National College of Ireland

National College of Ireland

Project Submission Sheet

Student Name:	Shivansh Bhatnagar		
Student ID:	X23237252		
Programme:	Msc in Data Analytics (MSCDAD_A)	**Year:**	Sept 2024-25
Module:	Statistics and Optimization		
Lecturer:	Hicham Rifai		
Submission Due Date:	2nd Dec 2024		
Project Title:	Project Report on Multiple Linear Regression and Time Series Analysis		
Word Count:	2832		

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: Shivansh Bhatnagar

Date: 2nd Dec .

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):	

AI Acknowledgement Supplement

Statistics & Optimisation

Project Report on Multiple Linear Regression & Time Series Analysis

Your Name/Student Number	Course	Date
Shivansh Bhatnagar/x23237252	Msc in Data Analytics	2nd Dec 2024

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool
Chatgpt	Took reference to divide the sections according to given format for report.	https://chatgpt.com

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

Chat Gpt	
Reference to divide sections according to report.	
divide my multiple linear regression to explain for report according to below report format.I have performed EDA, Data cleaning, Data Preprocessing , Visualize variables,outliar handling,correlation matrix,Model fitting,lasso, ridge, linear regression model,ols summary, r2, mse,mae using a test data set	**Introduction: Briefly state the analysis objective to predict the dependent variable using independent variables and understand their relationships.** **EDA: Present descriptive statistics, correlation matrix, and visualizations (e.g., scatterplots, boxplots) to explore variable patterns and detect outliers.** **Data Preparation: Explain data cleaning, handling missing values, outlier treatment, feature selection, and any transformations applied for model readiness.** **Modelling: Describe building multiple linear regression models, using Lasso and Ridge for regularization, and finalizing the model with OLS summary (interpreting coefficients, p-values, R^2).**

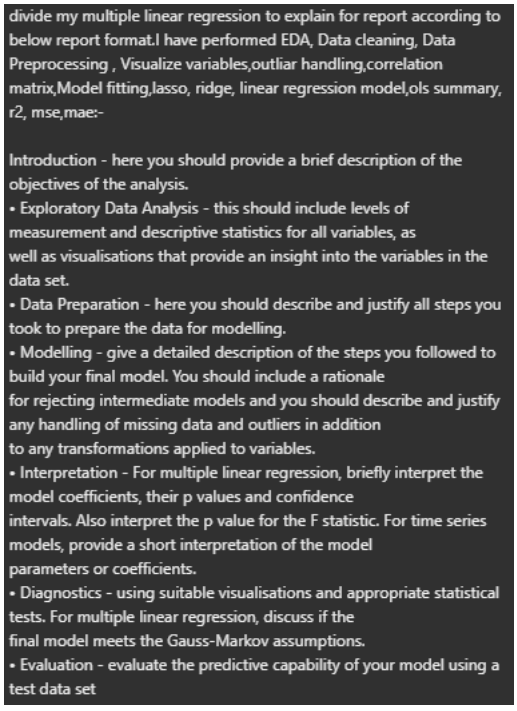
	Diagnostics & Evaluation: Assess Gauss-Markov assumptions, validate model performance with R^2, MSE, and MAE, and evaluate predictive capability using test data.

Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:

Below is my prompt:-



Additional Evidence:

Below is output:-

Data Preparation

Detail your preprocessing steps:

1. **Data Cleaning:**
   - Handled missing values (e.g., using mean/median imputation or dropping).
   - Dealt with duplicates, if any.
2. **Outlier Handling:**
   - Techniques applied: Z-scores, IQR method, etc.
   - Justification for their removal or retention.
3. **Transformations:**
   - Log transformations for skewed data.
   - Scaling or standardization (if required).
4. **Feature Selection:**
   - Criteria for variable inclusion/exclusion based on EDA or statistical tests.

Modelling

1. **Model Building:**
   - Explain step-by-step processes:
     - Baseline model.
     - Incorporation of Lasso and Ridge regression for feature selection and regulariz
   - The rationale for selecting the final linear regression model.
2. **OLS Summary:**
   - Key statistics like coefficients, p-values, R^2, and adjusted R^2.
   - Interpretation of significant predictors.
3. **Rejected Models:**
   - Briefly discuss why certain intermediat odels were unsuitable (e.g., overfitting,

Introduction

Provide a brief overview of your project, including:

- The objective: Predicting the target variable and understanding the relationships between independent and dependent variables.
- The importance of the problem: Why this analysis is relevant or beneficial.

Example:
"This analysis aims to predict [dependent variable] using multiple independent variables from a [dataset description]. By applying multiple linear regression, we strive to identify significant predictors and evaluate the model's performance."

Exploratory Data Analysis

Discuss the dataset and provide insights into variable relationships:

1. **Levels of Measurement:** Specify the measurement type (nominal, ordinal, interval, ratio) for each variable.
2. **Descriptive Statistics:** Include measures such as mean, median, standard deviation, and range.
3. **Visualizations:** Provide plots like:
   - Histograms for distributions.
   - Boxplots for outlier identification.
   - Scatterplots for relationships between variables.
4. **Correlation Matrix:** Present the correlation heatmap to understand variable interdependencies.

Diagnostics

1. **Gauss-Markov Assumptions:**
   - Linearity: Residual vs. Fitted plot.
   - Homoscedasticity: Breusch-Pagan test or visualizations.
   - Normality of residuals: Q-Q plot or Shapiro-Wilk test.
   - Independence: Durbin-Watson statistic.
2. **Multicollinearity:**
   - Variance Inflation Factor (VIF) and its thresholds.
3. **Model Fit:**
   - Residual diagnostics and leverage plots.

