esade

Python Final Presentation Group 3

Do Good. Do Better.

Preprocessing Steps – Cleaning and Preparing the Data

Exercise 1 – Dataset Overview

Exercise 2 – Close Price Trends for 2023

Exercise 3 – Highest Close Prices

Exercise 4 – Comparing monthly average close prices for selected companies

Exercise 5 – Computing and comparing yearly average close prices across all companies

Exercise 6 – Visualizing the range of prices for each month for each company

Exercise 7 – Volume – Close Price Relation

Exercise 8 – Top Volume Month

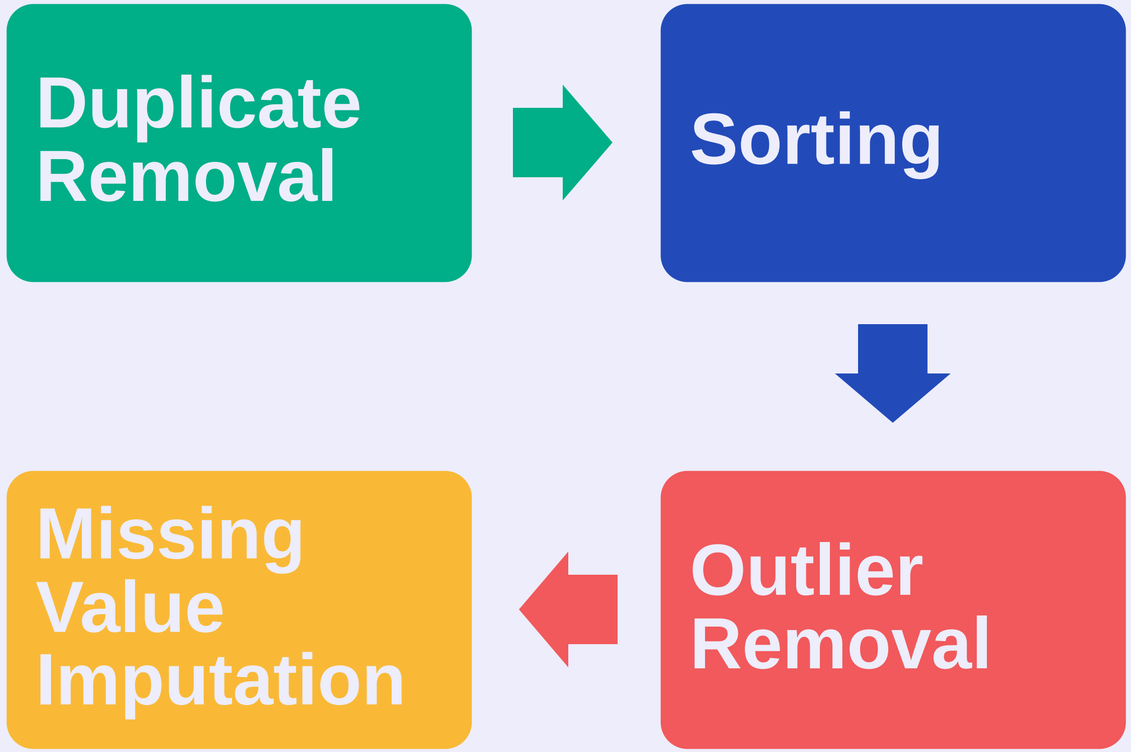
Exercise 9 – Yearly Aggregated Combined Dataset with no Missing Values

Exercise 10 – For Each Company: Daily Spread + Overall Mean Spread

Preprocessing Steps – Cleaning and Preparing the Data

Steps taken:

- Removed duplicate rows based on the **Date** column to avoid redundancy.
- Sorted the datasets chronologically by the **Date** column for time-series analysis.
- Used the Interquartile Range (IQR) method to identify and replace outliers in the **Close** column with NaN.
- Filled missing values in the **Close** column using linear interpolation, followed by forward and backward fill methods.



These steps ensured the datasets were clean, structured, and ready for reliable analysis.
Addressing outliers and missing values focused on the Close column, which is central to the exercises.

Exercise 1 – Dataset Overview

Steps taken:

- Loaded each dataset and checked for consistency in structure.
- Counted the number of rows and columns for each dataset.
- Extracted the column names and their data types.

