Abubacker Siddhique I
PM
Feb 18, 2024

Predictive Modelling

Table of Contents

List Of Figures

List Of Tables:

Problem 1

Define the problem and perform exploratory Data Analysis :

- The university department wants to monitor various activities of a computer system that people use for diverse tasks, such as internet access, file editing, and CPU-intensive programs.
- USR (User Mode Percentage) represents the portion of time (%) that CPUs run in user mode (executing user-level application code). High user mode time indicates active application processing.
- We have to build a linear regression model to predict the USR using different attributes collected by the Sun Sparcstation 20/712.

Data Description:

- The comp-activ database comprises activity measures of computer systems.
- There are 8192 rows and 22 columns.
- Of the 22 available columns,13 are of data type float,8 are integer columns, and 1 categorical column.

```
---  ------    --------------  -----
 0   lread     8192 non-null   int64
 1   lwrite    8192 non-null   int64
 2   scall     8192 non-null   int64
 3   sread     8192 non-null   int64
 4   swrite    8192 non-null   int64
 5   fork      8192 non-null   float64
 6   exec      8192 non-null   float64
 7   rchar     8088 non-null   float64
 8   wchar     8177 non-null   float64
 9   pgout     8192 non-null   float64
 10  ppgout    8192 non-null   float64
 11  pgfree    8192 non-null   float64
 12  pgscan    8192 non-null   float64
 13  atch      8192 non-null   float64
 14  pgin      8192 non-null   float64
 15  ppgin     8192 non-null   float64
 16  pflt      8192 non-null   float64
 17  vflt      8192 non-null   float64
 18  runqsz    8192 non-null   object
 19  freemem   8192 non-null   int64
 20  freeswap  8192 non-null   int64
 21  usr       8192 non-null   int64
dtypes: float64(13), int64(8), object(1)
memory usage: 1.4+ MB
```

Figure 1

Data Dictionary:

Description
lread - Reads (transfers per second) between system memory and user memory
lwrite - writes (transfers per second) between system memory and user memory
scall - Number of system calls of all types per second
sread - Number of system read calls per second .
swrite - Number of system write calls per second .
fork - Number of system fork calls per second.fork measures the number of process creation requests per second.
exec - Number of system exec calls per second.exec counts the number of program execution requests per second
rchar - Number of characters transferred per second by system read calls.
wchar - Number of characters transfreed per second by system write calls.
pgout - Number of page out requests per second.(moving data from RAM to disk)
ppgout - Number of pages, paged out per second.
pgfree - Number of pages per second placed on the free list.
pgscan - Number of pages checked if they can be freed per second.
atch - Number of page attaches (satisfying a page fault by reclaiming a page in memory) per second.
pgin - Number of page-in requests per second.
ppgin - Number of pages paged in per second.
pflt - Number of page faults caused by protection errors (copy-on-writes).
vflt - Number of page faults caused by address translation .
runqsz - Process run queue size (The number of kernel threads in memory that are waiting for a CPU to run. Typically, this value should be less than 2. Consistently higher values mean that the system might be CPU-bound.
freemem - Number of memory pages available to user processes
freeswap - Number of disk blocks available for page swapping.
usr - Portion of time (%) that cpus run in user mode

Table 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

Figure 2

- We can see that there are null values present in rchar and wchar.
- The mean USR is around 84%.
- Most of the columns have a minimum value of 0.
- Freeswap has the highest standard deviation among the other variables and fork with the least.

Univariate analysis:

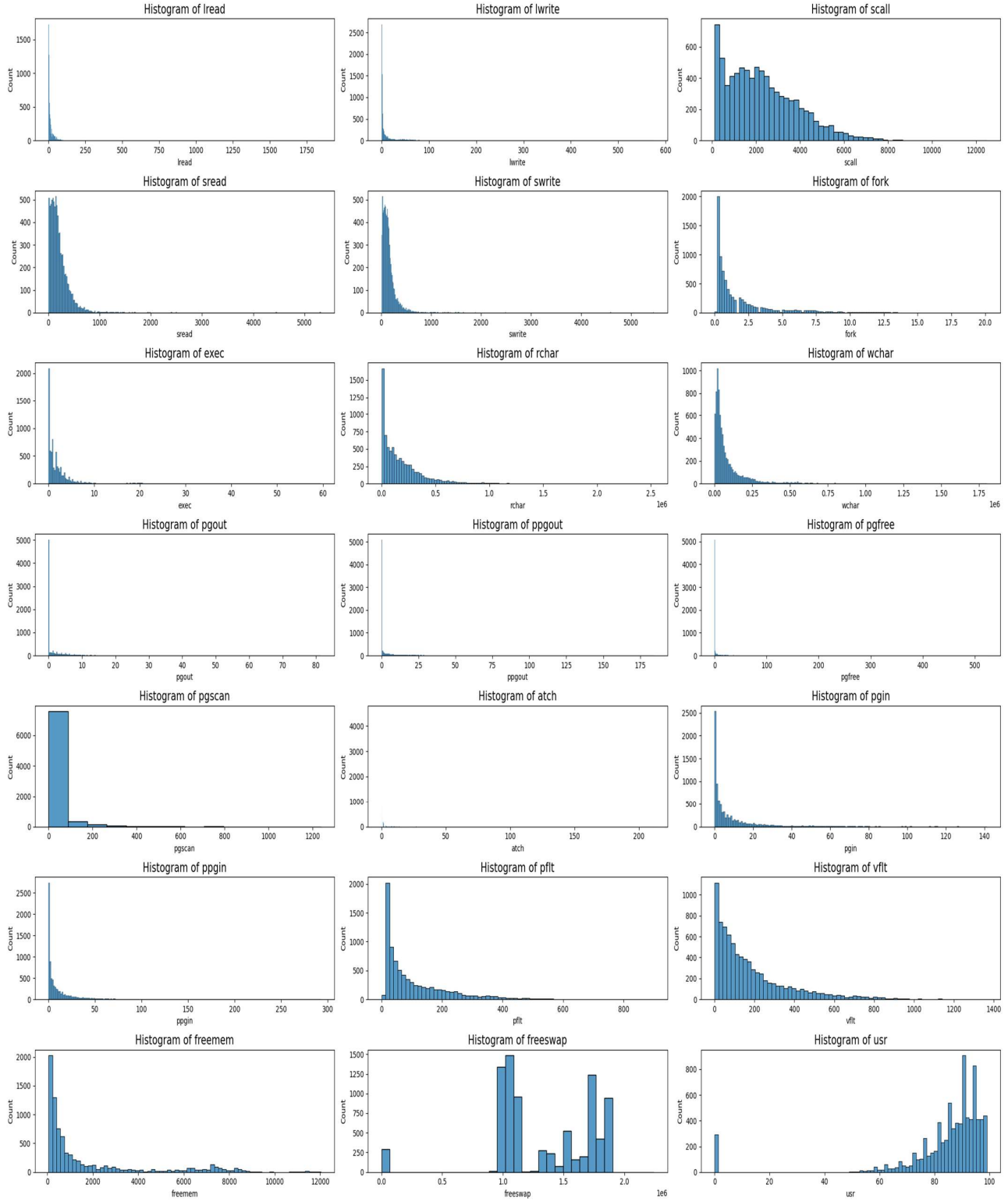

Figure 3

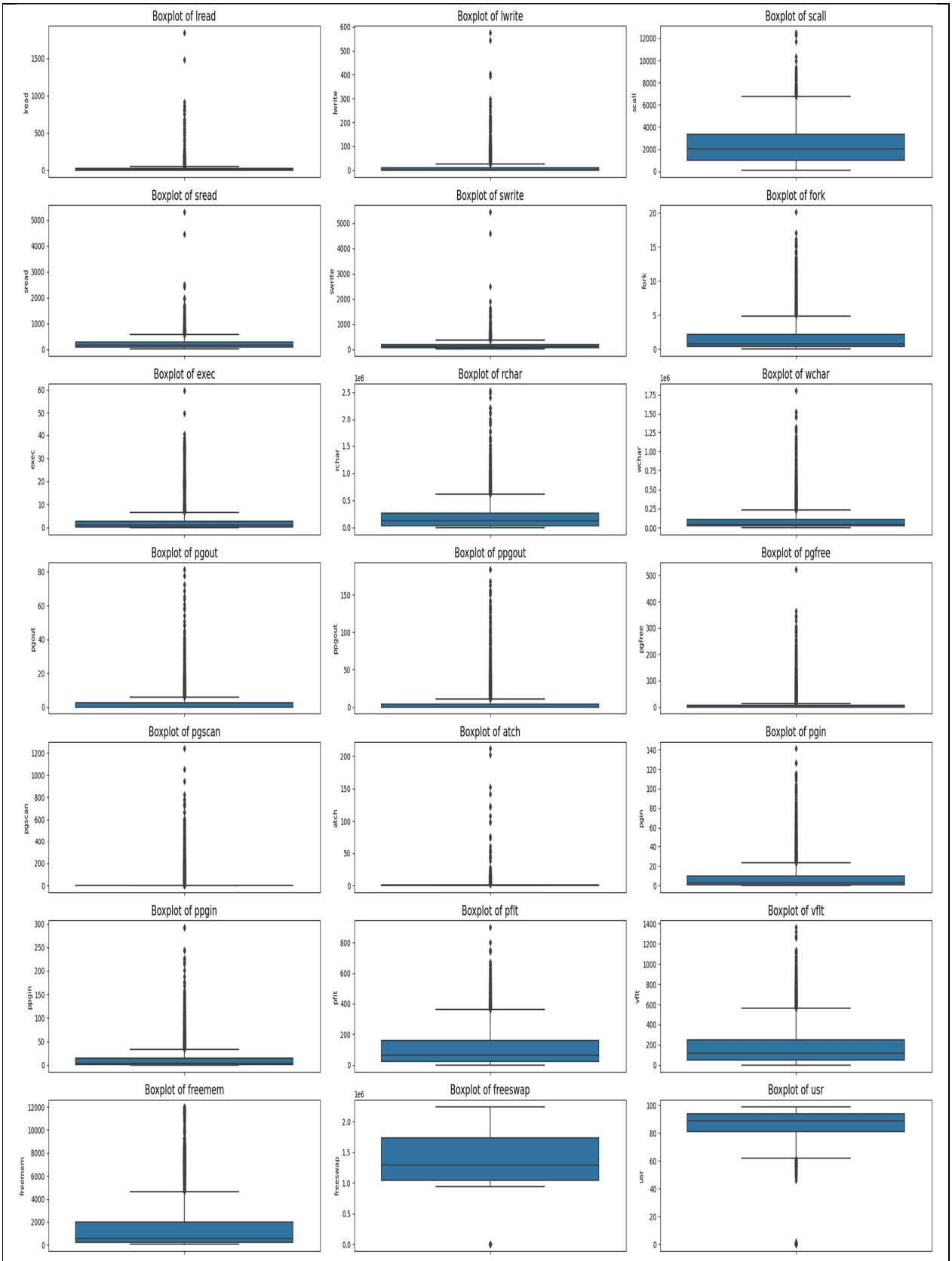
Boxplot of lread
lread
Boxplot of lwrite
lwrite
Boxplot of scall
scall
Boxplot of sread
sread
Boxplot of swrite
swrite
Boxplot of fork
fork
Boxplot of exec
exec
Boxplot of rchar
1e6
rchar
Boxplot of wchar
1e6
wchar
Boxplot of pgout
pgout
Boxplot of ppgout
ppgout
Boxplot of pgfree
pgfree
Boxplot of pgscan
pgscan
Boxplot of atch
atch
Boxplot of pgin
pgin
Boxplot of ppgin
ppgin
Boxplot of pflt
pflt
Boxplot of vflt
vflt
Boxplot of freemem
freemem
Boxplot of freeswap
1e6
freeswap
Boxplot of usr
usr

Figure 4

- There are outliers present in almost all the columns.
- The median of USR lies around 90.
- The distribution of page attaches is close to zero. This is because 50% of the values are zero, and 75% of the values fall below 0.600.
- Most of the distributions are right skewed.

