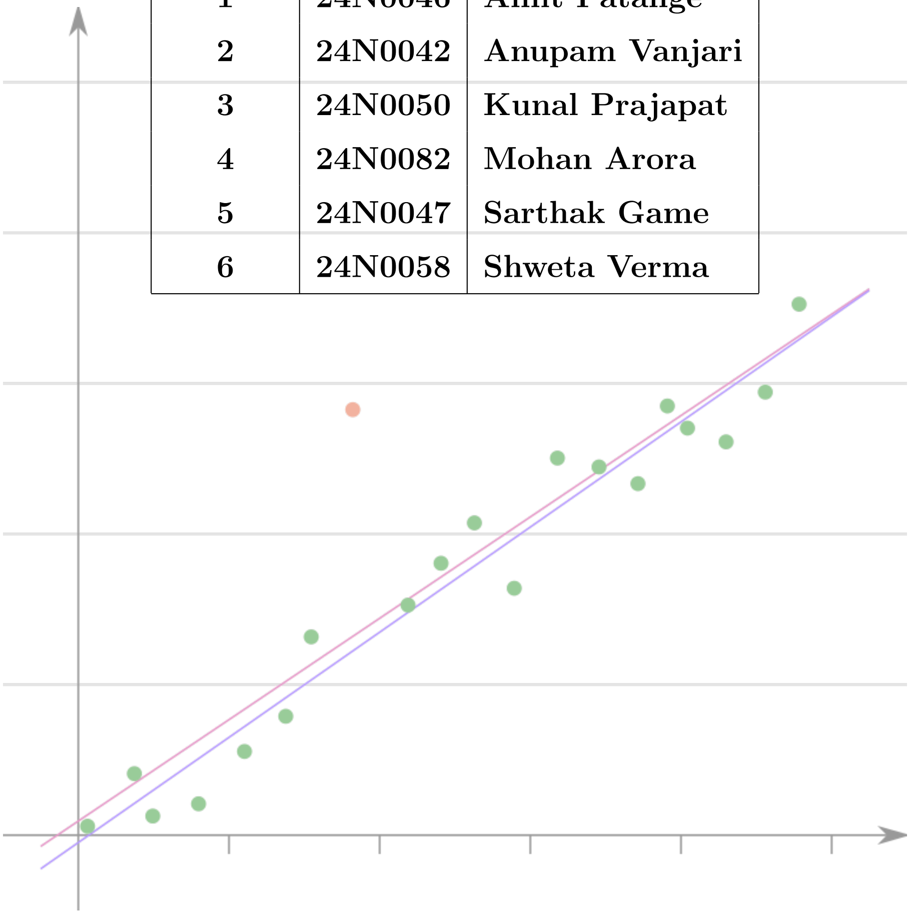
Department of Mathematics, IIT Bombay

Regression Analysis and Diagnostics

Course Project: Regression Analysis (SI-422)

Guide: Prof. Monika Bhattacharjee

Sr. No.	Roll No.	Name
1	24N0046	Amit Patange
2	24N0042	Anupam Vanjari
3	24N0050	Kunal Prajapat
4	24N0082	Mohan Arora
5	24N0047	Sarthak Game
6	24N0058	Shweta Verma



Contents

0.1 Introduction and Motivation . . . 2
0.2 Decoding Financial Metrics . . . 3
0.3 Objective and Scope . . . 4
0.4 Dataset Overview . . . 4
0.5 Feature Engineering . . . 5
0.6 Exploratory Data Analysis . . . 6
0.7 Outlier Detection . . . 10
0.8 Assumptions Validation . . . 12
0.9 Limitations and Future Scope . . . 14

0.1 Introduction and Motivation

Financial markets operate under uncertainty, driven by a multitude of economic and financial metrics that influence investment performance and market dynamics. The complexity of these markets stems from the interconnected nature of various factors including macroeconomic indicators, company-specific fundamentals, sector performance, and investor sentiment, all of which contribute to the volatility and unpredictability of stock returns. In this challenging environment, investors and financial analysts continuously seek reliable methods to predict investment outcomes and optimize portfolio performance.