Evaluation

1. **Performance Metrics:**
   - Train-test split method.
   - Metrics such as R^2, Mean Squared Error (MSE), and Mean Absolute Error (MAE).
2. **Validation:**
   - Cross-validation results (if performed).
3. **Comparison:**
   - Comparison of predicted vs. actual values with a visualization (e.g., scatterplot or residual plot).

Project Report on Multiple Linear Regression & Time Series Analysis(Statistics and Optimization)

Name :- Shivansh Bhatnagar
Roll No. :- x23237252
Institution:- National College of Ireland, Dublin
Course:- Msc in Data Analytics

***Abstract*—This is a Project Report of detailed analysis and steps performed in Multiple Linear Regression(MLR) and Time Series Analysis to predict forecasts, and trends using two provided datasets: mlr2.csv and ts2.csv. The MLR section focuses on evaluation distribution between variables, Data Cleaning, Data Preprocessing, Modelling and Model Optimization including Ridge and Lasso Models then performed prediction using required statistical metrics and techniques to validate the model and it's forecasting. The Time-Series Analysis addresses Trends, stationary checks, transformations to stabilise the variances, ARIMA, SES, HOLT, and HYBRID modelling for forecasting. Diagnostic and validation steps of the model then performed forecasting using the above models and compared the models using statistical techniques.**

Keywords—mlr, time series, model, ols, predictions, forecast, plot, residual, correlation, matrix, Arima, r2, Mae, Rmse, Mape, diagnostic, evaluation, pandas.

I. INTRODUCTION

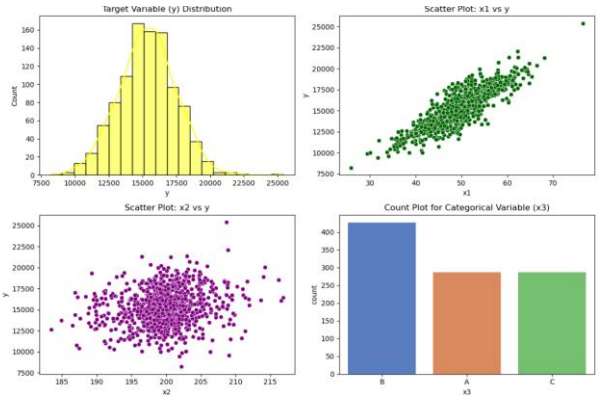
Data Analytics performs a predictive analytical approach by observing previous trends to extract meaningful insights and to make decisions future-oriented. This report is to briefly explains and justifies below Statistical Techniques:-

- **Multiple Linear Regression(MLR)**:- MLR used for prediction using target(dependent) variable using various independent variables inorder to understand relationships, predict values and validate model.
- **Time Series Analysis**:- Time series analysis identifies trends, seasonality and patterns in historical data for future forecasts.

II. MULTIPLE LINEAR REGRESSION ANALYSIS

2.1 INTRODUCTION

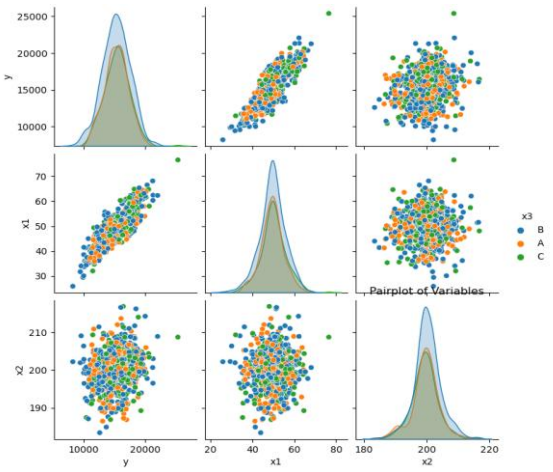
The dataset for this analysis selected on basis of last digit of student ID, mlr2.csv which contains 3 numerical variables and 1 categorical variable. And our target variable is 'y' because it represents continuous values. Other 3 columns are features/ independent variables. 'x1' and 'x2' are numerical features. Which used to calculate scores and 'x3' is categorical feature different categorical groups of 'y'. While testing split and training phase I have used Random seed in random state, whose value is equal to my roll number 23237252. Our goal of the project is to use these variables, clean data, apply mlr model, validate model using various statistics metrics then predict using previous trends. Then compare Predicted vs Actual value. (below image show casing description of variables.)



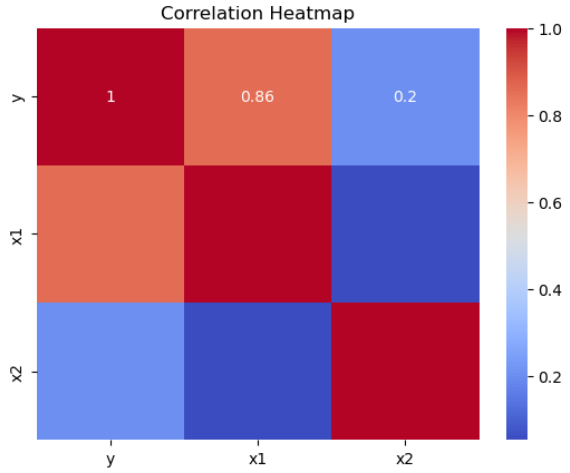
2.2 Exploratory Data Analysis(EDA)

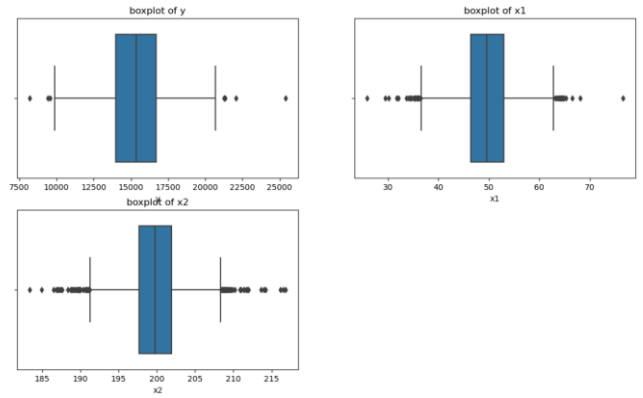
EDA is used to analyze, summarize variables in data and using EDA we finds relationships and patterns between each variables. Or we can see EDA briefs data characteristics which is most important to identify inorder to perform further advance analysis.

