PM PROJECT REPORT

Rohit Nagarahalli

Problem 1

Executive Summary:

The comp-activ database comprises activity measures of computer systems. Data was gathered from a Sun Sparcstation 20/712 with 128 Mbytes of memory, operating in a multi-user university department. Users engaged in diverse tasks, such as internet access, file editing, and CPU-intensive programs.

Introduction:

The aim is to establish a linear equation for predicting 'usr' (the percentage of time CPUs operate in user mode). Also, to analyse various system attributes to understand their influence on the system's 'usr' mode.

Data shape:

The data shape by default is said to have 8192 records and 22 variables. However, variables such as 'rchar' and 'wchar' have 8088 and 8017 records respectively considering the rest as missing values/records which shall be addressed later.

Data types:

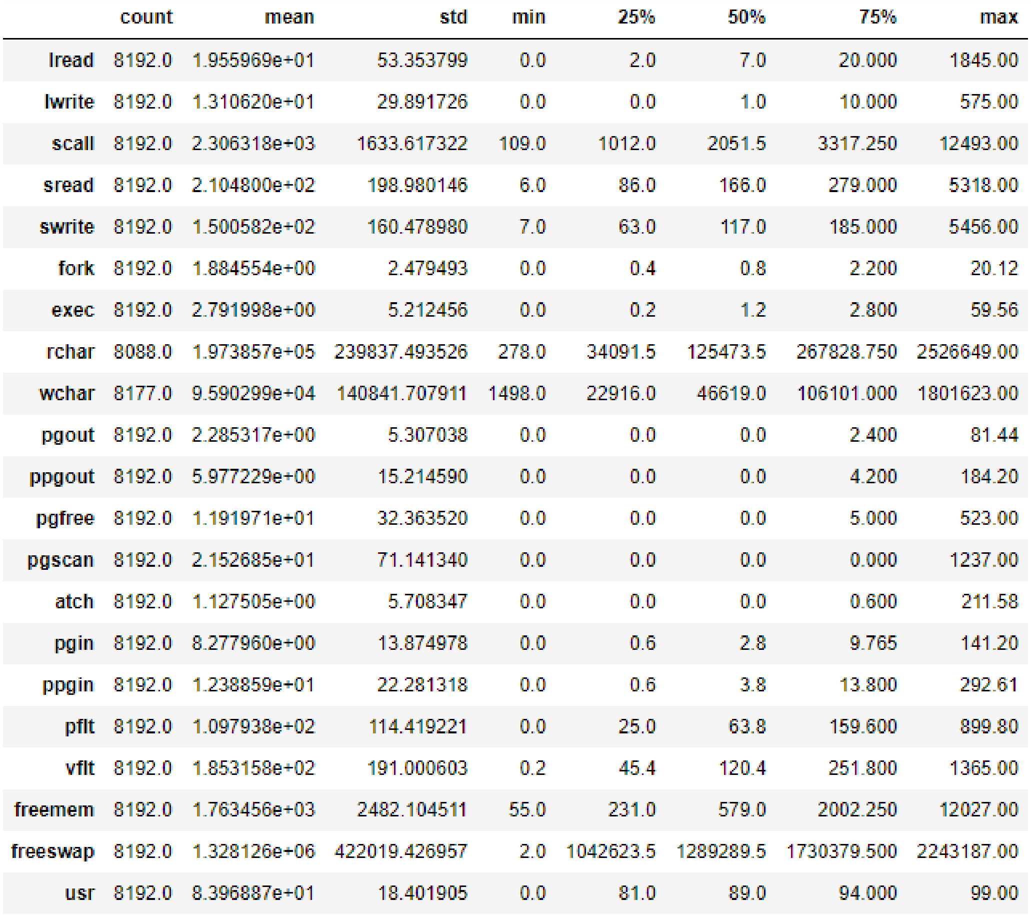
```
#   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   lread     8192 non-null   int64
 1   lwrite    8192 non-null   int64
 2   scall     8192 non-null   int64
 3   sread     8192 non-null   int64
 4   swrite    8192 non-null   int64
 5   fork      8192 non-null   float64
 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
```

```
#   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 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
```

Table 1 Data Types

From the above table 1 it can be noted that there are a total of 21 variables that belong to numerical data type (int64 or float64). The variable 'runqsz' is an object which is actually a category of CPU_Bound and Not_CPU_Bound. However, the runqsz variable will later be converted to a numeciral category (0 or 1).

Statistical Summary:



	count	mean	std	min	25%	50%	75%	max
lread	8192.0	1.955969e+01	53.353799	0.0	2.0	7.0	20.000	1845.00
lwrite	8192.0	1.310620e+01	29.891726	0.0	0.0	1.0	10.000	575.00
scall	8192.0	2.306318e+03	1633.617322	109.0	1012.0	2051.5	3317.250	12493.00
sread	8192.0	2.104800e+02	198.980146	6.0	86.0	166.0	279.000	5318.00
swrite	8192.0	1.500582e+02	160.478980	7.0	63.0	117.0	185.000	5456.00
fork	8192.0	1.884554e+00	2.479493	0.0	0.4	0.8	2.200	20.12
exec	8192.0	2.791998e+00	5.212456	0.0	0.2	1.2	2.800	59.56
rchar	8088.0	1.973857e+05	239837.493526	278.0	34091.5	125473.5	267828.750	2526649.00
wchar	8177.0	9.590299e+04	140841.707911	1498.0	22916.0	46619.0	106101.000	1801623.00
pgout	8192.0	2.285317e+00	5.307038	0.0	0.0	0.0	2.400	81.44
ppgout	8192.0	5.977229e+00	15.214590	0.0	0.0	0.0	4.200	184.20
pgfree	8192.0	1.191971e+01	32.363520	0.0	0.0	0.0	5.000	523.00
pgscan	8192.0	2.152685e+01	71.141340	0.0	0.0	0.0	0.000	1237.00
atch	8192.0	1.127505e+00	5.708347	0.0	0.0	0.0	0.600	211.58
pgin	8192.0	8.277960e+00	13.874978	0.0	0.6	2.8	9.765	141.20
ppgin	8192.0	1.238859e+01	22.281318	0.0	0.6	3.8	13.800	292.61
pflt	8192.0	1.097938e+02	114.419221	0.0	25.0	63.8	159.600	899.80
vflt	8192.0	1.853158e+02	191.000603	0.2	45.4	120.4	251.800	1365.00
freemem	8192.0	1.763456e+03	2482.104511	55.0	231.0	579.0	2002.250	12027.00
freeswap	8192.0	1.328126e+06	422019.426957	2.0	1042623.5	1289289.5	1730379.500	2243187.00
usr	8192.0	8.396887e+01	18.401905	0.0	81.0	89.0	94.000	99.00

Table 2 Statistical Summary

It can be inferred from the statistical that most of the variable's statistic is abnormal and this is because of the presence of zeros in most of the columns. The importance of zero and whether or not they need to imputed will be evaluated approach wise. Since, there has been no domain information on zero model needs to be built considering both the scenarios.

Univariate Analysis:

There are a total of 21 columns which are numerical and continuous. A violin plot would do justice to the analysis since it not only describes the distribution but also the outlying points, skewness and also the normality.