This comprehensive report aims to provide a robust and scientifically rigorous approach to predicting nominal returns for random investments at the New York Stock Exchange using advanced regression techniques and statistical modeling. The study leverages a substantial dataset comprising over 400,000 simulated investment transactions spanning the last decade, encompassing diverse financial instruments across multiple sectors including Banking, Fast-Moving Consumer Goods (FMCG), Retail, and Technology. This extensive dataset provides a rich foundation for developing predictive models that can capture the complex relationships between various financial metrics and investment returns.

The research employs sophisticated modeling strategies that go beyond traditional linear regression approaches, incorporating advanced techniques such as Principal Component Analysis (PCA) for dimensionality reduction, multi-layered outlier detection methods, and regularization techniques including Ridge and Lasso regression. These methodologies are designed to enhance both prediction accuracy and model interpretability, ensuring that the resulting insights are not only statistically sound but also practically applicable for investment decision-making.

The study addresses critical challenges in financial modeling, including multicollinearity among predictor variables, the presence of influential outliers, and the need to validate fundamental regression assumptions. Through rigorous statistical testing and model validation procedures, this research aims to develop a reliable framework for predicting nominal returns that can withstand the scrutiny of academic peer review while providing actionable insights for financial practitioners.

Furthermore, the analysis incorporates Environmental, Social, and Governance (ESG) factors alongside traditional financial metrics, reflecting the modern investment landscape's increasing emphasis on sustainable investing. This holistic approach ensures that the predictive models capture contemporary market dynamics and investor preferences, making the findings relevant to current and future investment strategies.

The ultimate goal of this research is to bridge the gap between theoretical statistical modeling and practical investment applications, providing a comprehensive toolkit for predicting stock returns that combines academic rigor with real-world applicability. By employing multiple validation techniques and comparing various modeling approaches, this study aims to identify the most effective methods for predicting nominal returns while maintaining transparency in methodology and acknowledging the inherent limitations of any predictive model in the complex domain of financial markets.

0.2 Decoding Financial Metrics

Understanding financial metrics is crucial for effective investment analysis and portfolio management. These quantitative measures provide insights into company performance, risk assessment, and investment attractiveness. The following key financial metrics form the foundation of our regression analysis and serve as primary predictors for nominal return estimation:

- **Nominal Return**: The percentage gain or loss on an investment without adjusting for inflation, taxes, or other external factors. This raw return metric represents the actual monetary gain or loss experienced by investors and serves as our primary target variable. Nominal returns are essential for understanding the absolute performance of investments over specific time periods.

- **Sharpe Ratio**: A risk-adjusted measure that quantifies excess return per unit of volatility, calculated as (portfolio return - risk-free rate) divided by portfolio standard deviation. This metric helps investors understand how much additional return they receive for the extra volatility endured, making it invaluable for comparing investments with different risk profiles.

- **Volatility Scale**: Represents the degree of variation in investment returns over a period, commonly measured by standard deviation. Higher volatility indicates greater price fluctuations and investment uncertainty, while lower volatility suggests more stable returns. This metric is fundamental for risk assessment and portfolio diversification strategies.

- **Expected Yearly Return**: The anticipated average return on an investment over a year, based on historical data, statistical models, or predictive algorithms. This forward-looking metric helps investors set realistic expectations and make informed allocation decisions across different asset classes and time horizons.

- **Price-to-Earnings (P/E) Ratio**: A valuation metric calculated by dividing the current market price per share by earnings per share (EPS). This ratio indicates how much investors are willing to pay for each dollar of earnings, providing insights into market expectations about future growth. Lower P/E ratios may suggest undervaluation, while higher ratios might indicate growth expectations or overvaluation.

