Multiple Linear Regression and Time Series Analysis

Rebecca Santana de Pina
Msc in Data Analytics
National College of Ireland
Dublin, Ireland
x23436786@student.ncirl.ie

I. Multiple Linear Regression

A. Introduction

Multiple linear regression is a statistical method used to analyze the relationship between a dependent variable and independent variables. It is important for making predictions. But for the model to be valid, it must satisfy the Gauss-Markov assumptions, which are as follows: linearity, homoscedasticity, independence of errors, absence of multicollinearity, and normality of residuals.

This report analyses a dataset which contains 1,000 observations, using multiple linear regression to predict a continuous dependent variable (y) based on three independent variables (x1,x2,x3).

B. Explanatory Data Analysis

The dataset used contains 1,000 rows and 4 columns:

- *y*: continuous dependent variable
- *x1,x2*: continuous independent variable
- *x3*: categorical independent variable with three levels (A,B,C)

TABLE I
Descriptive Statistics for Variables

Variable	y	x1	x2	$x3_B$	$x3_C$
Count	1000	1000	1000	1000	1000
Mean	15,461.12	49.99	199.92	0.403	0.291
Std Dev	1,795.13	4.91	4.35	0.490	0.454
Min	7,972.73	33.15	186.23	0	0
25	14,434.33	47.27	197.44	0	0
Median	15,521.01	49.96	199.91	0	0
75	16,538.97	53.00	202.46	1	0
Max	21,565.78	69.42	214.46	1	1

Key observations:

- The categorical variable $x3$ was one-hot encoded into binary columns $x3_B$ and $x3_C$, with $x3_A$ as the reference category.

The correlation matrix in Fig.1 reveals strong linear relationships between y, x1 and a small relationship between y and x2. No multicollinearity concerns are observed, as correlations between predictors are low.

A histogram of y in Fig.2 indicates a normal distribution, though a few extreme values can be observed on the left and right ends of the distribution, indicating potential outliers.

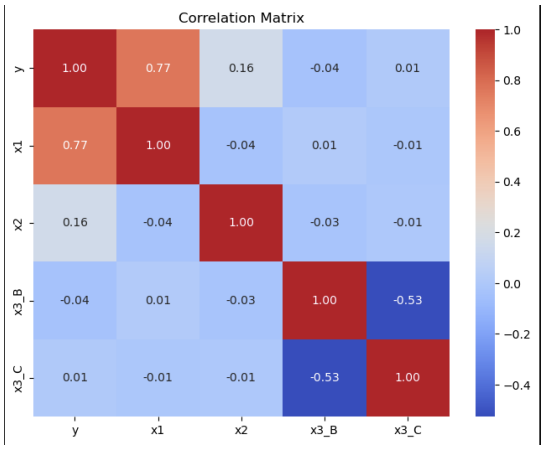

Fig. 1. Correlation Heatmap

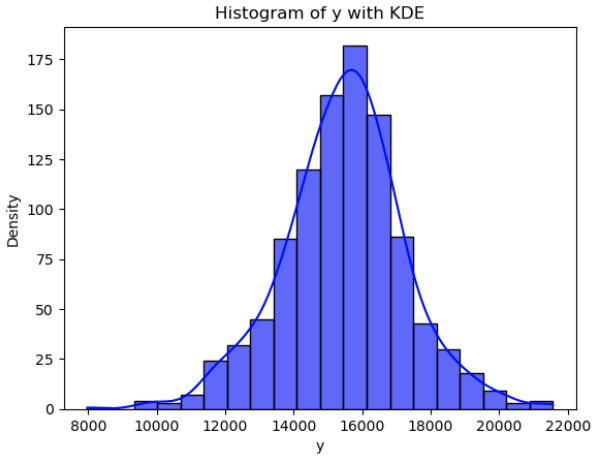

Fig. 2. Histogram of y with KDE Overlay

C. Data Preparation

1) Missing values: Upon inspection, no missing values were detected in the dataset, as shown by the isnull() function. Therefore, no values were removed or replaced.
2) Encoding Categorical Variables: The independent variable $x3$ was categorical with three levels (A,B,C). To incorporate it into the regression model, one-hot en-