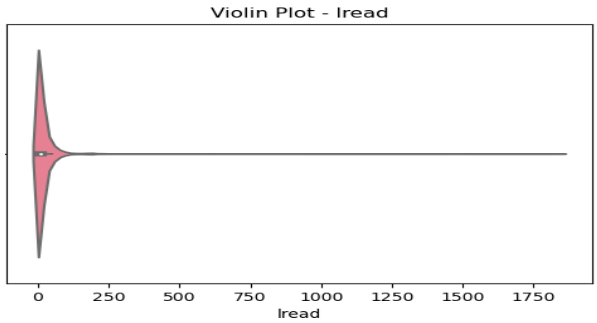


Figure 1 lread

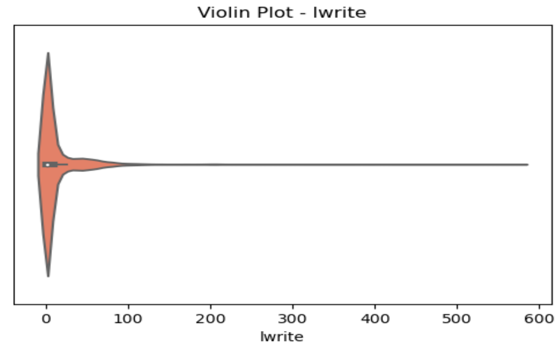


Figure 2 lwrite

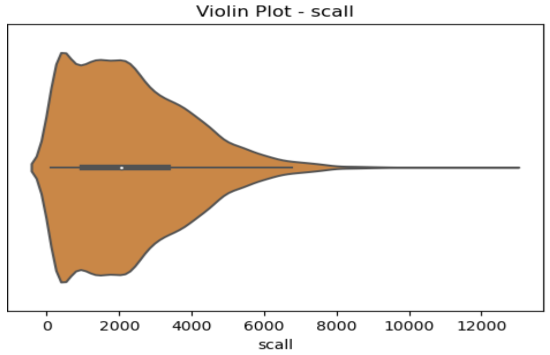


Figure 3 scall

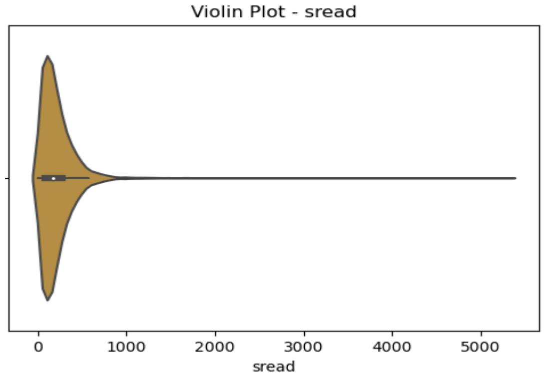


Figure 4 sread

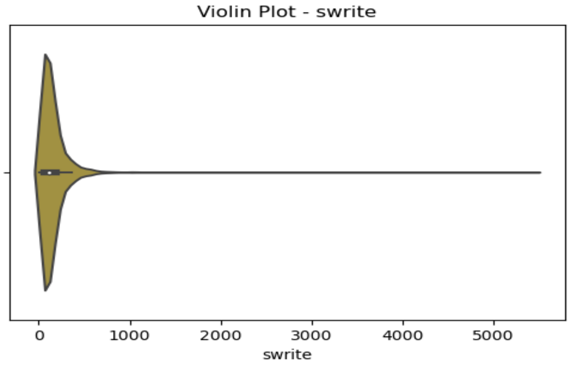


Figure 5 swrite

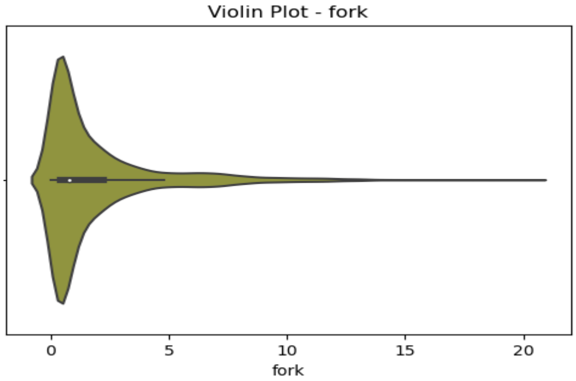


Figure 6 fork

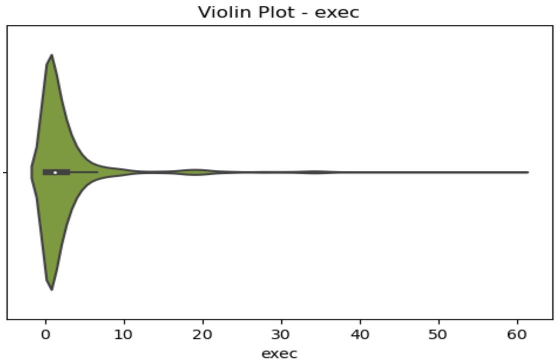


Figure 7 exec

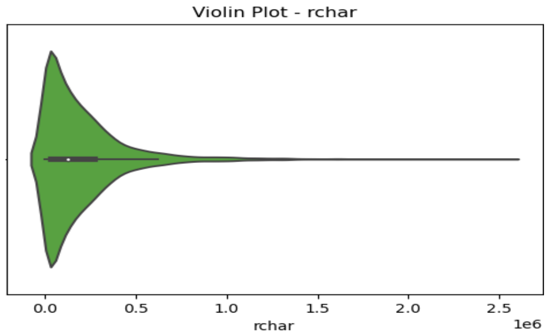


Figure 8 rchar

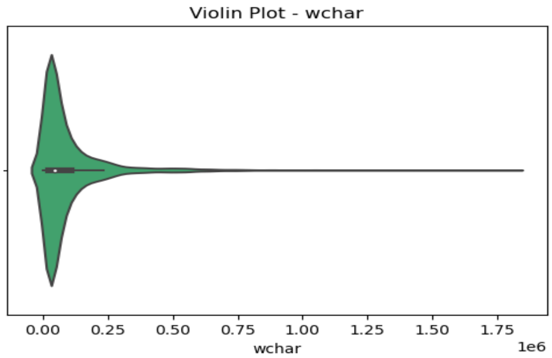


Figure 9 wchar

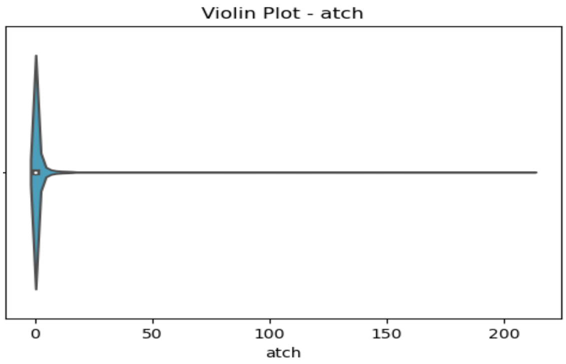


Figure 10 atch

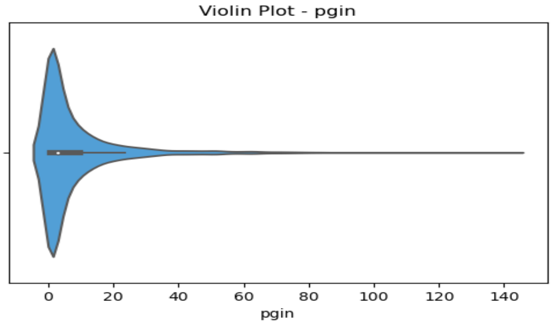


Figure 11 pgin

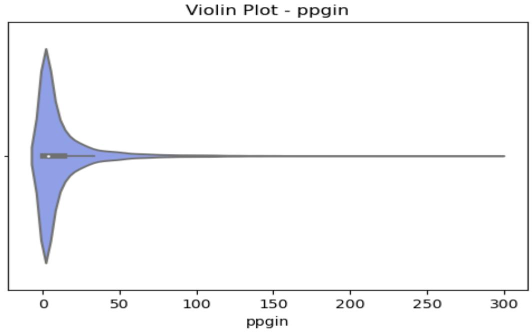


Figure 12 ppgin

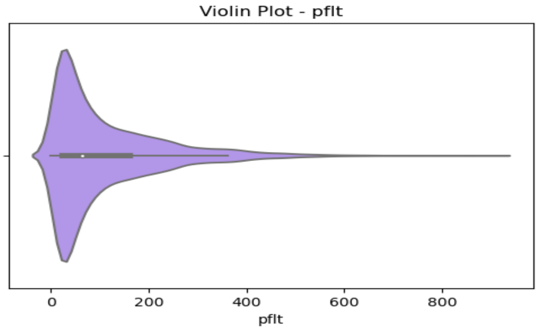


Figure 13 pfit

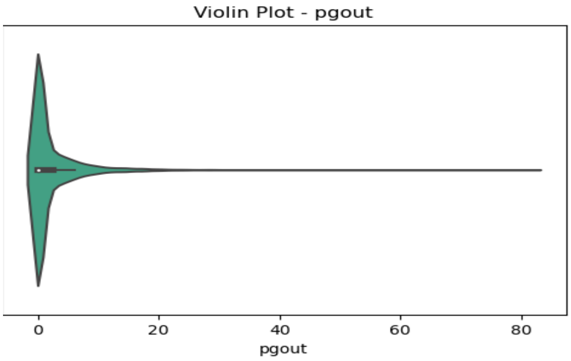


Figure 14 pgout

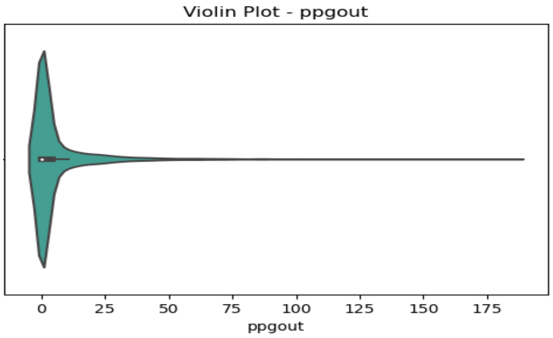


Figure 15 ppgout

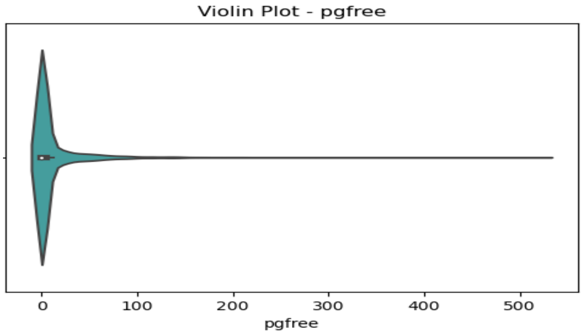


Figure 16 pgfree

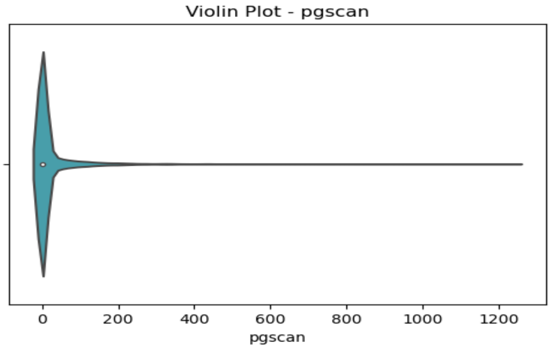


Figure 17 pgscan

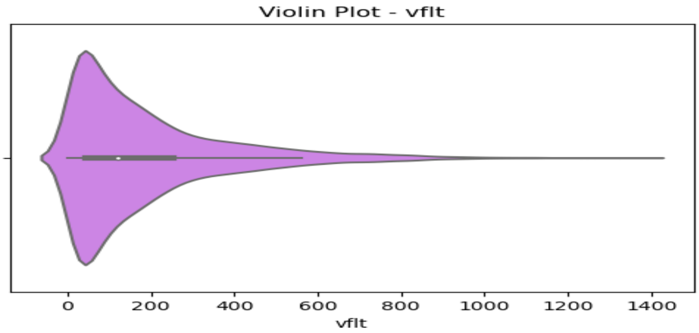


Figure 18 vflt

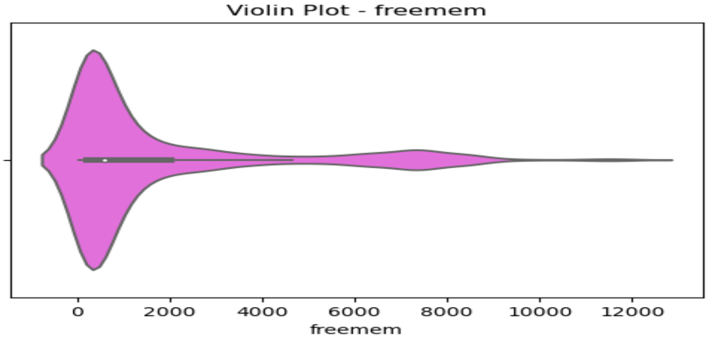


Figure 19 freemem

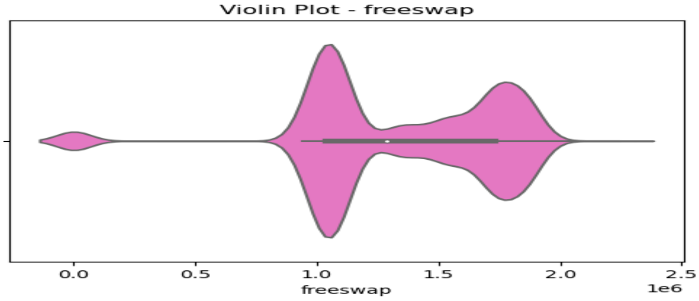


Figure 20 freeswap

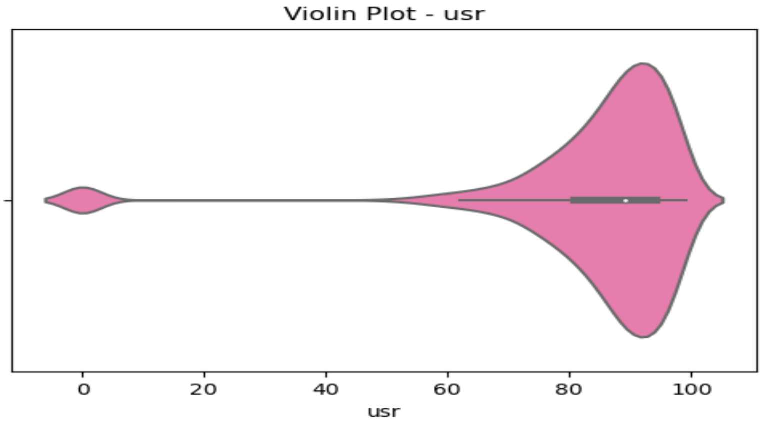


Figure 21 usr

The above plots from Figure 1 to Figure 21 represents the univariate analysis of Continuous variables. It can be noted that from figure 1 to figure 19 all the variables follow a pattern.

They all are skewed towards their right. As for figure 20 representing freeswap appears to have a trimodal distribution as there are 3 peaks in the plot. The target variable

which is “**usr**” is more of a left skewed and a very small peak could also be observed to its left where the smaller values belong. For all the right skewed a log transformation approach

could be an ideal solution. However, a model with and without transformation will be built ahead to differentiate the performances of the same.

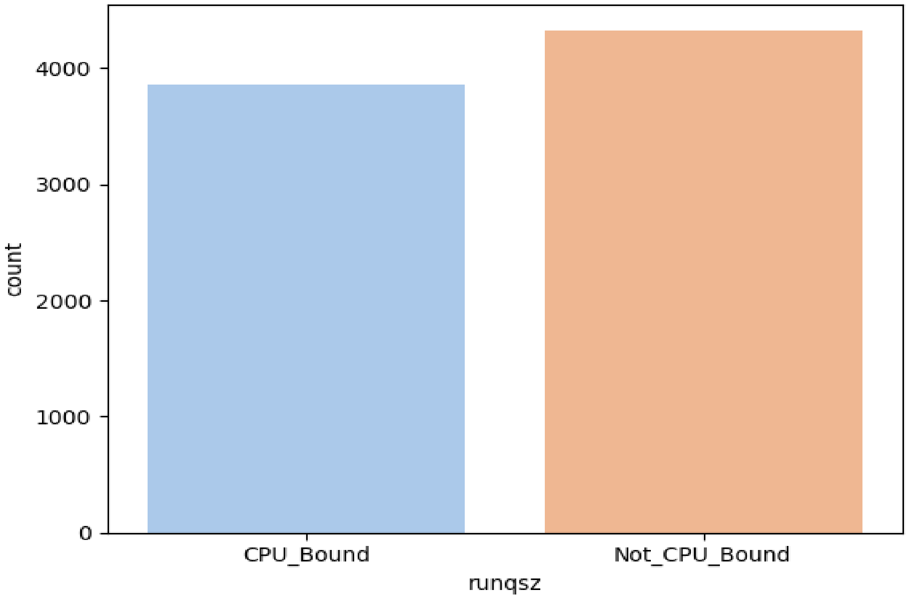


Figure 22 runqsz count

The above plot for the variables "runqsz" indicate there is a balance between the the two values (CPU_Bond and Not_CPU_Bound). Further indicating absence of Imbalanced data for this particular variable

Multivariate Analysis:

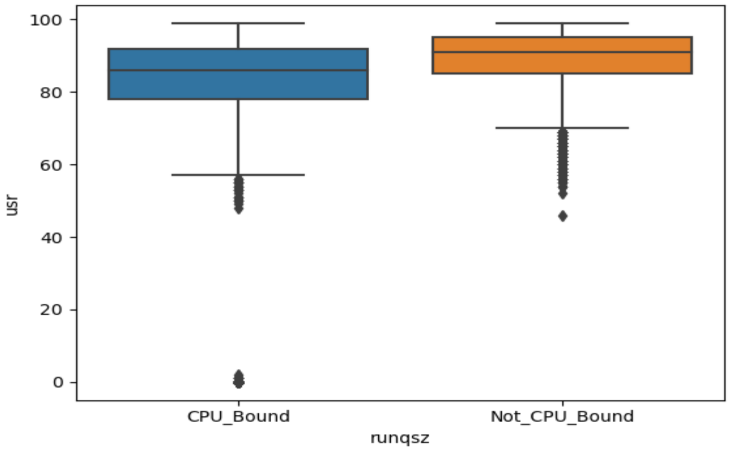


Figure 23 runqsz distribution w.r.t target

Figure 22 represents the distribution of the variable runqsz over target. It can be noted that though there is a slight difference in the mean the distribution does not vary hugely. CPU Bound has few outliers on the very lower end. This also implies that the only categorical variable present in the data does not provide cautious insight with at least the **target**.

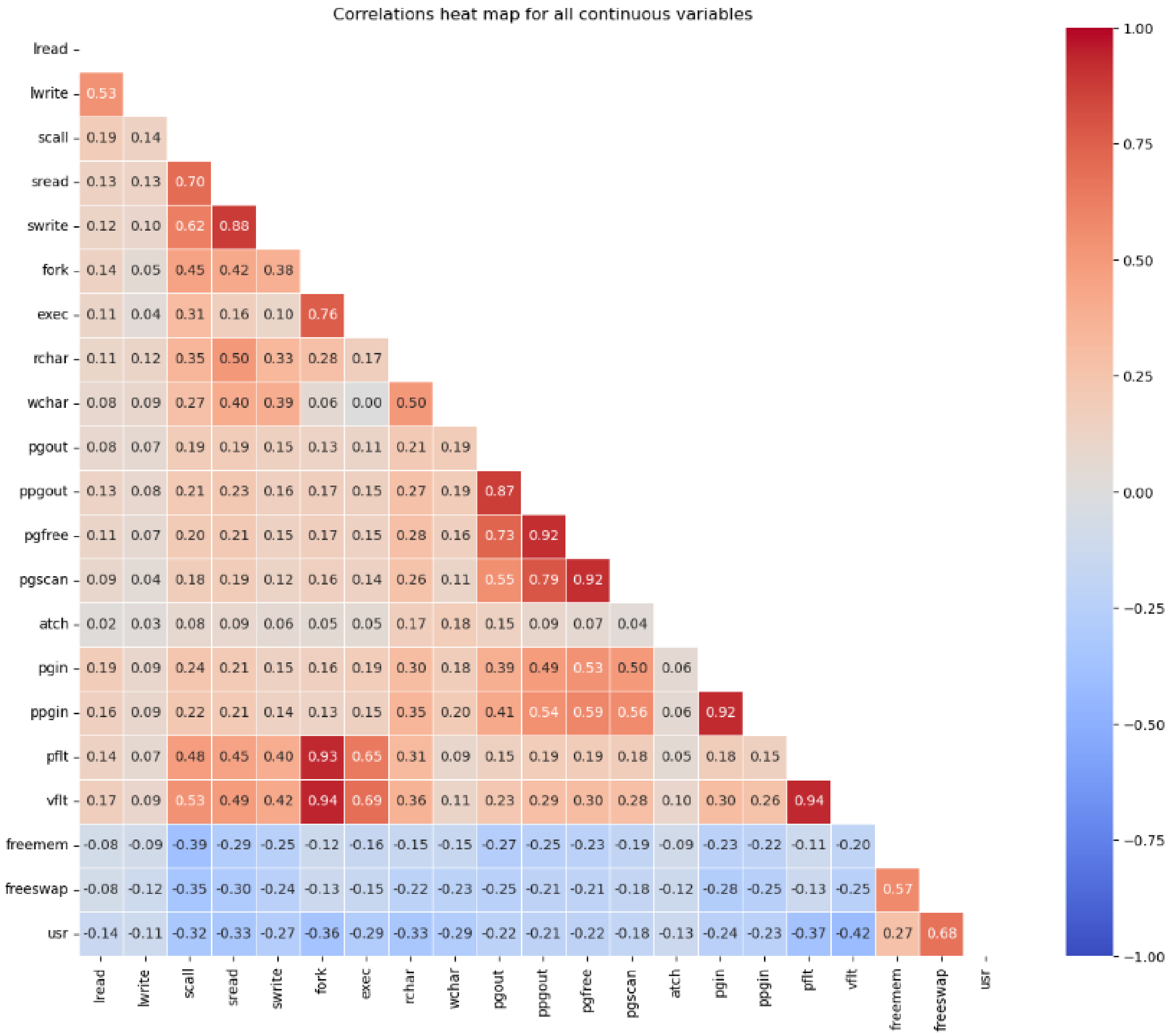


Figure 24 Raw Data Correlations

It can be inferred from the figure 23 that there are ample number of independent variables that are correlated with each other. For example, if we could set a modulus of correlation of a variable and a threshold of at least 0.5, a total of 15 combinations of variables can be noted to have correlations above 0.5(50%). Another interesting thing can be noted with respect to the target variable is that all the variables except the runqsz(categorical), freemem and freeswap are negatively correlated and freeswap is said to have a strong positive correlation with the target(usr).

Key Meaningful Observations:

- From the raw data univariate analysis, it can be noted that except for the variable freeswap and usr rest all are skewed towards the right. Also, freeswap is observed to have trimodal distribution.

- The statistical summary came out to be uneven, this observation is a direct impact of presence of large number of zeros in the variables. This also indicated presence of outliers.

- The target although appears to have a bimodal distribution at the left end of the tail is however a left tailed distribution. The correlation of independent variables alongside the target indicates all the variables are negatively correlated to the target except freemem and freeswap. Of the two positive correlation with the target freeswap has a strong positive correlation with the target.

- It has been noted since there are a large number of zeros in the data, zero can turn out be a problem especially in linear regression. Since, there has not been any mentions about whether or not presence of zeros is legitimate by the domain we will consider building multiple model one without and another with zeros.

Data Preprocessing:

It was noted that there were 119 missing values(nan) in the dataset. Of which 104 were in the rchar and 15 in the wchar. These missing values were handled using KNN Imputer whenever a model was built. Since the approach of building the model were bifurcated with zeros and non-zero models, the missing values were treated along with zeros if the model was built considering zeros are of bluff values adding least importance. If not, only the initial 119 missing values were handled considering zeros as the legitimate values.

Similar to the presence of zeros, there were also outliers in the data. Though it seemed like outliers, investigating the data came out to be legitimate and building the model with legitimate outliers did not bring any harm to the model.

Though outliers can bring in error slightly more than what the non-outlier model would bring in, it is still best suited to go ahead with outliers if the outlying values turn out to be legitimate.

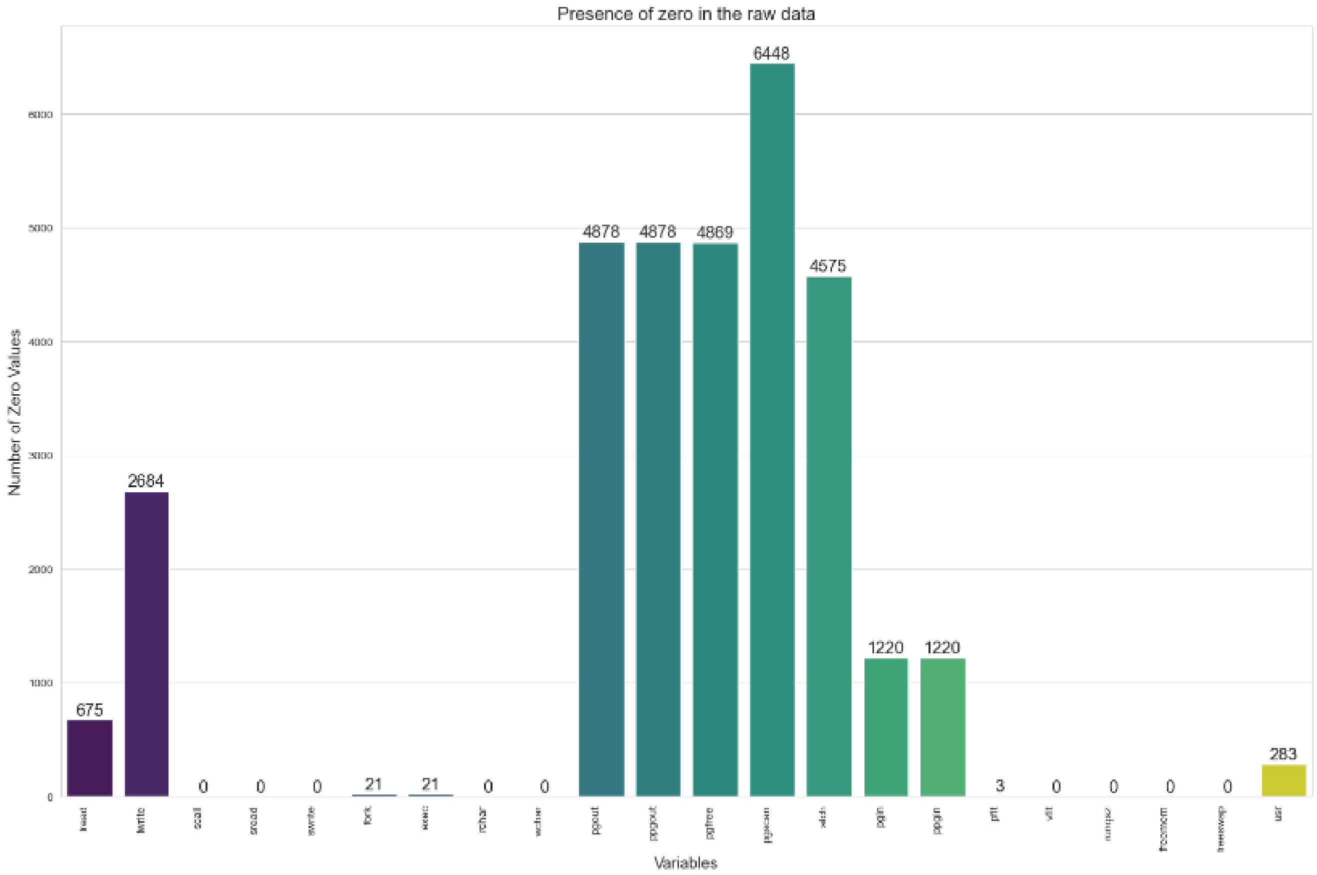


Figure 25 Count of Zeros

The above plot represents the count of zeros in the raw data. We need to keep in mind the data is hugely affected by the presence of large number of zeroes, zero inflated datasets ruin the regression model. Also, it is evidently visible that the data is skewed as seen in the univariate analysis. Multiple approaches for building the model could possibly provide a solution since a single ad-hoc approach might not be the right solution.

It was observed that the variable "pgscan" consists of 79% of zero's it should cause no harm if it is dropped from the dataset. The presence of such large number of zero would not be suitable to consider a variable for model building.

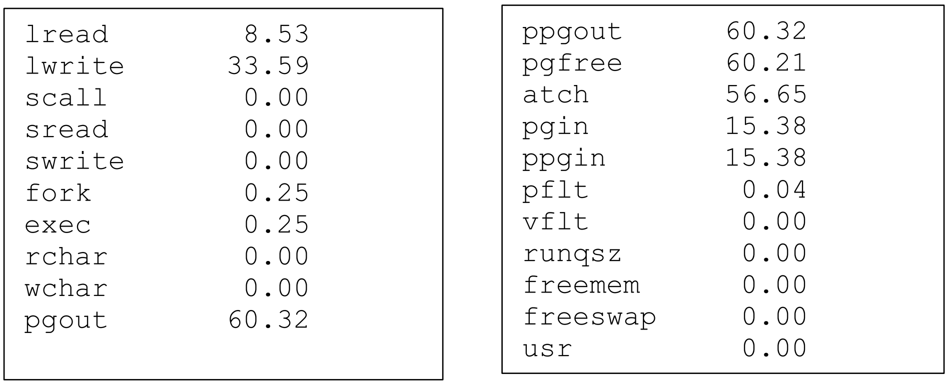
Multiple approach for building the model has been implemented like building the model with zeros, without zeros, with log transformation in the presence of zeros, penalising the variables coefficient using Regularization methods.

