BUSINESS REPORT ON PREDICTIVE MODELLING PROJECT

- RAHUL SHARMA

A Problem1: Comp-active database

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.

Being an aspiring data scientist, our aim to establish a linear equation for predicting 'usr' (the percentage of time CPUs operate in user mode). Your goal is to analyze various system attributes to understand their influence on the system's 'usr' mode.

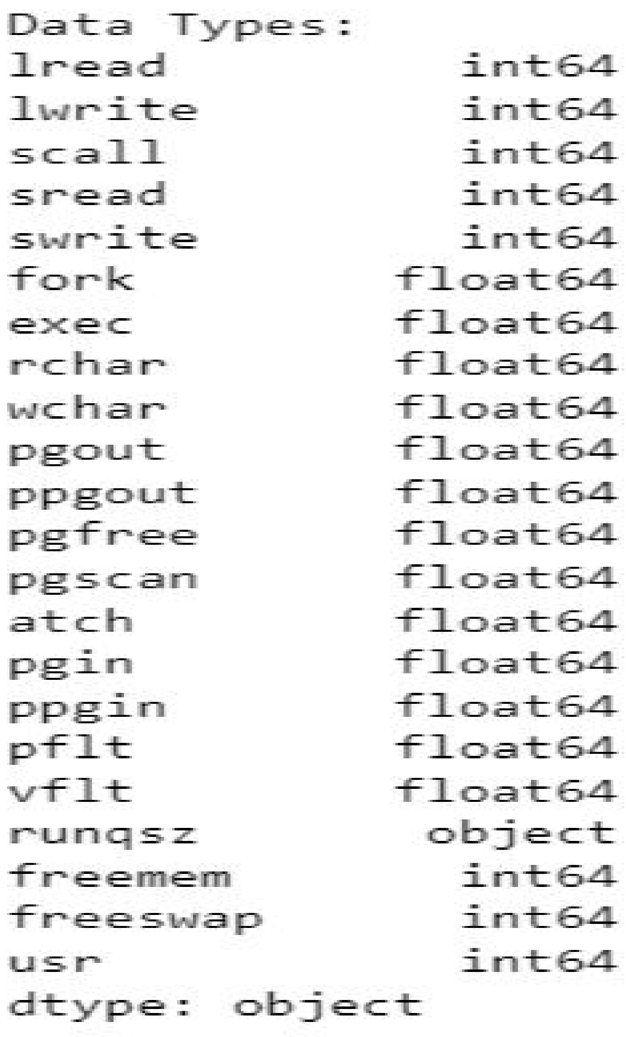
1.1 - Define the problem and perform exploratory Data Analysis

Problem definition - Check shape, Data types, statistical summary - Univariate analysis - Multivariate analysis - Use appropriate visualizations to identify the patterns and insights - Key meaningful observations on individual variables and the relationship between variables.

Problem definition:-

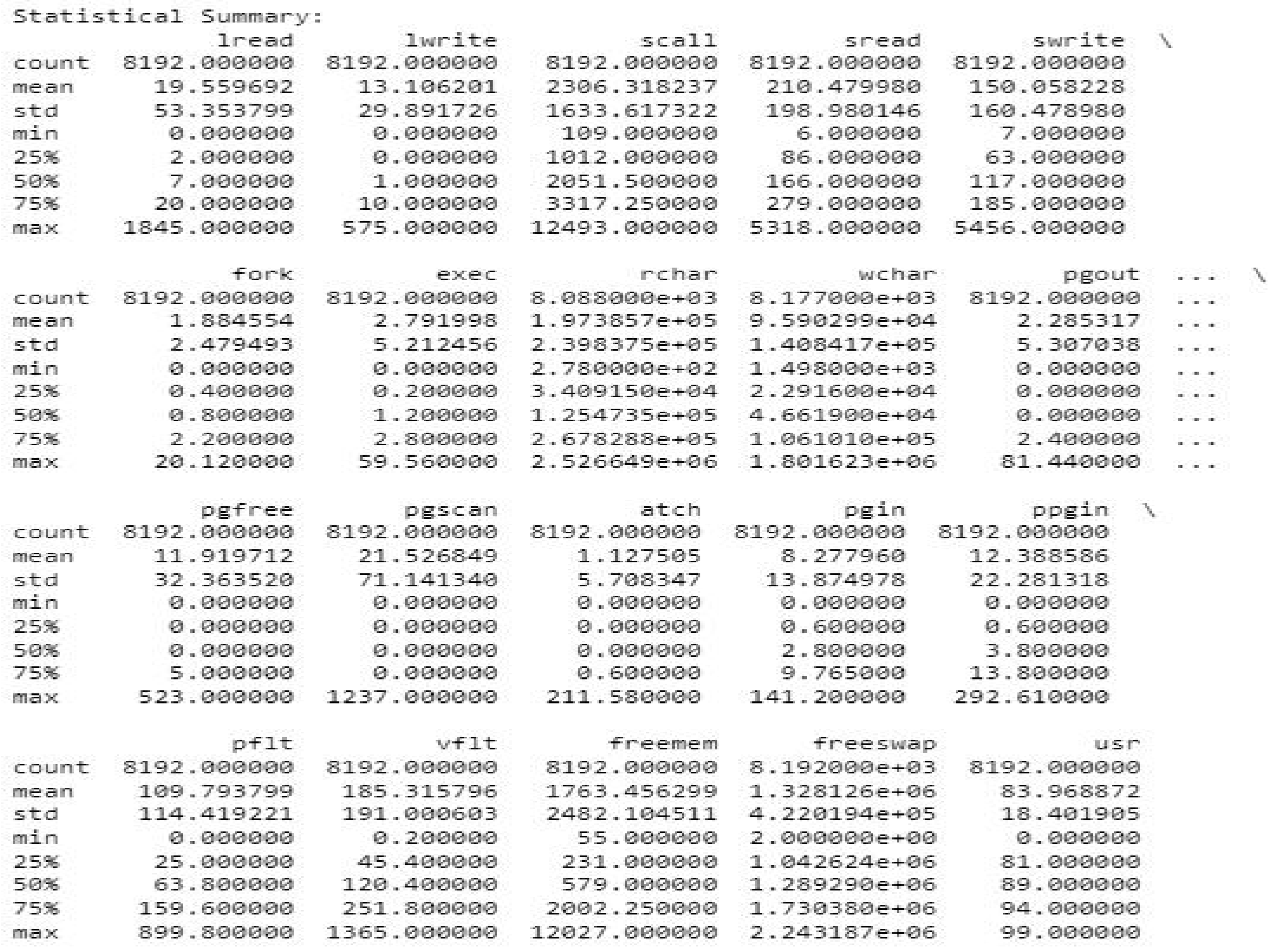
Check shape:- 8192 rows x 22columns

Data types:-



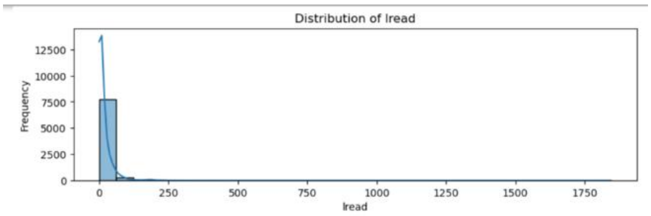
```
Data Types:
lread           int64
lwrite          int64
scall           int64
sread           int64
swrite          int64
fork          float64
exec          float64
rchar         float64
wchar         float64
pgout         float64
ppgout        float64
pgfree        float64
pgscan        float64
atch          float64
pgin          float64
ppgin         float64
pflt          float64
vflt          float64
runqsz         object
freemem         int64
freeswap        int64
usr             int64
dtype: object
```

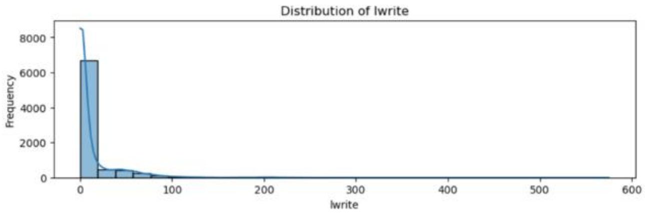
Statistical Summary:-

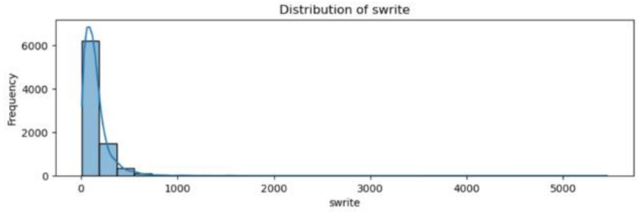


```
Statistical Summary:
             lread       lwrite         scall        sread       swrite  \
count  8192.000000  8192.000000   8192.000000  8192.000000  8192.000000
mean     19.559692    13.106201   2306.318237   210.479980   150.058228
std      53.353799    29.891726   1633.617322   198.980146   160.478980
min       0.000000     0.000000    109.000000     6.000000     7.000000
25%       2.000000     0.000000   1012.000000    86.000000    63.000000
50%       7.000000     1.000000   2051.500000   166.000000   117.000000
75%      20.000000    10.000000   3317.250000   279.000000   185.000000
max    1845.000000   575.000000  12493.000000  5318.000000  5456.000000

              fork         exec         rchar         wchar        pgout  ...  \
count  8192.000000  8192.000000  8.088000e+03  8.177000e+03  8192.000000  ...
mean      1.884554     2.791998  1.973857e+05  9.590299e+04     2.285317  ...
std       2.479493     5.212456  2.398375e+05  1.408417e+05     5.307038  ...
min       0.000000     0.000000  2.780000e+02  1.498000e+03     0.000000  ...
25%       0.400000     0.200000  3.409150e+04  2.291600e+04     0.000000  ...
50%       0.800000     1.200000  1.254735e+05  4.661900e+04     0.000000  ...
75%       2.200000     2.800000  2.678288e+05  1.061010e+05     2.400000  ...
max      20.120000    59.560000  2.526649e+06  1.801623e+06    81.440000  ...

            pgfree       pgscan         atch         pgin        ppgin  \
count  8192.000000  8192.000000  8192.000000  8192.000000  8192.000000
mean     11.919712    21.526849     1.127505     8.277960    12.388586
std      32.363520    71.141340     5.708347    13.874978    22.281318
min       0.000000     0.000000     0.000000     0.000000     0.000000
25%       0.000000     0.000000     0.000000     0.600000     0.600000
50%       0.000000     0.000000     0.000000     2.800000     3.800000
75%       5.000000     0.000000     0.600000     9.765000    13.800000
max     523.000000  1237.000000   211.580000   141.200000   292.610000

              pflt         vflt       freemem      freeswap          usr
count  8192.000000  8192.000000   8192.000000  8.192000e+03  8192.000000
mean    109.793799   185.315796   1763.456299  1.328126e+06    83.968872
std     114.419221   191.000603   2482.104511  4.220194e+05    18.401905
min       0.000000     0.200000     55.000000  2.000000e+00     0.000000
25%      25.000000    45.400000    231.000000  1.042624e+06    81.000000
50%      63.800000   120.400000    579.000000  1.289290e+06    89.000000
75%     159.600000   251.800000   2002.250000  1.730380e+06    94.000000
max     899.800000  1365.000000  12027.000000  2.243187e+06    99.000000
```

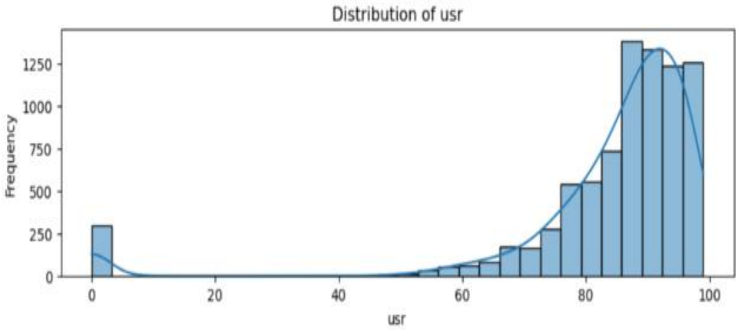
Uni-variate analysis:-





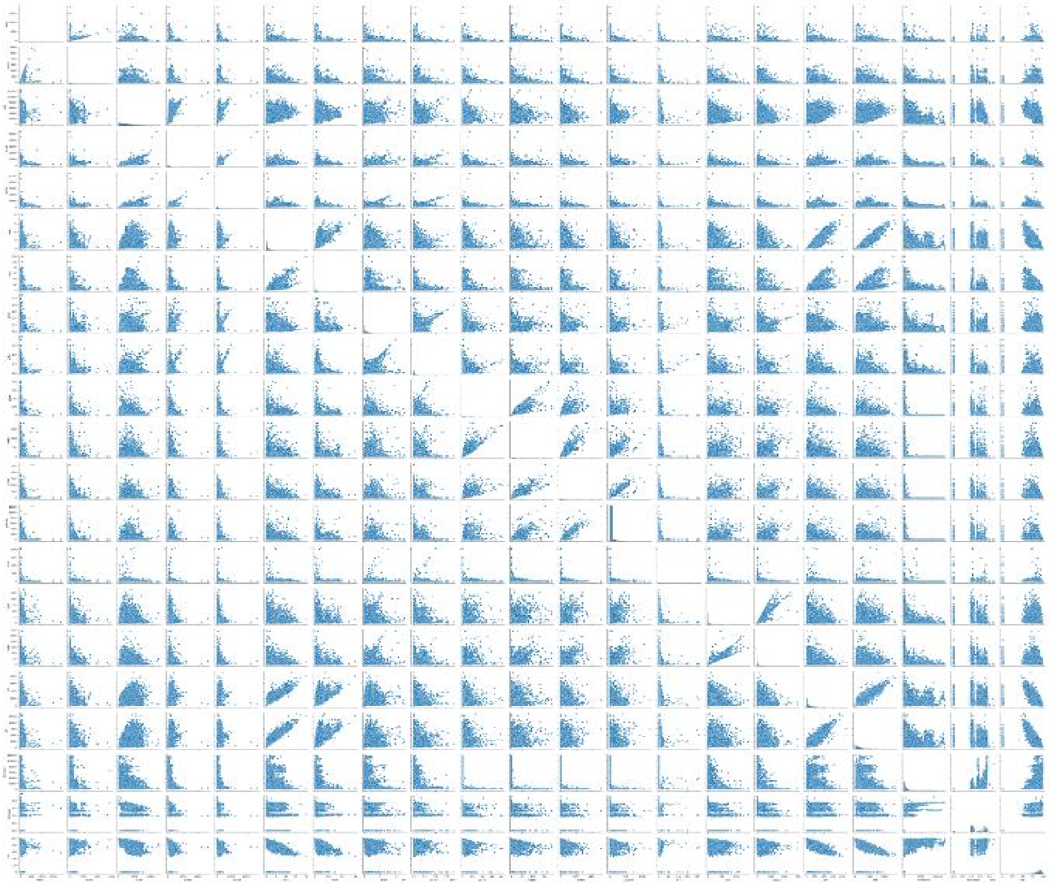


- The transfers per second for both reading and writing are brisk, with the majority occurring at a rapid pace.
- Most transactions are swiftly processed by the system, with a read-write rate that is generally quick, typically under 5%.
- The current situation suggests a relative absence of ongoing activities.



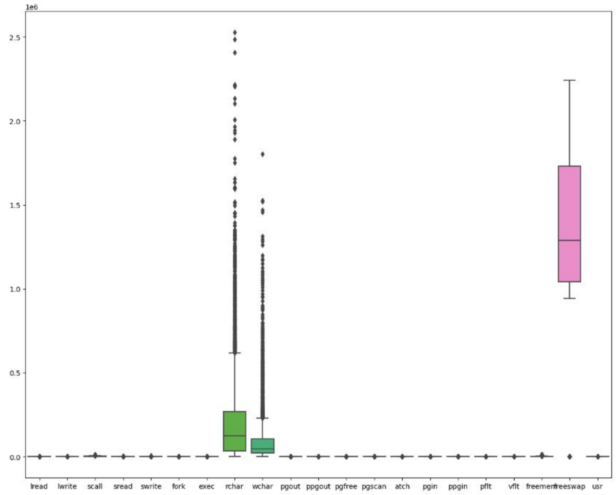
CPU able to run in user mode b/w 80- 99% times & its ideal.