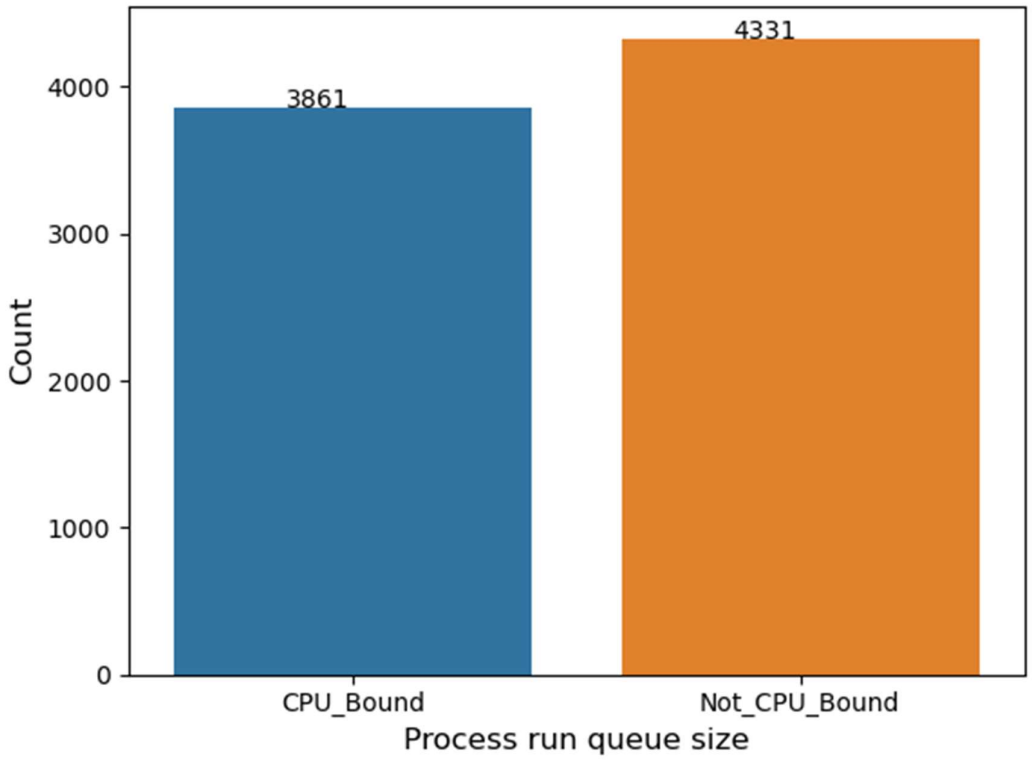

Figure 5

The process that are not limited by the processing capacity of the CPU are comparatively higher than those that are CPU bound.

Multivariate analysis:

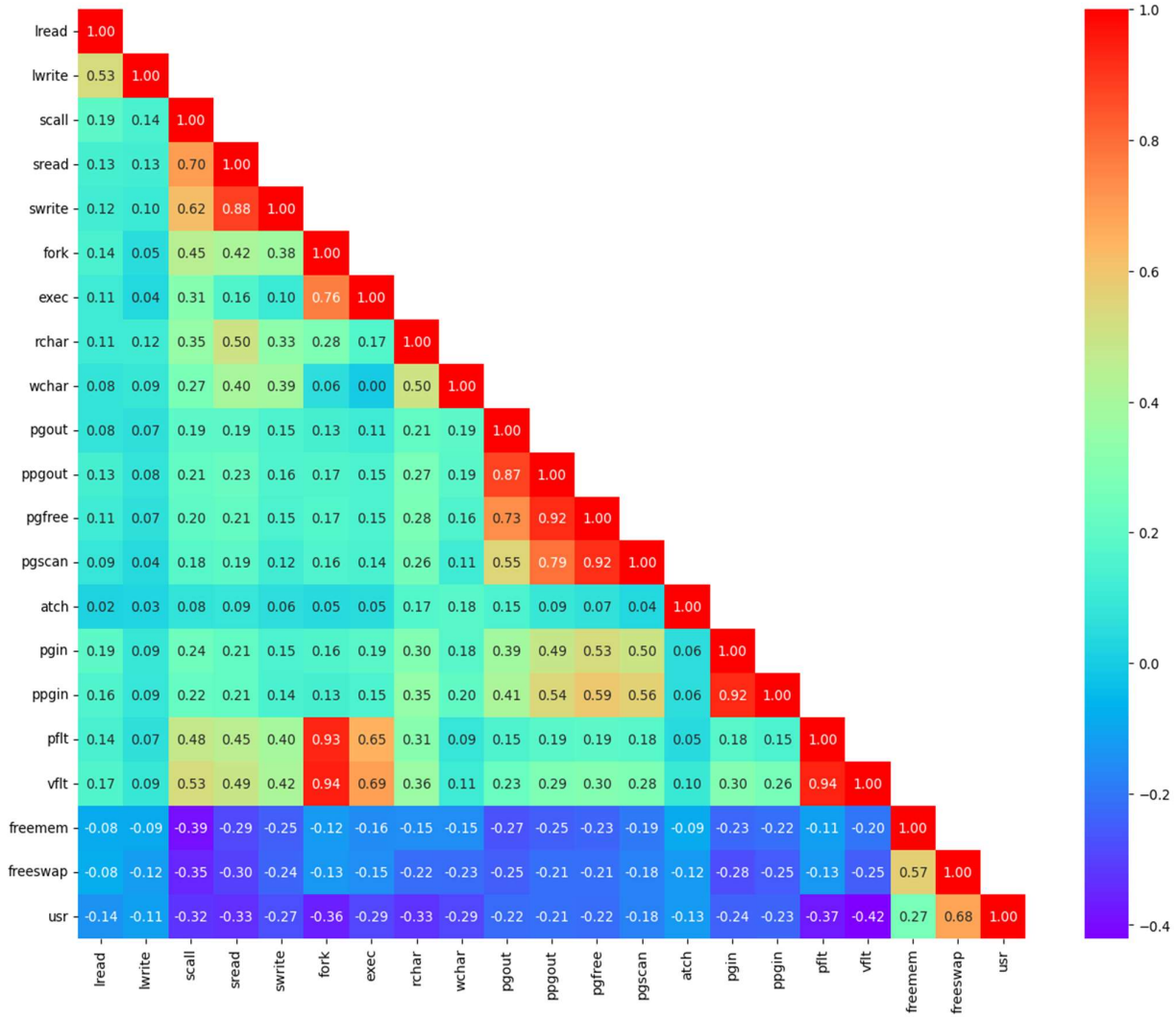

Figure 6

From the above figure, it is evident that there is multicollinearity among the different variables. Let us separate those variables with score greater than 0.7.