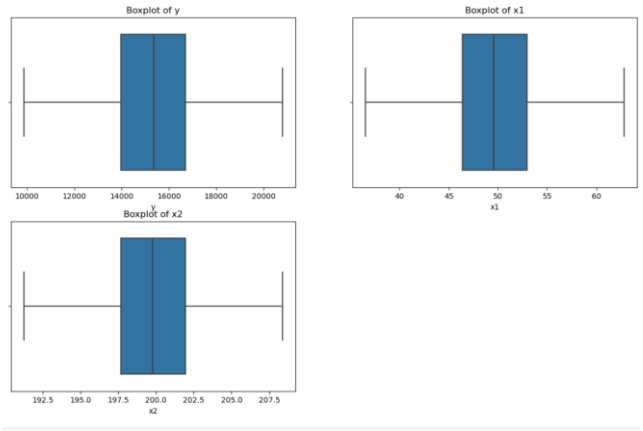
a. Here we have prepared chart to show pairplot of Variables to provide multi dimensional image to identify patterns and relevance between variables. My pairplot reflecting distribution of each variable along with diagonal plots. Scatters identifying pattern between pair of variables. Color coding reflecting how categorical variable 'x3' influencing on each variable with their different categories.



b. The correlation matrix showcases the relationship between variables which gives direction to select features on the basis of highly correlated variables. By Analyzing correlation we can detect multicollinearity. Where we select variables to enhance regression model accuracy. Where scale ranges between -1 to 1, 1 means negetive correlation, 0 means no correlation and 1 means high correlation. In my below correlation matrix. ‘y’ and ‘x1’ reflecting strong positive correlation indication 0.86 value. ‘y’ and ‘x2’ showing weak correlation with 0.2 value. ‘x1’ and ‘x2’ indicated average correlation with 0.4 value.







2.3 Data Preparation

Data Preparation is very crucial step on any Statistical / Data Analysis because it helps to clean the data and removes unnecessary anomalies from data to make data easily usable.

a. In my mlr analysis first I have performed the data cleaning step where I have checked null values and dropna step inorder to increase accuracy of my regression analysis.

b. Detect and Handled outliars by calculating Lower(LW) and upper bounds(UP) to identify interquartile range(IQR.)

outliar directly impacts on predictive model results by increasing distortions in values, so I have replaced the calculated bounds with their allocated boundaries. This step balances the propotion of data to make model more reliable.

c. Performed Additional Outliar handling step by removing rows where z-score of ‘x1’ and ‘x2’ is greater than 3 standard deviations of mean because 99.7% data points captures at 3 SD of mean, which removes extreme anamolies.

These steps removed my outliers distortions in data and made data smooth. Below boxplots shows data before outlier handling and after outlier handling. (1- Before, 2- After)

d. Standardized the numerical variables and performed one hot encoding on categorical variables to make it usable in linear regression. Used standard scalar method to transform ‘x2’, and ‘x2’ to get mean=0 and Standard Deviation=1. One hot encoded divided ‘x3’ and 3 different categories with Boolean values (True/False) so we can fit the regression model, As the Statistics model required numerical values.

e. Then we used Log-Transform on target variable ‘y’ to normalise the distribution. This method stabilises the variance by removing skewness and make data normally distributed. I have added 1 to handle zero or negative value error in logarithm. Formulae used (data['y'] = np.log(data['y'] + 1)).

After Performing outlier handling and log transform steps, I have shown Raw data vs Clean data(Image below). In which we can clearly see the changes in ‘y’ Variances are stabilized which improved the normality of target variable. ’y’ values scaled down to 9-10 from very high value. Which made data ready to fit in multiple linear regression and perform analysis.

```
Original data dimensions: <bound method NDFrame.describe of              y         x1          x2 x3
0     18546.674932  58.620100  202.906299  B
1     16616.498903  56.019437  192.718794  B
2     12773.312433  47.857470  189.915798  A
3     14613.038340  46.831493  199.289231  B
4     16549.372517  51.656511  201.500167  C
..             ...        ...         ... ..
995   14460.386067  48.609211  200.377106  C
996   15042.981309  48.798773  198.014740  B
997   15754.974480  49.966635  200.496834  A
998   18129.404048  59.425034  198.509978  A
999   16380.784037  46.423217  201.724777  C

[1000 rows x 4 columns]>
Cleaned data dimensions: <bound method NDFrame.describe of           y         x1          x2 x3
0     9.828100  58.620100  202.906299  B
1     9.718212  56.019437  192.718794  B
2     9.455192  47.857470  191.242679  A
3     9.589738  46.831493  199.289231  B
4     9.714164  51.656511  201.500167  C
..         ...        ...         ... ..
995   9.579237  48.609211  200.377106  C
996   9.618733  48.798773  198.014740  B
997   9.664975  49.966635  200.496834  A
998   9.805346  59.425034  198.509978  A
999   9.703925  46.423217  201.724777  C

[1000 rows x 4 columns]>
```

f. Created 2 new values for separating dependent('y') and independent variables('x1','x2','x3') to make model ready for training. X captures all values except 'y' and Y contains 'y'.