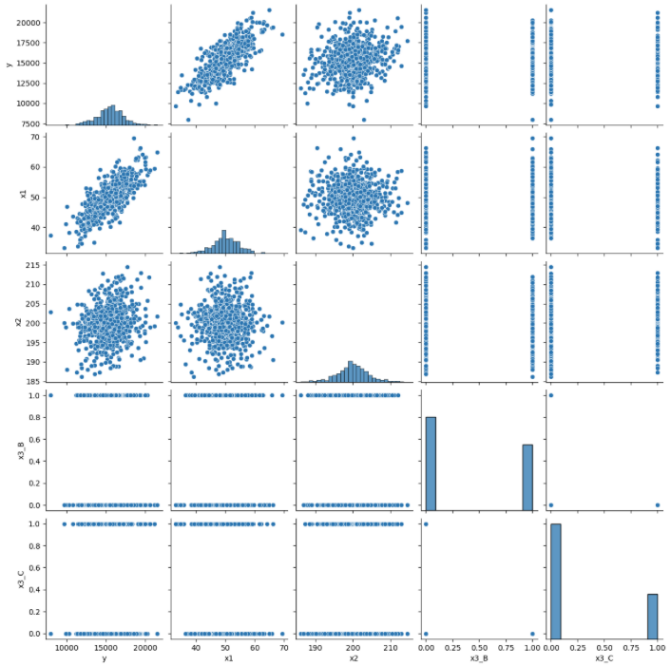

Fig. 3. Pairplot of the train data

coding was applied. Using the getdummies() method, the categorical variables was transformed into binary indicators and later for integers. This approach avoids multicollinearity issues and allows comparisons of the effects of $x3_B$ and $x3_C$ relative to $x3_A$

3) Outlier Detection: Two methods were used to detect outliers in the continuous variables (y,x1,x2):
   a) Z-score Method: Outliers were identified as data points with absolute z-score greater than 3. This method detected 16 outliers.
   b) Interquartile Range (IQR) Method: Outliers were identified as data points outside the range [Q1 - 1.5 x IQR, Q3 + 1.5 x IQR]. This method detected 68 outliers. After evaluating the impact of removing outliers on model performance, the Z-Score Method was chosen as it retained more data and improved model stability.
4) Data Splitting: The dataset was split into training (80) and test (20) subsets using the traintestsplit() function, with the student number as the random seed for reproducibility. This split ensures that the model can be trained and evaluated on separate data to assess its generalization capability.

D. Modelling

This section details the development of the multiple linear regression model using the prepared dataset.

An initial Ordinary Least Squares (OLS) regression model was built using all the variables, including outliers. The results showed:

- Significant p-values for $x1, x2$ and $x3_B$ ($p < 0.05$)
- Some predictors, such as $x3_C$, were not statistically significant ($p > 0.05$)

1) Building the Final Model: Built the OLS Model using the cleaned training dataset (Z-Score method).
   - R^2: the final model achieved an R^2 of 0.628, indicating it explained 62.8 percent of the variation in y.
   - Significant Predictors:
     - $x1$ and $x2$ were highly significant ($p < 0.001$)
     - $x3_B$ was significant ($p = 0.004$)
     - $x3_C$ was not significant ($p = 0.289$)
2) Intermediate Models and Their Rejection:
   - Baseline Model (No Cleaning):
     - R^2= 0.650, indicating that 65 percent of the variance in y was explained by the independent variables ($x1, x2, x3_B, x3_C$).
     - Residual plots indicated potential outliers.
     - Retained for comparison but rejected as the final model due to the presence of outliers
   - IQR-Cleaned Model:
     - Achieved a lower R^2 (0.576).
     - Removed a large number of observations, reducing the model's variability.
     - Rejected in favor of the Z-Score-cleaned model.

