```
                                        PREDICTIVE MODELLING PTOJECT - BUSINESS
REPORT
```

Introduction :-

Problem Statement 1:- Linear Regression

Question 1:- Read the data and do exploratory data analysis. Describe the data briefly. (Check the Data types, shape, EDA, 5 point summary). Perform Univariate, Bivariate Analysis, Multivariate Analysis.

Introduction : This report presents an analysis of the Dataset ["compactiv.xlsx"], focusing on the data's
basic characteristics, including data summary, missing values, and duplicate values. The objective of this analysis is to gain insights into the dataset and prepare a Linear Model.

Basic information of the Data.

0 lread - int64

1 lwrite - int64

2 scall - int64

3 sread - int64

4 swrite - int64

5 fork - float64

6 exec - float64

7 rchar - float64

8 wchar - float64

9 pgout - float64

10 ppgout - float64

11 pgfree - float64

12 pgscan - float64

13 atch - float64

14 pgin - float64

15 ppgin - loat64

16 pflt - float64

17 vflt - float64

18 runqsz - object

19 freemem - int64

20 freeswap - int64

21 usr - int64

1. The Dataset has 8192 Rows and 12 Columns.
2. The Dataset has 13 Float datatypes, 8 integer datatypes and 1 object datatype.
3. The Columns 'rchar' and 'wchar' has missing values.

Data Preprocessing:
Based on the initial data exploration, the following preprocessing steps may be considered:
Handling missing values: [Describe how missing values will be addressed]
Removing duplicates: [Explain how duplicates will be handled]
Feature scaling/normalization: [If applicable, mention if features will be scaled or normalized]
This initial analysis provides a foundational understanding of the dataset, highlighting potential areas for further investigation. As the analysis progresses, more insights will be gained, leading to meaningful business outcomes and recommendations.

UniVarient Analysis of Categorical variable:-

Univarient analysis is performed on "runqsz" column as it is the only categorical column in the dataset

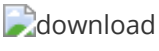
The analysis says that 'NOT_CPU_BOUND' is more dominant and appeared more frequently than 'CPU_BOUND' , so hence the the data is more bound towards 'NOT_CPU_BOUND'.

Bivarient Analysis :-

The Analysis between Two variables is called Bivarient analysis, Heat map is used to perform the Bivarient analysis to check correlation between the variales.

In summary, bivariate analysis using a heatmap in linear regression helps you explore the relationships and correlations between independent variables. It is particularly useful for detecting multicollinearity, which can affect the stability and interpretation of the regression model. By visualizing these relationships, you can make informed decisions about variable selection and model development.

Figure 1



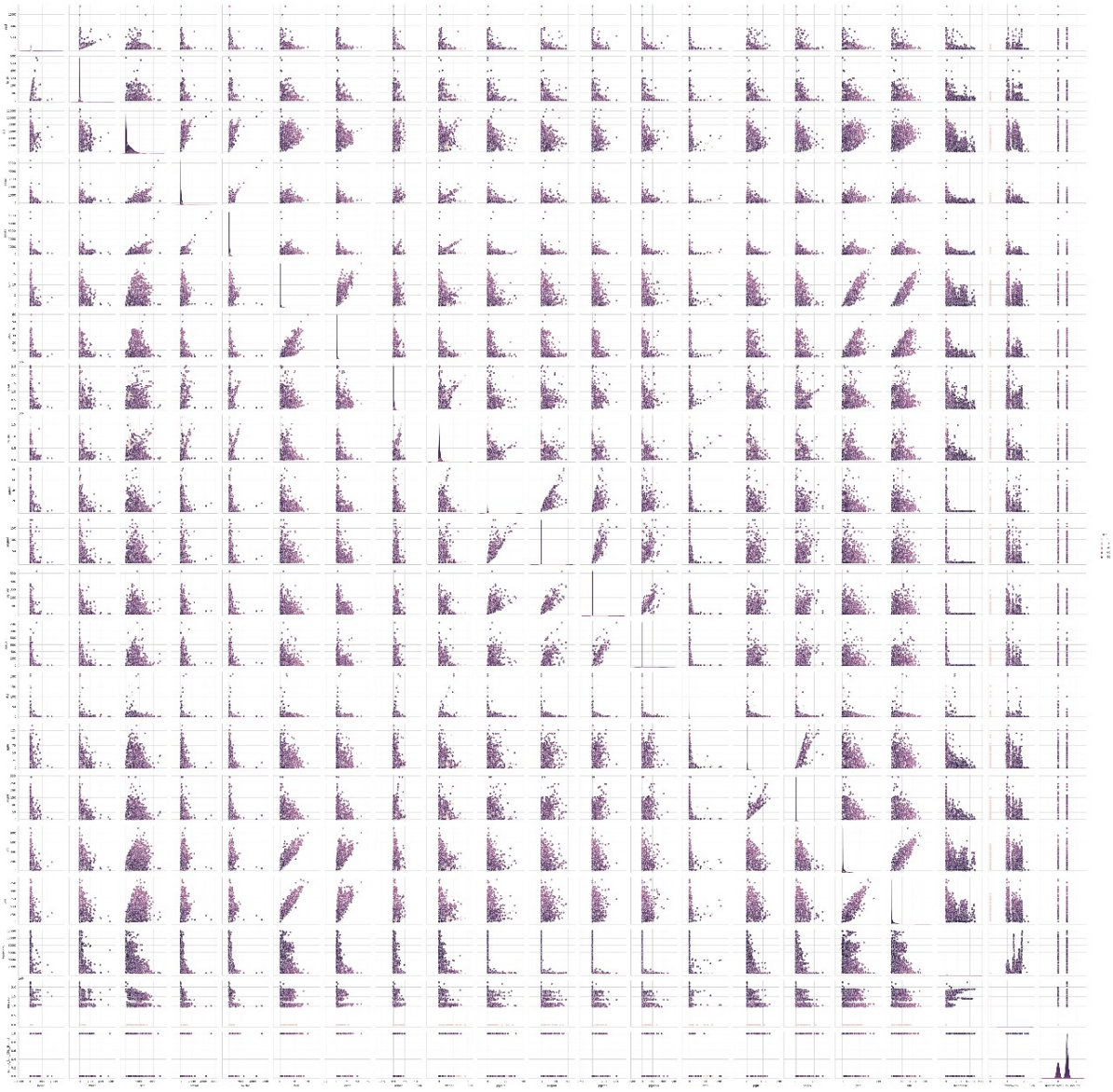
From the Heatmap we found the bivarient analysis bewteen the variables.
There are correlations between many varibles from the heatmap,
The following pair of variables have corrilation between them
1)Iread and Iwrite, 2) Scall and Sread, 3) scall and Swrite, 4) Scall and Vfit, 5) Sread and Swrite, 6) sread and rchar, 7) Sread and Vfit, 8) fork and exec, 9) fork and pfit, 10) fork and vfit, 11) exec and pfit, 12) exec and vfit, 13) rchar and wchar, 14) pgout and ppgout, 15) pgout and pgfree, 16) pgout

and pgscan, 17) ppgout and pgfree, 18) ppgout and pgscan, 19) ppgout and ppgin, 20)pgfree and pgscan, 21) pgfree and pgin, 22)pgfree and ppgin 23)pgscan adn pgin, 24)pgscan and ppgin, 25)pgin and ppgin, 26)pfit and fork, 27)pfit and exec 28)pfit and vfit, 29)vfit and fork, 30)vfit and exec 31)freeman and freeswap, 32)freeswap and usr

Multivarient Analysis:-

The Analysis between Three or more variables is called multi-varient analysis, Pair plot is used to perform MultiVarient analysis between the variables, to check relation and multicollinearity between the variables.

