National College of Ireland

National College of Ireland

Project Submission Sheet

Student Name:	Zuu Zuu Kyaw Shwe		
Student ID:	24106585		
Programme:	MSCDAD_C	**Year:**	2024
Module:	Statistics and Optimisation		
Lecturer:	John Kelly		
Submission Due Date:	2nd Dec 2024		
Project Title:	Continuous Assessment		
Word Count:	4467		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	*Zuu*
Date:	2nd Dec 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Statistics & Optimisation

Continuous Assessment

Zuu Zuu Kyaw Shwe
School of Computing (MSCDAD-C)
National College of Ireland
Dublin, Ireland
x24106585@student.ncirl.ie

***Abstract*—This project applies multiple linear regression and time series analysis to develop predictive models. The work draws attention to the importance of systematic data preparation and evaluation in the construction of the trustable models yet suggests to search for more advanced techniques for better predictive performance and forecasts.**

***Index Terms*—multiple linear regression, time series analysis**

I. Multiple Linear Regression

A. Introduction

The purpose of this analysis is to utilize multiple linear regression (MLR) in the analysis and prediction of a specific target variable that is dependent with respect to several independent variables. Multiple Linear Regression (MLR) is one of the most used statistical methods used for modelling. The outcome figures with respect to multiple factors to determine the individual's contribution and the total contributions of the model. The analysis consists of dataset exploration and preparation for modelling, detection of issues like multicollinearity and influential points. Furthermore, the analysis also intends to verify the strength of the model through appropriate statistical tests and its ability to forecast unseen data.

B. Exploratory Data Analysis (EDA)

The dataset used in this analysis contains a dependent variable (y) and three independent variables (x1, x2, and x3). Two of them, x1 and x2 are numbers while x3 is a categorical variable and has three categories that is A, B and C. From the preliminary data analysis, there were no missing values in the provided dataset.

As for numeric variables, Boxplots made it clear that there were remarkable outliers among the dependent variable (y) while the predictors (x1 and x2) did not indicate an outlier. To integrate the categorical variable (x3), one hot encoding was used and transformed it into two dummy variables $x3_B$ and $x3_C$, and $x3_A$ was made the reference category to avoid multicollinearity.

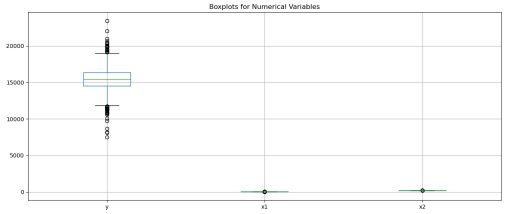

Fig. 1: Boxplots for Numerical Variables

The next step was to split the dataset into the training and testing sets bearing the ratio of 80:20 and ensuring that a random state was set to allow reproducibility of the results. The splits therefore consisted of 800 training observations and 200 testing observations with each containing four features. This configuration provides the necessary basis for further analysis and modelling work.

The exploratory data analysis (EDA) phase served to enhance the understanding of the linkage of variables, the distributions of these variables, and helped to reveal relationships and dependencies in the dataset. Some of the important conclusions based on the analyses are:

Pair Plot Analysis: The pair plot displays the pairwise relations and distributions of the predictors(x1, x2, $x3_B$, $x3_C$) and the target variable (y). Some of the insights gained are that it can be observed the relationship between y and x1 appear approximately linear, with positive trend which means when x1 increases, y also increases. However, y and x2, $x3_B$, and $x3_C$ do not appear to interact significantly.

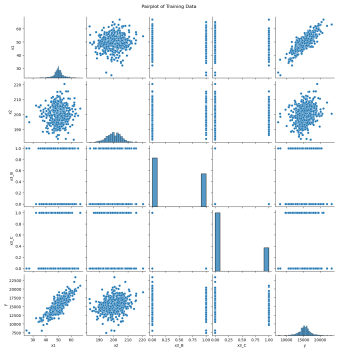

Fig. 2: Pairplot of Training Data

Correlation Heatmap: The heatmap lets us have a more integrated quantitative picture of the relationships among the numeric variables: y and x1 have a high positive correlation of (0.81) supporting the linear dependence we saw from the pair plot. Evidence suggests that x2 is only weakly correlated (0.25) with y which is consistent with the previous findings that x2 has a very limited role in the explanation of the variability of y. Also, the categorical variables ($x3_B$ and $x3_C$) are weakly related with y as well as with the other predictors.