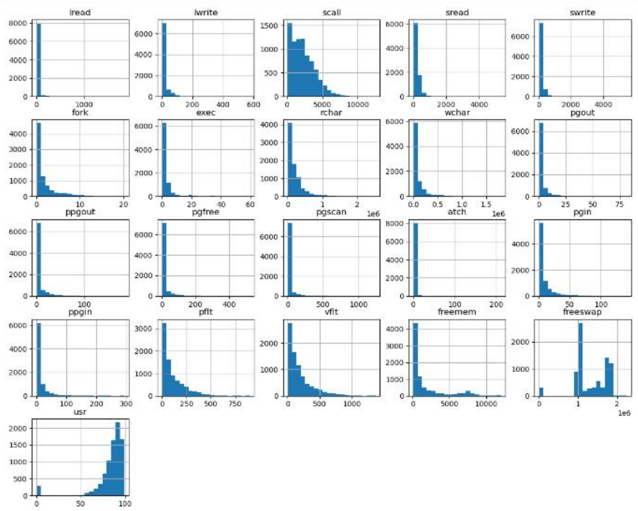
Multivariate analysis:-

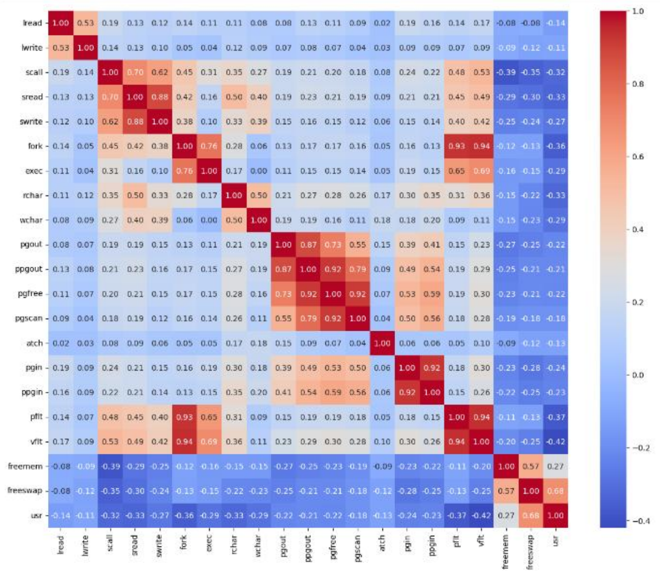


- A correlation can be observed between 'vflt,' 'pflt,' and 'fork,' suggesting that an increase in fork calls is associated with a rise in page faults.
- Likewise, there is a strong correlation between the number of page out requests per second and the number of pages paged out per second.

Use appropriate visualizations to identify the patterns and insights:-







- The read system call is the most frequently used call, with an average of 53 calls per second. This is likely because it is used to read data from files and devices.
- The write system call is the second most frequently used call, with an average of 39 calls per second. This is likely because it is used to write data to files and devices.

- The fork system call is the third most frequently used call, with an average of 24 calls per second. This is likely because it is used to create new processes.
- The sread system call is the fourth most frequently used call, with an average of 21 calls per second. This is likely because it is used to read data from sockets.
- The swrite system call is the fifth most frequently used call, with an average of 15 calls per second. This is likely because it is used to write data to sockets.

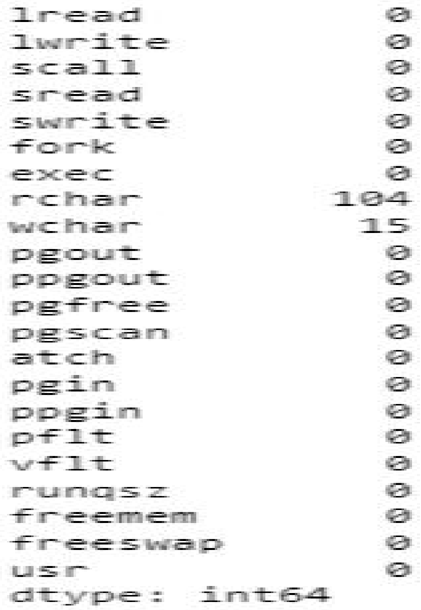
Key meaningful observations on individual variables and the relationship between variables:-

- Memory Metrics Tango: The amount of available memory (freemem) and its companions are closely connected. When the system needs to use the swap space (a backup memory area), it's like a dance, but a bit more structured.
- I/O, the Lone Wolf: Input and output operations (I/O), represented by sread and swrite, follow their own rhythm. They're less connected to the overall system, moving to their unique beat.
- PFIT Playing Ping-Pong: The page fitting process (pfit) plays a game of ping-pong. It makes fewer mistakes on its own, allowing other processes more freedom to move and operate smoothly.
- CPU, the Independent Actor: The Central Processing Unit (CPU) acts independently. When it executes (exec) or forks, it does so on its own stage, less dependent on other parts of the system.
- System, a Grand Ensemble: The entire system is like a grand ensemble. Many intricate connections exist, and when one metric makes a move (twirls), it affects the entire dance. Everything is interconnected, and each part influences the whole performance.

1.2 Data Pre-processing

Prepare the data for modelling: - Missing Value Treatment (if needed) - Outlier Detection (treat, if needed) - Feature Engineering - Encode the data - Train-test split

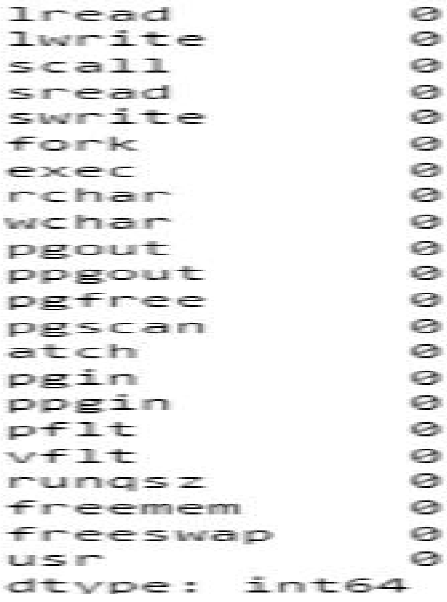
Missing Value Treatment (if needed)



```
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
dtype: int64
```

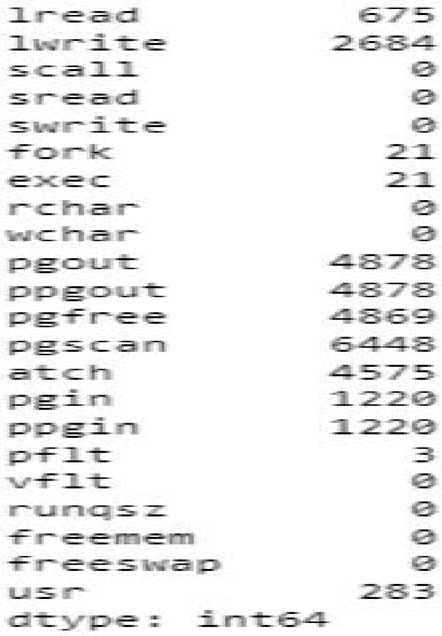
There are 104 missing values present at rchar & 15 at wchar

AFTER TREATMENT:-



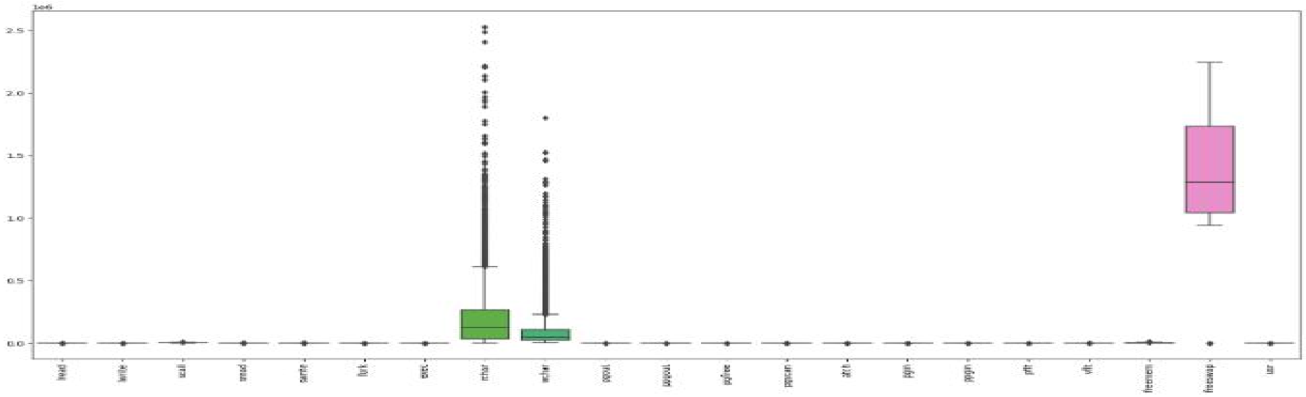
```
lread        0
lwrite       0
scall        0
sread        0
swrite       0
fork         0
exec         0
rchar        0
wchar        0
pgout        0
ppgout       0
pgfree       0
pgscan       0
atch         0
pgin         0
ppgin        0
pflt         0
vflt         0
runqsz       0
freemem      0
freeswap     0
usr          0
dtype: int64
```

Outlier Detection (treat, if needed):-



```
lread        675
lwrite      2684
scall          0
sread          0
swrite         0
fork          21
exec          21
rchar          0
wchar          0
pgout       4878
ppgout      4878
pgfree      4869
pgscan      6448
atch        4575
pgin        1220
ppgin       1220
pflt           3
vflt           0
runqsz         0
freemem        0
freeswap       0
usr          283
dtype: int64
```

- **There are total 31775 outliers present**
- **All the outliers are treated by adjusting them to the lower and upper bound values calculated by the IQR value.**

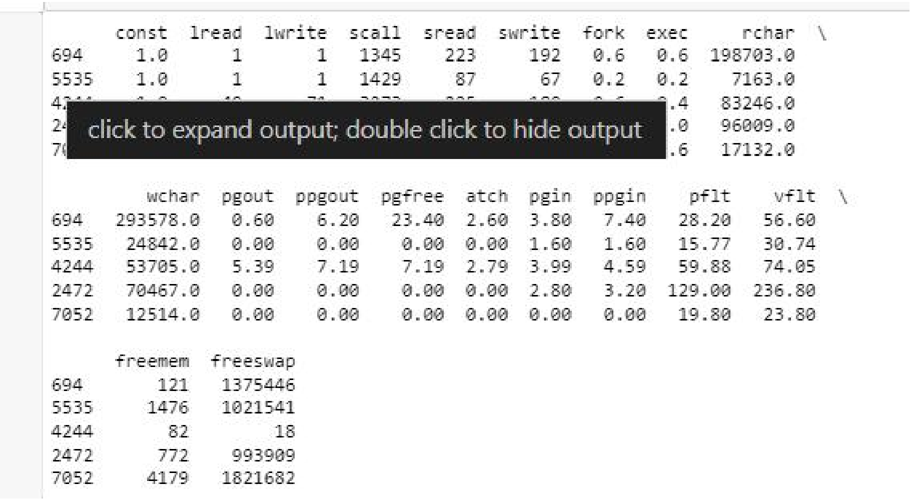


Feature Engineering:-

- New features - no. of page rate & page requests rate have been added/created with the variables pgin, pgout, ppgin & ppgout.
- Although, these new features has not given any significant output, as the majority of the values are in form of 0 or inf.

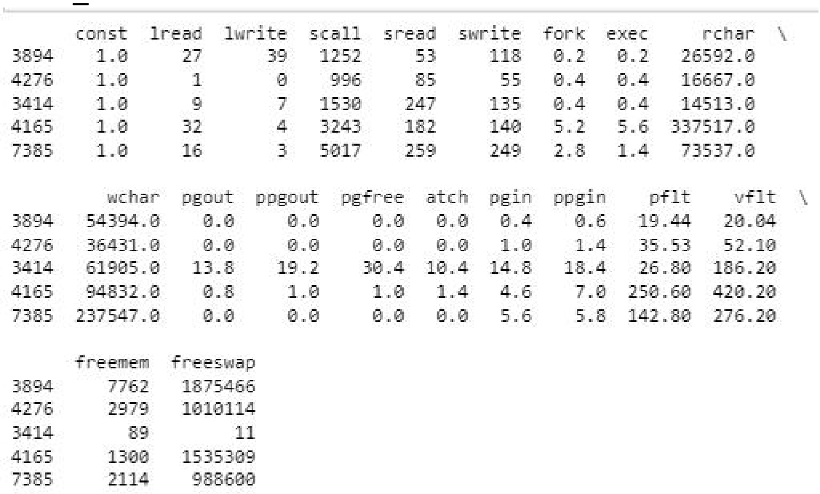
Encode the data - Train-test split:-

- After the encoded the data, the data-set has split-ted into training and testing in the 70:30 ratio.
- X_TRAIN 1st 5 rows:-



```
      const  lread  lwrite  scall  sread  swrite  fork  exec     rchar  \
694     1.0      1       1   1345    223     192   0.6   0.6  198703.0
5535    1.0      1       1   1429     87      67   0.2   0.2    7163.0
42                                                            .4   83246.0
24                                                            .0   96009.0
70                                                            .6   17132.0

         wchar  pgout  ppgout  pgfree  atch  pgin  ppgin    pflt    vflt  \
694   293578.0   0.60    6.20   23.40  2.60  3.80   7.40   28.20   56.60
5535   24842.0   0.00    0.00    0.00  0.00  1.60   1.60   15.77   30.74
4244   53705.0   5.39    7.19    7.19  2.79  3.99   4.59   59.88   74.05
2472   70467.0   0.00    0.00    0.00  0.00  2.80   3.20  129.00  236.80
7052   12514.0   0.00    0.00    0.00  0.00  0.00   0.00   19.80   23.80

      freemem  freeswap
694       121   1375446
5535     1476   1021541
4244       82        18
2472      772    993909
7052     4179   1821682
```

- X_TEST 1ST 5 rows:-

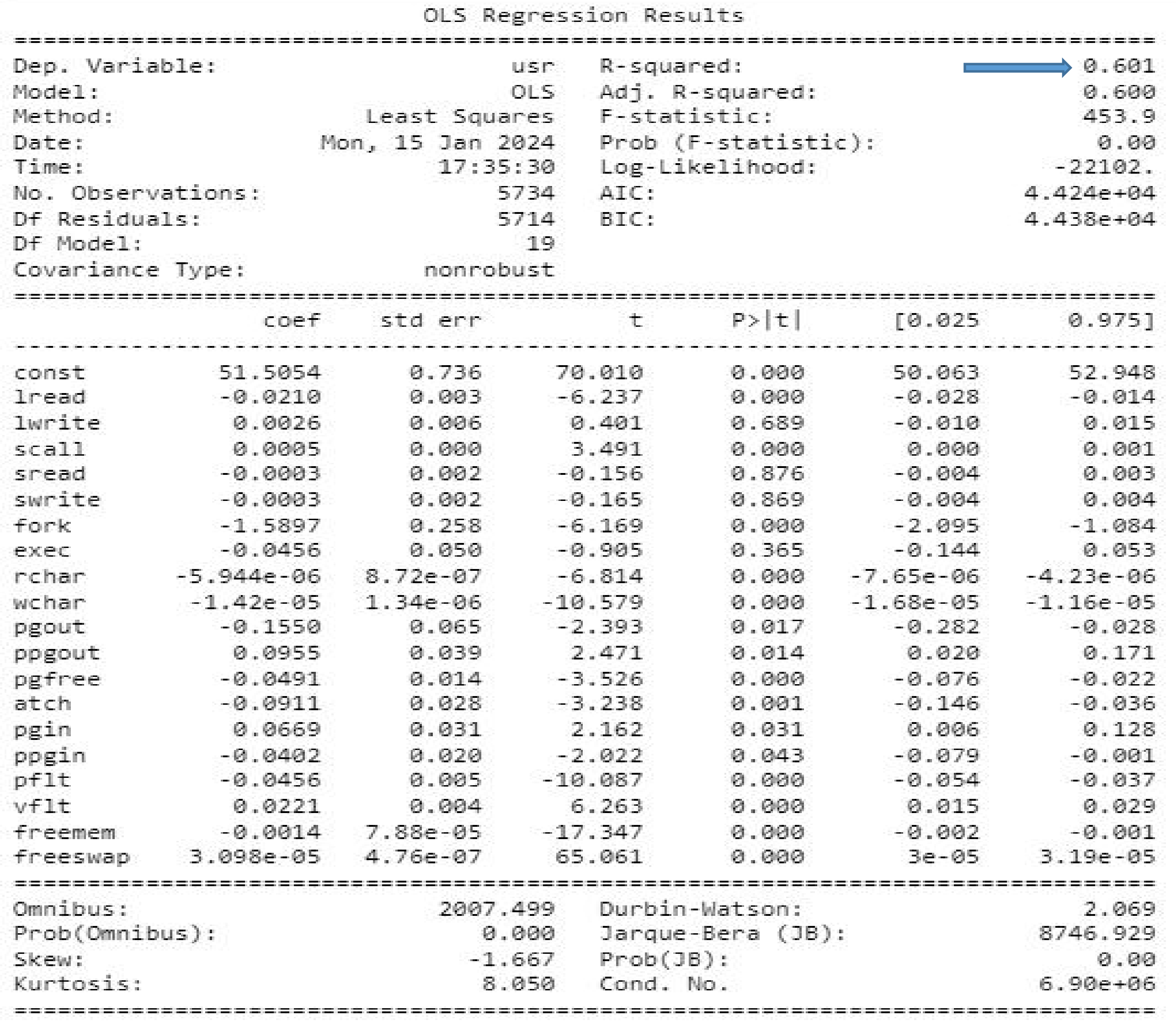


```
      const  lread  lwrite  scall  sread  swrite  fork  exec     rchar  \
3894    1.0     27      39   1252     53     118   0.2   0.2   26592.0
4276    1.0      1       0    996     85      55   0.4   0.4   16667.0
3414    1.0      9       7   1530    247     135   0.4   0.4   14513.0
4165    1.0     32       4   3243    182     140   5.2   5.6  337517.0
7385    1.0     16       3   5017    259     249   2.8   1.4   73537.0

         wchar  pgout  ppgout  pgfree  atch  pgin  ppgin    pflt    vflt  \
3894   54394.0    0.0     0.0     0.0   0.0   0.4    0.6   19.44   20.04
4276   36431.0    0.0     0.0     0.0   0.0   1.0    1.4   35.53   52.10
3414   61905.0   13.8    19.2    30.4  10.4  14.8   18.4   26.80  186.20
4165   94832.0    0.8     1.0     1.0   1.4   4.6    7.0  250.60  420.20
7385  237547.0    0.0     0.0     0.0   0.0   5.6    5.8  142.80  276.20

      freemem  freeswap
3894     7762   1875466
4276     2979   1010114
3414       89        11
4165     1300   1535309
7385     2114    988600
```

