PART 1: TSP (Optimization Problem)

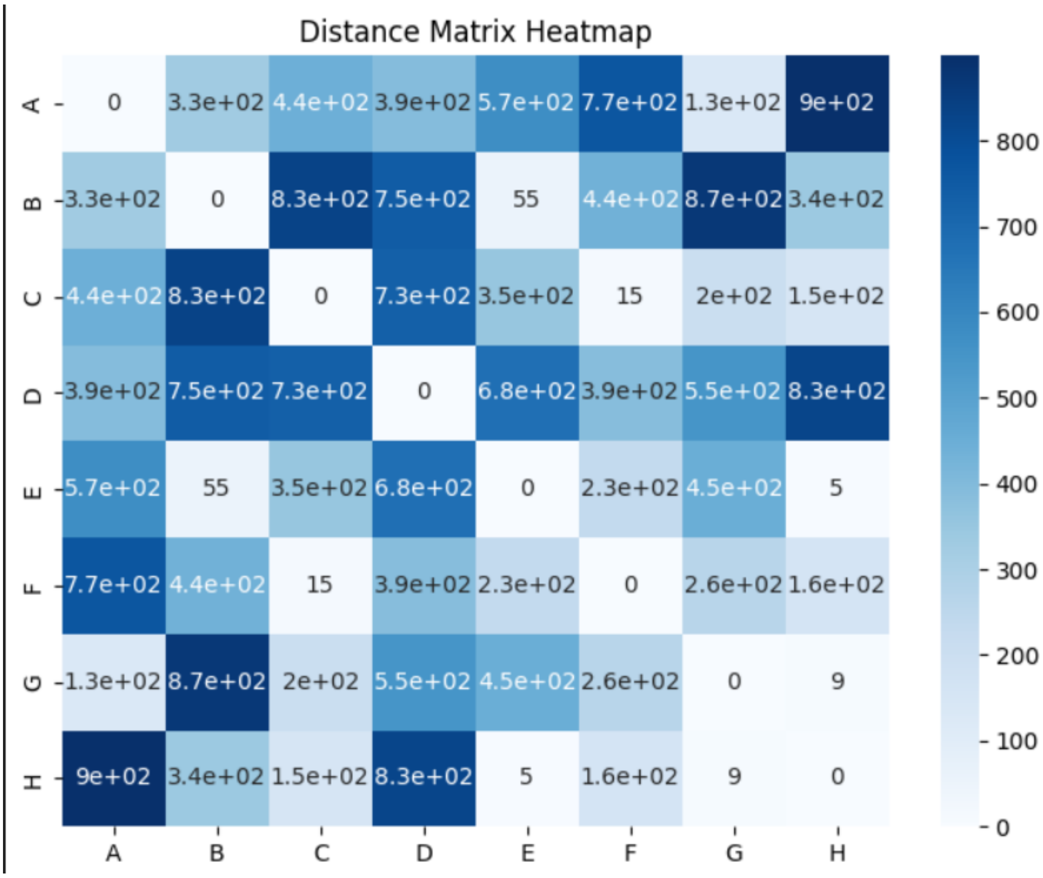
Introduction

In this section, the delivery worker visits a number of cities to deliver the packages using efficient delivery routes. Both time-based and item-based restrictions are taken into consideration for each relevant city. The objective is to determine the delivery person's optimal path while taking the constraints into consideration.

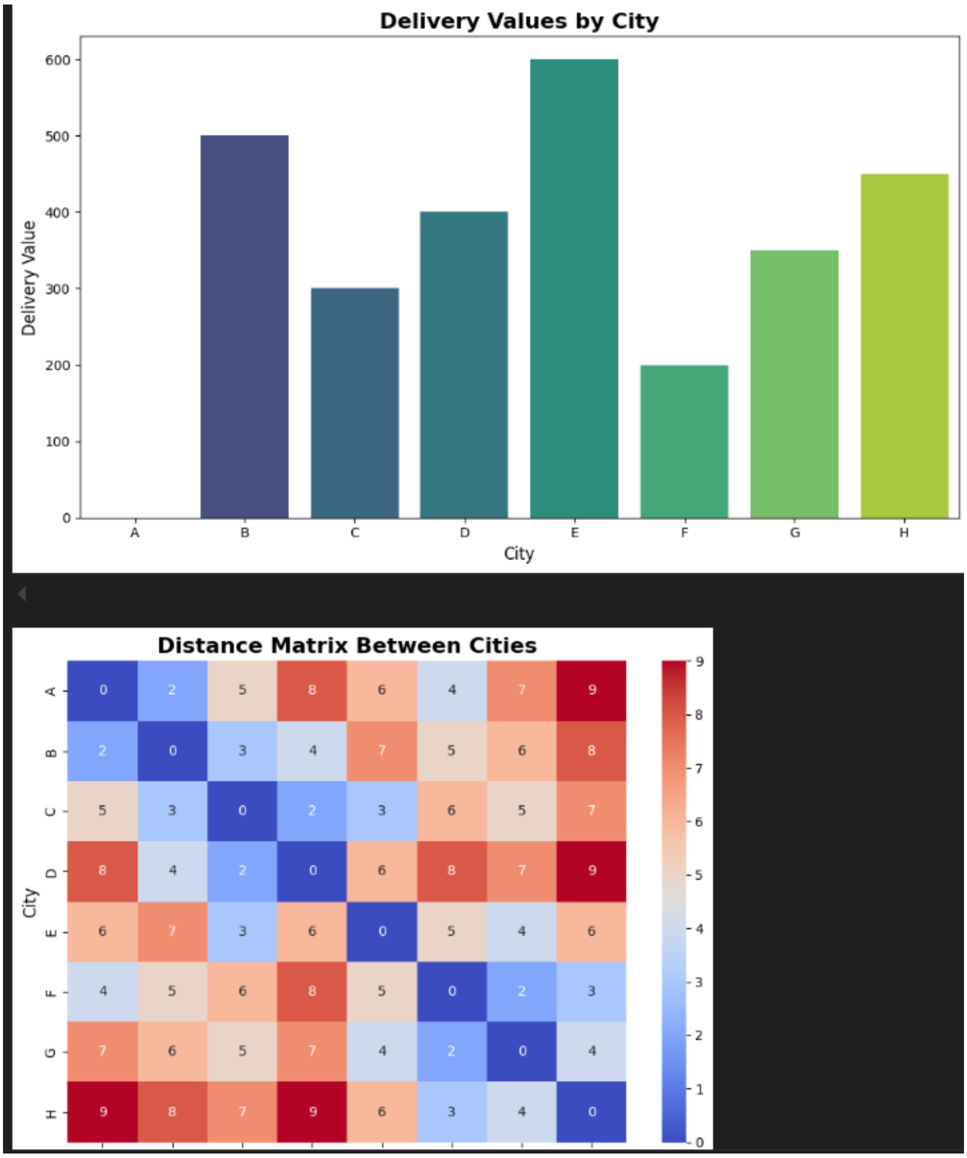
A. Define the decision variables and formulate the mathematical model for the above problem

EDA was carried out in the below steps:

a) Distance Matrix: As seen in the figure below, a distance matrix was created using a random seed to show the separations between various cities and aid in determining the proximity and relationships of the current variables.