E. Interpretation

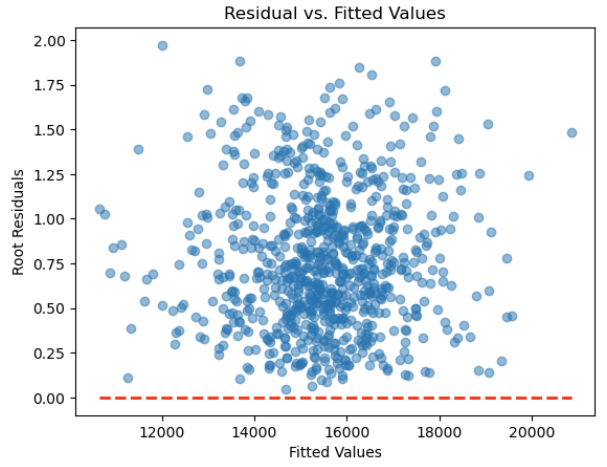

Fig. 4. Residual vs. Fitted

In the Fig. 4 the residuals appear randomly scattered around 0. There doesn't appear to be a clear systematics pattern. The variance of the residuals appears constant across the range of the fitted values, supporting the assumption of *homoscedasticity*. A few residuals deviate significantly from the rest, suggesting potential outliers.

In the Fig. 5 the residuals generally follow the red line, indicating that they are approximately normally distributed. But, there are some deviations at both ends of the plot, suggesting the presence of a few outliers.

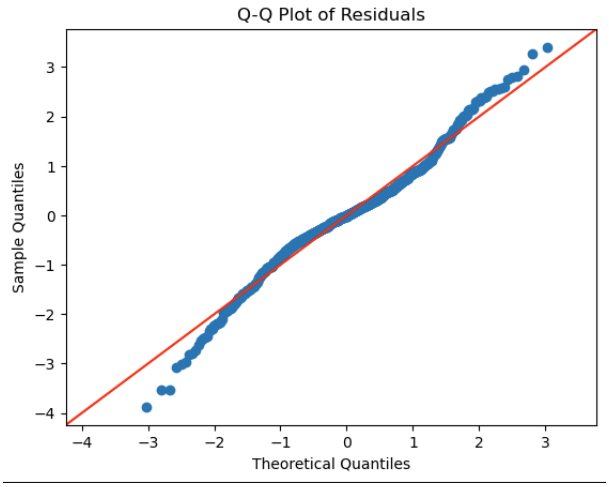

Fig. 5. Q-Q Plot of Residuals

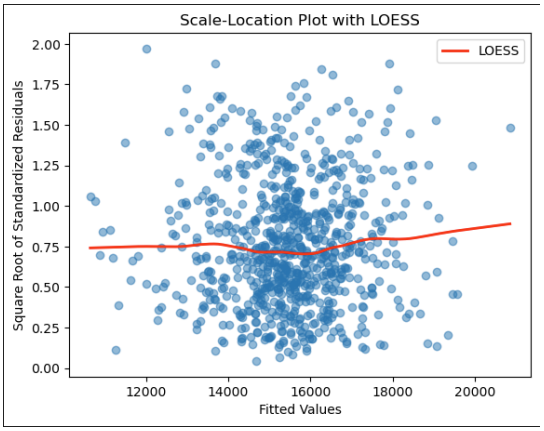

Fig. 6. Scale-Location Plot with LOESS

In the Fig. 6 the LOESS curve in the second plot is relatively flat, though it shows a slight upward trend at higher fitted values. This could suggest a increase in variance but is not strong. The plots suggest that the residuals generally satisfy the assumption of homoscedasticity.

In the Fig.7 there are several large spikes in Cook's Distance plot, indicating the presence of influential observations. More attention and futher investigation required.

F. Diagnostics

1) Homoscedasticity - The Breusch-Pagan test:
   - Lagrange Multiplier Statistic: 0.21187216164513245
   - p-value: 0.9947697531205559
   - F-statistic: 0.05265093424278067
   - F-test p-value: 0.9948187960355617

   Since the p-value is 0.9948, which is much greater than the 0.05, there is no evidence of heteroscedasticity on the model. The residuals appear to have a constant variance, satisfying the homoscedasticity assumption.
2) Multicollinearity:

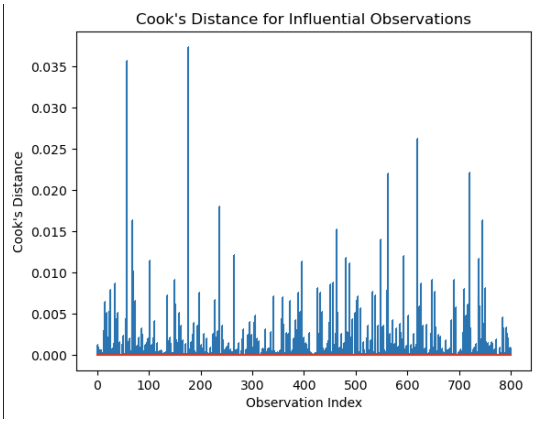

Fig. 7. Cook's Distance

TABLE II
VIF TABLE

Variable	VIF
const	2281.571599
x1	1.002543
x2	1.002694
$x3_B$	1.378945
$x3_C$	1.380044

Even that the const VIF its high, its normal because the correlation with the others variables, but doesn't affect the model's validity. All predictors have VIF values below 5, indicating no multicollinearity concerns.