2.4 Modelling

a. Generated random_seed variable whose value is nothing but my Student ID(23237252) to keep its value equal to random_state in the train and test data splitting stage.

b. Split the data into Training and testing sets to make the model ready to fit in the regression model and to ensure that we can evaluate the model's performance on unseen data. The random seed ensures reproducibility.

c. Used Ridge Regression to increase performance and reduce risk of overfitting because it helps coefficients to shrink.

d. Used Lasso Regression to more Optimize the performance because the Lasso Regression technique makes some coefficients 0. Both Ridge regression and Lasso Regression evaluated using Mean Square error(MSE) and R2 Metrics. [3]

e. Implemented and fit Standard Multiple Linear Regression Method as it is our main goal of this task to perform on MLR with more accurate and optimize way.

f. Made prediction on data that we kept in testing set which is one of the most important step of our MLR model because this step allows us to determine the ability of fitted model to make accurate predictions on the dataset.

2.5 Interpretation

a. Intercept: This is the expected value of the dependent variable y when all independent variables are equal to zero.

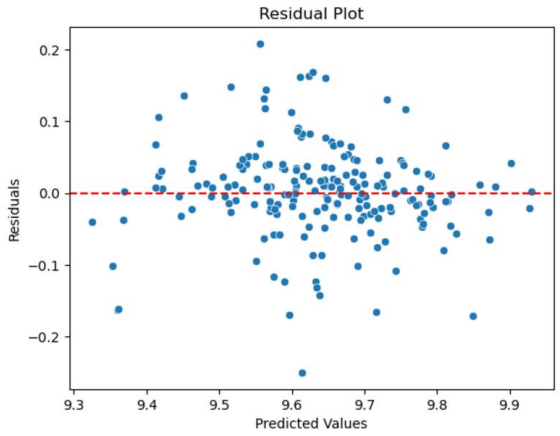
b. Coefficients: Each one tells the partial change of the dependent variable 'y' by changing one unit of the respective independent variable, while the other variables are held constant.[4]

c. Output of Coefficients and Interpretation :-

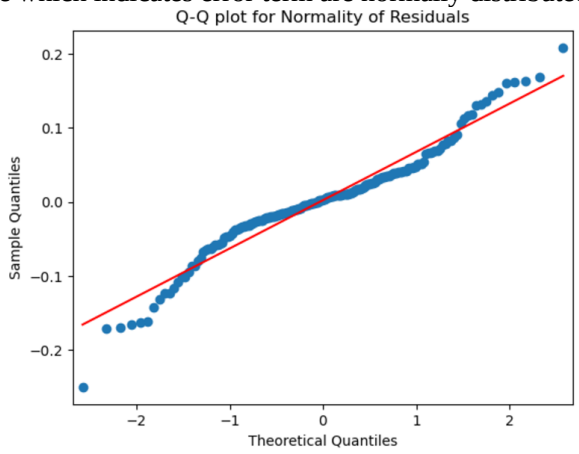
- x1: 0.02086612 implies that for every one unit increase in x1, y increases approximately by 0.021 units.
- x2: 0.00546459 implies that for every one unit increase in x2, y increases approximately by 0.0055 units.
- x3_A: 0.00550045 is the effect of the category A of x3 on y compared to the reference category.
- x3_B: -0.00744192 is the effect of the category B of x3 on y compared to the reference category.
- x3_C: 0.00194147 is the effect of the category C of x3 on y compared to the reference category.

2.6 Diagnostics

a. Plotted Residual plot of linearity which shows Residual vs Predicted values. My chart shows reflects linear relationship between test and prediction.



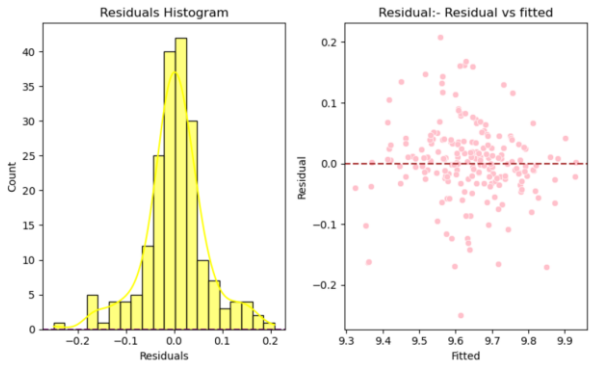
b. Plotted QQ plot of residual which is reflecting 45 degree line which indicates error term are normally distributed.




```
Durbin-Watson statistic: 2.1871622361222998
```

c. Durbin-Watson static is coming 2.18 which indicates residuals are independent.

d. Plotted Residual vs Model Fitted chart which shows assumptions of linearity, homoscedasticity, and independence of residuals are likely to satisfy MLR model. Which validates model is a good fit and residual does not exhibit any unsymmetric pattern.



e. Implemented Variance Inflation Factor(VIF) to check multi collinearity and dropping x3_A because it has high VIF value, and other variables value is around 1(VIF<10 means multicollinearity is not an issue.)

f. Gauss-Markov assumptions:- As per the above outputs result meets Gauss-Markov assumptions and this model seems to be a good fit.

2.7 Evaluation

a. Calculated R2, MSE,MAE value of linear regression model.

- R2=0.7456 which means 74.6% variance on 'y'. Which means the Model is a very good fit.
- Mean Squared Error(MSE)= 0.004249784231650473. Lower MSE shows a better performance model. It is best when closest to 0.
- Mean Absolute Error(MAE)= 0.04546740554710334. Lower MAE confidences accurate prediction. It is best when closest to 0.

b. Model Comparison (Linear and Ridge Performing almost equally well whereas Lasso Regression performance is comparatively lower but not bad.)

