Stock Price Time Series Forecast: Exploratory Data Analytics

Group 8: Ahmad Javed, Steven Sullivan, Robert Lignowski, Uditi Shah

Project Summary/ Recap

Our project details the acquisition and preprocessing of historical stock data for Ford, Volkswagen, Toyota, and Tesla from Investing.com. We selected these companies to analyze trends in both traditional and electric vehicle markets. The acquisition process involved downloading structured CSV files, ensuring consistency, and standardizing formats for accurate analysis. During preprocessing, we handled missing values, inconsistent volume formatting, and data noise using statistical methods like median imputation and numerical conversions. Standardizing date formats enabled effective time-series analysis, ensuring data integrity for predictive modeling and financial insights.

Exploratory Data Analysis

Data Overview:

- Source: Investing.com & Microsoft Excel
- Date Range: Jan 2, 2020 – Jan 2, 2025
- Features: Close, Open, High, Low, Volume, % Change

Data Cleaning & Processing:

- Converted date to datetime format
- Cleaned Volume & % Change columns
- Ensured no missing values

Exploratory Data Analysis

Key Insights:

- Stock Trend: Fluctuations observed over five years
- Volume Trends: Spikes during major market events
- Closing Price Distribution: Normal with outliers
- Correlation: High correlation among price features

Visualizations:

- Line Charts: Closing price & volume trends
- Histograms: Price & volume distribution
- Box Plot: Outlier detection
- Heatmap: Feature correlations

Purpose of Feature Engineering

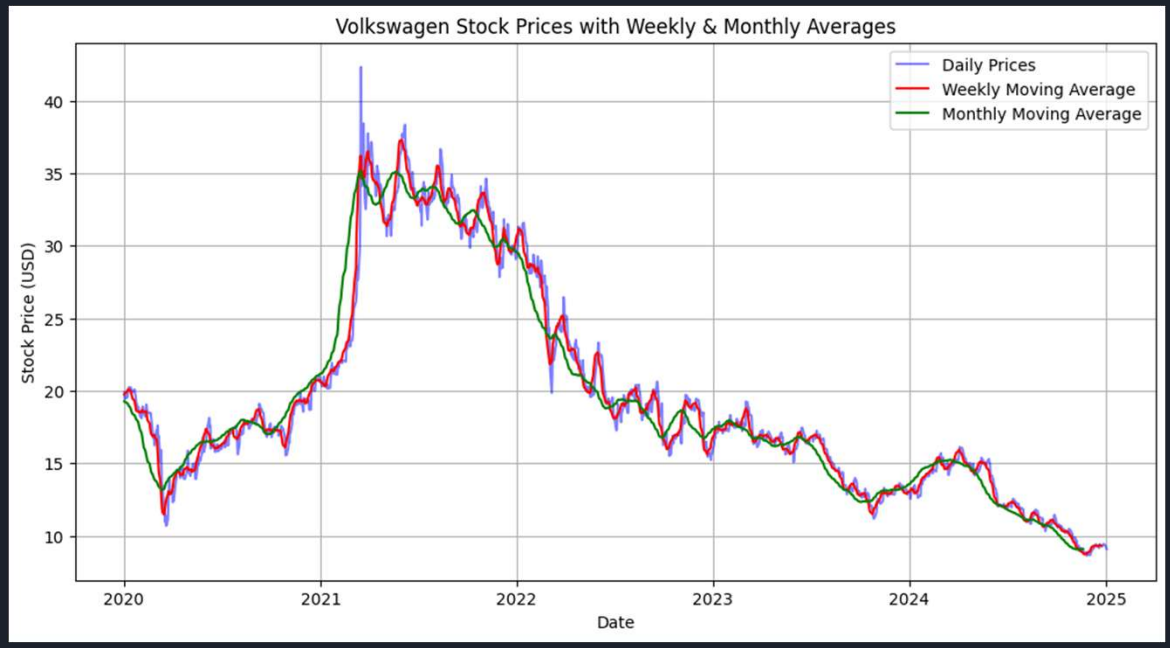
Why Feature Engineering?

- Transforms raw stock data into more meaningful inputs.
- Helps models recognize trends, seasonality, and volatility.
- Reduces redundancy by removing highly correlated features.