Figure 2



Question 1.2 :- Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of creating new features if required. Also check for outliers and duplicates if there.

Response:-

Null Value Treatment :- Treating null (missing) values is an essential step in data preprocessing. The appropriate method for handling missing values depends on the nature of the data and the specific problem you are working on.

We should Be cautious about the choice of imputation method, as it can affect the analysis and model performance.

Null Values can be treated by Replacing The Null values with mean, median, mode, or predictive modeling of that particular Column.

In The analysis, There are null values in "rchar" and "Wchar"

104 and 15 null values are present respectively in "rchar" and "wchar"

Zero Value Treatment:- Treating zero values is an important aspect of data preprocessing, as zero values can affect statistical analysis, machine learning models, and the interpretation of the data. The approach to handling zero values depends on the context and the nature of the data, for example zero in the 'AGE' column is an anamaly

Impute zero values with meaningful estimates. Consider using imputation methods such as mean, median, or mode imputation for numerical data.

Here, In the Dataset The columns lread, lwrite, fork, exec, pgout, ppgout, pgfree, pgscan, atch, pgin, ppgin, pflt, usr have zero values.
In this case we cannot remove the zeros as it can mean that there was no action at that time or no usage of CPU's at that time. So, in my opinion they should be ignored.

Check for Duplicates

Duplicate values in a dataset are identical or nearly identical rows that appear more than once. Handling duplicate values is important in data preprocessing to avoid issues with data integrity and analysis.

In some cases, duplicate rows are meaningful, and you may want to retain them. For instance, in transactional data, multiple entries for the same transaction may be valid.

Duplicates can be treated By dropping them by using "drop_duplicates()" function

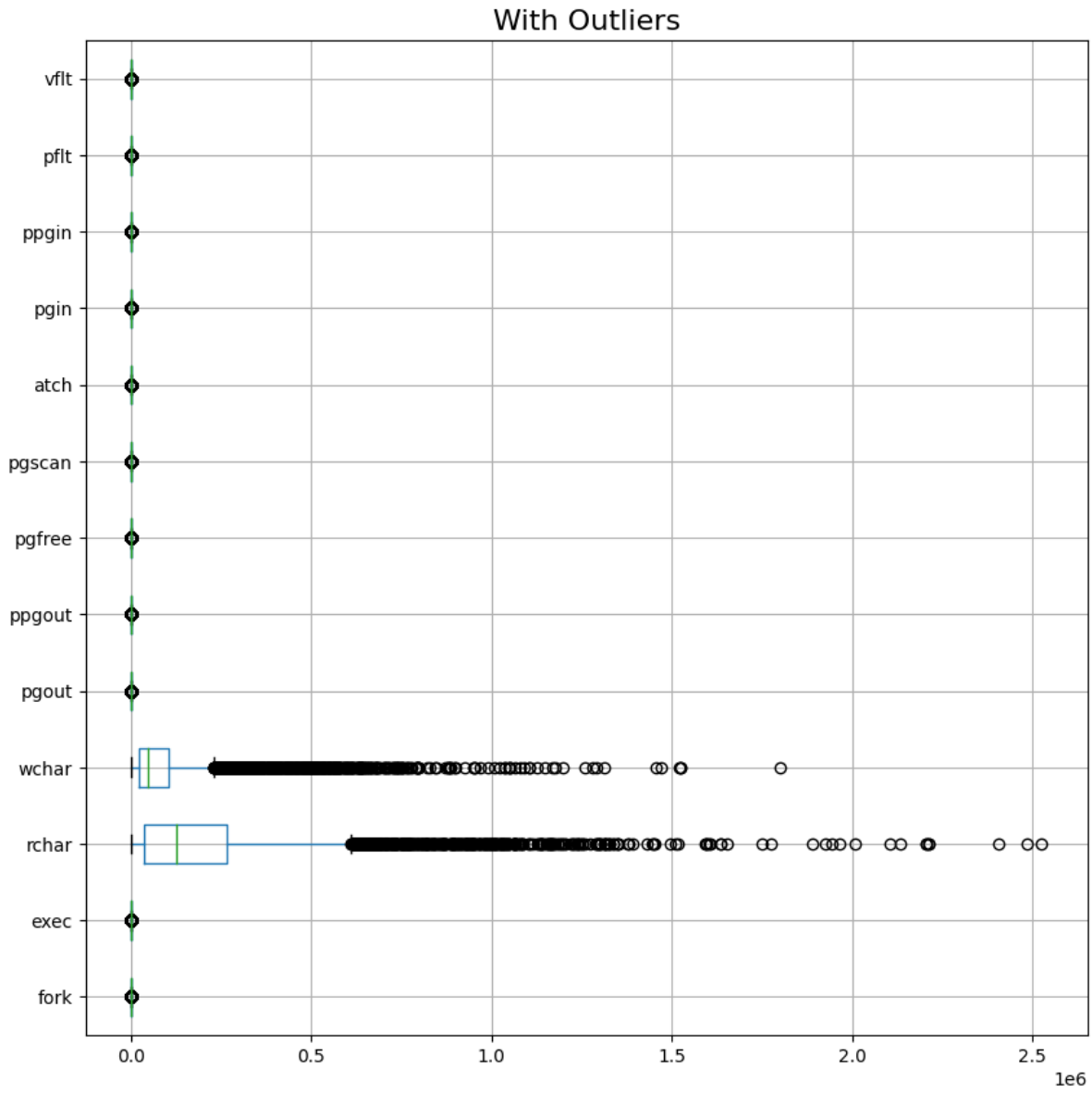
There are no Duplicate rows or columns in the dataset

Outlier Treatment:-

Introduction:- Outliers are data points that significantly deviate from the majority of the data in a dataset. Treating outliers is an essential step in data preprocessing to ensure that they don't unduly influence statistical analyses or machine learning models.

