PREDICTIVE MODELLING

Project Report

By,

Shobha Kumari Choudhary

3rd September 2023

CONTENTS

LIST OF DIAGRAMS

Problem 1: Linear Regression

The comp-active databases are a collection of a computer systems activity measures. The data was collected from a Sun Sparcstation 20/712 with 128 Mbytes of memory running in a multi-user university department. Users would typically be doing a large variety of tasks ranging from accessing the internet, editing files, or running very CPU-bound programs.

As you are a budding data scientist you thought to find out a linear equation to build a model to predict 'usr'(Portion of time (%) that CPUs run in user mode) and to find out how each attribute affects the system to be in 'usr' mode using a list of system attributes.

1.1 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.

❖ **Reading the Data:**

Reading 1st five Rows of the Dataset

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	
0	1	0	2147	79	68	0.2	0.2	40671.0	53995.0	0.0	...	0.0	0.0	1.6	2.6	16.00	26.40	C
1	0	0	170	18	21	0.2	0.2	448.0	8385.0	0.0	...	0.0	0.0	0.0	0.0	15.63	16.83	Not_C
2	15	3	2162	159	119	2.0	2.4	NaN	31950.0	0.0	...	0.0	1.2	6.0	9.4	150.20	220.20	Not_C
3	0	0	160	12	16	0.2	0.2	NaN	8670.0	0.0	...	0.0	0.0	0.2	0.2	15.60	16.80	Not_C
4	5	1	330	39	38	0.4	0.4	NaN	12185.0	0.0	...	0.0	0.0	1.0	1.2	37.80	47.60	Not_C

5 rows × 22 columns

Reading last five Rows of the Dataset

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt
8187	16	12	3009	360	244	1.6	5.81	405250.0	85282.0	8.02	...	55.11	0.6	35.87	47.90	139.28	270.74
8188	4	0	1596	170	146	2.4	1.80	89489.0	41764.0	3.80	...	0.20	0.8	3.80	4.40	122.40	212.60
8189	16	5	3116	289	190	0.6	0.60	325948.0	52640.0	0.40	...	0.00	0.4	28.40	45.20	60.20	219.80
8190	32	45	5180	254	179	1.2	1.20	62571.0	29505.0	1.40	...	18.04	0.4	23.05	24.25	93.19	202.81
8191	2	0	985	55	46	1.6	4.80	111111.0	22256.0	0.00	...	0.00	0.2	3.40	6.20	91.80	110.00

5 rows × 22 columns

❖ **Info of the Dataset:**

- The dataset contains 8192 Rows and 22 columns.
- There are 13 float, 8 int64 and 1 object datatype present in the data.
- There are Null values present in rchar and wchar columns of the dataset.

```
RangeIndex: 8192 entries, 0 to 8191
Data columns (total 22 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   lread     8192 non-null   int64
 1   lwrite    8192 non-null   int64
 2   scall     8192 non-null   int64
 3   sread     8192 non-null   int64
 4   swrite    8192 non-null   int64
 5   fork      8192 non-null   float64
 6   exec      8192 non-null   float64
 7   rchar     8088 non-null   float64
 8   wchar     8177 non-null   float64
 9   pgout     8192 non-null   float64
 10  ppgout    8192 non-null   float64
 11  pgfree    8192 non-null   float64
 12  pgscan    8192 non-null   float64
 13  atch      8192 non-null   float64
 14  pgin      8192 non-null   float64
 15  ppgin     8192 non-null   float64
 16  pflt      8192 non-null   float64
 17  vflt      8192 non-null   float64
 18  runqsz    8192 non-null   object
 19  freemem   8192 non-null   int64
 20  freeswap  8192 non-null   int64
 21  usr       8192 non-null   int64
```

❖ **Five-point Summary:**

- There are many Zero entries present in the columns.
- Also, the data seems to be Skewed mostly Right Skewed and usr column is left skewed.
- There is huge difference in the Min and Max values which shows the presence of Outliers.

Univariate Analysis:

❖ **Histogram of all Continuous Variables:**

Observations:

- We can clear see that none of the column is Normally distributed.
- Almost all the column is Right Skewed and Shows the Presence of Outliers in the data.
- The Target column 'Usr' is Left Skewed.

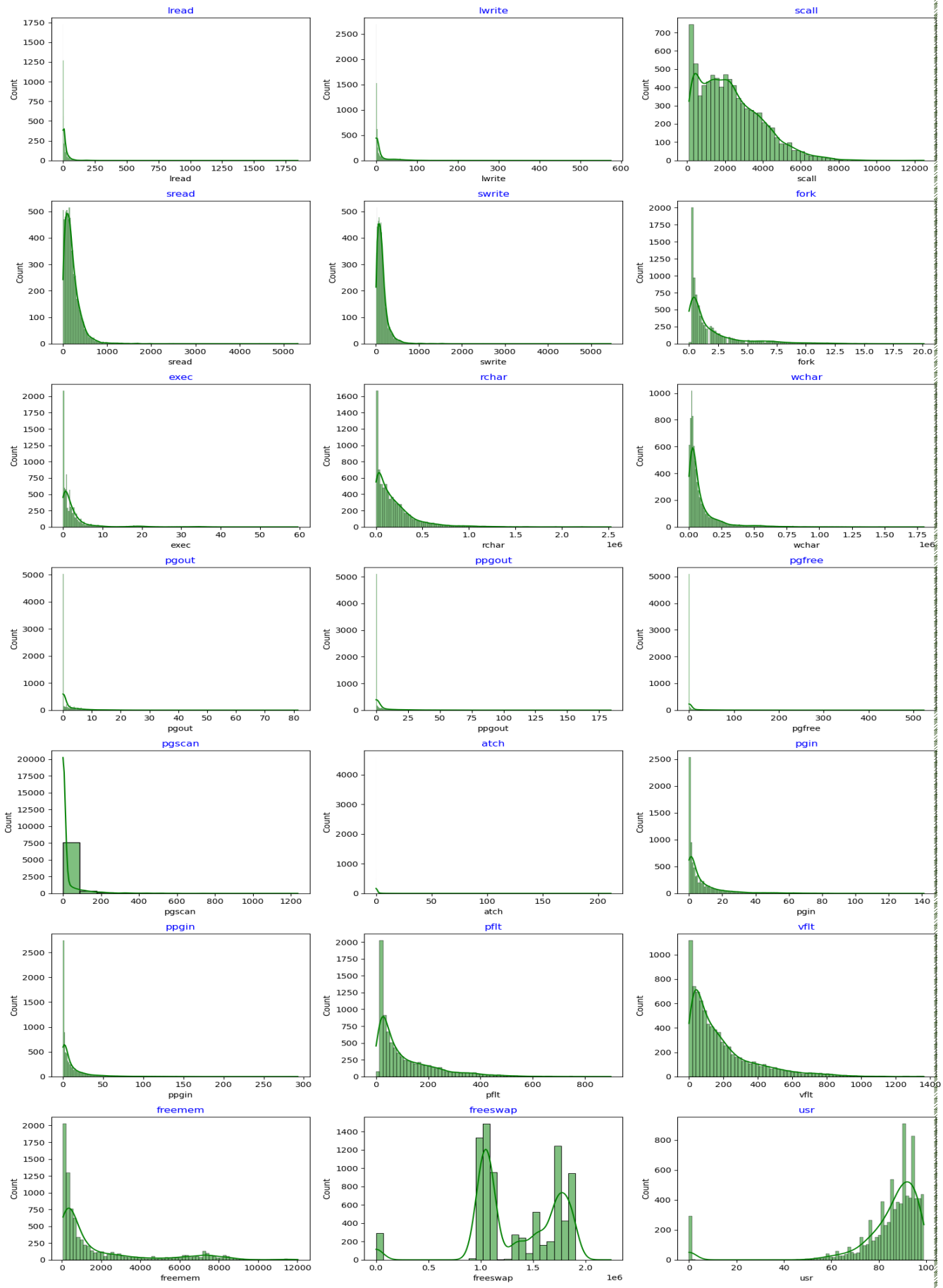


Fig 1: Histogram of all Continuous Variables

❖ **Count plot of categorical variable:**

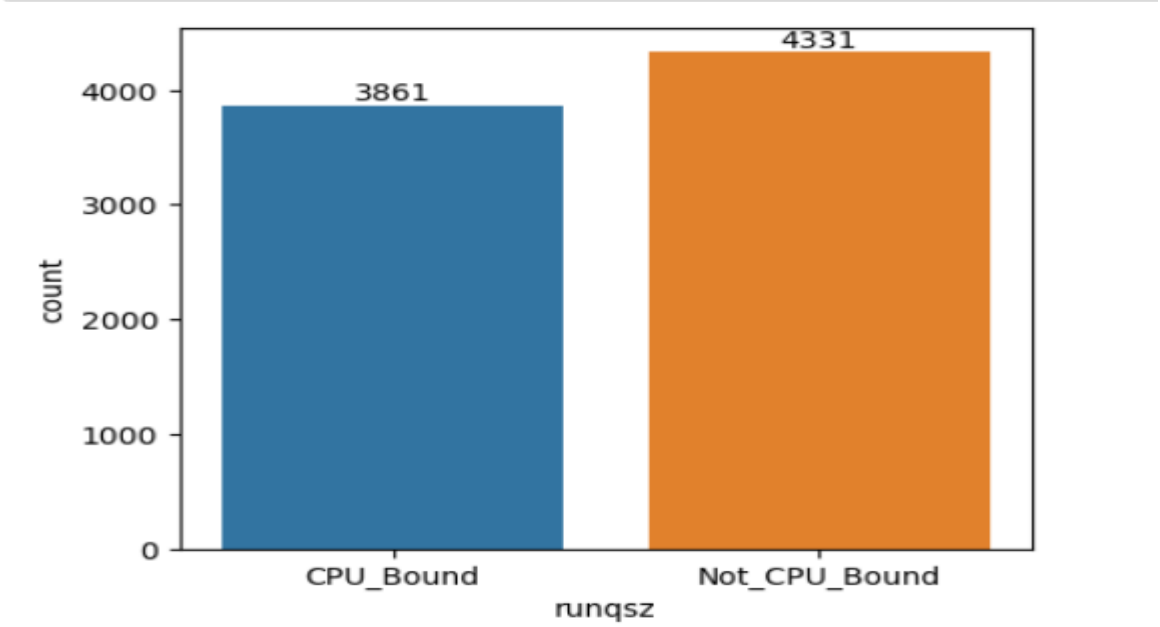


Fig 2: Count plot of categorical variable

Observations:

- Process run queue size; this value should be less than 2 means the system is Not_CPU_Bound.
- System being Not_CPU_Bound is more in numbers as compared to CPU_Bound.

❖ **Boxplot of all Continuous variables:**

Observations:

- There are Outliers present in all the Numeric Columns.
- All the Columns are Right Skewed Except the columns freeswap and Usr which are Left Skewed.

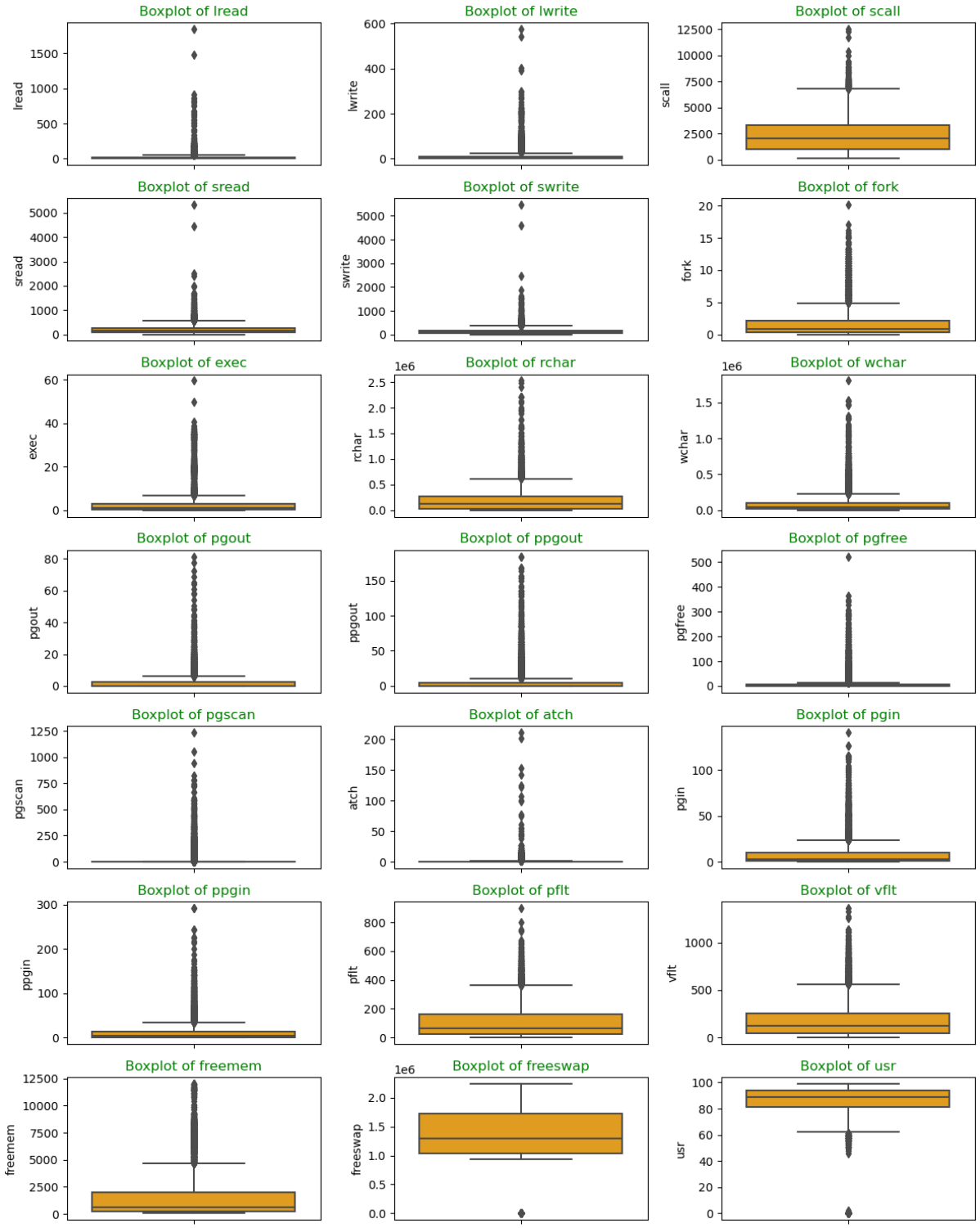


Fig 3: Boxplot of all Continuous variables

Bivariate Analysis:

❖ Plot between lread and usr:

- Two unusual trends can be seen when comparing the number of reads per second with the CPUs running in user mode.