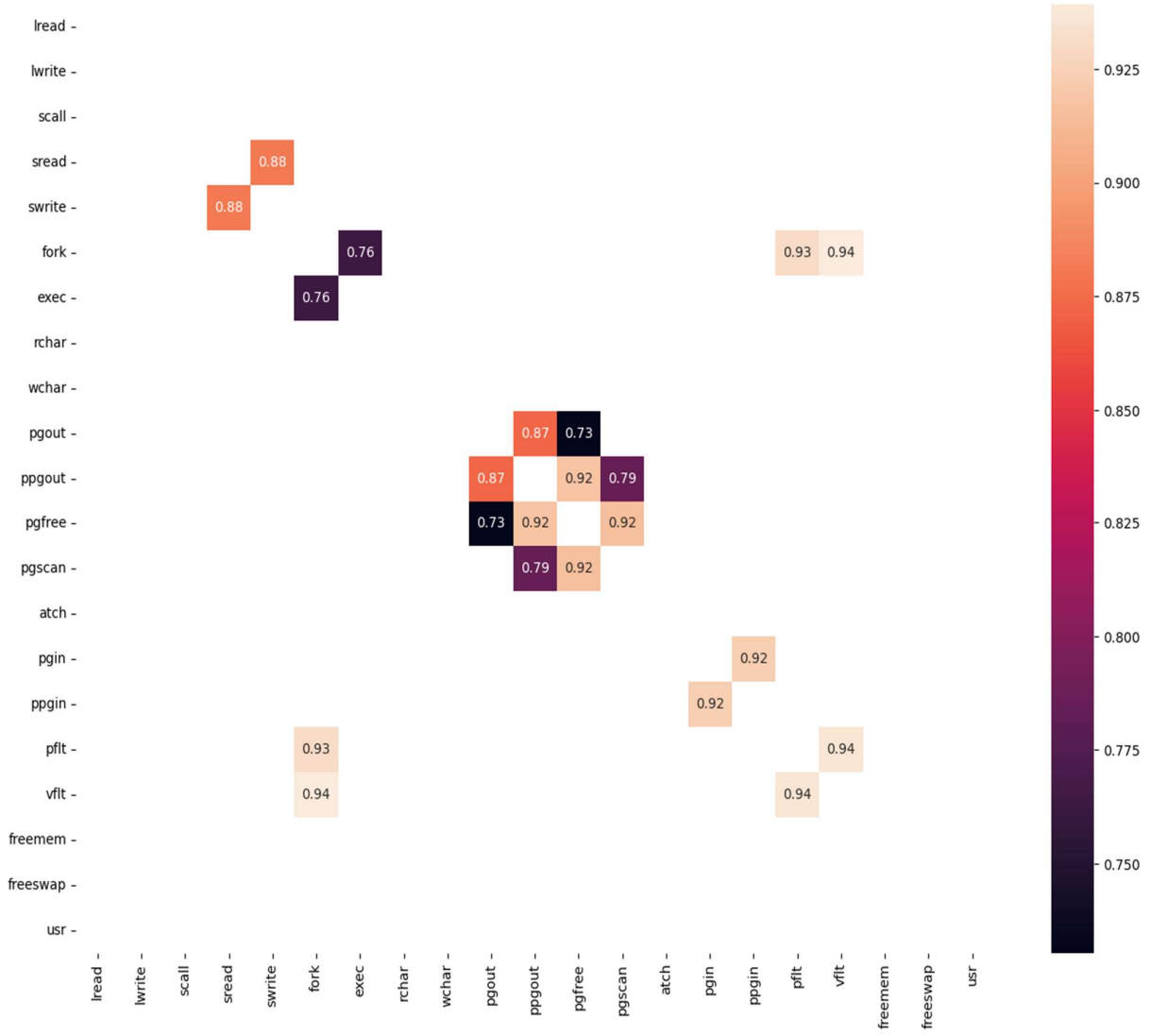

Figure 7

Variables with Strong positive correlation:

- Fork vs vflt
- Fork vs pflt
- sread vs swrite
- ppgout vs pgfree
- pgfree vs pgscan
- ppgin vs pgin
- pflt vs vflt