- The choice of treatment depends on the nature of the data and the objectives of your analysis. Outliers can be treated in various ways.
- **Removal**: Remove the outlier data points from the dataset.
- **Transformation**: Apply mathematical transformations (e.g., log, square root) to make the data less sensitive to outliers.

Figure 3



From the above Box plots we can clearly see that there are outliers present in the columns "wchar" and "rchar", these outliers have to be treated to get proper and right understanding of data which can be used improve business and make predictions.

Figure 4

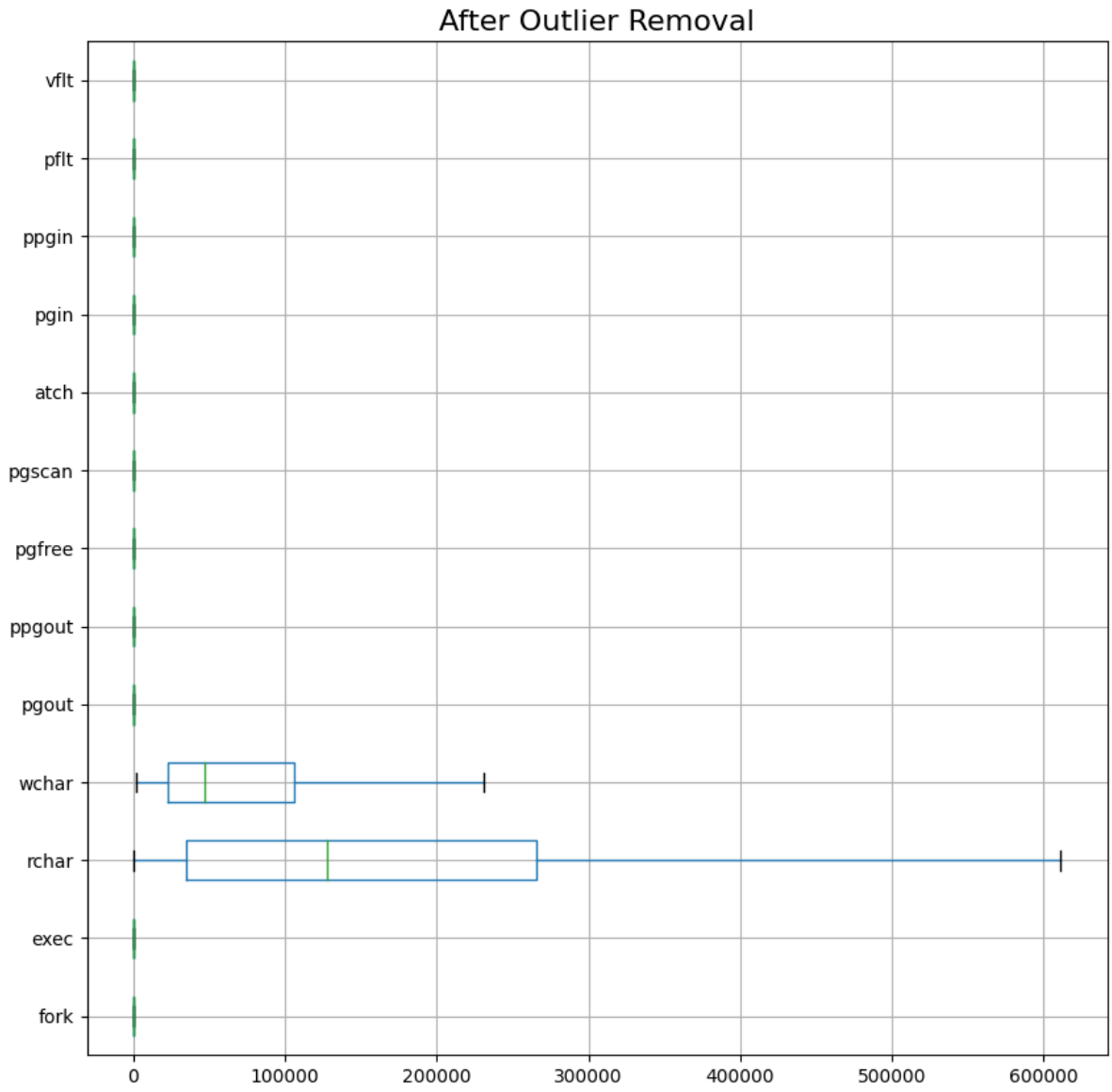


Figure 2 shows that all the outliers are treated and the data is clean without any outliers, we can now proceed with the next steps in the data analysis.

Question 1.3 :- Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.

Response:- The train and test is a machine learning technique which is used to know the performance of the Model, for example if we have 100 observations or records, we take 70 records to train the model and 25 records to test the model to make the predictions.

The following can be inferred from the Model Summary

1. The R-squared (R^2) value in a model summary provides information about the goodness of fit of the model to the data. R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables in the model. It is a value between 0 and 1.

R-square = 1- sum of square error / sum of square total

(or)

R-square = sum of square of reseduals / sum of square total

$R^2 = 0$: The model does not explain any of the variance in the dependent variable. It's a poor fit to the data.

$R^2 = 1$: The model perfectly explains all the variance in the dependent variable. It's an excellent fit to the data.

1)From the model summary we found the value of R-square as 0.626, that means we have ±60% tolerance in the predictions, it also means that after we fit the model and run the variation between the observed and the predicted value is 60%

2)Adjusted R^2 is almost same as the R^2 value so there is not much to be interpreted from the adjusted R^2

3)F-statistics is 478.6 from the output

std-error:- its the mean absolute error, it is a measure of the variability or dispersion of the estimated coefficients (parameters) of the independent variables in a regression model.

Smaller standard errors generally indicate more precise estimates, making it more likely that the associated independent variables are statistically significant in explaining the variation in the dependent variable.

confidence Intervals: Standard errors are used to calculate confidence intervals around the coefficient estimates. A confidence interval provides a range of values within which the true population parameter is likely to fall. The standard error is used to determine the width of the confidence interval.

To find the confidence interval they plot multiple number of samplings and try multiple combinatins and take multiple means and find the coefficient and to fing the range of confidence interval

From the model summary,The confidence interval rangers from 41.564(0.25 quartile) to 44.585(0.95 quartile)

Hypothesis Testing :- The p-value (denoted as p|t| or p-value for the t-statistic) associated with each coefficient's t-statistic is used to determine whether you should reject the null hypothesis. Here's what different p-values mean in hypothesis testing for linear regression.

Variation Inflation Factor:-

The Variance Inflation Factor (VIF) is a statistical measure used to assess multicollinearity in a regression model. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other Calculation of VIF: VIF is calculated for each independent variable in a multiple regression model. For each variable, it is determined by regressing that variable against all the other independent variables. The formula for VIF is:

$VIF = 1 / (1 - R^2)$

Where:

VIF: Variance Inflation Factor
R^2: The coefficient of determination for the regression of the variable against all other variables.