1.3 Model Building - Linear Regression

Apply linear Regression using Sklearn - Using Statsmodels Perform checks for significant variables using the appropriate method - Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare.

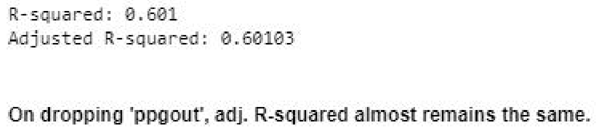
a) Standard errors assume that the convenience metrics of the errors is correctly specified.
b) The condition number is large, 6.9 e +06. This might indicate that there are strong multicollinearity or other numerical problems.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.601
Model:                            OLS   Adj. R-squared:                  0.600
Method:                 Least Squares   F-statistic:                     453.9
Date:                Mon, 15 Jan 2024   Prob (F-statistic):               0.00
Time:                        17:35:30   Log-Likelihood:                -22102.
No. Observations:                5734   AIC:                         4.424e+04
Df Residuals:                    5714   BIC:                         4.438e+04
Df Model:                          19
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         51.5054      0.736     70.010      0.000      50.063      52.948
lread         -0.0210      0.003     -6.237      0.000      -0.028      -0.014
lwrite         0.0026      0.006      0.401      0.689      -0.010       0.015
scall          0.0005      0.000      3.491      0.000       0.000       0.001
sread         -0.0003      0.002     -0.156      0.876      -0.004       0.003
swrite        -0.0003      0.002     -0.165      0.869      -0.004       0.004
fork          -1.5897      0.258     -6.169      0.000      -2.095      -1.084
exec          -0.0456      0.050     -0.905      0.365      -0.144       0.053
rchar      -5.944e-06   8.72e-07     -6.814      0.000   -7.65e-06   -4.23e-06
wchar       -1.42e-05   1.34e-06    -10.579      0.000   -1.68e-05   -1.16e-05
pgout         -0.1550      0.065     -2.393      0.017      -0.282      -0.028
ppgout         0.0955      0.039      2.471      0.014       0.020       0.171
pgfree        -0.0491      0.014     -3.526      0.000      -0.076      -0.022
atch          -0.0911      0.028     -3.238      0.001      -0.146      -0.036
pgin           0.0669      0.031      2.162      0.031       0.006       0.128
ppgin         -0.0402      0.020     -2.022      0.043      -0.079      -0.001
pflt          -0.0456      0.005    -10.087      0.000      -0.054      -0.037
vflt           0.0221      0.004      6.263      0.000       0.015       0.029
freemem       -0.0014   7.88e-05    -17.347      0.000      -0.002      -0.001
freeswap    3.098e-05   4.76e-07     65.061      0.000       3e-05    3.19e-05
==============================================================================
Omnibus:                     2007.499   Durbin-Watson:                   2.069
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             8746.929
Skew:                          -1.667   Prob(JB):                         0.00
Kurtosis:                       8.050   Cond. No.                     6.90e+06
==============================================================================
```

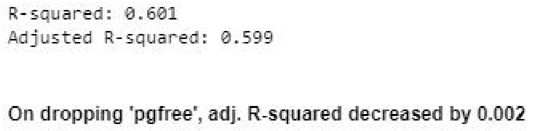
- Interpretation of R-squared
- R-squared value can shows 60.1% of the variance in the training set.

By dropping multicollinear columns one by one, we observe that some almost remain same And there is quite only 0 .001 and 0.002 Downwards difference.



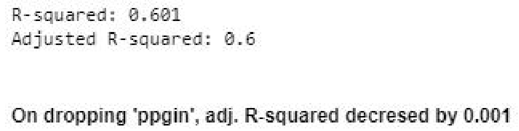
```
R-squared: 0.601
Adjusted R-squared: 0.60103
```

On dropping 'ppgout', adj. R-squared almost remains the same.



```
R-squared: 0.601
Adjusted R-squared: 0.599
```

On dropping 'pgfree', adj. R-squared decreased by 0.002

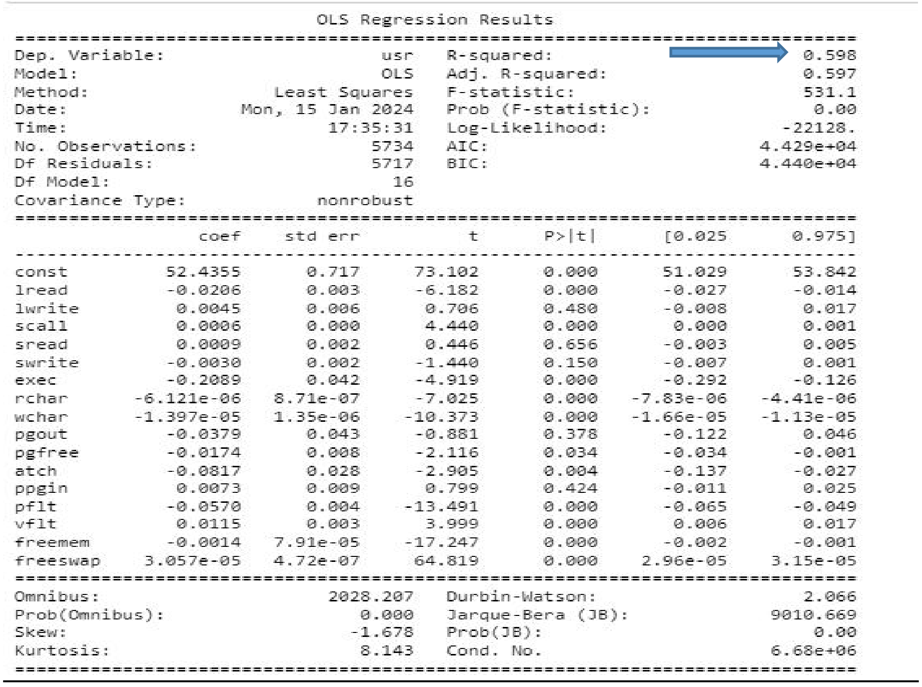


```
R-squared: 0.601
Adjusted R-squared: 0.6
```

On dropping 'ppgin', adj. R-squared decresed by 0.001

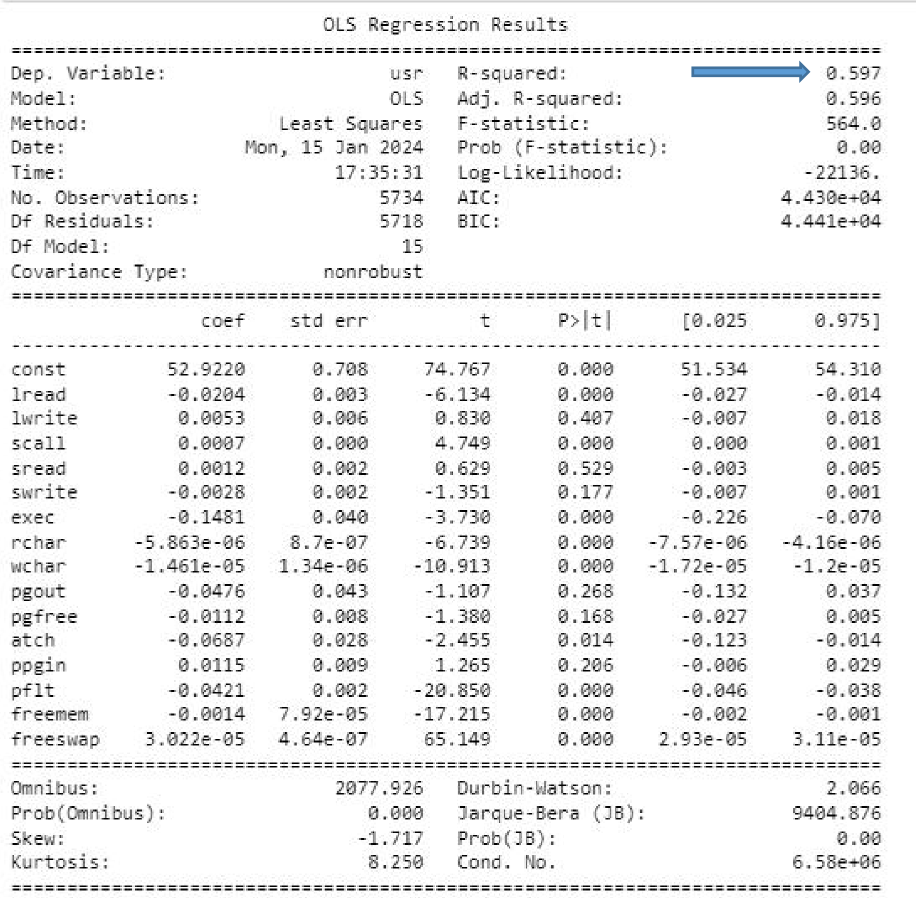
<u>**SO ON…..**</u>

- **There is no effect on adj. R-squared after dropping the 'ppgout' column, and it has highest number in value of variance influence factor, so we remove it from the training set.**
- **Since there is ALSO no effect on adj. R-squared after dropping the 'pgin' column, and it has highest number in value of variance influence factor, so we remove it from the training set.**



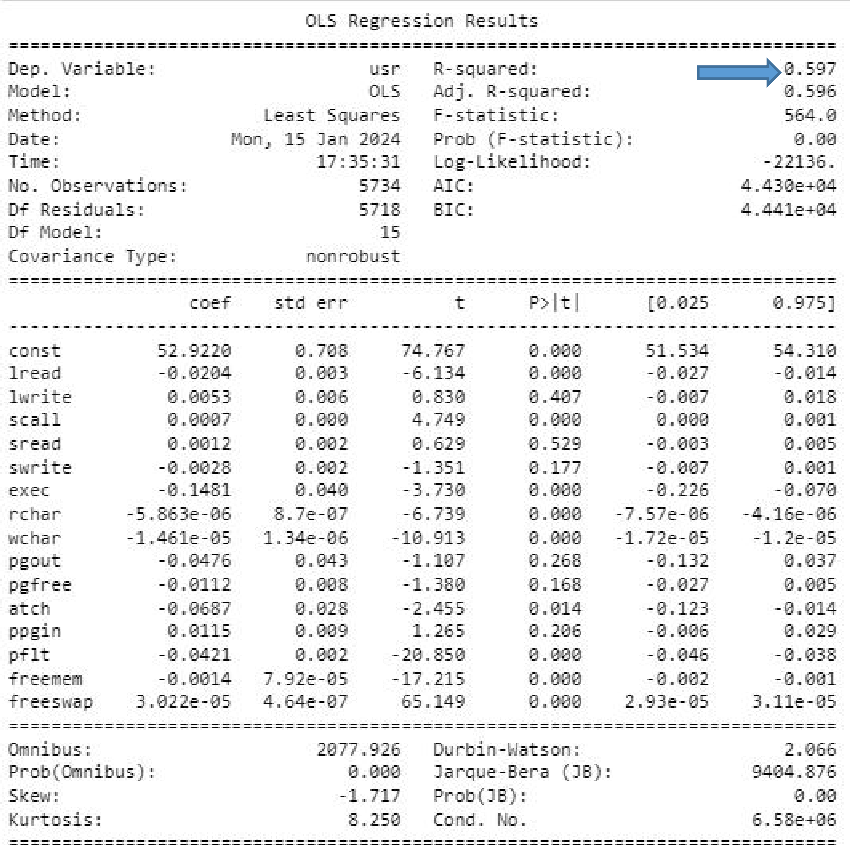
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.598
Model:                            OLS   Adj. R-squared:                  0.597
Method:                 Least Squares   F-statistic:                     531.1
Date:                Mon, 15 Jan 2024   Prob (F-statistic):               0.00
Time:                        17:35:31   Log-Likelihood:                -22128.
No. Observations:                5734   AIC:                         4.429e+04
Df Residuals:                    5717   BIC:                         4.440e+04
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         52.4355      0.717     73.102      0.000      51.029      53.842
lread         -0.0206      0.003     -6.182      0.000      -0.027      -0.014
lwrite         0.0045      0.006      0.706      0.480      -0.008       0.017
scall          0.0006      0.000      4.440      0.000       0.000       0.001
sread          0.0009      0.002      0.446      0.656      -0.003       0.005
swrite        -0.0030      0.002     -1.440      0.150      -0.007       0.001
exec          -0.2089      0.042     -4.919      0.000      -0.292      -0.126
rchar      -6.121e-06   8.71e-07     -7.025      0.000   -7.83e-06   -4.41e-06
wchar      -1.397e-05   1.35e-06    -10.373      0.000   -1.66e-05   -1.13e-05
pgout         -0.0379      0.043     -0.881      0.378      -0.122       0.046
pgfree        -0.0174      0.008     -2.116      0.034      -0.034      -0.001
atch          -0.0817      0.028     -2.905      0.004      -0.137      -0.027
ppgin          0.0073      0.009      0.799      0.424      -0.011       0.025
pflt          -0.0570      0.004    -13.491      0.000      -0.065      -0.049
vflt           0.0115      0.003      3.999      0.000       0.006       0.017
freemem       -0.0014   7.91e-05    -17.247      0.000      -0.002      -0.001
freeswap    3.057e-05   4.72e-07     64.819      0.000    2.96e-05    3.15e-05
==============================================================================
Omnibus:                     2028.207   Durbin-Watson:                   2.066
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9010.669
Skew:                          -1.678   Prob(JB):                         0.00
Kurtosis:                       8.143   Cond. No.                     6.68e+06
==============================================================================
```

- **As we see, There is little bit effect on adj. R-squared after dropping the 'fork' column.**