- **Price-to-Book (P/B) Ratio**: The ratio of a company's market capitalization to its book value, reflecting how much investors are paying relative to the company's net asset value. This metric is particularly useful for value investing strategies, as companies trading below their book value (P/B ¡ 1) may represent potential bargains, though this requires careful fundamental analysis.

- **Return on Equity (ROE)**: A profitability metric measuring how efficiently a company generates profits from shareholders' equity, calculated as net income divided by shareholders' equity. Higher ROE values indicate more efficient use of investor capital and generally suggest stronger management performance and competitive positioning within the industry.

- **Current Ratio**: A liquidity metric that measures a company's ability to pay short-term obligations, calculated as current assets divided by current liabilities. A current ratio above 1.0 indicates that the company has more current assets than current liabilities, suggesting good short-term financial health and operational efficiency.

- **ESG Ranking**: Environmental, Social, and Governance score that evaluates a company's sustainability practices and ethical business conduct. ESG factors have become increasingly important in investment decisions, as companies with higher ESG scores often demonstrate better long-term risk management, stakeholder relationships, and regulatory compliance, potentially leading to more sustainable returns.

These metrics collectively provide a comprehensive framework for evaluating investment opportunities across multiple dimensions including profitability, valuation, risk, liquidity, and sustainability. The interplay between these variables forms the basis of our predictive modeling approach, allowing us to capture the complex relationships that drive nominal returns in modern financial markets. These metrics, along with sectoral aspects and engineered features, provide the building blocks of the analysis.

0.3 Objective and Scope

The primary objective of this research is to develop and validate robust regression models capable of accurately predicting nominal returns for stock investments at the New York Stock Exchange, leveraging comprehensive financial ratios, market volatility indicators, and sector-specific characteristics. The scope encompasses systematic evaluation of multiple regression techniques including Ordinary Least Squares (OLS), Ridge regression, and Lasso regression, coupled with advanced feature engineering through Principal Component Analysis to address multicollinearity issues. Rigorous diagnostic testing protocols are implemented to validate fundamental regression assumptions including linearity, independence of residuals, homoscedasticity, and normality of error terms. The study employs sophisticated outlier detection methodologies utilizing Cook's distance, Mahalanobis distance, and Isolation Forest algorithms to ensure model robustness and reliability. Model selection is conducted through comprehensive comparison using statistical criteria such as Mallow's Cp, cross-validation techniques, and performance metrics to identify the optimal predictive framework. The ultimate goal is to deliver a scientifically validated, practically applicable tool for investment decision-making that bridges academic rigor with real-world financial market applications..

0.4 Dataset Overview

This study utilizes a rich dataset comprising 405,258 simulated investment transactions executed on the New York Stock Exchange between 2015 and 2025. Each record captures critical trade parameters and comprehensive financial metrics to support predictive modeling of nominal returns. The core data fields include investment horizon (measured in days) and transaction amount, which together characterize the temporal and monetary dimensions of each investment scenario. Risk and performance indicators are represented by separate measures of volatility for buy and sell orders, along with the Sharpe ratio, providing insights into price fluctuations and risk-adjusted return. The dataset also records expected yearly return forecasts, applied inflation rates, and the realized nominal return for each simulation, enabling analysis of real versus anticipated performance. Fundamental valuation metrics span Price-to-Earnings (P/E) and Price-to-Book (P/B) ratios, Earnings Per Share (EPS), Price-to-Sales (PS) ratio, Net Profit Margin, and key profitability ratios such as Return on Assets (ROA) and Return on Equity (ROE). Liquidity and solvency are captured via the current ratio, while sustainability considerations are integrated through standardized Environmental, Social, and Governance (ESG) rankings. Advanced engineered features, including Principal Component Analysis–derived components capturing combined Sharpe/volatility effects and PS/ROA/net-profit interactions, further enrich the dataset for robust regression analysis.