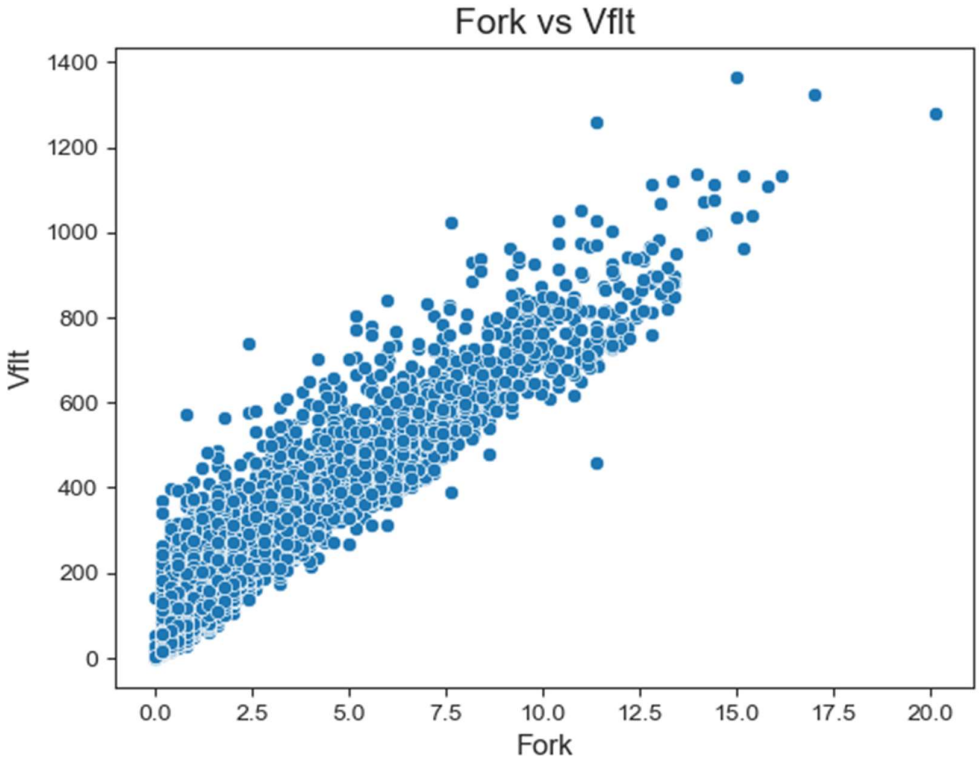

Figure 8

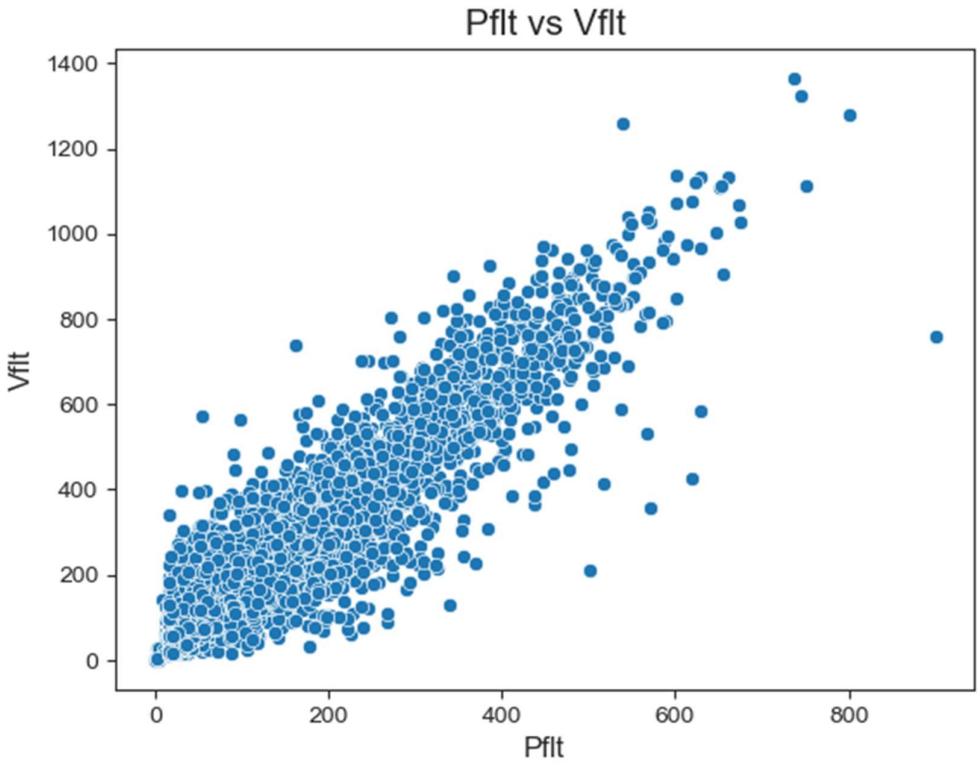

Figure 9

The USR has the highest correlation of 0.68 with Freeswap.Let us visualise it using scatter plot.