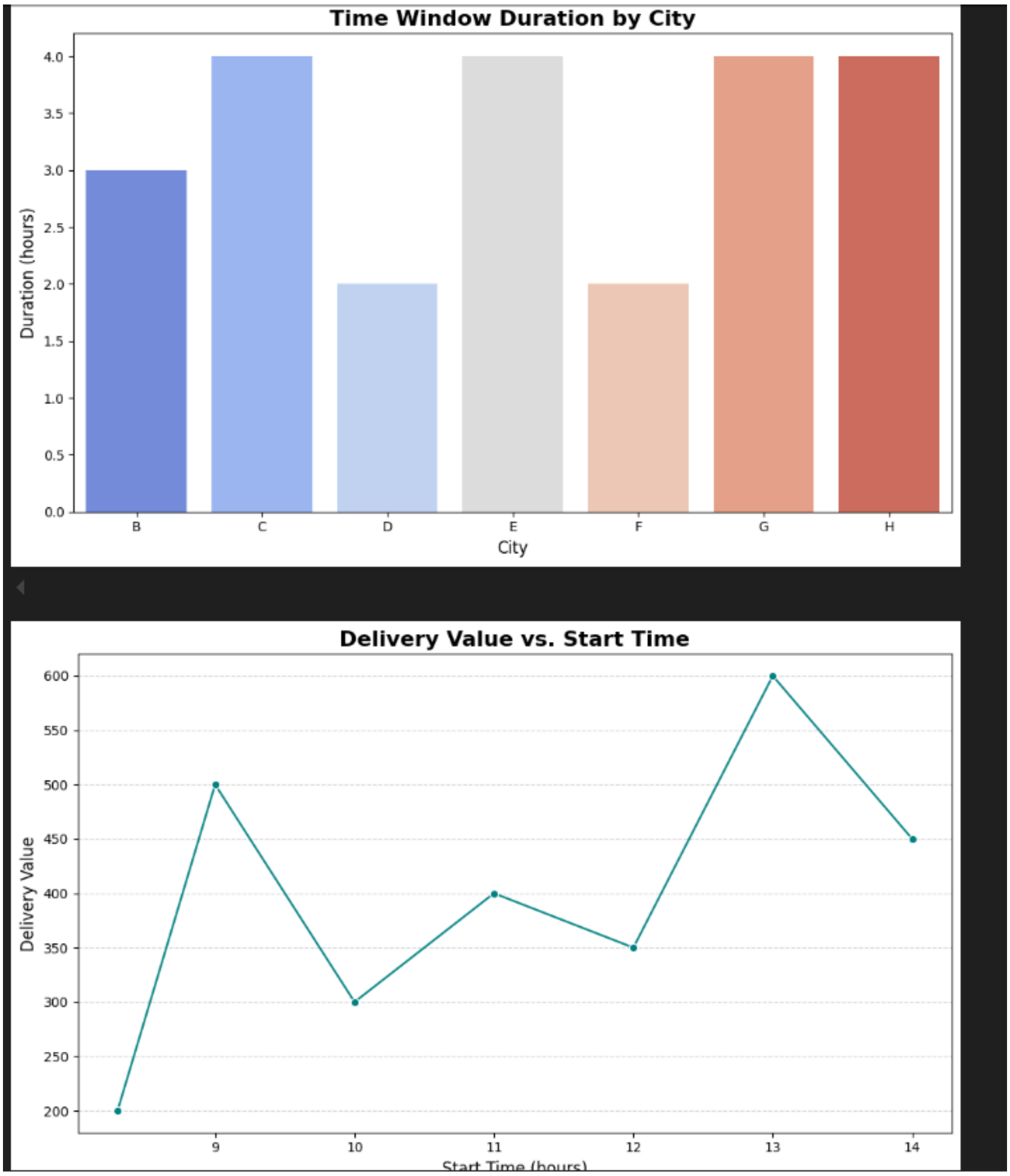


a) Data preparation: This stage involves preparing the data for the next model, which includes delivery values and time windows.This aids in identifying the crucial variables that should be given priority. Delivery values are displayed as a bar chart that shows the demand for deliveries in each city.To determine the logical paths between cities and optimize the routes between them, another heatmap was made.It uses a range of graphs (fig. 2) to illustrate this stage, as seen below:



A	0	2	5	8	6	4	7	9
B	2	0	3	4	7	5	6	8
C	5	3	0	2	3	6	5	7
D	8	4	2	0	6	8	7	9
E	6	7	3	6	0	5	4	6
F	4	5	6	8	5	0	2	3
G	7	6	5	7	4	2	0	4
H	9	8	7	9	6	3	4	0

Fig 2



b) Solve the problem using the Exact methods (Pulp) or using a heuristic method

Both the pulp and the heuristic methods are used in this stage, and I compared them to determine which one worked best.
In order to determine the best course of action and achieve 100% accuracy in terms of both distance and time, the pulp method takes into account every potential solution and every conceivable combination.
The greedy technique selects the next nearest city after making local optimal choices in a given city. This does not offer the plan or the worldwide optimal solution.

In contrast to the exact method, the comparison reveals that this approach only achieved 66.6% distance accuracy, 48% value accuracy, and an overall accuracy of 57.7%.
The precision

C) Remodel the problem so that we have two objective functions one to maximise the profit and the other to minimize the total distance travelled

- Implemented and checked Best possible route, Total Value and Total Distance.

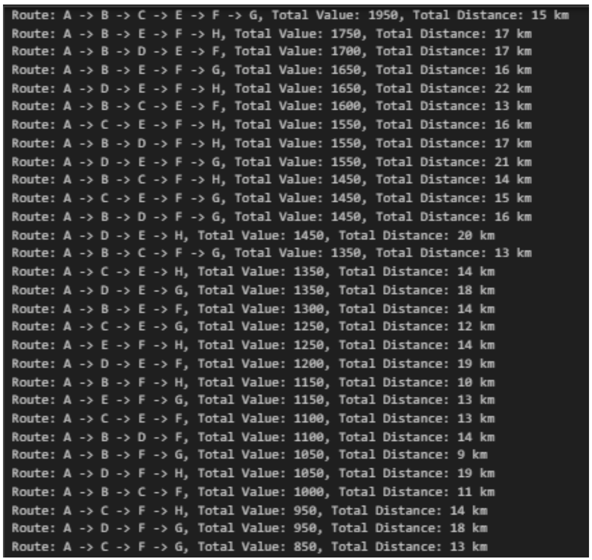
By weighing all possible city combinations according to limitations, total delivery value, and trip distance, this algorithm optimizes delivery routes. It

Creates Routes: Beginning with city 'A', itertools.combinations generates every route from the city list, taking into account three to five cities every route.

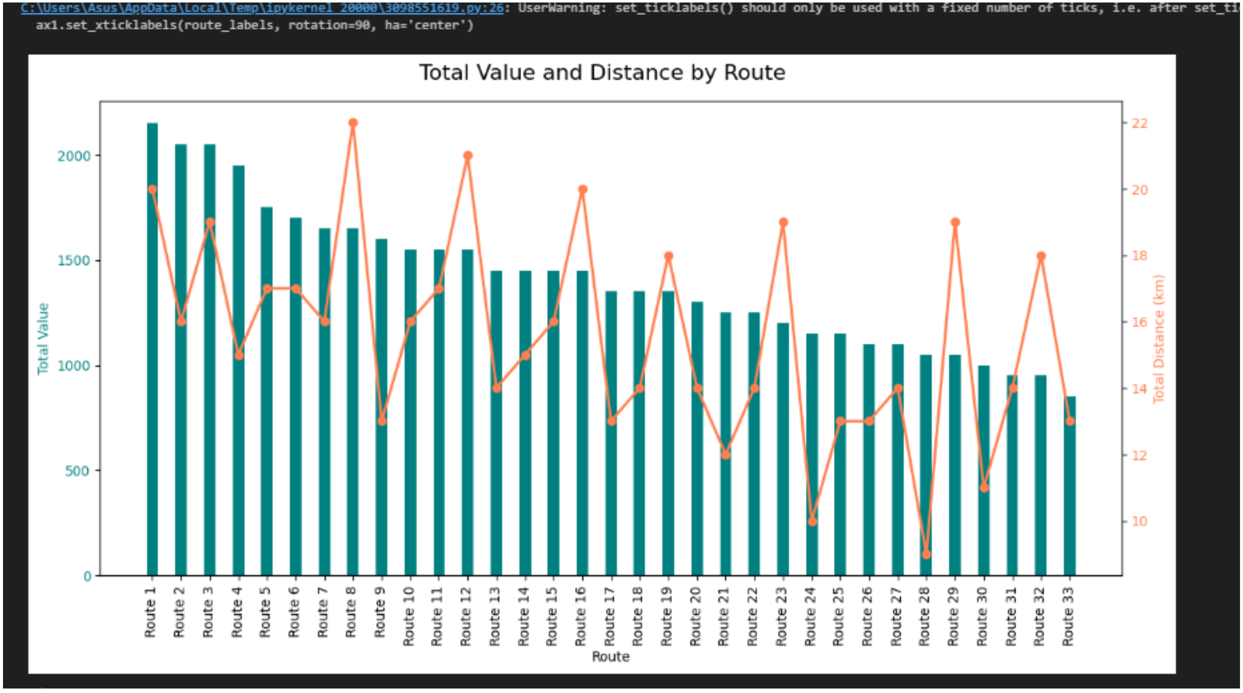
computes Metrics: Functions use a related value list to calculate the total delivery value and a predetermined distance matrix to calculate the total distance.

Relevant Limitations: filters routes according to logical restrictions, like conditional dependencies (e.g., if city B is visited, F must also be included) or mutual exclusivity (e.g., either city D or C).