3) Condition Number: 1.7934583745508974. The condition number can indicate the presence of multicollinearity, in this case, its below 10, so suggest that multicollinearity its not a significant problem and the model is stable.
4) Shapiro-Wilk: The test statistic is 0.978. The p-value is 0.000. The W-statistic (0.978) is close to 1, which suggests that the residuals are approximately normally distributed. A p-value of 0.000 indicates that you reject the null hypothesis. This means that the residuals of the model do not follow a normal distribution.

G. Evaluation

1) Performance Metrics
   - Mean Absolute Error (MAE):
     - measures the average magnitude of the errors between actual and predicted values
     - lower MAE indicates better predictive performance.

   The MAE for the model is 911.5279174892966.

The Multiple Linear Regression model demonstrates strong predictive performance on the test dataset, with low error metrics and close alignment between predicted and actual values.

The model satisfies the Gauss-Markov assumptions, validating its reliability for both inference and prediction. The

small deviations in residual normality are not expected to substantially impact the model's utility.

II. Time Series

A. Introduction

The time series is used to capture patterns in sequential data over time, considering trends, seasonality, and autocorrelation. This report analyses a dataset which contains 401 observations.

B. Explanatory Data Analysis

The dataset contains 401 rows and 2 columns:

- *Unnamed: 0*
- *x*: numerical variable

TABLE III
Descriptive Statistics for Variable

Variable	x
Count	401
Mean	215.988071
Std Dev	44.187797
Min	108.099998
25	184.309998
Median	214.240005
75	251.449997
Max	309.070007

Key observations:

- the column "unnamed: 0" was deleted, and a new index was created, with the name Date, starting in 2024-01-01.

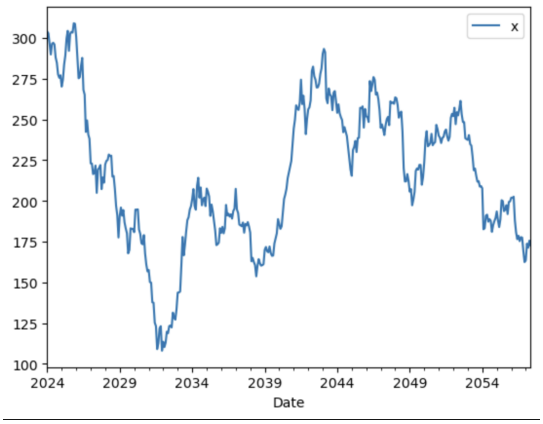

Fig. 8. Time Series Data Plot

C. Data Preparation

1) Data Spliting: the data set was split into training (80) and test (20).
2) Missing values: Upon inspection, no missing values were detected in the dataset, as show by the isnull() function.
3) Duplicated values: no duplicated values were detected in the index of the dataset.

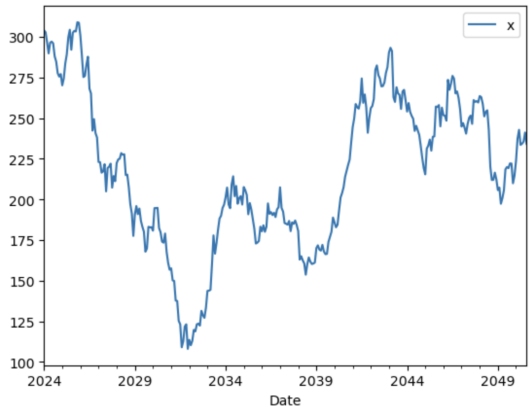

Fig. 9. Train Data Plot

D. Modelling

The train data don't look to have a pattern in the plot. So we continue with the analysis to verify stationarity.