Interpretation of VIF:

1)A VIF of 1 indicates no multicollinearity. This means that the variable is not correlated with any other independent variable in the model.
2)A VIF greater than 1 suggests the presence of multicollinearity. The higher the VIF, the more the variable is correlated with other independent variables. Typically, a VIF above 5 or 10 is considered high and indicates a problematic level of multicollinearity.

3)As few predictors have higher VIF value that means that there is multicollinearity in the data.

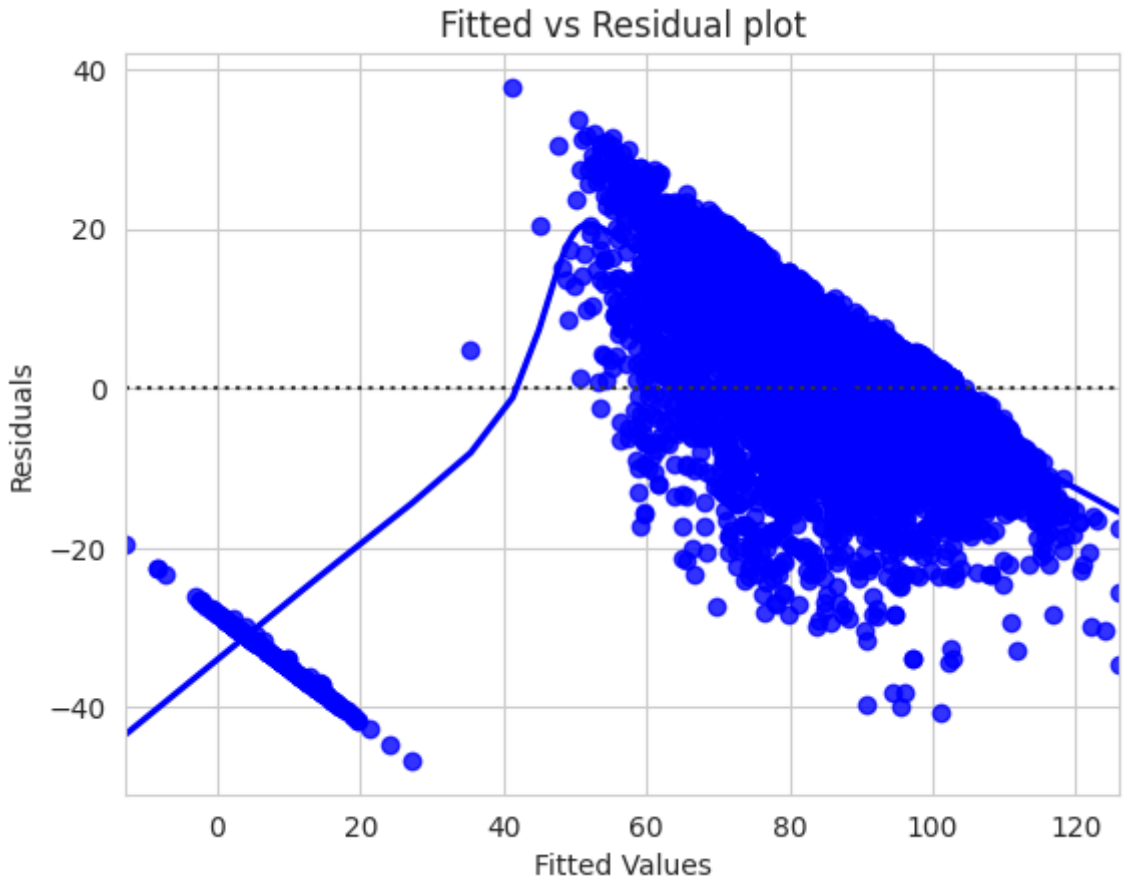
We remove the columns which have higher VIF as they show multicollinearity and drop the ones which have least impact on the R-square and adj R-square.

We have dropped the columns sread, fork, pgout, ppgout, pgin, ppgin, pflt and vflt one after the other to check the Change in the R-square and adjusted r-square, if there is no change in the R-square then we can drop the variables if there is an impact on the R-square then we can't drop the columns as that will impact the regression model and effect the predictions.

Testing the Assumptions of Linear Regression
For Linear Regression, we need to check if the following assumptions hold:-
1)Linearity
2)Independence
3)Homoscedasticity
4)Normality of error terms
5)No strong Multicollinearity

1)Linearity and Independence

Figure 5



3)Normality of error terms:

The p-value=1.3342254403286478e-15, here as p-value is > 0.05, the residuals are normal according to shapiro test

4)Homoscedasticity:-

The p-value = 0.9643898365671506, Since the p0-value is > 0.05 we can say that the residulas are homoscendastic.

The Equation of the Linear Regression Model:

```
usr = -0.028942338273987726 + 0.03176924339553025 * ( lwrite ) +
0.0029652381838927984 * ( scall ) +  0.0068350456294324225 * ( sread ) +
-0.010076545779152313 * ( swrite ) +  -2.506067162084726 * ( fork ) +
0.19964318281069857 * ( rchar ) +  -4.818739105143903e-06 * ( wchar ) +
1.5581391443999104e-05 * ( pgout ) +  -0.7952303600579108 * ( pgscan ) +
0.24003918278446518 * ( atch ) +  5.232789891809855e-15 * ( pflt ) +
3.2450895592571736 * ( vflt ) +  0.7194454541042051 * ( freemem ) +
-0.3283646666031982 * ( freeswap ) +  -0.07205965306945604 * ( pflt )
```

Question 1.4 Inference: Basis on these predictions, what are the business insights and recommendations.
Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Response:-From the predictions above, the basic information we found is that,
The variables "runqsz", "atch" and "fork" has the highest coefficient values so that means that they haave a huge impact on the predictions.
The r-square and adj r-square both of them are almost same so that means that there is not much to infer from them, as the rsquare value is 0.974 it means that the predictors have + or - 97% varience.

Problem Statement 2 :- Logistic Regression, LDA and CART

You are a statistician at the Republic of Indonesia Ministry of Health and you are provided with a data of 1473 females collected from a Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of the survey.

The problem is to predict do/don't they use a contraceptive method of choice based on their demographic and socio-economic characteristics.

Introduction :- This report presents an analysis of the Dataset ["compactiv.xlsx"], focusing on the data's
basic characteristics, including data summary, missing values, and duplicate values. The objective of this analysis is to gain insights into the dataset and prepare and perform Logistic Regression, LDA and CART.