Datasets	Rows
BRK-A	11222
DNUT	810
DPZ	5083
LKNCY	1345
MCD	14652
PZZA	7878
QSR	2459
SBUX	8117
WEN	11187
YUM	6796

Column	Type
Date	Datetime64
Open	Float64
High	Float64
Low	Float64
Close	Float64
Adj Close	Float64
Volume	Float64

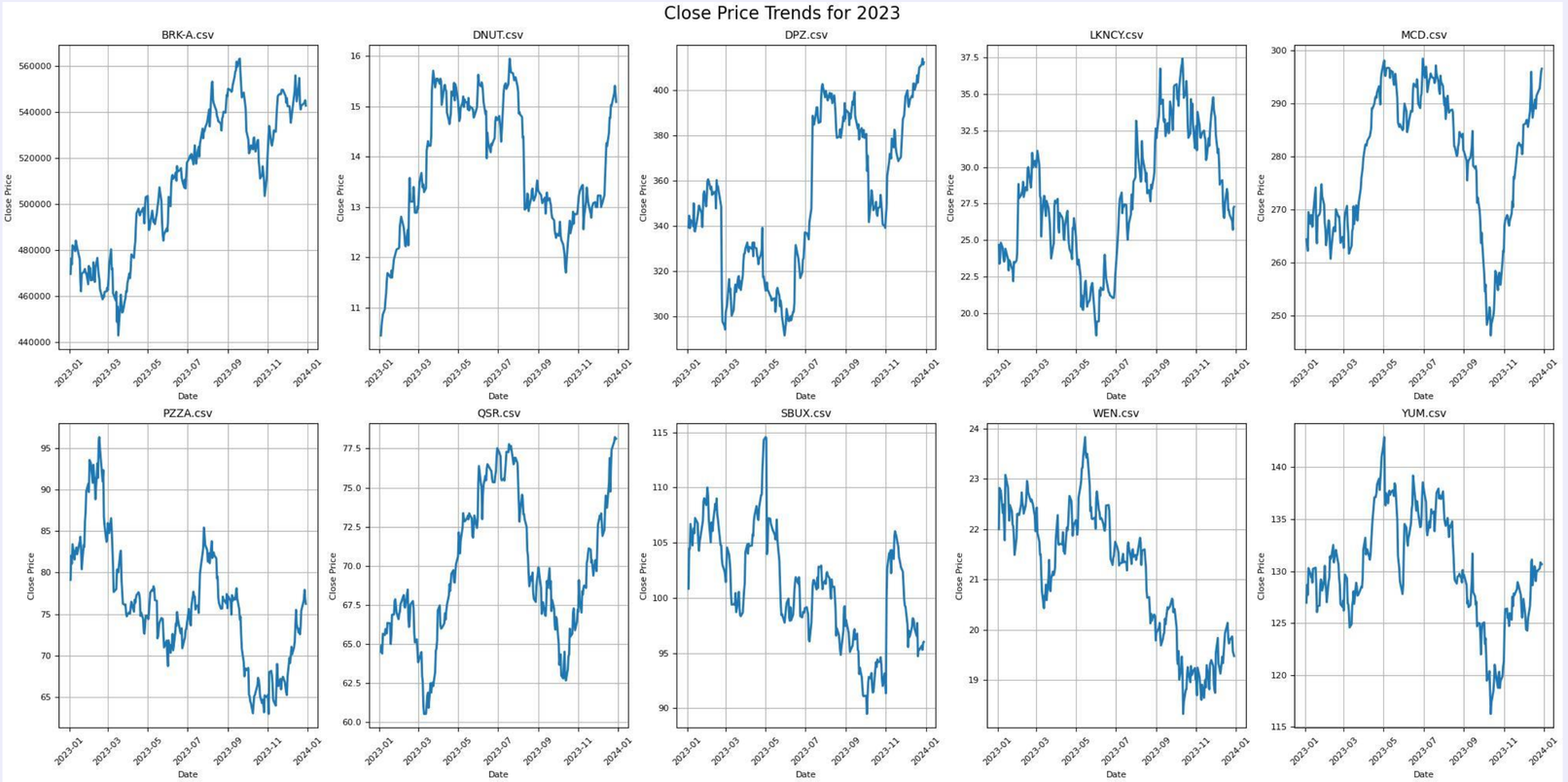
The datasets are consistent, with 7 columns in each file.
The Close column is critical for analysis, and the Date column ensures chronological alignment.

Exercise 2 – Close Price Trends for 2023

Steps taken:

- Filtered the data to include only rows where the year is 2023.
- Plotted the **Close** price trends for each company as line plots.

Line plots best show trends over time and allow easy comparisons across companies.



Several companies, like BRK-A and DPZ, experienced a dip around March 2023, possibly due to market-wide factors.
Many recovered by October 2023, reflecting improved conditions. Fast food-related stocks (e.g., MCD) showed greater stability than others.

Exercise 3 – Highest Close Prices

Steps taken:

- Iterated through each dataset to locate the row with the highest Close price.
- Recorded the highest price and its corresponding date for each company.