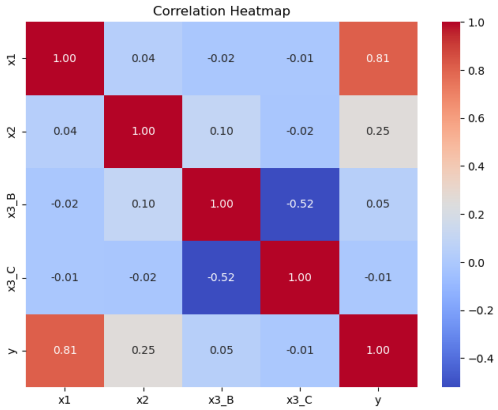

Fig. 3: Correlation Heatmap

C. Data Preparation

The data preparation phase consisted of several different steps which are aimed at cleaning the dataset and making it appropriate for regression analysis. An exploratory data analysis (EDA) phase was conducted to study relationships between the variables, their distributions as well as possible outliers. In the case of the dependent variable (y) boxplots indicated some significant presence of outliers but none for independent variables (x1, x2).

The EDA analysis also included the pairwise plots; the correlation heatmap showed a reasonable high positive correlation for y and x1. After the EDA, one of the categorical variables (x3) was converted into two dummy variables ($x3_B$ and $x3_C$), and $x3_A$ was selected as the reference category to avoid multicollinearity. Thus, eighty percent of the dataset was assigned for the training subset and twenty for the testing subset for purposes of proper model development and evaluation. These steps ensured the dataset was clean, structured, and ready for modelling.

D. Modelling

In this section, we present the steps which were undertaken to construct the multiple linear regression model with an aim of predicting the dependent variable y. The whole procedure was marked by the processes of model building, dealing with identified high leverage points, and subsequently, a comprehensive model check was done after fitting the model to confirm the statistical validity of the model.

1) Initial Model Development: The first model was developed based on all the independent variables which were available in the dataset: dependent variable (y) and independent variables (x1, x2, $x3_B$ and $x3_C$). A regression model based on Ordinary Least Squares (OLS) approach was used for fitting the model, after which R-squared, adjusted R-squared, and p-values were analyzed to evaluate the reliability of the model. Also, the model's diagnostic tests and residuals and diagnostic plots were used to assess the Gauss-Markov assumptions.

2) Identification and Handling of High-Influence Points: Cook's Distance and leverage diagnostics were employed to help identify high-influence observations which is beyond a value of 4/n, n being the number of observations in the dataset, were marked as high leverage points. In doing this, training observations such as high leverage points were dropped from the training dataset, and the model was refitted.

3) Refitted Model Development: A new model was built after removing the identified high-influence points in the previous step. The diagnostic process was repeated with the refitted model in order to test whether the last step has indeed led to an improvement. When comparing the R-squared and adjusted R-squared values of the fitted model, a slight increase in these values can be seen in comparison to the initial model. Nevertheless, diagnostic measures of the residual vs fitted plots, QQ plot and scale-location plot did not show major improvements with the refitted model.

4) Intermediate Models and Rejection Rationale: Residuals vs. Fitted plot and Breusch-Pagan test (which evaluate linearity and homoscedasticity assumptions) demonstrated that the specified relationships formulated by the model were acceptable. However, the Shapiro-Wilk test confirmed a violation of the normality assumption for residuals. Even so attempts were made to use the Box-Cox transformation or a logarithm of the remaining variables, these two presented only slight improvement for this variable's dimensionality.

Several models were tested for optimal performance, with the initial model showing favorable diagnostic results. The refitted model, which removed high influential points, did not improve diagnostics significantly. The model scope was deemed inappropriate, and data points were excluded. The initial model was found to be low complexity, making it easier to interpret and deploy compared to alternative models.

5) Final Model Selection: Although to a certain degree, the model suffers from overfitting, it still can be considered useful in predicting the dependent variable since the training R-squared is 70.9% which is a decent fit for seen data. Within the attempts to minimize overfitting, the deletion of high influencers or the application of transformations did not yield encouraging results in terms of the predictive models or their diagnostic measures. Furthermore, other models didn't show any meaningful benefits with respect to the accuracy and robustness of the prediction.

E. Interpretation of the Final Model

The interpretation of the final model involves examining the coefficients, p-values, and confidence intervals for each

predictor, along with the p-value of the F-statistic to assess the overall significance of the model.