"usr" variable which is the dependent represents the percentage of time CPUs operate in user mode. If the values are zero in this variable it suggests that either the CPU is not operating or might be miscalculated. Since this could affect our model building and also the quantity is small, we will drop the records and associated with it wherever there are zero, one and two since even they are very less in count and sound inappropriate from the problem statement.

Model Building

Appraoch-1

In this approach we will drop the variables that have more than 50% of zeros in their column, and convert the other zeroes in to NaN and impute it using KNN Imputer.



```
lread           8.53
lwrite         33.59
scall           0.00
sread           0.00
swrite          0.00
fork            0.25
exec            0.25
rchar           0.00
wchar           0.00
pgout          60.32
```

```
ppgout         60.32
pgfree         60.21
atch           56.65
pgin           15.38
ppgin          15.38
pflt            0.04
vflt            0.00
runqsz          0.00
freemem         0.00
freeswap        0.00
usr             0.00
```

Table 3 Percentage of 0

From the table 3, we will be be dropping the variables pgout, ppgout. Pgfree and atch from the dataset and proceed further for model building.

NOTE: The variables that are being dropped here because of the presence of zero will be considered as it is in the other approaches. Since, there has been no domain information on the importance of the feature and zeros associated to it we will also build the model considering that the features that are dropped here might contain crucial information and are not suitable to drop.

Further, the feature having less than 50% of zeros are converted to nan and are imputed using KNN Imputation.

Why KNN Imputation?

- KNN imputation considers relationships between variables. It takes into account the similarity between data points and imputes missing values based on the values of the nearest neighbours.
- KNN imputation is not limited by assumptions of linearity. It can capture non-linear relationships between variables, making it suitable for datasets where the relationships are more complex and cannot be adequately represented by a linear model.
- In datasets where variables are interrelated, KNN imputation can be more effective. It considers the overall patterns in the data and can provide better estimates for missing values when variables interact with each other.

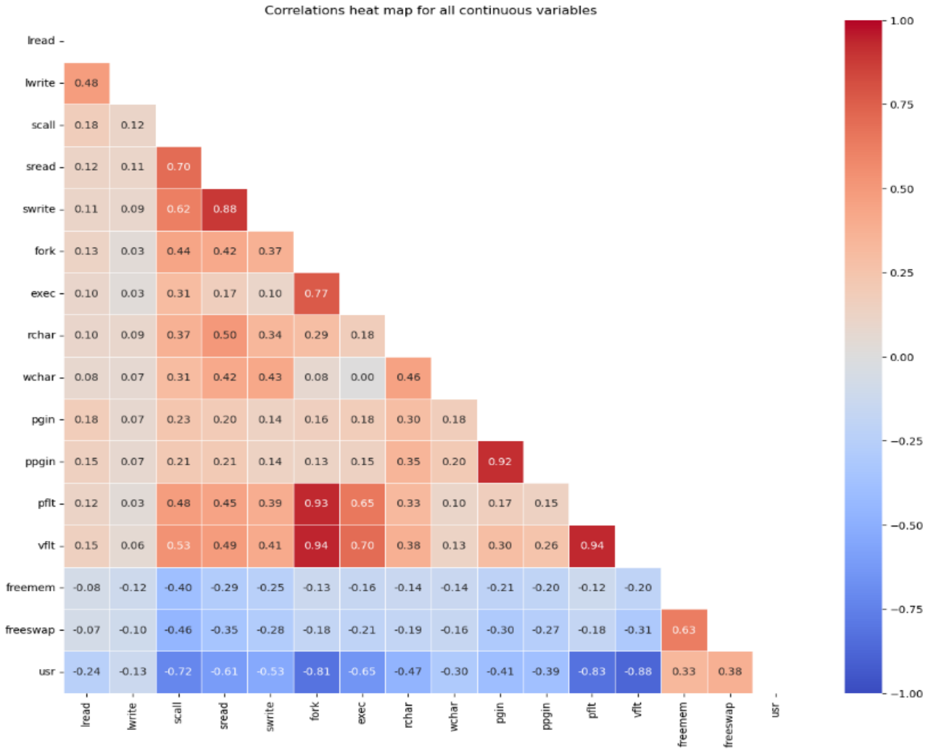
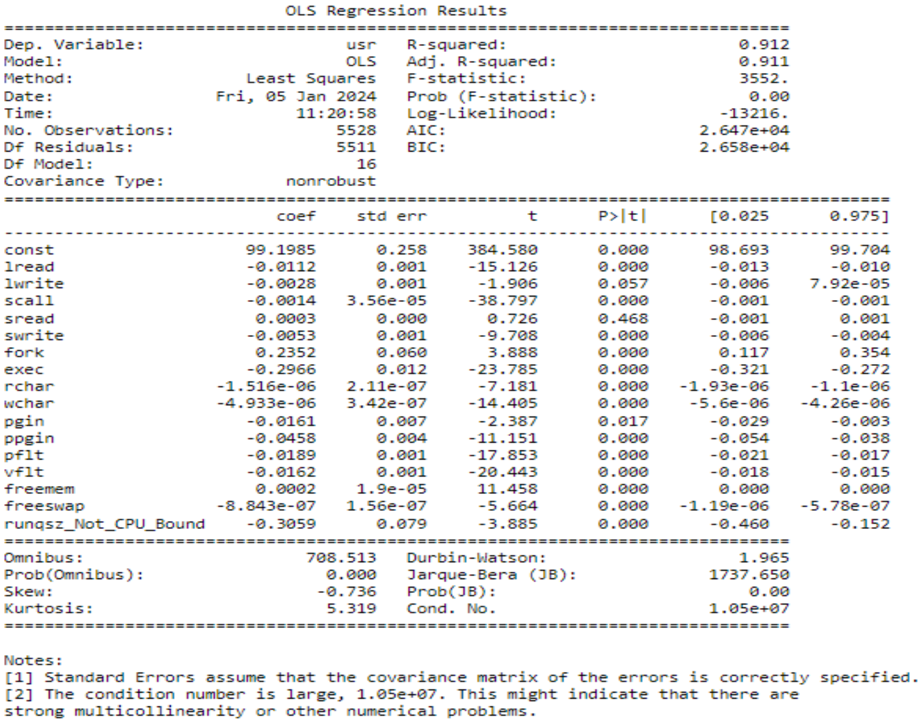


Figure 26 Correlations post data cleaning



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.912
Model:                            OLS   Adj. R-squared:                  0.911
Method:                 Least Squares   F-statistic:                     3552.
Date:                Fri, 05 Jan 2024   Prob (F-statistic):               0.00
Time:                        11:20:58   Log-Likelihood:                -13216.
No. Observations:                5528   AIC:                         2.647e+04
Df Residuals:                    5511   BIC:                         2.658e+04
Df Model:                          16
Covariance Type:            nonrobust
=========================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                  99.1985      0.258    384.580      0.000      98.693      99.704
lread                  -0.0112      0.001    -15.126      0.000      -0.013      -0.010
lwrite                 -0.0028      0.001     -1.906      0.057      -0.006    7.92e-05
scall                  -0.0014   3.56e-05    -38.797      0.000      -0.001      -0.001
sread                   0.0003      0.000      0.726      0.468      -0.001       0.001
swrite                 -0.0053      0.001     -9.708      0.000      -0.006      -0.004
fork                    0.2352      0.060      3.888      0.000       0.117       0.354
exec                   -0.2966      0.012    -23.785      0.000      -0.321      -0.272
rchar               -1.516e-06   2.11e-07     -7.181      0.000   -1.93e-06    -1.1e-06
wchar               -4.933e-06   3.42e-07    -14.405      0.000    -5.6e-06   -4.26e-06
pgin                   -0.0161      0.007     -2.387      0.017      -0.029      -0.003
ppgin                  -0.0458      0.004    -11.151      0.000      -0.054      -0.038
pflt                   -0.0189      0.001    -17.853      0.000      -0.021      -0.017
vflt                   -0.0162      0.001    -20.443      0.000      -0.018      -0.015
freemem                 0.0002    1.9e-05     11.458      0.000       0.000       0.000
freeswap            -8.843e-07   1.56e-07     -5.664      0.000   -1.19e-06   -5.78e-07
runqsz_Not_CPU_Bound   -0.3059      0.079     -3.885      0.000      -0.460      -0.152
==============================================================================
Omnibus:                      708.513   Durbin-Watson:                   1.965
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1737.650
Skew:                          -0.736   Prob(JB):                         0.00
Kurtosis:                       5.319   Cond. No.                     1.05e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.05e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Table 4 Initial OLS model

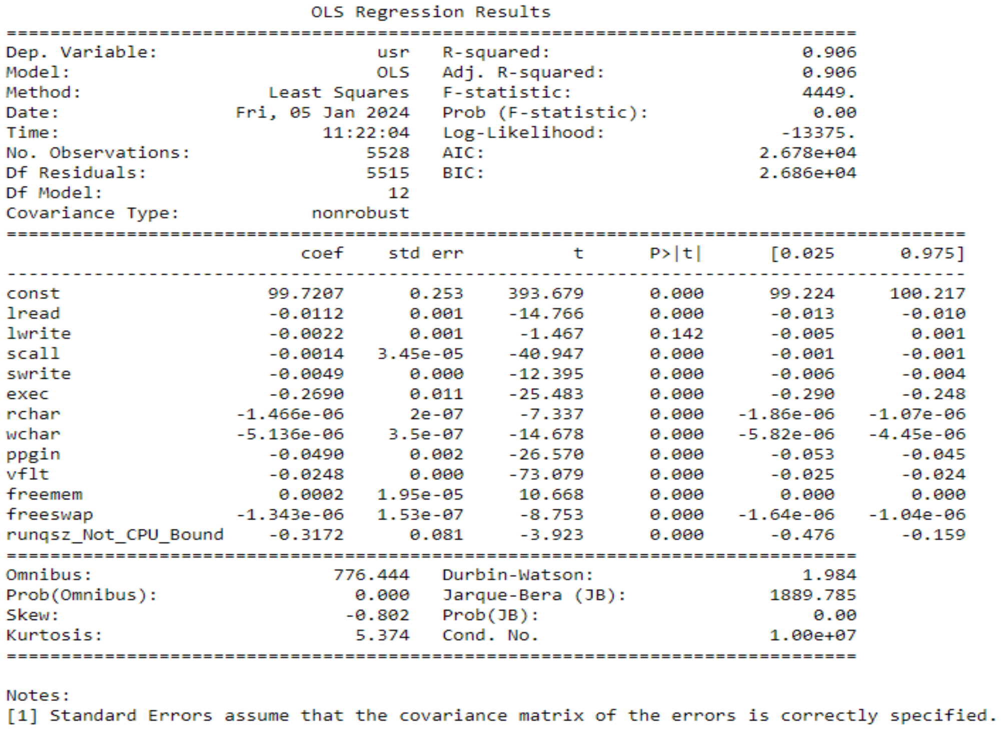
The data was split into X and y for Independent and dependent variables respectively, it was further distributed into respective train and test with a test size of 30% and a random state as 7 to obtain the model shown in the table 4. The random state 7 will be used for all the models.

From the table 4, the initial model (assumptions unchecked) infers few key observations. The R-squared and adjusted R-squared can be noted as 0.912 and 0.91 respective. This can be termed as a good observation as the model explains 91% of variance. However, the model is yet to be evaluated for its assumptions as we can see the presence of multicollinearity.

Since there is a presence of strong multicollinearity, the approach to solve the same was employed by VIF (Variance Inflation Factor). The Vif was calculated for the initial model and the variables having VIF larger than 10 was noted down and eliminated one by one and check for R-squared and adj. R-squared was conducted to make note of any changes in them.

Noting down the R-squared and the Adjusted R-squared dropping the variable **"fork"** which was contributing towards multicollinearity would bring less or no decrease in the model efficiency. The R-squared and adj. R-Squared came out to be 0.911 and 0.911 respectively post eliminating “fork”

Since there was presence of strong multicollinearity post eliminating “fork”, further VIF was checked and variables were dropped one by one so as to eliminate the multicollinearity. Columns such as **sread**, **pflt** and **pgin** were dropped from the model one by one and the model with less/no multicollinearity was built.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.906
Model:                            OLS   Adj. R-squared:                  0.906
Method:                 Least Squares   F-statistic:                     4449.
Date:                Fri, 05 Jan 2024   Prob (F-statistic):               0.00
Time:                        11:22:04   Log-Likelihood:                -13375.
No. Observations:                5528   AIC:                         2.678e+04
Df Residuals:                    5515   BIC:                         2.686e+04
Df Model:                          12
Covariance Type:            nonrobust
=========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                   99.7207      0.253    393.679      0.000      99.224     100.217
lread                   -0.0112      0.001    -14.766      0.000      -0.013      -0.010
lwrite                  -0.0022      0.001     -1.467      0.142      -0.005       0.001
scall                   -0.0014   3.45e-05    -40.947      0.000      -0.001      -0.001
swrite                  -0.0049      0.000    -12.395      0.000      -0.006      -0.004
exec                    -0.2690      0.011    -25.483      0.000      -0.290      -0.248
rchar                -1.466e-06      2e-07     -7.337      0.000   -1.86e-06   -1.07e-06
wchar                -5.136e-06    3.5e-07    -14.678      0.000   -5.82e-06   -4.45e-06
ppgin                   -0.0490      0.002    -26.570      0.000      -0.053      -0.045
vflt                    -0.0248      0.000    -73.079      0.000      -0.025      -0.024
freemem                  0.0002   1.95e-05     10.668      0.000       0.000       0.000
freeswap             -1.343e-06   1.53e-07     -8.753      0.000   -1.64e-06   -1.04e-06
runqsz_Not_CPU_Bound    -0.3172      0.081     -3.923      0.000      -0.476      -0.159
==============================================================================
Omnibus:                      776.444   Durbin-Watson:                   1.984
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1889.785
Skew:                          -0.802   Prob(JB):                         0.00
Kurtosis:                       5.374   Cond. No.                     1.00e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

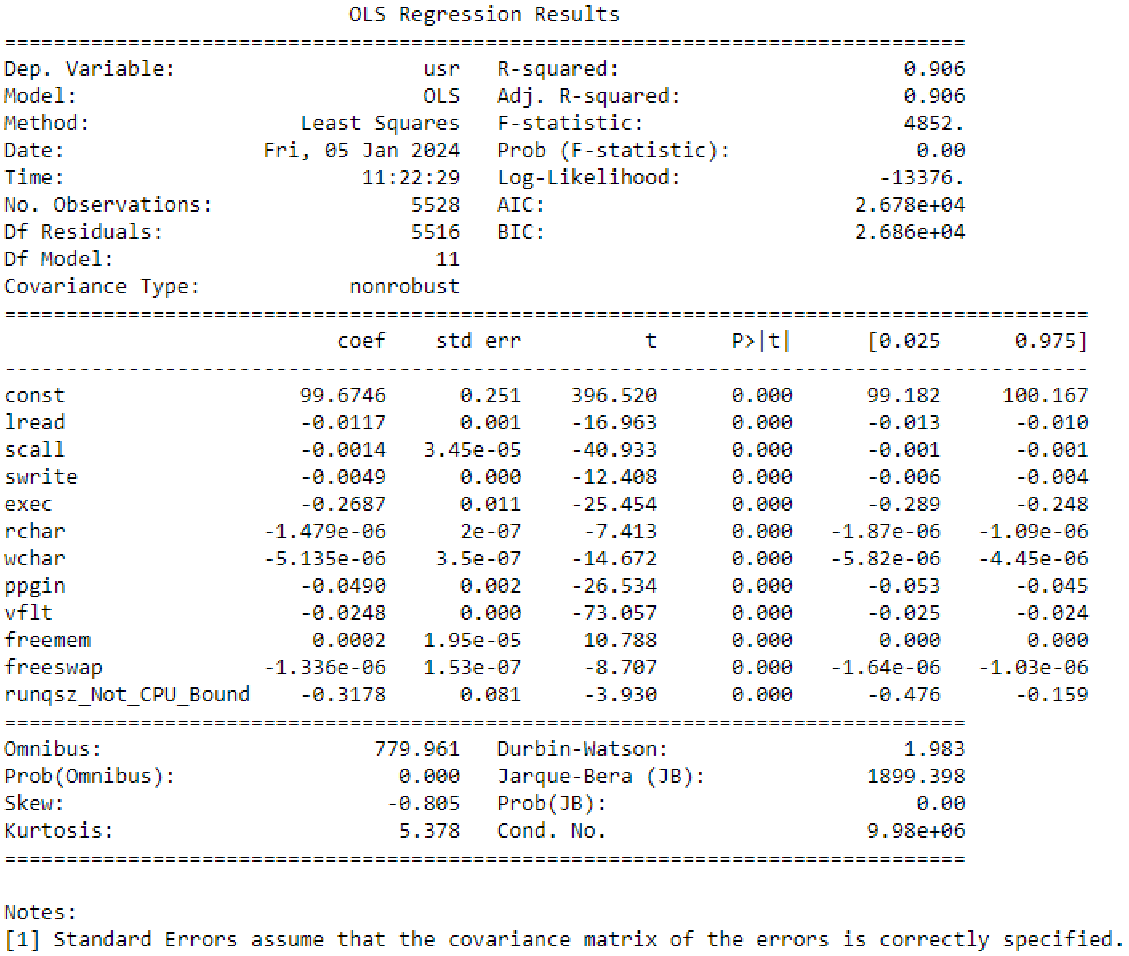
Table 5 OLS model post VIF treatment

The concept of multicollinearity was resolved by VIF. The threshold has been set 5 and now all the variables have variability below 5. The table 5 represents the model post treatment of multicollinearity with R-squared and adj. R-squared of 0.906 respectively. However, this can not be considered as the final model as we can see the variable **lwrite** with **p-value > 0.05**

P-value is the test for importance of feature and whether or not the features importance is zero.