Optimizes and Sorts: Chooses and arranges viable routes by minimizing distance and optimizing delivery value, then shows the top routes along with their distances and values.



```
Route: A -> B -> C -> E -> F -> G, Total Value: 1950, Total Distance: 15 km
Route: A -> B -> E -> F -> H, Total Value: 1750, Total Distance: 17 km
Route: A -> B -> D -> E -> F, Total Value: 1700, Total Distance: 17 km
Route: A -> B -> E -> F -> G, Total Value: 1650, Total Distance: 16 km
Route: A -> D -> E -> F -> H, Total Value: 1650, Total Distance: 22 km
Route: A -> B -> C -> E -> F, Total Value: 1600, Total Distance: 13 km
Route: A -> C -> E -> F -> H, Total Value: 1550, Total Distance: 16 km
Route: A -> B -> D -> F -> H, Total Value: 1550, Total Distance: 17 km
Route: A -> D -> E -> F -> G, Total Value: 1550, Total Distance: 21 km
Route: A -> B -> C -> F -> H, Total Value: 1450, Total Distance: 14 km
Route: A -> C -> E -> F -> G, Total Value: 1450, Total Distance: 15 km
Route: A -> B -> D -> F -> G, Total Value: 1450, Total Distance: 16 km
Route: A -> D -> E -> H, Total Value: 1450, Total Distance: 20 km
Route: A -> B -> C -> F -> G, Total Value: 1350, Total Distance: 13 km
Route: A -> C -> E -> H, Total Value: 1350, Total Distance: 14 km
Route: A -> D -> E -> G, Total Value: 1350, Total Distance: 18 km
Route: A -> B -> E -> F, Total Value: 1300, Total Distance: 14 km
Route: A -> C -> E -> G, Total Value: 1250, Total Distance: 12 km
Route: A -> E -> F -> H, Total Value: 1250, Total Distance: 14 km
Route: A -> D -> E -> F, Total Value: 1200, Total Distance: 19 km
Route: A -> B -> F -> H, Total Value: 1150, Total Distance: 10 km
Route: A -> E -> F -> G, Total Value: 1150, Total Distance: 13 km
Route: A -> C -> E -> F, Total Value: 1100, Total Distance: 13 km
Route: A -> B -> D -> F, Total Value: 1100, Total Distance: 14 km
Route: A -> B -> F -> G, Total Value: 1050, Total Distance: 9 km
Route: A -> D -> F -> H, Total Value: 1050, Total Distance: 19 km
Route: A -> B -> C -> F, Total Value: 1000, Total Distance: 11 km
Route: A -> C -> F -> H, Total Value: 950, Total Distance: 14 km
Route: A -> D -> F -> G, Total Value: 950, Total Distance: 18 km
Route: A -> C -> F -> G, Total Value: 850, Total Distance: 13 km
```



```
C:\Users\Asus\AppData\Local\Temp\ipykernel_20000\3098551619.py:26: UserWarning: set_ticklabels() should only be used with a fixed number of ticks, i.e. after set_ti
  ax1.set_xticklabels(route_labels, rotation=90, ha='center')
```

d) Describe the different steps followed in NSGA II algorithm and use it to Solve the Multi objective optimisation problem formulated in the previous question.

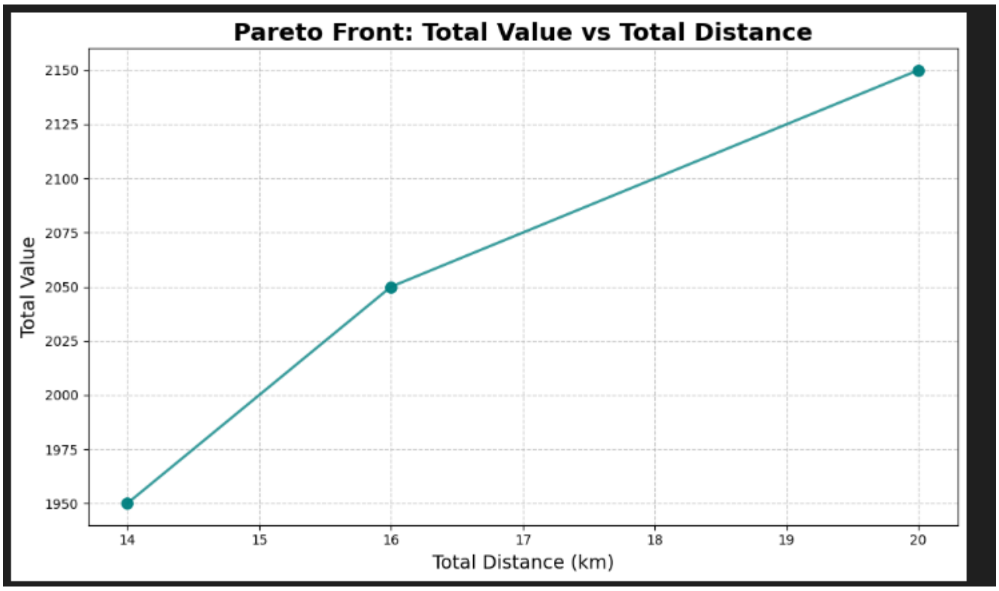
- In order to balance the goals of maximizing delivery value and lowering distance, this method optimizes delivery routes using a Genetic Algorithm (NSGA-II). This is a synopsis:

Setup and Initialization: Using DEAP tools, a population and individuals (city sequences) are created, and a multi-objective fitness function (maximize value, minimize distance) is defined.

Evaluation: Determines the distance and total delivery value for a specified route while imposing restrictions such as the number of cities and city dependencies (visit B only if F is present, for example).

Genetic Operations: NSGA-II evolves a varied Pareto front of optimal solutions by using crossover, mutation, and selection operators designed for routing issues.

Results and Execution: The evolutionary algorithm is run, 100 generations are assessed, and the Pareto front is printed, displaying trade-offs between distance and delivery value for the optimal



PART 2:

1.Write the logistic regression equation relating independent variables to dependent variable y

$P(y=1|X)=1+e-(\beta0+\beta1x1+\beta2x2+\ldots+\beta nxn)1$ is the equation that expresses the relationship between the independent variables (features) and the dependent variable (tumor malignancy). The probability that the tumor is malignant is represented by the equation $P(y = 1 \mid X)$, which also makes use of the data from X.The logistic regression used in the equation shows the characteristics from 0 to 1 likelihood. The log-odds of a tumor being malignant while the other fertaures stay at zero are shown using β0, an intercept term.The coefficients that reflect the independent variable independently are denoted by the terms β1, β2,…, βn, and are expressed as x1, x2,…, xn. The aforementioned coefficients show how each characteristic may impact a tumor's likelihood of being malignant. During the training phase, the logistic regression model forecasts the coefficients.

2.Load the Breast Cancer Dataset and perform an exploration analysis of the data. Is there any class imbalance?

Overview
I used the provided dataset, which included thirty numerical characteristics and one binary as the target variable, in part 2. After loading the dataset into the Python environment, the additional required libraries were added.

EDA
In order to determine the feature distribution, relationships between the variables, outlier detection, and class imbalance, EDA techniques are performed to the previously imported dataset in this part. I started by looking for missing values in the dataset and getting a thorough explanation of it using summary statistics like mean, median, standard deviation, and quartile values.The contributions of the s "mean radius," "mean texture," "mean smoothness," "mean symmetry," and "mean fractal dimension" are presented in Fig. 1.

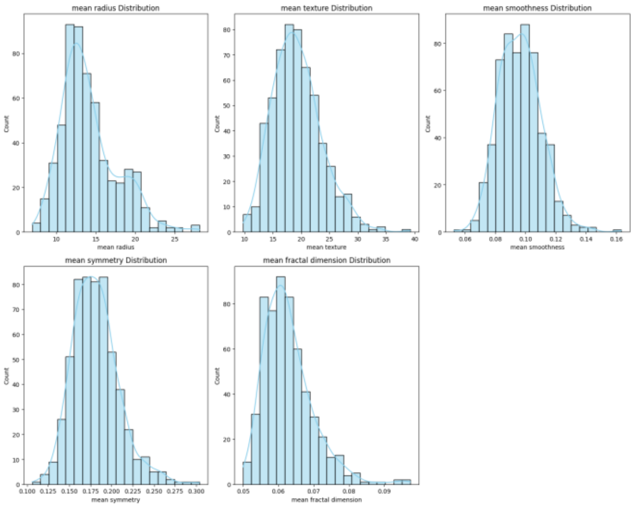


Fig.1

Additionally,I created boxplots for the visualization of the outliers in the above mentioned features as it can be seen in Fig.2.