Linear Regression: MSE = 0.004249784231650473, R^2 = 0.7461
Ridge Regression: MSE = 0.004249657347240059, R^2 = 0.7461
Lasso Regression: MSE = 0.00477867178213764, R^2 = 0.7145

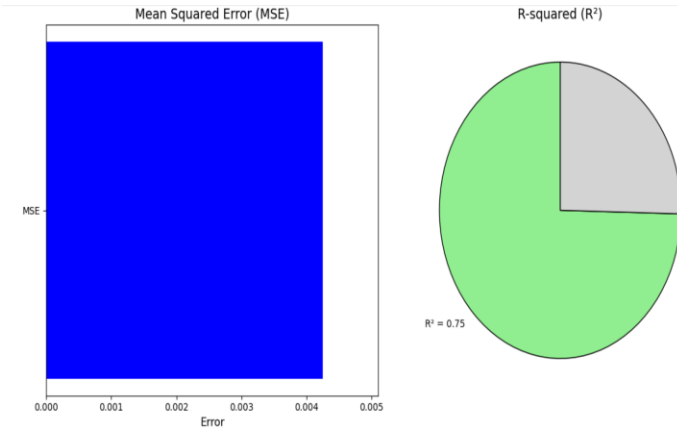
c. Imputed (Ordinary Least Square)OLS Regression Summary:- From the diagnostic analysis and model evaluation metrics, the Multiple Linear Regression model passes the Gauss-Markov assumptions. This model is robust with high predictive accuracy from both the training and test data. The linear and Ridge regression models have somewhat comparable performances, whereas the Lasso regression gives a little worse accuracy. Overall, this model fits well and makes reliable predictions.

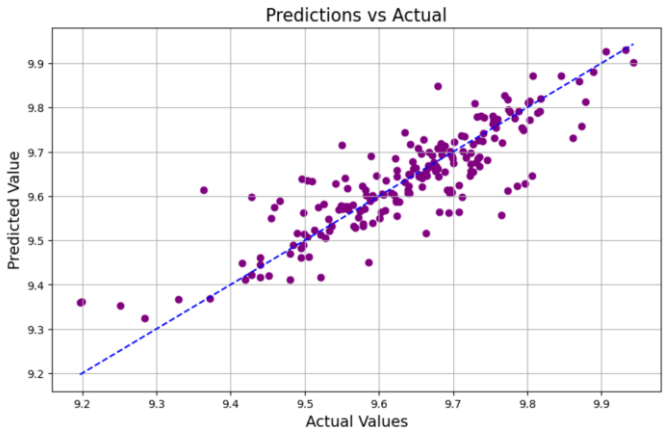
OLS Regression Results

Dep. Variable:	y	R-squared:	0.745
Model:	OLS	Adj. R-squared:	0.743
Method:	Least Squares	F-statistic:	579.7
Date:	Sun, 01 Dec 2024	Prob (F-statistic):	6.00e-234
Time:	21:31:50	Log-Likelihood:	978.14
No. Observations:	800	AIC:	-1946.
Df Residuals:	795	BIC:	-1923.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.6260	0.096	58.346	0.000	5.437	5.815
x1	0.0209	0.000	46.913	0.000	0.020	0.022
x2	0.0055	0.001	8.556	0.000	0.004	0.007
x3	1.8808	0.032	58.143	0.000	1.817	1.944
x4	1.8679	0.033	57.430	0.000	1.804	1.932
x5	1.8773	0.032	58.352	0.000	1.814	1.940

Omnibus:	21.944	Durbin-Watson:	2.077
Prob(Omnibus):	0.000	Jarque-Bera (JB):	45.372
Skew:	-0.096	Prob(JB):	1.40e-10
Kurtosis:	4.151	Cond. No.	1.67e+18

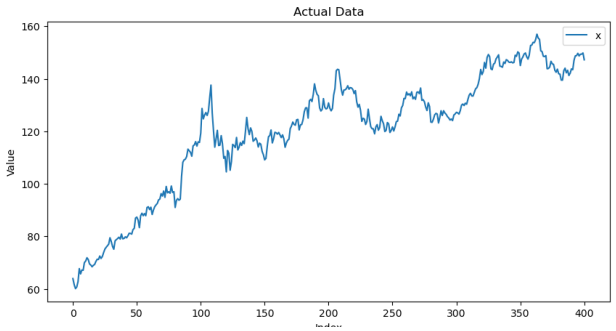




III. Multiple Linear Regression Analysis

3.1 INTRODUCTION

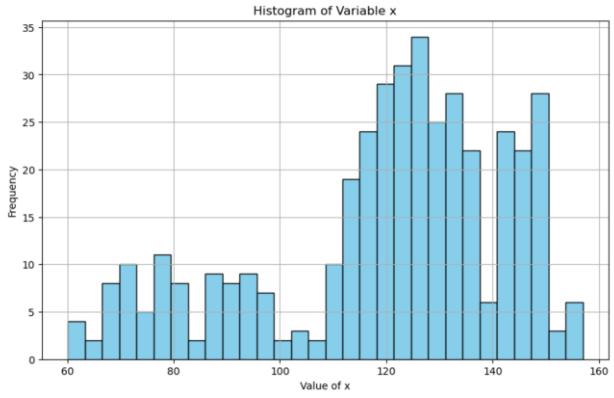
Time Series deployed to understand temporal trends and future forecasts using various models like ARIMA, SARIMA, Simple Exponential Smoothening(SES), HOLT, and HYBRID. My dataset(ts2.csv) is selected on the basis of the last digit of the roll number which is 2. The dataset contains 2 columns 1 in unnamed column which is only a sequence of data and other 'x' is our main target numerical variable for time series analysis. Which is non-stationary Our goal is to calculate forecast. Which we have performed a box-cox transform to stabilize the variance. Auto-Arima, SES forecast, holt forecast, and Hybrid Forecast models for forecasting then diagnosed all models. Perform ACF residual and check if it looks like white noise then show which model is giving the best-forecasted value and forecast value accordingly. (below image shows description of our dataset.)