- Intercept (Constant): -14,470 (Baseline value when all predictors are zero).
- x1: Coefficient = 313.66. Keeping all other factors equal, a one-unit increase in x1 corresponds to a 313.66 increase in y. (Highly significant, p<0.001).
- x2: Coefficient = 70.57. Keeping all other factors equal, a one-unit increase in x2 corresponds to a 70.57 increase in y. (Highly significant, p<0.001).
- $x3_B$: Coefficient = 195.78. Being in category B is associated with a 195.78 increase in y, compared to category A. (Statistically significant, p=0.013).
- $x3_C$: Coefficient = 142.31. Being in category C is associated with a 142.31 increase in y, compared to category A. (Not significant, p=0.094).
- R-squared: 0.709. About 70.9% of the variance in y is explained by the predictors.
- Adjusted R-squared: 0.707. Adjusted for the number of predictors, maintaining a strong fit.
- Prob (F-statistic): 4.09 x 10^{-211}. This infinitesimally small value indicates that the regression model as a whole is statistically significant and provides a much better fit than a model with no predictors.
- Confidence Intervals: The confidence intervals for x1,x2, and $x3_B$ do not contain zero, supporting the reliability of the estimates and their statistical significance while $x3_C$ contain zero, meaning it is not statistically significant at the 5% level.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.709
Model:	OLS	Adj. R-squared:	0.707
Method:	Least Squares	F-statistic:	483.2
Date:	Fri, 29 Nov 2024	Prob (F-statistic):	4.09e-211
Time:	21:53:10	Log-Likelihood:	-6597.9
No. Observations:	800	AIC:	1.321e+04
Df Residuals:	795	BIC:	1.323e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.447e+04	1315.817	-10.998	0.000	-1.71e+04	-1.19e+04
x1	313.6630	7.479	41.942	0.000	298.983	328.343
x2	70.5682	6.401	11.025	0.000	58.004	83.133
x3_B	195.7821	78.795	2.485	0.013	41.111	350.453
x3_C	142.3086	84.935	1.676	0.094	-24.414	309.031

Omnibus:	43.069	Durbin-Watson:	2.037
Prob(Omnibus):	0.000	Jarque-Bera (JB):	135.153
Skew:	-0.143	Prob(JB):	4.49e-30
Kurtosis:	4.993	Cond. No.	8.29e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.29e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Fig. 4: OLS Regression Results

F. Diagnostics

A wide range of diagnostic and visual plots were used to check that the constructed multiple linear regression model met Gauss-Markov assumptions. For the OLS regression to be valid; these assumptions including linearity, autocorrelations, homoscedasticity, residual normality, and no multicollinearity have to be met conditions.

1) Linearity: The scatter plot of residuals vs. fitted values was examined. The residuals appear to be somewhat randomly distributed around the zero-neck line. It follows that the model seems capable of incorporating the linearity assumption. While the variance of the residuals is constant for the most part, they appear to become somewhat more spread out at higher fitted values. This would indicate a mild form of heteroscedasticity in other words the variability of the residuals increases with the fitted values. To conclude, while the model appears to be consistent with the homoscedasticity assumption, some heteroscedasticity would be appreciable towards extreme fitted values.

2) Autocorrelation: The Durbin-Watson test was used to check for autocorrelation in the residuals. From the OLS regression results table, the Durbin-Watson statistic was close to 2 indicating no significant autocorrelation among the residuals, thus supporting the independence assumption.

3) Heteroskedasticity: A Scale-Location plot and The Breusch-Pagan test was conducted to check for homoscedasticity in the model. The points are randomly scattered around, indicating no clear pattern or trend. No funneling patterns were observed also. The p-value of the Breusch-Pagan test (0.281) is greater than 0.05, failing to reject the null hypothesis of the residuals have constant variance (homoscedasticity). The residuals do not show significant heteroscedasticity. The assumption of homoscedasticity is satisfied.

```
bp_test = het_breuschpagan(residuals, X_train_const)
bp_stat, bp_pvalue = bp_test[0], bp_test[1]
print(f"Breusch-Pagan Test Statistic: {bp_stat}, P-value: {bp_pvalue}")
```

```
Breusch-Pagan Test Statistic: 5.056405673390074, P-value: 0.2815587339495312
```

Fig. 5: The Breusch-Pagan Test

4) Normality of Residuals: A Q-Q plot was used to assess the normality of residuals. Most of the points lie close to the red diagonal line, indicating that the residuals are approximately normally distributed. It is also evident from the graph that at both the lower tail (left) and the upper tail (right), there are departures from the diagonal line. This indicates that the residuals could be having more weight towards the tail than that of normal distribution i.e., there can be outliers or extreme values which may not be normally distributed. Additionally, the Shapiro-Wilk test was performed and the value was much lower than 0.05, rejecting the null hypothesis of normality for residuals.

```
shapiro_test_stat, shapiro_p_value = shapiro(residuals)
residual_diagnostics = {
    "Shapiro-Wilk Test Statistic": shapiro_test_stat,
    "Shapiro-Wilk P-Value": shapiro_p_value
}
residual_diagnostics
```

```
{'Shapiro-Wilk Test Statistic': 0.9715442495100705,
 'Shapiro-Wilk P-Value': 2.2538133528670752e-11}
```