Data Preprocessing:
Based on the initial data exploration, the following preprocessing steps may be considered:
Handling missing values: [Describe how missing values will be addressed]
Removing duplicates: [Explain how duplicates will be handled]
Feature scaling/normalization: [If applicable, mention if features will be scaled or normalized]
This initial analysis provides a foundational understanding of the dataset, highlighting potential areas for further investigation. As the analysis progresses, more insights will be gained, leading to meaningful business outcomes and recommendations.

Question 2.1 :- Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, check for duplicates and outliers and write an inference on it. Perform Univariate and Bivariate Analysis and Multivariate Analysis.

Response:-

Basic Information of the Data:-

1. The Shape of the Dataset is (1473, 10), that implies the data has 1473 rows and 10 columns
2. The dataset has 2float datatype, 1 integer datatype and 7 object data types

Wife_age - Float datatype

Wife_education - object datatype

Husband_education - object datatype

No_of_children_born - Float datatype

wife_religion - Object datatype

wife_working - object datatype

Husband_occupation - integer datatype

standard_of_living - object datatype

media_exposure - object datatype

contraceprive_ menthods_used - object datatype

Data Description:-

1. Wife age has a mean of 32.6062 and it ranges from minimum value of 16 years and maximum value of 49 years
2. Number of children born has a mean of 3.254232 and the range is from 0.000 to 16.000 as min and max values respectively
3. Husband Occupation has the mean of 2.137014 and its range is from 1.000 to 4.000

Null Value Treatment :- Treating null (missing) values is an essential step in data preprocessing. The appropriate method for handling missing values depends on the nature of the data and the specific problem you are working on.

We should Be cautious about the choice of imputation method, as it can affect the analysis and model performance.

Null Values can be treated by Replacing The Null values with mean, median, mode, or predictive modeling of that particular Column. "isnull()" function is used to check for null values

There are null values in the columns , 71 null values are there in "wife_age" and 21 null values are present in "no_of_children_born"

Check for Duplicates

Duplicate values in a dataset are identical or nearly identical rows that appear more than once. Handling duplicate values is important in data preprocessing to avoid issues with data integrity and analysis.

In some cases, duplicate rows are meaningful, and you may want to retain them. For instance, in transactional data, multiple entries for the same transaction may be valid.

Duplicates can be treated By dropping them by using "drop_duplicates()" function.

There are 80 duplicate rows present in the dataset.

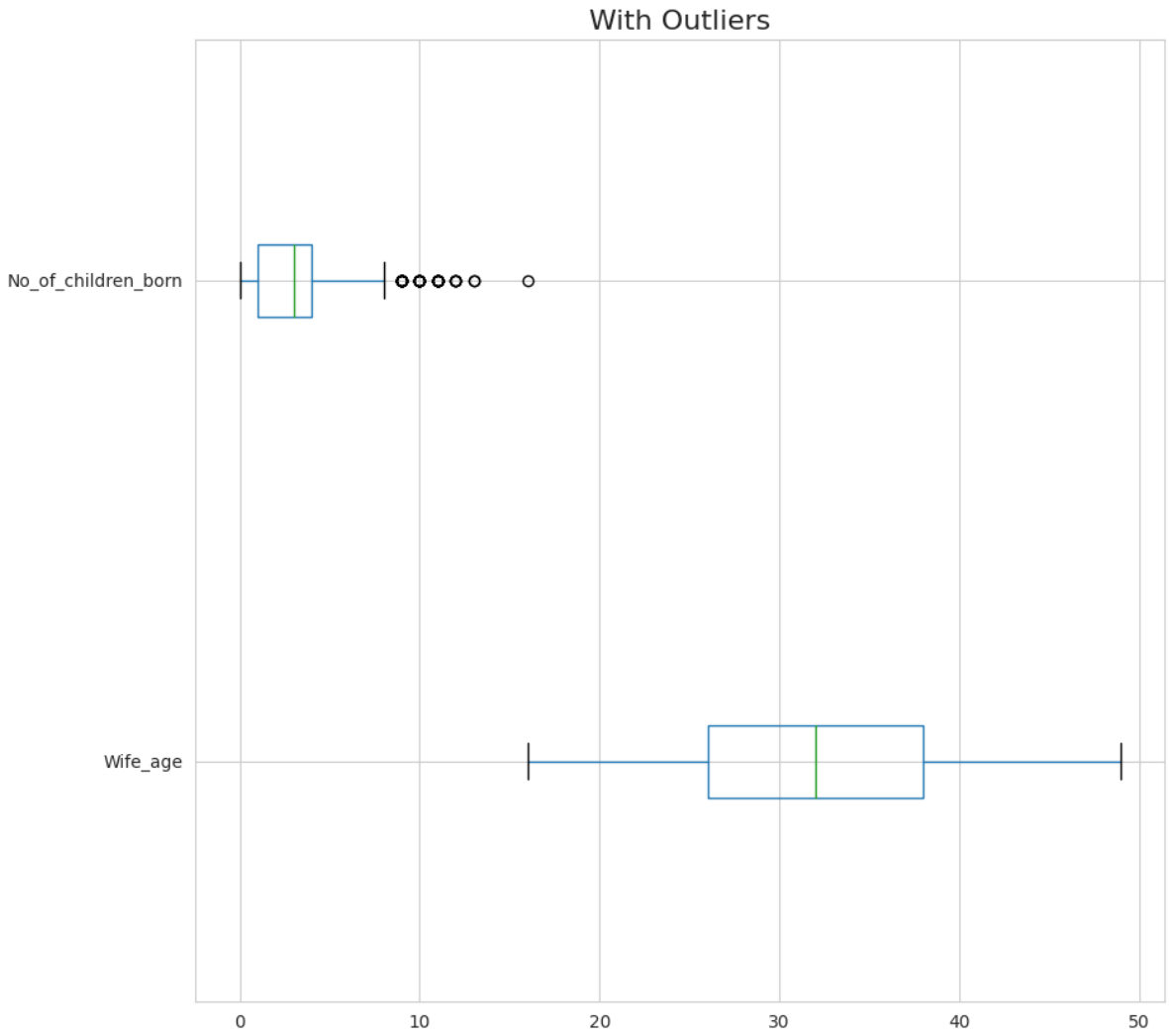
we have removed the duplicates and the cleaned the dataset for further analysis. after removing the duplicates the shape of the dataset is (1393, 10) that is 1393 rows and 10 columns.

Outlier Treatment:-

Introduction:- Outliers are data points that significantly deviate from the majority of the data in a dataset. Treating outliers is an essential step in data preprocessing to ensure that they don't unduly influence statistical analyses or machine learning models.