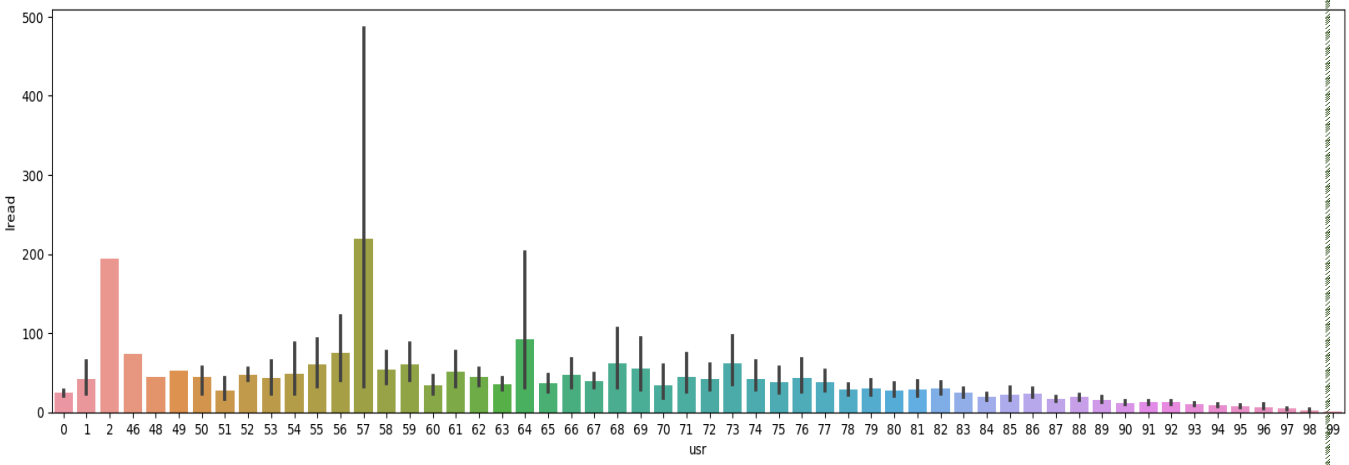


Fig 4: Plot between lread and usr

❖ **Plot between lwrite and usr**:

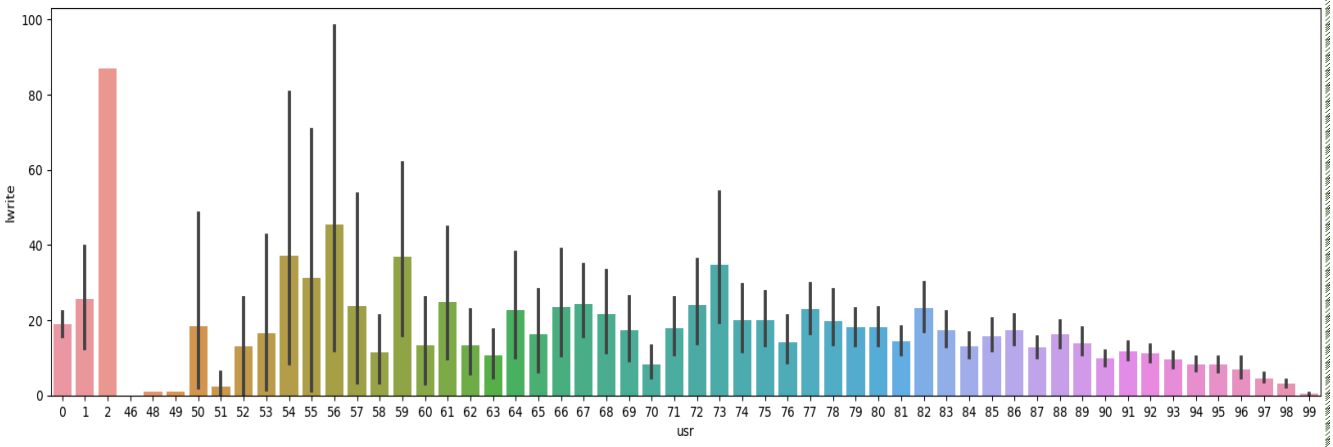


Fig 5: Plot between lwrite and usr

Observations:

- ❖ We can observe that when the number or write is high, only 2% of the CPU runs in user mode.
- ❖ This indicates that when the read/write is high, most of the CPU does not run in user mode.

Multivariate Analysis:

❖ **Heatmap:**

Observations:

- ❖ variable exec Pfit and Vfit are highly Correlated to fork having a correlation of 0.76, 0.93 and 0.94 respectively.
- ❖ swrite is correlated to sread with 0.88 magnitude.
- ❖ From the analysis we can see the presence of correlations.
- ❖ Both the page fault variables – Pfit & vflt are highly correlated with the fork variable.

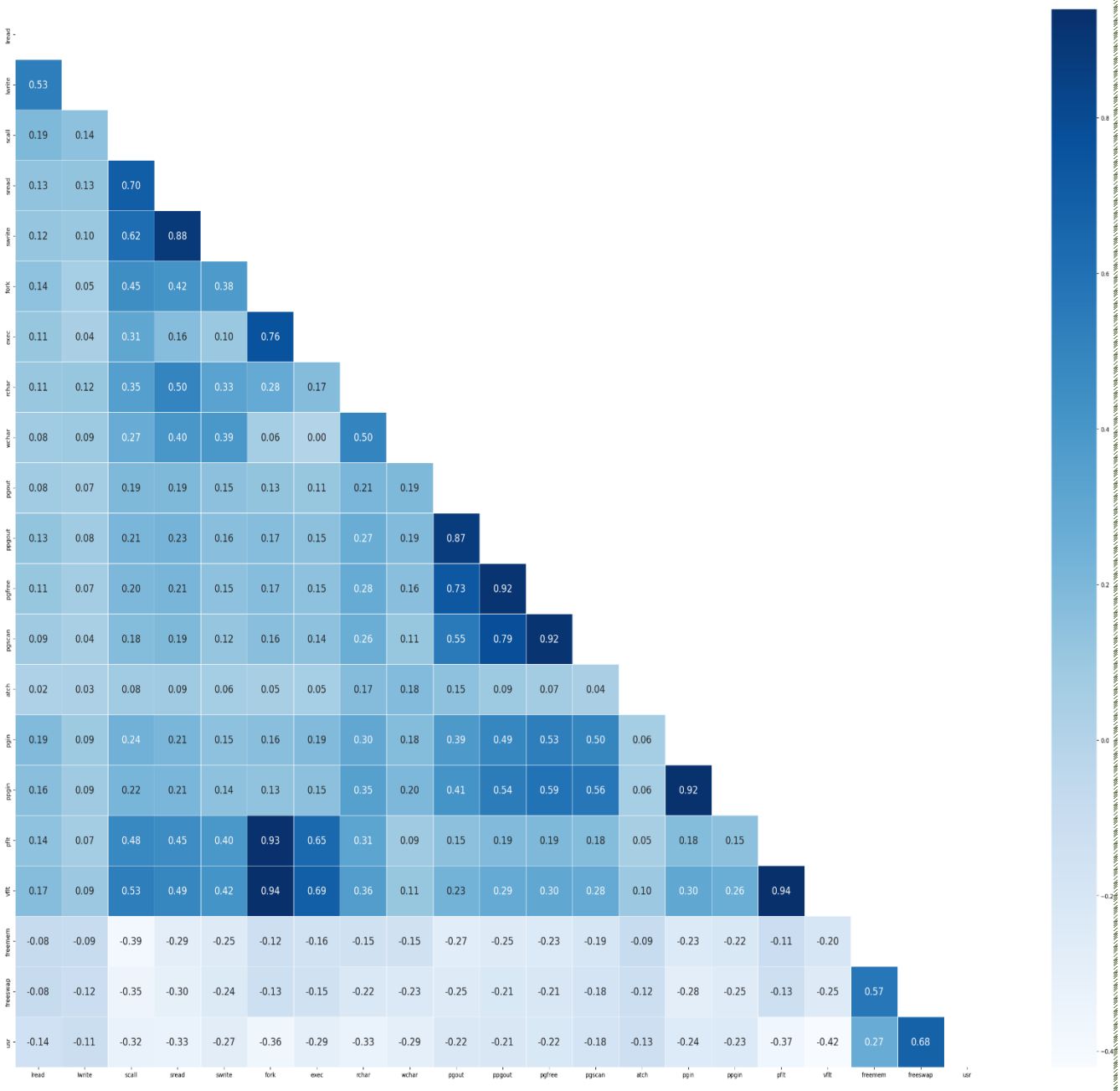


Fig 6: Heatmap

❖ Pairplot

Observations:

- There is strong positive correlation between the target variable usr and the predictor variables freemem and freeswap.
- Linear correlation can be seen between ‘vflt’, ‘pflt’ & ‘fork’. If the fork calls increase, number of page faults also tend to increase.
- Similarly, Number of page out requests per second is also highly correlated to the number of pages, paged out per second.

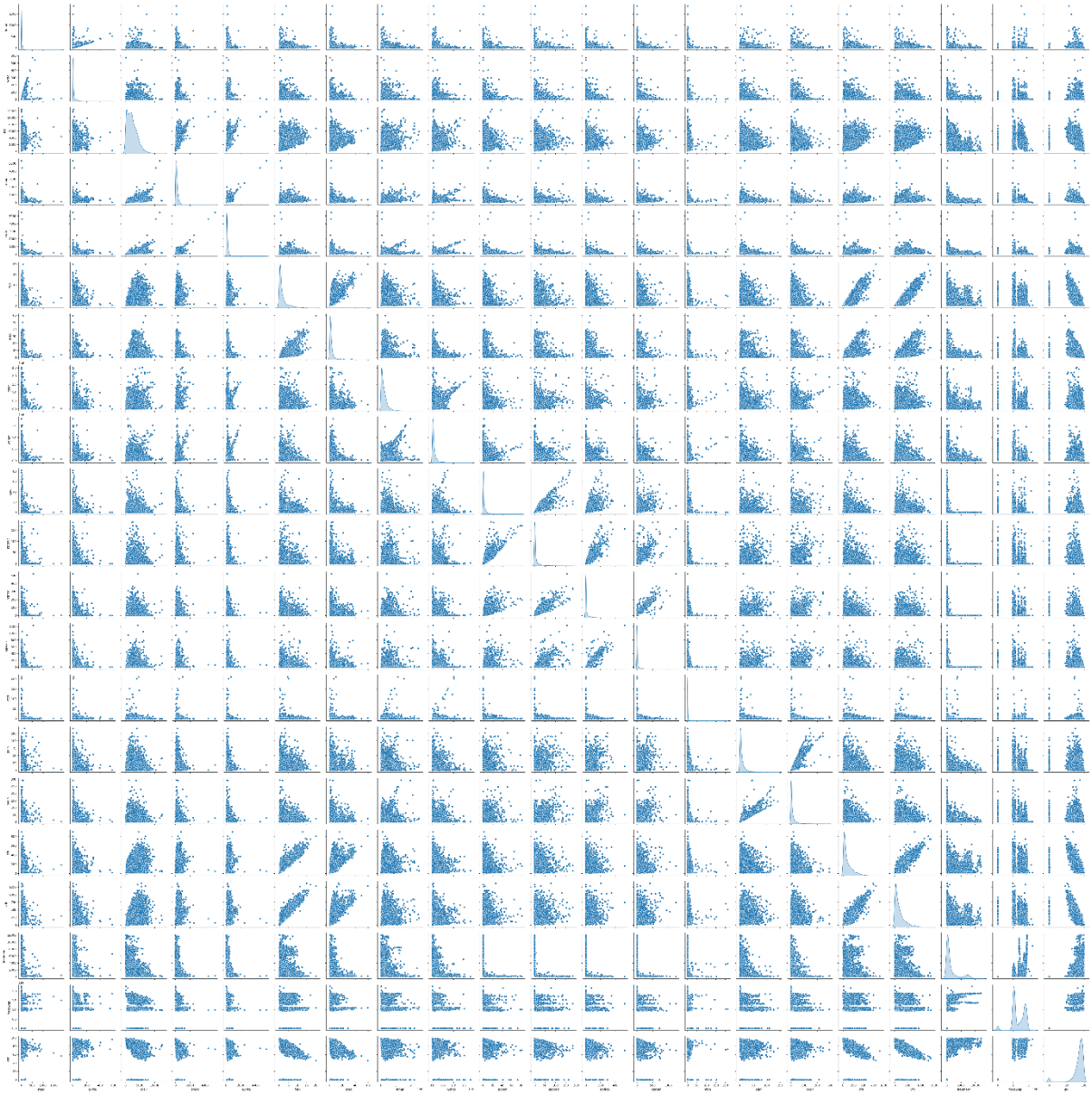
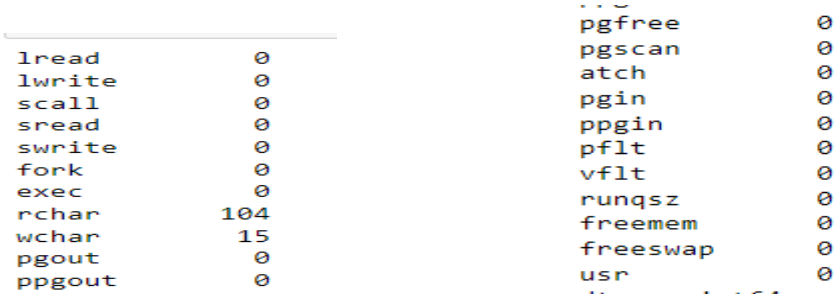


Fig 7: Pairplot

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.

❖ **Checking NULL Vaalues:**



lread	0
lwrite	0
scall	0
sread	0
swrite	0
fork	0
exec	0
rchar	104
wchar	15
pgout	0
ppgout	0
pgfree	0
pgscan	0
atch	0
pgin	0
ppgin	0
pflt	0
vflt	0
runqsz	0
freemem	0
freeswap	0
usr	0

- There are Null Values present in rchar and wchar columns.
- We can impute the Null values with the median values.
- We can't use mean to impute Null values since Outliers are present in the column.

❖ **Null Value Treatment:**

- Let's replace the missing values with median values of the columns.
- Respective column's missing value is replaced with that column's median respectively.

```
compactive[['rchar','wchar']].isnull().sum()
```

```
rchar    0
wchar    0
dtype: int64
```

❖ **check for the values which are equal to zero:**

- From the Statistical Summary above It has been observed that there are 0s present for many columns.
- We do not need to drop all the rows with Zero values.
- These are all valid values as it is related to the activities being done in the computer system.
- we also observe that column pgscan is having all the vales as Zero from min to 75% percentile and there is only max value of 1237.00 which we need to consider for deleting since after the Outlier treatment all the values will be Zero then.

