BIG DATA ANALYTICS

ETL Pipeline BDA MID TERM

Iqra Irshad (DS-041/24-25)

ETL Pipeline BDA MID TERM

# Project Title: Spending Trend and Usage Analysis of Debit Card Users

## 1. Introduction

This project presents an ETL (Extract, Transform, Load) pipeline applied to stock market data with the aim of analyzing volatility, trends, and statistical properties. The primary dataset includes daily historical stock data, and the analysis includes the S&P 500 index along with multiple company stocks.

## 2. Data Extraction

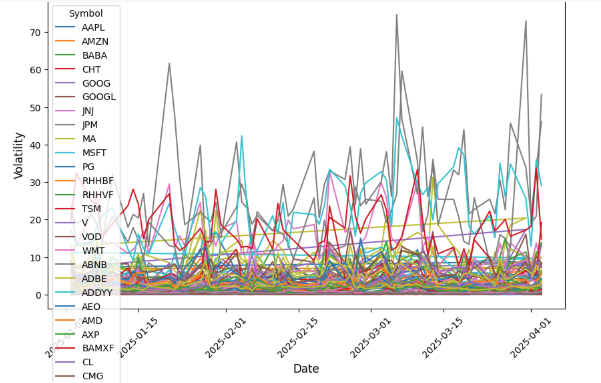
The data is extracted using Python libraries such as `yfinance`, retrieving historical stock prices for various companies and indices like the S&P 500 (^GSPC). The extracted data contains fields such as:  
- Open  
- Close  
- High  
- Low  
- Volume  
- Ticker  
- Date

## 3. Data Overview

The initial examination of the dataset involves methods like `.info()`, `.describe()`, `.head()`, and `.tail()`.

## 4. Volatility Analysis

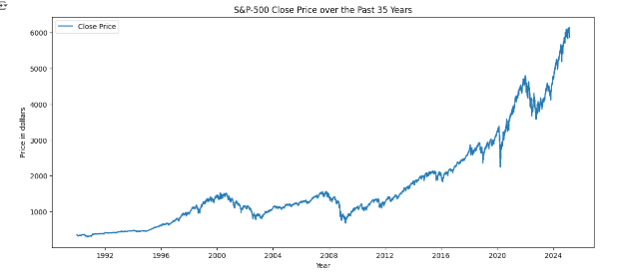
Volatility is calculated and visualized for different companies to understand stock price fluctuations over time.



This line graph displays the daily volatility of selected companies between January and April 2025. Stocks like CMG and CHT show higher volatility.

## 5. S&P 500 Historical Trend

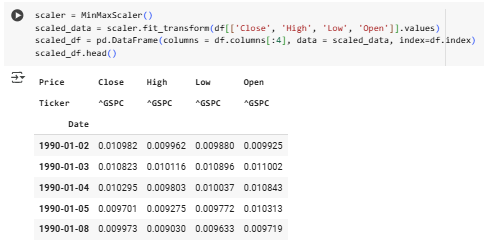
The S&P 500 index data from 1990 to 2025 is plotted to analyze long-term market trends.



This figure demonstrates a clear upward trend with significant dips around the 2008 financial crisis and the COVID-19 pandemic in 2020.

## 6. Feature Scaling

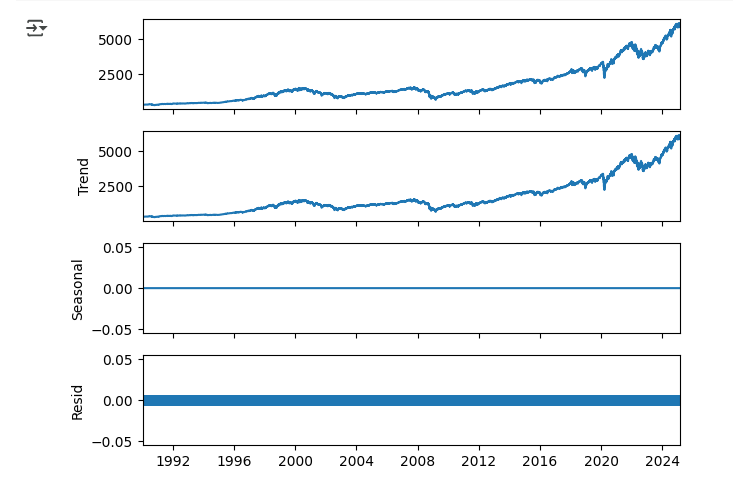
The features (Close, High, Low, Open) are scaled using MinMaxScaler for better model performance and uniformity.  
  
```python  
scaler = MinMaxScaler()  
scaled\_data = scaler.fit\_transform(df[['Close', 'High', 'Low', 'Open']].values)  
```



This snippet and output preview show the normalized values for better comparison.

## 7. Distribution Analysis

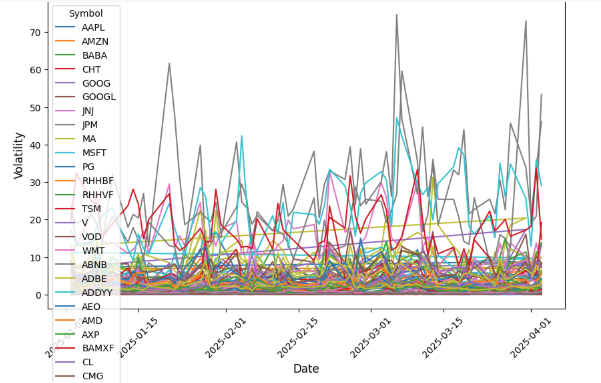
KDE (Kernel Density Estimation) plots help in understanding the distribution of features like Open, High, Low, Close, and Volume.



The plots indicate a right-skewed distribution for prices, typical of stock market data.

## 8. Time Series Decomposition

To analyze the components of the S&P 500 time series, the data is decomposed into Trend, Seasonal, and Residual parts.



This shows a strong trend component while the seasonal and residual components are negligible or zero, indicating consistent long-term growth.

## 9. Conclusion

This ETL pipeline efficiently extracts, transforms, and loads stock market data to uncover key insights.  
It begins by extracting historical data using APIs like yfinance, followed by data cleaning, normalization, and statistical analysis to highlight trends and volatility. Visualizations such as volatility plots, KDE charts, and time series decomposition help in identifying market behavior patterns.

By processing raw financial data into structured, visual formats, the pipeline enables analysts to make data-driven decisions, understand risk factors, and observe long-term market trends effectively.

## 10. Appendix: Code

The accompanying Python script includes all steps of the ETL pipeline, from extraction to visualization.  
- File: `ETL PIPELINE.py`