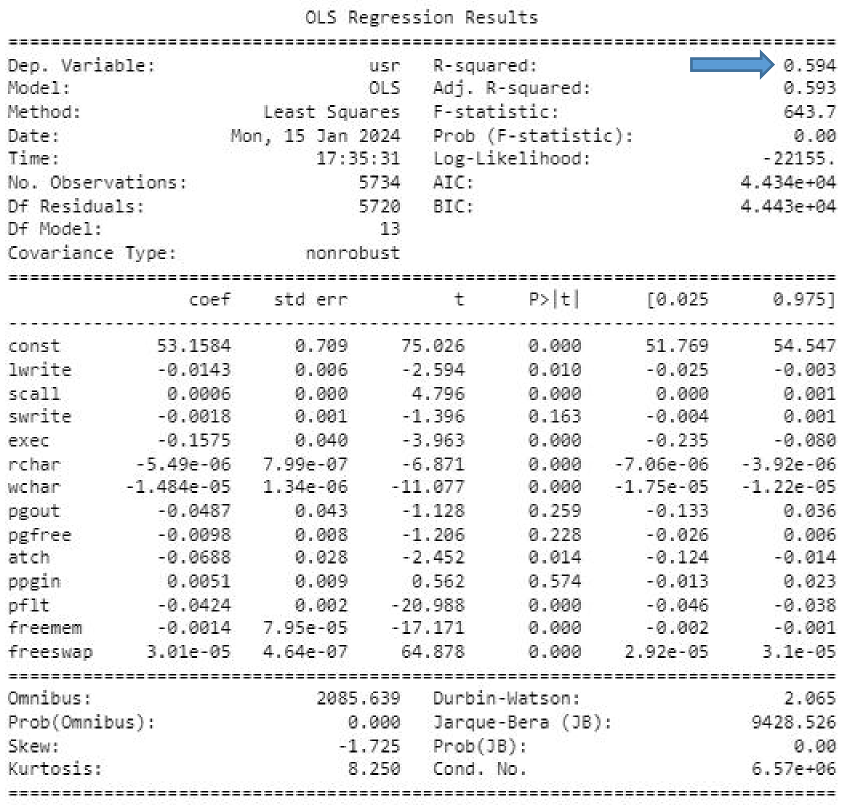
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.597
Model:                            OLS   Adj. R-squared:                  0.596
Method:                 Least Squares   F-statistic:                     564.0
Date:                Mon, 15 Jan 2024   Prob (F-statistic):               0.00
Time:                        17:35:31   Log-Likelihood:                -22136.
No. Observations:                5734   AIC:                         4.430e+04
Df Residuals:                    5718   BIC:                         4.441e+04
Df Model:                          15
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         52.9220      0.708     74.767      0.000      51.534      54.310
lread         -0.0204      0.003     -6.134      0.000      -0.027      -0.014
lwrite         0.0053      0.006      0.830      0.407      -0.007       0.018
scall          0.0007      0.000      4.749      0.000       0.000       0.001
sread          0.0012      0.002      0.629      0.529      -0.003       0.005
swrite        -0.0028      0.002     -1.351      0.177      -0.007       0.001
exec          -0.1481      0.040     -3.730      0.000      -0.226      -0.070
rchar      -5.863e-06    8.7e-07     -6.739      0.000   -7.57e-06   -4.16e-06
wchar      -1.461e-05   1.34e-06    -10.913      0.000   -1.72e-05    -1.2e-05
pgout         -0.0476      0.043     -1.107      0.268      -0.132       0.037
pgfree        -0.0112      0.008     -1.380      0.168      -0.027       0.005
atch          -0.0687      0.028     -2.455      0.014      -0.123      -0.014
ppgin          0.0115      0.009      1.265      0.206      -0.006       0.029
pflt          -0.0421      0.002    -20.850      0.000      -0.046      -0.038
freemem       -0.0014   7.92e-05    -17.215      0.000      -0.002      -0.001
freeswap    3.022e-05   4.64e-07     65.149      0.000    2.93e-05    3.11e-05
==============================================================================
Omnibus:                     2077.926   Durbin-Watson:                   2.066
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9404.876
Skew:                          -1.717   Prob(JB):                         0.00
Kurtosis:                       8.250   Cond. No.                     6.58e+06
==============================================================================
```

- **As we see, There is also little bit effect on adj. R-squared after dropping the 'vflt' column.**



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.597
Model:                            OLS   Adj. R-squared:                  0.596
Method:                 Least Squares   F-statistic:                     564.0
Date:                Mon, 15 Jan 2024   Prob (F-statistic):               0.00
Time:                        17:35:31   Log-Likelihood:                -22136.
No. Observations:                5734   AIC:                         4.430e+04
Df Residuals:                    5718   BIC:                         4.441e+04
Df Model:                          15
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         52.9220      0.708     74.767      0.000      51.534      54.310
lread         -0.0204      0.003     -6.134      0.000      -0.027      -0.014
lwrite         0.0053      0.006      0.830      0.407      -0.007       0.018
scall          0.0007      0.000      4.749      0.000       0.000       0.001
sread          0.0012      0.002      0.629      0.529      -0.003       0.005
swrite        -0.0028      0.002     -1.351      0.177      -0.007       0.001
exec          -0.1481      0.040     -3.730      0.000      -0.226      -0.070
rchar      -5.863e-06    8.7e-07     -6.739      0.000   -7.57e-06   -4.16e-06
wchar      -1.461e-05   1.34e-06    -10.913      0.000   -1.72e-05    -1.2e-05
pgout         -0.0476      0.043     -1.107      0.268      -0.132       0.037
pgfree        -0.0112      0.008     -1.380      0.168      -0.027       0.005
atch          -0.0687      0.028     -2.455      0.014      -0.123      -0.014
ppgin          0.0115      0.009      1.265      0.206      -0.006       0.029
pflt          -0.0421      0.002    -20.850      0.000      -0.046      -0.038
freemem       -0.0014   7.92e-05    -17.215      0.000      -0.002      -0.001
freeswap    3.022e-05   4.64e-07     65.149      0.000    2.93e-05    3.11e-05
==============================================================================
Omnibus:                     2077.926   Durbin-Watson:                   2.066
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9404.876
Skew:                          -1.717   Prob(JB):                         0.00
Kurtosis:                       8.250   Cond. No.                     6.58e+06
==============================================================================
```

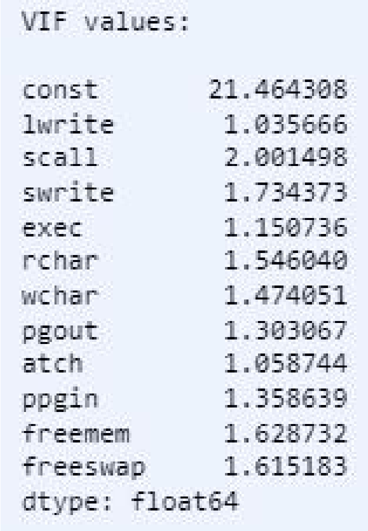
- **<u>There is no effect on adj. R-squared after dropping the 'sread', ‘lread’,’pgfree’ column</u>**



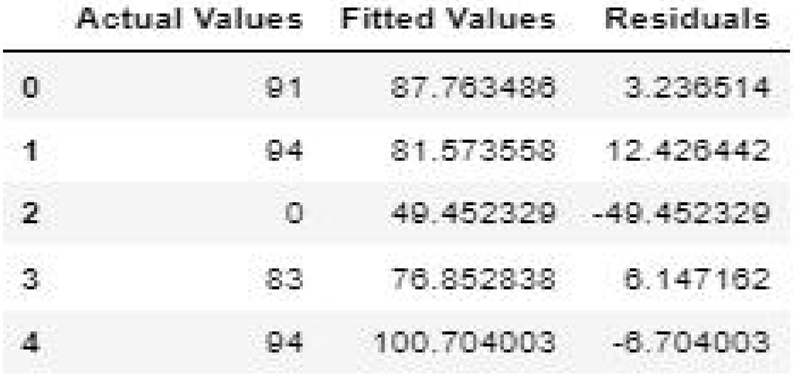
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.594
Model:                            OLS   Adj. R-squared:                  0.593
Method:                 Least Squares   F-statistic:                     643.7
Date:                Mon, 15 Jan 2024   Prob (F-statistic):               0.00
Time:                        17:35:31   Log-Likelihood:                -22155.
No. Observations:                5734   AIC:                         4.434e+04
Df Residuals:                    5720   BIC:                         4.443e+04
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         53.1584      0.709     75.026      0.000      51.769      54.547
lwrite        -0.0143      0.006     -2.594      0.010      -0.025      -0.003
scall          0.0006      0.000      4.796      0.000       0.000       0.001
swrite        -0.0018      0.001     -1.396      0.163      -0.004       0.001
exec          -0.1575      0.040     -3.963      0.000      -0.235      -0.080
rchar       -5.49e-06   7.99e-07     -6.871      0.000   -7.06e-06   -3.92e-06
wchar      -1.484e-05   1.34e-06    -11.077      0.000   -1.75e-05   -1.22e-05
pgout         -0.0487      0.043     -1.128      0.259      -0.133       0.036
pgfree        -0.0098      0.008     -1.206      0.228      -0.026       0.006
atch          -0.0688      0.028     -2.452      0.014      -0.124      -0.014
ppgin          0.0051      0.009      0.562      0.574      -0.013       0.023
pflt          -0.0424      0.002    -20.988      0.000      -0.046      -0.038
freemem       -0.0014   7.95e-05    -17.171      0.000      -0.002      -0.001
freeswap     3.01e-05   4.64e-07     64.878      0.000    2.92e-05     3.1e-05
==============================================================================
Omnibus:                     2085.639   Durbin-Watson:                   2.065
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9428.526
Skew:                          -1.725   Prob(JB):                         0.00
Kurtosis:                       8.250   Cond. No.                     6.57e+06
==============================================================================
```

- **<u>As we see, There is little bit effect on adj. R-squared after dropping the 'pflt' column.</u>**

<u>After dropping the features causing strong multicollinearity and the statistically insignificant ones, our model performance hasn't dropped sharply. This shows that these variables did not have much predictive power.</u>

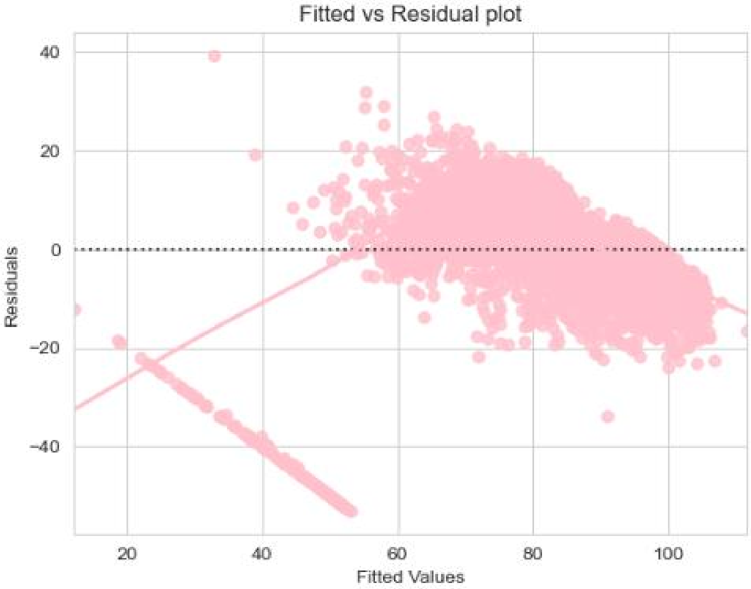


```
VIF values:

const       21.464308
lwrite       1.035666
scall        2.001498
swrite       1.734373
exec         1.150736
rchar        1.546040
wchar        1.474051
pgout        1.303067
atch         1.058744
ppgin        1.358639
freemem      1.628732
freeswap     1.615183
dtype: float64
```

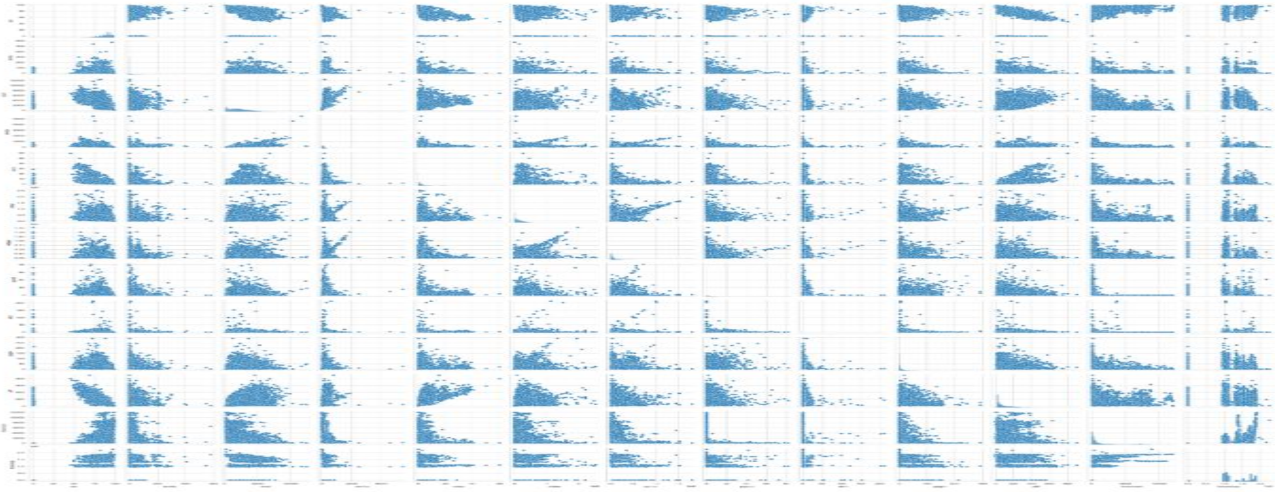


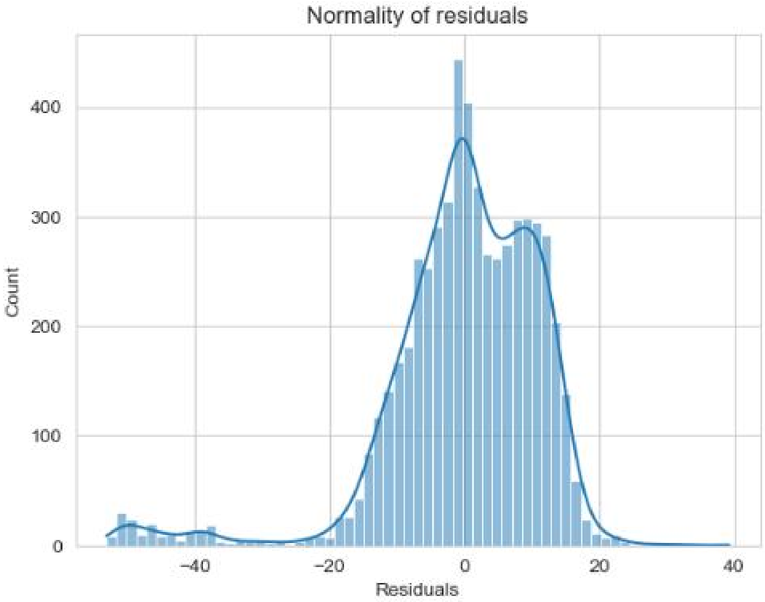
	Actual Values	Fitted Values	Residuals
0	91	87.763486	3.236514
1	94	81.573558	12.426442
2	0	49.452329	-49.452329
3	83	76.852838	6.147162
4	94	100.704003	-6.704003

- **VIF for all features is <3**
- VIF method can be used for identifying important variables & eliminating/removing the ones that may not significant and have high multicollinearity.

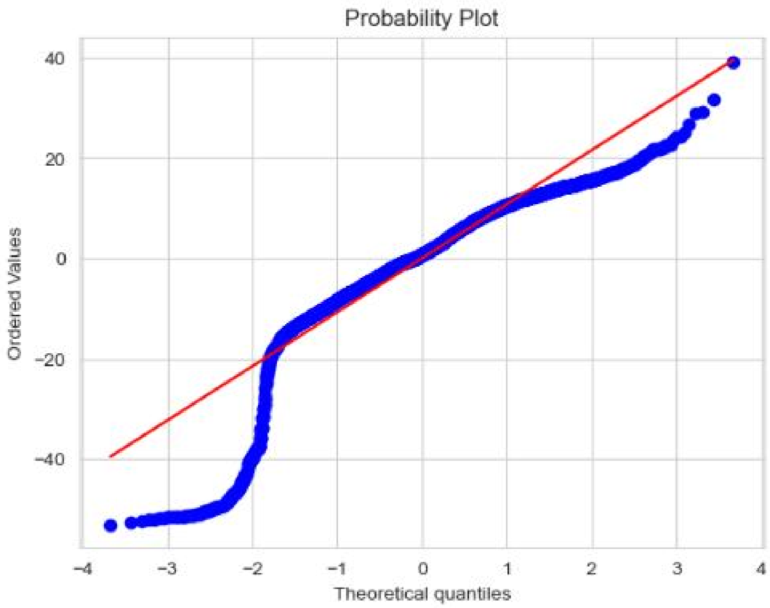


- **We observe that the pattern has slightly decreased and that Data points seems to be randomly distributed.**





- **The QQ plot of residuals can be used to visually check the normally assumptions.**
- **The normally probability plot of residual should approximately follow a straight line.**



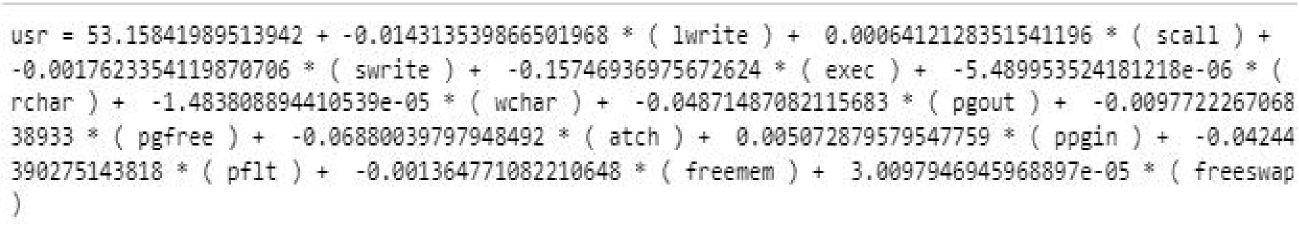
- **Partially, the points are laying on the straight line in QQ plot.**