From the OLS table above it can be noted that there is a 14.2% chance for variable **"lwrite"** to be least important among the group since the hypothesis test assumes that the important variable be less than or equal to the threshold of 0.05.

Another and the final model with less/no multicollinearity and all variable importance was built as shown below.



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.906
Model:                            OLS   Adj. R-squared:                  0.906
Method:                 Least Squares   F-statistic:                     4852.
Date:                Fri, 05 Jan 2024   Prob (F-statistic):               0.00
Time:                        11:22:29   Log-Likelihood:                -13376.
No. Observations:                5528   AIC:                         2.678e+04
Df Residuals:                    5516   BIC:                         2.686e+04
Df Model:                          11                                         
Covariance Type:            nonrobust                                         
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                    99.6746      0.251    396.520      0.000      99.182     100.167
lread                    -0.0117      0.001    -16.963      0.000      -0.013      -0.010
scall                    -0.0014   3.45e-05    -40.933      0.000      -0.001      -0.001
swrite                   -0.0049      0.000    -12.408      0.000      -0.006      -0.004
exec                     -0.2687      0.011    -25.454      0.000      -0.289      -0.248
rchar                 -1.479e-06      2e-07     -7.413      0.000   -1.87e-06   -1.09e-06
wchar                 -5.135e-06    3.5e-07    -14.672      0.000   -5.82e-06   -4.45e-06
ppgin                    -0.0490      0.002    -26.534      0.000      -0.053      -0.045
vflt                     -0.0248      0.000    -73.057      0.000      -0.025      -0.024
freemem                   0.0002   1.95e-05     10.788      0.000       0.000       0.000
freeswap              -1.336e-06   1.53e-07     -8.707      0.000   -1.64e-06   -1.03e-06
runqsz_Not_CPU_Bound     -0.3178      0.081     -3.930      0.000      -0.476      -0.159
==============================================================================
Omnibus:                      779.961   Durbin-Watson:                   1.983
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1899.398
Skew:                          -0.805   Prob(JB):                         0.00
Kurtosis:                       5.378   Cond. No.                     9.98e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Table 6 final OLS model

The R-squared and adj. R-squared remain un-altered post dropping **lwrite** as it is clearly visible in the Table 6 consisting of final OLS model. In this OLS model the assumption of multicollinearity was eliminated and the feature importance was taken into consideration.

The final **R-squared and Adj. R-squared** remain **0.96**, explaining **96%** of the **variance**. There are few more assumptions that needs to be satisfied so as to conclude the satisfaction of the result the model will predict.

Assumptions:

Linearity and Independence test:

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

- 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.

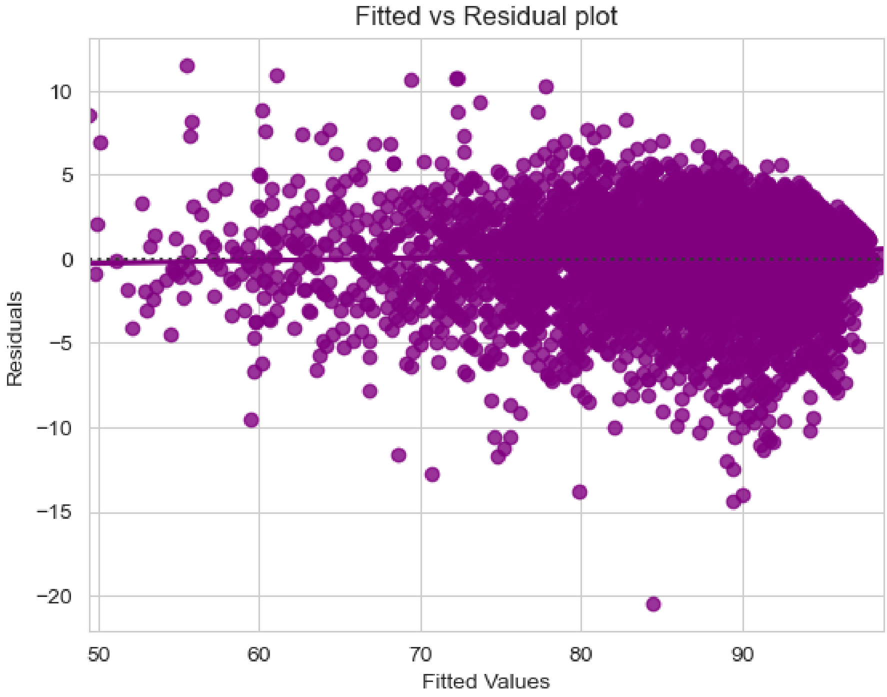


Figure 27 Linearity and Independence test

The above plot (Figure 27) of fitted values vs residuals don't follow any pattern (the curve is a straight line), then we say the model is almost linear. Achieving this can be a challenging task as perfect linear sometimes can be impossible. However, a near perfect linear like the above plot is achievable as they represent randomly distributed.

Test for Normality:

- Error terms/residuals should be normally distributed.

- If the error terms are not normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares.

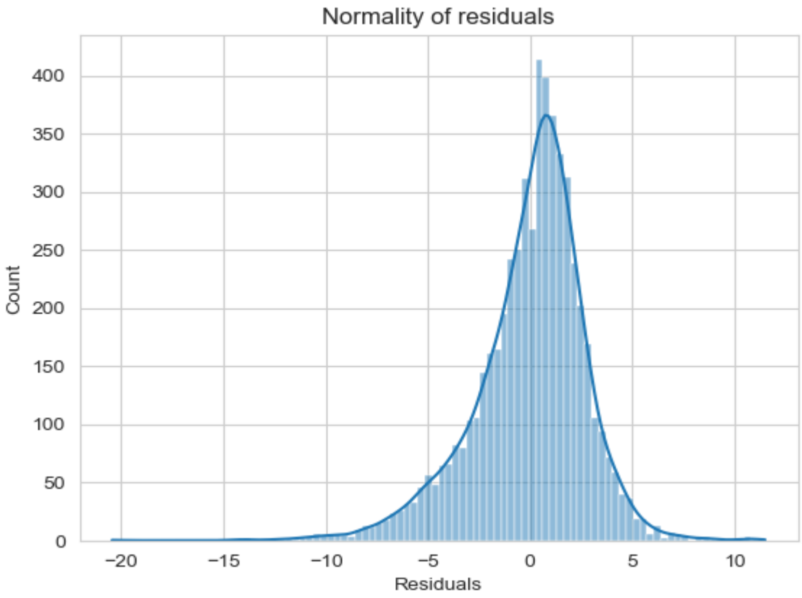


Figure 28 Normality of residuals

The above visual representation tells that the errors are normally distributed. It also suggests to some extent it is skewed towards its left. Since the model is built without outlier treatment considering the outliers are a legitimate value, the slight skewness could be a result of it. However, we will try solving the data with other approaches.

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.

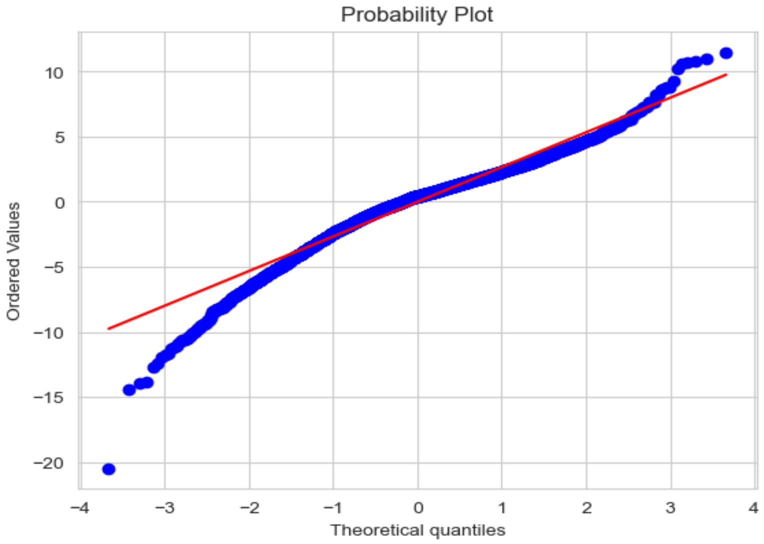
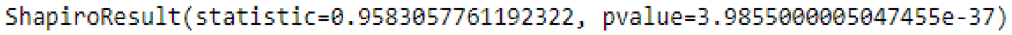


Figure 29 QQ Plot for normality

Most of the points are lying on the straight line in QQ plot. There can be few exceptions as suggested earlier getting a full perfect model can be highly challenging especially without domain intervention. However, the above QQ plot could satisfy the need.

The Shapiro-Wilk test can also be used for checking the normality. The null and alternate hypotheses of the test are as follows:

- Null hypothesis - Data is normally distributed.
- Alternate hypothesis - Data is not normally distributed.



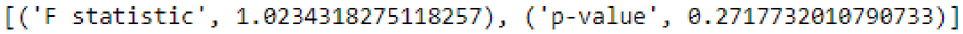
```
ShapiroResult(statistic=0.9583057761192322, pvalue=3.9855000005047455e-37)
```

- 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

Test for Homoscedasticity:

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

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



```
[('F statistic', 1.0234318275118257), ('p-value', 0.2717732010790733)]
```

- Since p-value > 0.05 we can say that the **residuals are homoscedastic.**

Linear Regression Equation:

usr **=** 99.67456476140191 + -0.011698042141953 * (**lread**) + -0.0014128167946733641 * (**scall**) + -0.004938179830042236 * (**swrite**) + -0.2686742094564625 * (**exec**) + -1.4794325809918855e-06 * (**rchar**) + -5.134623999262638e-06 * (**wchar**) + -0.04895505808532618 * (**ppgin**) + -0.024796525604816134 * (**vflt**) + 0.0002103487045647368 * (**freemem**) + -1.335542105397542e-06 * (**freeswap**) + -0.3178231678893986 * (**runqsz_Not_CPU_Bound**)

Equation 1 Linear Regression

RMSE Score:

Training Score: 2.7203264821013207

Test Score: 2.780337238168565

- We can see that RMSE on the train and test sets are comparable. So, our model is not suffering from overfitting.
- Hence, we can conclude the model "OLS" is good for prediction as well as inference purposes.

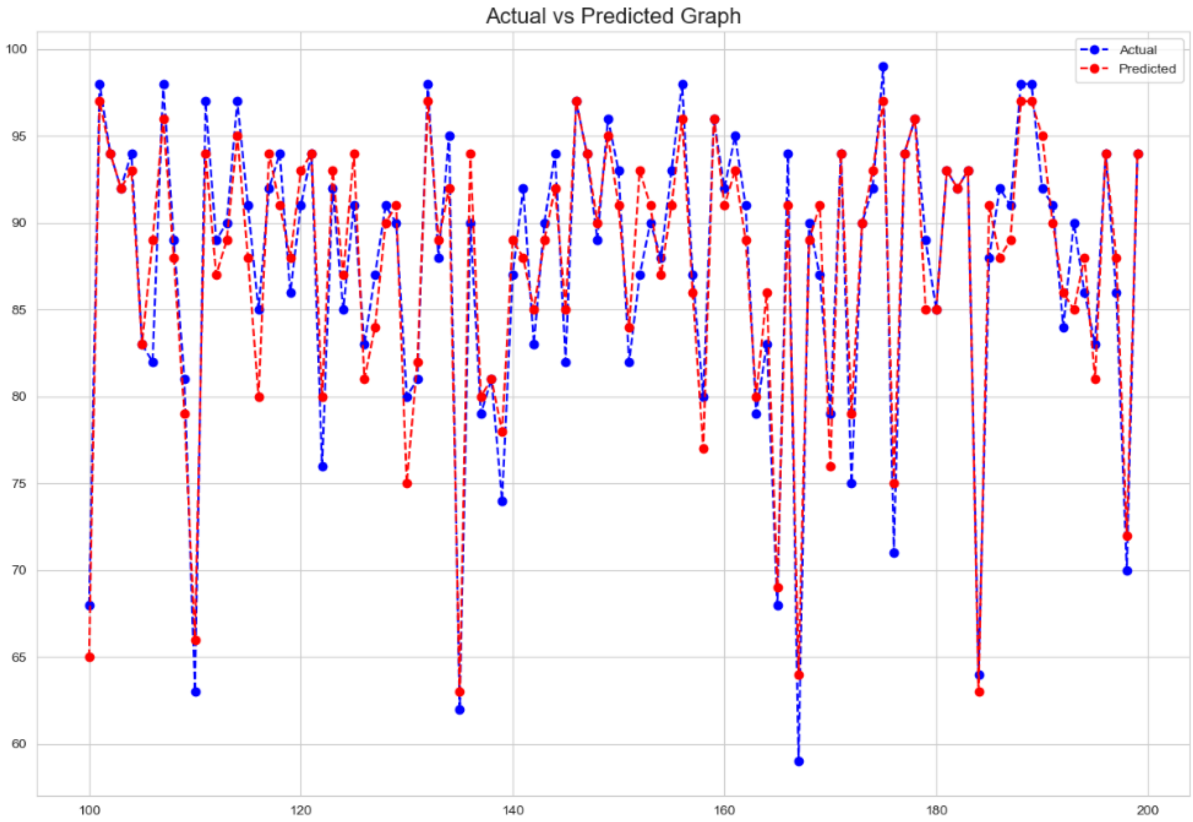


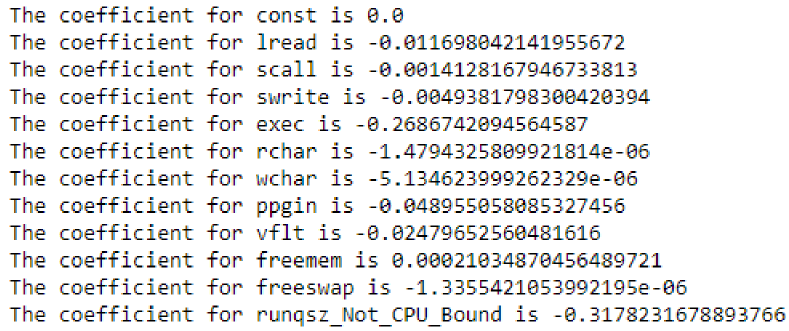
Figure 30 Actual vs Predicted

The plot represents the **Actual vs Predicted** graph of a randomly chosen 100 records. The blue represents the Actual records and the red represents the Predicted records. We can now see a very minimal deviation when compared each other representing the model is doing a perfect job in predicting the values.

Observations for Approach-1

- R-squared of the model is 0.906 and adjusted R-squared is 0.906, which shows that the model is able to explain ~90% variance in the data. This is quite phenomenal.
- A unit increase in the exec will result in a -0.2687 unit decrease in the usr, all other variables remaining constant.
- A unit increase in the independent variable will result in a const. unit increase/decrease in the usr.
- The usr of a Not CPU Bound in runqsz will be -0.3178 units lesser than a runqsz of CPU Bound, all other variables remaining constant.
- As we move forward, we will build few more model with data considering zeros as a legitimate value, with Linear Regression and Regularization.

Linear Regression using sklearn.



```
The coefficient for const is 0.0
The coefficient for lread is -0.011698042141955672
The coefficient for scall is -0.001412816794673381З
The coefficient for swrite is -0.0049381798300420394
The coefficient for exec is -0.2686742094564587
The coefficient for rchar is -1.4794325809921814e-06
The coefficient for wchar is -5.134623999262329e-06
The coefficient for ppgin is -0.048955058085327456
The coefficient for vflt is -0.02479652560481616
The coefficient for freemem is 0.00021034870456489721
The coefficient for freeswap is -1.3355421053992195e-06
The coefficient for runqsz_Not_CPU_Bound is -0.3178231678893766
```

Table 7 Coefficient

The **intercept** for our model is 99.67456476141315

R-squared:

Training: 0.9063271600466553

Testing: 0.9030559276940191

90% of the variation in the **usr** is explained by the predictors in the model for train and test set.

RMSE:

Training: 2.7203264821013207

Testing: 2.780337238168478

Approach-2 Regularization Methods

Ridge and Lasso regularization involve adding a penalty term to the linear regression cost function. These penalty terms are based on the magnitude of the coefficients, and they can be sensitive to the scale of the input features. Therefore, scaling is recommended to ensure that all features contribute equally to the regularization term.

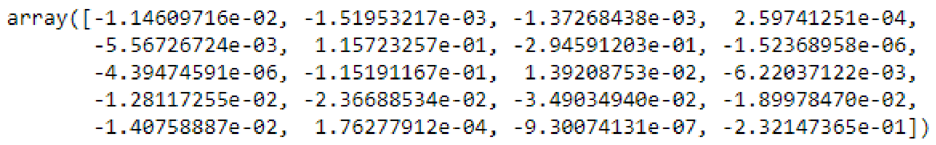
In this approach let us consider zero to be valid in our data. Since regularization method does a feature selection part within it, may be the variables with large zero will be punished in the presence of lambda (alpha) function.

All the features except **pgscan** shall be considered for this approach. Hence, 21 features is what we will be proceeding with for model building.