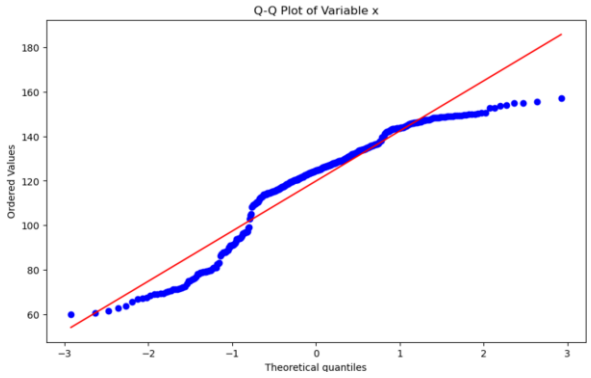
3.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis involves understanding the underlying patterns, distributions, and characteristics of the data.

a. I have described dataset to know characteristics of data. Below chart explains 'x' has indicating multimodal distribution.



b. Then I have Plotted QQ plot which shows extreme values in data which is not a sign of normal distribution. It indicating potential issues with normality. This insight guides to perform data transformations and further analysis.If data is normally distributed then all blue dots should lie along red line. But points are deviating which means variable facing normality.



c. Performed Shapiro-wilk test which interpreted Test Static=0.92 which is closer to 1 implies close to normality. P-value=1.19. This is extremely insignificant which means data is not normally distributed. Which is clear sign that data need transformation to work on further analysis.

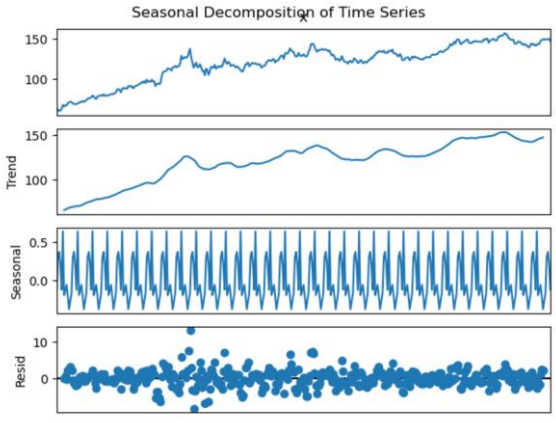
3.3 Data Prepration

a. Performed Missing value and duplicate value test, there is no missing value or duplicate value in target variable.

b. Made 1st unnamed column as index and 'x' main variable for time series analysis.

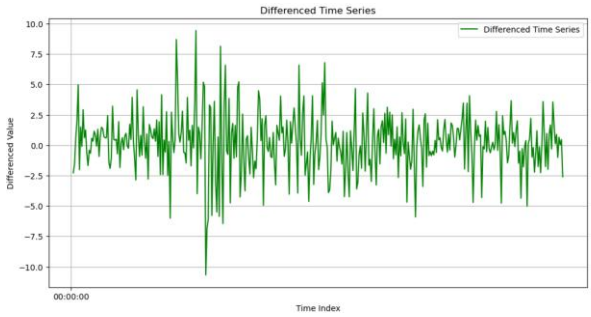
c. Performed Augmented Dickey Fuller Test, which checks whether time series in data is stationary or not. In result of this P-value=0.14 which is bigger than 0.05 and ADF static value is -2.38 which means data is not stationary.

d. Plotted Seasonal decomposition chart in which trend shows upward trends that indicated steady growth overtime. Strong regular seasonal patterns suggests some factors that directly impacts data. And residuals reflecting model is effectively decomposing meaningful time series components.



e. Applied Box Cox Transformation to make variance in stabilize form and differencing(time_series_diff =

ts_data['x'].diff().dropna()) to make the time series stationary in data. Because Shapiro-wilk test suggests transformation to handle normality issue, and ADF showed non stationary pattern. So I have applied Box-Cox and differencing to fix these issues.



f. Splitted the data into training and testing sets by assigning 80% to training and 20% to testing set.