Fig. 6: The Shapiro-Wilk test

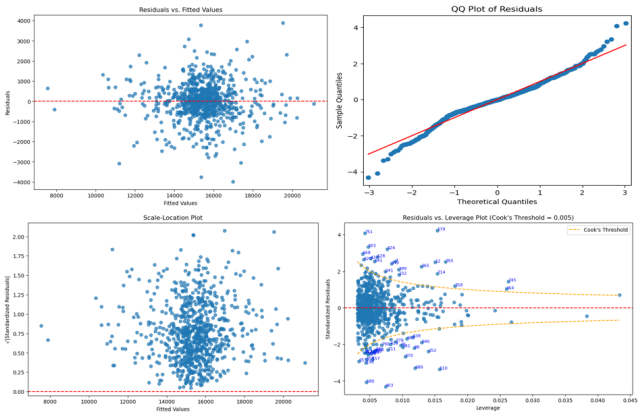

Fig. 7: Regression Diagnostics Plots

5) No Multicollinearity: The Variance Inflation Factors (VIF) were calculated for each predictor. None of the predictors exhibit multicollinearity, as all VIF values are well below the threshold of 5 or 10.

```
vif_data = pd.DataFrame()
vif_data["Predictor"] = X_train_const.columns
vif_data["VIF"] = [variance_inflation_factor(X_train_const.values, i)
                   for i in range(X_train_const.shape[1])]
vif_data
```

	Predictor	VIF
0	const	1612.869340
1	x1	1.002876
2	x2	1.014272
3	x3_B	1.390941
4	x3_C	1.376579

Fig. 8: Variance Inflation Factor (VIF)

6) High Influence Points: Cook's distance and leverage plots were evaluated to identify high-influence points. The revised model identified and eliminated observations when Cook's distance was more than the threshold. However, the refitted model was rejected due to no significant diagnostic improvements and underperformed on the test set than the initial model.

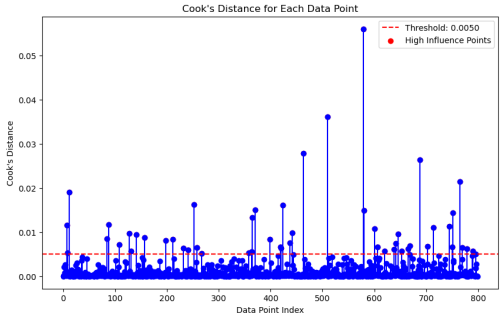

Fig. 9: Cook's Distance for Each Data Point

Even though model possesses important Gauss-Markov assumptions including linearity, homoscedasticity, autocorrelation, however, the Shapiro-Wilk test indicated that residuals do not follow a normal distribution. This might compromise the validity of tests for the coefficient hypotheses. Future models may incorporate alternative modelling strategies in order to overcome this limitation.

G. Evaluation

In the final stage, prediction capabilities of the developed model were tested using a dataset that encompassed 20% of the original data. The performance measures were computed so as to assess the accuracy of the model as well as its general reliability. In this case, the test data set was used to make further predictions about the value of the unilateral variable.

For the evaluation, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (Test) was calculated. From the calculated results, the model exhibits an average level of predictiveness on the test set as seen from the R squared value (0.57), explaining 57% of the variance. The prediction error measures (MAE and RMSE) are relatively high and the drop from the training R-squared (70.9%) suggests overfitting.

```
# Add a constant to the test predictors
X_test_const = sm.add_constant(X_test)

# Generate predictions for the test set using the rinitial model
y_test_pred = mlr_model.predict(X_test_const)

# Calculate evaluation metrics
mae = mean_absolute_error(y_test, y_test_pred)
mse = mean_squared_error(y_test, y_test_pred)
rmse = np.sqrt(mse)
r2_test = r2_score(y_test, y_test_pred)

# Compile the metrics
test_evaluation_metrics = {
    "Mean Absolute Error (MAE)": mae,
    "Mean Squared Error (MSE)": mse,
    "Root Mean Squared Error (RMSE)": rmse,
    "R-squared (Test)": r2_test
}
test_evaluation_metrics
```

```
{'Mean Absolute Error (MAE)': 745.9702409545089,
 'Mean Squared Error (MSE)': 1074709.204139512,
 'Root Mean Squared Error (RMSE)': 1036.681823965055,
 'R-squared (Test)': 0.5748723485510679}
```

Fig. 10: Test Performance Metrics

H. Conclusions and Recommendations

The multiple linear regression model seems to have an adequate predictive capability. The essential assumptions such as linearity, homoscedasticity and autocorrelation were also all satisfied. However, the normality of residuals assumption was not satisfied and the residuals were not normally distributed given the results of the Shapiro-Wilk test. Also, it was evident that the model was overfitting due to the R-squared statistics where the training set registered about 0.70 and only 0.57 for the test set. Other transformations (Box-Cox, logarithmic) that were included in the models to alleviate the situation did not have significant boosting on the performance of the models and hence were not incorporated. The inclusion of non-significant predictor $x3_C$ may explain why the model performs so poorly. The model can be simplified by removing these predictors and using the non-linear relationship, or interaction terms could be effective in identifying the pattern in the data.