Ridge Regularization:

Best alpha found: 10

Method employed: Grid Search CV (To Identify best alpha).



```
array([-1.14609716e-02, -1.51953217e-03, -1.37268438e-03,  2.59741251e-04,
       -5.56726724e-03,  1.15723257e-01, -2.94591203e-01, -1.52368958e-06,
       -4.39474591e-06, -1.15191167e-01,  1.39208753e-02, -6.22037122e-03,
       -1.28117255e-02, -2.36688534e-02, -3.49034940e-02, -1.89978470e-02,
       -1.40758887e-02,  1.76277912e-04, -9.30074131e-07, -2.32147365e-01])
```

Table 8 Ridge model Coefficient

Ridge Model Score

Training Score: 0.9152828363177401

Testing Score: 0.9093091434469945

The Root Mean Squared Error on Train Set: 2.587020614213069

The Root Mean Squared Error on Test Set: 2.6891721118726006

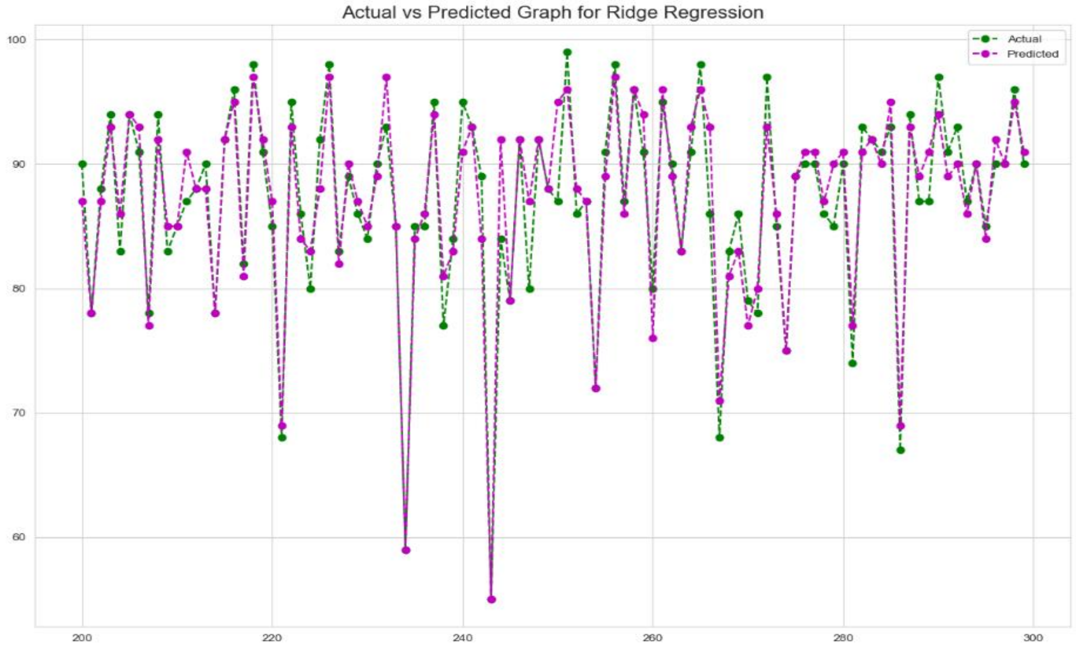


Figure 31 Ridge Actual vs Predicted

The above plot represents the Actual vs Predicted graph of a randomly chosen 100 records for the ridge model. The green represents the Actual records and the magenta represents the Predicted records. We can ne see a very minimal deviation when compared each other representing the model is doing a perfect job in predicting the values.

Lasso Regularization:

Best alpha found: 0.01

Method employed: Grid Search CV (To Identify best alpha).

Lasso Model Score

Training Score: 0.9152708482960568

Testing Score: 0.9093835593687756

The Root Mean Squared Error on Train Set: 2.587203647737826

The Root Mean Squared Error on Test Set: 2.6880685921944107

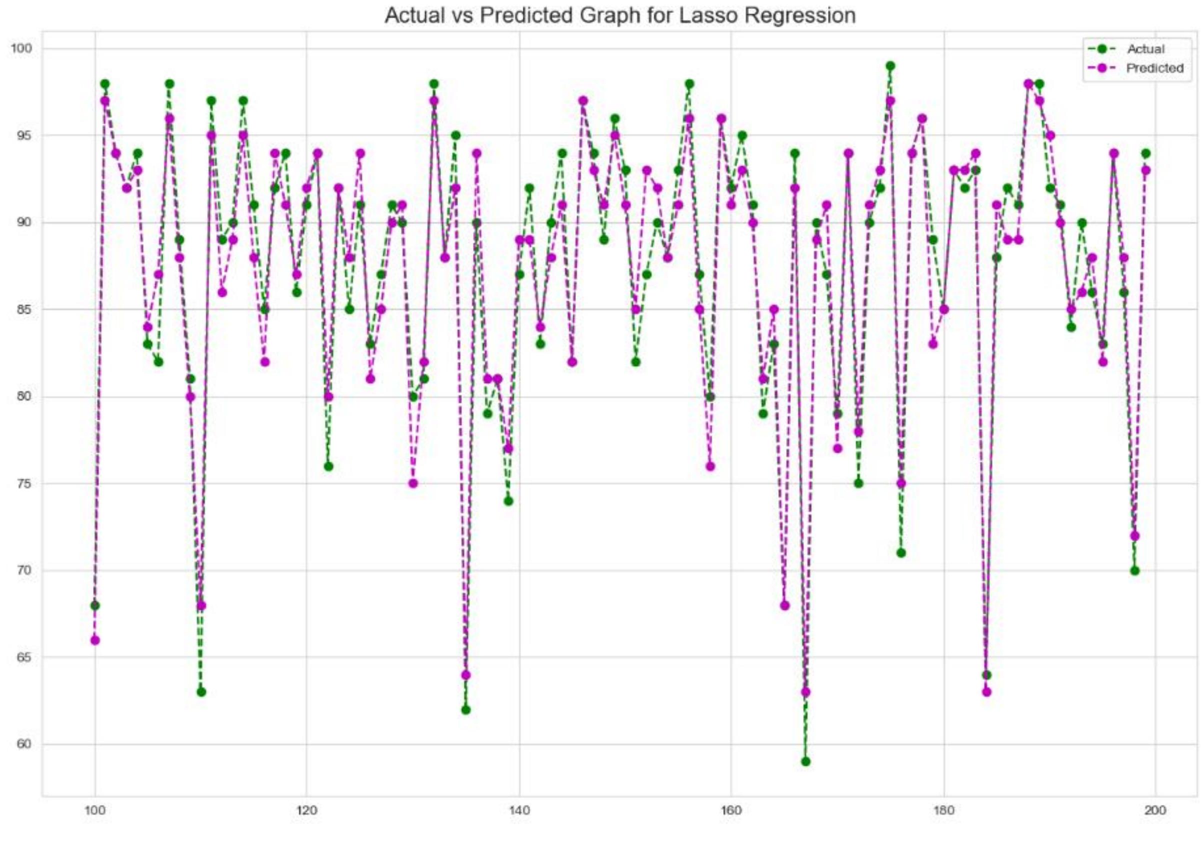


Figure 32 Lasso Actual vs Predicted

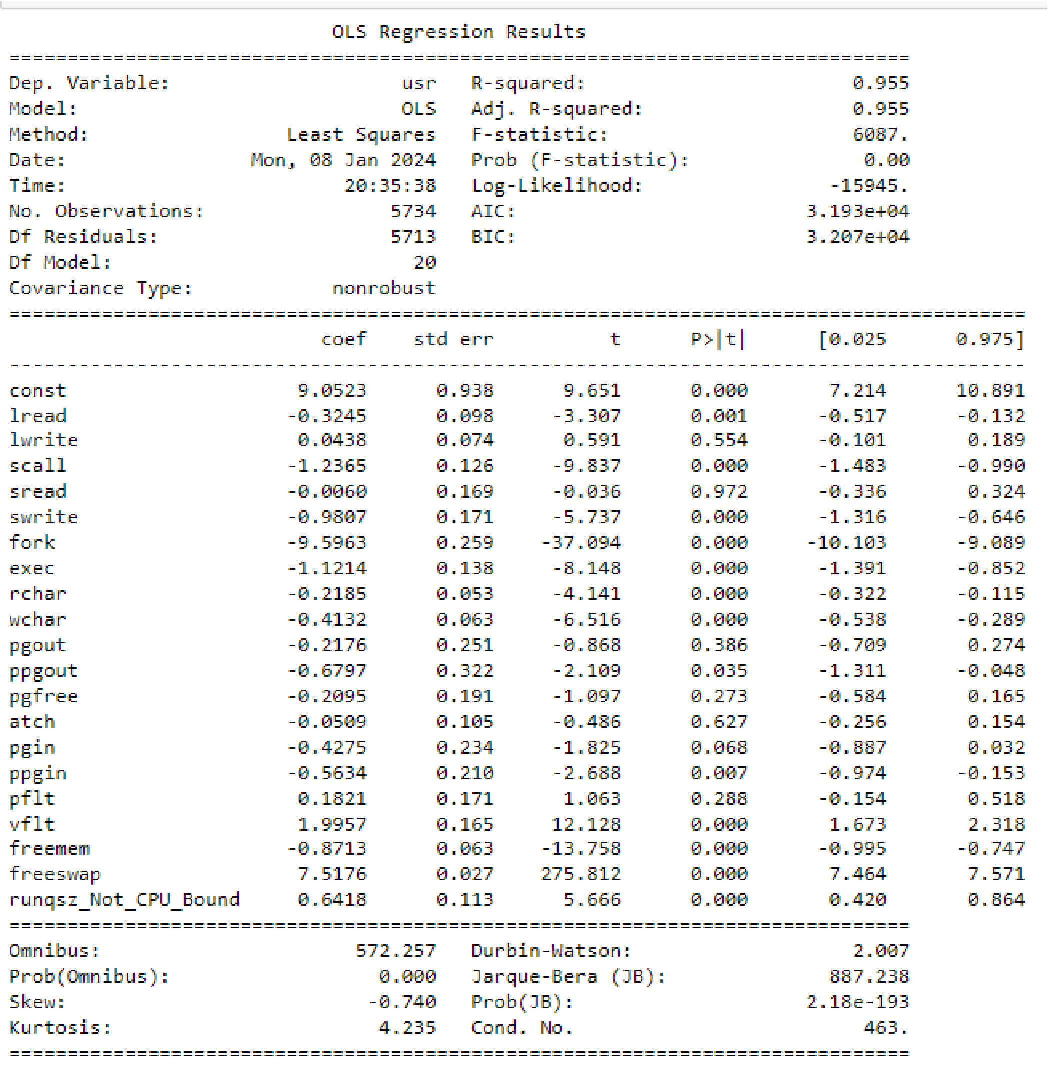
The above plot represents the Actual vs Predicted graph of a randomly chosen 100 records for the lasso model. The green represents the Actual records and the magenta represents the Predicted records. We can ne see a very minimal deviation when compared each other representing the model is doing a perfect job in predicting the values.

Approach-3

EDA (Univariate Analysis) evidently displays that there is a skewness in the data that too almost all variables having their values skewed towards right. Another approach of building the model would be by log transformation also not excluding the columns with significant number of zeros except "pgscan"

The reason for including **log transformation**, most of the variables were right (positive) skewed.

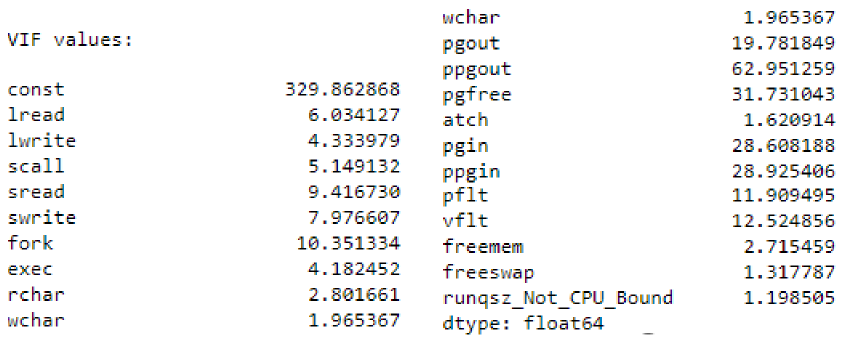
Initial Model



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.955
Model:                            OLS   Adj. R-squared:                  0.955
Method:                 Least Squares   F-statistic:                     6087.
Date:                Mon, 08 Jan 2024   Prob (F-statistic):               0.00
Time:                        20:35:38   Log-Likelihood:                -15945.
No. Observations:                5734   AIC:                         3.193e+04
Df Residuals:                    5713   BIC:                         3.207e+04
Df Model:                          20                                         
Covariance Type:            nonrobust                                         
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                    9.0523      0.938      9.651      0.000       7.214      10.891
lread                   -0.3245      0.098     -3.307      0.001      -0.517      -0.132
lwrite                   0.0438      0.074      0.591      0.554      -0.101       0.189
scall                   -1.2365      0.126     -9.837      0.000      -1.483      -0.990
sread                   -0.0060      0.169     -0.036      0.972      -0.336       0.324
swrite                  -0.9807      0.171     -5.737      0.000      -1.316      -0.646
fork                    -9.5963      0.259    -37.094      0.000     -10.103      -9.089
exec                    -1.1214      0.138     -8.148      0.000      -1.391      -0.852
rchar                   -0.2185      0.053     -4.141      0.000      -0.322      -0.115
wchar                   -0.4132      0.063     -6.516      0.000      -0.538      -0.289
pgout                   -0.2176      0.251     -0.868      0.386      -0.709       0.274
ppgout                  -0.6797      0.322     -2.109      0.035      -1.311      -0.048
pgfree                  -0.2095      0.191     -1.097      0.273      -0.584       0.165
atch                    -0.0509      0.105     -0.486      0.627      -0.256       0.154
pgin                    -0.4275      0.234     -1.825      0.068      -0.887       0.032
ppgin                   -0.5634      0.210     -2.688      0.007      -0.974      -0.153
pflt                     0.1821      0.171      1.063      0.288      -0.154       0.518
vflt                     1.9957      0.165     12.128      0.000       1.673       2.318
freemem                 -0.8713      0.063    -13.758      0.000      -0.995      -0.747
freeswap                 7.5176      0.027    275.812      0.000       7.464       7.571
runqsz_Not_CPU_Bound     0.6418      0.113      5.666      0.000       0.420       0.864
==============================================================================
Omnibus:                      572.257   Durbin-Watson:                   2.007
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              887.238
Skew:                          -0.740   Prob(JB):                    2.18e-193
Kurtosis:                       4.235   Cond. No.                         463.
==============================================================================
```

Table 9 Approach-3 Initial OLS Model

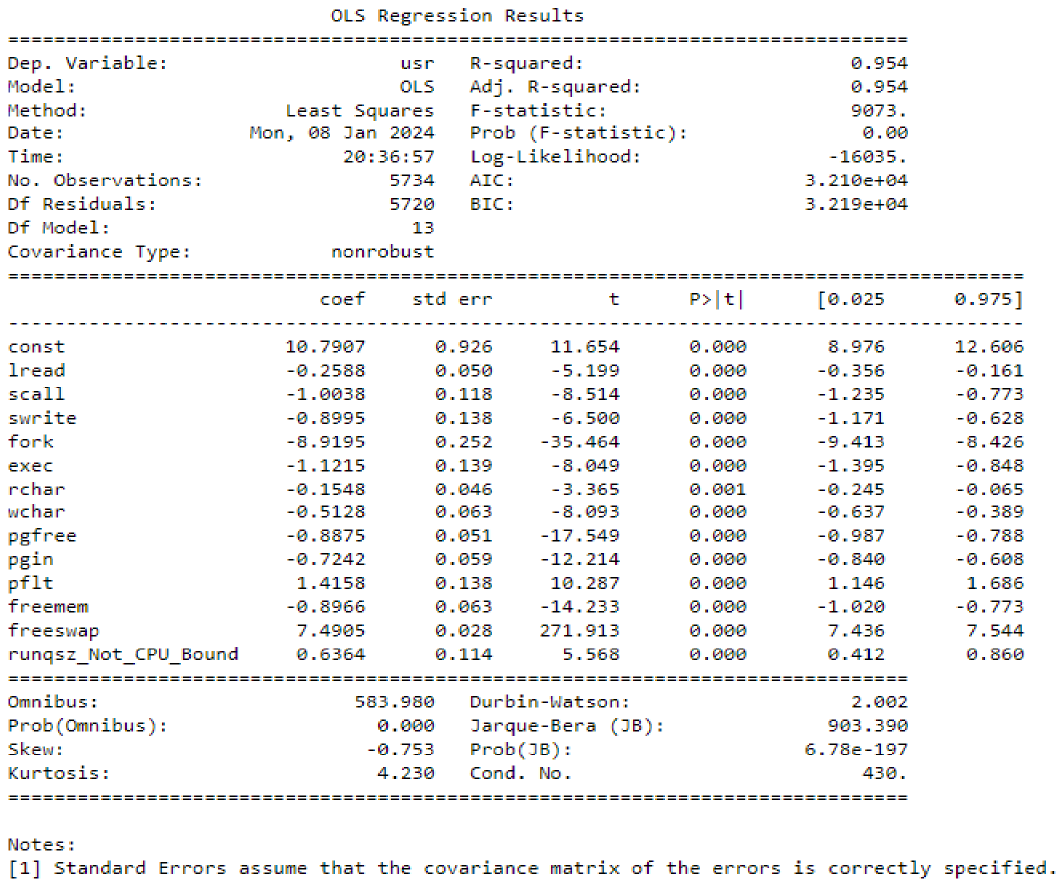
The R-squared and Adj. R-squared comes out to be 0.955 respectively. Since there is a presence of strong multicollinearity, we would look to solve it using VIF.



