strong Positive Correlations

Several variables have strong positive relationships, meaning they tend to move together. Some of the strongest positive correlations include:

- ni (Net Income) and ebit (Earnings Before Interest & Taxes): Very strong correlation (~0.90+), which is expected since EBIT is a precursor to net income.
- mrc1 and mrc3: Very strong correlation (~0.80), these variables represent Rental Commitments – Minimum for 1-year and 3-year so it would be expected for them to be correlated.
- pstkr (Preferred Stock Redeemable) and inxt (Interest and Related Expense Total): Significant correlation (~0.70). This is an interesting relation; it suggests that firms with higher preferred stock redeemable also report higher tax-related figures.
- rect (Receivables Total) and ebit (Earnings Before Interest and Tax): Have a significant correlation of 0.65, suggesting that firms with high receivables report higher amounts of EBIT.

If multiple variables highly correlate, PCA can help reduce redundancy in financial modeling. We could also choose to drop one of those variables for modeling.

4. Handling Categorical Data

Bar chart of Active/Inactive Status Marker

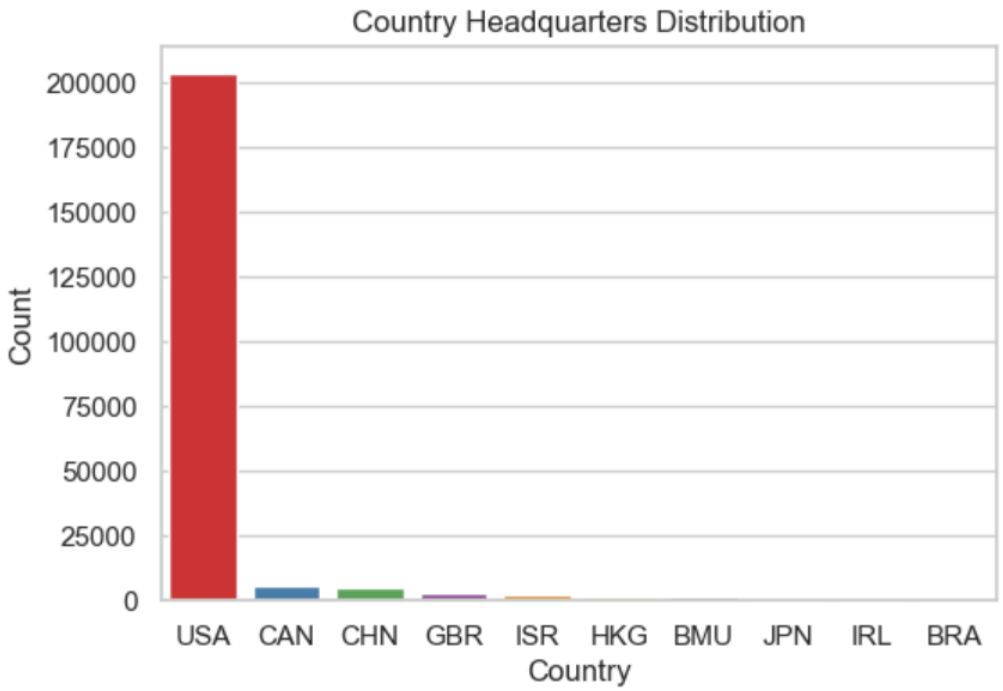
We can see from the plot that the number of active companies (141,254) is higher than the number of inactive companies (94,742). Nonetheless, there is a significant number of inactive companies present in the dataset.

Bar chart of Country Headquarters Distribution

This bar chart represents the distribution of company headquarters across different countries, with the number of companies (Count) on the y-axis and countries (ISO codes) on the x-axis.
The USA has an overwhelming majority of company headquarters, accounting for nearly all companies in the dataset (203,858). Countries like Canada (5,758), China (4,974), and the UK (2,656) have a small fraction of headquarters compared to the US. Even major global economies like Japan (JPN) and Germany (DEU) are not prominent, indicating a potential US-centric dataset bias.

Key Takeaways

- The USA dominates company headquarters, likely due to a US-centric dataset (compustat).
- Non-US companies are significantly underrepresented, suggesting the dataset may not be globally comprehensive.
- Presence of financial hubs like Hong Kong, Bermuda, and Ireland indicates some global diversification in corporate registrations.


Country Headquarters Distribution
200000
175000
150000
125000
100000
75000
50000
25000
0
Count
USA
CAN
CHN
GBR
ISR
HKG
BMU
JPN
IRL
BRA
Country