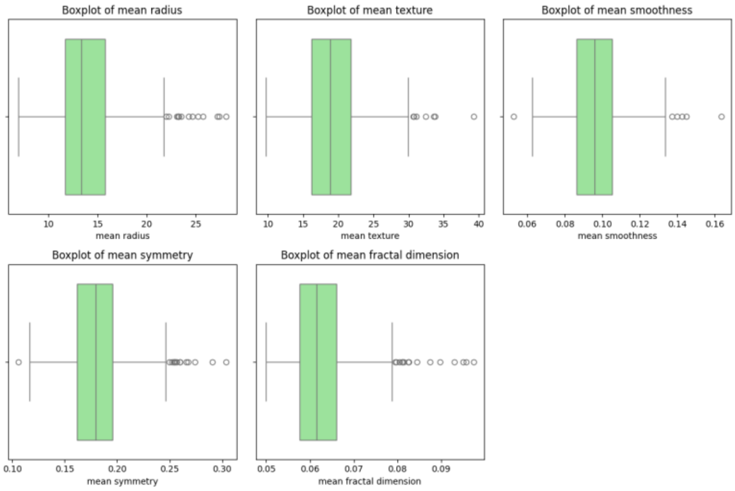


Fig.2

To monitor the relationship among varibales while crerating a model is vital for which a correlation matrix was created using different colour gradients for different correlation strength. It can also be used to keep the multicollinearity under check as it can be seen in fig.3.

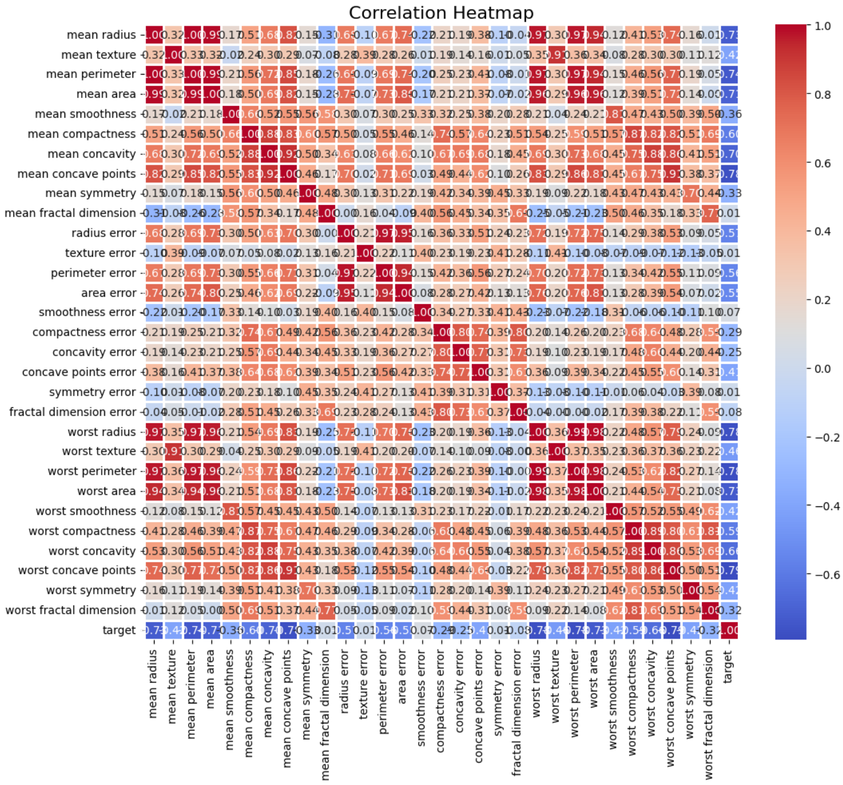


Fig 3

From the correlation matrix,I selected the varibales but to check the pairwise relation of these with the target variable Pairplots were created as shown in fig 4 recognizing the patterns and the interactions of the same.

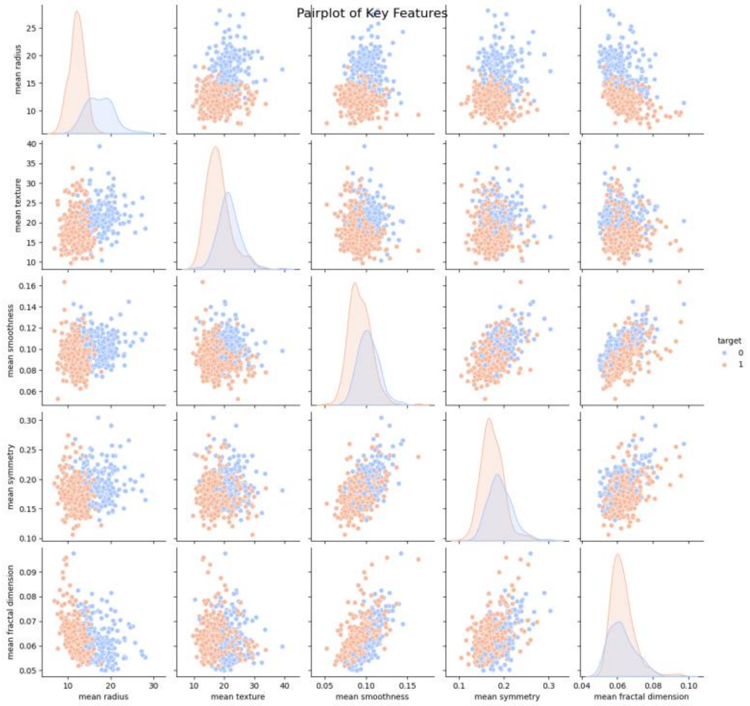


Fig.4

Finally,it was vital to check the class imbalance.For the determination of benign (0) and malignant (1) cases the target variable also refered to as Target was used.Upon examination,slight imbalance was concluded with with 357 malignant cases and 212 benign cases.The count plot (fig 5) was used as an evidence for the same visualizing a higher occurrence of malignant tumors in the dataset.

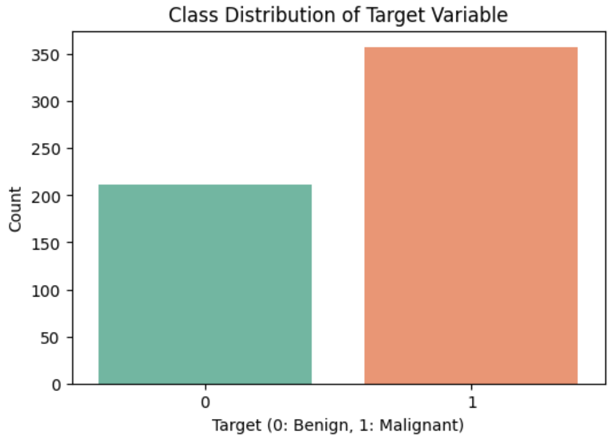


Fig.5

3.Use both independent variables and software to compute the estimated model equation by

I. Divide the data into testing and training sets.
Model Construction and Assessment
Splitting, model training, evaluation, and interpretation are some of the components that make up the model building process.
—Model training and splitting
The independent variables (features) and the dependent variable (goal) comprised the two primary components of the dataset.Thirty traits made up the independent variables, while the target variable determines whether the tumor is benign (0) or malignant (1). For proper model building in this instance, the next step is to divide the data into training and testing portions. For this, I've thought of an 8:2 ratio.
Numerous sophisticated machine learning liberaries are employed to train the logistic regression model utilizing predetermined training data. The remaining data split is used to test the model, in which

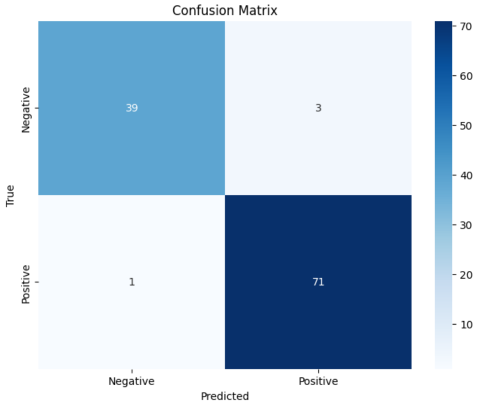
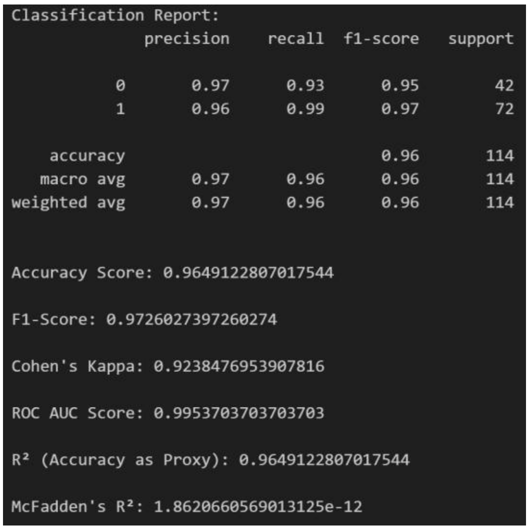


Fig.6

a) Classification Report: This offers a far more thorough examination of the model and comprises:
o Precision: If the malignant instances found are, in fact, malignant cases.

o Recall (Sensitivity): The capacity to identify malignant cases among all cases.
o F1-Score: This represents the harmonomic mean of 0.97 for both precision and recall. A high score denotes an exceptional model since it shows a perfect balance between the two.
o Accuracy: This score of 96.49% indicates the proportion of cases that were correctly identified overall.It also demonstrates how well the model classified the variables.
o A propotional agreement between the expected and actual class labels is indicated by a Cohen's Kappa:score of 0.92.