0.5 Feature Engineering

To address multicollinearity among closely related financial metrics, we employed Principal Component Analysis (PCA) to generate composite features that capture the underlying variance while preserving critical information. Specifically, two new PCA-derived features were created:

- **Combined Sharpe/Volatility**: Synthesizes the Sharpe ratio and buy/sell volatility measures into a single principal component that reflects overall risk-adjusted variability.
- **Combined PS/ROA/Net Profit**: Integrates the Price-to-Sales ratio, Return on Assets (ROA), and Net Profit Margin into a unified component representing fundamental profitability and valuation synergy.

The combined features showed high correlation and elevated VIF scores, but rather than removing them and risking information loss, we combined them using PCA to preserve their explanatory power. This approach reduces dimensionality, mitigates multicollinearity, and retains the maximum possible variance explained by the original variables. As a result, the regression models benefit from improved stability and interpretability, with the principal components capturing the dominant patterns of volatility and profitability in a more compact form.

0.6 Exploratory Data Analysis

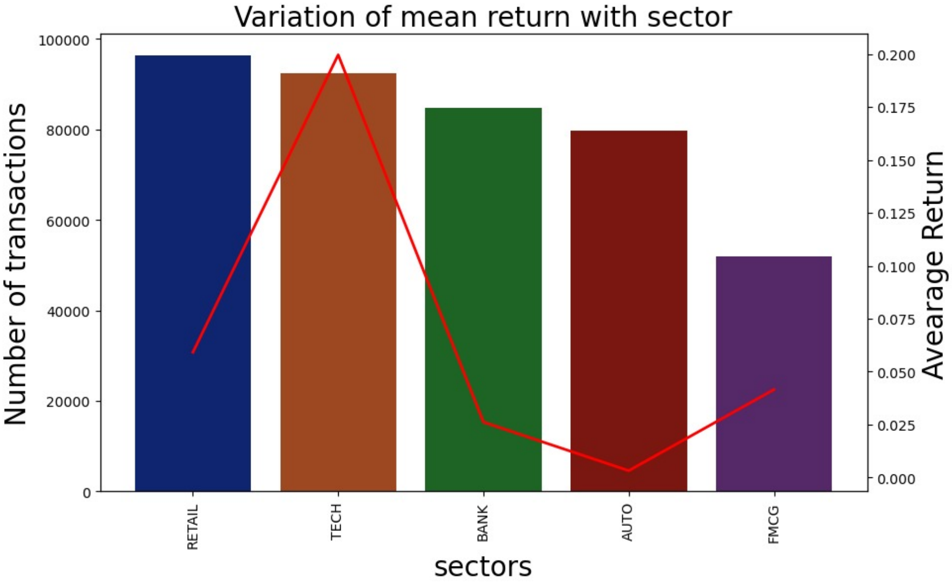


Figure 1: Variation of mean return across different sectors

Figure 1 overlays two key dimensions for each sector: the total number of transactions (left axis) and the average nominal return (right axis). The Retail sector shows the highest trading activity with nearly 96,000 transactions but yields a moderate average return of around 0.07. The Technology sector follows closely in transaction count (92,000) yet achieves the highest profitability, with an average return of approximately 0.20. Banking and Auto exhibit substantial volumes (85,000 and 80,000 transactions respectively) but lower mean returns of about 0.02 and 0.005. Despite having the fewest transactions (52,000), the FMCG sector maintains a respectable average return near 0.10, highlighting a divergence between trading frequency and investment performance across sectors.