```
                                            wchar                     1.965367
VIF values:                                 pgout                    19.781849
                                            ppgout                   62.951259
const                       329.862868      pgfree                   31.731043
lread                         6.034127      atch                      1.620914
lwrite                        4.333979      pgin                     28.608188
scall                         5.149132      ppgin                    28.925406
sread                         9.416730      pflt                     11.909495
swrite                        7.976607      vflt                     12.524856
fork                         10.351334      freemem                   2.715459
exec                          4.182452      freeswap                  1.317787
rchar                         2.801661      runqsz_Not_CPU_Bound      1.198505
wchar                         1.965367      dtype: float64
```

Table 10 VIF values

From the table 10, it can be noted that many columns have high VIF value. The elimination of high VIF features was done in a phased manner. The high VIF feature was dropped and the R-squared and Adj. R-squared was noted. Drastic change in Adj. R-squared would indicate the drop in feature was unnecessary to drop. This process continues until we eliminated the VIF features also being aware of importance of feature and dip in the Variance of the model.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.954
Model:                            OLS   Adj. R-squared:                  0.954
Method:                 Least Squares   F-statistic:                     9073.
Date:                Mon, 08 Jan 2024   Prob (F-statistic):               0.00
Time:                        20:36:57   Log-Likelihood:                -16035.
No. Observations:                5734   AIC:                         3.210e+04
Df Residuals:                    5720   BIC:                         3.219e+04
Df Model:                          13
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                   10.7907      0.926     11.654      0.000       8.976      12.606
lread                   -0.2588      0.050     -5.199      0.000      -0.356      -0.161
scall                   -1.0038      0.118     -8.514      0.000      -1.235      -0.773
swrite                  -0.8995      0.138     -6.500      0.000      -1.171      -0.628
fork                    -8.9195      0.252    -35.464      0.000      -9.413      -8.426
exec                    -1.1215      0.139     -8.049      0.000      -1.395      -0.848
rchar                   -0.1548      0.046     -3.365      0.001      -0.245      -0.065
wchar                   -0.5128      0.063     -8.093      0.000      -0.637      -0.389
pgfree                  -0.8875      0.051    -17.549      0.000      -0.987      -0.788
pgin                    -0.7242      0.059    -12.214      0.000      -0.840      -0.608
pflt                     1.4158      0.138     10.287      0.000       1.146       1.686
freemem                 -0.8966      0.063    -14.233      0.000      -1.020      -0.773
freeswap                 7.4905      0.028    271.913      0.000       7.436       7.544
runqsz_Not_CPU_Bound     0.6364      0.114      5.568      0.000       0.412       0.860
==============================================================================
Omnibus:                      583.980   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              903.390
Skew:                          -0.753   Prob(JB):                    6.78e-197
Kurtosis:                       4.230   Cond. No.                         430.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Table 11 Final OLS Approach-3

NOTE: The final model obtained (Table-11) for Approach-3 did go through a series of eliminating high VIF feature one by one, checking R-squared and further dropping features that had p-values > 0.05. Only the screenshot of final model was put up in the report so as to reduce the complexity of the report.

Test for Normality (Approach-3):

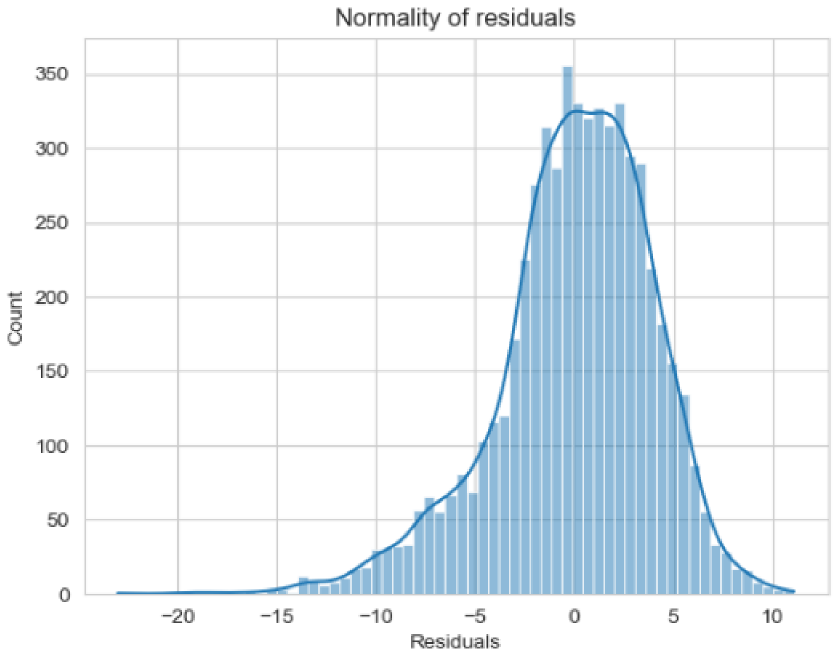


Figure 33 Normality Test

The above visual representation tells that the errors are normally distributed. It also suggests to some extent it is skewed towards its left. Since the model is built without outlier treatment considering the outliers are a legitimate value, the slight skewness could be a result of it.

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.

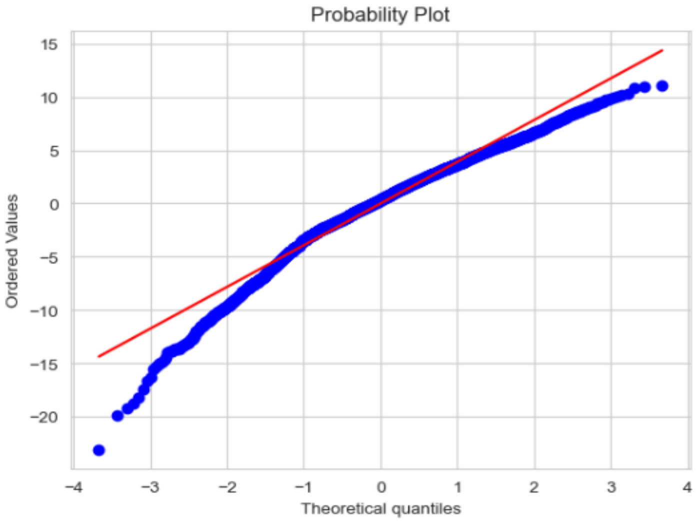
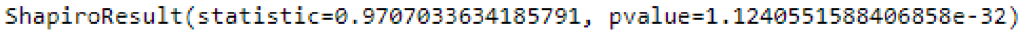


Figure 34 QQ Plot (Approach-3)

From the figure 34, it can be inferred that points do lie on the straight line and also equal number of points that lie outside the line and few being closer and few little farther. However, the QQ plot in approach-1 seemed better than the approach-3 QQ-plot.

The Shapiro-Wilk test can also be used for checking the normality. The null and alternate hypotheses of the test are as follows:

- Null hypothesis - Data is normally distributed.
- Alternate hypothesis - Data is not normally distributed.



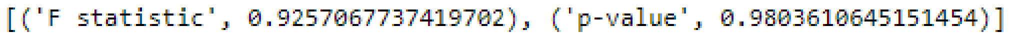
```
ShapiroResult(statistic=0.9707033634185791, pvalue=1.1240551588406858e-32)
```

- 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

Test for Homoscedasticity:

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

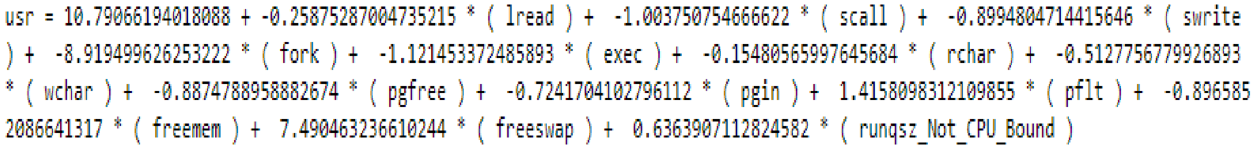
- Null hypothesis: Residuals are homoscedastic
- Alternate hypothesis: Residuals have heteroscedasticity.



```
[('F statistic', 0.9257067737419702), ('p-value', 0.9803610645151454)]
```

- Since p-value > 0.05 we can say that the residuals are homoscedastic.

Linear Regression Equation and comment on variables:



```
usr = 10.79066194018088 + -0.25875287004735215 * ( lread ) +  -1.003750754666622 * ( scall ) +  -0.8994804714415646 * ( swrite
) +  -8.919499626253222 * ( fork ) +  -1.121453372485893 * ( exec ) +  -0.15480565997645684 * ( rchar ) +  -0.5127756779926893
* ( wchar ) +  -0.8874788958882674 * ( pgfree ) +  -0.7241704102796112 * ( pgin ) +  1.4158098312109855 * ( pflt ) +  -0.896585
2086641317 * ( freemem ) +  7.490463236610244 * ( freeswap ) +  0.6363907112824582 * ( runqsz_Not_CPU_Bound )
```

Equation 2 Linear Regression Equation (Approach-3)

Comments:

- From the Equation-2, a unit increase in the exec will result in a -1.1214 unit decrease in the usr, all other variables remaining constant.

- a unit increase in the freemem will result in a 7.4904 unit increase in the usr, all other variables remaining constant.

- The usr of a Not CPU Bound in runqsz will be -0.3178 units lesser than a runqsz of CPU Bound, all other variables remaining constant.

RMSE Score

Training: 3.965124845143364

Testing: 3.9354229764644773

MAE

Training: 3.0475993791394114

Testing: 3.0323486562705573

- 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 2.0 units on the test data.

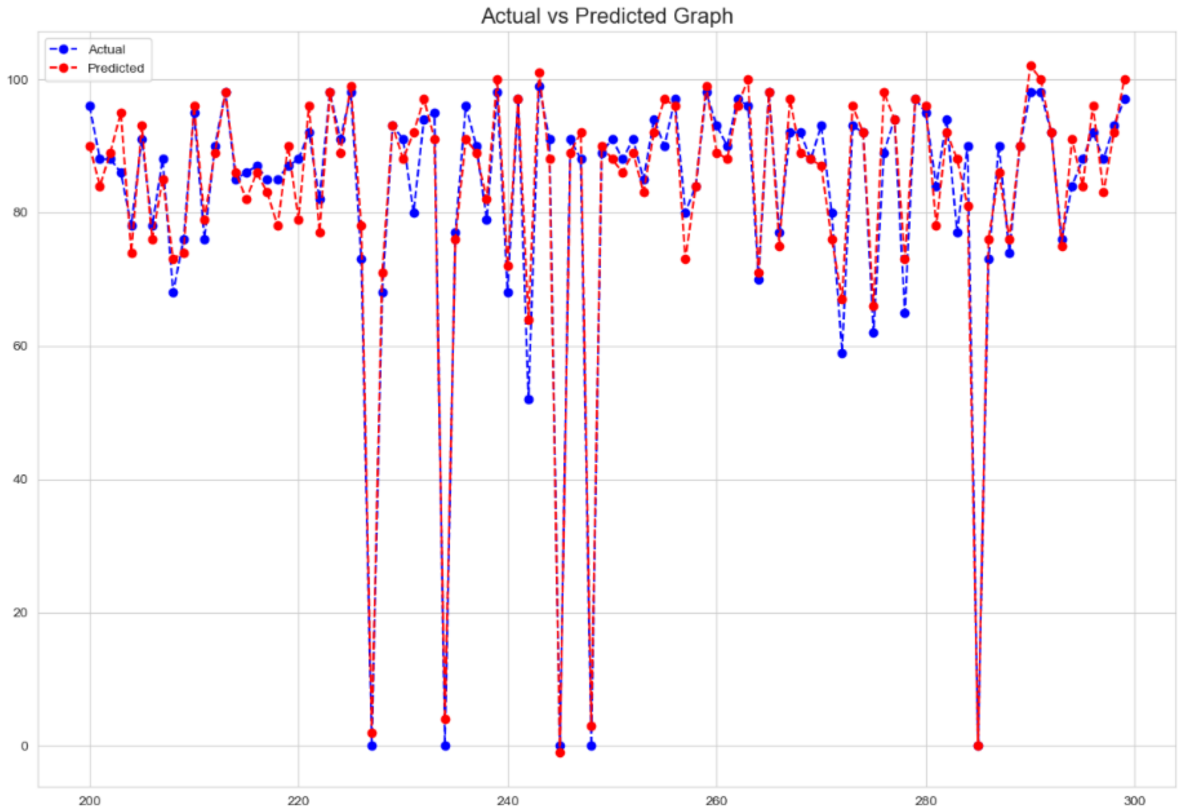


Figure 35 Actual vs Predicted Approach-3

The above plot represents the Actual vs Predicted graph of a randomly chosen 100 records. The blue represents the Actual records and the red represents the Predicted records. We can ne see a very minimal deviation when compared each other representing the model is doing a perfect job in predicting the values.

Key Observations

- In the Univariate Analysis of raw data, the observations of all continuous variables except for the target and freeswap, all others displayed a pattern which exhibits right (positive) skew. It was also observed the variable freeswap was a trimodal distribution as three peaks were observed. A detail data check by the domain could possibly explain such patterns and the target "usr" came out to be left skewed indicating higher values forming a peak and extreme lower value being lesser and dragging towards its left.

- As an exercise of going through multivariate analysis, it was observed few independent variables had a very high positive correlation among them. This also suggests that presence of such high correlation could lead us to multicollinearity and it was later handled while building the model.

- With respect to the target, all the variables except freeswap and freemem had a negative correlation. However, freeswap exhibited a correlation value of 0.27 and freemem exhibited a correlation value of 0.68 with the target.

- Since the data was observed to have zeros which in turn dominated few variables, the model building was done considering both (with and without) zeros. This method was employed since there was no domain information on the same. However, for all the approach the variable "pgscan" was dropped (not considered) for the presence of around 79% of zero in that column alone.

- The first approach involved in building model had variables that had percentage of zeros less than 50%. The NaN value was imputed using KNN Imputer. Initial model was built observed to have multicollinearity, less important features which were eliminated step by step and the final model was estimated to have explained around 90% of variance. Most of the assumptions were satisfied.

- Regularization techniques such as Ridge and Lasso model were built so as to check if it could further better the model. The performance and prediction remained less changed indicating the first approach did the right model building.

- The last approach was to consider the presence of zero as legit and build model on it. All the variables except pgscan were considered. The same methods such as initial model building, multicollinearity check and p-value were employed. This in turn explained 95% variance, better than the first approach. However, this has its own drawback when it comes to handling zeros in the production and the normality for the approach-1 looked better. Both these approaches did provide best results and considering the approach depends on the domain.

Problem-2

Executive Summary:

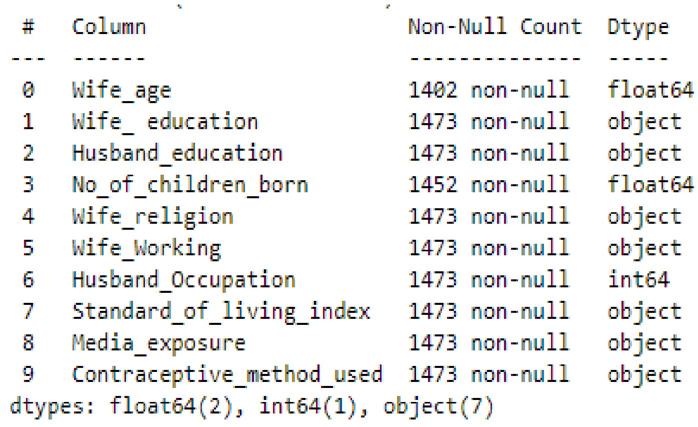
Republic of Indonesia Ministry of Health, has entrusted us with a dataset containing information from a Contraceptive Prevalence Survey. This dataset encompasses data from 1473 married females who were either not pregnant or were uncertain of their pregnancy status during the survey.

Introduction:

Expectation involves predicting whether these women opt for a contraceptive method of choice. This prediction will be based on a comprehensive analysis of their demographic and socio-economic attributes.

Data shape:

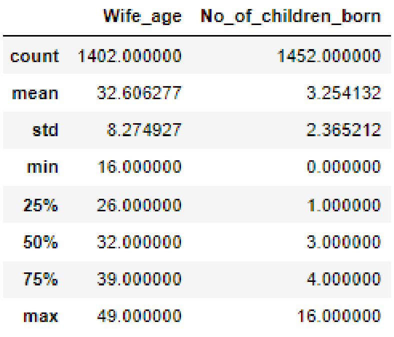
The data shape by default is said to have 1473 records and 10variables. There are a total of 4 categorical variables, 2 numerical, 3 binary and the target being binary class.