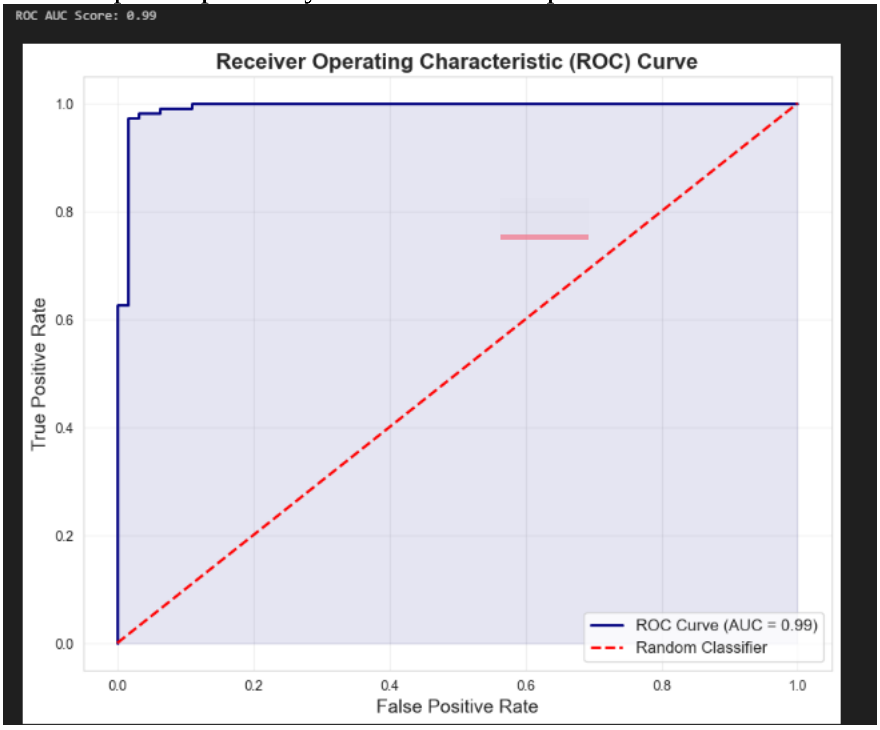


```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.93      0.95        42
           1       0.96      0.99      0.97        72

    accuracy                           0.96       114
   macro avg       0.97      0.96      0.96       114
weighted avg       0.97      0.96      0.96       114


Accuracy Score: 0.9649122807017544

F1-Score: 0.9726027397260274

Cohen's Kappa: 0.9238476953907816

ROC AUC Score: 0.9953703703703703

R² (Accuracy as Proxy): 0.9649122807017544

McFadden's R²: 1.8620660569013125e-12
```

Fig.7

III.Present interpretation of your model

- Interpretation
- With a score of 0.995, as shown in, I produced a ROC curve for this section that shows outstanding performance in predicting the likelihood of benign and malignant tumors. Though it might have seemed a little odd, the model's unusually strong performance in keeping the classes apart is primarily to blame for this problem.



```
ROC AUC Score: 0.99
```

Fig.8

4.Use α = 0.05 to determine whether each of the independent variables is significant.

Statistical testing was used to determine the significance of the independent variables, yielding the p-values for each one.The variables with p-values less than 0.05 were deemed statistically significant, as shown by the specified value of α=0.05. Multicollinearity was taken into account using the Variance Inflation Factor (VIF) prior to model fitting. Each variable had its own VIF value, and variables with more than 10 were eliminated in order to reduce multicollinearity.

The model was refitted with variables with a CIF value less than 10 following the multicollinearity check. The model's p-values were then determined based on their statistical significance, and it was discovered that the intercept had the sole significant feature following the other features, with a p-value of 9.725701×10−8.

5.What is the estimated odds ratio for five variables of your choice? Interpret it.

I calculated the odds for the five variables listed below in order to interpret the regression model:The coefficient for the variable "Mean Radius" was 1.50, yielding a ratio of 4.51. This indicates that a one-unit increase in tumor size is proportional to a 4.51-fold rise in tumor size, resulting in greater mean radii.

The coefficient for the variable "Mean Radius" was 0.512435, yielding a ratio of 1.669351. This suggests that an increase of one unit raises the likelihood of malignancy by roughly 1.67 times, indicating greater tumor diversity.
—The coefficient for the variable "Radius Error" was 0.279208, yielding a ratio of 1.322083.Accordingly, tumors with greater worst radii—roughly 1.60 times larger—have higher

6.A patient's tumor has the following feature values (use mean radius, mean texture, mean smoothness, and mean symmetry):

14.5 is the mean radius.
18.0 is the mean texture.
0.095 is the mean smoothness.
Symmetry mean = 0.180
Determine whether this tumor is likely to be benign or malignant using the training model. Display your work and describe the outcome.
Using the provided metrics, determine the likelihood that the patient's tumor is benign or malignant.The patient's data was first examined for any anomalies or missing values; if any were discovered, they were filled in with 0 because the model does not allow null values.The patient's data was then run through the trained logistic model, which was trained on the first stage to determine the likelihood that the tumor would be malignant. According to the model, the tumor is malignant. This forecast is predicated on characteristics that are related to