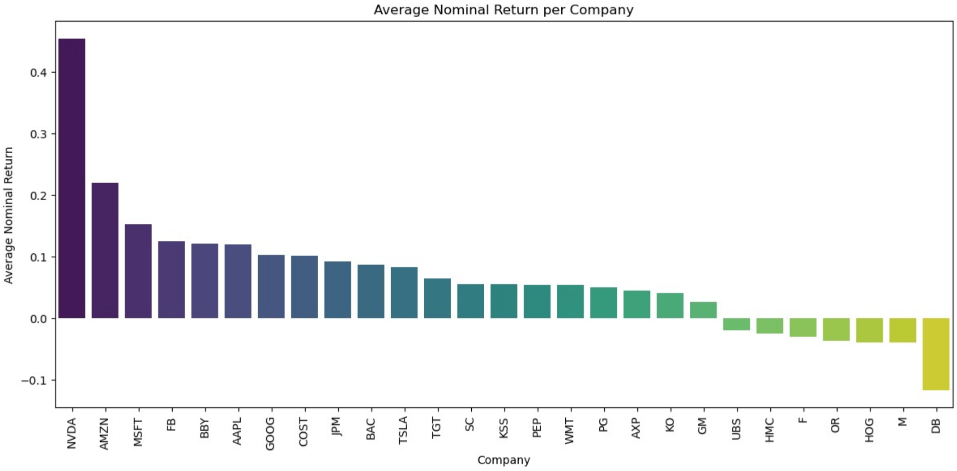


Figure 2: Average Nominal Return per Company

Figure 2 displays the top 20 companies ranked by their average nominal return over the 2015–2025 period. NVIDIA (NVDA) leads with an average return exceeding 0.45, reflecting its exceptional growth and market performance. Technology giants such as

Amazon (AMZN) and Microsoft (MSFT) follow with average returns around 0.22 and 0.15, respectively. Mid-tier performers include Facebook (FB), BlackBerry (BBY), and Apple (AAPL), each achieving returns between 0.12 and 0.13. Companies like Deutsche Bank (DB) and Macy's (M) exhibit negative average returns (around –0.11 and –0.02), highlighting significant underperformance and volatility in certain sectors. This ranking underscores the pronounced dispersion in stock performance and the importance of company-level analysis in investment decision-making.

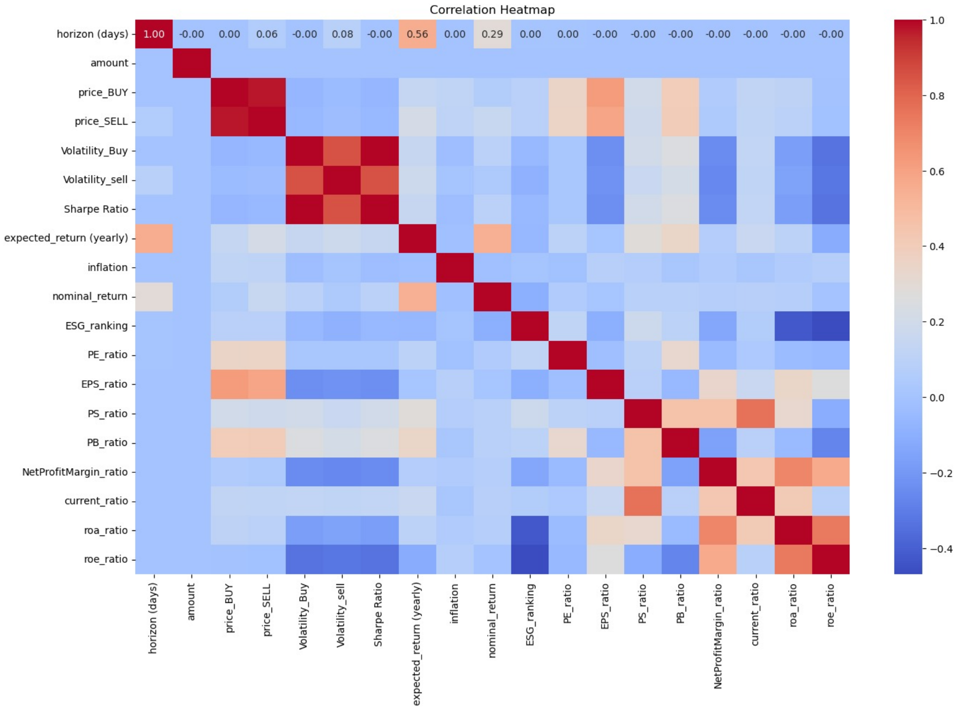


Figure 3: Correlation Heatmap of financial and trading variables

The figure above displays the pairwise Pearson correlation coefficients between various financial and trading-related variables. Let us denote by $r_{X,Y}$ the correlation coefficient between variables X and Y. The entries satisfy $-1 \leq r_{X,Y} \leq 1$, where