The chosen model has its drawbacks but is worthy of use as a predictive baseline for the dependent variable. In this regard, future work could implement more sophisticated methods such as machine learning models, including decision trees, or the comprehensive use of several models, which would allow increasing the predictive ability of the model in cases of more complicated dependencies. Additionally, bootstrapping techniques might be useful to deliver confidence intervals and p-values which can withstand violations of normality. These improvements would make the model more comprehensible and robust for the out-of-sample application.

II. Time Series Analysis

A. Introduction

This analysis focuses on determining and fitting the most suitable time series model onto one specific dataset. As far as the time series data is concerned it involves the identification of the recurrent structures of data with respect to time which are trends along with the seasonal deviations and their application in prediction of the future. Two strategies have been employed in the course of the report:

- A non-seasonal ARIMA model built using the auto_arima function, which pinpointed the ARIMA(0,1,0) model as the optimal choice for the dataset.
- A seasonal (1,1,0,12) SARIMA model, created to accommodate seasonal variations apparent from the series decomposition.

The analysis looks at the two most models in considerable detail and it comes out clearly that the simplicity of ARIMA(0,1,0) model works best for this dataset.

B. Exploratory Data Analysis (EDA)

The dataset has 401 observations, the mean value is 250.40, the standard deviation is 71.71, the minimum value is 108.10, the maximum value is 409.97. The median value is equal to 241.87 which suggests that the data might be slight right-skew. In terms of the procedure to prepare the dataset for modelling, it was divided into two parts; the training set which contained 80% of the data and the testing set which made up 20% of the total data.

In order to gain insight into the underlying dynamic nature of the time series, the time series was decomposed into three components: trend, seasonal, and residual. The trend component showed first a decline which was later followed by some recovery and the seasonal one was periodic in nature with frequency of 12 periods. However, further analysis for seasonality conducted with auto_arima did not reveal significant seasonality.

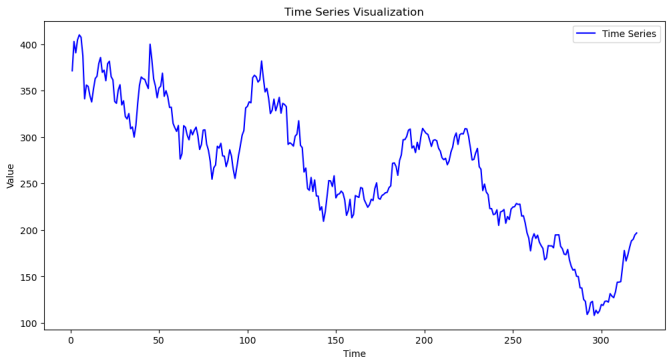

Fig. 11: Time Series Plot

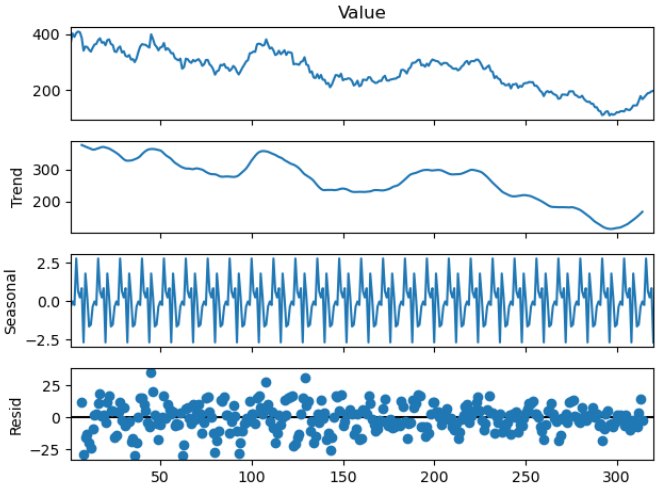

Fig. 12: Additive Model Decomposition Plot

Stationarity was assessed on the training dataset using the Augmented Dickey-Fuller (ADF) test. The first test illustrates that this variable was non-stationary since the p-value was equal to 0.468 which means that the null hypothesis of a unit root was not rejected. This was tackled through first order differencing which was able to successfully convert the series to the stationary form and this was validated again by the ADF test.

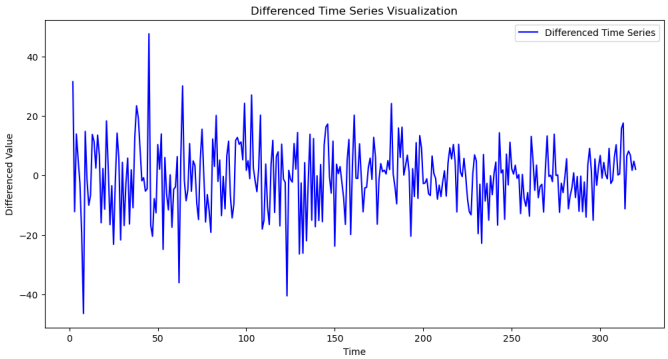

Fig. 13: Differenced Time Series Plot

To potentially identify the parameters of the model, the autocorrelation and the partial autocorrelation functions (ACF and PACF) were plotted for the differenced series. The PACF and ACF plots revealed that there was significant autocorrelation at initial lags and thus AR and MA terms are needed to model this series. These findings were instructive in the ARIMA and SARIMA model parameter selection during the subsequent modeling.