Datasets	Highest Close Price	Date
BRK-A	715910.0	2024-09-03
DNUT	21.0	2021-07-01
DPZ	564.33	2021-12-31
LKNCY	50.02	2020-01-17
MCD	300.53	2024-01-19
PZZA	140.01	2021-11-04
QSR	82.75	2024-03-13
SBUX	126.06	2021-07-26
WEN	32.0	1993-09-13
YUM	143.19	2024-04-29

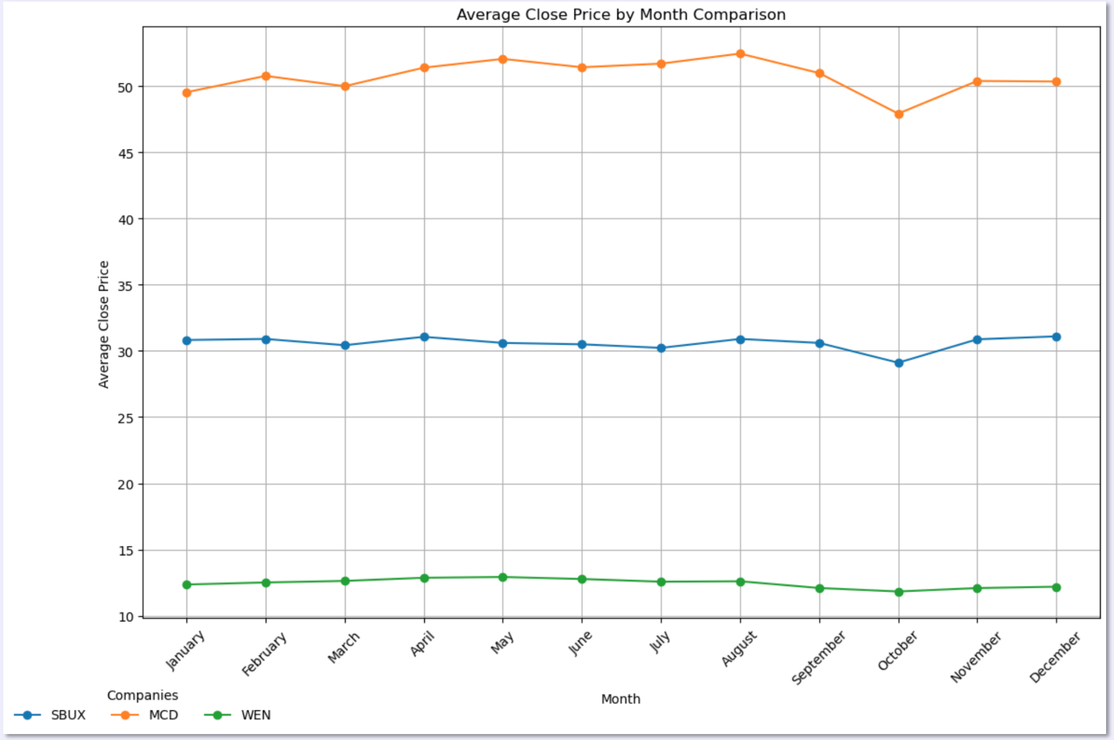
The datasets are consistent, with 7 columns in each file.
The Close column is critical for analysis, and the Date column ensures chronological alignment.

Exercise 4: Comparing monthly average close prices for selected companies grouping the data by month

Steps taken:

- Grouped the data by month for each company
- Extracted the month from the Date column
- Calculated the average Close price for each month across all years
- Stored the monthly averages for each company in a dictionary
- Created a line plot
- Formatted the plot