1) Stationarity: the Augmented Dickey-Fuller(ADF) test was applied, The p-value is 0.230, failing to reject the null hypothesis (the time series is non-stationary). This confirms that the train data series is non-stationary, necessitating differentiation to achieve stationarity.
2) First Order Differencing: The ADF test reapplied to the transformed series, producing stationarity. Since the first difference of the series is stationary, d = 1.

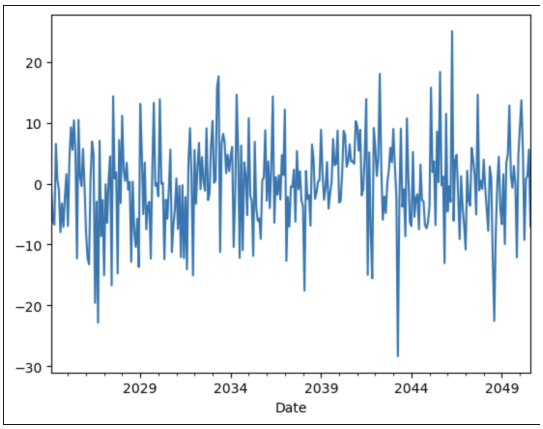

Fig. 10. First Difference Plot

The lag 0 represents the autocorrelation of the series with itself, which is always 1. There is only one significant spike at lag 4, so q=1.

The number of significant spikes in the pacf plot indicate the p. In this case is 1, so, p=1. To identify the appropriate ARIMA model parameters(p,d,q), the following steps were undertaken:

1) Differencing (d): Firs-order differencing (d=1) was applied to achieve stationarity, as verified in the stationarity subitem.
2) Autoregressive (p) and Moving Average (q): The parameters p and q were determined using Autocorrelation