$$r_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \, \sigma_Y}.$$

- **Strong Positive Correlations:**
  - $_{\text{BUY}}$ and $_{\text{SELL}}$ exhibit $r \approx +0.95$, indicating they move almost in lockstep.
  - $_{\text{Buy}}$ and $_{\text{Sell}}$ have $r \approx +0.80$, showing market turbulence affects buy-side and sell-side volatility similarly.
  - The fundamental ratios ROE and ROA correlate at $r \approx +0.75$, reflecting that firms with higher return on assets also tend to have higher return on equity.
  - PE ratio and EPS ratio correlate positively around $r \approx +0.45$, since higher earnings per share often coincide with higher price-to-earnings valuations.
- **Strong Negative Correlations:**

  - ROA and the Sharpe Ratio show $r \approx -0.40$, suggesting that assets with higher accounting returns may not always deliver commensurately risk-adjusted performance.
  - Sharpe Ratio versus $_{\text{Sell}}$ has $r \approx -0.35$, indicating more volatile sell signals tend to have lower risk-adjusted returns.

- **Moderate and Weak Correlations:**
  - (holding period) correlates moderately with nominal return ($r \approx +0.30$), implying longer horizons slightly increase raw returns.
  - Inflation shows negligible correlations ($|r| < 0.10$) with most trading variables, meaning short-term portfolio metrics are largely orthogonal to macro inflation in this dataset.
  - ESG_ranking has near-zero correlation with all price and volatility measures, indicating environmental, social, and governance scores are independent of immediate market metrics.

- **Key Insights:**
  1. Price and volatility variables are internally consistent (high mutual correlations), justifying dimensionality reduction when modeling.
  2. Fundamental ratios (PE, PB, EPS, PS) exhibit interdependencies, which should be accounted for to avoid multicollinearity in regression models.
  3. Risk-adjusted measures (Sharpe Ratio) can behave counterintuitively relative to pure accounting returns (ROA, ROE).
  4. Macroeconomic (inflation) and ESG scores present as orthogonal factors, potentially serving as exogenous features in predictive models.

Methodology: Multiple Linear Regression

Multiple linear regression (MLR) is a statistical technique used to model the linear relationship between a continuous dependent variable Y and a set of p independent (predictor) variables $X_1, X_2, \dots, X_p$. The general form of the MLR model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon,$$

where:

- β_0 is the intercept term.
- β_j for $j = 1, \dots, p$ are the regression coefficients quantifying the change in Y associated with a one-unit change in X_j, holding all other predictors constant.
- ε is the random error term, assumed to capture all variability in Y not explained by the linear combination of the predictors.

Estimation of Coefficients

The regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\mathsf{T}$ are estimated via ordinary least squares (OLS), which minimizes the sum of squared residuals:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \bigl(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij}\bigr)^2.$$

In matrix notation, let

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Then the OLS solution is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\, \mathbf{X}^\mathsf{T}\mathbf{Y}.$$

Assumptions of Multiple Linear Regression

For the OLS estimates to be unbiased, efficient, and valid for inference, the following assumptions must hold:

1. **Linearity:** The true relationship between each predictor X_j and the response Y is linear:
$$\mathbb{E}[Y \mid X_1, \dots, X_p] = \beta_0 + \sum_{j=1}^{p} \beta_j X_j.$$
2. **Independence:** The error terms ε_i are independent across observations: $(\varepsilon_i, \varepsilon_{i'}) = 0$ for $i \neq i'$.

3. **Homoscedasticity:** The error variance is constant for all levels of the predictors:
$$(\varepsilon_i) = \sigma^2 < \infty \quad \forall i.$$
4. **Normality of Errors (for inference):** The errors are normally distributed:
$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \text{for } i = 1, \ldots, n.$$
5. **No Perfect Multicollinearity:** The predictor matrix $\mathbf{X}$ has full column rank, i.e.,
$$\det(\mathbf{X}^\mathsf{T}\mathbf{X}) \neq 0,$$
ensuring that $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ exists.