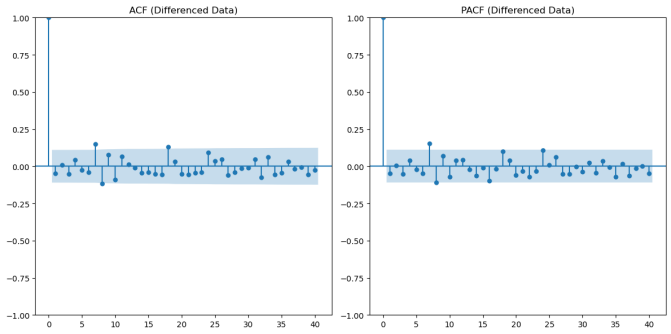

Fig. 14: ACF & PACF Plots

C. Data Preparation

The dataset was prepared in a way that would make it suitable for modelling. Once the data had been loaded, it was found that there were no missing values and only two rows that were deemed to be duplicates and were kept due to their minimal effect. Then, the data was then partitioned into two parts for training and testing model.

The ADF test was used to evaluate stationarity, which is an important prerequisite for performing time-series modelling. The first test performed indicated a non-stationary condition (p-value=0.468) and this was rectified by first order differencing and reconfirmed by time series analysis with a second ADF test.

To aggregate various knowledge about the structure of the data, the series was decomposed into trend, seasonal and residual components. A downward trend with a slight increase and seasonal with periodicity of 12 was also noticed. Later, auto_arima function suggest that the seasonality was not statistically significant though. The differenced series of ACF and PACF plots gave an significant correlations at initial lags, hence modelling will require the integration of AR and MA components.

D. Modelling

In the course of the modelling activities, several models were developed and appraised with an aim of coming up with the best model for forecasting. Outcomes achieved include providing justifications on the model selected, coverage of the data, and discarding the intermediate models.

1) Initial ARIMA Model: The modelling process started with auto_arima which proposed ARIMA(0,1,0) model as optimal. This is a basic model that lacks an AR and MA term and simply views the series as a slightly fluctuating random walk.

2) Exploration of SARIMA Models: A seasonal ARIMA model, SARIMA(0,1,0)(1,1,0,12), was developed to address the periodicity that could exist given the seasonality observed in the decomposition process. This model included one seasonal AR term of period 12. There was no significant autocorrelation in the residuals, confirming that this model provided a good fit to the data in terms of capturing both trends and seasonality. The SARIMA model was however later rejected because in the test datasets its performance was poor, registering considerably high RMSE and MAE than that of ARIMA(0,1,0).

3) Intermediate Models: The ACF and PACF analysis indicated some intermediate models with considerations to ARIMA contains (1,1,1), (1,1,2), and (2,1,1) models. The only drawback is that these models do not improve the AIC value over the ARIMA(0,1,0) model. Due to their high complexity combined with lack of focus gain so these were also discarded later.

4) Missing Data and Transformations: There were no such missing values in the data set and there were none of the outliers which were needed to be removed. The first order differencing was used to remove the trend and static was achieved which was subsequently validated using ADF test.

5) Final Model Selection: The final model chosen was the ARIMA(0,1,0) model, which is quite straightforward and produced superior performance and residual diagnostics on test data. The primary focus was to capture seasonal patterns through the SARIMA model, but the model was less effective due to statistically insignificant seasonalities.

E. Interpretation

In the ARIMA(0,1,0) method selected, emphasis is placed on the model parameters and analysis of the residuals. The key insights from the model are as follows:

1) Model Parameters: The ARIMA(0,1,0) model does not contain any autoregressive (p) or moving average (q) components and applies first order (sigma) of 131.22 was also observed from the model in estimating the variance of residuals. The z-score for the sigma2 coefficient is 17.113, which is very large, and the corresponding p-value is 0.000, which is extremely small. This indicates that the estimated variance of the error term (sigma2) is significantly different from zero, meaning the model does well in terms of capturing the variations within the series.