Line plot: to compare the monthly average Close prices for the companies



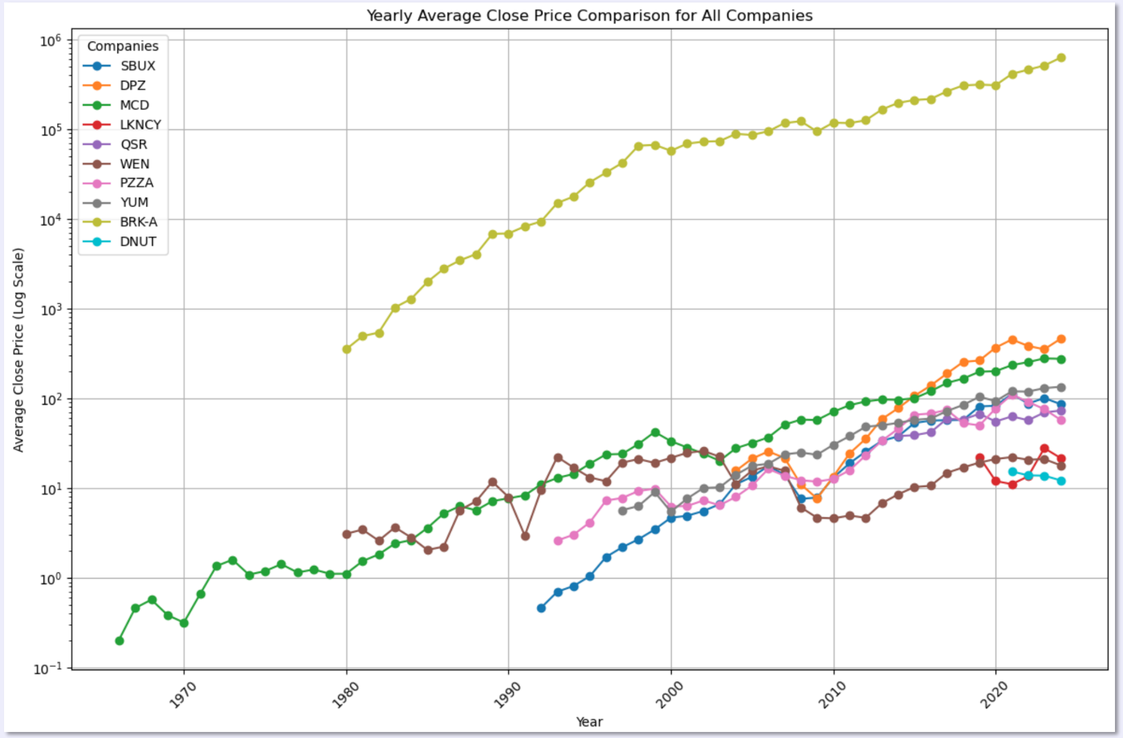
MCD consistently has the highest average close price. WEN shows the lowest average across all months.
There are minimal monthly variation for all companies.

Exercise 5: Computing and comparing yearly average close prices across all companies

Steps taken:

- Grouped the data by year for each company
- Calculated the yearly average Close price for each company
- Stored the yearly averages in a dictionary
- Created a line plot with logarithmic y-scale to handle large variations in Close prices among companies
- Formatted the plot

Line plot with a logarithmic y-scale: highlights disparities in value and stability



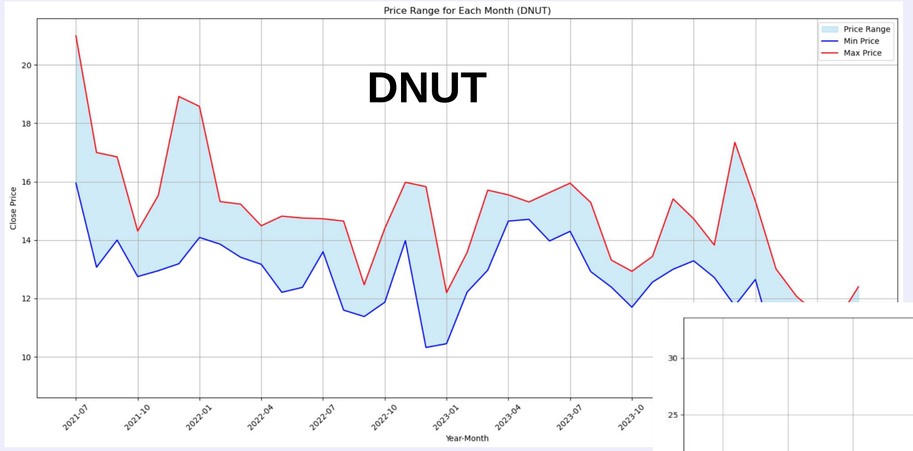
Companies like BRK-A have significantly higher average closing prices and sustained growth, reflecting their dominance and market value, while smaller companies like DNUT and WEN exhibit lower values and more modest growth. Most companies exhibit upward trends in recent years, reflecting broader market growth or recovery from economic downturns.

Exercise 6: Visualizing the range of prices for each month for each company

Steps taken:

- Grouped the data by year-month for each company
- Extracted the YearMonth value from the Date column
- Calculated the range of Close prices (min and max) for each month
- Created a plot with the min and max values, highlighting the price range using a filled area

Filled line plot: variability in prices across months (min, max, and range)