```
 #   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)
```

Table 12 Shape

Statistical Summary:



	Wife_age	No_of_children_born
count	1402.000000	1452.000000
mean	32.606277	3.254132
std	8.274927	2.365212
min	16.000000	0.000000
25%	26.000000	1.000000
50%	32.000000	3.000000
75%	39.000000	4.000000
max	49.000000	16.000000

Table 13 Summary

Univariate Analysis

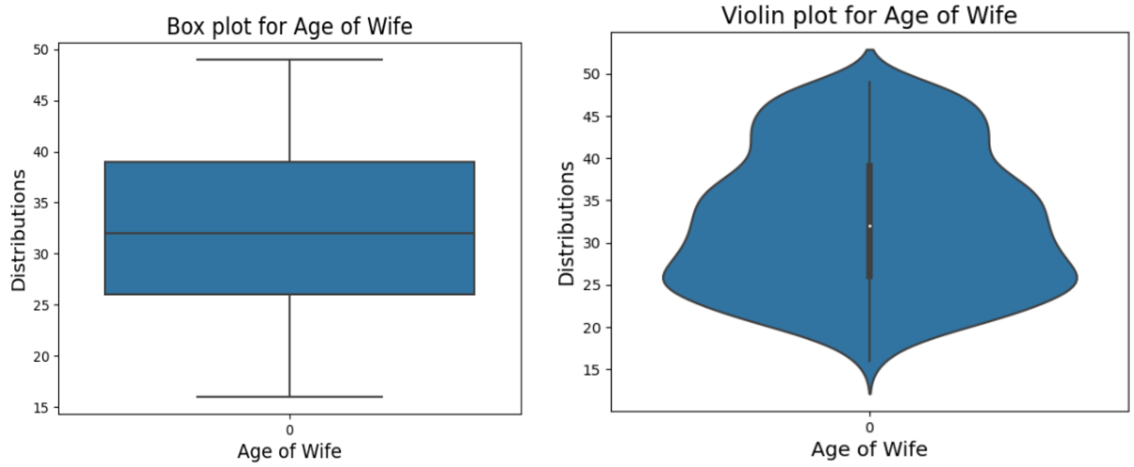


Figure 36 Age of Wife

The box plot represents presence of no outliers in the distribution and the violin plot describes a slightly right skewed distribution but does not require any transformation as the appearance is not abnormal.

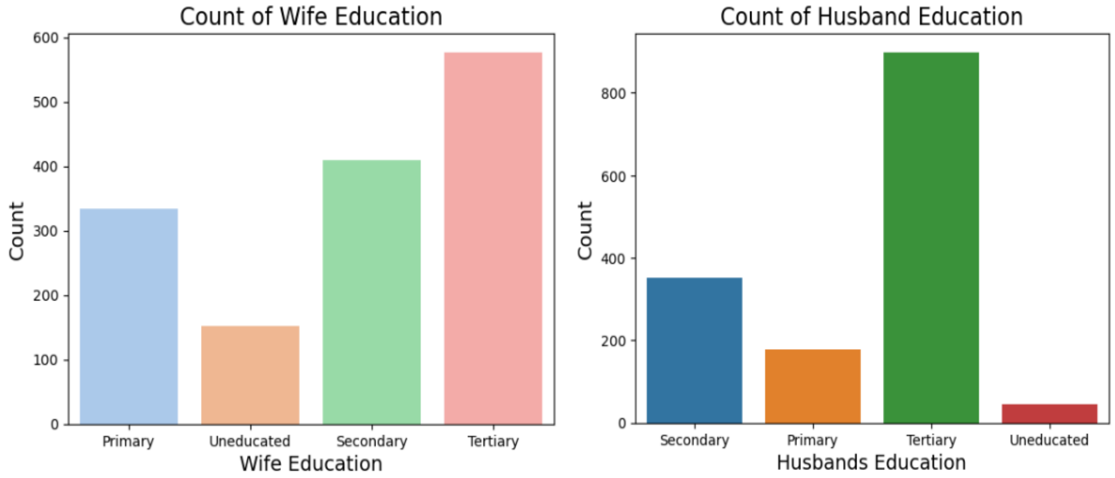


Figure 37 Wife and Husband Education

The Education status of the Husband and Wife, Both the genders tend to have higher count of Tertiary education level. Least is the Uneducated section. The Wife and Husband Education tend to have similar pattern as the count of that particular level appears more or less the same.

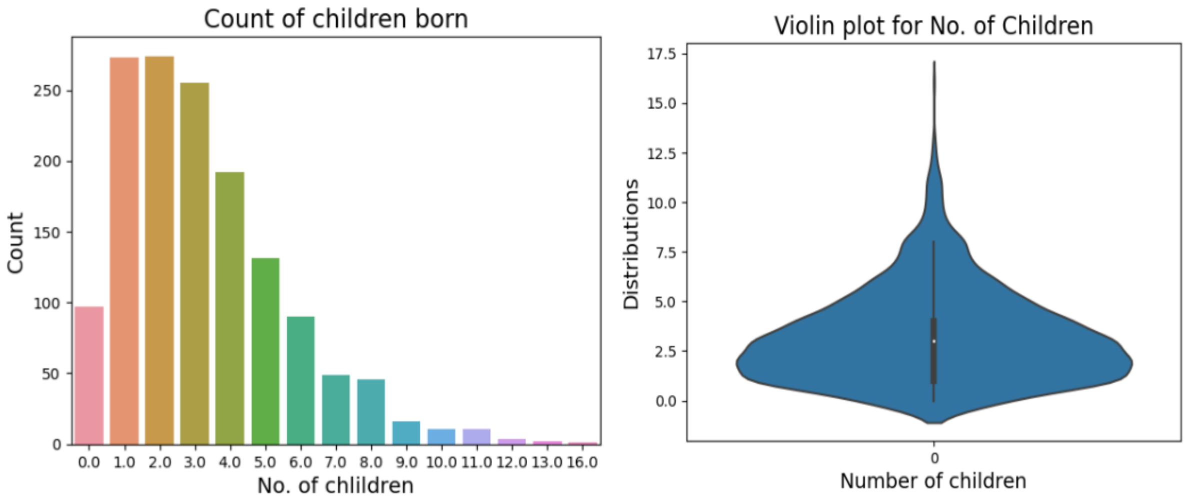


Figure 38 No. of Children

Though the number of children has been categorized as a numerical variable, it can also be visualized as a category. The left plot of Figure 38 is a categorical visual of the variable where Number of children greater than 5 has comparatively lesser count and the right plot representing the violin plot indicates presence of outliers as the plot is skewed towards right.

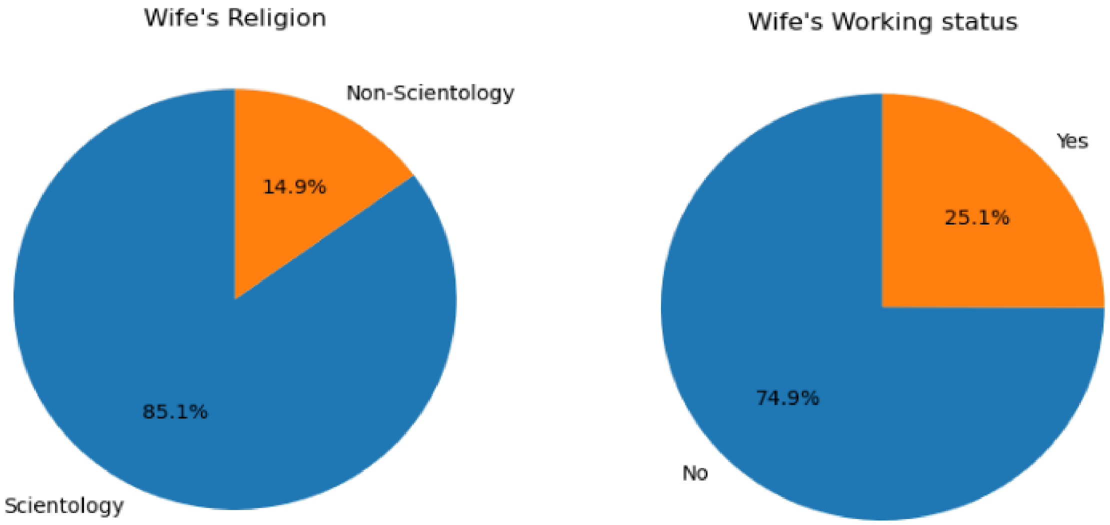


Figure 39 Wife's Religion and Working status

Figure 39 represents the pie chart of two variables wife’s religion and their working status. The wife’s religion when it is Scientology dominated the dataset as 85% of the data consists the same whereas the non-scientology is only around 15% of records in the data. Also, around 80% of the Wife’s are not working despite the education of the same dominating over uneducated.

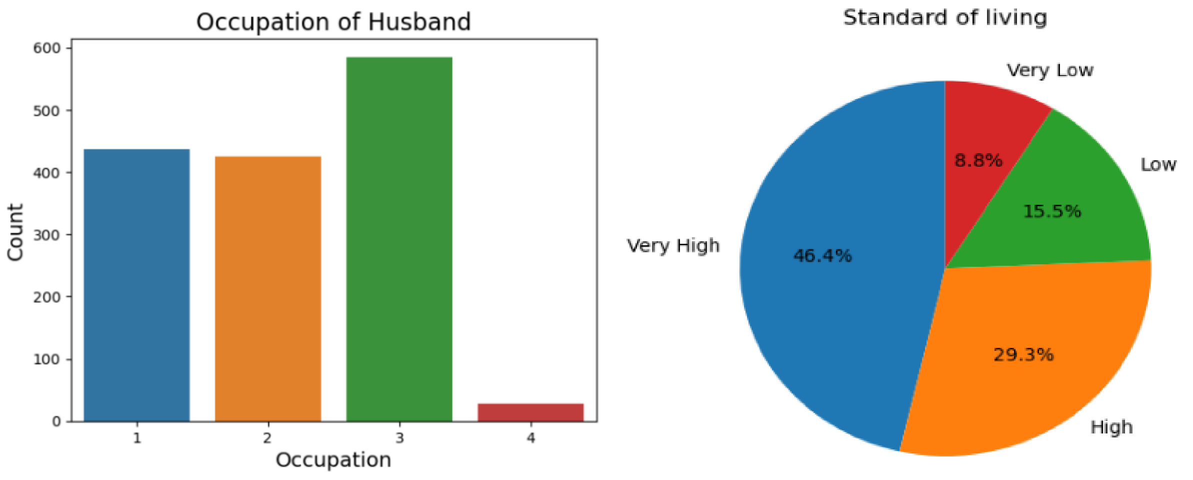


Figure 40 Husband Occupation and Standard of living

The left side plot of Figure 40 represents the Occupation. The Occupation category 3 which represents high is the highest among the category whereas the Occupation 4 is the least among the group. Right side plot of Figure 40 represents that 46.5% of the data represents the family tends to have a Very high standard of living and only 8.8% tend to have very low.

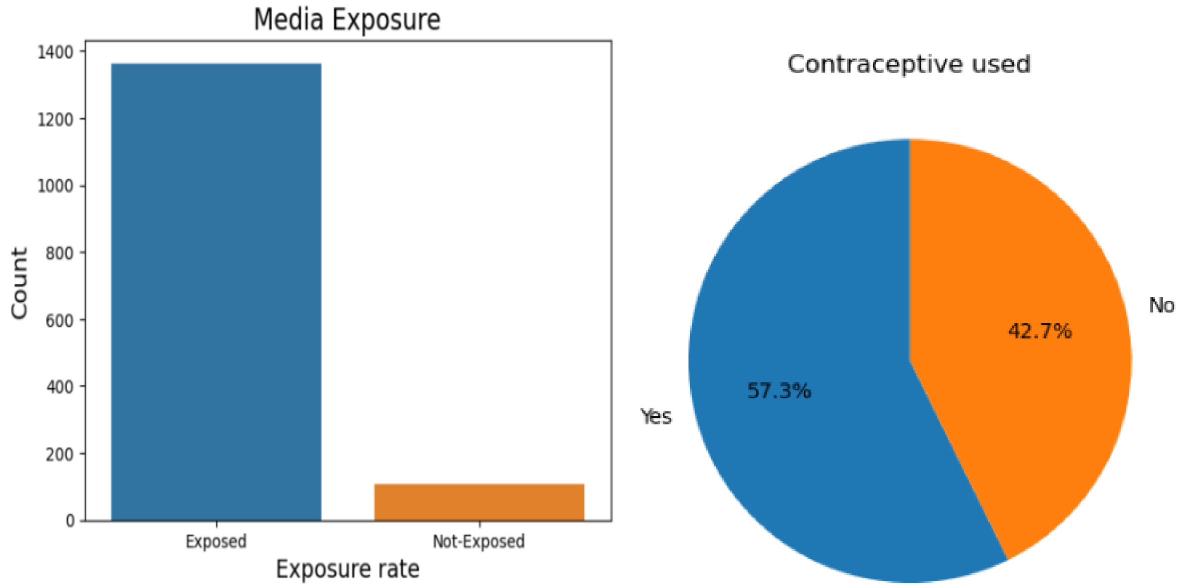


Figure 41 Media Exposure and Contraceptive used

Figure 41 to it left has the details of Media Exposure and the right side has the details of Contraceptive used. It can be noted that Exposed exposure rate is very high and dominating toan extent we can say the variable is oversampled towards Exposed. The variable contraceptive used being the target (dependent) variable eliminates the concept of oversampling as the class target is sampled if not perfectly but near perfect samples.

Multivariate Analysis

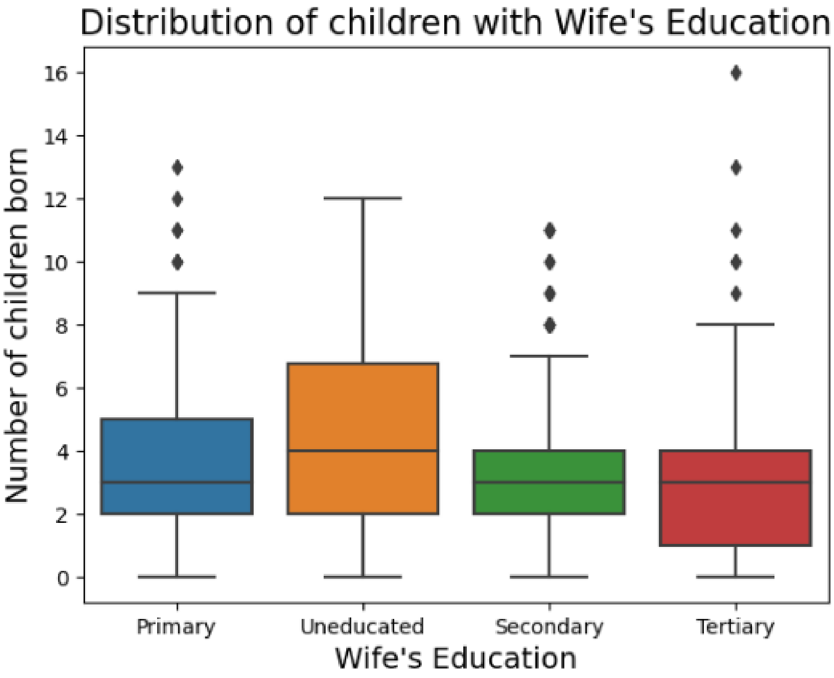


Figure 42 Distribution of children vs Wife's Education

The wife's education category appears to have same median across all category except Uneducated indicating higher among the category. The tertiary education has outliers at extreme ends, such outliers can be miscalculation or could be legit. A domain consultation could help identify such cases if not it can be considered as miscalculated outlier and then be capped.

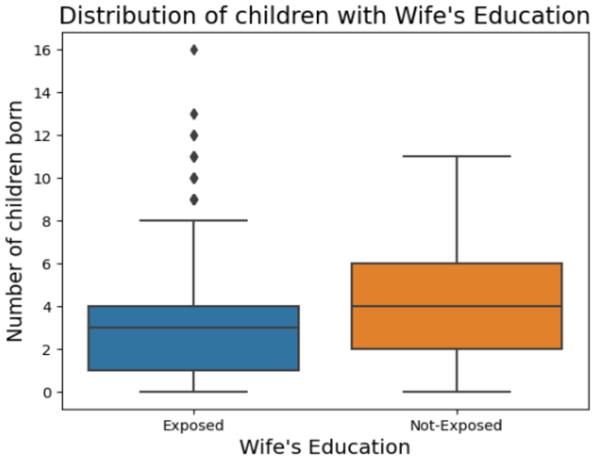


Figure 43 Distribution of children with Wife's Education

The mean of media not-exposed category is slightly higher than that of Exposed despite the Media exposed category having extreme outliers. The Not-Exposed category also indicates the distribution not having any outliers.

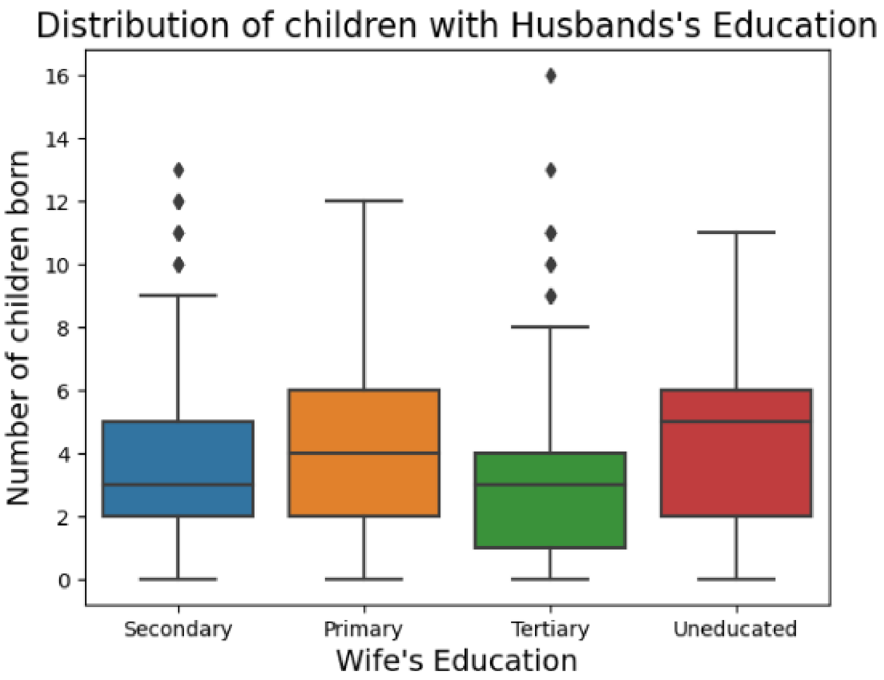
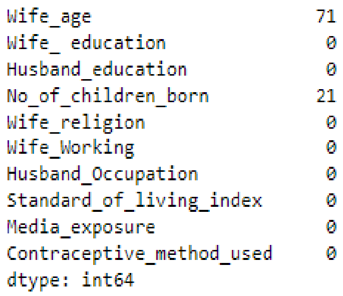


Figure 44 Children per Husband's Education

From the figure 44 it can be noted that despite the secondary and tertiary category having outliers the median of uneducated is higher compared to others in the same category. Uneducated husbands tend to have a greater number of children and a proper education on the contraceptives could reduce this significantly.

Data Preprocessing

Missing Values



```
Wife_age                    71
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
```

Table 14 Missing Values

It was noted that there were missing values in the Wife_age and No_of_children_born variables with 71 and 21 respectively.

The variables Wife age follows a normal distribution. The mean and median appear to be same. This indicates we can either use mean or median for imputing nan values.

The number of children born was imputed with the mode, the most appearing values in that particular column.

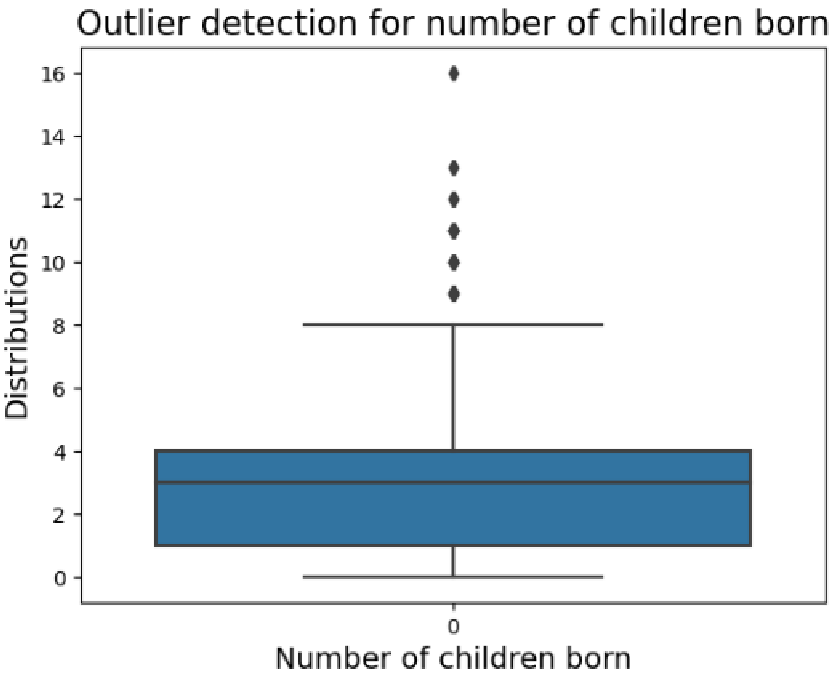


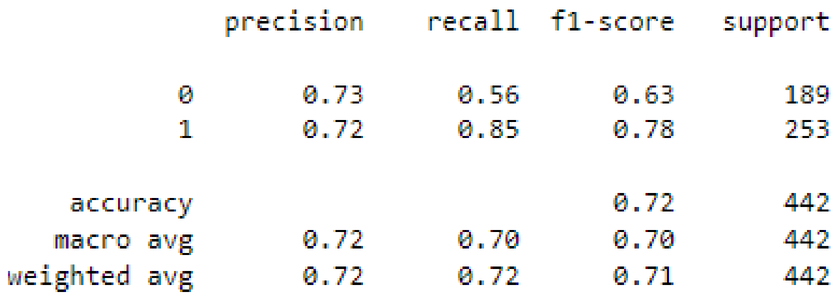
Figure 45 Outlier Detection

There are Outliers in the Number of children variable. Logistic regression can be sensitive to outliers, and their presence can influence the estimation of coefficients and impact the performance of the model. However, a two-model approach was considered one with and another without outliers. The model without outliers proved to have better accuracy, precision and recall compared to others.

The outliers were treated with the capping value. Any value above 5 number of children were categorized to fall under 5. The model was then split into train, test and split with a random state of 7 the test sample was taken to be 30%.

Logistic Regression:

Model Score: 72%