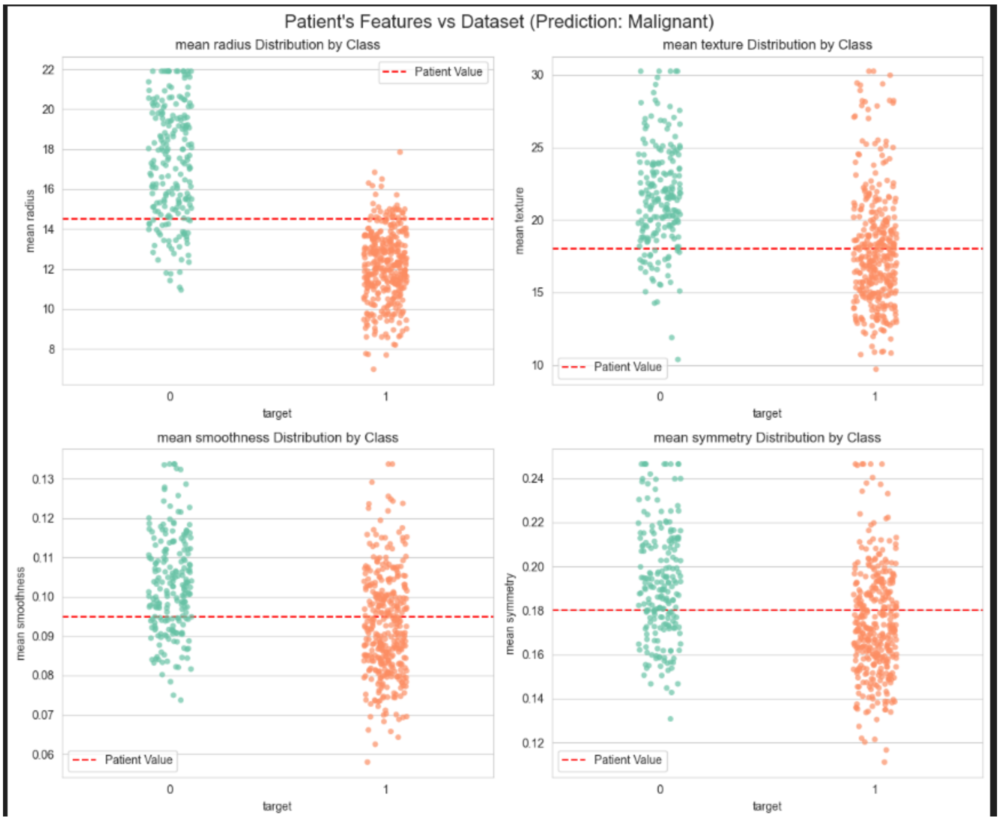
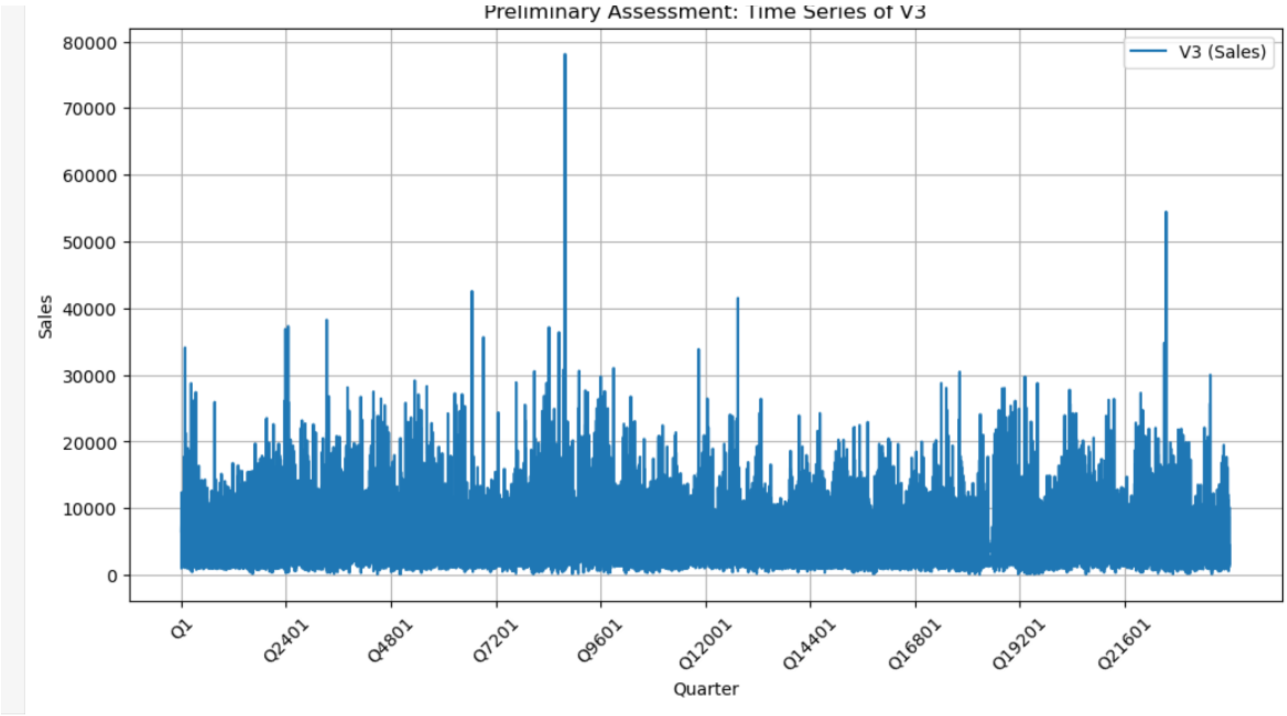


Fig.9

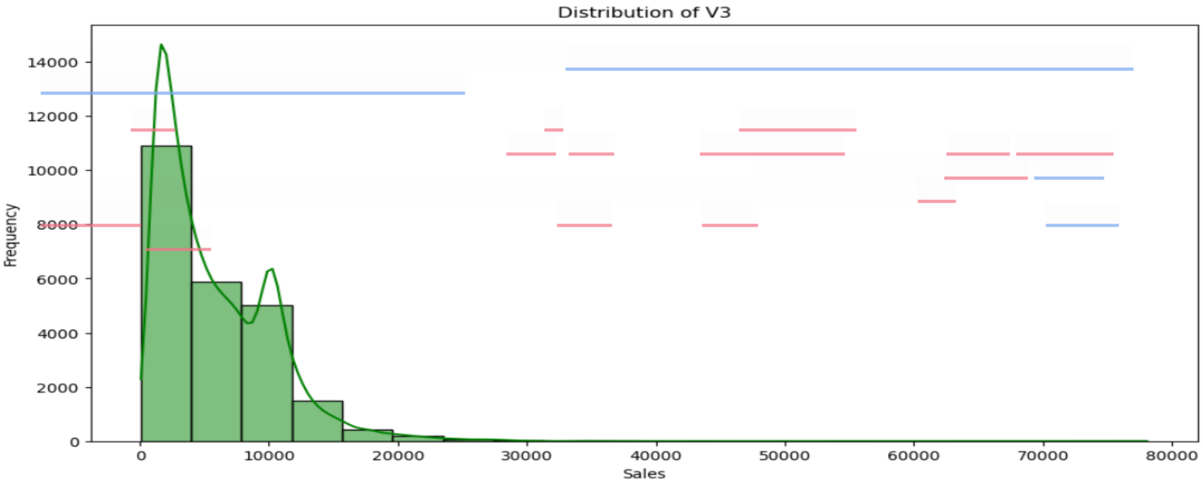
PART 3: Time Series

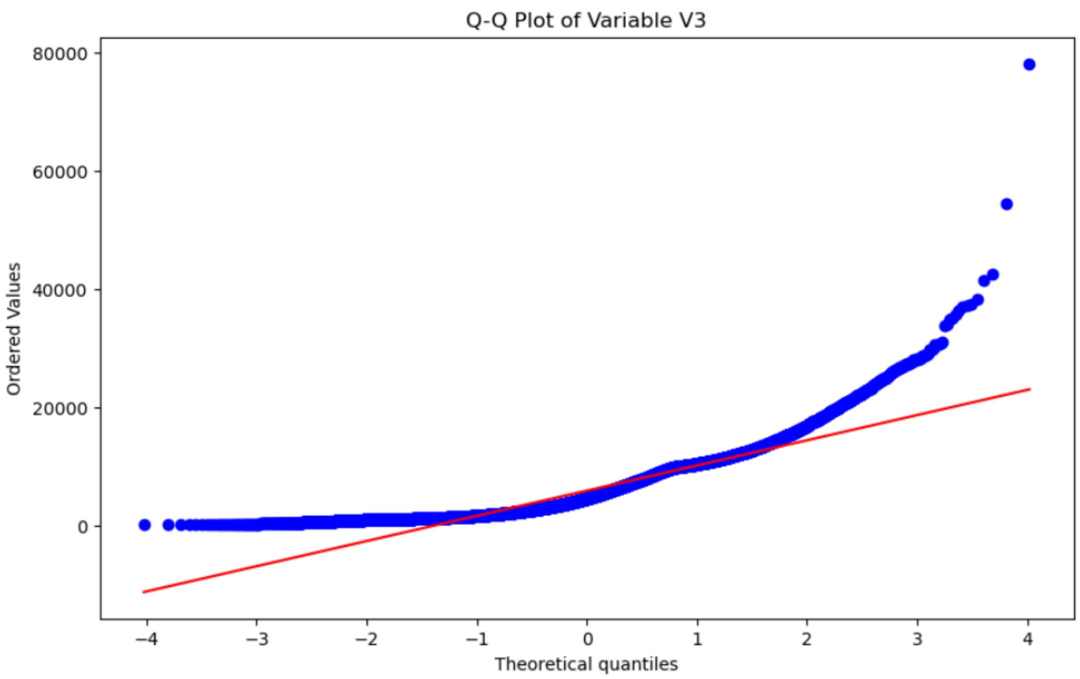
1. Perform a preliminary assessment of the nature and components of the raw time series, using visualisations as appropriate.

- In this part Important Libraries and Dataset is imported and I have have described the dataset inorder to know the nature of variables in data. And As per Requirement given in TABA briefing , I have targeted variable V3 because my roll number ends with 2. After Describing dataset I have noticed that V1 has different quarters of historical sales of any specific product.So I have imputed Quarter= V1 and Target Variable = V3. Then I have plotted Time series pattern of V3. It is clearly visible in below figure.



- Then I have applied a script to check missing values and duplicate values. But there is not Missing, Duplicate value in dataset and checked distribution of V3 to work on further analysis. And displayed Q-Q plot and Shapiro-Wilk Test to check if data is normally distributed and normality is fine in data. I have noticed That data is normally distributed and normalized properly. Below figure and values(static and P-value) are verification of that.