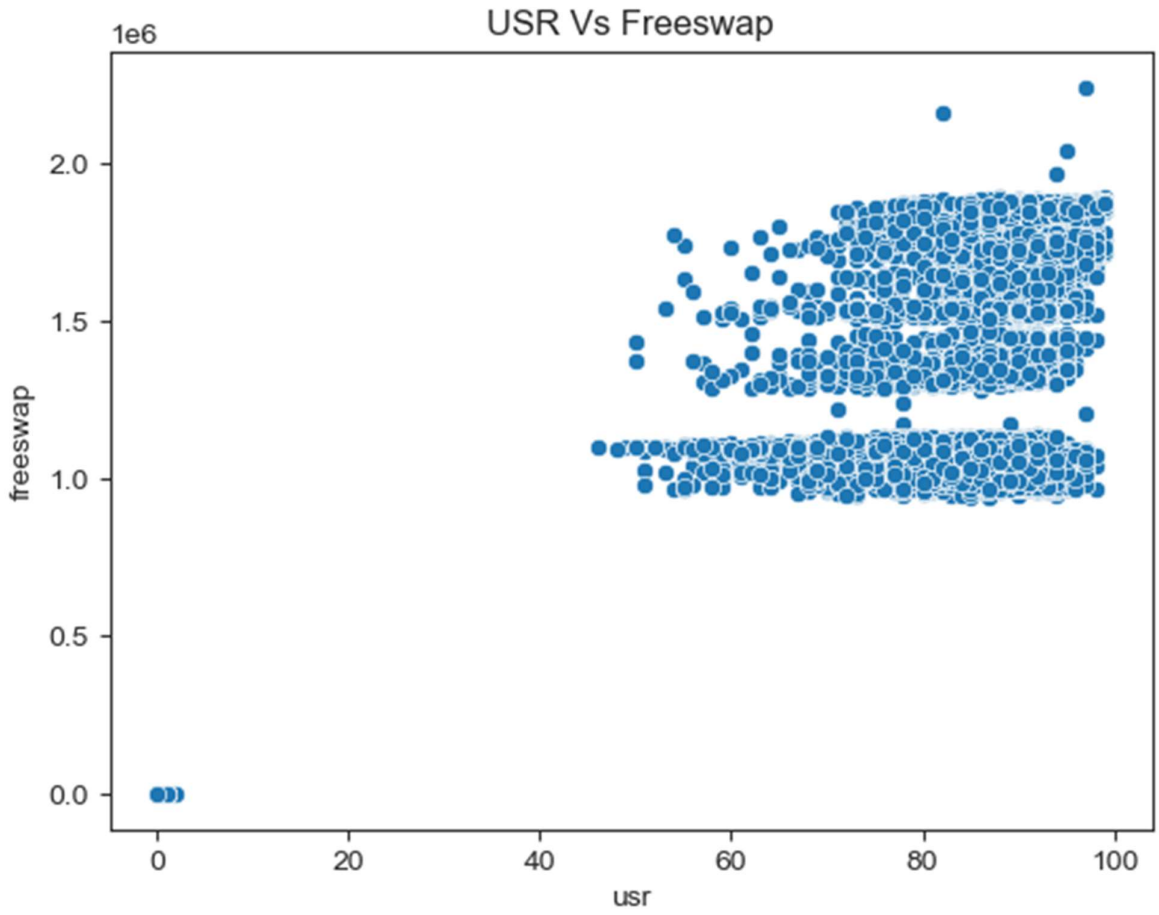

Figure 10

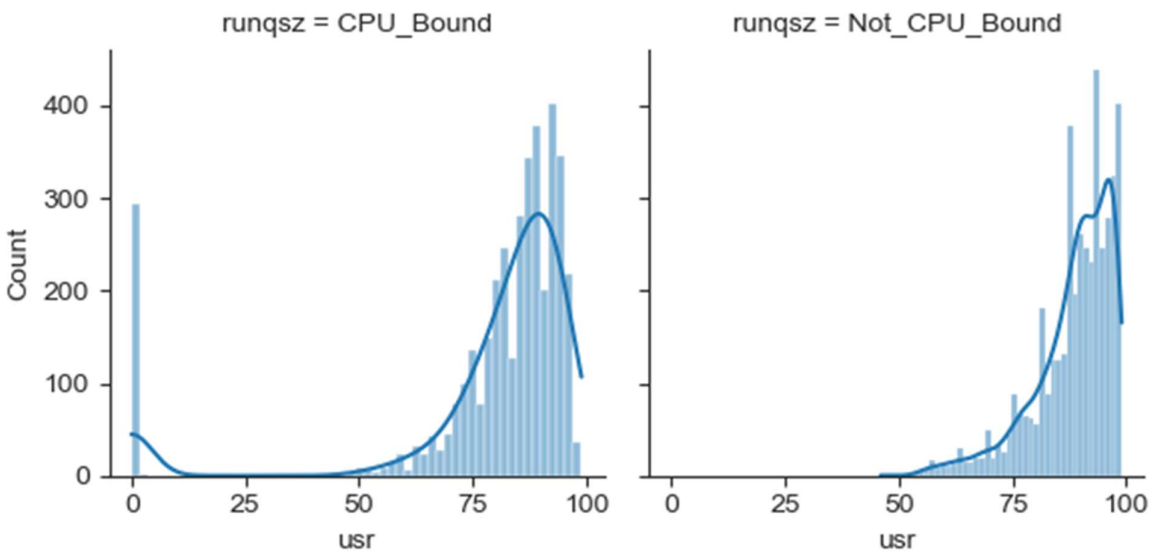

Figure 11

- From the distribution we can see that the programs that are Not CPU bound (i.e.) not limited by the CPU speed, have high USR.
- High USR values indicate that your programs are actively doing their work, like processing data or responding to user input.

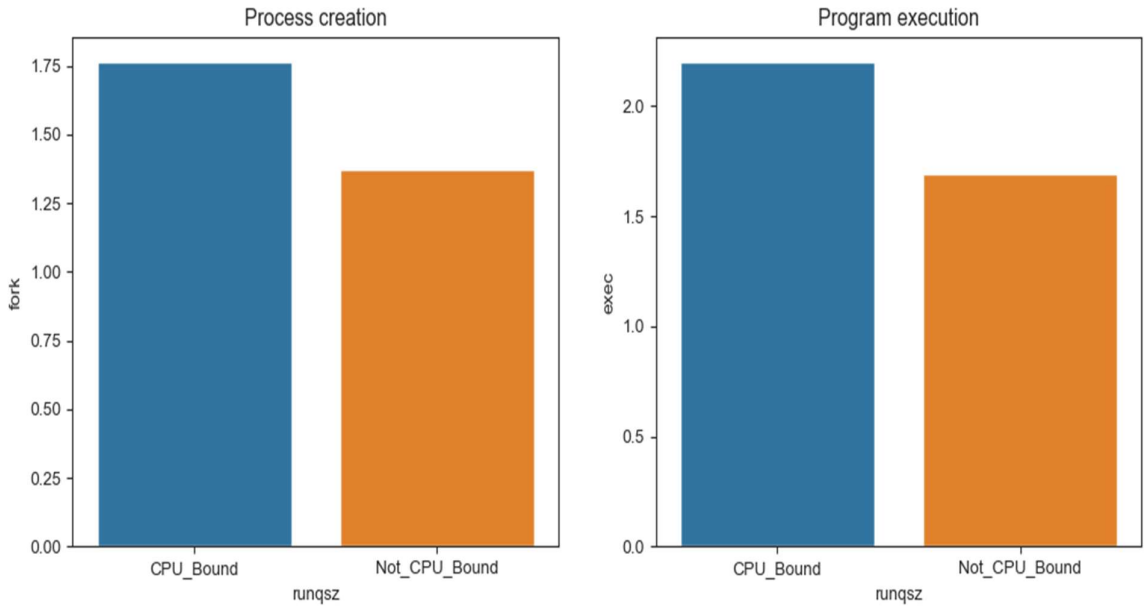

Figure 12

- The number of processes that are CPU bound are higher for both fork and exec.
- It is evident that the number of process being executed is higher that the number of new process created.

Data Pre-processing :

There are null values present in rchar and wchar. Let us impute those missing values using median.

Before imputing

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

After imputing

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

Figure 13

Data Encoding:

- The column runqsz is of binary object data type consisting of CPU bound and Not CPU bound.
- We are going to impute those values using dummy variables. This process will create a new column with values 1 & 0 with respect to the CPU Not bound condition.

Before Imputing

```
runqsz
Not_CPU_Bound    4331
CPU_Bound        3861
Name: count, dtype: int64
```

Figure 14

After Imputing

```
runqsz_Not_CPU_Bound
1    4331
0    3861
Name: count, dtype: int64
```

Figure 15

Splitting the dataset into Train-Test:

Let us split the dataset with predictor variable and dependent variable as a separate dataset.

- Shape of X Train: (5734,21)
- Shape of X Test: (2458,21)

Outlier Treatment:

An observation is considered to be an outlier if that particular observation has been mistakenly captured in the data set. Treating outliers sometimes results in the models having better performance but the models lose out on generalization. So, a good way to approach this would be to build models with and without treating outliers and then report the results.

Model Building- Linear regression:

Linear Regression using statsmodel(OLS) :

Model Summary without treating the outliers:

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.643
Model:	OLS	Adj. R-squared:	0.642
Method:	Least Squares	F-statistic:	489.6
Date:	Fri, 09 Feb 2024	Prob (F-statistic):	0.00
Time:	23:15:06	Log-Likelihood:	-21788.
No. Observations:	5734	AIC:	4.362e+04
Df Residuals:	5712	BIC:	4.377e+04
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	44.6380	0.746	59.831	0.000	43.175	46.101
lread	-0.0199	0.003	-6.214	0.000	-0.026	-0.014
lwrite	0.0048	0.006	0.795	0.427	-0.007	0.017
scall	0.0010	0.000	7.451	0.000	0.001	0.001
sread	-0.0005	0.002	-0.257	0.797	-0.004	0.003
swrite	-0.0020	0.002	-1.018	0.309	-0.006	0.002
fork	-1.7222	0.244	-7.052	0.000	-2.201	-1.244
exec	-0.0896	0.048	-1.879	0.060	-0.183	0.004
rchar	-4.062e-06	8.29e-07	-4.898	0.000	-5.69e-06	-2.44e-06
wchar	-1.164e-05	1.28e-06	-9.118	0.000	-1.41e-05	-9.14e-06
pgout	-0.1739	0.064	-2.717	0.007	-0.299	-0.048
ppgout	0.0989	0.037	2.701	0.007	0.027	0.171
pgfree	-0.0703	0.020	-3.508	0.000	-0.110	-0.031
pgscan	0.0086	0.006	1.362	0.173	-0.004	0.021
atch	-0.0786	0.027	-2.949	0.003	-0.131	-0.026
pgin	0.0913	0.029	3.103	0.002	0.034	0.149
ppgin	-0.0594	0.019	-3.128	0.002	-0.097	-0.022
pflt	-0.0415	0.004	-9.697	0.000	-0.050	-0.033
vflt	0.0223	0.003	6.665	0.000	0.016	0.029
freemem	-0.0016	7.53e-05	-21.489	0.000	-0.002	-0.001
freeswap	3.219e-05	4.54e-07	70.985	0.000	3.13e-05	3.31e-05
runqsz_Not_CPU_Bound	7.7908	0.303	25.693	0.000	7.196	8.385