3.4 Modelling

a. Implemented the ARIMA(Auto-Regressive Integrated Moving Average) model to effectively capture the trend seasonality for solid accurate forecasting. And used Auto Arima which auto selects optimal parameters (p, d, qp , d, qp , d , q) using AIC and BIC tests. It shows the best seasonal order that can be fitted in main ARIMA / SARIMA models. As my model showed best order according to my data is (2,1,2)(0,0,0)[12].

```
Performing stepwise search to minimize aic
 ARIMA(2,1,2)(1,0,1)[12] intercept   : AIC=6605.808, Time=2.40 sec
 ARIMA(0,1,0)(0,0,0)[12] intercept   : AIC=6612.343, Time=0.03 sec
 ARIMA(1,1,0)(1,0,0)[12] intercept   : AIC=6616.033, Time=0.13 sec
 ARIMA(0,1,1)(0,0,1)[12] intercept   : AIC=6616.008, Time=0.19 sec
 ARIMA(0,1,0)(0,0,0)[12]             : AIC=6612.355, Time=0.02 sec
 ARIMA(2,1,2)(0,0,1)[12] intercept   : AIC=6604.435, Time=1.68 sec
 ARIMA(2,1,2)(0,0,0)[12] intercept   : AIC=6604.625, Time=0.82 sec
 ARIMA(2,1,2)(0,0,2)[12] intercept   : AIC=6605.947, Time=4.19 sec
 ARIMA(2,1,2)(1,0,0)[12] intercept   : AIC=6604.723, Time=1.95 sec
 ARIMA(2,1,2)(1,0,2)[12] intercept   : AIC=inf, Time=6.84 sec
 ARIMA(1,1,2)(0,0,1)[12] intercept   : AIC=6614.012, Time=0.58 sec
 ARIMA(2,1,1)(0,0,1)[12] intercept   : AIC=6613.484, Time=0.33 sec
 ARIMA(3,1,2)(0,0,1)[12] intercept   : AIC=6604.778, Time=2.00 sec
 ARIMA(2,1,3)(0,0,1)[12] intercept   : AIC=6604.764, Time=1.42 sec
 ARIMA(1,1,1)(0,0,1)[12] intercept   : AIC=6616.868, Time=0.47 sec
 ARIMA(1,1,3)(0,0,1)[12] intercept   : AIC=6614.914, Time=0.72 sec
 ARIMA(3,1,1)(0,0,1)[12] intercept   : AIC=6615.481, Time=0.62 sec
 ARIMA(3,1,3)(0,0,1)[12] intercept   : AIC=6604.753, Time=2.17 sec
 ARIMA(2,1,2)(0,0,1)[12]             : AIC=6603.812, Time=1.30 sec
 ARIMA(2,1,2)(0,0,0)[12]             : AIC=6603.800, Time=0.78 sec
 ARIMA(2,1,2)(1,0,0)[12]             : AIC=6603.993, Time=1.63 sec
 ARIMA(2,1,2)(1,0,1)[12]             : AIC=6605.635, Time=2.27 sec
 ARIMA(1,1,2)(0,0,0)[12]             : AIC=6613.032, Time=0.37 sec
 ARIMA(2,1,1)(0,0,0)[12]             : AIC=6612.498, Time=0.40 sec
 ARIMA(3,1,2)(0,0,0)[12]             : AIC=6604.764, Time=1.20 sec
 ARIMA(2,1,3)(0,0,0)[12]             : AIC=6604.745, Time=0.60 sec
 ARIMA(1,1,1)(0,0,0)[12]             : AIC=6614.816, Time=0.20 sec
 ARIMA(1,1,3)(0,0,0)[12]             : AIC=6613.777, Time=0.27 sec
 ARIMA(3,1,1)(0,0,0)[12]             : AIC=6614.461, Time=0.56 sec
 ARIMA(3,1,3)(0,0,0)[12]             : AIC=6605.995, Time=1.03 sec

Best model:  ARIMA(2,1,2)(0,0,0)[12]
Total fit time: 37.217 seconds
```

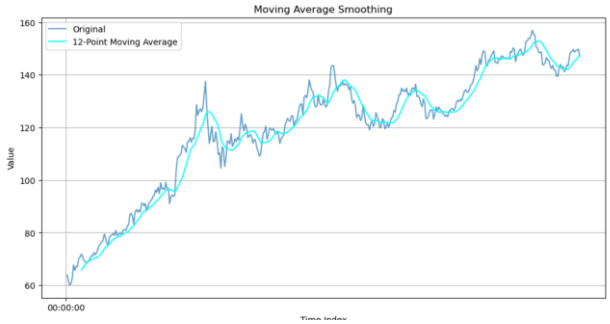
b. Now fitted the best seasonal order in ARIMA model using Auto Arima result. We have used OIM covariance in model fitting for good parameter estimation.

c. Performed predictions using Arima forecast on transformed data. Which will calculate confidence intervals for the forecasting.

d. Performed Inverse Box Cox transformation, to convert predicted value back to original form and inputed that in new variables to make it ready to perform on other forecasting model other than ARIMA. Implemented Confidential Intervals for lower and upper intervals (arima_conf_int = arima_forecast.conf_int()) to tranfsorm the confidence interval bounds.

e. Implemented Simple exponential smoothening model. It helps to smoothen series and give short term forecast.I have chosen this model because this model can be used on data without strong trends and seasonality

f. Performed Moving Average metrics of statistics which is ideally used to reduce noise and underlying patterns of dataset. I have taken window equal to 12 which will take average of last 12 data tabs consecutively.



g. Then I performed Holt's forecasting method which is the extension to simple exponential smoothening techniques which relies trends and level of data.

h. Then I performed Hybrid Model which is taking average of above three models and showing predictions according to this model (hybrid_predictions = (arima_predictions + ses_predictions + holt_predictions) / 3). Which main target is to consolidate predictions of ARIMA, SES, Holt's to improve the forecasts by pulling out strength of each model. [2]

3.5 Interpretation

Summarized values, outputs and coefficients of ARIMA model to analyze there job in model. Used Sarimax model summary method for this. Dependent Variable: y (Target variable with 320 observations)[1]

- Model Type: ARIMA — two AR Terms with the first-order difference and two MA terms.
- Log-Likelihood: -3296.900 — fit statistic; higher values are considered better.
- Akaike information criterion (AIC): 6603.800: Quality metric considering both the fit and complexity of a model (lower value indicates better performance)
- Bayesian information criterion: 6622.626: another general quality criterion with heavier penalty for complex models.
- AR Terms (ar.L1 and ar.L2): Significant with $p < 0.001$, reflecting the effect of past values. Coefficients reveal strong negative lags.

- MA Terms (ma.L1 and ma.L2): Significant with $p < 0.001$, accounting for error terms of previous steps. Coefficients are positive and large.
- Variance (sigma2): High residual variance, suggesting considerable variability in the data that is not accounted
- Ljung-Box test=0.39 indicated value greater than 0.05 which means residuals looks like white noise and forecast is ready to shoot.
- Jarque-Bera test shows some non normality in residuals.

All these values concludes non normality but this model is good fit on basis of AIC and BIC scores.