❖ **Checking** for **Duplicates:**

- Number of Duplicate Rows Present: 0

❖ **Outlier Treatment:**

- The linear regression model is quite sensitive to outliers in the data set. If we do not treat outliers, our model is likely to make inaccurate predictions.
- Treating Outlier using IQR

❖ **Boxplot After Outlier Treatment:**

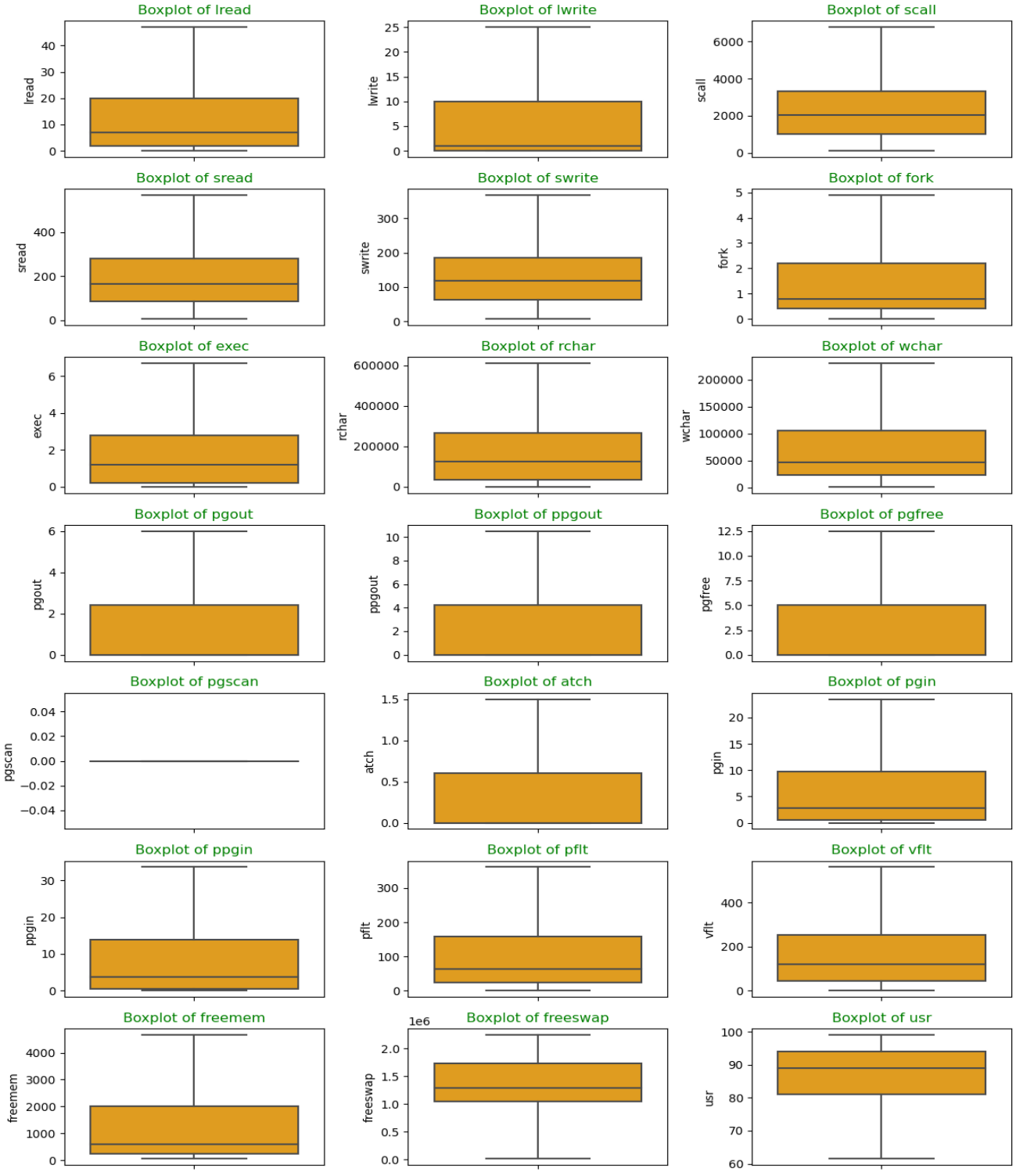


Fig 8: Boxplot after Outlier Treatment

Observations:

- After the Outlier treatment we observe that the column pgscan have all the vales as Zeros, so we need to drop this column before model building.

```
compactive['pgscan'].describe()
```

```
count    8192.0
mean        0.0
std         0.0
min         0.0
25%         0.0
50%         0.0
75%         0.0
max         0.0
Name: pgscan, dtype: float64
```

❖ **Dropping pgscan column which is having zero values present in all rows.**

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 stats model. 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.

❖ **Data Encoding:**

- Encoding the data using 'get_dummies' method.

xec	rchar	wchar	pgout	...	pgfree	atch	pgin	ppgin	pflt	vflt	freemem	freeswap	usr	runqsz_Not_CPU_Bound
0.2	40671.0	53995.0	0.0	...	0.0	0.0	1.6	2.6	16.00	26.40	4659.125	1730946.0	95.0	0
0.2	448.0	8385.0	0.0	...	0.0	0.0	0.0	0.0	15.63	16.83	4659.125	1869002.0	97.0	1
2.4	125473.5	31950.0	0.0	...	0.0	1.2	6.0	9.4	150.20	220.20	702.000	1021237.0	87.0	1
0.2	125473.5	8670.0	0.0	...	0.0	0.0	0.2	0.2	15.60	16.80	4659.125	1863704.0	98.0	1
0.4	125473.5	12185.0	0.0	...	0.0	0.0	1.0	1.2	37.80	47.60	633.000	1760253.0	90.0	1

- 0 stands for CPU_Bound and 1 stands for Not_CPU_Bound

❖ **Split the data into train and test (70:30):**

Shape of the data after Train and test split

```
Number of rows and columns of the training set for the independent variables: (5734, 21)
Number of rows and columns of the training set for the dependent variable: (5734, 1)
Number of rows and columns of the test set for the independent variables: (2458, 21)
Number of rows and columns of the test set for the dependent variable: (2458, 1)
```

Linear Regression using Sklearn method:

The Linear Regression model is built and fitted into the Training dataset using

Sklearn Library.

- The coefficients for each of the independent attributes are given below:

```
The coefficient for const is 0.0
The coefficient for lread is -0.06348150618197021
The coefficient for lwrite is 0.04816128709146338
The coefficient for scall is -0.0006638280111666028
The coefficient for sread is 0.0003082521031322379
The coefficient for swrite is -0.0054218222976218314
The coefficient for fork is 0.029312727249262423
The coefficient for exec is -0.3211664838987394
The coefficient for rchar is -5.166841759456434e-06
The coefficient for wchar is -5.4028752354273735e-06
The coefficient for pgout is -0.36881906387327296
The coefficient for ppgout is -0.07659768212738544
The coefficient for pgfree is 0.08448414470556936
The coefficient for atch is 0.6275741574815779
The coefficient for pgin is 0.019987907678725117
The coefficient for ppgin is -0.06733383975700702
The coefficient for pflt is -0.03360282937751041
The coefficient for vflt is -0.005463668798530671
The coefficient for freemem is -0.0004584671879475078
The coefficient for freeswap is 8.83184026303531e-06
The coefficient for runqsz_Not_CPU_Bound is 1.615297848824913
```

- **Let us check the intercept for the model:**

  The intercept for our model is: 84.1217407953198

- **R square on training data:** 0.796108610127457

  - 79.61% of the variation in the 'usr' is explained by the predictors in the model for train set.

- **R square on testing data:** 0.7677318597936396
  - This shows that almost 77% of the variance of the testing data set was captured by the model.

- RMSE on Training data: 4.419536092979902
- RMSE on Testing data: 4.652295704192376

Inferences:

- The model seems to be neither overfitted nor under-fitted since the R-squared value of Train and Test data are very close to each other i.e., comparable.
- Therefore, we can say that this is a good model to go with.

- However, let's see if there is any improvement with the stats model approach.

Linear Regression using stats model (OLS)

❖ Regression summary of the model:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.796
Model:                            OLS   Adj. R-squared:                  0.795
Method:                 Least Squares   F-statistic:                     1115.
Date:                Mon, 28 Aug 2023   Prob (F-statistic):               0.00
Time:                        14:19:46   Log-Likelihood:                -16657.
No. Observations:                5734   AIC:                         3.336e+04
Df Residuals:                    5713   BIC:                         3.350e+04
Df Model:                          20
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                      84.1217      0.316    266.106      0.000      83.502      84.741
lread                      -0.0635      0.009     -7.071      0.000      -0.081      -0.046
lwrite                      0.0482      0.013      3.671      0.000       0.022       0.074
scall                      -0.0007   6.28e-05    -10.566      0.000      -0.001      -0.001
sread                       0.0003      0.001      0.305      0.760      -0.002       0.002
swrite                     -0.0054      0.001     -3.777      0.000      -0.008      -0.003
fork                        0.0293      0.132      0.222      0.824      -0.229       0.288
exec                       -0.3212      0.052     -6.220      0.000      -0.422      -0.220
rchar                  -5.167e-06   4.88e-07    -10.598      0.000   -6.12e-06   -4.21e-06
wchar                  -5.403e-06   1.03e-06     -5.232      0.000   -7.43e-06   -3.38e-06
pgout                      -0.3688      0.090     -4.098      0.000      -0.545      -0.192
ppgout                     -0.0766      0.079     -0.973      0.330      -0.231       0.078
pgfree                      0.0845      0.048      1.769      0.077      -0.009       0.178
atch                        0.6276      0.143      4.394      0.000       0.348       0.908
pgin                        0.0200      0.028      0.703      0.482      -0.036       0.076
ppgin                      -0.0673      0.020     -3.415      0.001      -0.106      -0.029
pflt                       -0.0336      0.002    -16.957      0.000      -0.037      -0.030
vflt                       -0.0055      0.001     -3.830      0.000      -0.008      -0.003
freemem                    -0.0005   5.07e-05     -9.038      0.000      -0.001      -0.000
freeswap                8.832e-06    1.9e-07     46.472      0.000    8.46e-06     9.2e-06
runqsz_Not_CPU_Bound       1.6153      0.126     12.819      0.000       1.368       1.862
==============================================================================
Omnibus:                     1103.645   Durbin-Watson:                   2.016
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2372.553
Skew:                          -1.119   Prob(JB):                         0.00
Kurtosis:                       5.219   Cond. No.                     7.74e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.74e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

❖ **Interpretation of R-squared**

- The R-squared value tells us that our model can explain 79.6% of the variance in the training set.
- Adj. R-squared: 0.795

❖ **Interpretation of Coefficients**

- The coefficients tell us how one unit change in X can affect y.
- The sign of the coefficient indicates if the relationship is positive or negative.

- In this data set, for example, an increase of 1 lread - Reads (transfers per second) between system memory and user memory occurs with 0.0635 decrease in usr(Portion of time (%) that CPUs run in user mode)
- Similarly, a unit increase in lwrite occurs with 0.0482 increase in usr.

❖ **Multicollinearity:**

- Multicollinearity occurs when predictor variables in a regression model are correlated. This correlation is a problem because predictor variables should be independent. If the collinearity between variables is high, we might not be able to trust the p-values to identify independent variables that are statistically significant.
- When we have multicollinearity in the linear model, the coefficients that the model suggests are unreliable.

❖ **Interpretation of p-values (P > |t|):**

- (P > |t|) gives the p-value for each predictor variable to check the null hypothesis.
- If the level of significance is set to 5% (0.05), the p-values greater than 0.05 would indicate that the corresponding predictor variables are not significant.

❖ **How to check for Multicollinearity**

- There are different ways of detecting (or testing) multicollinearity. One such way is Variation Inflation Factor.
- Variance Inflation factor: Variance inflation factors measure the inflation in the variances of the regression coefficients estimates due to collinearities that exist among the predictors. It is a measure of how much the variance of the estimated regression coefficient βk is "inflated" by the existence of correlation among the predictor variables in the model.
- General Rule of Thumb:
  - **If VIF is 1**, then there is no correlation among the kth predictor and the remaining predictor variables, and hence, the variance of βk is not inflated at all.
  - **If VIF exceeds 5**, we say there is moderate VIF, and if it is 10 or exceeding 10, it shows signs of high multi-collinearity.
  - The purpose of the analysis should dictate which threshold to use.

❖ **let's check the VIF of the predictors:**

```
VIF values:

const                    29.229332
lread                     5.350560
lwrite                    4.328397
scall                     2.960609
sread                     6.420172
swrite                    5.597135
fork                     13.035359
exec                      3.241417
rchar                     2.133616
wchar                     1.584381
pgout                    11.360363
ppgout                   29.404223
pgfree                   16.496748
atch                      1.875901
pgin                     13.809339
ppgin                    13.951855
pflt                     12.001460
vflt                     15.971049
freemem                   1.961304
freeswap                  1.841239
runqsz_Not_CPU_Bound      1.156815
dtype: float64
```

Inferences:

- Considering the Threshold value for VIF as 5, we will check the features with VIF value > 5 for dropping the column with multicollinearity.
- Most of the variables has VIF values more than 5. It means there are high multi collinearity present in the data. This will affect the model's prediction.
- The VIF values indicate that the features Iread, sread, swrite, fork, pgout, ppgout, pgfree, pgin, ppgin, pflt, vflt, are correlated with one or more independent features.
- Multicollinearity affects only the specific independent variables that are correlated. Therefore, in this case, we can trust the p-values of Iwrite, scall, exec, rchar, wchar, atch, freemem, freeswap and runqsz_Not_CPU_Bound variables.
- To treat multicollinearity, we will have to drop one or more of the correlated features (Iread, sread, swrite, fork, pgout, ppgout, pgfree, pgin, ppgin, pflt, vflt).
- We will drop the variable that has the least impact on the adjusted R-squared of the model.

Treating multicollinearity:

- **Let's remove/drop multicollinear columns one by one and observe the effect on our predictive model.**

We will remove all the variables that has the least impact on the adjusted R-squared.
We dropped all the features that caused multicollinearity in the data and having VIF value > 5.

- **The following predictors are dropped:**
  ppgout, vflt, ppgin, sread, pgfree, fork, lread.

After dropping all the above features, Let's check if multicollinearity is still present in the data.

```
VIF values:

const                    28.319429
lwrite                    1.051857
scall                     2.647850
swrite                    3.011789
exec                      2.819082
rchar                     1.672367
wchar                     1.533942
pgout                     2.029125
atch                      1.859694
pgin                      1.454496
pflt                      3.254499
freemem                   1.945632
freeswap                  1.761784
runqsz_Not_CPU_Bound      1.144436
dtype: float64
```

- Now there is no multicollinearity present in the data.
- VIF < 5, the threshold value for all the predictors.
- Also, the p-value is significant for all the predictors.
- p-value is < 0.5 for the features.

❖ Our Final Model is model_51:

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.793
Model:	OLS	Adj. R-squared:	0.792
Method:	Least Squares	F-statistic:	1684.
Date:	Mon, 28 Aug 2023	Prob (F-statistic):	0.00
Time:	14:21:56	Log-Likelihood:	-16702.
No. Observations:	5734	AIC:	3.343e+04
Df Residuals:	5720	BIC:	3.353e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	83.9943	0.313	267.987	0.000	83.380	84.609
lwrite	-0.0347	0.007	-5.325	0.000	-0.047	-0.022
scall	-0.0007	5.98e-05	-11.434	0.000	-0.001	-0.001
swrite	-0.0059	0.001	-5.567	0.000	-0.008	-0.004
exec	-0.3937	0.049	-8.118	0.000	-0.489	-0.299
rchar	-5.539e-06	4.35e-07	-12.740	0.000	-6.39e-06	-4.69e-06
wchar	-4.52e-06	1.02e-06	-4.416	0.000	-6.53e-06	-2.51e-06
pgout	-0.3554	0.038	-9.278	0.000	-0.431	-0.280
atch	0.5867	0.143	4.096	0.000	0.306	0.868
pgin	-0.0942	0.009	-10.132	0.000	-0.112	-0.076
pflt	-0.0415	0.001	-39.927	0.000	-0.044	-0.039
freemem	-0.0005	5.09e-05	-9.195	0.000	-0.001	-0.000
freeswap	8.998e-06	1.87e-07	48.055	0.000	8.63e-06	9.37e-06
runqsz_Not_CPU_Bound	1.6886	0.126	13.376	0.000	1.441	1.936

Omnibus:	1034.508	Durbin-Watson:	2.010
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2155.036
Skew:	-1.067	Prob(JB):	0.00
Kurtosis:	5.114	Cond. No.	7.61e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.61e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

After dropping the features causing strong multicollinearity and the statistically insignificant features, our model performance hasn't dropped sharply (adj. R-squared has dropped from 0.795 to 0.792 and R-squared dropped from 0.796 to 0.793). This shows that these variables did not have much predictive power.

❖ check model performance: -

- ♦ RMSE on the train data: 4.354992605513227
- ♦ RMSE on the test data: 4.551650464226139
- ♦ MAE on the train data: 3.1918138704684407
- ♦ MAE on the test data: 3.2619385835954215

Conclusions:

- We can see that RMSE on the train and test sets are comparable. So, our model is not suffering from overfitting.
- MAE indicates that our current model is able to predict 'usr' within a mean error of 3.2 units on the test data.
- Hence, we can conclude the model "olsres_52" is good for prediction as well as inference purposes.