Rolling Averages & Trend Smoothing

Key Features Created:

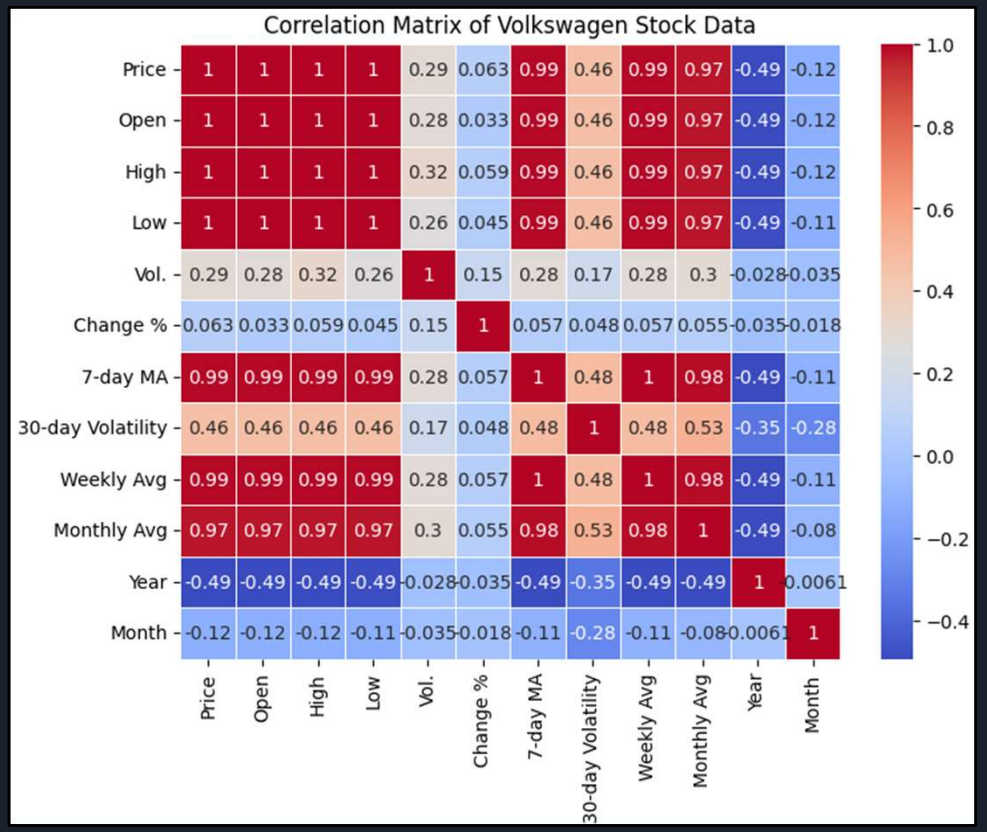
- Weekly & Monthly Moving Averages: Smoothed price trends.
- Trend Extraction: Identified long-term stock movements.
- Volatility Measures: Quantified price fluctuations.



Correlation Analysis & Feature Selection

Findings from Correlation Analysis:

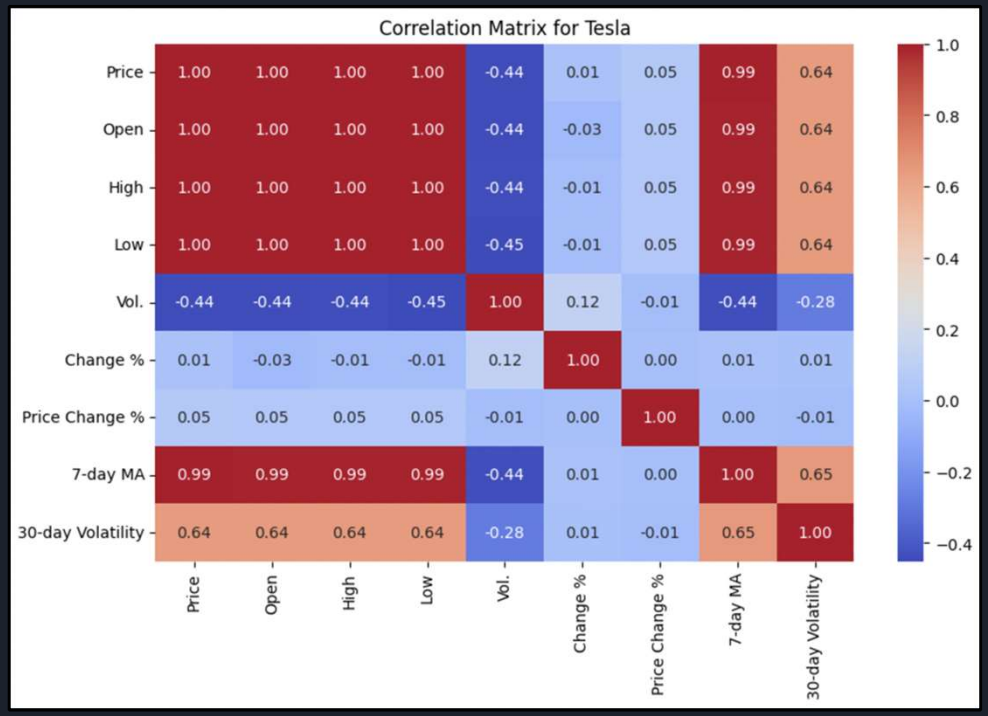
- High correlation among Open, High, Low, and Close prices → Some features removed.
- Trading Volume had weak correlation with stock prices.
- 7-day Moving Average & 30-day Volatility retained as meaningful features.