Model Diagnostics

After fitting the model, verify assumptions by:

- Plotting residuals vs. fitted values (check homoscedasticity and linearity).
- Quantile–quantile plot of residuals (check normality).
- Variance inflation factors (VIF) to detect multicollinearity:
$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2},$$
where R_j^2 is the R^2 from regressing X_j on the other predictors.
- Durbin–Watson statistic to test for autocorrelation of residuals.

0.7 Outlier Detection

To enhance model accuracy and stability, a three-stage sequential outlier detection and filtering procedure was implemented. This process significantly improved the model's adjusted coefficient of determination (R^2_{adj}) from 0.41 to 0.75, reflecting a substantial reduction in noise and improved explanatory power.

Stage 1: Cook's Distance — Influential Point Removal

Cook's Distance is a regression diagnostic that measures the influence of each observation on the fitted regression coefficients. For observation i, Cook's distance is defined as:
$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{(p+1)\,\hat{\sigma}^2},$$
where $\hat{Y}_{j(i)}$ is the predicted value for observation j when the i-th observation is removed, p is the number of predictors, and $\hat{\sigma}^2$ is the mean squared error. Observations with $D_i > \frac{4}{n}$ are typically considered influential and were removed in this stage to prevent distortion of subsequent analyses.

Stage 2: Mahalanobis Distance — Multivariate Outlier Removal

Mahalanobis distance identifies points that are extreme in the multivariate feature space while accounting for variable correlations. For a data vector $\mathbf{x}$, it is computed as:

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})},$$

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix of the predictors. Observations exceeding the chi-square critical value $\chi^2_{p,\,\alpha}$ were flagged as multivariate outliers and removed.

Stage 3: Isolation Forest — Final Cleanse

The Isolation Forest algorithm is a tree-based anomaly detection method that isolates observations by randomly selecting features and split values. Anomalies are isolated more quickly, resulting in shorter average path lengths in the ensemble of trees. This method does not assume a particular statistical distribution and is effective for detecting complex, non-linear anomalies. In the final stage, the Isolation Forest identified subtle anomalies that were not captured by traditional statistical diagnostics.

Overall Impact

By applying these methods sequentially—first removing highly influential observations, then addressing multivariate extremes, and finally detecting non-linear anomalies—the dataset was progressively refined. This rigorous filtering boosted R^2_{adj} from 0.41 to 0.75, substantially enhancing model fit and reliability.

Stage	Description
1: Cook's Distance	Detects and removes influential observations that disproportionately affect regression coefficients. Threshold: $D_i > \frac{4}{n}$.
2: Mahalanobis Distance	Identifies multivariate outliers by measuring distance from the mean vector while accounting for covariance. Threshold based on $\chi^2_{p,\alpha}$.
3: Isolation Forest	Tree-based anomaly detection method that isolates anomalies with fewer splits. Captures non-linear and complex outliers without distributional assumptions.

Table 1: Three-stage outlier filtering process improving R^2_{adj} from 0.41 to 0.75.

0.8 Assumptions Validation

To ensure the validity of the multiple linear regression results, standard diagnostic checks were conducted. In real-world datasets, it is often challenging to achieve perfect adherence to the classical regression assumptions. Minor deviations are generally acceptable, provided they do not substantially affect the model's predictive accuracy or interpretability.

No Multicollinearity

Variance Inflation Factor (VIF) values for all predictors were well below the commonly used threshold of 10, indicating no significant multicollinearity. This ensures that predictor variables are not linearly dependent, preserving the interpretability and stability of the coefficient estimates.

Normality of Errors

The histogram and kernel density estimate (KDE) of residuals (Figure 4) show a symmetric, approximately bell-shaped distribution centered at zero. While the curve is not a perfect Gaussian—something common in applied datasets—it is close enough to justify the normality assumption for inference.