❖ **Assumptions of Linear Regression:**

These assumptions are essential conditions that should be met before we draw inferences regarding the model estimates or use the model to make a prediction.

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

❖ **TEST FOR LINEARITY AND INDEPENDENCE:**

- Linearity describes a straight-line relationship between two variables, predictor variables must have a linear relation with the dependent variable.

How to check linearity?

- Make a plot of fitted values vs residuals. If they don't follow any pattern (the curve is a straight line), then we say the model is linear otherwise model is showing signs of non-linearity.

	Actual_values	Fitted_values	Residuals
0	91.0	90.985214	0.014786
1	94.0	91.740525	2.259475
2	61.5	75.127299	-13.627299
3	83.0	80.546759	2.453241
4	94.0	97.487993	-3.487993

❖ **let us plot the fitted values vs residuals:**

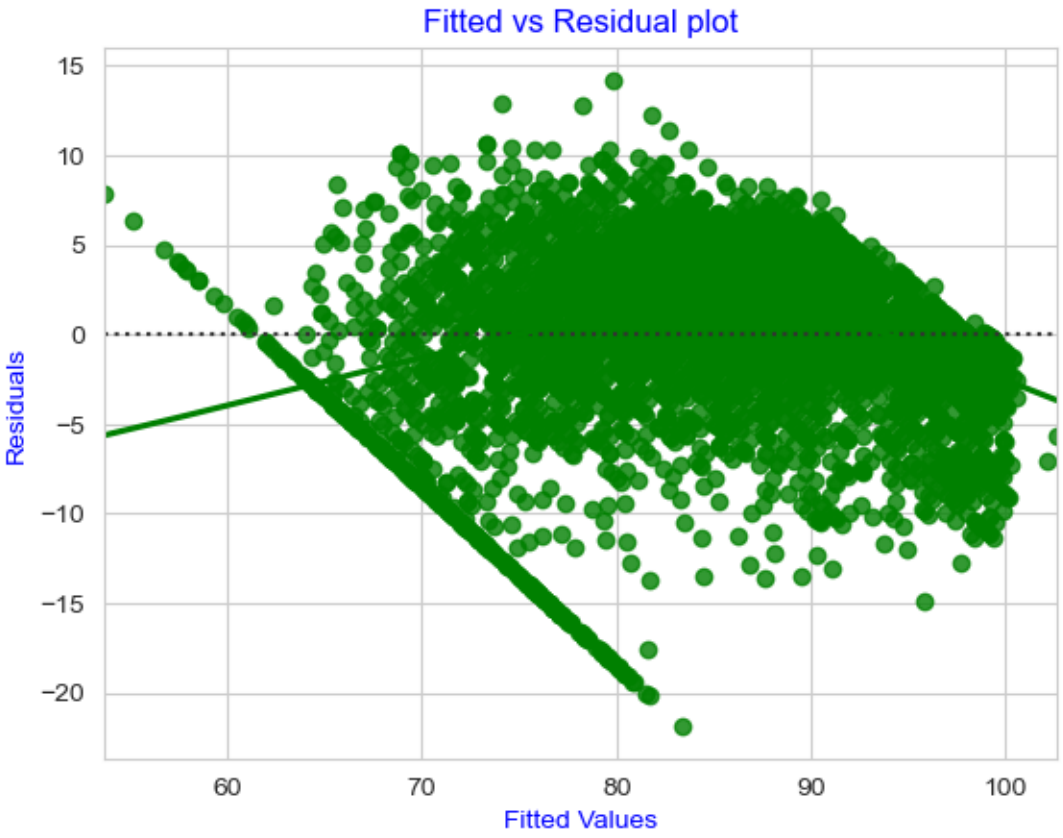


Fig 9: Fitted vs Residuals plot

- **No pattern in the data thus the assumption of linearity and independence of predictors satisfied.**

Test for Normality

❖ **Shapiro Test:**

Shapiro Result:
(statistic=0.9425298571586609, p_value=1.4881789691129557e-42)

- Since p-value < 0.05, the residuals are not normal as per Shapiro test.
- Strictly speaking - the residuals are not normal. However, as an approximation, we might be willing to accept this distribution as close to being normal.

❖ **Transformation: using square transformation**

- Since the Residuals doesn't seems to be normal, we can transform one of the predictor scall to scall_sq to bring the Residuals to Normal.

After variable Transformation we noticed that the value of R-squared and Adjusted R-squared are also increased drastically which is a good sign for our model.

- R-squared value = 0.802
- Adj. R-square – 0.802

```
print(olsres_52.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.802
Model:                            OLS   Adj. R-squared:                  0.802
Method:                 Least Squares   F-statistic:                     1655.
Date:                Mon, 28 Aug 2023   Prob (F-statistic):               0.00
Time:                        14:22:33   Log-Likelihood:                -16573.
No. Observations:                5734   AIC:                         3.318e+04
Df Residuals:                    5719   BIC:                         3.328e+04
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                         81.0901      0.355    228.583      0.000      80.395      81.786
lwrite                        -0.0336      0.006     -5.274      0.000      -0.046      -0.021
scall                          0.0015      0.000     10.300      0.000       0.001       0.002
swrite                        -0.0085      0.001     -8.079      0.000      -0.011      -0.006
exec                          -0.4446      0.048     -9.354      0.000      -0.538      -0.351
rchar                     -5.443e-06   4.25e-07    -12.803      0.000   -6.28e-06   -4.61e-06
wchar                     -4.556e-06      1e-06     -4.552      0.000   -6.52e-06   -2.59e-06
pgout                         -0.3196      0.038     -8.518      0.000      -0.393      -0.246
atch                           0.5637      0.140      4.024      0.000       0.289       0.838
pgin                          -0.1073      0.009    -11.761      0.000      -0.125      -0.089

 pflt                         -0.0413      0.001    -40.620      0.000      -0.043      -0.039
 freemem                      -0.0003   5.05e-05     -6.507      0.000      -0.000      -0.000
 freeswap                  9.477e-06   1.85e-07     51.105      0.000    9.11e-06    9.84e-06
 runqsz_Not_CPU_Bound         1.9173      0.124     15.433      0.000       1.674       2.161
 scall_sq                 -3.405e-07    2.1e-08    -16.252      0.000   -3.82e-07   -2.99e-07
 ==============================================================================
 Omnibus:                      984.083   Durbin-Watson:                   2.009
 Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2031.464
 Skew:                          -1.023   Prob(JB):                         0.00
 Kurtosis:                       5.078   Cond. No.                     7.73e+07
 ==============================================================================

 Notes:
 [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
 [2] The condition number is large, 7.73e+07. This might indicate that there are
 strong multicollinearity or other numerical problems.
```

Observations

- R-squared of the model is 0.802 and adjusted R-squared is also 0.802, which shows that the model is able to explain ~80% variance in the data. This is quite good.
- A unit increase in the lwrite will result in a 0.0336 unit decrease in the usr, all other variables remaining constant.
- The usr of Not CPU Bounded will be 1.9173 units higher than a usr of CPU Bounded, all other variables remaining constant.

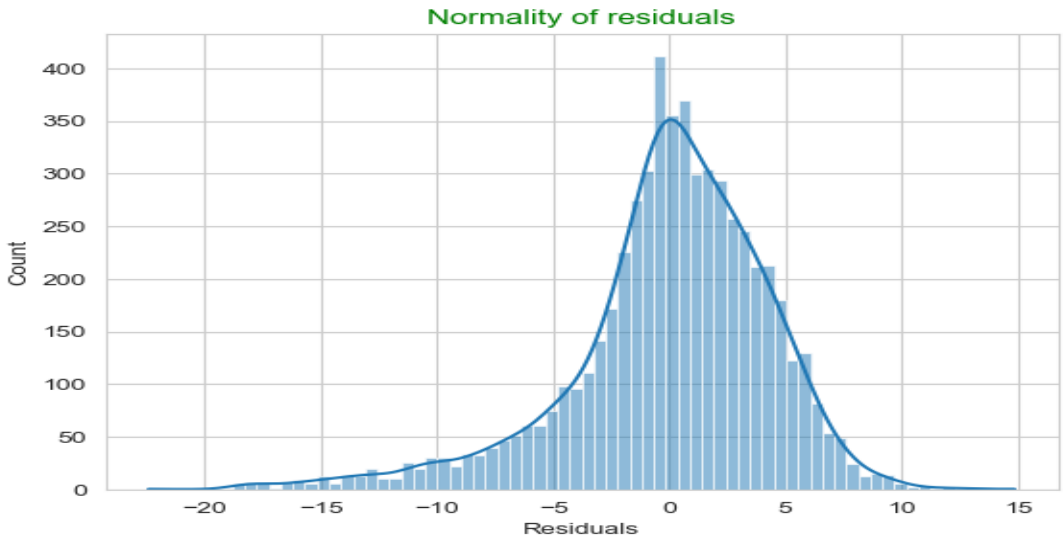


Fig 10: Normality of Residuals plot

- The above histpot shows the Residuals are Normally distributed now after transformation.

The QQ plot of residuals can be used to visually check the normality assumption. The normal probability plot of residuals should approximately follow a straight line.

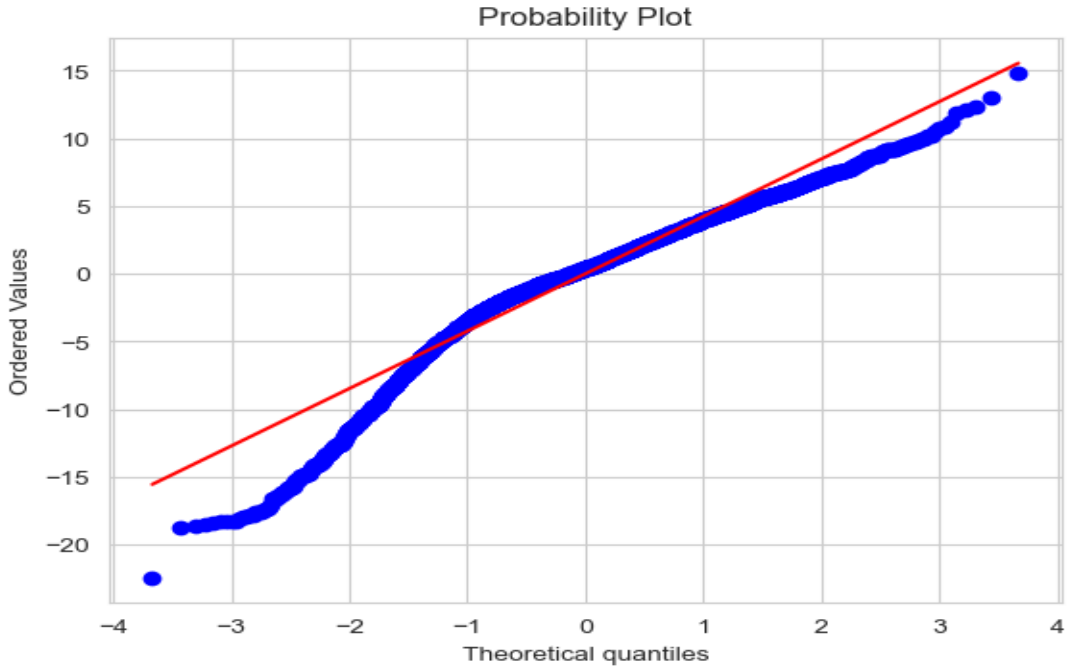


Fig 11: QQ Plot

- Data points are closer to the line in QQ plot.
- Very few data points are lying on the straight line in QQ plot.

❖ **TEST FOR HOMOSCEDASTICITY**

- **Homoscedasticity** - If the variance of the residuals is symmetrically distributed across the regression line, then the data is said to homoscedastic.

- **Heteroscedasticity** - If the variance is unequal for the residuals across the regression line, then the data is said to be heteroscedastic. In this case the residuals can form an arrow shape or any other non-symmetrical shape.

How to check if model has Heteroscedasticity?

- Can use the Goldfeld Quandt test. If we get p-value > 0.05 we can say that the residuals are homoscedastic, otherwise they are heteroscedastic.

The null and alternate hypotheses of the goldfeldquandt test are as follows:

- Null hypothesis: Residuals are homoscedastic.
- Alternate hypothesis: Residuals have heteroscedasticity.

The simplest way to detect heteroscedasticity is by creating a fitted value vs. residual plot.

- The scatterplot below shows a typical fitted value vs. residual plot in which the residuals have constant variance at every level of x.
- Thus, we can say that the residuals are homoscedastic.

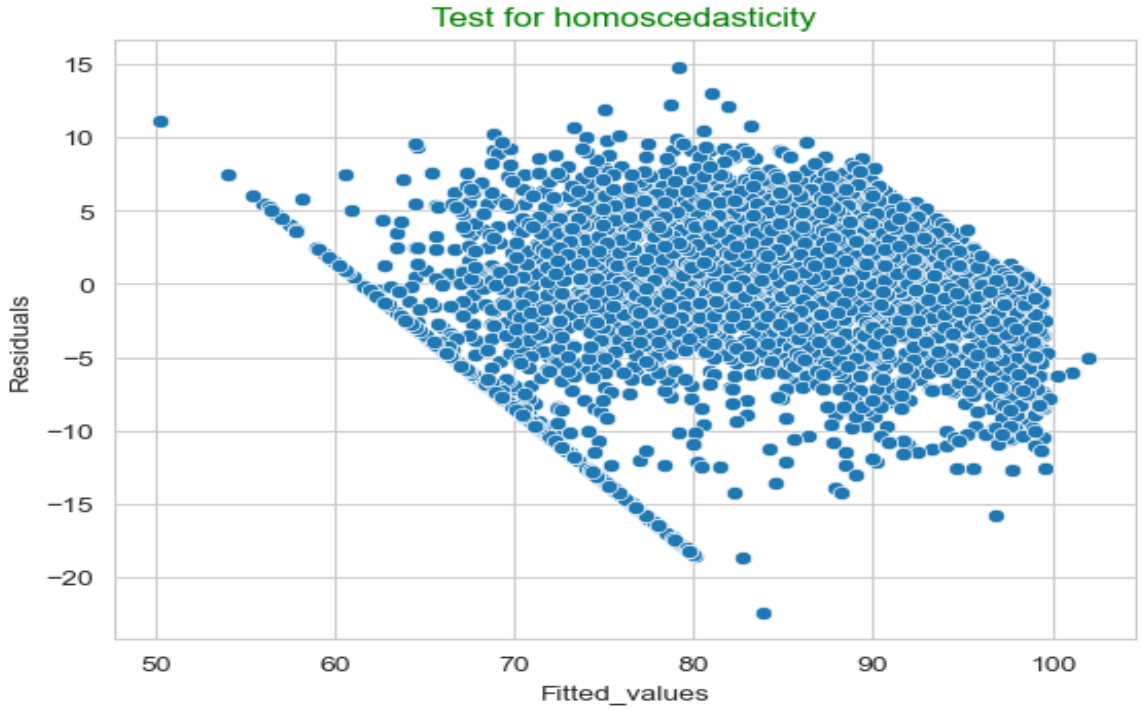


Fig 12: Test for homoscedasticity

Goldfeld Quandt test:

[('F statistic', 1.0988966818089467), ('p_value', 0.00590963103997442)]

- **All the assumptions of linear regression are now satisfied.**

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.**

Regression Plot between Actual Vs Fitted values:

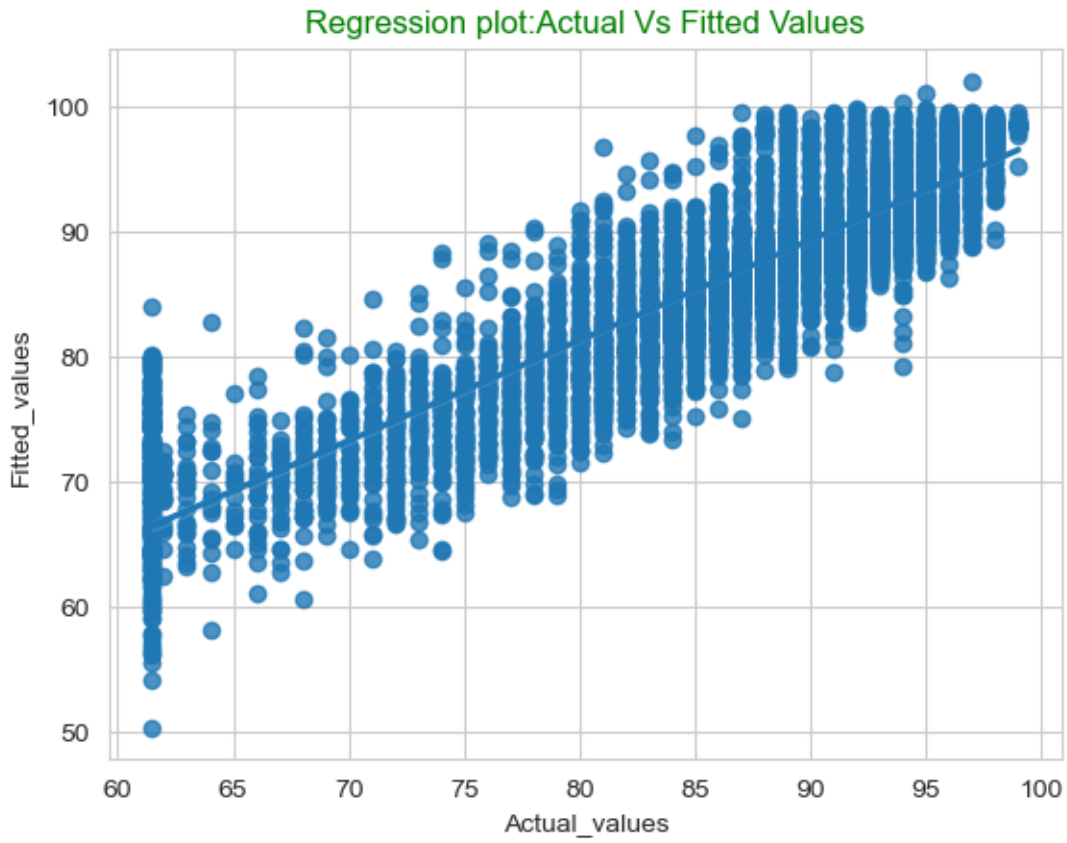


Fig 13: Regression Plot

Inferences:

- Since this is regression model, we have plotted the plot between predicted y values and the actual y values for the test dataset.
- From the plot, it is visible that the actual and the predicted values are close enough.
- This shows that the model performed well on the data.

❖ **Let us write the equation of linear regression:**

We get the following Linear Regression equation from the final model:

```
usr = 81.09011593776927 + -0.03359964065954595 * ( lwrite ) +  0.0015236175844430323 * ( scall ) +  -
0.00847434700315516 * ( swrite ) +  -0.4445753808970566 * ( exec ) +  -5.442968448708459e-06 * ( rcha
r ) +  -4.555519240157551e-06 * ( wchar ) +  -0.31964472067471217 * ( pgout ) +  0.5636772219282964 *
( atch ) +  -0.10732589945944634 * ( pgin ) +  -0.04128531280093563 * ( pflt ) +  -0.0003285562364779
736 * ( freemem ) +  9.476903926771167e-06 * ( freeswap ) +  1.917260136795174 * ( runqsz_Not_CPU_Bou
nd ) +  -3.4048825066243266e-07 * ( scall_sq )
```

❖ **Inferences:**

- When lwrite increases by 1 unit, usr decrease by 0.0335 units keeping all the predictors constant.
- When pgout increases by 1-unit usr increase by 0.5636 units keeping all the predictors constant.
- Similarly, we can write inference for all other predictors.
- runqsz_Not_CPU_Bound seems an important predictor as 1 unit increase in runqsz_Not_CPU_Bound usr increases by 1.917 units keeping all the predictors constant. (IMPORTANT ATTRIBUTE)

❖ **Business Insights and Recommendations:**

- The important Attributes are runqsz_Not_CPU_Bound, scall, exec, rchar and atch.
- Portion of time (%) that CPUs run in user mode and how each features effects can be predicted by the final Linear regression equation.
- runqsz_Not_CPU_Bound and atch seems an important predictor.

❖ **Various Steps performed in building Linear Regression Model:**

- 1st we load the dataset and performed data analysis by using info() and describe function().
- We Visualized the data using Univariate, Bivariate and Multivariate analysis.
- Done all the Missing value treatment, Outlier treatment, dummy encoding and converted the datatype to integer for our algorithm to perform model building.
- Train-Test split is done on the data for training and testing the model.
- We used both the sklearn and stats model algorithm to build the model.
- Variables are dropped to treat multicollinearity.
- Predictors having VIF > 5 are dropped such that it doesn't affect the adj. R-squared value.
- Assumptions of Linear Regression analysis done.
- Calculated the Performance metrices RMSE, MAE, Accuracy.

❖ **R-squared value of Final model:**

- R-squared value = 0.802
- Adj. R-squared value = 0.802

Problem 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.

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.

❖ Read the dataset:

1st five rows of the dataset:

	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard
0	24.0	Primary	Secondary	3.0	Scientology	No	2	
1	45.0	Uneducated	Secondary	10.0	Scientology	No	3	
2	43.0	Primary	Secondary	7.0	Scientology	No	3	
3	42.0	Secondary	Primary	9.0	Scientology	No	3	
4	36.0	Secondary	Secondary	8.0	Scientology	No	3	

Last five rows of the dataset:

	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standa
1468	33.0	Tertiary	Tertiary	NaN	Scientology	Yes	2	
1469	33.0	Tertiary	Tertiary	NaN	Scientology	No	1	
1470	39.0	Secondary	Secondary	NaN	Scientology	Yes	1	
1471	33.0	Secondary	Secondary	NaN	Scientology	Yes	2	
1472	17.0	Secondary	Secondary	1.0	Scientology	No	2	

❖ Info of the Dataset:

- There are 1473 Rows, and 10 columns present in the Dataset.
- Datatype of Husband_occupation column is Numeric here which we don't need to change it to Categorical column bcoz for model building we need numeric column only.
- Null Values are present in Wife age and No_of_children_born columns.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1473 entries, 0 to 1472
Data columns (total 10 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Wife_age                    1402 non-null   float64
 1   Wife_ education             1473 non-null   object
 2   Husband_education           1473 non-null   object
 3   No_of_children_born         1452 non-null   float64
 4   Wife_religion               1473 non-null   object
 5   Wife_Working                1473 non-null   object
 6   Husband_Occupation          1473 non-null   int64
 7   Standard_of_living_index    1473 non-null   object
 8   Media_exposure              1473 non-null   object
 9   Contraceptive_method_used   1473 non-null   object
dtypes: float64(2), int64(1), object(7)
memory usage: 115.2+ KB
```

❖ **Five-point summary:**

	count	mean	std	min	25%	50%	75%	max
Wife_age	1402.0	32.606277	8.274927	16.0	26.0	32.0	39.0	49.0
No_of_children_born	1452.0	3.254132	2.365212	0.0	1.0	3.0	4.0	16.0
Husband_Occupation	1473.0	2.137814	0.864857	1.0	1.0	2.0	3.0	4.0

Observations:

- Maximum no of children born is 16, which doesn't seem normal distribution in the data.
- This column shows the presence of outliers.
- Also, the column is Right skewed.

❖ **Check for duplicate data:**

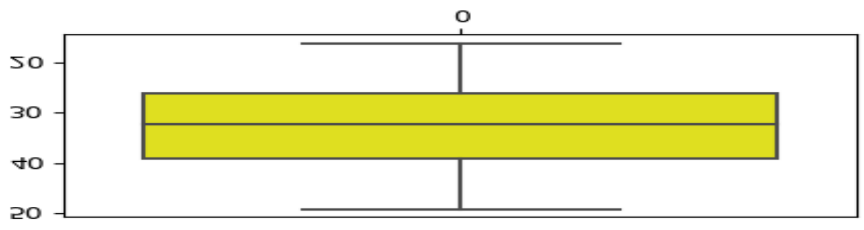
Number of duplicate rows = 80
(1473, 10)

❖ **Removing Duplicate Rows**

- It is necessary to treat duplicated rows as it will create bias in our model.
- It is better to remove the corresponding rows present in our dataset.
- Number of duplicate rows = 0
- (1393, 10)

❖ **Checking the spread of the data using boxplot for the continuous variables.**

Boxplot of Wife age:



Boxplot of No_of_children_born:

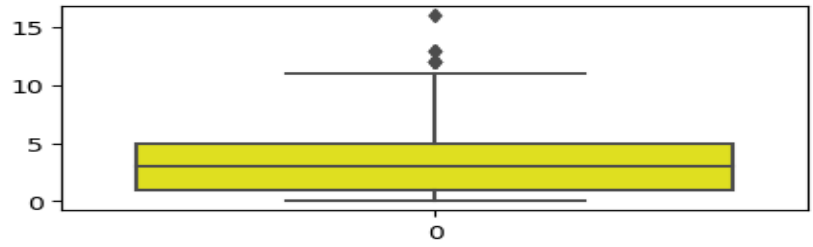


Fig 14: Boxplots

- Outlier is present in column 'No_of_children_born'.

❖ **Checking for Null values:**

```
Wife_age                        67
Wife_ education                  0
Husband_education                0
No_of_children_born             21
Wife_religion                    0
Wife_Working                     0
Husband_Occupation               0
Standard_of_living_index         0
Media_exposure                   0
Contraceptive_method_used        0
dtype: int64
```

❖ **Null Value Treatment:**

- We can impute the null values for 'wife age' column using mean method, since there are no Outliers present in this column as can be seen from the boxplot shown above.
- For 'No_of_children_born' column we can use median method to impute the Null values because of Outliers present in the column.