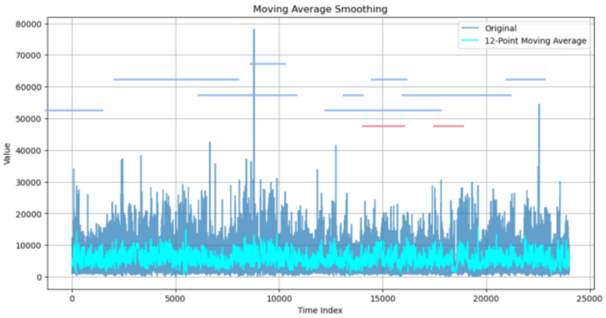


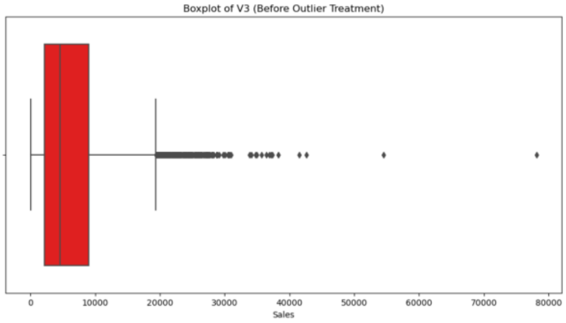


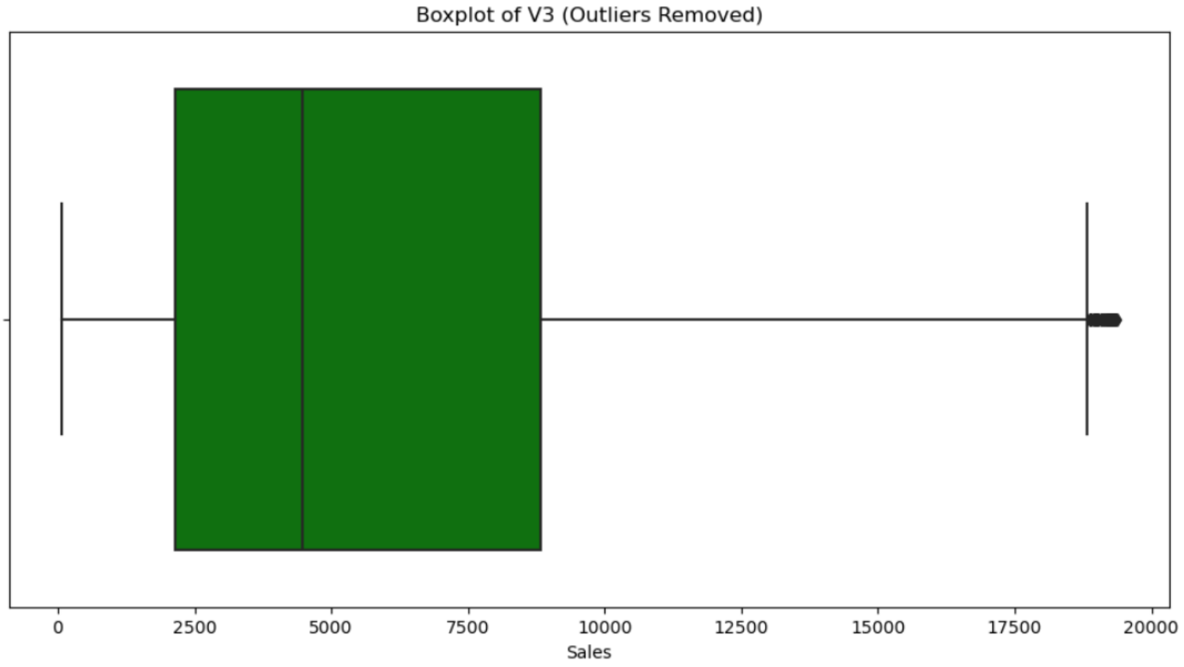
```
Shapiro-Wilk test statistic: 0.8775734349730997, p-value: 2.2978629880701606e-85
```

- In order to check seasonality, Noise, Outliar I have framed Moving Average smoothing. I have noticed that data has some outliars and little noise as well. And to check and verify outliars further I have implemented Boxplot of outliars as well and found outliars then I have Performed IQR test by taking Q1= 0.25, Q3= 0.75 quantile and reduced function of these Q3 - Q1 using lower and upper bound bands . Which reduced Outliars and make data more smoothen.

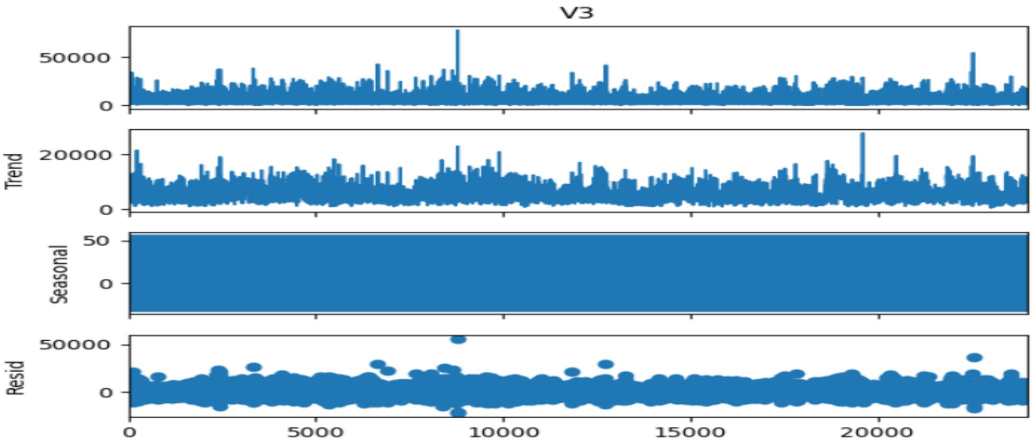
Below Figures shows outliars before and after Handling by IQR methods.







- To Understand pattern after outliar handling and improve forecasting I have visualized Seasonality decomposition chart. Which shows complete noisy trend in ‘V3’.



- Then I have applied Augmented Dickey Filler test (ADF) test to check if data is stationary. P-value must be less than 0.05 which signifies data is stationary and my values are as below, which clearly indicate data is statistically stationary.

- ADF Test Results:
- ADF Statistic: -15.362434098721376
- p-value: 3.597616190802098e-28
- Data is stationary.

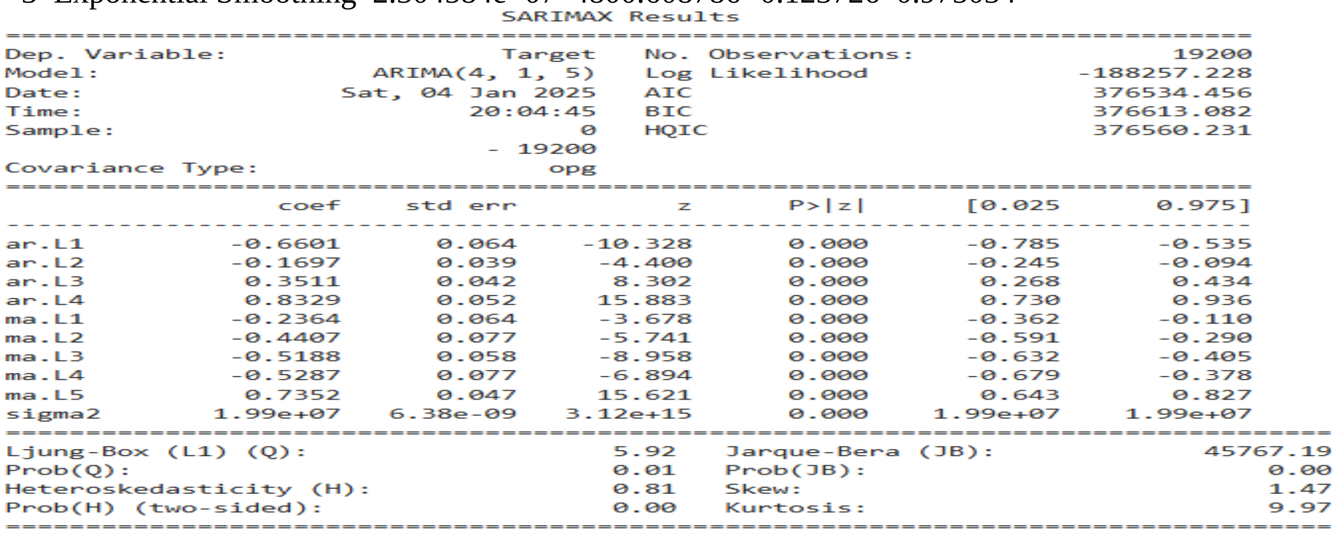
- Then Applied Box Cox Transformation using parameter alpha. It should be near to 0 and it balances ske wness and non-constance variances in dataset and make data ready for further analysis. And my valu e is 0.16 .