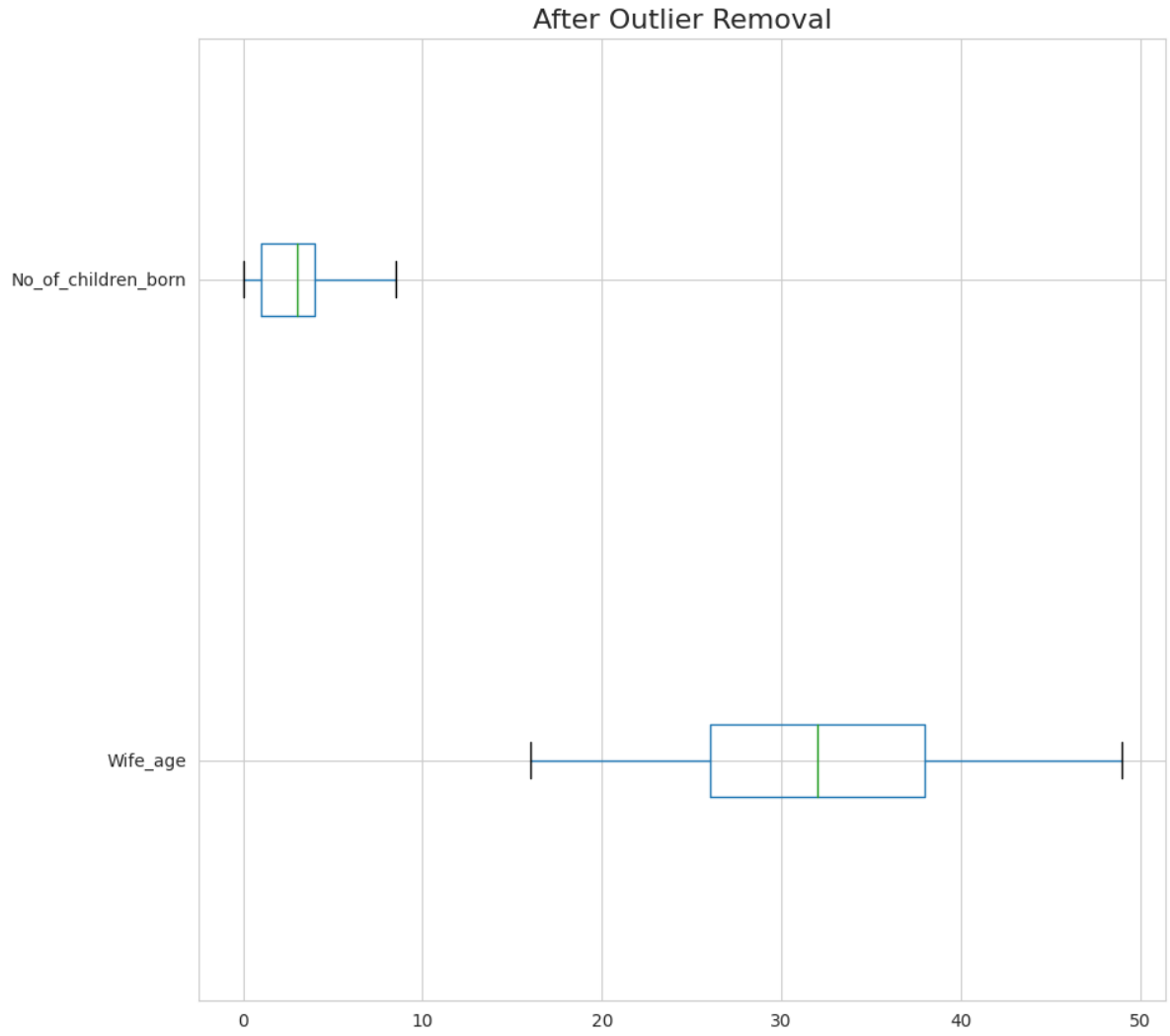
- The choice of treatment depends on the nature of the data and the objectives of your analysis. Outliers can be treated in various ways.
- **Removal**: Remove the outlier data points from the dataset.
- **Transformation**: Apply mathematical transformations (e.g., log, square root) to make the data less sensitive to outliers.

Figure 6



There are outliers in the column no. of children born which have to be treated.

Figure 7



From the boxplot above we can see that the outliers have been treated successfully.

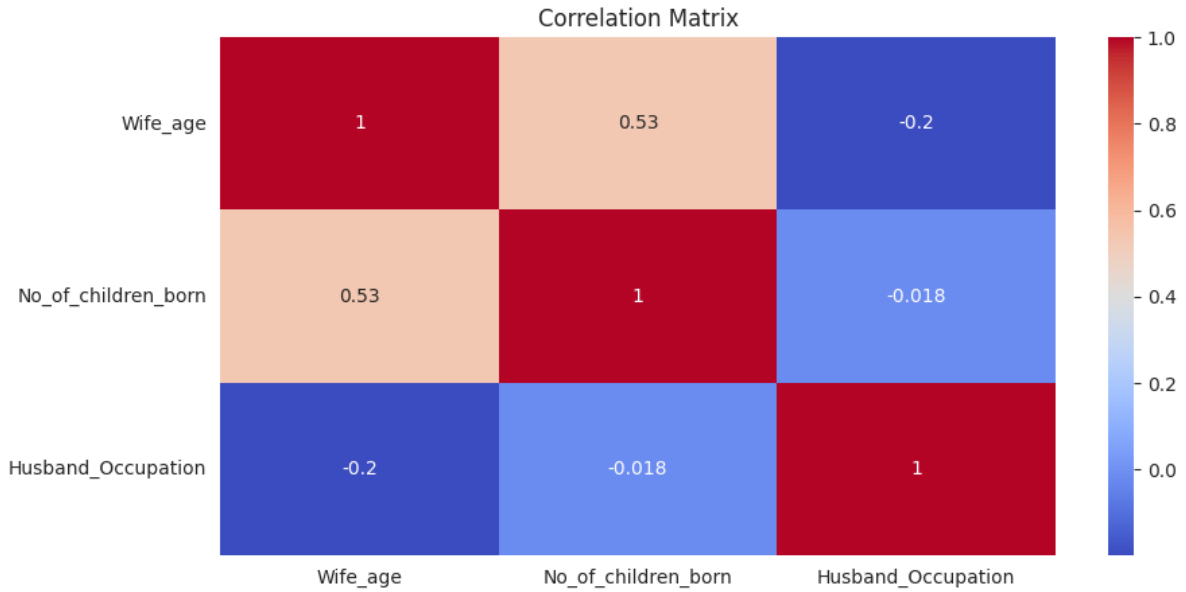
Univarient Analysis of categorical variables:-

There are seven categorical variables present in the dataset on which we perform Univarient Analysis

Bivarinet Analysis:-The Analysis between Two variables is called Bivarient analysis, Heat map is used to perform the Bivarient analysis to check correlation between the variales.