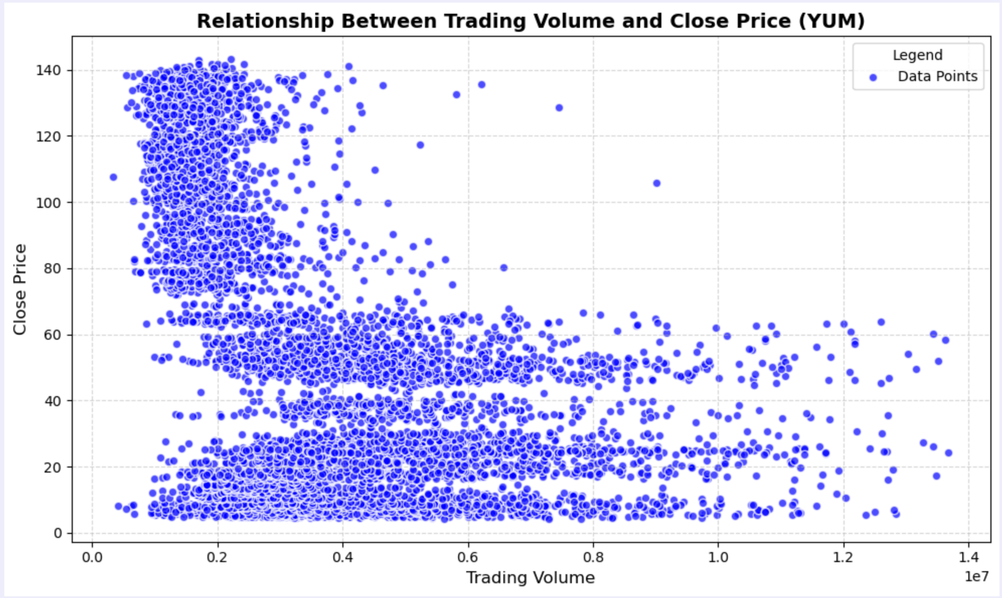
DNUT: relatively consistent price range with notable spikes in volatility during late 2023 and early 2024, possibly influenced by market adjustments or company-specific factors, limited historical range
WEN: significant price range spikes during economic crisis: 2008 financial crisis, and the COVID-19 pandemic in 2020

Exercise 7: Volume – Close Price Relation

Steps taken:

- Calculated the **correlation** between trading volume and close price for all datasets to identify the strongest relationship.
- Selected **YUM** as it had the highest correlation (**0.47**) among all companies.
- Created a **scatter plot** for YUM to visualize the relationship between trading volume and close price.

Company	Correlation
BRK-A	-0.25
DNUT	-0.04
DPZ	0.05
LKNCY	0.28
MCD	0.04
PZZA	-0.14
QSR	0.08
SBUX	-0.45
WEN	-0.08
YUM	-0.47



Higher trading volumes in YUM are moderately correlated with significant price changes, indicating that increased trading activity often coincides with market events or investor reactions. Most trades occur at lower volumes, with higher volumes showing greater price variability, reflecting potential market catalysts or extraordinary events.

Exercise 8: Volume – Top Volume Month

High Trading Volumes Often Align with Market or Company Milestones

Steps taken:

- Grouped trading volume data by year and month for all companies.
- Identified the month with the highest total trading volume for each company.
- Created a summary table highlighting the peak volume month and total trading volume per company.

Company	Month	Volume
BRK-A	October 2008	2,821,350
DNUT	April 2024	56,887,750
DPZ	August 2011	26,953,250
LKNCY	January 2020	198,657,000
MCD	October 2008	282,264,600
PZZA	August 2018	37,345,150
QSR	March 2020	78,798,500
SBUX	January 2008	782,555,000
WEN	June 2009	175,335,200
YUM	October 2009	175,288,047

LKNCY's 198M volume in January 2020 reflects its scandal and early COVID-19 impact. BRK-A and MCD spikes in October 2008 show resilience during the financial crisis. SBUX's 782M in January 2008 highlights confidence amid market volatility. PZZA (August 2018) and YUM (October 2009) reflect fast-food growth and recovery strategies.

Exercise 9: Yearly Aggregated Combined Dataset with no Missing Values

Time-Series Back-casting with Enhanced ARIMA Adjustments:

Inclusive of Automated Parameters and Additional Adjustments

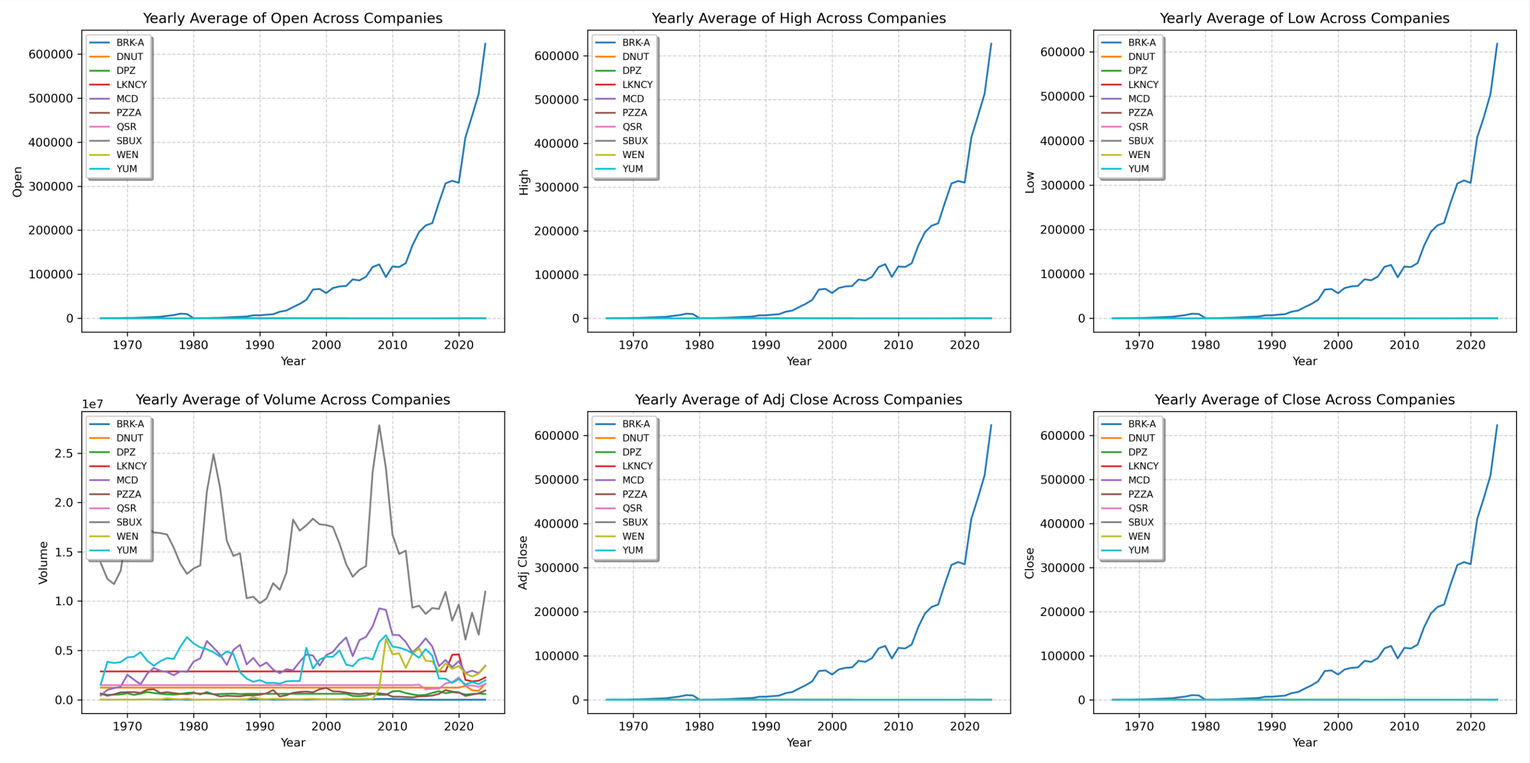
Steps taken:

1. For each Company, grouped data by year and dropped all non-stock related columns
2. Combined yearly data from all companies with Outer Join on Year
3. Imputed Missing Values using ARIMA with modifications.
4. **Reasons for ARIMA:**

- Model past values and errors for realistic backcast in time-series.
- Removes trends and focuses on the data's core structure.

Additional Adjustments:

- Non-negative floor before the 1st available date with data
- Smoothing Mechanism



ARIMA (Auto-Regressive Integrated Moving Average) contains 3 parameters which we automated

1. **Auto-Regressive Order (p):** # Lagged Observations → Captures the effect of future values on earlier data points
2. **Differencing Order (d):** # Times Data is Differenced → Removes noise by focusing on underlying structure of the data via adjacent values
3. **Moving-Average Order (q):** # Lagged Backcast Errors → Captures the effect of present errors on earlier predictions

Key Insight: BRK-A's values are significantly higher than others in all Prices Relating to Stock