2) Statistical Significance: The model predictions of ARIMA (0,1,0) were of a simple nature and did not accommodate AR or MA coefficients, thus making their significance testing irrelevant. Mostly the evaluation of the model performance is through its residuals diagnostics and its prediction accuracy.

```
# Fit an ARIMA(0,1,0) model
arima_model_0_1_0 = ARIMA(train_data['Value'], order=(0, 1, 0))
arima_result_0_1_0 = arima_model_0_1_0.fit()

arima_summary_0_1_0 = arima_result_0_1_0.summary()
arima_summary_0_1_0
```

SARIMAX Results

Dep. Variable:	Value	**No. Observations:**	320
Model:	ARIMA(0, 1, 0)	**Log Likelihood**	-1230.493
Date:	Fri, 29 Nov 2024	**AIC**	2462.986
Time:	23:46:53	**BIC**	2466.751
Sample:	0	**HQIC**	2464.489
	- 320		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	131.2170	7.668	17.113	0.000	116.189	146.245

Ljung-Box (L1) (Q):	0.71	**Jarque-Bera (JB):**	37.51
Prob(Q):	0.40	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.34	**Skew:**	-0.16
Prob(H) (two-sided):	0.00	**Kurtosis:**	4.65

Fig. 15: ARIMA Model Results

3) Residual Diagnostics: The residuals are independent (no significant autocorrelation), zero-centered, and these properties of the residuals were shown to be the appropriate characteristics of an estimated ARIMA model. The Q-Q plot and Jarque-Bera test showed some mild degrees of normality.

4) SARIMA Model Interpretation: To some extent, in comparison, the SARIMA(0,1,0)(1,1,0,12) model also had a seasonal AR term with a coefficient of -0.5391. This was statistically significant ($p < 0.01$), suggesting cyclical behavior with a periodicity of 12. Nevertheless, the test results of this model against the test data set did not impress since high RMSE and MAE measures were registered. This implies that the detected seasonal structure did not enhance the forecast performance.

```
# Fit an SARIMA(0,1,0)(1,1,0,12) model
sarima_model = ARIMA(train_data['Value'], order=(0, 1, 0), seasonal_order=(1, 1, 0, 12))
sarima_result = sarima_model.fit()

sarima_summary = sarima_result.summary()
sarima_summary
```

SARIMAX Results

Dep. Variable:	Value	**No. Observations:**	320
Model:	ARIMA(0, 1, 0)x(1, 1, 0, 12)	**Log Likelihood**	-1231.766
Date:	Fri, 29 Nov 2024	**AIC**	2467.533
Time:	23:47:47	**BIC**	2474.987
Sample:	0	**HQIC**	2470.514
	- 320		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.S.L12	-0.5391	0.044	-12.207	0.000	-0.626	-0.453
sigma2	176.4801	11.273	15.655	0.000	154.386	198.575

Ljung-Box (L1) (Q):	1.01	**Jarque-Bera (JB):**	43.67
Prob(Q):	0.31	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.39	**Skew:**	-0.55
Prob(H) (two-sided):	0.00	**Kurtosis:**	4.49

Fig. 16: SARIMA Model Results

F. Diagnostics

To study how far the tested model ARIMA(0,1,0) parts from the main three features of any time series including independence, zero-mean residuals, and normality, the diagnostics of model parameters was performed. The results are the following:

1) Residual Independence: The residual independence process was evaluated using the Autocorrelation Function (ACF) and the Ljung-Box test. The autocorrelation of residuals ACF plot has zero residual autocorrelation at all lag duration which proves residuals being self dependent which was tested by using ACF plot. Ljung-Box on the contrary sustained last bullet but also said that this is partly the case with high value of (p=0.94) providing evidence for no large presence of long run dependencies overall. This assures that dependencies in data were captured well by the ARIMA model.

2) Zero-Mean Residuals: Looking at the time series residuals plot and say that the residuals are concentrated around zero and look random in terms of no systematic patterns. This explains why the assumption that mean of model errors does not stray from zero in average is easily fulfilled.

3) Normality of Residuals: The normality of residuals was evaluated utilising:

- Histogram: The picture which the histogram paints is that of distribution that looked like a normal distribution but it was skewed quite a little.
- Q-Q Plot: The tail indicates as there was some deviation from the straight line that showed normality of the rest of the graph which appeared more or less normal.
- Jarque-Bera Test: The test rejected the null hypothesis of normality ($p<0.01$), suggesting that the residuals are not perfectly normal.

4) Homoscedasticity: In the absence of obvious heteroscedasticity, the residuals looked visually quite appropriate. The distribution of residuals continued to be the same over the periods in question, hence the assumption of constant variance is met.