```
ShapiroResult(statistic=0.870936930179596, pvalue=0.0)
```

If p-value is < 0.05, the residuals are rejected in shapiro test. but the tested value is greater than 0.05

1.4 Business Insights & Recommendations

Comment on the Linear Regression equation from the final model and impact of relevant variables (atleast 2) as per the equation - Conclude with the key takeaways (actionable insights and recommendations) for the business



```
usr = 53.15841989513942 + -0.014313539866501968 * ( lwrite ) +  0.0006412128351541196 * ( scall ) +
-0.0017623354119870706 * ( swrite ) +  -0.15746936975672624 * ( exec ) +  -5.489953524181218e-06 * (
rchar ) +  -1.483808894410539e-05 * ( wchar ) +  -0.04871487082115683 * ( pgout ) +  -0.0097722267068
38933 * ( pgfree ) +  -0.06880039797948492 * ( atch ) +  0.005072879579547759 * ( ppgin ) +  -0.04244
390275143818 * ( pflt ) +  -0.001364771082210648 * ( freemem ) +  3.0097946945968897e-05 * ( freeswap
)
```

- RMSE on the train data = 11.5289
- MAE on the train data = 8.1244

- RMSE on the train and test sets are comparable. So, our model may not suffer from over-fitting.
- MAE indicates that our current model able to predict mpg within a mean error of the test data.
- Therefore, we can assume the model "fitres-42" is good for prediction as well as inference purposes.

Key Influence of Process Run Queue Size:
The CPU run-time in user mode shows a significant dependency on the Process run queue size. Understanding and managing the size of the queue for running processes are crucial for optimizing CPU performance.

Sensitivity to CPU Bound Queue Size:
A noteworthy finding is that increasing the CPU bound queue size by just 1 unit leads to a substantial 33.5 times increase in the percentage of time the CPU runs in user mode. This suggests that proper management of CPU-bound tasks in the queue is vital for improving user mode run-time.

Impact of Non-CPU Bound Queue Size:
Similarly, the non-CPU bound queue size has a significant impact, with a 32.7 times increase in CPU run-time in user mode for every 1-unit increase. Balancing and optimizing I/O-bound tasks in the queue are important considerations for overall system performance.

Cumulative Effect of Process Run Queue Size:
When considering both CPU and non-CPU bound queues, the overall impact on the percentage of time the CPU runs in user mode is substantial, approximately 132 times, including the Intercept. This underscores the holistic influence of the process run queue size on CPU behavior.

Constant Impact of Other Features:
The analysis suggests that, while the process run queue size has a substantial impact, the other features considered in the model do not significantly affect CPU run-time. This could guide resource allocation efforts, focusing primarily on optimizing the process run queue size.

B Problem2: Contraceptive Method Data-set

In your role as a statistician at the Republic of Indonesia Ministry of Health, you have been entrusted 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.

NOW, we 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.

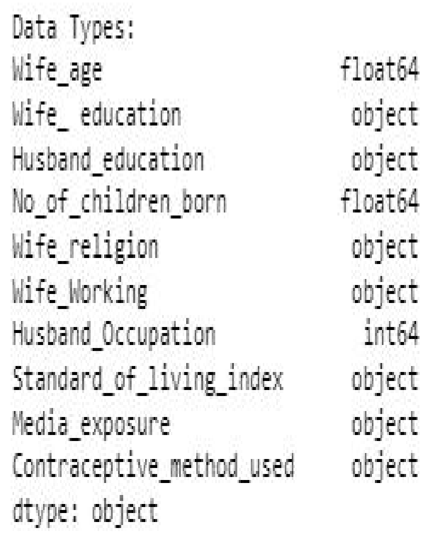
2.1 Define the problem and perform exploratory Data Analysis

Problem definition - Check shape, Data types, statistical summary - Univariate analysis - Multivariate analysis - Use appropriate visualizations to identify the patterns and insights - Key meaningful observations on individual variables and the relationship between variables.

Problem definition:-

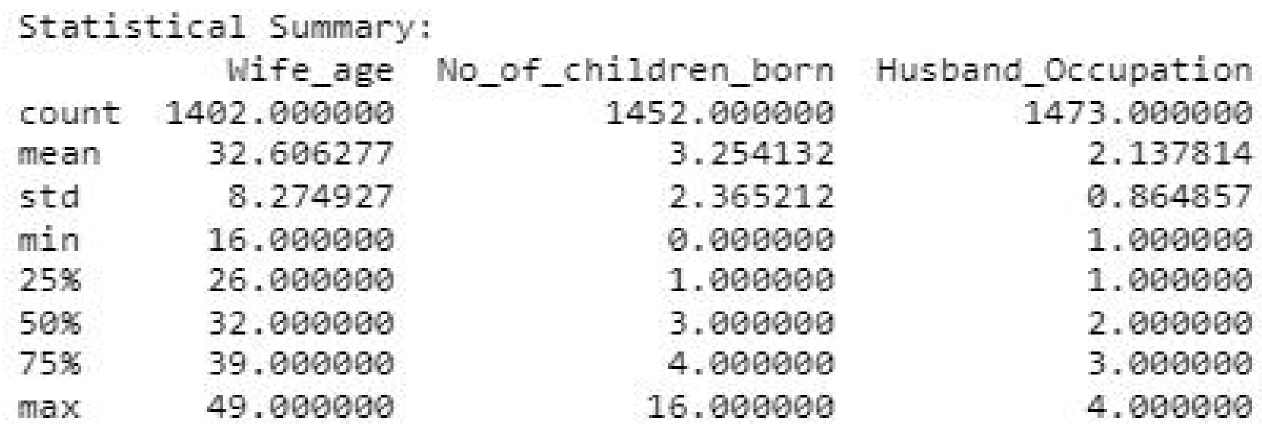
Check shape:- 1473 rows x 10columns

Data types:-



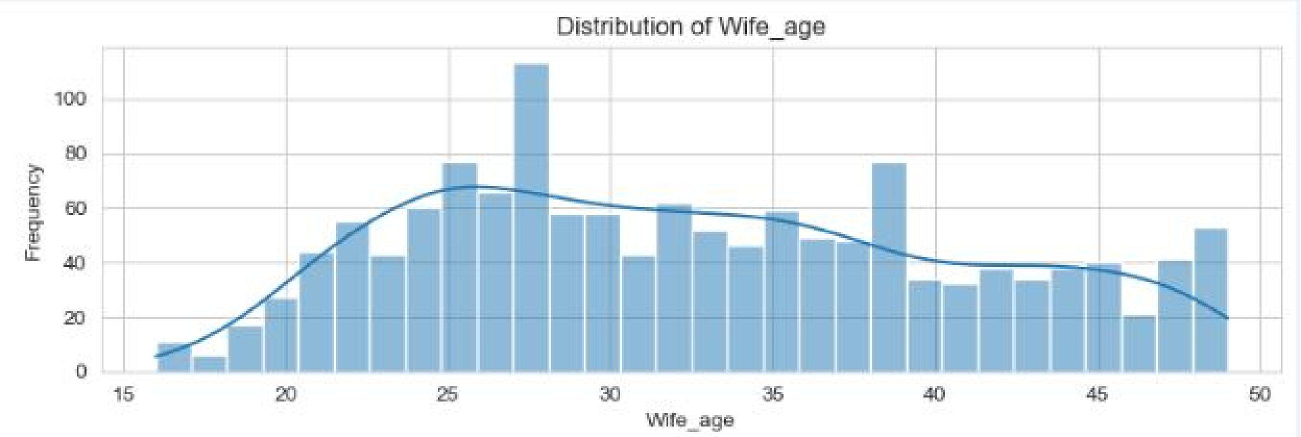
```
Data Types:
Wife_age                     float64
Wife_ education               object
Husband_education             object
No_of_children_born          float64
Wife_religion                 object
Wife_Working                  object
Husband_Occupation             int64
Standard_of_living_index      object
Media_exposure                object
Contraceptive_method_used     object
dtype: object
```

Statistical Summary:-

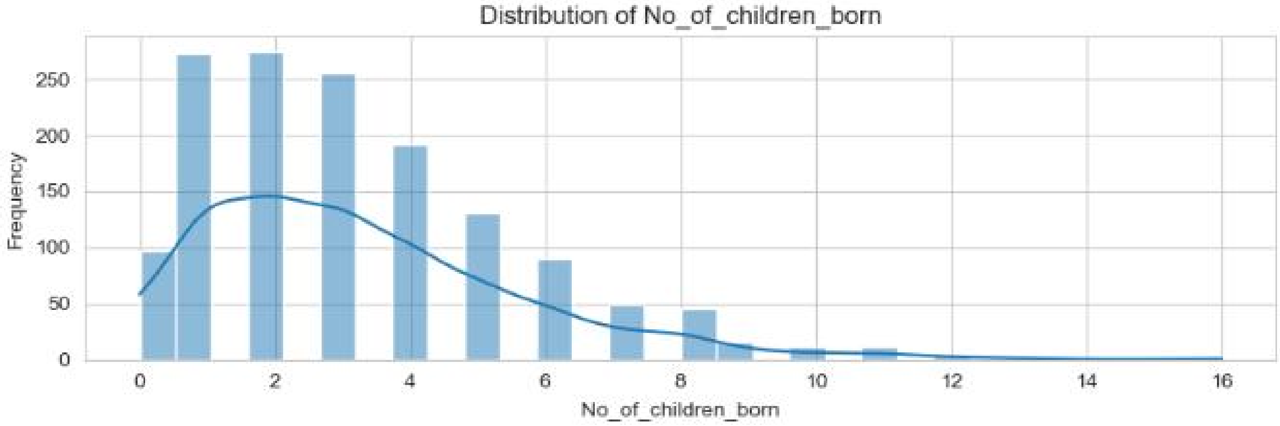