Omnibus:	1507.319	Durbin-Watson:	2.057
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4768.238
Skew:	-1.333	Prob(JB):	0.00
Kurtosis:	6.585	Cond. No.	7.48e+06

Figure 16

Model Summary After treating the outliers:

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.796
Model:	OLS	Adj. R-squared:	0.795
Method:	Least Squares	F-statistic:	1115.
Date:	Sun, 18 Feb 2024	Prob (F-statistic):	0.00
Time:	12:06:56	Log-Likelihood:	-16657.
No. Observations:	5734	AIC:	3.336e+04
Df Residuals:	5713	BIC:	3.350e+04
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.1217	0.316	266.106	0.000	83.502	84.741
lread	-0.0635	0.009	-7.071	0.000	-0.081	-0.046
lwrite	0.0482	0.013	3.671	0.000	0.022	0.074
scall	-0.0007	6.28e-05	-10.566	0.000	-0.001	-0.001
sread	0.0003	0.001	0.305	0.760	-0.002	0.002
swrite	-0.0054	0.001	-3.777	0.000	-0.008	-0.003
fork	0.0293	0.132	0.222	0.824	-0.229	0.288
exec	-0.3212	0.052	-6.220	0.000	-0.422	-0.220
rchar	-5.167e-06	4.88e-07	-10.598	0.000	-6.12e-06	-4.21e-06
wchar	-5.403e-06	1.03e-06	-5.232	0.000	-7.43e-06	-3.38e-06
pgout	-0.3688	0.090	-4.098	0.000	-0.545	-0.192
ppgout	-0.0766	0.079	-0.973	0.330	-0.231	0.078
pgfree	0.0845	0.048	1.769	0.077	-0.009	0.178
pgscan	5.192e-14	2.39e-16	216.826	0.000	5.15e-14	5.24e-14
atch	0.6276	0.143	4.394	0.000	0.348	0.908
pgin	0.0200	0.028	0.703	0.482	-0.036	0.076
ppgin	-0.0673	0.020	-3.415	0.001	-0.106	-0.029
pflt	-0.0336	0.002	-16.957	0.000	-0.037	-0.030
vflt	-0.0055	0.001	-3.830	0.000	-0.008	-0.003
freemem	-0.0005	5.07e-05	-9.038	0.000	-0.001	-0.000
freeswap	8.832e-06	1.9e-07	46.472	0.000	8.46e-06	9.2e-06
runqsz_Not_CPU_Bound	1.6153	0.126	12.819	0.000	1.368	1.862

Omnibus:	1103.645	Durbin-Watson:	2.016
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2372.553
Skew:	-1.119	Prob(JB):	0.00
Kurtosis:	5.219	Cond. No.	2.92e+22

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.34e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Figure 17

- We can see that the R square has significantly improved after treating the outliers from 64% to 79%.
- Therefore, we are proceeding with the model with outliers being treated.

Variance Inflation Factor (VIF):

```
VIF values:

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

Figure 18

- It is evident that multicollinearity is present, as the variables have high VIF.

- Let's drop the variables one by one to see if the model performance improves.

Dropping ppgout:

After dropping the ppgout ,the R square and the adjusted R squared remains the same.

```
VIF values:

const                    29.021961
lread                     5.350387
lwrite                    4.328325
scall                     2.960379
sread                     6.420135
swrite                    5.597025
fork                     13.027305
exec                      3.239231
rchar                     2.133614
wchar                     1.580894
pgout                     6.453978
pgfree                    6.172847
pgscan                         NaN
atch                      1.875553
pgin                     13.784007
ppgin                    13.898848
pflt                     12.001460
vflt                     15.966865
freemem                   1.959267
freeswap                  1.838167
runqsz_Not_CPU_Bound      1.156421
dtype: float64
```

Figure 19

Now Vflt has the highest VIF and pgscan is NAN.

Dropping pgscan and Vflt:

- R-squared: 0.796 .
- Adjusted R-squared: 0.795.

```
VIF values:

const                  28.641818
lread                   5.335455
lwrite                  4.327130
scall                   2.952947
sread                   6.374687
swrite                  5.595777
fork                   10.089700
exec                    3.235396
rchar                   2.123783
wchar                   1.558923
pgout                   6.450724
pgfree                  6.149223
atch                    1.864254
pgin                   13.602134
ppgin                  13.898845
pflt                    9.131802
freemem                 1.957966
freeswap                1.787695
runqsz_Not_CPU_Bound    1.156363
dtype: float64
```

Figure 20

Dropping ppgin:

- R-squared: 0.795.A drop of 0.001 which is not significant.
- Adjusted R-squared: 0.795.

```
VIF values:

const                  28.594882
lread                   5.304009
lwrite                  4.316362
scall                   2.951826
sread                   6.374556
swrite                  5.595670
fork                   10.074886
exec                    3.235387
rchar                   2.090401
wchar                   1.558921
pgout                   6.445478
pgfree                  6.093623
atch                    1.863536
pgin                    1.529142
pflt                    9.131545
freemem                 1.957713
freeswap                1.785393
runqsz_Not_CPU_Bound    1.155990
dtype: float64
```

Figure 21

Now Fork has the highest VIF.

Dropping fork:

- R-squared: 0.795.
- Adjusted R-squared: 0.794. A drop of 0.001 which is not significant.

```
VIF values:

const                 28.440419
lread                  5.285069
lwrite                 4.298019
scall                  2.914853
sread                  6.373458
swrite                 5.390263
exec                   2.856973
rchar                  2.089364
wchar                  1.550686
pgout                  6.445377
pgfree                 6.093041
atch                   1.862553
pgin                   1.526800
pflt                   3.458168
freemem                1.957226
freeswap               1.782829
runqsz_Not_CPU_Bound   1.155448
dtype: float64
```

Figure 22

Dropping sread:

- R-squared: 0.795.
- Adjusted R-squared: 0.794.