In summary, bivariate analysis using a heatmap in linear regression helps you explore the relationships and correlations between independent variables. It is particularly useful for detecting multicollinearity, which can affect the stability and interpretation of the regression model. By visualizing these relationships, you can make informed decisions about variable selection and model development.

Figure 8



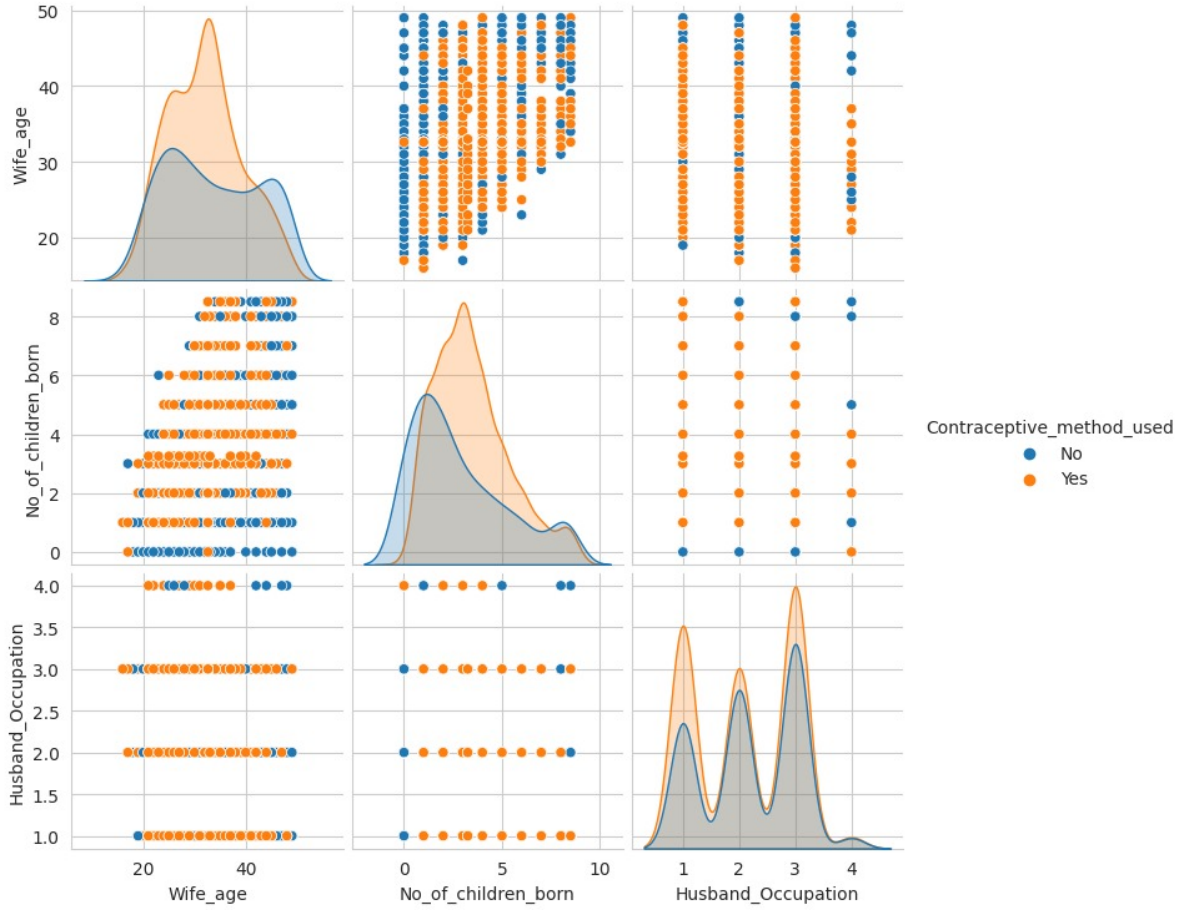
From the correlation plot we can clearly see that wife_age and no_of_children have a correlation between them.

There is no correlation between wige_age and husband_occupation and between husband_occupation and no_of_children_born.

Multivarient Analysis:-

The Analysis between Three or more variables is called multi-varient analysis, Pair plot is used to perform MultiVarient analysis between the variables, to check relation and multicollinearity between the variables.

Figure 8



Question 2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis) and CART.

Response:- In this report, we present the evaluation results of three different classification models applied to our dataset. The models evaluated are Logistic Regression, Linear Discriminant Analysis (LDA), and Classification and Regression Trees (CART). These models aim to predict the target variable, and we have assessed their accuracy and classification performance.

Results of Logistic Regression

- **Accuracy**: 67.87%
- Classification Report
    - Precision for Class 0: 70%
    - Recall for Class 0: 47%
    - F1-score for Class 0: 56%
    - Precision for Class 1: 67%
    - Recall for Class 1: 84%
    - F1-score for Class 1: 75%
- **Analysis**: The Logistic Regression model achieved an accuracy of 67.87%. The model performed reasonably well in distinguishing between the two classes, with a balanced precision and recall trade-off.

LDA Report

- **Accuracy**: 67.42%
- Classification Report
    - Precision for Class 0: 70%
    - Recall for Class 0: 45%
    - F1-score for Class 0: 55%
    - Precision for Class 1: 66%
    - Recall for Class 1: 85%
    - F1-score for Class 1: 74%
- **Analysis**: The LDA model achieved an accuracy of 67.42%. It demonstrated a slightly lower precision and recall for Class 0 compared to Logistic Regression but performed well in predicting Class 1.