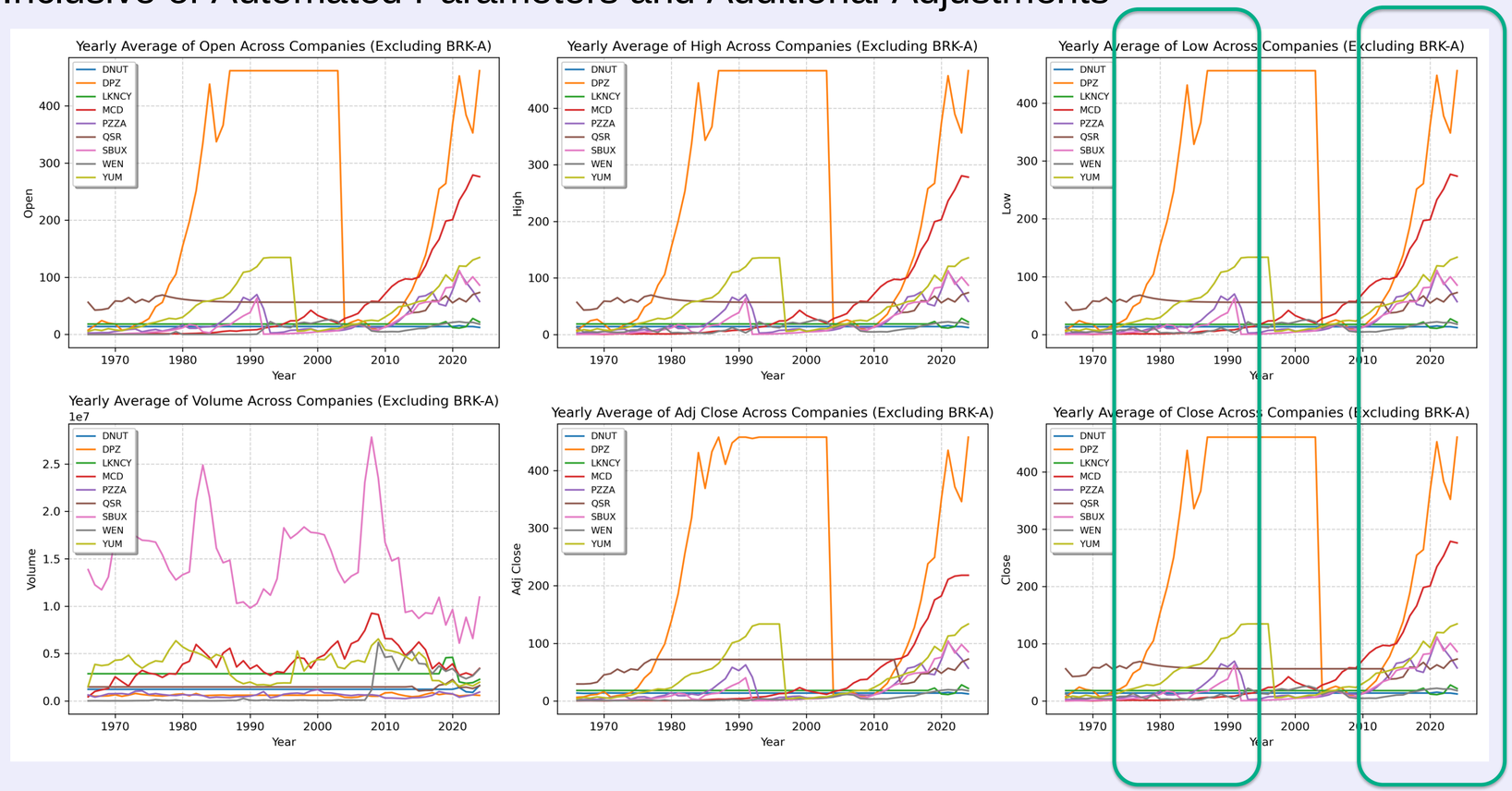
Exercise 9: Yearly Aggregated Combined Dataset with no Missing Values