```
df[['Wife_age','No_of_children_born']].isnull().sum()
```

```
Wife_age              0
No_of_children_born    0
dtype: int64
```

Univariate Analysis:

❖ **Count plot of all the Categorical Variables.**

Observations:

- Tertiary level of education is higher for both Husband and Wife.
- There are a greater number of Educated men and Women as compared to Uneducated one.
- There are more Uneducated wives than Husbands.
- Most of them follow Scientology as religion.
- Non-working women are much higher as compared to working women.
- Major portion of the people are from the areas where the standard of living is Very High and High.

- Almost 227 people have 'Low' standards of living and 129 have 'very low' level of living.
- Majority of the women have used a contraceptive method, however there is a good proportion as well who have not used any.
- Majority of the people are Exposed to Media almost 1248 people as compared to 109 people who are not exposed to media.

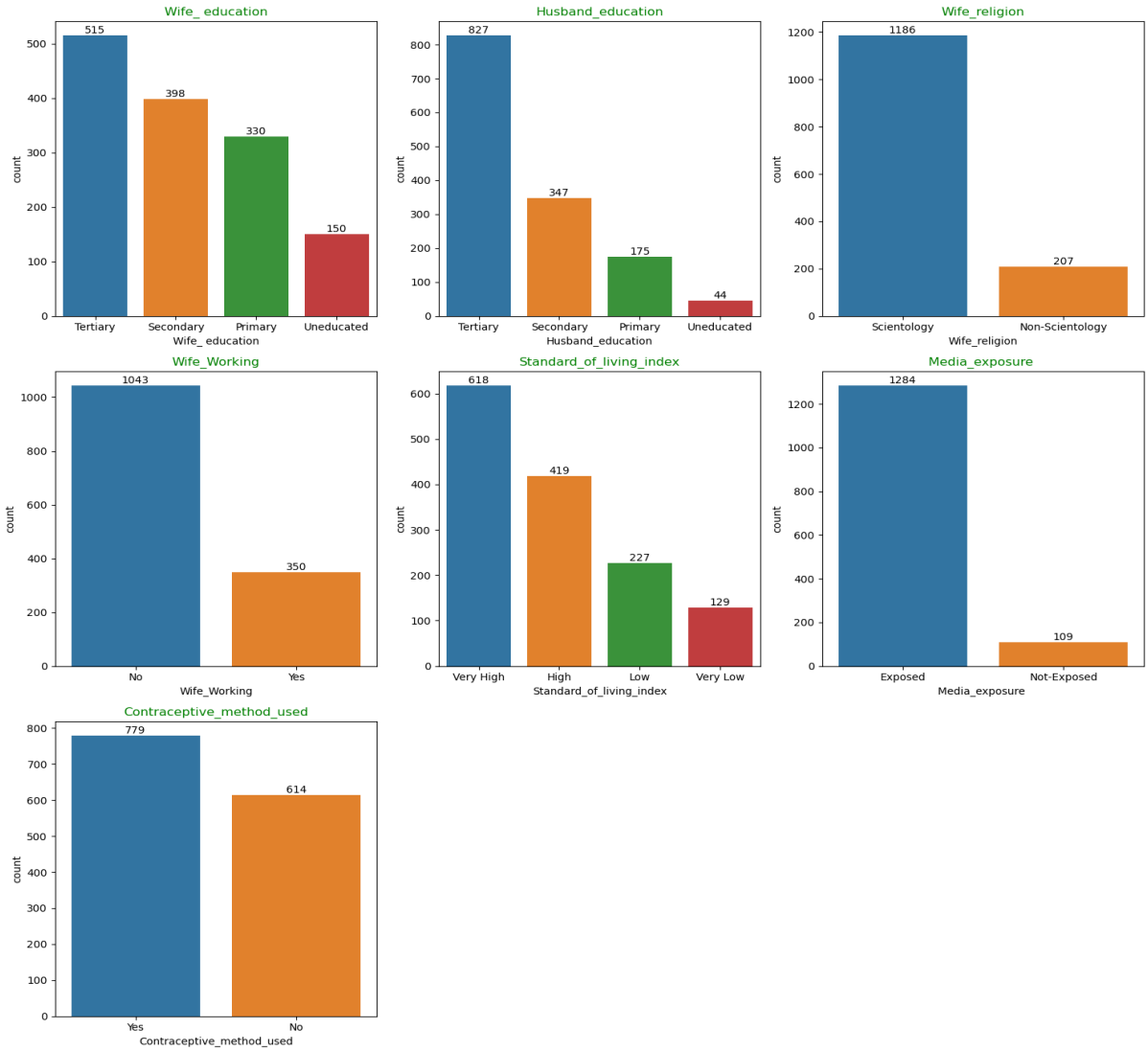


Fig 15: Countplots

❖ **Histogram of all continuous variables of the data frame:**

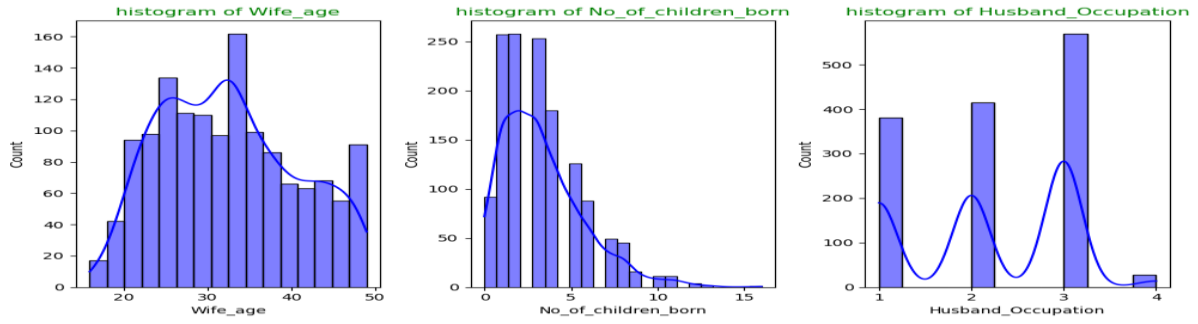


Fig 16: Histograms

Observations:

- 'No_of_children_born' columns is Right Skewed, showing the presence of Outliers.
- The age group of women is from 16 to 49.
- Majority of the people have 1 or 4 children, but a few have more than 10 children as well and max is 16.

Bivariate Analysis:

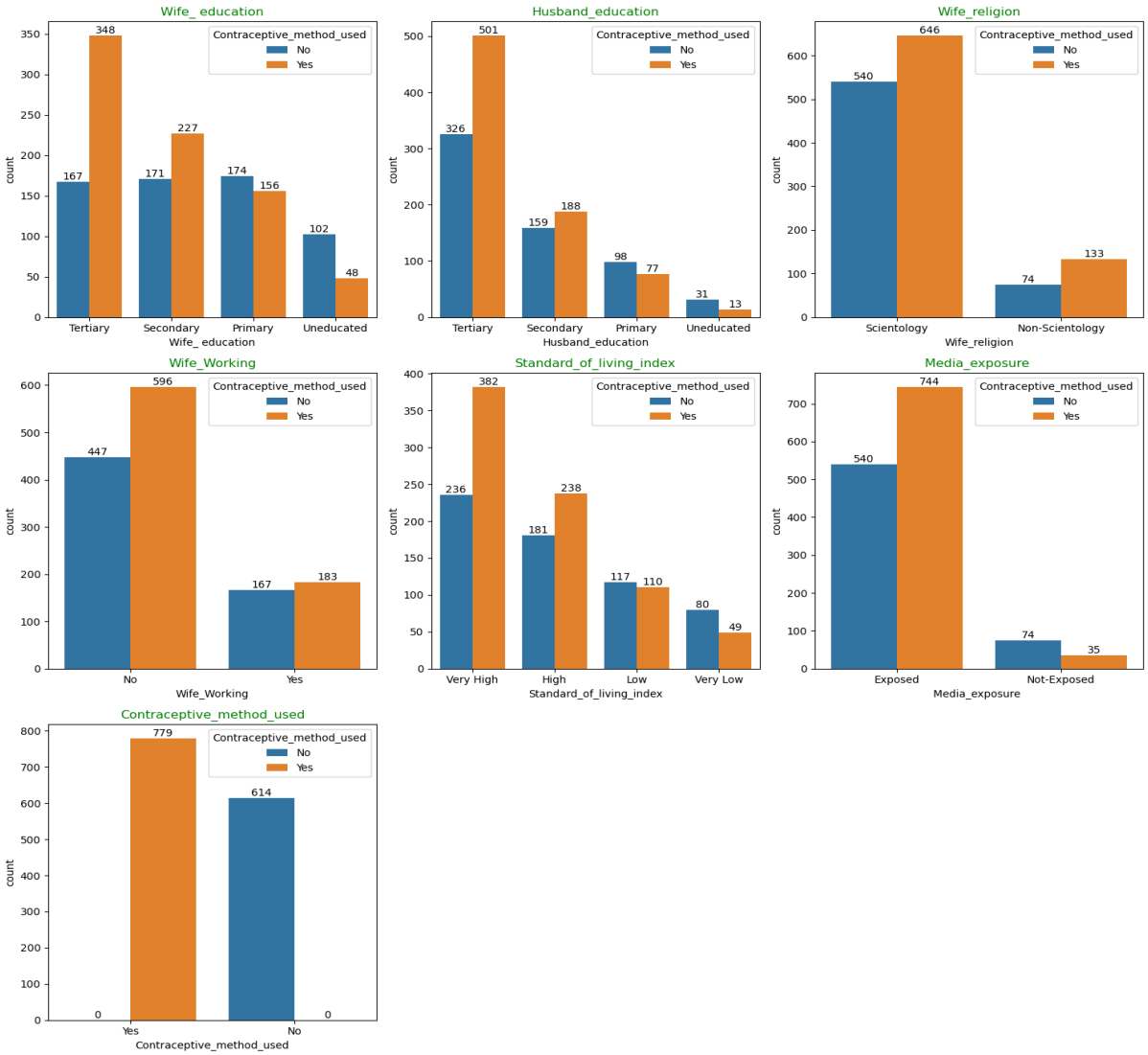


Fig 17: Bivariate Analysis

Observations:

- Women who have completed their secondary and Tertiary education use contraceptive methods more as compared to the others.
- Uneducated Women tends to use fewer contraceptive methods.
- Similar finding can be seen based on the Husband's education level.
- Women belonging to Scientology Religion tends to use more Contraceptives.
- Non-Working women uses more Contraceptives.
- People belonging to Very high standard of living use more contraceptives.

❖ Media Exposure plays an important role in using contraceptives.

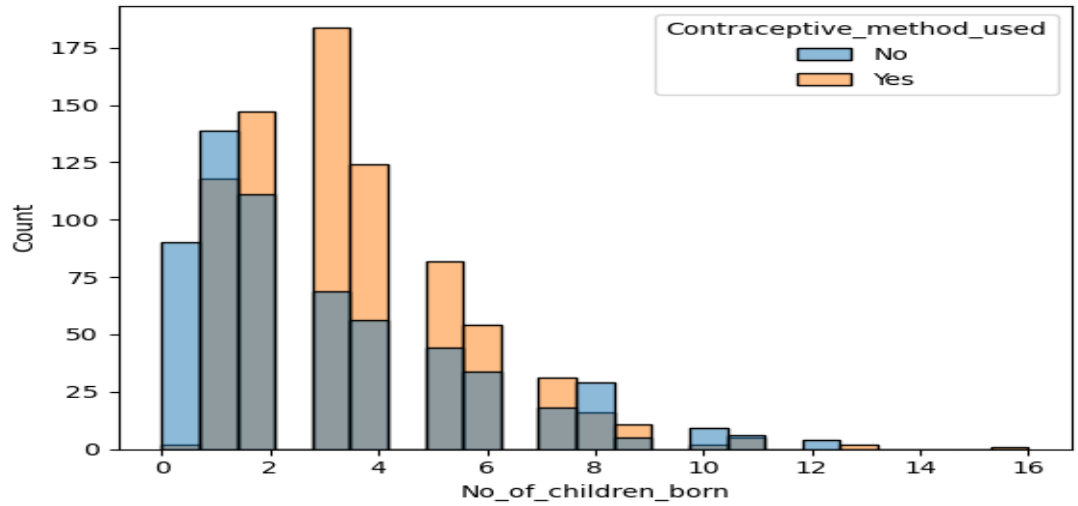


Fig 18 : Histograms

- Majority of the women are using contraceptives after 3 children.

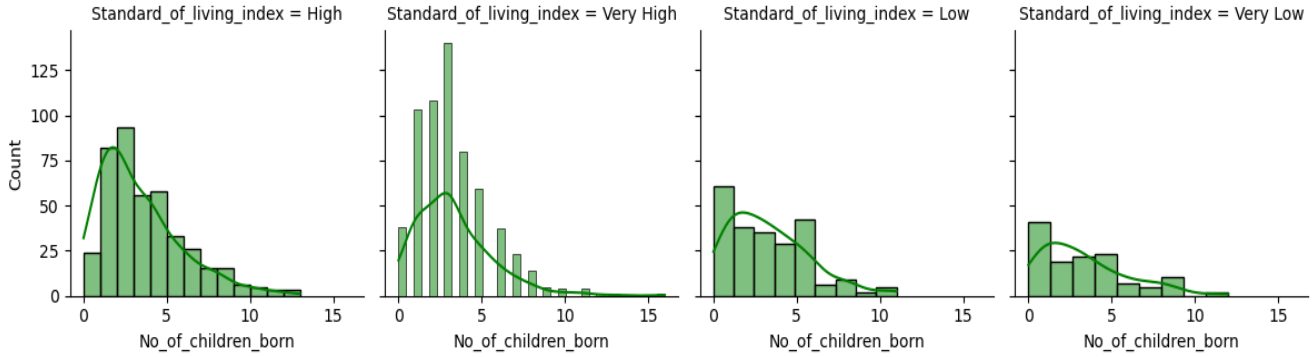


Fig 19: FacetGrid

- No_of_children_born is maximum whose standard of living index is very High.

❖ **Checking for Correlations:**

	Wife_age	No_of_children_born	Husband_Occupation
Wife_age	1.000000	0.528918	-0.185913
No_of_children_born	0.528918	1.000000	-0.024213
Husband_Occupation	-0.185913	-0.024213	1.000000

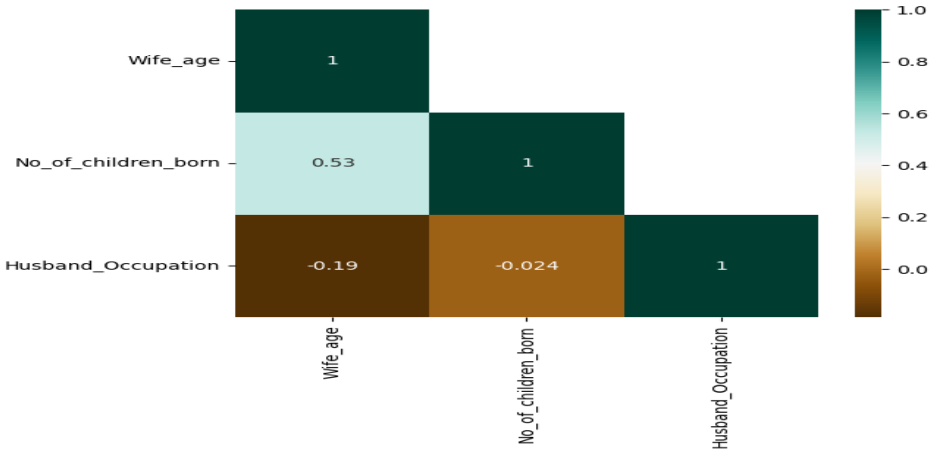


Fig 20: Heatmap-2

- There is some Correlation between Wife age and No_of_children_born.

Multivariate Analysis: Pair plot:

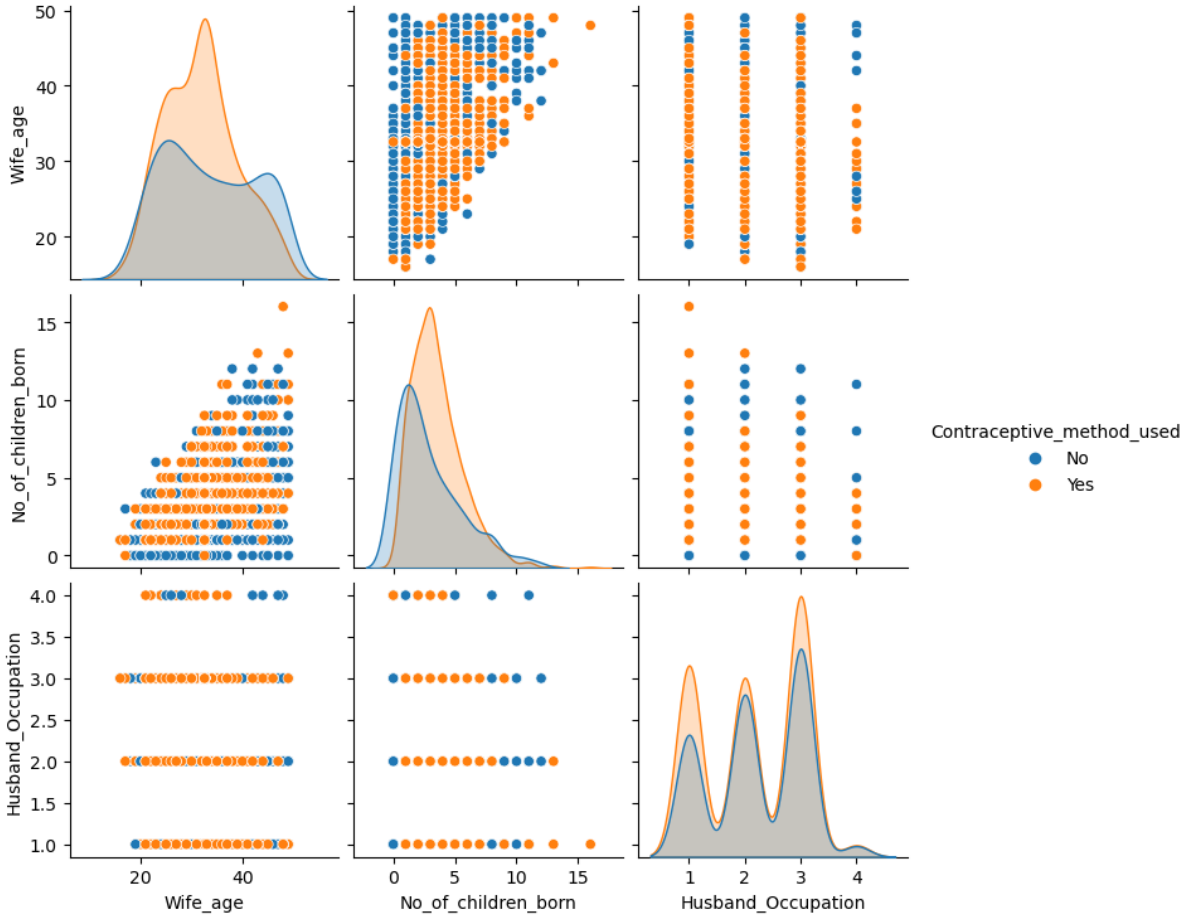


Fig 21: Pairplot-2

Observations:

- ❖ We can observe that as the Wife's age in increasing the number of children born is also increasing.
- ❖ We can also observe that child born is more in the case of Contraceptive method used which doesn't seem normal.
- ❖ This case may be due to the wrong data entered in the column as we can see there are also Outliers present in this column.

❖ **Outlier Treatment**

- We can treat Outliers with the following code. We will treat the outliers for the 'No_of_children_born' variable only.
- Treating Outlier using IQR.

❖ **Boxplot after Outlier Treatment:**

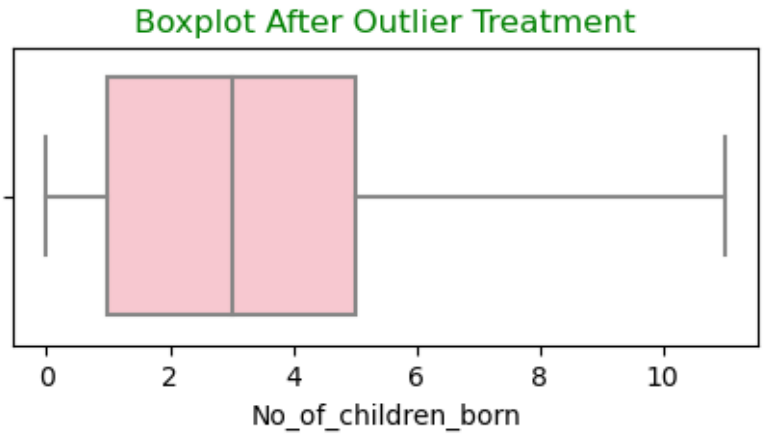


Fig 22: Boxplot after outlier Treatment-2

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.

❖ **Encoding The Data: -**

Converting all objects to categorical codes: Giving Ordinal numbers to these columns-->

- Wife's education (categorical) 1=uneducated, 2=Primary, 3=Secondary, 4=Tertiary
- Husband's education (categorical) 1=uneducated, 2=Primary, 3=Secondary, 4=Tertiary
- Standard-of-living index (categorical) 1=Very Low, 2=Low, 3=High, 4=Very High

❖ **Encoding Target Variable to 0 and 1:**

Dummy Variable Encoding:

- Converting the other 'object' type variables as dummy variables

Sample of Dataset after Encoding

	Wife_age	No_of_children_born	Husband_Occupation	Contraceptive_method_used	Wife_education_2	Wife_education_3	Wife_education_4	H
0	24.0	3.0	2	0	1	0	0	
1	45.0	10.0	3	0	0	0	0	
2	43.0	7.0	3	0	1	0	0	
3	42.0	9.0	3	0	0	1	0	
4	36.0	8.0	3	0	0	1	0	

- Now the Datatype of all the variables are Numeric and now we can proceed with Model Building steps.

❖ **Train Test Split:**