- Then I have performed Feature engineering by taking test size 0.8 to make model ready to fit in Time se ries Prediction models.

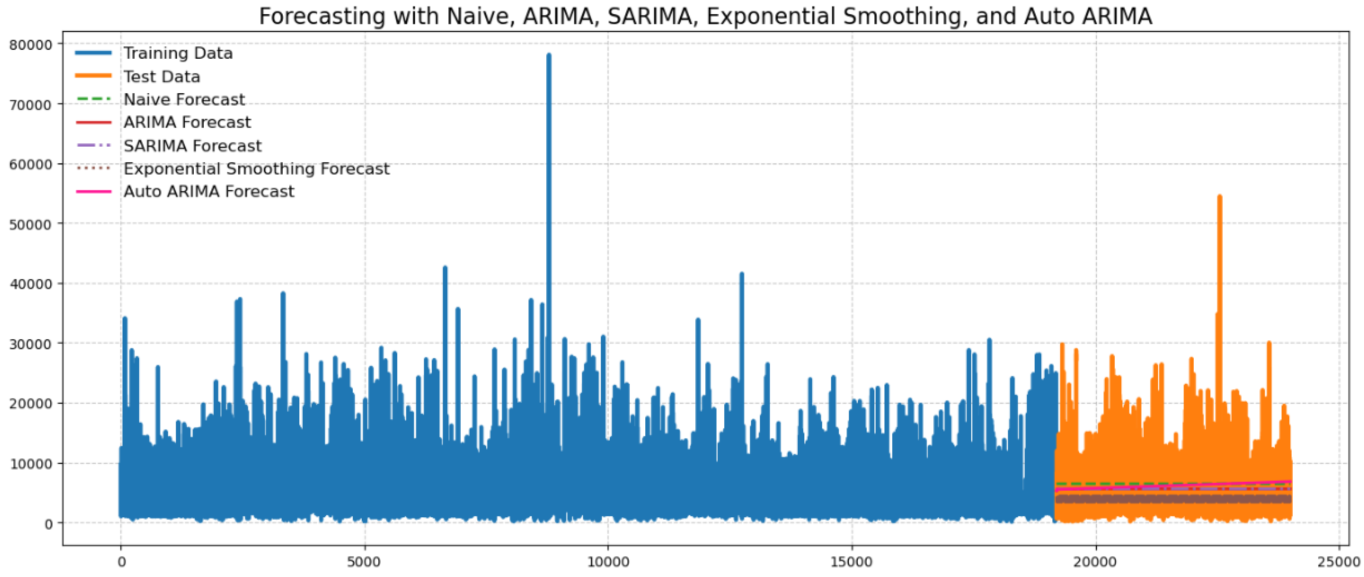
2. Estimate suitable time series models from the categories covered in class and assess the adequacy of each model.

- I have chosen to apply Auto Arima, Naïve, SARIMAX, Smoothening exponential method and compared all these 4 models in order to decide which is performing best. Selected ARIMA / SARIMA because of seasonal component handling feature and organized best model of auto arima in both. Chosen SME because of its focused observation quality and seasonality handling feature. And Naïve baised is selected because it acts as baseline for comparison, forecasting solely on last observed value.
- Below is the summary of all models, which clearly shows Naïve predicting poorest accuracy, ARIMA, SARIMA is somewhat similar but ARIMA is quite better and SME has lowest MAPE which made its performance poor. SARIMA is chosen for final prediction:-

```
Model          MSE        RMSE       R²      MAPE
0              Naive  2.115278e+07  4599.215032 -0.031420  1.602317
1              ARIMA  2.050978e+07  4528.771938 -0.000066  1.355398
2             SARIMA  2.050978e+07  4528.771938 -0.000066  1.355398
3  Exponential Smoothing  2.304584e+07  4800.608786 -0.123726  0.975034
```



```
                               SARIMAX Results
==============================================================================
Dep. Variable:                 Target   No. Observations:                19200
Model:                 ARIMA(4, 1, 5)   Log Likelihood             -188257.228
Date:                Sat, 04 Jan 2025   AIC                         376534.456
Time:                        20:04:45   BIC                         376613.082
Sample:                             0   HQIC                        376560.231
                              - 19200
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6601      0.064    -10.328      0.000      -0.785      -0.535
ar.L2         -0.1697      0.039     -4.400      0.000      -0.245      -0.094
ar.L3          0.3511      0.042      8.302      0.000       0.268       0.434
ar.L4          0.8329      0.052     15.883      0.000       0.730       0.936
ma.L1         -0.2364      0.064     -3.678      0.000      -0.362      -0.110
ma.L2         -0.4407      0.077     -5.741      0.000      -0.591      -0.290
ma.L3         -0.5188      0.058     -8.958      0.000      -0.632      -0.405
ma.L4         -0.5287      0.077     -6.894      0.000      -0.679      -0.378
ma.L5          0.7352      0.047     15.621      0.000       0.643       0.827
sigma2       1.99e+07   6.38e-09   3.12e+15      0.000    1.99e+07    1.99e+07
===================================================================================
Ljung-Box (L1) (Q):                   5.92   Jarque-Bera (JB):             45767.19
Prob(Q):                              0.01   Prob(JB):                         0.00
Heteroskedasticity (H):               0.81   Skew:                             1.47
Prob(H) (two-sided):                  0.00   Kurtosis:                         9.97
===================================================================================
```



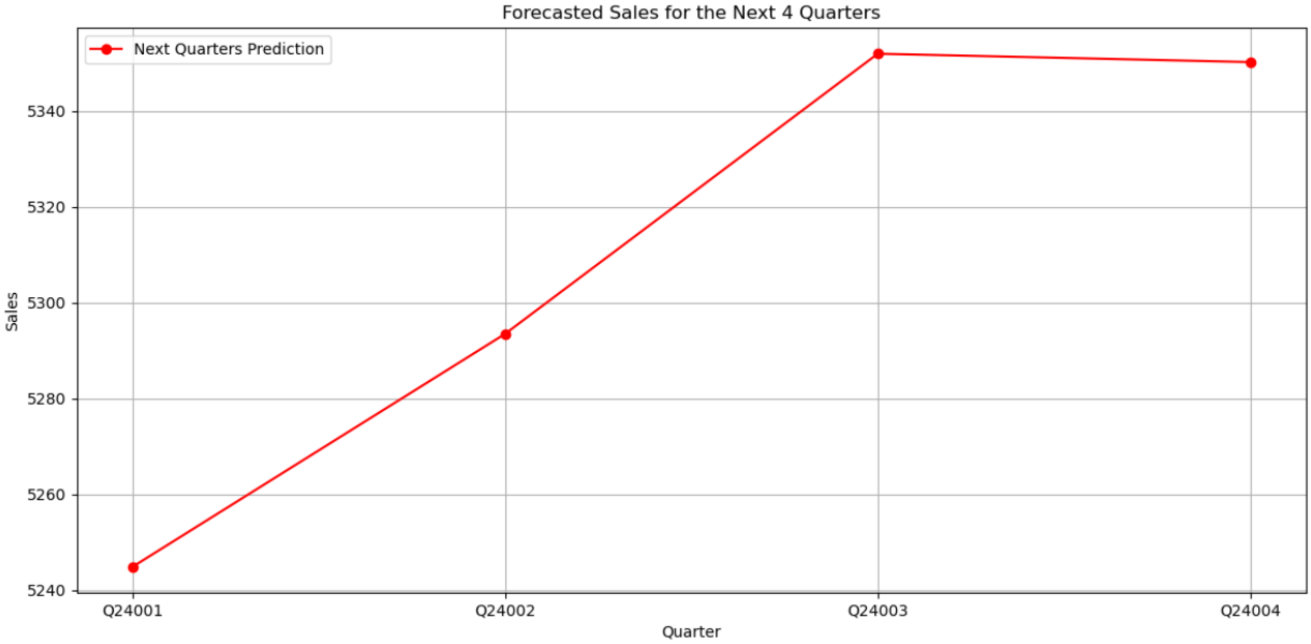
3. Using the best model in the previous question, perform forecasting for the next four quarters.

- I have predicted next 4 quarters using ARIMA model and values comes like below.

```
Quarter  Predicted_Sales
19200  Q24001      5244.878063
19201  Q24002      5293.502970
19202  Q24003      5351.954586
```

```
19203  Q24004      5350.208502
```

- Forecasted Sales for Next 4 Quarters shown in below chart which clearly shows it is constantly increasing till Q3 and after Q3 it decreased.


Forecasted Sales for the Next 4 Quarters
Next Quarters Prediction
Sales
Q24001
Q24002
Q24003
Q24004
Quarter