```
Statistical Summary:
          Wife_age  No_of_children_born  Husband_Occupation
count  1402.000000          1452.000000         1473.000000
mean     32.606277             3.254132            2.137814
std       8.274927             2.365212            0.864857
min      16.000000             0.000000            1.000000
25%      26.000000             1.000000            1.000000
50%      32.000000             3.000000            2.000000
75%      39.000000             4.000000            3.000000
max      49.000000            16.000000            4.000000
```

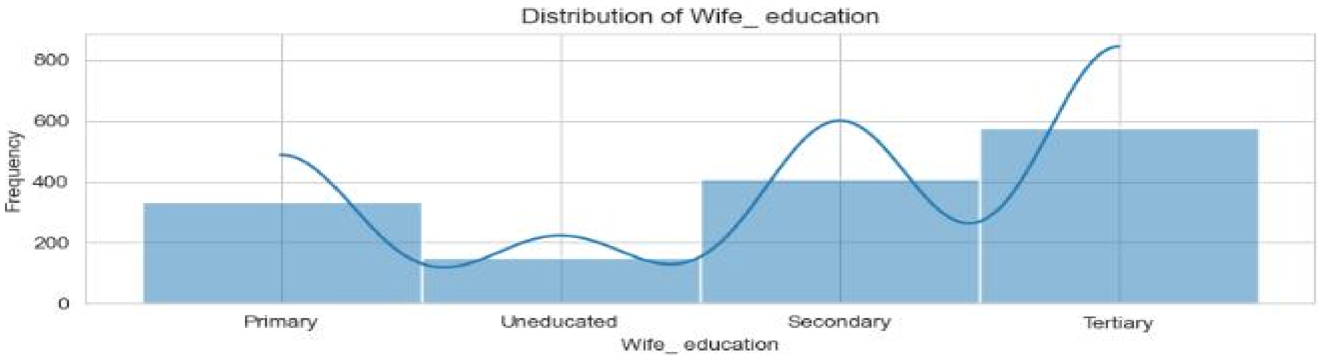
- **Uni-variate analysis:-**

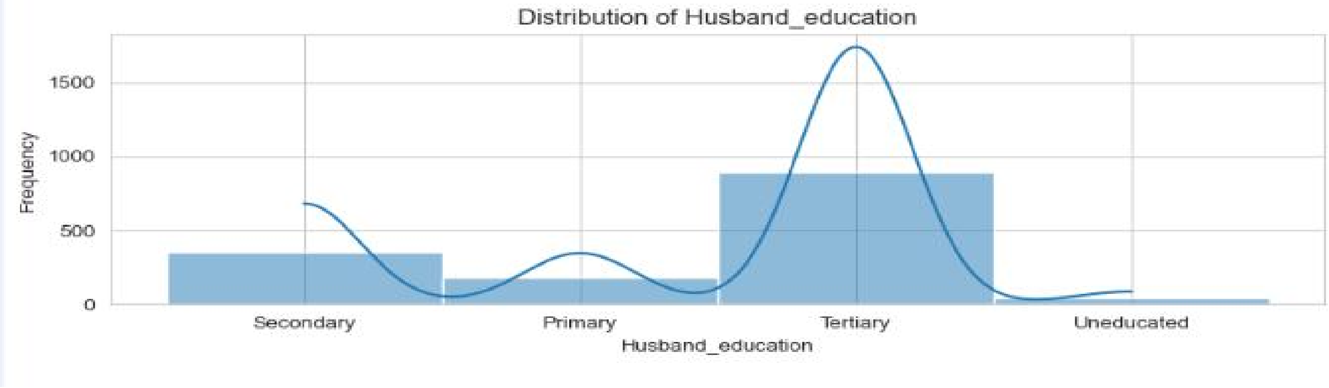


- The age of the wives B\W 17 - 49 years, where mostly they are in 28's and mid 20s - early 50s.

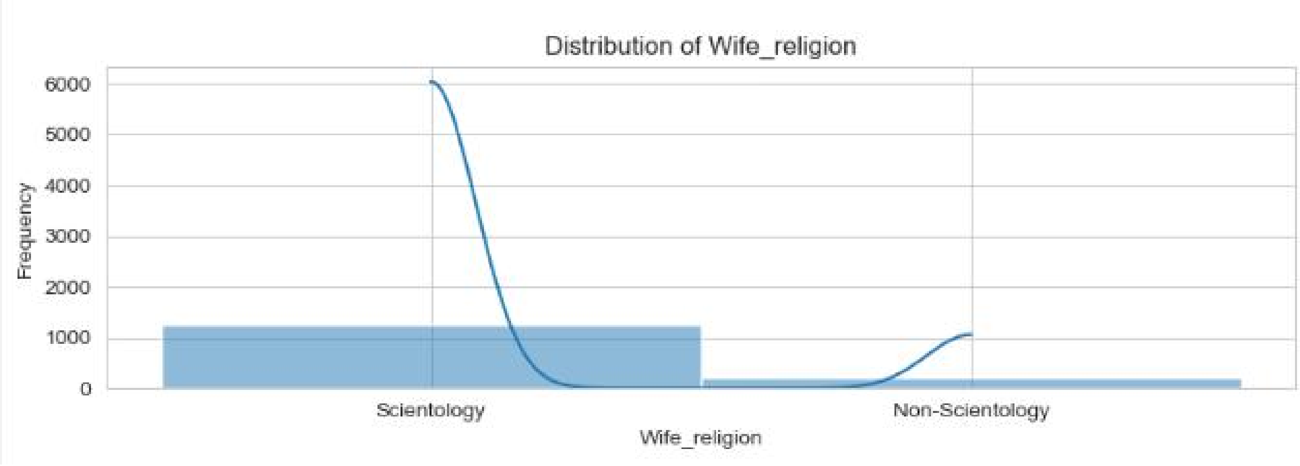


- Majority of the people have 1 or 2 children but a few people have more than 15 children as well.

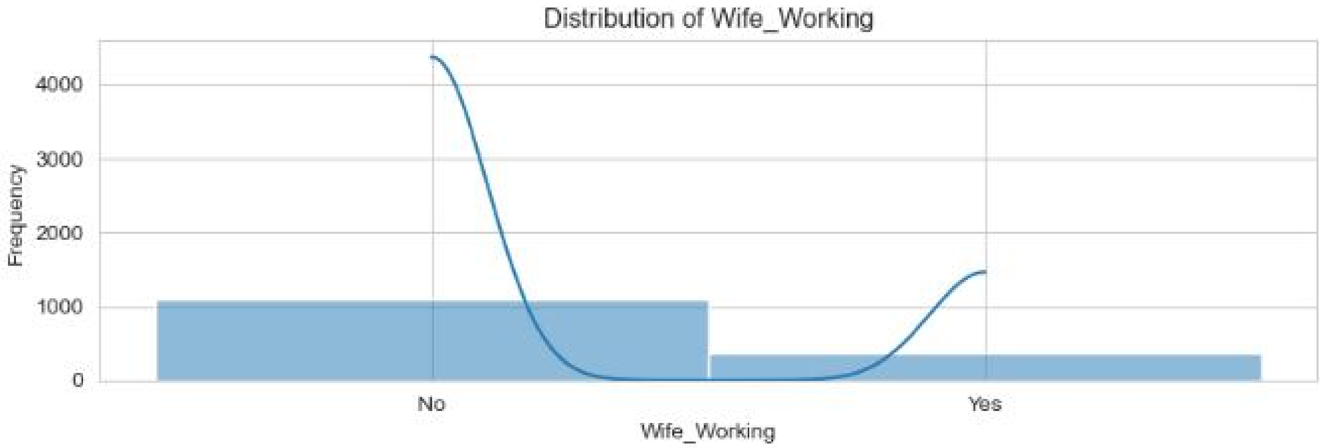




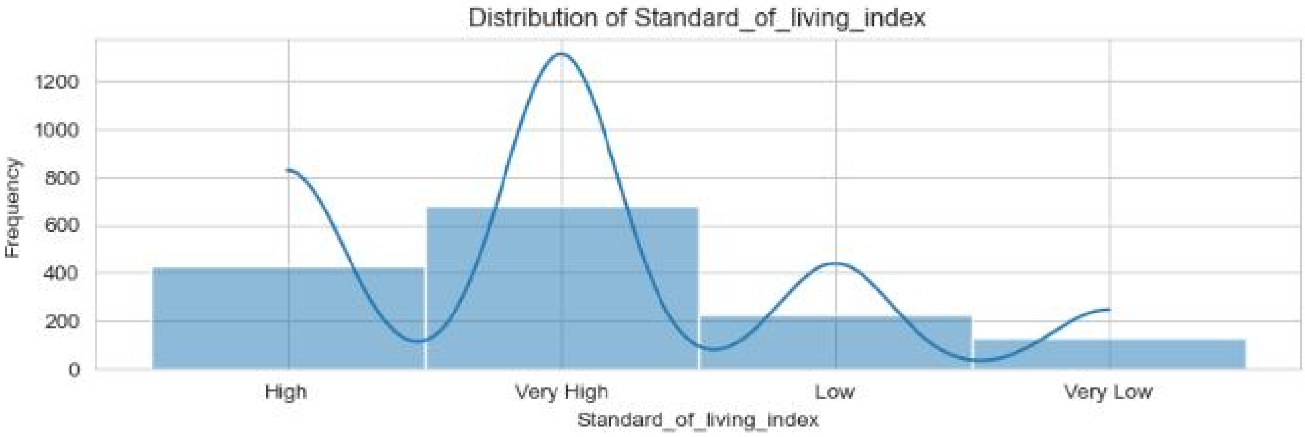
- Wives who have done their secondary and Tertiary education have used contraceptive methods more as compared to the others.
- Wives who are not educated or only completed Primary education are not to use any contraceptive methods.
- Commonly same thing find on the Husband’s education.
- Fewer Husbands are uneducated as compared to the wives.



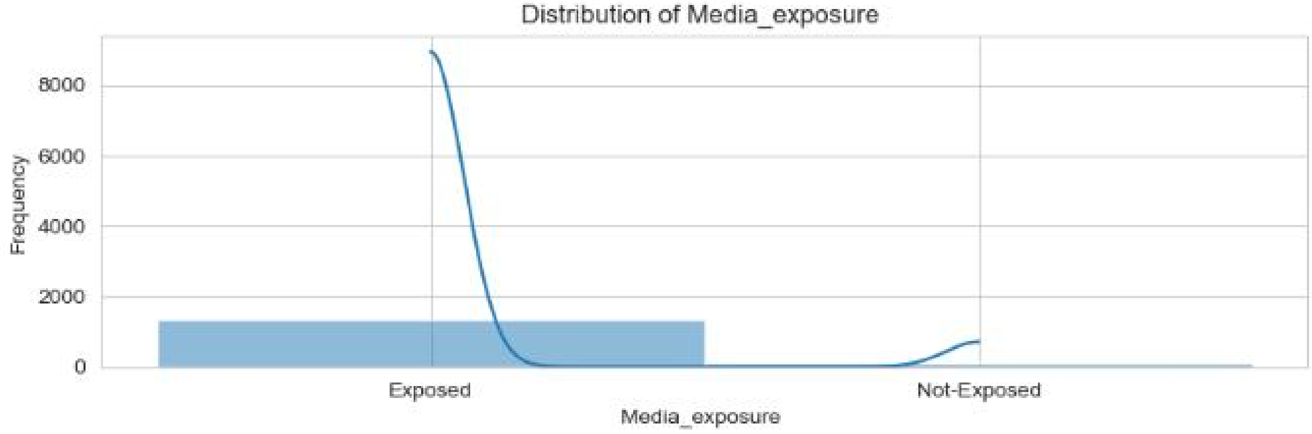
- Scientology is playing wider role in wife region.



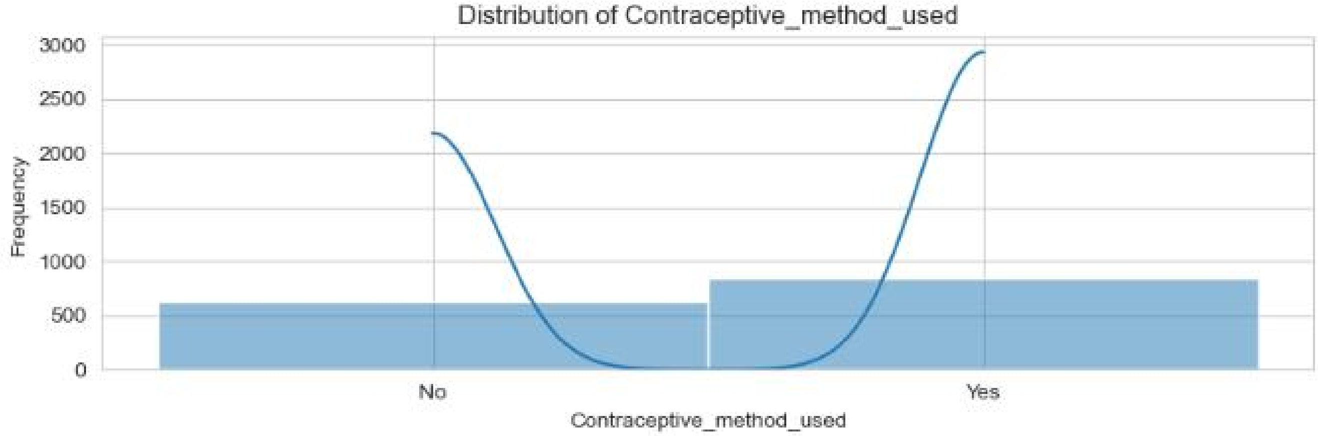
- Mostly Wives are not in working professional.



- Mostly people are belonging the areas where the standard of living is Very High and High.
- Nearly less than 250 people are belonging with Low and Very low standard of living index.

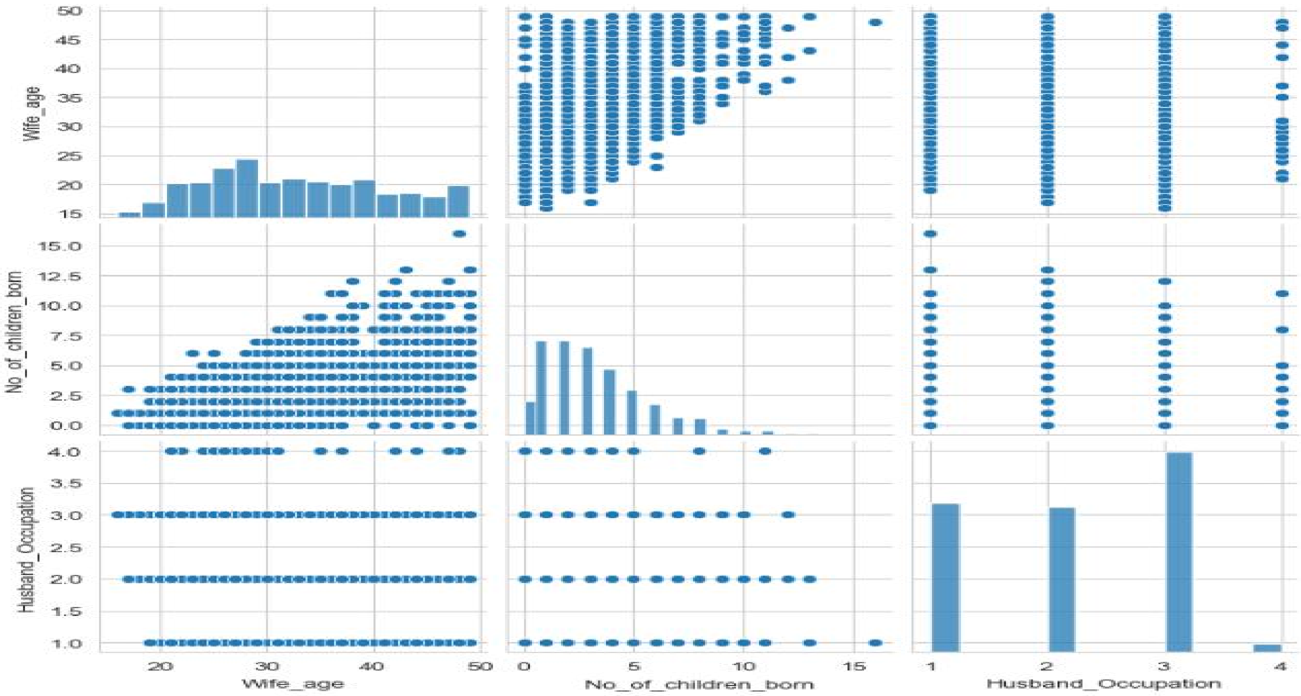


- Distribution of media exposure is quite better, its more than 1000.



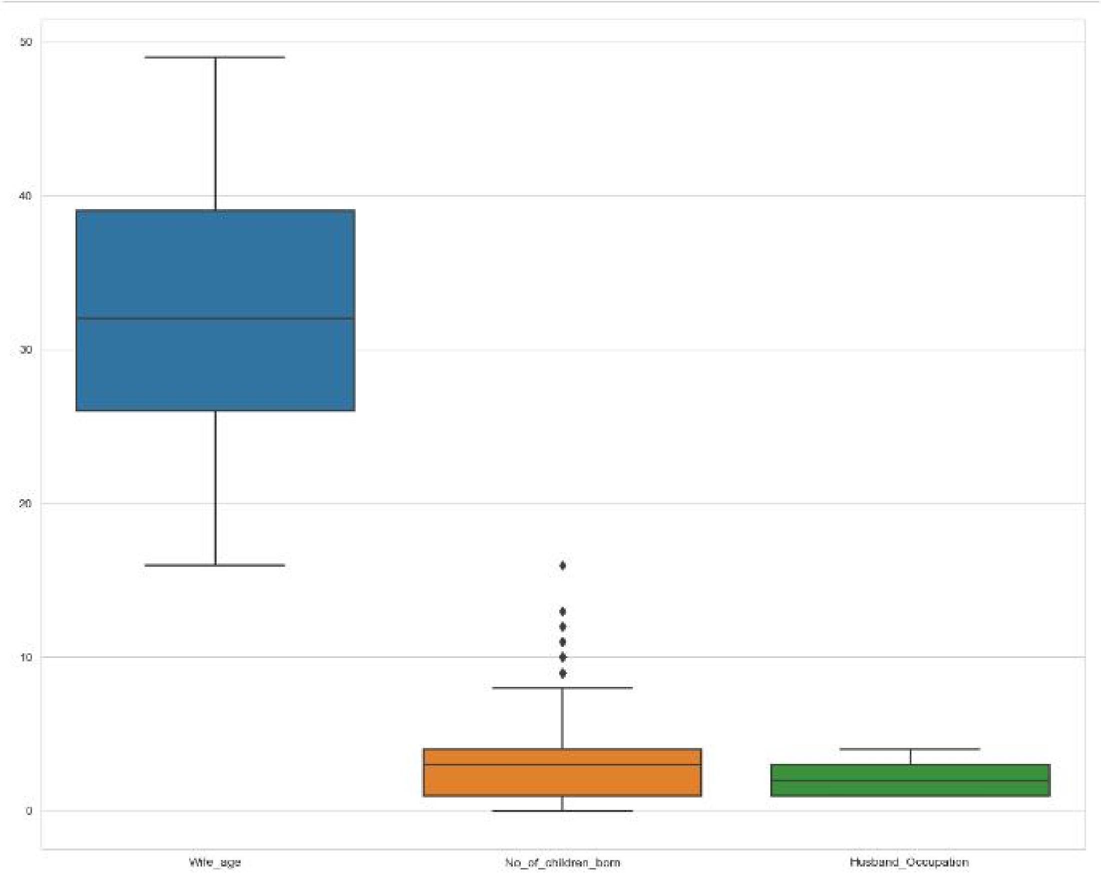
- As we already knew that, the mostly wives have used a contraceptive method, however there is a good proportional as well who have not used any.

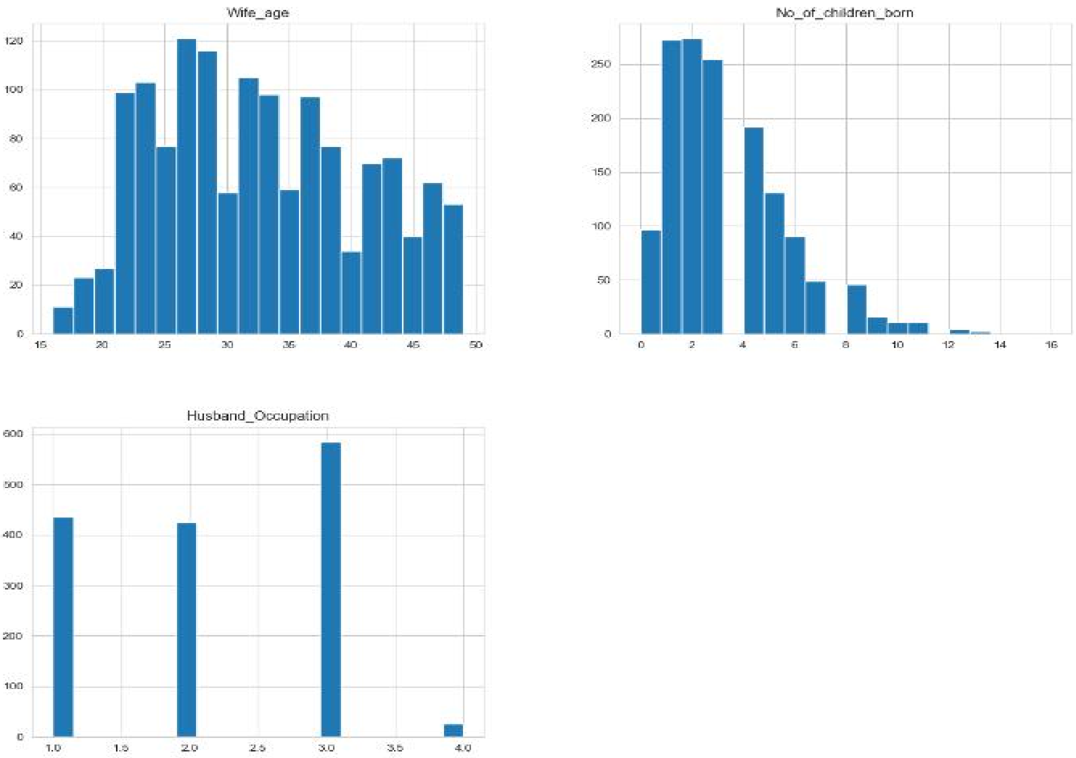
Multivariate analysis:-

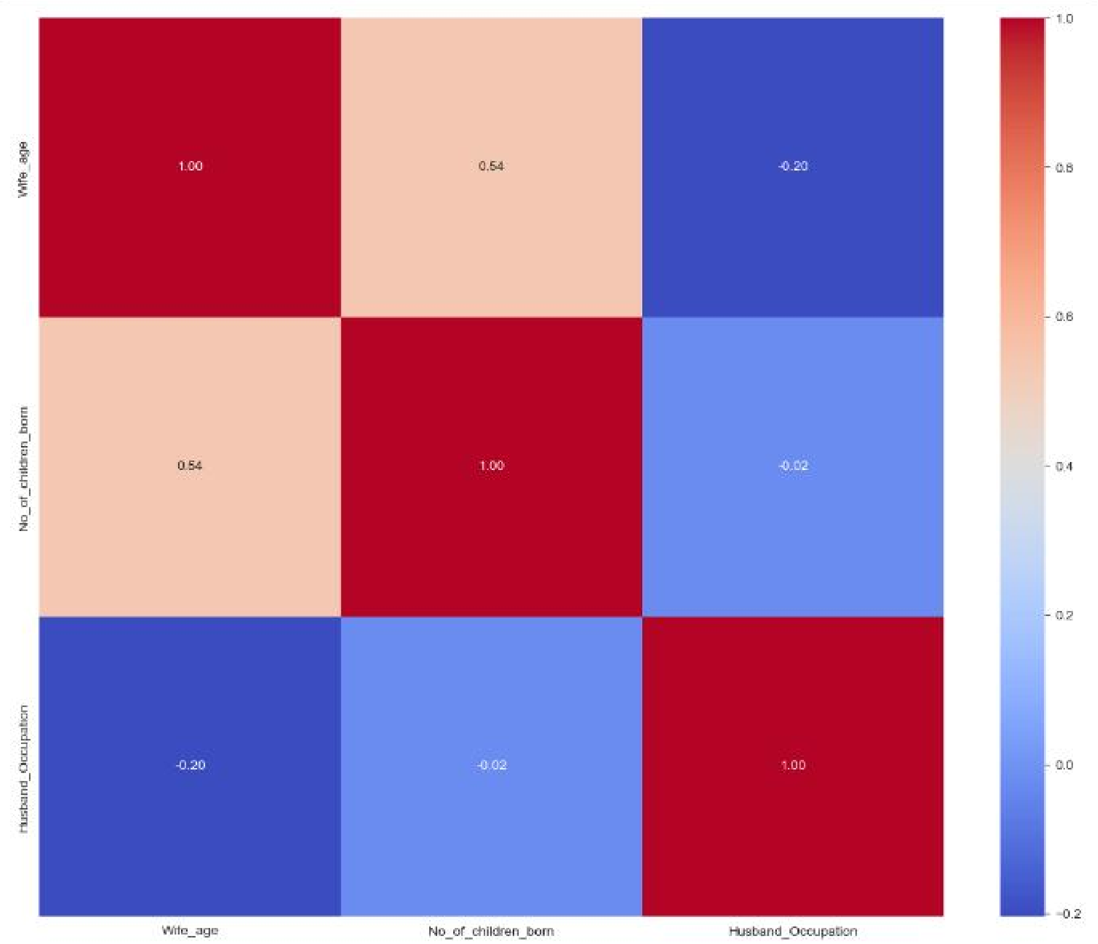


- This plot does not identify any major trend/correlation between the variables.
- Very Few of the variables are available in the pair-plot, they don't have the classes of well separated. They will not be a good predictors.

Use appropriate visualizations to identify the patterns and insights:-







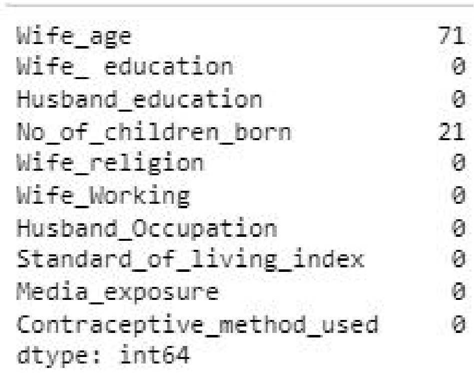
- Strong positive correlation shows b/w wife's age and husband's occupation.
- Strong negative correlation shows b/w number of children born and wife's age.
- Based on the above heat-map, it shows that couples where the wife was younger tended to have more children than couples where the wife was older. There are also a few with have much higher number of children born.

2.2 Data Pre-processing

Prepare the data for modelling: - Missing value Treatment (if needed) - Outlier Detection(treat, if needed) - Feature Engineering (if needed) - Encode the data - Train-test split

Prepare the data for modelling: -

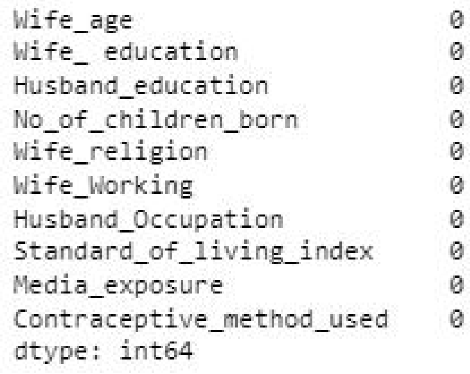
Missing value Treatment (if needed)