```
Number of rows and columns of the training set for the independent variables: (975, 15)
Number of rows and columns of the training set for the dependent variable: (975,)
Number of rows and columns of the test set for the independent variables: (418, 15)
Number of rows and columns of the test set for the dependent variable: (418,)
```

Logistic Regression Model

We built the Logistic Regression model with the Following Parameters:

```
[Parallel(n_jobs=2)]: Using backend LokyBackend with 2 concurrent workers.
[Parallel(n_jobs=2)]: Done   1 out of   1 | elapsed:    4.4s finished
```

```
LogisticRegression(max_iter=10000, n_jobs=2, penalty=None, solver='newton-cg',
                   verbose=True)
```

❖ **Model Evaluation:**

Finding Accuracy of Training and Test Data.

- Accuracy of Training Data: 0.6758974358974359
- Accuracy of Test Data: 0.6483253588516746

Building LDA Model

- All the Pre-processing step has already been performed above.
- Like EDA, Missing value treatment, Outlier treatment, Data Encoding and Train-Test split.
- Now we can proceed with Model Building i.e. LDA Model.
- Training Data Class Prediction with a cut-off value of 0.5
- Test Data Class Prediction with a cut-off value of 0.5

❖ **Accuracy of Train and Test data:**

- Accuracy of Train Data: 0.6738461538461539
- Accuracy of Test Data: 0.6435406698564593

Building CART Model

Note:

- Decision tree in Python can take only numerical / categorical columns. It cannot take string / object types.
- The data type of the dataset is converted into integer and now we can proceed with the Model Building.

❖ A CART model is also built using the following parameters:

- criterion = 'gini',
- max_depth = 7,
- min_samples_leaf=30,
- min_samples_split=50

❖ **Accuracy of the Train and Test Dataset:**

- Accuracy of Training data: 0.7425641025641025
- Accuracy of Test data: 0.6602870813397129

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

❖ **Performance Metrices:**

Logistic Regression

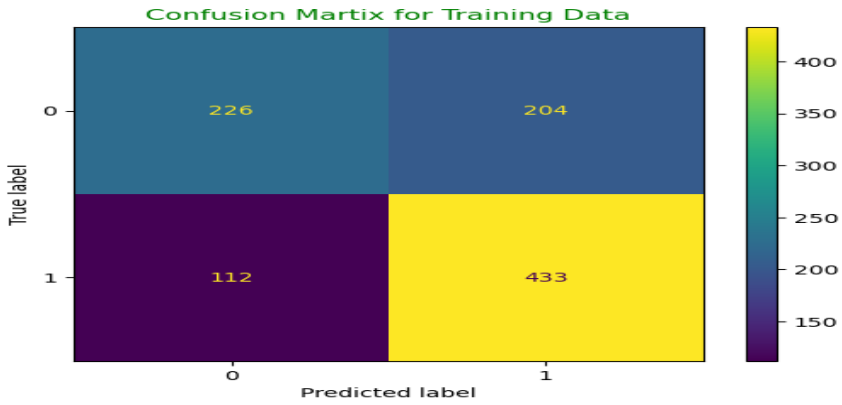
Classification Report of Training Data:

	precision	recall	f1-score	support
0	0.67	0.53	0.59	430
1	0.68	0.79	0.73	545
accuracy			0.68	975
macro avg	0.67	0.66	0.66	975
weighted avg	0.67	0.68	0.67	975

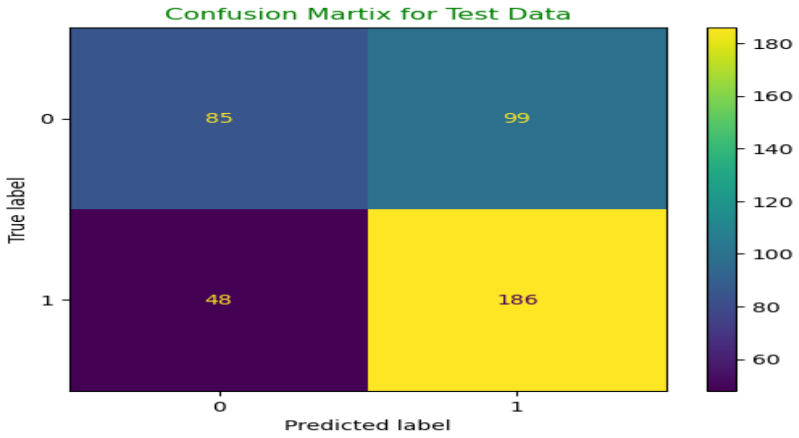
Classification Report of Test Data:

	precision	recall	f1-score	support
0	0.64	0.46	0.54	184
1	0.65	0.79	0.72	234
accuracy			0.65	418
macro avg	0.65	0.63	0.63	418
weighted avg	0.65	0.65	0.64	418

❖ **Confusion Matrix for the Training data:**



❖ **Confusion Matrix for the Test data:**



Inferences:

- The confusion matrix of train data shows that the model predicted 433 True positives values, 226 True negative, 204 false positive and 112 false Negative.
- The confusion matrix of test data shows that the model predicted 186 True positives values, 85 True negative,99 false positive and 48 false Negative.

AUC and ROC:

❖ AUC Value closer to 1 tells that there is good separability between the predicted classes and thus the model is good for prediction.

- ROC Curve visually represents the above concept where the plot should be as far as possible from the diagonal.
- ROC is a probability curve and AUC represents the classification model's ability to separate the two classes.
- The higher the AUC, the more powerful is the model to predict true class membership.

AUC and ROC for the training data : AUC: 0.720

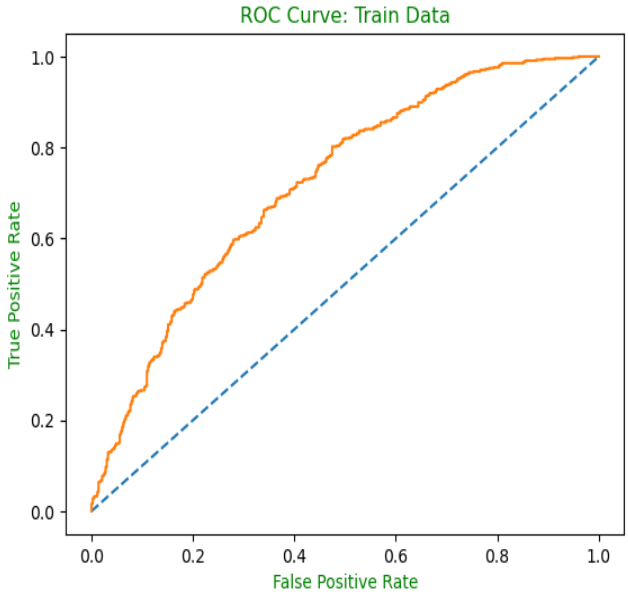


Fig 23: ROC train LR

AUC and ROC for the test data: AUC: 0.665

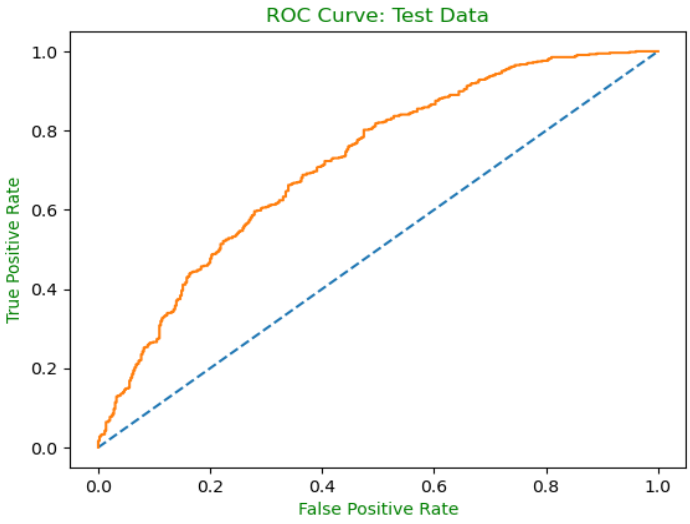


Fig 24: ROC test LR

Observations:

- AUC for Train data is 72% and AUC for Test data is 66%.

- From the AUC values & ROC curve for the Train and Test data, it shows that Test data is not covering a large area as compared to the train data.
- Therefore, there is a need to optimize this model.

❖ **Important Features and their Coefficients:**

```
The coefficient for Wife_age is -0.07521379242680283
The coefficient for Wife_ education is 0.5213124033028135
The coefficient for Husband_education is 0.04294381833842849
The coefficient for No_of_children_born is 0.33245722288877055
The coefficient for Husband_Occupation is 0.14547576593766215
The coefficient for Standard_of_living_index is 0.31190496477110663
The coefficient for Wife_religion_Scientology is -0.4332499416554644
The coefficient for Wife_Working_Yes is -0.16919269842352977
The coefficient for Media_exposure _Not-Exposed is -0.3479222248439086
```

Optimized Logistic Regression Model:

❖ Applying GridSearchCV for Logistic Regression: using following parameters

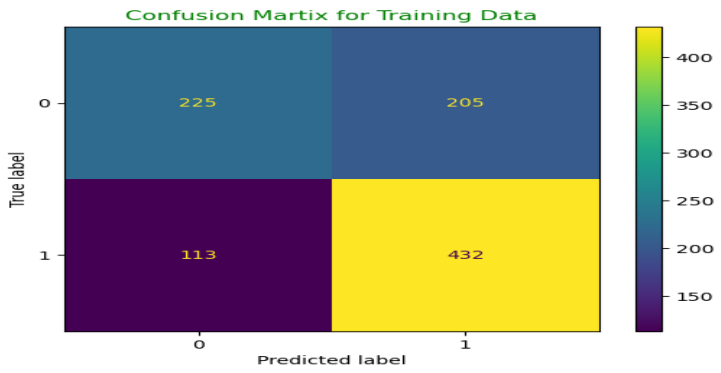
```
GridSearchCV(cv=3, estimator=LogisticRegression(max_iter=10000, n_jobs=2),
             n_jobs=-1,
             param_grid={'penalty': ['l2', 'none'], 'solver': ['lbfgs', 'sag'],
                         'tol': [0.0001, 1e-05]},
             scoring='f1')
```

After GridSearchCV we got the following hyperparameters for the best model.

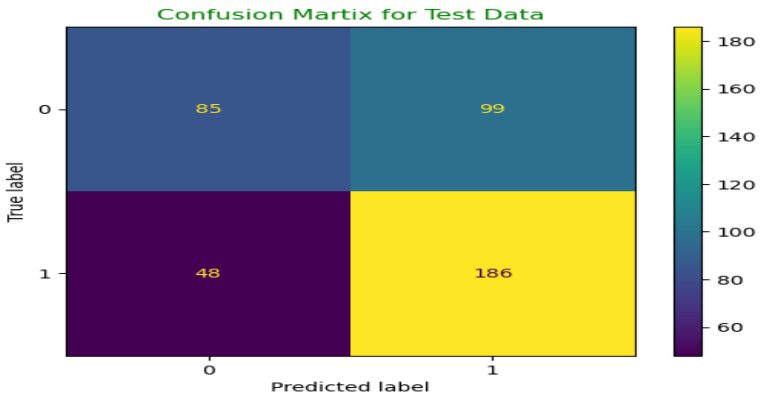
```
{'penalty': 'l2', 'solver': 'sag', 'tol': 0.0001}