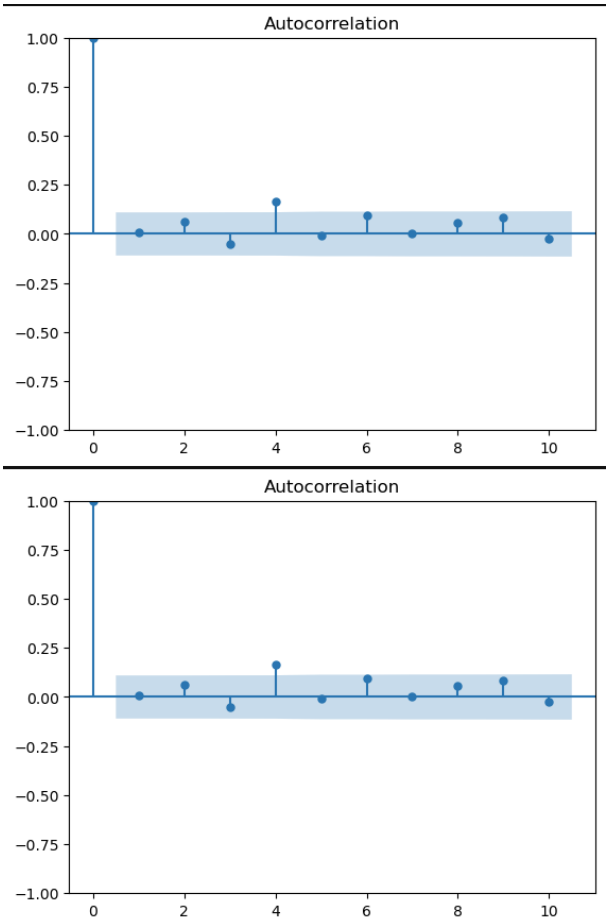

Fig. 11. ACF Plot

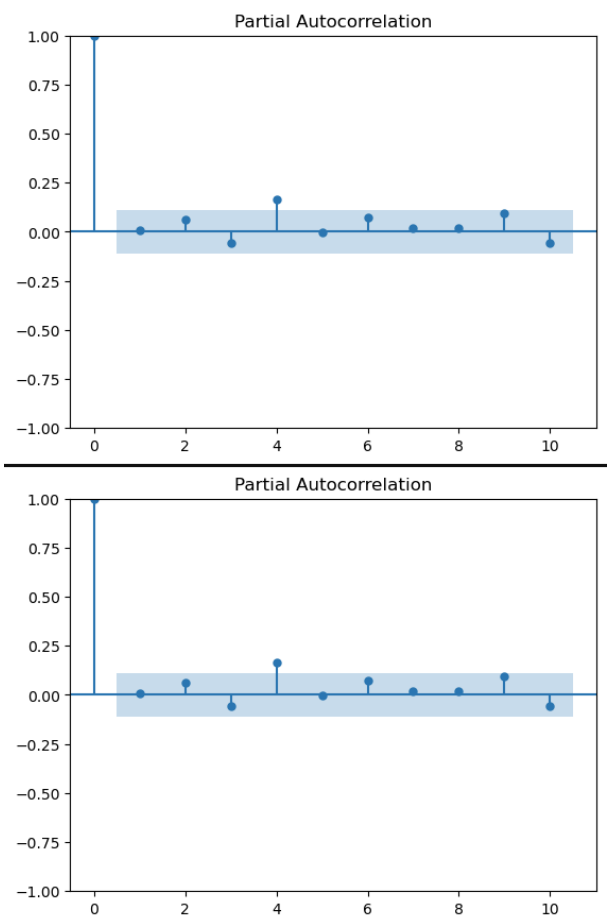

Fig. 12. PACF Plot

Function (ACF) and Partial Autocorrelation Function (PACF) plots of the differenced series:

- ACF Plot: Significant autocorrelation at lag 1 suggested q=1.
- PACF Plot: Significant partial autocorrelation at lag 4 suggested p=1.

E. Interpretation

The autoarima() function was used to automatically determine the model order. The result was that the best model fitted was ARIMA(0,1,0). But the difference was minimal, so ARIMA(1,1,1) was selected for further analysis.

The ARIMA(1,1,1) model includes three types of parameters:

- AR(Autoregressive) Coefficient (p): Measures the influence of the previous time steps (lags) on the current value. The current value is negatively influenced by 86.17 percent of its lags - 1 value. The p-value is statistically significant (p¡0.001)
- MA(Moving Average) Coefficient (q): Models the relationship between current values and past forecast erros. The current value is positively influenced by 81.17 percent of the error at lag-1. The p-value is statistically significant (p-value ¡0.001)
- $Sigma^2$(Variance of Residuals): The variance of residuals is 56.0197, a lower value generally indicates better model fit. The variance is statistically significant (p¡0.001).
- AIC (2195.510) and BIC (2206.806) were used to compare alternative ARIMA models. The selected model has low values, indicating a good model complexity and fit.

F. Diagnostics

The residuals appear to be close to white noise. The autocorrelation function(ACF) plot of the residuals Fig. 14 shows that all lags fall within the confidence intervals (blue shaded area). This indicates that the residuals are uncorrelated, behaving like the white noise.

The Ljung-Box test was applied to further confirm the absence of autocorrelation in the residuals. The test evaluates the null hypothesis (H_0) that the residuals are not autocorrelated.

- p-values for Lags 1 to 10: All p-values were greater than 0.05, confirming no significant autocorrelation in the residuals.