```
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
```

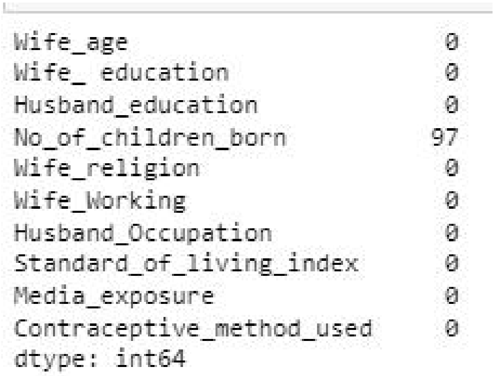
- There are 71 missing values are present in "wife_age" and 21 in "no_of_children_born". So now we treat the missing values.

AFTER TREATMENT:



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

Outlier Detection(treat, if needed)



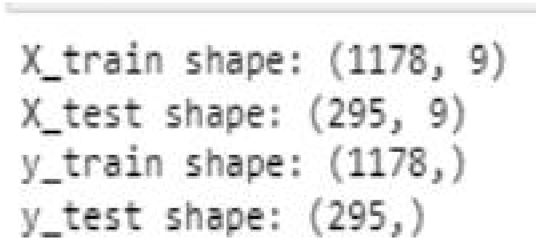
```
Wife_age                     0
Wife_ education              0
Husband_education            0
No_of_children_born         97
Wife_religion                0
Wife_Working                 0
Husband_Occupation           0
Standard_of_living_index     0
Media_exposure               0
Contraceptive_method_used    0
dtype: int64
```

- 97 Outliers are present. So, it has to treat the outliers.
- Now'Husband_Occupation' has been also changed to Object data type as it is a categorical variable.
- There are 85 duplicate which can be dropped from the dataset.

Encode the data - Train-test split

- Data has string & categorical variables, these variables must be encoded so that the Machine Learning model understands the data.
- In the targeted variable, "No" is switched to 0 and "Yes" is switched to 1.
- Likewise, other no.'s are given to the values in variables Wife_ education, Husband_education & Standard_of_living_index.
- After this, dummy encoding used to encode the data for the rest of the columns.
- After the encoded the data, the data-set has split-ted into training and testing in the 70:30 ratio.

Accuracy = 0.7152

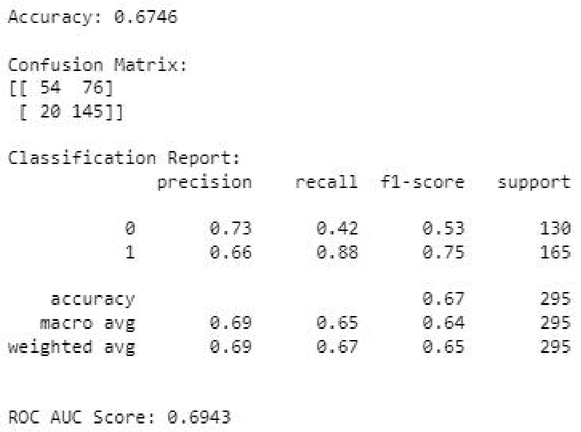


```
X_train shape: (1178, 9)
X_test shape: (295, 9)
y_train shape: (1178,)
y_test shape: (295,)
```

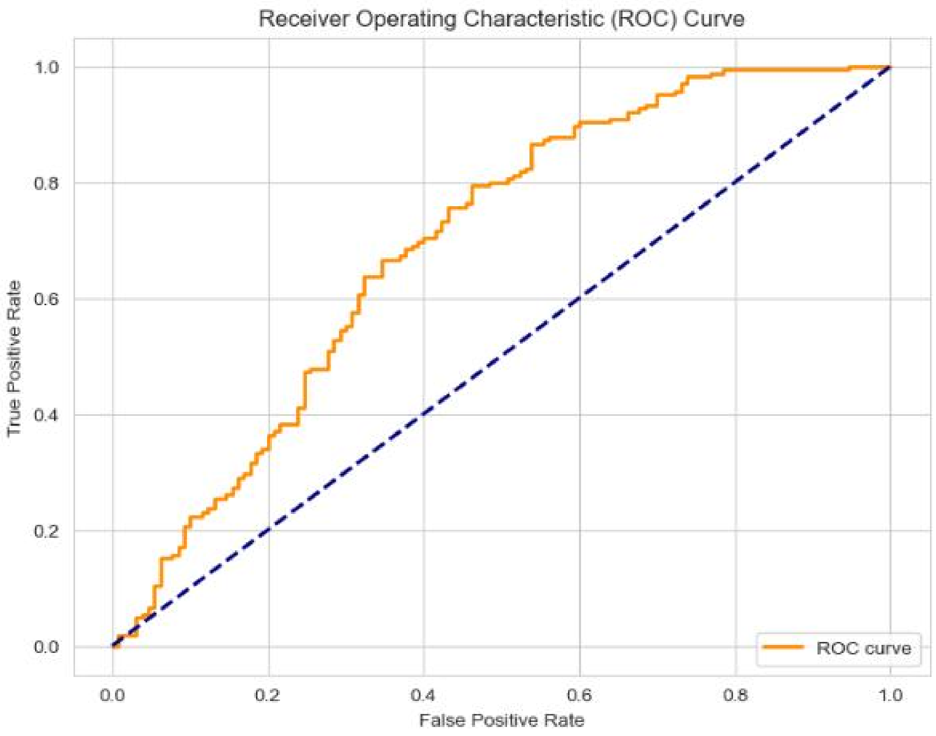
2.3 Model Building and Compare the Performance of the Models:-

Build a Logistic Regression model - Build a Linear Discriminant Analysis model - Build a CART model - Prune the CART model by finding the best hyper parameters using Grid Search - Check the performance of the models across train and test set using different metrics - Compare the performance of all the models built and choose the best one with proper rationale

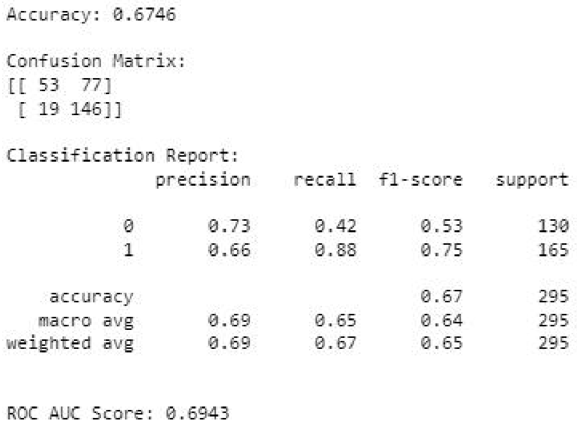
Build a Logistic Regression model:-



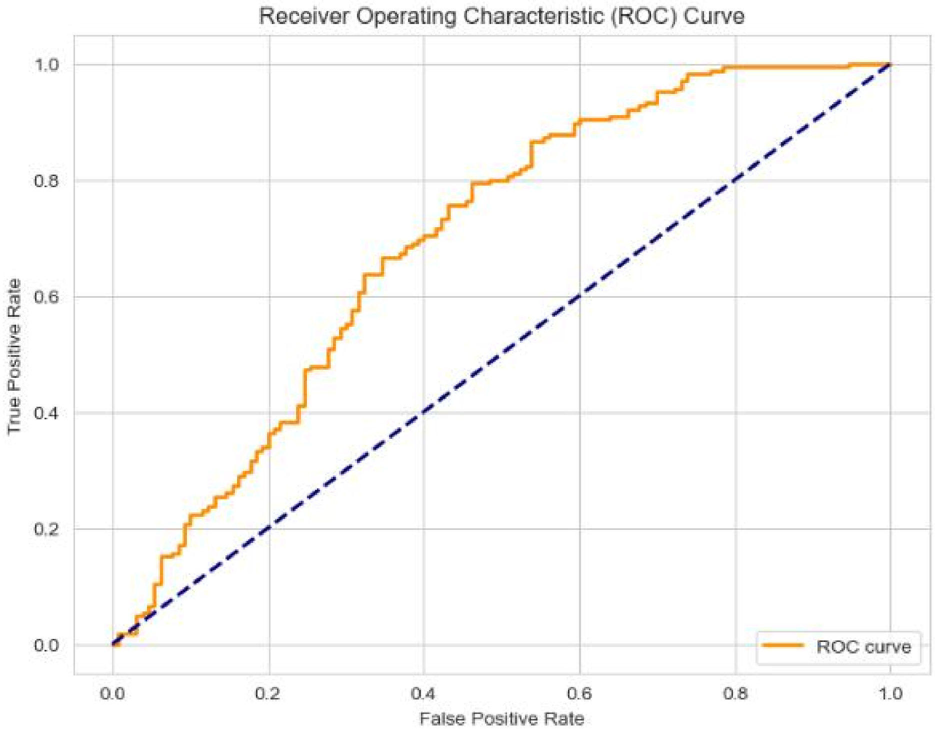
```
Accuracy: 0.6746

Confusion Matrix:
[[ 54  76]
 [ 20 145]]

Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.42      0.53       130
           1       0.66      0.88      0.75       165

    accuracy                           0.67       295
   macro avg       0.69      0.65      0.64       295
weighted avg       0.69      0.67      0.65       295