LogisticRegression(max_iter=10000, n_jobs=2, solver='sag')
```

- **Confusion matrix on the training data:**



❖ **Confusion matrix on the test data:**



❖ **Classification Report of Training Data:**

	precision	recall	f1-score	support
0	0.67	0.52	0.59	430
1	0.68	0.79	0.73	545
accuracy			0.67	975
macro avg	0.67	0.66	0.66	975
weighted avg	0.67	0.67	0.67	975

❖ **Classification Report of Test Data:**

	precision	recall	f1-score	support
0	0.64	0.46	0.54	184
1	0.65	0.79	0.72	234
accuracy			0.65	418
macro avg	0.65	0.63	0.63	418
weighted avg	0.65	0.65	0.64	418

Inferences:

- The confusion matrix of train data shows that the model predicted 433 True positives values, 225 True negative, 205 false positive and 112 false Negative.
- The confusion matrix of test data shows that the model predicted 186 True positives values, 84 True negative,100 false positive and 48 false Negative.

AUC-ROC Curve:

- AUC Score Train Data: 0.720
- AUC Score Test Data: 0.665

Inferences:

For predicting Contraceptive_method_used = yes (label 1)

- **Precision (65%)** -- 65% of the Women predicted are using Contracepltive_method out of all the women predicted that use Contraceptive.
- **Recall (79%)** -- Out of all the Women using Contraceptives 79% of the Women have been predicted correctly.

For predicting Contraceptive_method_used = No (label 0)

- **Precision (64)** -- 64% of the Women predicted are not using Contracepltive_method out of all the women predicted that don't use Contraceptive.
- **Recall (46%)** -- Out of all the Women not using Contraceptives 46% of the Women have been predicted correctly.

Overall accuracy of the model – 65 % of total predictions are correct.

Accuracy, AUC, Precision and Recall for test data is almost inline with training data. This proves no overfitting or underfitting has happened, and overall, the model is a good model for classification.

LDA: Performance Matrix

Prediction and Evaluation on both Training and Test Set using Confusion Matrix, Classification Report and AUC-ROC.

❖ **Classification Report:**

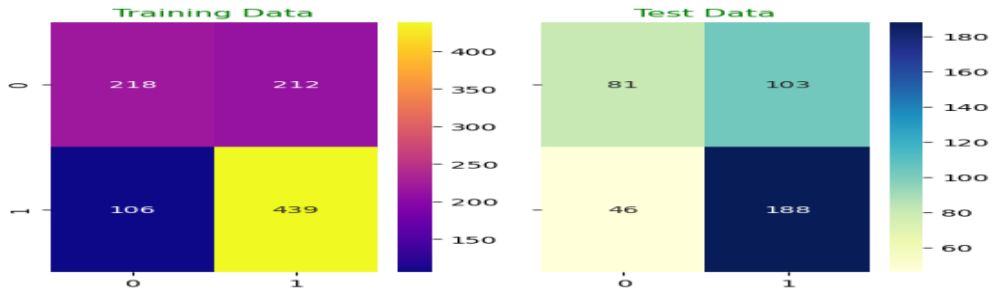
Classification Report of the training data:

	precision	recall	f1-score	support
0	0.67	0.51	0.58	430
1	0.67	0.81	0.73	545
accuracy			0.67	975
macro avg	0.67	0.66	0.66	975
weighted avg	0.67	0.67	0.67	975

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.64	0.44	0.52	184
1	0.65	0.80	0.72	234
accuracy			0.64	418
macro avg	0.64	0.62	0.62	418
weighted avg	0.64	0.64	0.63	418

Confusion Matrix for the Training data and Test Data:



AUC and ROC for the training and Test data ~ Repeater Operator Curve

- AUC for the Training Data: 0.720
- AUC for the Test Data: 0.663

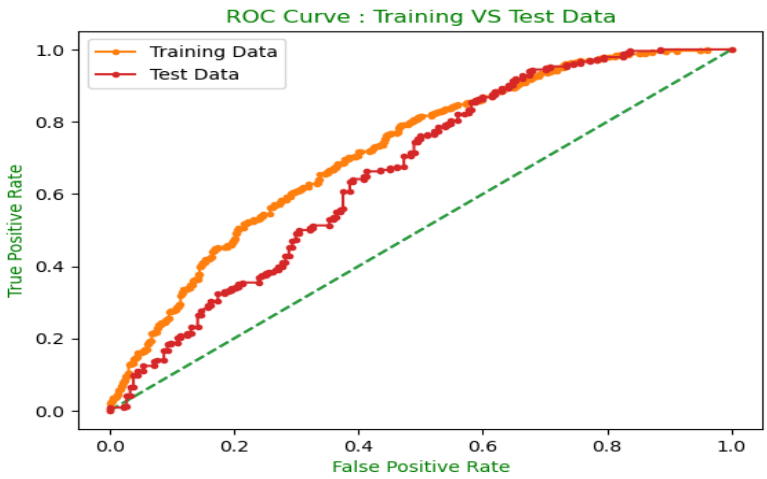


Fig 25: ROC LDA

Inferences:

For predicting Contraceptive_method_used = yes (label 1)

- Precision (65%) -- 65% of the Women predicted are using Contracepltive_method out of all the women predicted that use Contraceptive.
- Recall (80%) -- Out of all the Women using Contraceptives 79% of the Women have been predicted correctly.

For predicting Contraceptive_method_used = No (label 0)

- Precision (64) -- 64% of the Women predicted are not using Contracepltive_method out of all the women predicted that don't use Contraceptive.
- Recall (44%) -- Out of all the Women not using Contraceptives 46% of the Women have been predicted correctly.

➢ **Overall accuracy of the model – 67 % of total predictions are correct.**

CART: Performance Matrix

Prediction and Evaluation on both Training and Test Set using Confusion Matrix, Classification Report and AUC-ROC.

- **Accuracy of the Train and Test Dataset:**
  - Accuracy of Training data: 0.7425641025641025
  - Accuracy of Test data: 0.6602870813397129

- **Variable Importance: important features of the model are:**

	Imp
Wife_age	0.296765
Wife_ education	0.161951
Husband_education	0.038190
No_of_children_born	0.461566
Husband_Occupation	0.015188
Standard_of_living_index	0.026341
Wife_religion_Scientology	0.000000
Wife_Working_Yes	0.000000
Media_exposure _Not-Exposed	0.000000

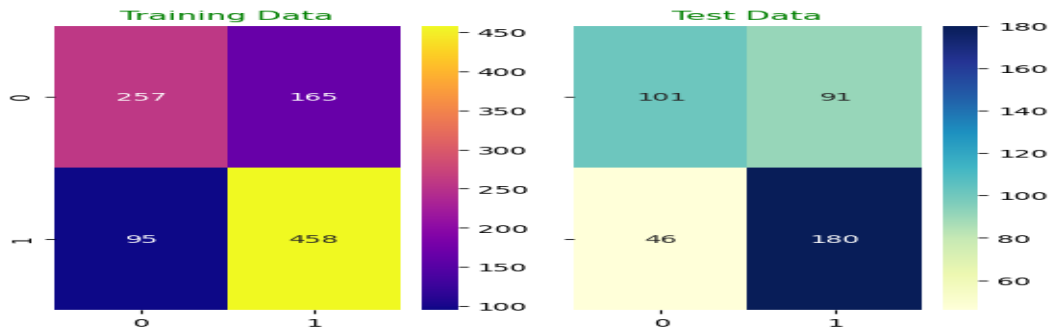
- **Classification Report of the Training data:**

	precision	recall	f1-score	support
0	0.73	0.61	0.66	422
1	0.74	0.83	0.78	553
accuracy			0.73	975
macro avg	0.73	0.72	0.72	975
weighted avg	0.73	0.73	0.73	975

- **Classification Report of the Test data:**

	precision	recall	f1-score	support
0	0.69	0.53	0.60	192
1	0.66	0.80	0.72	226
accuracy			0.67	418
macro avg	0.68	0.66	0.66	418
weighted avg	0.67	0.67	0.67	418

❖ **Confusion Matrix for Training and Test data:**



❖ **Measuring AUC-ROC Curve - Training Data:**

AUC of Training Data: 0.801

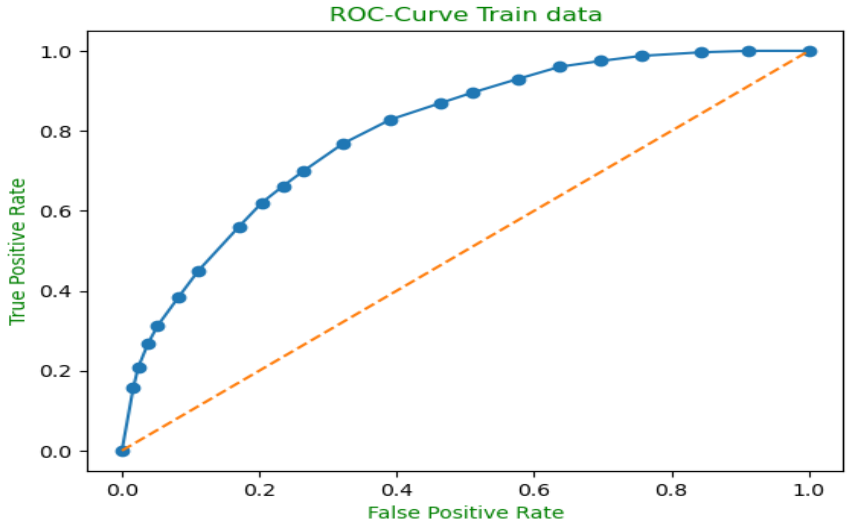


Fig 26: ROC Train CART

AUC of Test Data: 0.721

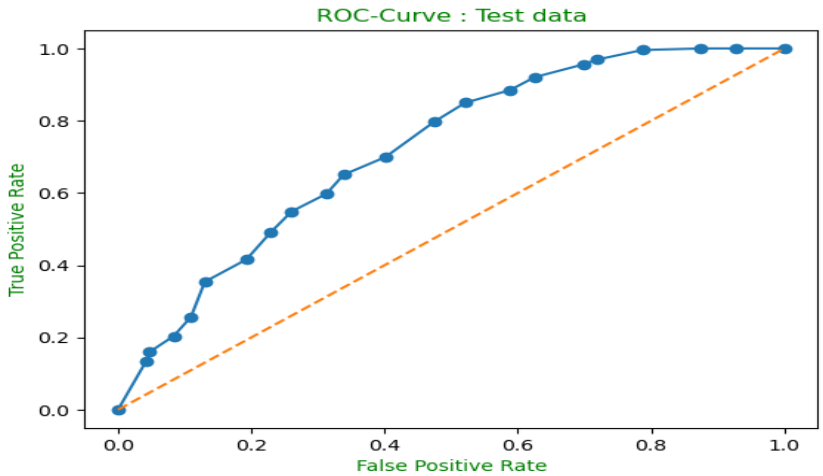


Fig 27: ROC Test CART

Conclusion:

- Accuracy on the Training Data: 73.33%
- Accuracy on the Test Data: 67.22%

- AUC on the Training Data: 80%
- AUC on the Test: 72%

- Accuracy, AUC, Precision and Recall for test data is almost in line with training data.
- Wife age, No_of_children_born, wife education and Husband_education are the most important features in determining the Contraceptive method used.

Inferences:

For predicting Contraceptive_method_used = yes (label 1)

- Precision (66%) -- 66% of the Women predicted are using Contracepltive_method out of all the women predicted that use Contraceptive.
- Recall (80%) -- Out of all the Women using Contraceptives 80% of the Women have been predicted correctly.

For predicting Contraceptive_method_used = No (label 0)

- Precision (69%) -- 69% of the Women predicted are not using Contracepltive_method out of all the women predicted that don't use Contraceptive.
- Recall (53%) -- Out of all the Women not using Contraceptives 53% of the Women have been predicted correctly.

➢ Overall accuracy of the model – 67 % of total predictions are correct.

Optimizing CART Model:

❖ **Applying GridSearchCV on CART Model:**

```
GridSearchCV(cv=3, estimator=DecisionTreeClassifier(), n_jobs=-1,
             param_grid={'criterion': ['gini', 'entropy', 'log_loss'],
                         'max_depth': [10, 20, 30],
                         'min_impurity_decrease': [0.001, 0.0001],
                         'min_samples_leaf': [2, 3, 4, 5],
                         'min_samples_split': [2, 3, 5, 7]},
             scoring='accuracy', verbose=1)
```

After applying the GridSearchCV we got the following best parameter for our model:

```
DecisionTreeClassifier
DecisionTreeClassifier(criterion='entropy', max_depth=10,
                       min_impurity_decrease=0.0001, min_samples_leaf=3,
                       min_samples_split=5)
```

❖ **Accuracy of the Training and Test data:**

- Accuracy of Train data: 0.8143589743589743
- Accuracy of Test data: 0.6363636363636364

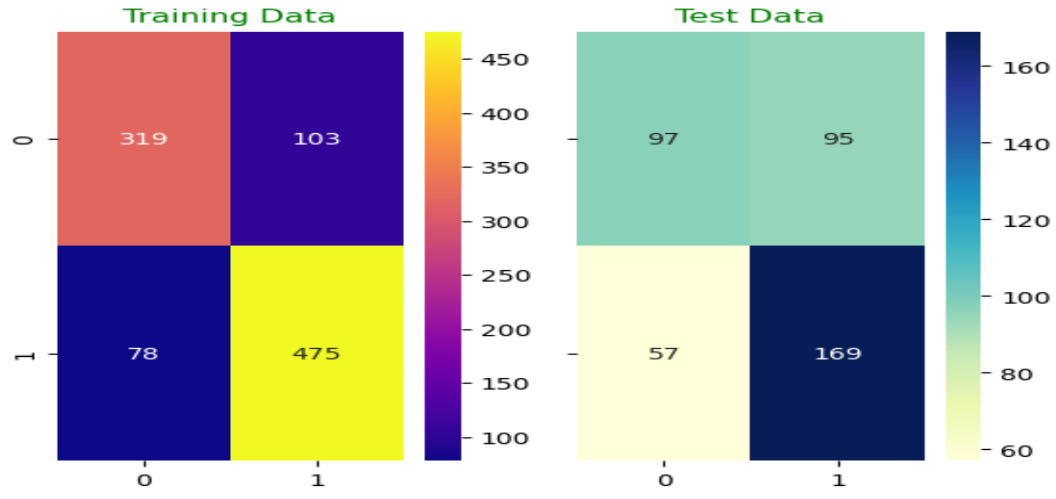
❖ **Classification Report of the training data:**

	precision	recall	f1-score	support
0	0.80	0.76	0.78	422
1	0.82	0.86	0.84	553
accuracy			0.81	975
macro avg	0.81	0.81	0.81	975
weighted avg	0.81	0.81	0.81	975

❖ **Classification Report of the test data:**

	precision	recall	f1-score	support
0	0.63	0.51	0.56	192
1	0.64	0.75	0.69	226
accuracy			0.64	418
macro avg	0.64	0.63	0.63	418
weighted avg	0.64	0.64	0.63	418

❖ **Confusion Matrix of Train and Test Data:**



❖ **AUC-ROC Curve:**

- AUC for the Training Data: 0.902
- AUC for the Test Data: 0.655

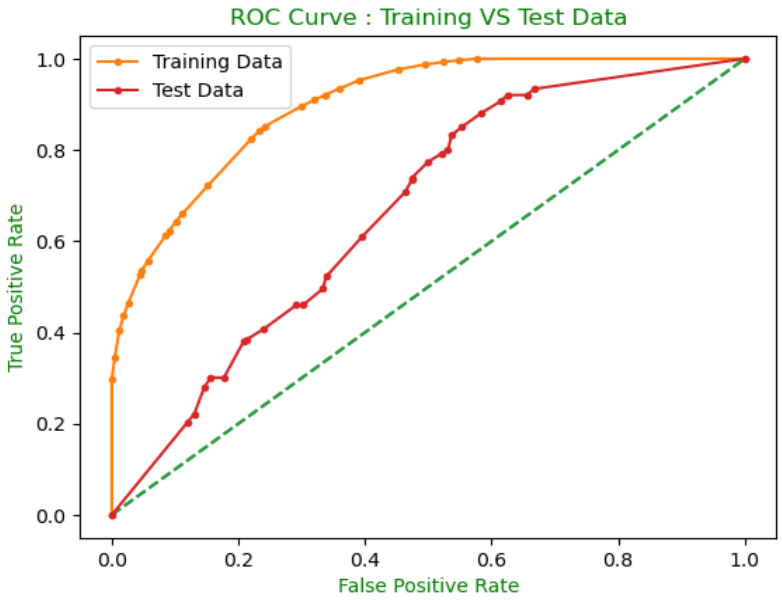


Fig 28: ROC Optimized CART

❖ **Variable Importance: Optimized Model**

	Imp
Wife_age	0.290017
Wife_ education	0.097229
Husband_education	0.072717
No_of_children_born	0.323562
Husband_Occupation	0.098554
Standard_of_living_index	0.070124
Wife_religion_Scientology	0.012375
Wife_Working_Yes	0.016265
Media_exposure _Not-Exposed	0.019156

- 'Wife age', 'No_of_children_born', 'wife education' and 'Husband_education'(in same order of preference) are the most important variables in determining if a women uses Contraceptive methods.

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.

Inferences:

❖ **Comparison Between Models:**

- Cart stands the top model with Accuracy of 73% on train data and 67% on test data.
- Optimized Logistic Regression Model's Accuracy is 67% on Train data and 65% on Test data same is for LDA model as well.
- CART with GridSearchCV model's Accuracy is 81% on Train data and 63* on test data, The accuracy of test data is not comparable with train data this may be due to Overfitting.

- **Comparison in Terms of Accuracy, Recall, Precision and AUC value:**
  - The CART model from all the other models seems to be performing the best in terms of Accuracy, Recall and Precision values.
  - The CART model also gives the most important features according to which the split in the Decision Tree was made.
  - 'Wife age', 'No_of_children_born', 'wife education' and 'Husband_education'(in same order of preference) are the most important variables in determining if a women uses Contraceptive methods.
  - If we look at the Recall value, the CART model is able to identify 80% of the true positives correctly. The LDA model also gives a similar Recall value of 79%.
  - The Accuracy of the CART model is slightly greater than the other two models, therefore it is better to consider the CART model prediction.
  - Similarly, we see that the Area Under the Curve (AUC) captured is 80% for train data and 72% for the test data.
  - AUC for the Optimized CART model is 90% for the Train data and 65% for the Test data.
  - These values are greater than the other two models.

- **Recommendations:**
  - We saw that wife age, No_of_children_born, Wife education and Husband_education is an important factor in deciding whether to use contraceptives or not.
  - Since Uneducated Women tends to use fewer contraceptive methods there should be a campaign run to educate these women on how and why to use contraceptives.
  - People belonging to Very high standard of living use more contraceptives.
  - Media Exposure plays an important role in using contraceptives methods.
- V**arious Steps performed in building Model's:**

  - We performed all the Data analysis steps that is required for the model building.
  - Train-Test Split is done.
  - Applied various Algorithm for model building.
  - Checked the performance of various model's using Accuracy, precision, Recall and AUC score.

Thank You