Impact of Feature Engineering

Key Takeaways from Feature Engineering:

- Improved Trend Detection: Rolling averages reduced short-term noise.
- Stronger Predictive Power: Selected key features while removing redundancies.
- More Efficient Model: Simplified dataset, making training faster.



Finalized Data

- Cleaned, and verified
  - Converted to datetime
  - Dropped missing variables
- Taken steps to deal with leakage
  - Created lagging features
  - Lagging features are important for time series forecasting
    - Help show trends and seasonality
    - Enable models to learn from past behavior
  - Data leakage interferes with lagging because it inserts data from the “future” into the analysis
    - Therefore, it is helpful to shift the data slightly so only data from before a certain point is used.
- Splitting
  - Splitting data into training (80%) and testing (20%)

Next Step: Pipeline Model

- Time Series Forecasting
  - Using historical data to predict future values
    - In this case, looking at trends in stock prices to predict future prices
- Different approaches to choose from
  - ARIMA - AutoRegressive Integrated Moving Average
    - Statistical model used for analyzing and forecasting time series data by combining autoregression, differencing, and moving averages to capture patterns over time.
  - Other Machine learning models
    - Linear Regression
    - Random forest
    - XGBoost
    - LSTM (Long Short-Term Memory)

Key Lessons: Data Acquisition and Preprocessing

- Ensuring the data is clean and consistent is crucial for accurate analysis. Missing values in key columns such as volume and price can distort results. Addressing this by filling missing data with median values helps maintain the integrity of the dataset.

- When working with multiple data sources (e.g., Ford, Volkswagen, Toyota, and Tesla), it's essential to standardize columns like volume and percentage change. This involves converting units (e.g., "M" for millions) and removing non-numeric characters to ensure consistency and comparability across datasets.

- Outliers, identified through visualizations like box plots, can heavily influence stock price analysis. It's important to identify and address outliers to improve the accuracy and reliability of the results.

Key Lessons: Exploratory Data Analysis (EDA)

- Visualizing data, such as stock prices and trading volumes over time, is key to identifying trends and spotting outliers. For example, spikes in trading activity and significant price movements on specific days provide valuable insights into market behavior.

- By analyzing the correlations between variables, we identified which factors were most relevant for predictive modeling. This process helps reduce redundancy in the model and ensures that the focus is on the most impactful features.

- Ensuring that the date column is formatted consistently is essential for time series analysis. Inconsistent date formats can lead to misalignment of data across different stocks, potentially disrupting the accuracy of the analysis.

Key Lessons: What Influences Car Stocks?

- Macroeconomic Factors
  - Interest Rates
  - Inflation
  - Consumer Spending
- Industry Specific Factors
  - Car Sales & Demands
  - EV Transition
  - Vehicle recalls
- Market & Competitor Factors
  - Earning Reports
  - Mergers & Acquisitions
  - Partnerships
- Government Policies & Regulations
  - Emission Standards
  - EV incentives (tax credits)
  - Trade policies & Tariffs
- Global Events & Disruptions
  - Geopolitical Conflicts
  - Trade Wars
  - Pandemics

Next Steps

- **Date-Time Indexing**
  - It's crucial to index a dataset by date and ensure the date column is in the correct datetime format. This step is vital for time series analysis, as it preserves the order of data points.

- **Stationarity Check**
  - ARIMA models require the time series to be stationary, meaning that its statistical properties, such as mean and variance, remain constant over time. Ensuring stationarity is a key step in preparing the data for accurate modeling.

- **Additional Analysis**
  - To enhance predictive power, adding external data sources (e.g., news article, social media trends) related to the companies can provide additional insights

- **Model Validation**:
  - It's important to validate our model using unseen future data (January - March 2025) to ensure that our model performs well in predicting outcomes which will result in providing long-term accuracy.