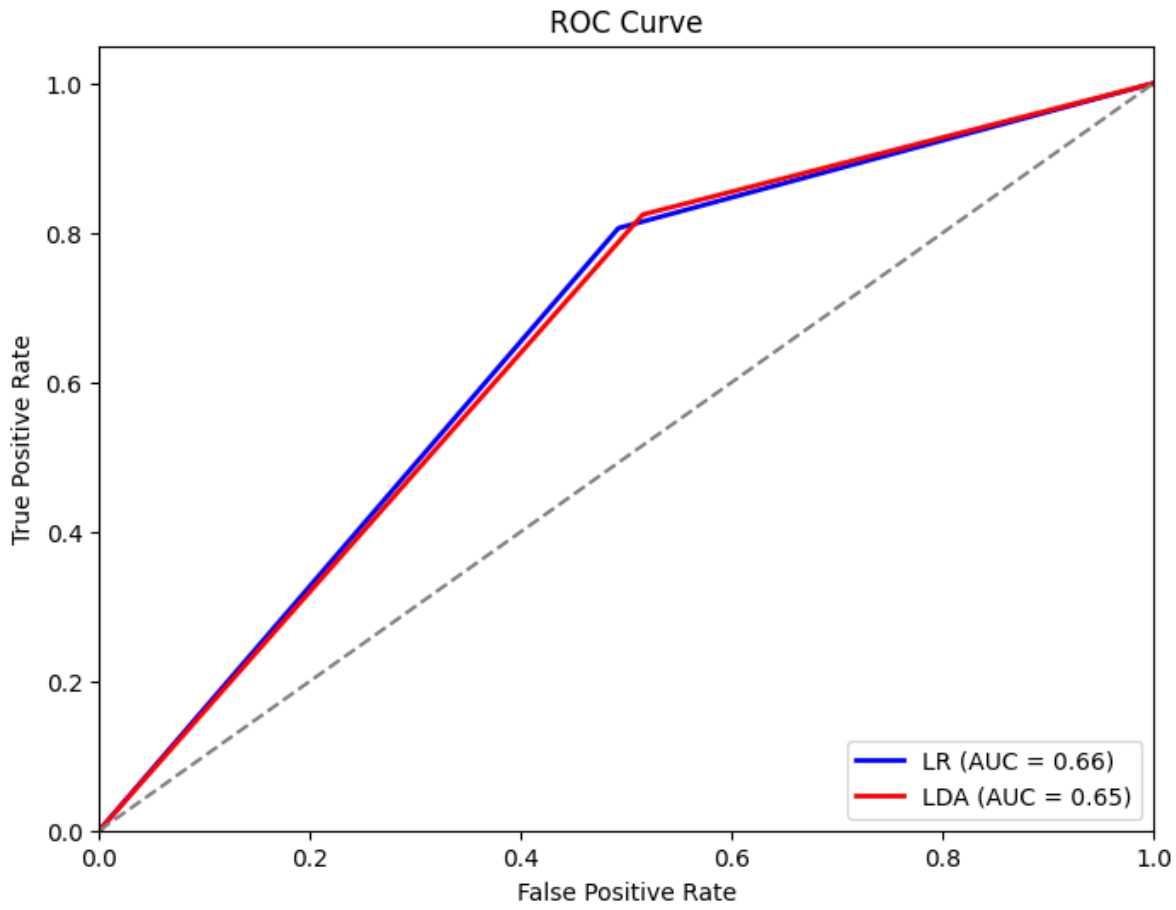
CART Report

- **Accuracy**: 70.36%
- Classification Report
    - Precision for Class 0: 65%
    - Recall for Class 0: 71%
    - F1-score for Class 0: 68%
    - Precision for Class 1: 75%
    - Recall for Class 1: 70%
    - F1-score for Class 1: 73%
- **Analysis**: The CART model outperformed the other models with an accuracy of 70.36%. It demonstrated a good balance between precision and recall for both classes, indicating its capability to predict both classes effectively.

In conclusion, we have evaluated three classification models on our dataset. The Classification and Regression Trees (CART) model has shown the highest accuracy and a balanced performance in predicting both classes. It is recommended for deployment due to its effectiveness.

Question 2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

Figure 9



Question 2.4 Inference: Basis on these predictions, what are the insights and recommendations. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Response:- Executive Summary

In this report, we present the evaluation results of three classification models—Logistic Regression, Linear Discriminant Analysis (LDA), and Classification and Regression Trees (CART)—applied to our dataset. Our goal is to predict the target variable and provide insights and recommendations based on the model evaluations.

The project involved the following key steps:

1. Data Ingestion: We loaded the dataset for analysis.
2. Data Preprocessing: We encoded categorical variables, split the data into training and testing sets, and evaluated three classification models.
3. Model Evaluation: We assessed model accuracy and classification performance.
4. Inference and Recommendations: We provided insights, recommendations, and guidance for model selection based on the business context.

Model Accuracy

1. **Logistic Regression**: Achieved an accuracy of 67.87%.
2. **LDA (Linear Discriminant Analysis)**: Achieved an accuracy of 67.42%.
3. **CART (Classification and Regression Trees)**: Outperformed the others with an accuracy of 70.36%.

Logistic Regression

- Achieved a balanced precision and recall trade-off.
- Performed reasonably well in distinguishing between the two classes.

LDA (Linear Discriminant Analysis)

- Demonstrated slightly lower precision and recall for Class 0 compared to Logistic Regression.
- Performed well in predicting Class 1.

CART (Classification and Regression Trees)

- Outperformed the other models.
- Demonstrated a good balance between precision and recall for both classes.

Model Selection

- **Logistic Regression**: While it achieved a decent accuracy, it may be less suitable for highly imbalanced datasets or when precise class prediction is critical. It is a good option for initial exploration but may not be the best choice in certain scenarios.
- **LDA (Linear Discriminant Analysis)**: Offers competitive performance and may be considered for cases where a simpler and interpretable model is preferred.

- **CART (Classification and Regression Trees)**: The recommended choice due to its highest accuracy and balanced performance in predicting both classes. CART offers versatility and effectiveness for this classification task.

Business Interpretation

- The choice of model should align with the specific business goals and requirements.
- CART offers a robust solution for our predictive tasks and should be considered for deployment.
- The interpretability of the model should be considered, especially if stakeholders require insights into the factors driving the predictions.

Actionable Insights

- Further analysis and fine-tuning of the selected model are recommended for enhanced performance.
- Consider collecting additional data or refining data preprocessing to improve model accuracy.
- Implement a monitoring system to assess model performance in real-world conditions and make necessary adjustments.

Conclusion

The choice of model is pivotal for successful predictions, and it should be aligned with the specific business context. In this evaluation, the Classification and Regression Trees (CART) model demonstrated the highest accuracy and balanced performance. We recommend further refinement and testing of this model for deployment.

The insights and recommendations presented in this report are designed to guide decision-making and enhance the effectiveness of predictive tasks. However, it's essential to consider domain knowledge and specific business requirements when making final decisions.

Model Accuracy

1. **Logistic Regression**: Achieved an accuracy of 67.87%.
2. **LDA (Linear Discriminant Analysis)**: Achieved an accuracy of 67.42%.
3. **CART (Classification and Regression Trees)**: Outperformed the others with an accuracy of 70.36%
4. ## Business Interpretation

    - The choice of model should align with the specific business goals and requirements.
    - CART offers a robust solution for our predictive tasks and should be considered for deployment.
    - The interpretability of the model should be considered, especially if stakeholders require insights into the factors driving the predictions.

      Conclusion

The choice of model is pivotal for successful predictions, and it should be aligned with the specific business context. In this evaluation, the Classification and Regression Trees (CART) model demonstrated the highest accuracy and balanced performance. We recommend further refinement and testing of this model for deployment.

The insights and recommendations presented in this report are designed to guide decision-making and enhance the effectiveness of predictive tasks. However, it's essential to consider domain knowledge and specific business requirements when making final decisions.