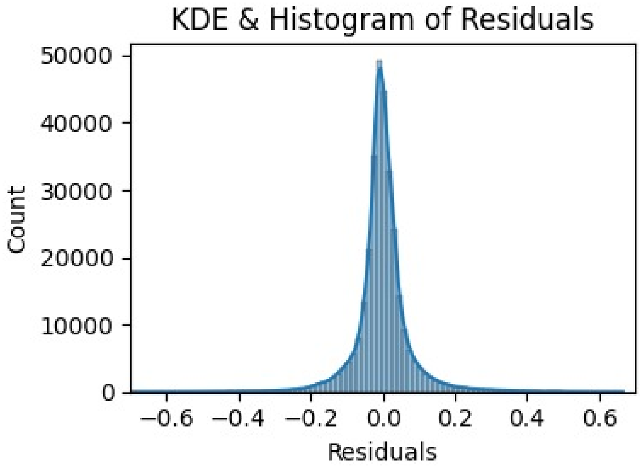


Figure 4: KDE and histogram of residuals showing approximate normal distribution. Minor deviations are acceptable in real-world data.

Independence of Errors

The residuals versus observation order plot (Figure 5) exhibits no clear trend, with residuals scattered randomly around zero. Although some clustering is visible, such patterns are often inevitable in observational datasets. Given the absence of systematic patterns, the independence assumption is reasonably satisfied, allowing the analysis to proceed.

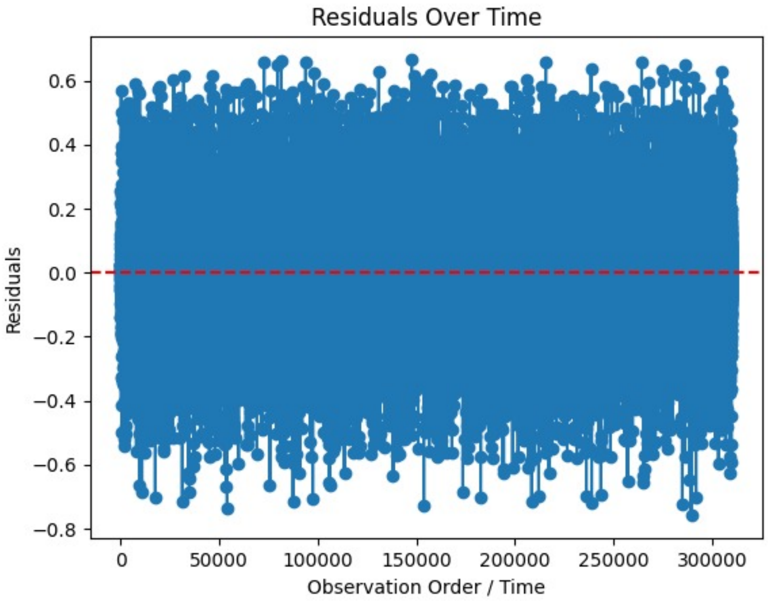


Figure 5: Residuals over time showing no systematic trend. Minor deviations are tolerated in applied settings.

0.9 Limitations and Future Scope

Limitations

- **Computational Constraints:** Due to the large size of the dataset and limited system resources, certain diagnostic tools such as DFBETAS and DFFITS could not be applied for outlier influence analysis.
- **Linearity Assessment:** Partial regression plots, which help assess the linearity of individual regressors, were not used for the same computational reasons.

Future Scope

- Future work can focus on incorporating DFBETAS, DFFITS, and partial regression plots to strengthen model diagnostics and improve interpretability.
- **Advanced Regression Methods:** Exploring regression techniques like Weighted Least Squares (WLS) may yield better results in datasets with heteroscedasticity or outliers.
- **Time Series Regression Integration:** Adapting time series regression approaches could further enhance forecasting accuracy, particularly when structural changes or temporal dependencies are present in the data.