SARIMAX Results

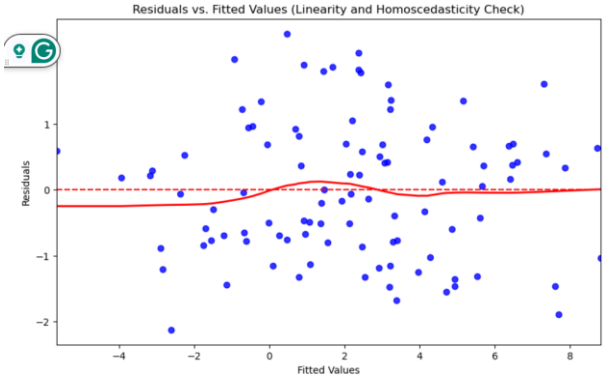
Dep. Variable:	y	No. Observations:	320
Model:	ARIMA(2, 1, 2)	Log Likelihood	-3296.900
Date:	Mon, 02 Dec 2024	AIC	6603.800
Time:	11:30:07	BIC	6622.626
Sample:	0	HQIC	6611.318
	- 320		
Covariance Type:	oim		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.1999	0.040	-29.789	0.000	-1.279	-1.121
ar.L2	-0.8976	0.039	-22.740	0.000	-0.975	-0.820
ma.L1	1.3038	0.028	46.984	0.000	1.249	1.358
ma.L2	0.9612	0.027	35.385	0.000	0.908	1.014
sigma2	5.864e+07	1.24e-11	4.74e+18	0.000	5.86e+07	5.86e+07

Ljung-Box (L1) (Q):	0.75	Jarque-Bera (JB):	103.36
Prob(Q):	0.39	Prob(JB):	0.00
Heteroskedasticity (H):	1.33	Skew:	-0.14
Prob(H) (two-sided):	0.14	Kurtosis:	5.77

3.6 Diagnostics

a. Performed Linearity and Homoscedasticity Check to check prediction levels. Linearity=Randomly scattered blue dots around red dotted line which is zero shows systematic patterns. Homoscedasticity= the residual variances remain const ant over the range of the fitted values, thus the assumption of homoscedasticity is met.



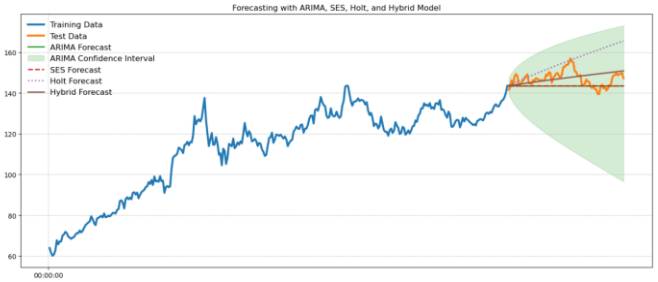
b. Performed Durbin-Watson statistic whose output is 1.9450573716040487 (Value between 0 and 2 considered good). This means model is fulfilling all assumptions of independent errors. Which validates the model is good fit.

3.7 Evaluation

I have performed model evaluation of all foresting model performed and calculated MAE(Mean Absolute error), RMSE(Root Mean Square error) and MAPE(Mean absolute percentage error) value test to validate the model. I have calculated these scores on basis of Hybrid Model because it is pulling strenngth of ARIMA, SES, Holt's forecast and forming new hybrind forecasting model.[1]

- Mean Absolute Error (MAE): 3.3760644565929625. This is measuring absolute difference between the actual and predicted values. And it is good outcome. Ideally this value must be closest as possible to 0.
- Root Mean Squared Error (RMSE): 4.318977075365975. It is showing quite large error which accounting large deviations constantly.
- Mean Absolute Percentage Error (MAPE): 2.30%. This shows average error percentage of actual values. It shows model has very good accuracy and model is reliable for forecasting with minimal deviation, which is good sign of good model fit and good forecasts.

In last I have visualized ARIMA, SES, Holt's and Hybrid model forecasts in one frame and reflected short summary of my model.



ARIMA Model Summary

Model: ARIMA(2, 1, 2)
AIC: 6603.8, BIC: 6622.63, HQIC: 6611.32 (model fit criteria suggest a reasonable fit)
Ljung-Box Test: p-value of 0.39 (no significant autocorrelation in residuals)

Conclusion

The hybrid model performs well with low error metrics (MAE, RMSE, MAPE).
The ARIMA model also fits the data reasonably well, with residuals behaving like white noise.
Therefore, both models provide a good fit to the data.

4. Conclusion

- Multiple Linear Regression:- Performed EDA, Data Cleaning, Preprocessing, transformation of data, to optimize model used Lasso and Ridge regression model for Linear regression and predicted with strong model and validate model by R2= 74.5 and MSE=0.004.
- Time Series Analysis:- Performed Forecasting using Hybrid model by making data stationary, normalize, transformed and smooth. Which resulted good model with MAPE 2.30% value.

5. Refrences

[1] " Gareth James, Daniela Witten," An Introduction to Statistical Learning with Application in python. Available: https://www.statlearning.com/. [Accessed: Dec. 02, 2024].

[2] A. Author:- Jin Xiano, Shouyang Wang, "A hybrid model for time series forecasting," ResearchGate. [Online]. Available: https://www.researchgate.net/publication/287569827_A_hybrid_model_for_time_series_forecasting/link/60c86c1aa6fdcc8267cf6d37/download?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19. [Accessed: Dec. 02, 2024].

[3] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, "Data Mining: Practical Machine Learning Tools and Techniques," 4th ed., San Francisco, CA, USA: Morgan Kaufmann, 2016. [Online]. Available: https://www.sciencedirect.com/book/9780123748560/data-mining-practical-machine-learning-tools-and-techniques. [Accessed: Dec. 02, 2024].

[4] R. Carter and M. Hill, "Chapter 4: Machine Learning," Eller College of Management, University of Arizona, 2024. [Online]. Available: https://eller.arizona.edu/sites/default/files/202404/carter%20hill%20ml%20chapter%204%2015%202024.pdf. [Accessed: Dec. 02, 2024].