```
              precision    recall  f1-score   support

           0       0.73      0.56      0.63       189
           1       0.72      0.85      0.78       253

    accuracy                           0.72       442
   macro avg       0.72      0.70      0.70       442
weighted avg       0.72      0.72      0.71       442
```

Table 15 Classification report

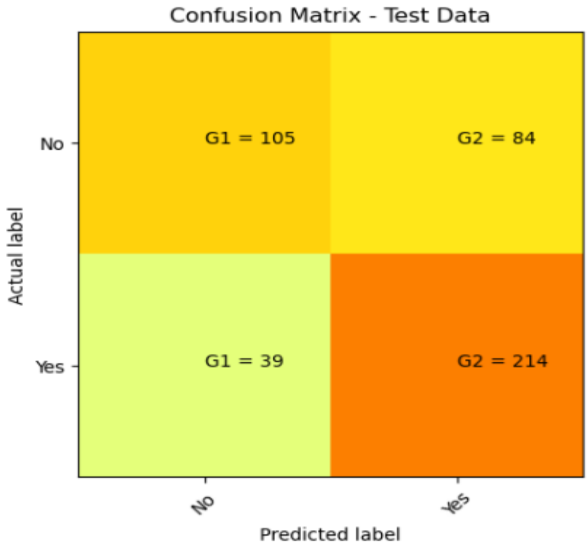


Figure 46 Confusion Matrix

Observations

Precision is the ratio of true positive predictions to the total number of positive predictions made by the model.

For class 0: Precision = 0.73

For class 1: Precision = 0.72

Interpretation: Out of all instances predicted as positive, 73% (class 0) and 72% (class 1) were actually positive.

Recall is the ratio of true positive predictions to the total number of actual positive instances in the dataset.

For class 0: Recall = 0.56

For class 1: Recall = 0.85

Interpretation: Out of all actual positive instances, the model captured 56% (class 0) and 85% (class 1).

F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics.

For class 0: F1-Score = 0.63

For class 1: F1-Score = 0.78

Interpretation: The F1-score considers both precision and recall, providing a single metric that balances the trade-off between false positives and false negatives.

```
AUC: 0.759
```

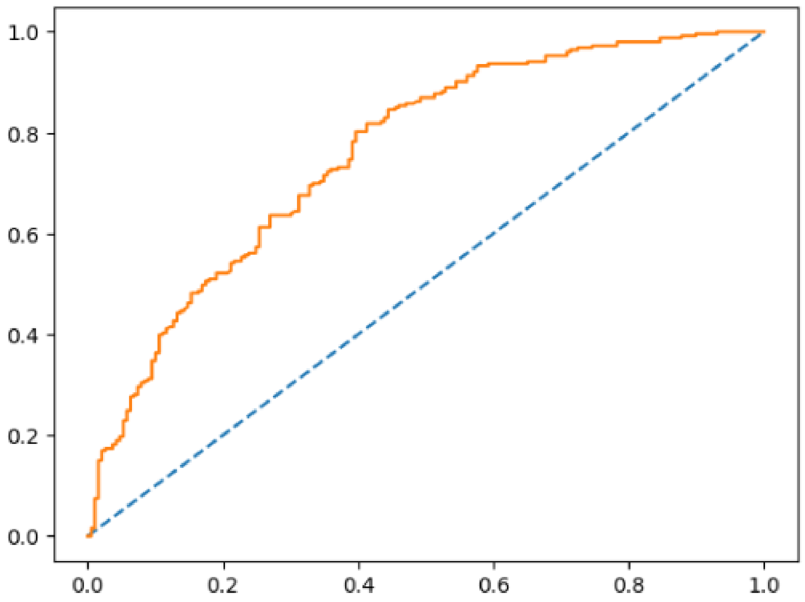


Figure 47 ROC AUC

The AUC score for the Logistic regression is 76% which appears pretty good since we have only 1473 records which is low. A data set with more records will eventually reflect in the model accuracy and predictions.

Linear Discriminant Analysis

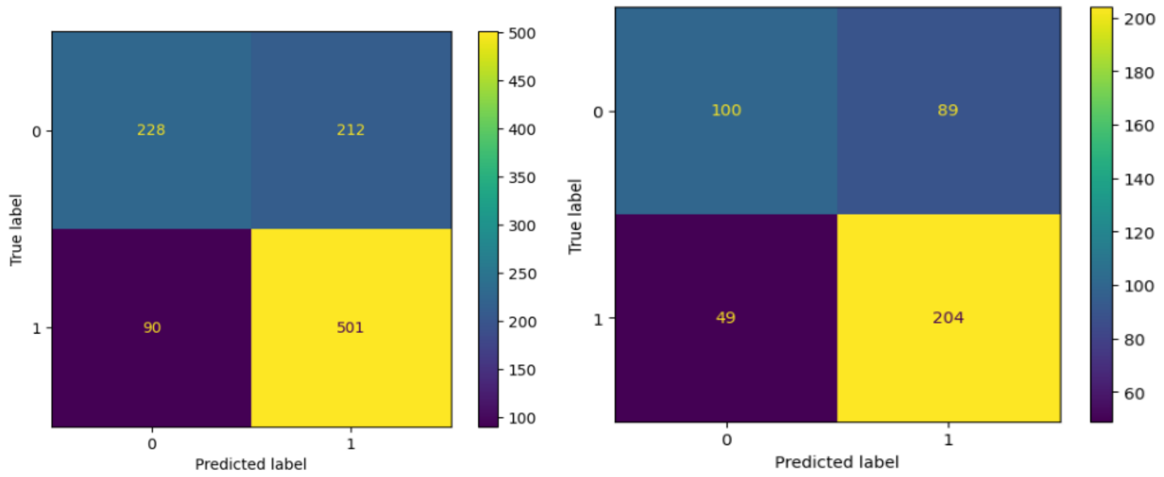
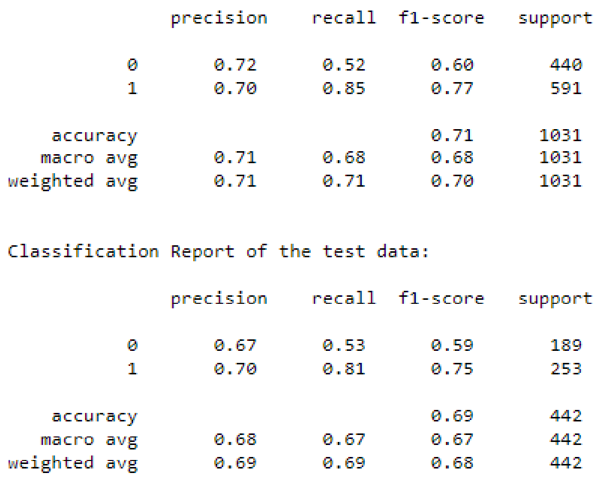


Figure 48 Training and Testing Confusion Matrix

Figure 48 represents Training and Testing prediction confusion matrix. The prediction of 1 (yes) which is a variable of interest is predicted by the model in the same proportion. This could possibly indicate the stability of the model. However, the model accuracy might not be satisfying considering the number of records we are dealing with.



```
              precision    recall  f1-score   support

           0       0.72      0.52      0.60       440
           1       0.70      0.85      0.77       591

    accuracy                           0.71      1031
   macro avg       0.71      0.68      0.68      1031
weighted avg       0.71      0.71      0.70      1031

Classification Report of the test data:

              precision    recall  f1-score   support

           0       0.67      0.53      0.59       189
           1       0.70      0.81      0.75       253

    accuracy                           0.69       442
   macro avg       0.68      0.67      0.67       442
weighted avg       0.69      0.69      0.68       442
```

Table 16 LDA Classification report

Observations

Precision is the ratio of true positive predictions to the total number of positive predictions made by the model.

For class 0: Precision = 0.72

For class 1: Precision = 0.70

Interpretation: Out of all instances predicted as positive, 72% (class 0) and 70% (class 1) were actually positive.

Recall is the ratio of true positive predictions to the total number of actual positive instances in the dataset.

For class 0: Recall = 0.52

For class 1: Recall = 0.85

Interpretation: Out of all actual positive instances, the model captured 52% (class 0) and 85% (class 1).

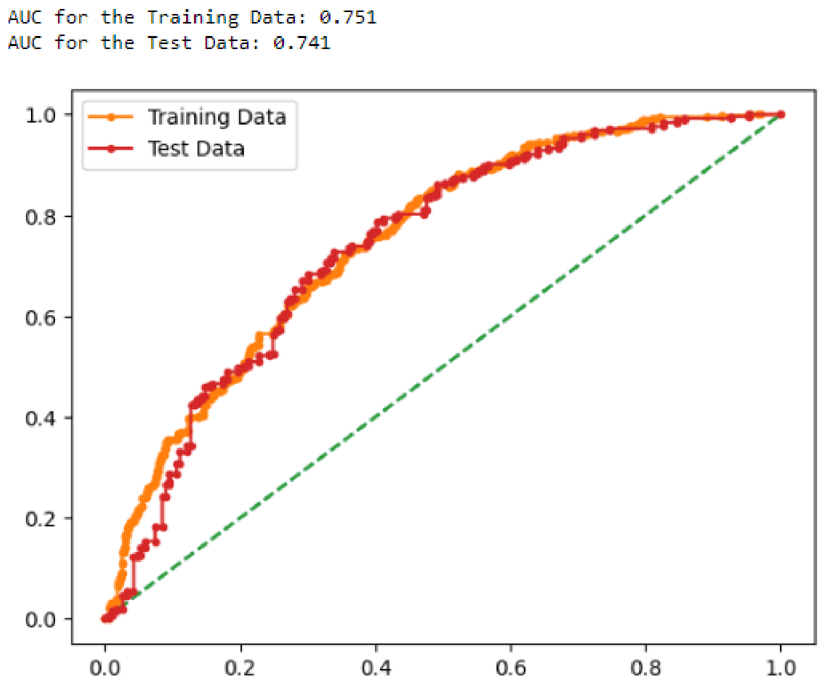
F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics.

For class 0: F1-Score = 0.59

For class 1: F1-Score = 0.75

Interpretation: The F1-score considers both precision and recall, providing a single metric that balances the trade-off between false positives and false negatives.

The **Model Accuracy** comes out to be 69%



```
AUC for the Training Data: 0.751
AUC for the Test Data: 0.741
```

Figure 49 AUC Curve LDA Model

The Curve for the Training and Testing looks similar. The model is quite stable so far but with a decrease in accuracy compared the logistic Regression. Emphasise on the data could bring even mode stability higher AUC, ROC curve and model accuracy.

$$LDF = 0.36725944 + X1 * -0.81 + X2 * (0.56) + X3 * (-0.01) + X4 * (0.99) + X5 * (-0.1) + X6 * (-0.02) + X7 * 0.02 + X8 * 0.29 + X9 * 0.13$$

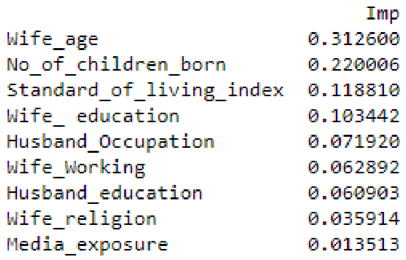
Equation 3 Linear Discriminant Function

By the above equation and the coefficients, it is clear that

- Predictor 'Number of children born' has the largest magnitude thus this helps in classifying the best
- Predictor 'Wife age' has the smallest magnitude thus this helps in classifying the least

CART Model

Independent and Dependent variables were split into X and y respectively. Initially the model was fit with Decision Tree Classifier with the gini as a criterion to split the root node. Important features were extracted as shown below.



```
                             Imp
Wife_age                0.312600
No_of_children_born     0.220006
Standard_of_living_index 0.118810
Wife_ education         0.103442
Husband_Occupation      0.071920
Wife_Working            0.062892
Husband_education       0.060903
Wife_religion           0.035914
Media_exposure          0.013513
```

Table 17 Important CART features

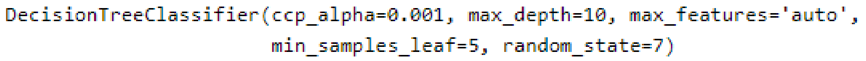
For the basic model that was fit the important features came out to be Wife age and Media exposure to be the least important feature.

Decision tree model scores:

Training: 98%

Testing: 62%

By looking at the training and testing score we can say that the model is overfitting, a result of high variance. To overcome this pruning method was employed by hyper tuning the parameters using Grid Search CV.



```
DecisionTreeClassifier(ccp_alpha=0.001, max_depth=10, max_features='auto',
                       min_samples_leaf=5, random_state=7)
```

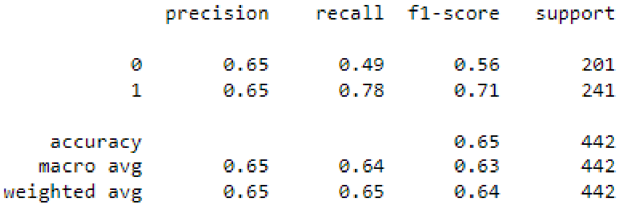
These were used as the best estimators and again the model was built.

Decision tree model scores (Hyper tuning the parameters):

Training: 80%

Testing: 65%

Though the training score is reduced the testing score is increased just by 2%. However, the model is still overfitting. This shows decision trees are prone to overfitting.



```
              precision    recall  f1-score   support

           0       0.65      0.49      0.56       201
           1       0.65      0.78      0.71       241

    accuracy                           0.65       442
   macro avg       0.65      0.64      0.63       442
weighted avg       0.65      0.65      0.64       442
```

Table 18 CART Classification matrix

The classification matrix for the testing data of the CART model post pruning the data yet appears dissatisfactory as the results are not up to the mark. A better approach with a different algorithm such as Random Forest etc. could be a better alternate.

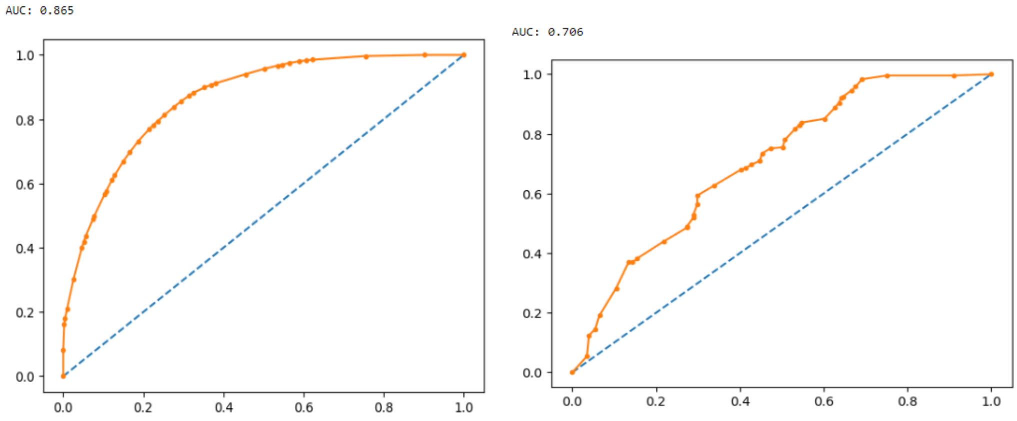


Figure 50 AUC ROC CART (Pruned)

The AUC for the training is better than the AUC for the testing data yet again indicating overfitting. The AUC score is 86% for training and 70% for testing. Emphasising more on the quantity of the data and advanced algorithms such as random forest, XG Boost etc. could make the model better.

Business Insights and Recommendations

Importance of feature based on best model

- It was observed for Linear Discriminant Analysis that the Variable Number of children born has the largest magnitude further helping in classifying.

- Linear discriminant Analysis focuses on the magnitude and coefficients are considered to do so. The higher the coefficients larger the magnitude and important the feature. Lower the coefficients lower the magnitude and least important the feature.

- In LDA, Wife age had the lesser magnitude and further specifying less importance I the model building.

- The CART model provides information on feature importance where Wife age is given the upper most importance and surprisingly the opposite of what LDA model suggested

- The least important came out to be Media exposure. To note the way Decision tree splits it would be better if there was absence of under sampling or else they have to carry crucial information.

Actionable Insights and recommendations

- The Univariate analysis was performed to analyse the pattern displayed by them. It can be noted the age of Wife has almost normal distribution with absence of outliers. Count of Education for both husband and wife have similar pattern where tertiary has the highest count and uneducated being the least.

- Multivariate Analysis was performed to check the distribution of continuous alongside the category. The missing value were identified and treated with proper imputation techniques.

- A two-fold approach was considered. The first one was to considered the outliers in Number of children born were legit and another was to consider them as exaggerated values where they will be capped to the nearest legit value. However, when the model was built for both the approach the second model provided better results and was considered.

- Logistic Regression was the first algorithm considered. The results provided by the model was pretty satisfying considering the number of records (less). The model accuracy, precision, recall and f1-score was calculated.

- Similarly, Linear Discriminant Analysis and CART (Decision tree) was built to check for their approach (feature importance) and accuracy. The LDA could divide the target as it is meant to and the accuracy was considerable.

- The decision tree model was built initially without tuning any parameters. The result was overfitting and variance was the cause of such results. Another model was built by pruning the tree and hyper tuning the parameters. Grid search CV was considered as the multi fold approach. The results did reduce the overfitting slightly but not significantly further proving that decision tree might not be the right option for the data provided.

- The Logistic Regression and LDA model provided better results compared to that of Decision tree. Considering either of Logistic regression or LDA could provide the desirable results. It is also important to note that the results can be astonishing if the quantity of data is more.