Structure of Combined Dataset With Proper Alignments:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59 entries, 0 to 58
Data columns (total 61 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Year             59 non-null     int32
 1   BRK-A_Open       59 non-null     float64
 2   BRK-A_High       59 non-null     float64
 3   BRK-A_Low        59 non-null     float64
 4   BRK-A_Close      59 non-null     float64
 5   BRK-A_Adj Close  59 non-null     float64
 6   BRK-A_Volume     59 non-null     float64
 7   DNUT_Open        59 non-null     float64
 8   DNUT_High        59 non-null     float64
 9   DNUT_Low         59 non-null     float64
 10  DNUT_Close       59 non-null     float64
 11  DNUT_Adj Close   59 non-null     float64
 12  DNUT_Volume      59 non-null     float64
 13  DPZ_Open         59 non-null     float64
 14  DPZ_High         59 non-null     float64
 15  DPZ_Low          59 non-null     float64
 16  DPZ_Close        59 non-null     float64
 17  DPZ_Adj Close    59 non-null     float64
 18  DPZ_Volume       59 non-null     float64
 19  LKNCY_Open       59 non-null     float64
...
 60  YUM_Volume       59 non-null     float64
dtypes: float64(60), int32(1)
memory usage: 28.0 KB
None
```

Time-Series Back-casting with Enhanced ARIMA Adjustments:

Inclusive of Automated Parameters and Additional Adjustments



Evaluating our Adjusted ARIMA Approach:

1. Avoided Negative Prediction of Values Before the Year of the First Available Original Data for a Company's Attribute
2. Smoothing Mechanism Partially Effective: Limitations with Data Volatility, Sparsity and Noise

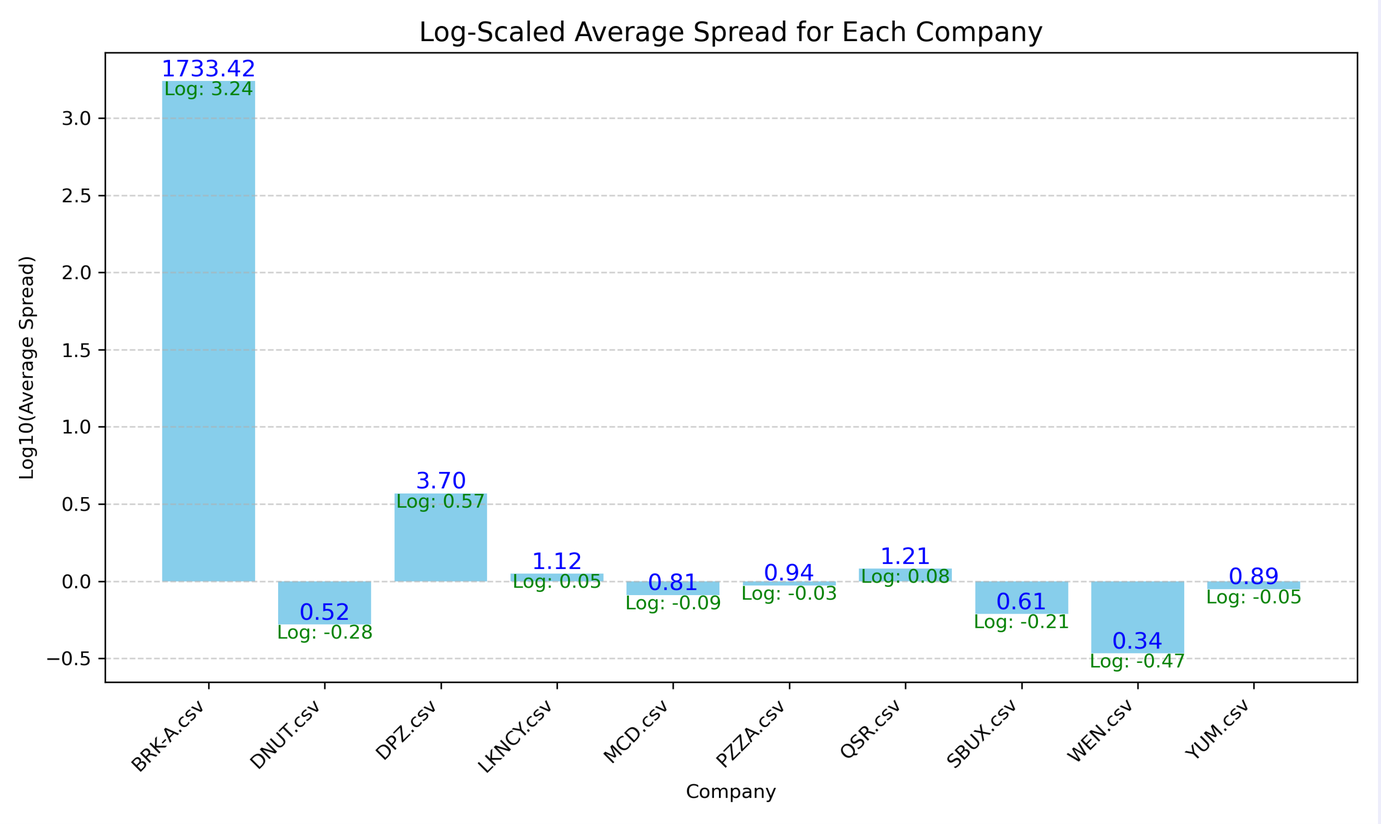
Exercise 10: For Each Company: Daily Spread + Overall Mean Spread

Bar plot: Log Scale on Mean Spread from Starting Year of Each Company

Steps taken:

1. Ensure positive spreads by creating a validation check:
- Apply modulus to ‘High’ if High < Low
- Only execute if ‘High’ > ‘Low’, else it remains 0 for ‘Spread’.

2. Create a Spread (High-Low) column to compute daily spread values for each company.
3. Calculate mean spread over the original data and plot using a logarithmic scale to accommodate BRK-A's magnitude.



- **BRK-A:** Due to its extraordinarily high base stock prices, it amplifies even minor percentage changes into large absolute spreads. This could be reflecting their strategy to attract long-term investors by maintaining exclusivity (e.g., no stock splits).

  DPZ: Due to higher intraday price volatility, it results in higher spreads. This may be due to investor expectations as a growth stock.

- **Companies with Smaller Spreads:** Due to lower stock prices and reduced market volatility, it then reflects nominal daily spreads, as seen in DNUT, QSR, and others.

esade

Do Good. Do Better.