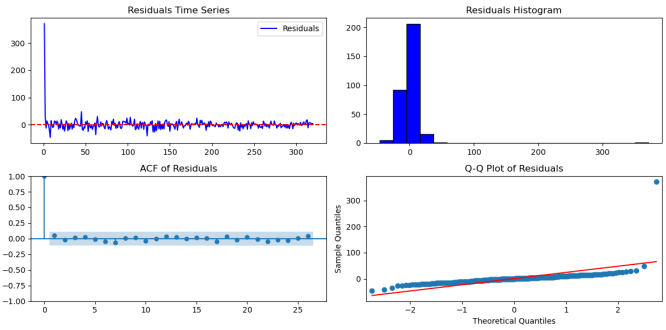

Fig. 17: Diagnostics Plots of ARIMA Model

G. Evaluation

The predictive ability employing the ARIMA(0,1,0) model was tested through the test dataset which was made up of twenty percent of the original observations. The metrics recorded were: Root Mean Square Error (RMSE): 18.86, and Mean Absolute Error (MAE): 15.21 which have not demonstrated a high degree of forecasts errors indicating the ability of ARIMA(0,1,0) model in forecasting values in the future.

```
# Combine results for comparison
comparison_metrics = {"ARIMA (0,1,0)": arima_metrics, "SARIMA (0,1,0)(1,1,0,12)": sarima_metrics}
comparison_metrics

{'ARIMA (0,1,0)': {'ARIMA RMSE': 18.860270674117732,
  'ARIMA MAE': 15.208888677903301},
 'SARIMA (0,1,0)(1,1,0,12)': {'SARIMA RMSE': 208.4712596505594,
  'SARIMA MAE': 180.27971935474451}}
```

Fig. 18: Comparative Metrics Between ARIMA & SARIMA

ARIMA (0,1,0) model makes prediction through use of the random walk with drift theory. Although, it is good in terms of the level of the data, it cannot however forecast well, hence the flat forecast depicted by the red dashed line that does not match up with the upward trend and the fluctuations that are evident in the actual test data which is green in color. Such observations indicate that such model is not able to account for the more complicated dynamics or trends that could be embedded in the series. This model is simple, and therefore may not be able to spot short term wiggles but rather describe the overall direction of the series.

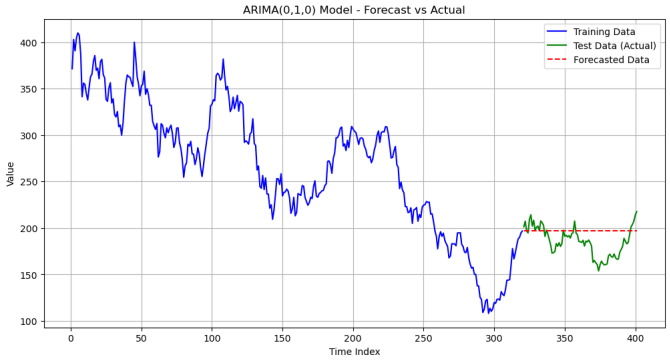

Fig. 19: ARIMA(0,1,0) Model - Forecast vs Actual

Apart from that, a model of kind SARIMA(0,1,0)(1,1,0,12) has been tested due to relevance but, unfortunately it scored significant higher number in RMSE: 208.47, and MAE: 180.28. The higher error metrics and poor alignment with the test data underscore the ineffectiveness of the SARIMA model, likely due to the absence of strong seasonality in the dataset.

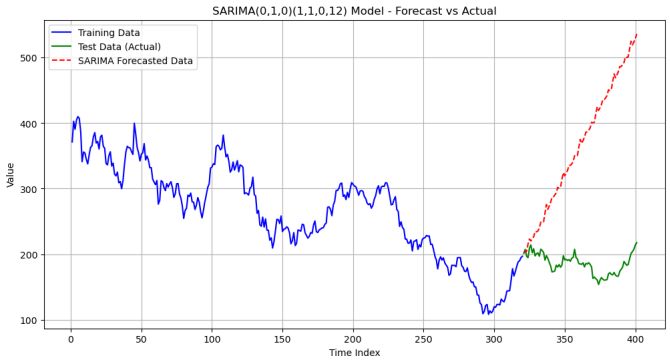

Fig. 20: SARIMA(0,1,0)(1,1,0,12) Model - Forecast vs Actual

H. Conclusions and Recommendations

The ARIMA (0, 1, 0) model performs the best in forecasting among the models analyzed, as evidenced by the test data set's forecasting errors. It did note the presence of a decomposition with a periodicity of 12, but statistical tests and the diagnostic evaluation of the SARIMA model indicated that seasonality is weak in this study which favored the application of ARIMA(0,1,0) model. No noticeable improvement was observed in the AIC values or in the accuracy of the forecast with the SARIMA model. Taking into consideration all aspects of the model, including its simplicity, interpretability and performance, ARIMA(0,1,0) was the best fitted model for this time series.

The accuracy of forecasting may be further improved by testing more advanced modeling techniques, such as ARIMAX, or machine learning based approaches like LSTM or Prophet, to account for nonlinear patterns and short term oscillations more appropriately. Expanding the dataset would also benefit the model by making its identification of trends more accurate and its forecast more precise.

Acknowledgment

I would like to thank my lecturer at the National College of Ireland for his help and support with this project assessment. An excellent basis for applying statistical and analytical ideas to practical issues was established by the lessons and resources.

List of Figures