Stock Price Data Cleaning, Outlier Treatment, and Correlation Analysis (Python)

Portfolio Source Report

Prepared by: Felix Stephen Aidoo
Date: January 28, 2026
Location: Berlin (Remote)

Table of Contents

1. EXECUTIVE SUMMARY 3

2. OBJECTIVES 3

3. DATASET SUMMARY 3

4. TOOLS USED 3

5. METHODOLOGY OVERVIEW 3

6. DATA CLEANING AND PREPROCESSING 3

6.1 Date Standardization 3
6.2 Missing Values 4
6.3 Duplicate Records 4
6.4 Outlier Detection and Treatment 4
6.5 Post-processing Verification 4

7. EXPLORATORY DATA ANALYSIS (EDA) 4

7.1 Summary Statistics 4
7.2 Correlation Analysis: Open vs Close 4
7.3 Correlation Analysis: High vs Low 4

8. KEY FINDINGS 5

9. RECOMMENDATIONS 5

10. APPENDIX 5

1. Executive Summary

This report documents a small end-to-end analysis of daily stock price data. The workflow focuses on improving data quality (missing values, data types, outliers) and generating descriptive insights through correlation analysis and visual reporting. Results show strong positive relationships between Open and Close prices and between High and Low prices, supporting consistent trend behavior across the period analyzed.

2. Objectives

- Import and inspect the dataset structure (shape, data types, summary statistics).
- Clean and preprocess the data (date parsing, missing values, duplicates, outliers).
- Perform exploratory data analysis (EDA) and visualize trends and relationships.
- Quantify correlations between Open vs Close and High vs Low prices.
- Summarize findings and provide practical recommendations.

3. Dataset Summary

Dataset characteristics used in this analysis:

- Type: Daily stock market price data.
- Size (before cleaning): ~3,019 rows × 7 columns.
- Key fields: Date, Open, High, Low, Close, Volume, Name.

4. Tools Used

- Python: pandas, numpy
- Visualization: matplotlib, seaborn
- Statistics: scipy/stats (Z-score for outlier detection)
- Environment: Jupyter Notebook

5. Methodology Overview

1. Load the dataset and review structure using head(), info(), and describe().
2. Convert the Date column to datetime for reliable time-based analysis.
3. Assess and handle missing values; check and remove duplicates if present.
4. Detect outliers in Open, High, Low, Close using Z-score ($|z| > 3$).
5. Treat outliers by replacing extreme values with the column mean (as a simple, transparent approach).
6. Run correlation analysis and visualize results using heatmaps, scatterplots, and trend lines.
7. Summarize findings and provide recommendations and next steps.

6. Data Cleaning and Preprocessing

6.1 Date Standardization

The Date column was converted to datetime format to support consistent sorting, grouping, and plotting.

6.2 Missing Values

A missing value check identified a small number of incomplete rows. Because the missing rate was low, incomplete rows were removed to preserve analysis consistency.

- Missing rows identified: 18
- Approximate missing rate: ~0.59%
- Rows after removal: ~3,001

6.3 Duplicate Records

Duplicate rows were checked, and no duplicate records were detected in the dataset.

6.4 Outlier Detection and Treatment

Outliers were identified for Open, High, Low, and Close prices using Z-scores. Values with $|z| > 3$ were considered outliers. To reduce distortion from extreme points, identified outliers were replaced with the mean of the corresponding feature.

- Outlier method: Z-score (threshold = 3)
- Treatment: Replace outliers with column mean
- Fields treated: Open, High, Low, Close

6.5 Post-processing Verification

After cleaning, summary statistics were recalculated to confirm expected value ranges and overall consistency.

7. Exploratory Data Analysis (EDA)

7.1 Summary Statistics

Descriptive statistics were generated for numeric columns to understand central tendency, spread, and ranges.

7.2 Correlation Analysis: Open vs Close

Open and Close prices exhibited an extremely strong positive correlation (near 1). This indicates that when the market opens higher, it typically closes higher as well (and vice versa).

- Evidence: correlation matrix and heatmap
- Visualization: scatterplot of Open vs Close
- Trend plot: Open and Close lines over time

7.3 Correlation Analysis: High vs Low

High and Low prices also showed a strong positive relationship. This suggests daily price ranges scale together across the observation period.

- Visualization: scatterplot of High vs Low
- Supporting chart: correlation heatmap

8. Key Findings

1. Open and Close prices move almost in parallel, showing a very strong positive correlation.
2. High and Low prices are strongly positively related, indicating consistent scaling of daily price ranges.
3. Time series plots show Open and Close share similar movement patterns over time.

9. Recommendations

- Track correlation and trend stability over time; meaningful deviations may indicate market shifts or unusual events.
- Pair quantitative patterns with external context (e.g., earnings/news) when interpreting abrupt changes.
- Extend the analysis with rolling volatility metrics and a simple baseline forecast to move from descriptive to predictive analytics.

10. Appendix

Artifacts produced for this project typically include:

- Jupyter Notebook containing data import, cleaning, and EDA steps.
- SQL-ready KPI logic (if replicated in SQL).
- Visual outputs: heatmap, scatterplots, and time series trend lines.
- One-page case study summary for quick recruiter review.