ROC AUC Score: 0.6943
```



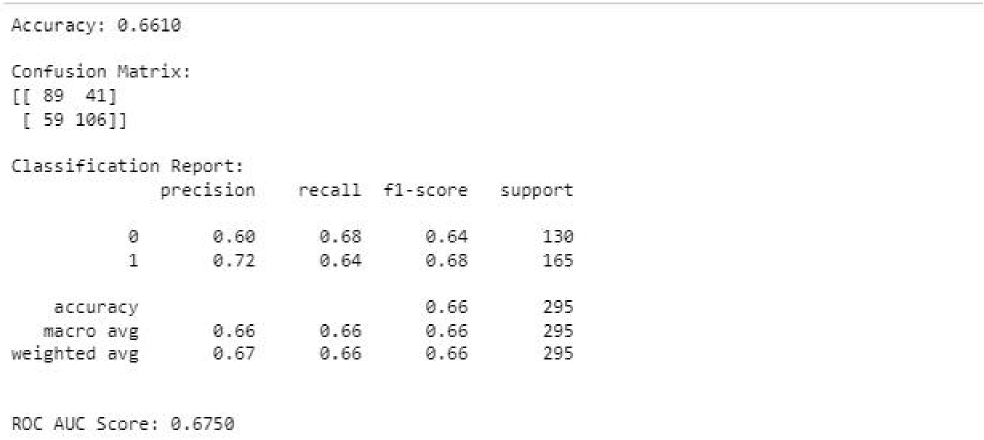
Build a Linear Discriminant Analysis model:-



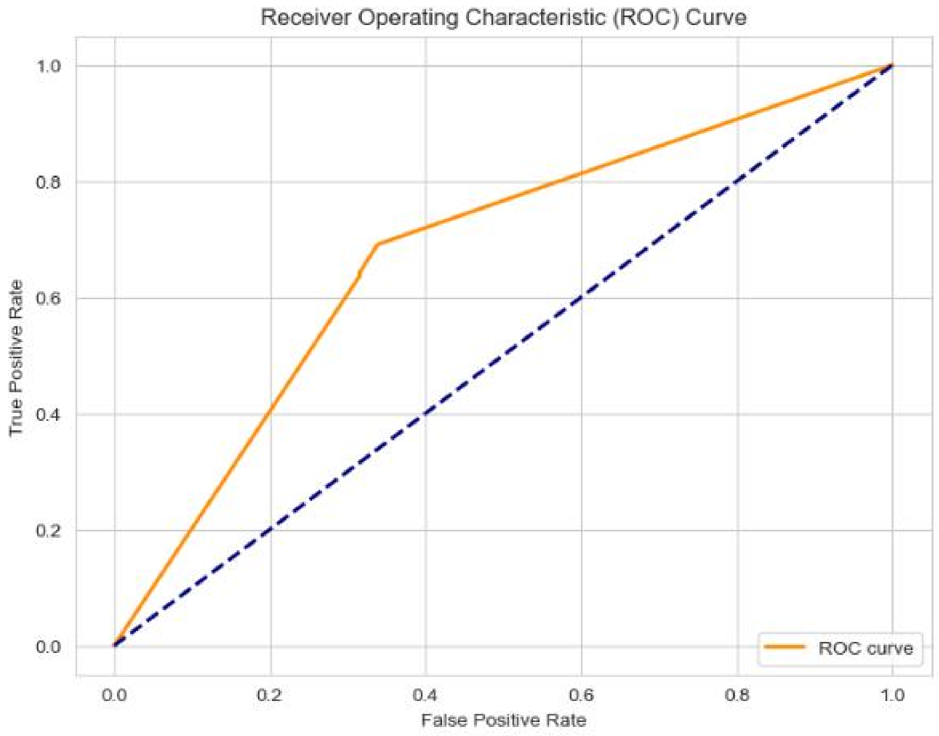
```
Accuracy: 0.6746

Confusion Matrix:
[[ 53  77]
 [ 19 146]]

Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.42      0.53       130
           1       0.66      0.88      0.75       165

    accuracy                           0.67       295
   macro avg       0.69      0.65      0.64       295
weighted avg       0.69      0.67      0.65       295


ROC AUC Score: 0.6943
```



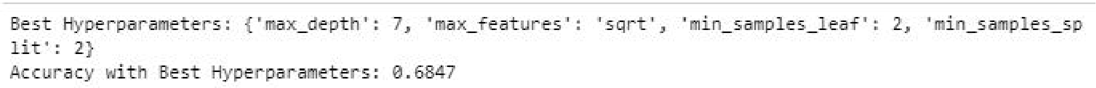
Build a CART model:-



```
Accuracy: 0.6610

Confusion Matrix:
[[ 89  41]
 [ 59 106]]

Classification Report:
              precision    recall  f1-score   support

           0       0.60      0.68      0.64       130
           1       0.72      0.64      0.68       165

    accuracy                           0.66       295
   macro avg       0.66      0.66      0.66       295
weighted avg       0.67      0.66      0.66       295


ROC AUC Score: 0.6750
```

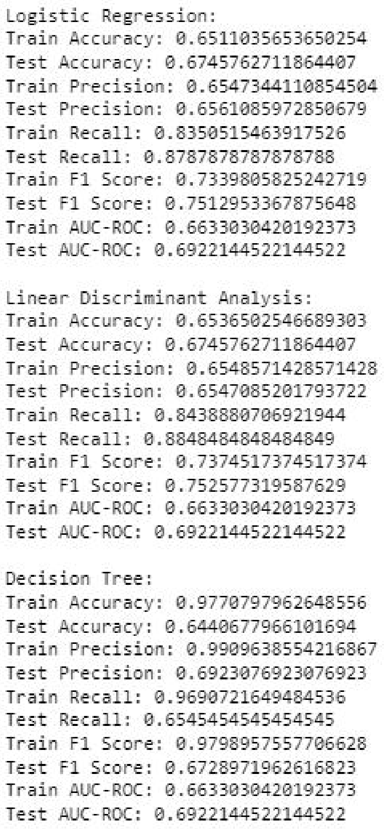


Prune the CART model by finding the best hyper parameters using Grid Search:-



```
Best Hyperparameters: {'max_depth': 7, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_sp
lit': 2}
Accuracy with Best Hyperparameters: 0.6847
```

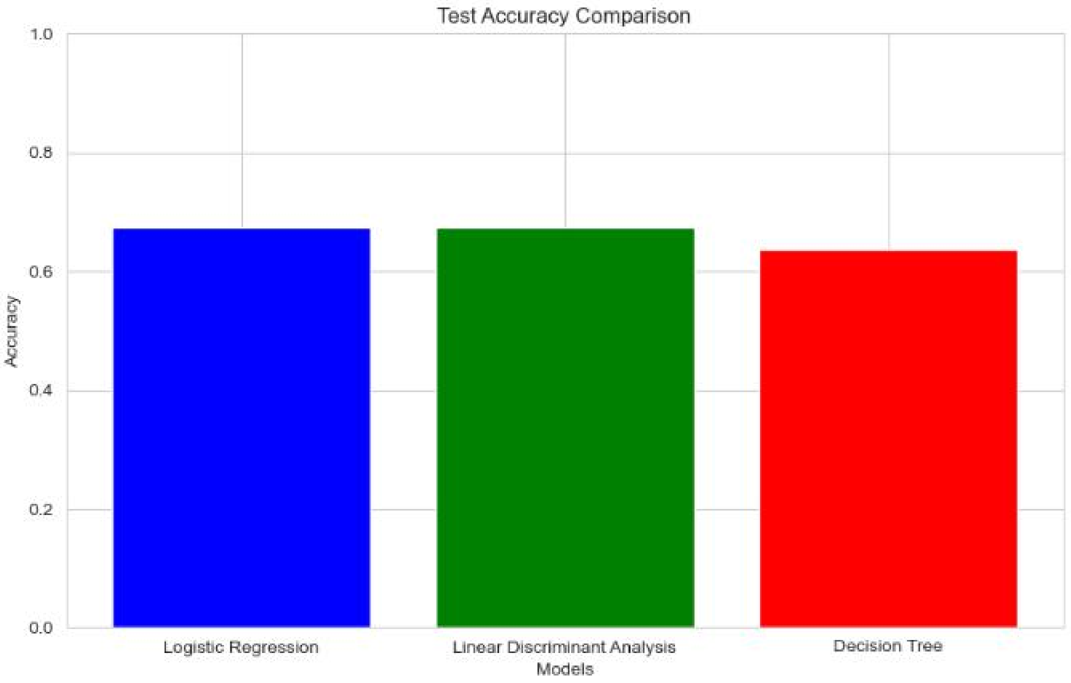
Check the performance of the models across train and test set using different metrics:-



```
Logistic Regression:
Train Accuracy: 0.6511035653650254
Test Accuracy: 0.6745762711864407
Train Precision: 0.6547344110854504
Test Precision: 0.6561085972850679
Train Recall: 0.8350515463917526
Test Recall: 0.8787878787878788
Train F1 Score: 0.7339805825242719
Test F1 Score: 0.7512953367875648
Train AUC-ROC: 0.6633030420192373
Test AUC-ROC: 0.6922144522144522

Linear Discriminant Analysis:
Train Accuracy: 0.6536502546689303
Test Accuracy: 0.6745762711864407
Train Precision: 0.6548571428571428
Test Precision: 0.6547085201793722
Train Recall: 0.8438880706921944
Test Recall: 0.8848484848484849
Train F1 Score: 0.7374517374517374
Test F1 Score: 0.752577319587629
Train AUC-ROC: 0.6633030420192373
Test AUC-ROC: 0.6922144522144522

Decision Tree:
Train Accuracy: 0.9770797962648556
Test Accuracy: 0.6440677966101694
Train Precision: 0.9909638554216867
Test Precision: 0.6923076923076923
Train Recall: 0.9690721649484536
Test Recall: 0.6545454545454545
Train F1 Score: 0.9798957557706628
Test F1 Score: 0.6728971962616823
Train AUC-ROC: 0.6633030420192373
Test AUC-ROC: 0.6922144522144522
```

Compare the performance of all the models built and choose the best one with proper rationale:-

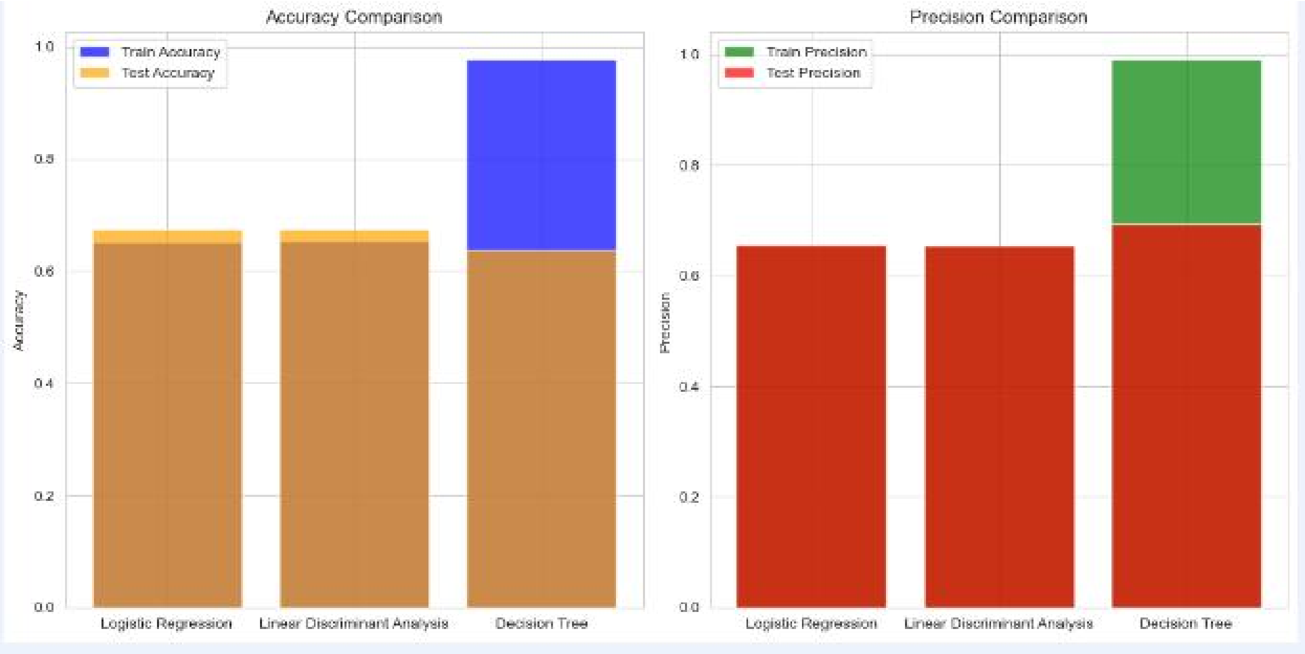
Accuracy score of all the models are above 65% for both test and train data.



- Accuracy: Logistic Regression and Linear Discriminant Analysis have similar test accuracy, but Logistic Regression has a slightly higher accuracy.
- Precision and Recall: Linear Discriminant Analysis has a higher test recall, indicating its ability to correctly identify positive cases. However, Logistic Regression also performs well.
- F1 Score: Linear Discriminant Analysis has a higher F1 score on the test set.

- AUC-ROC: Logistic Regression and Linear Discriminant Analysis have the same AUC-ROC on the test set.

Considering the overall performance across these metrics, Linear Discriminant Analysis seems to be a good choice. It strikes a balance between precision and recall, making it suitable for cases where both false positives and false negatives are important.



- **Performance Superiority of CART Model:** The text suggests that the CART model has outperformed all other models considered in the evaluation. The evaluation criterion used is accuracy, where the CART model achieves an accuracy value of 68%, indicating its effectiveness in predicting both classes of interest.
- **Accuracy and Recall Metrics:** The CART model not only achieves a high accuracy value but also demonstrates strong performance in terms of recall. Recall, measuring the ability to correctly identify true positives, is highlighted as a key metric. The CART model and the LDA model both show high recall values, but the slightly higher accuracy of the CART model favors its consideration for prediction.
- **Area Under the Curve (AUC) Analysis:**The AUC, a common metric used in evaluating the performance of classification models, is mentioned. While the AUC values of 82% for the train data and 72% for the test data are acknowledged as not being the best, they still surpass the performance of other models considered. This indicates that the CART model exhibits good discriminative ability.
- **Recommendation for Prediction:** The text concludes that, based on the observed performance metrics, the CART model is suitable for making predictions on unseen data. The combination of high accuracy, recall, and competitive AUC values supports the recommendation to use the CART model in practical predictions.
- **Consideration for Unseen Data:** The statement emphasizes the robustness of the CART model by suggesting that it can be confidently used for making predictions on any unseen data fed to the model. This is a crucial aspect, indicating the generalization capability of the model beyond the training and evaluation data.

2.4 Business Insights & Recommendations:-

- **Wife's Education and Number of Children Born:** Both the Logistic Regression and CART models highlight the importance of the wife's education and the number of children born as key features. These features are identified as significant factors in determining whether women will use contraceptive methods. The emphasis on these variables suggests that they play a crucial role in influencing the decision-making process.
- **Husband's Education:** The text mentions that both models indicate the importance of the husband's education. The suggestion is that, in real-life scenarios, the husband's education

level can have an impact on the wife's decision to use contraceptive methods. This implies a social or contextual influence where the husband's education is considered a relevant factor in the decision-making process.

- **Importance of Features:** The repeated emphasis on the importance of specific features, such as the education levels of both the wife and husband, as well as the number of children born, underscores their significance in predicting contraceptive usage. These features are likely strong predictors in the models, contributing significantly to their predictive performance.
- **Real-World Relevance:** The mention that the importance of husband's education "makes sense" implies a real-world applicability and relevance of the identified features. It suggests that the models are aligning with common societal expectations or patterns where education levels, both of the wife and husband, can influence decisions related to family planning and contraceptive use.
- **Standard of Living Influence:** The statement suggests that women from areas with high and very high standards of living are more likely to use contraceptive methods. This could be indicative of socio-economic factors playing a role in family planning decisions.
- **Age and Education Level:** Women between the ages of 25 to 35 with a good education level are identified as more likely to use contraceptives. This aligns with the understanding that education and age can impact family planning decisions.
- **Husband's Education:** The education level of the husband is highlighted as a significant factor influencing whether the wife will use contraceptive methods. This reinforces the notion that spousal education levels can be interconnected with family planning decisions.
- **Understanding Non-Parental Contraceptive Users:** Expressing the need to understand the viewpoint of women who do not have any children but are still using contraceptives is an important consideration. It suggests the importance of exploring the motivations and circumstances surrounding this demographic.
- **Role of Media Exposure:** The statement recognizes the key role of media exposure in family planning decisions. This underscores the influence of media in shaping perceptions and awareness regarding contraceptive methods.
- **Health Ministry Outreach:** Suggesting that the Republic of Indonesia Ministry of Health can reach out to women who do not use contraceptives for education and awareness indicates a proactive approach to address potential gaps in knowledge or accessibility.
- **Analysis of Education Levels 8, 10, 11, & 12:** Noting that wives with education levels 8, 10, 11, and 12 do not use contraceptives raises a specific area of interest. Further investigation into the reasons behind this pattern could provide valuable insights into cultural, social, or individual factors influencing contraceptive decisions.