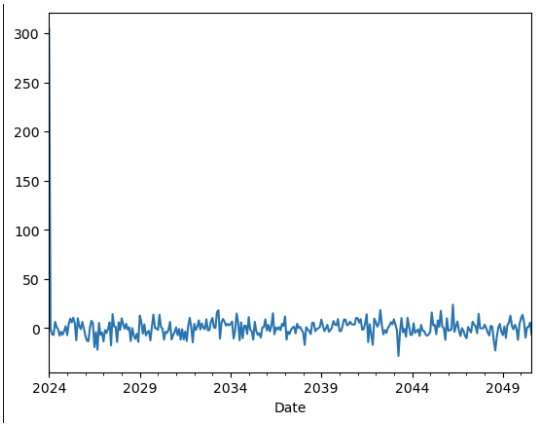

Fig. 13. Residuals Plot

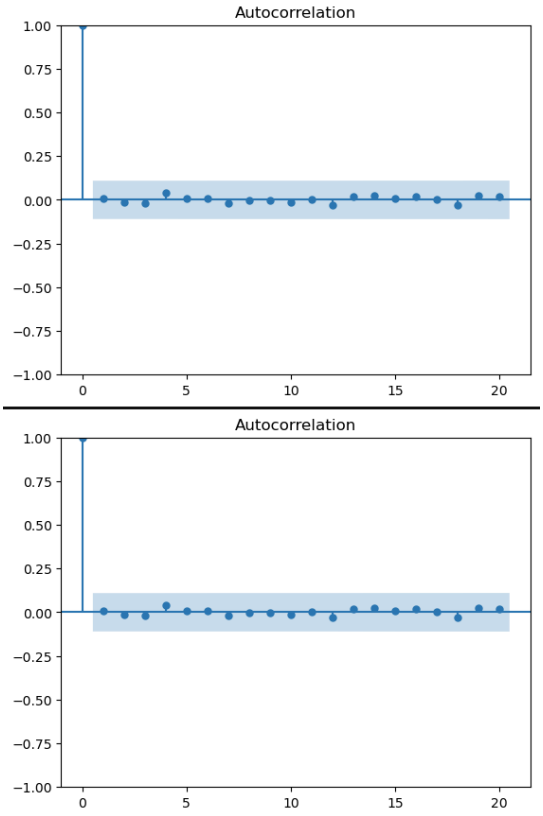

Fig. 14. Autocorrelation Residuals Plot

G. Evaluation

Producing a forecast using the fit model. 81 periods into the future (the length of the test series).

The ARIMA(1,1,1) model was used to forecast future values of the time series. The forecast results are visualized in Fig. 16, which shows the following components:

- Original Series (Black Lines): Represents the observed data from the training and testing sets.
- Fitted Values (Red Lines): Indicates the model's prediction on the training data, which align closely with the observed values.
- Forecast (Green Line): Represents the predicted values for the future time steps.

The ARIMA model produces a flat forecast. The red line (fitted values) closely tracks the black line (original series), indicating that ARIMA(1,1,1) model captures the underlying patterns in the training data effectively.

	lb_stat	lb_pvalue
1	0.024403	0.875865
2	0.092405	0.954849
3	0.226050	0.973278
4	0.746475	0.945476
5	0.769071	0.978964
6	0.790207	0.992334
7	0.883825	0.996491
8	0.887755	0.998863
9	0.888379	0.999655
10	0.941679	0.999869

Fig. 15. Ljung-Box Test

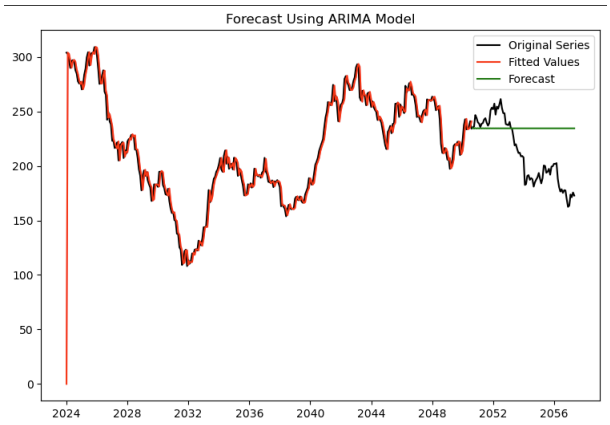

Fig. 16. Forecast using ARIMA Model

TABLE IV
MAE, RMSE, MAPE TABLE

Mean Absolute Error (MAE)	6.656
Root Mean Squared Error (RMSE)	18.564
Mean Absolute Percentage Error (MAPE)	3.1

Mean Absolute Error (MAE) - 6.656 its a relative low error, the model fit closely in the training data. Root Mean Squared Error (RMSE) - 18.564 reflects a reasonable fit the training data, even that its higher than the MAE. Mean Absolute Percentage Error (MAPE) - 0.031 indicate that the model's predictions are highly accurate.

The ARIMA(1,1,1) model effectively captures the temporal structure of the data and produces reliable forecasts.

REFERENCES

[1] Breusch, T., and Pagan, A. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. Econometrica, 47(5), 1287-1294.

[2] Durbin, J., and Watson, G. (1951). Testing for Serial Correlation in Least Squares Regression. II. Biometrika, 38(1/2), 159-177.

[3] Shapiro, S. S. and Wilk, M.B (1965). An analysis of variance test for normality (complete samples), Biometrika, Vol. 52, pp. 591-611.