```
VIF values:

const                 28.366808
lread                  5.277543
lwrite                 4.288733
scall                  2.657189
swrite                 3.013887
exec                   2.850220
rchar                  1.673113
wchar                  1.537416
pgout                  6.444663
pgfree                 6.092363
atch                   1.861273
pgin                   1.525797
pflt                   3.436271
freemem                1.956658
freeswap               1.769115
runqsz_Not_CPU_Bound   1.155441
dtype: float64
```

Figure 23

Dropping pgfree:

- R-squared: 0.795.
- Adjusted R-squared: 0.794.

```
VIF values:

const                  28.366778
lread                   5.272488
lwrite                  4.282984
scall                   2.653943
swrite                  3.012451
exec                    2.847353
rchar                   1.672481
wchar                   1.537067
pgout                   2.029172
atch                    1.860242
pgin                    1.497984
pflt                    3.436202
freemem                 1.945888
freeswap                1.767780
runqsz_Not_CPU_Bound    1.155214
dtype: float64
```

Figure 24

Dropping lwrite:

- R-squared: 0.794. A drop of 0.001 which is not significant.
- Adjusted R-squared: 0.794.

```
VIF values:

const                  28.299206
lread                   1.294870
scall                   2.650952
swrite                  3.012182
exec                    2.834855
rchar                   1.672218
wchar                   1.528722
pgout                   2.028322
atch                    1.859941
pgin                    1.468363
pflt                    3.300995
freemem                 1.944841
freeswap                1.767776
runqsz_Not_CPU_Bound    1.148773
dtype: float64
```

Figure 25

Now all the variables have VIF less than 5.

Model Summary:

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.794
Model:	OLS	Adj. R-squared:	0.794
Method:	Least Squares	F-statistic:	1700.
Date:	Sun, 18 Feb 2024	Prob (F-statistic):	0.00
Time:	12:06:58	Log-Likelihood:	-16681.
No. Observations:	5734	AIC:	3.339e+04
Df Residuals:	5720	BIC:	3.348e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.1528	0.312	269.584	0.000	83.541	84.765
lread	-0.0374	0.004	-8.429	0.000	-0.046	-0.029
scall	-0.0007	5.97e-05	-11.237	0.000	-0.001	-0.001
swrite	-0.0058	0.001	-5.512	0.000	-0.008	-0.004
exec	-0.3696	0.048	-7.627	0.000	-0.465	-0.275
rchar	-5.533e-06	4.33e-07	-12.774	0.000	-6.38e-06	-4.68e-06
wchar	-4.572e-06	1.02e-06	-4.491	0.000	-6.57e-06	-2.58e-06
pgout	-0.3572	0.038	-9.359	0.000	-0.432	-0.282
atch	0.6127	0.143	4.293	0.000	0.333	0.893
pgin	-0.0872	0.009	-9.373	0.000	-0.105	-0.069
pflt	-0.0405	0.001	-38.806	0.000	-0.043	-0.038
freemem	-0.0005	5.07e-05	-9.226	0.000	-0.001	-0.000
freeswap	8.916e-06	1.87e-07	47.713	0.000	8.55e-06	9.28e-06
runqsz_Not_CPU_Bound	1.6330	0.126	12.959	0.000	1.386	1.880

Omnibus:	1041.933	Durbin-Watson:	2.013
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2191.377
Skew:	-1.070	Prob(JB):	0.00
Kurtosis:	5.144	Cond. No.	7.61e+06

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

Figure 26

- All the P values are less than 0.05.
- 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.

Testing the Assumptions of Linear Regression

1.Linearity:

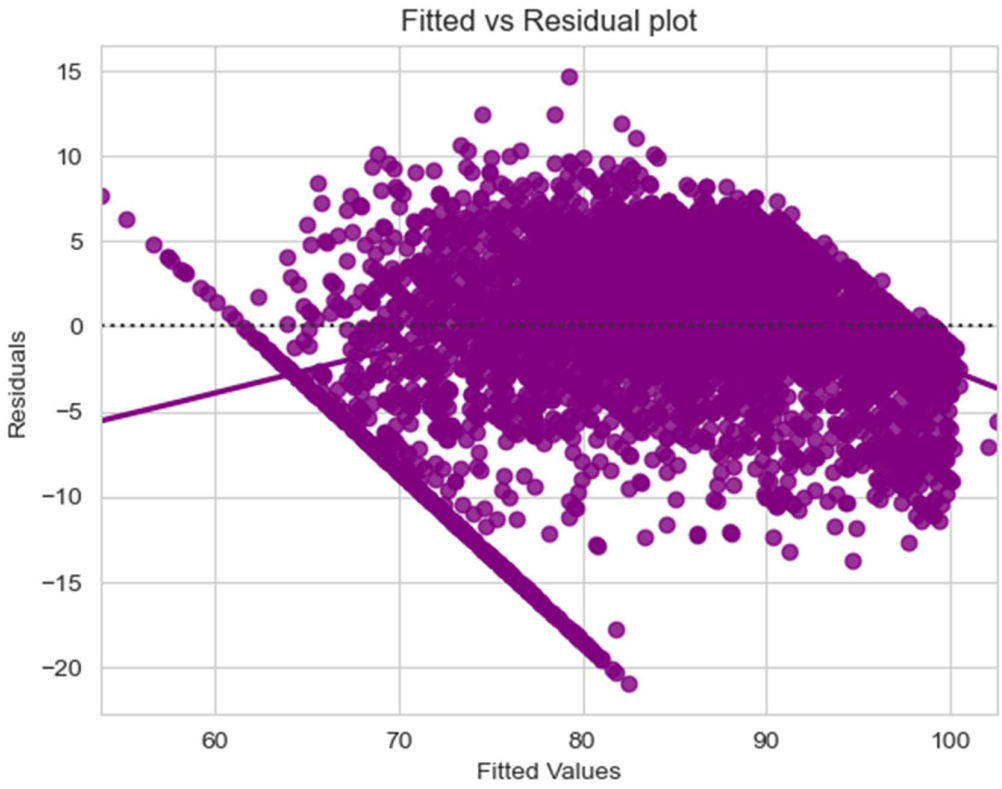

Figure 27

There seems no pattern in the data. Thus, we can confirm that the residuals are independent and the dependent and independent variables are linearly related.

2.Test For Normality:

Using Histogram and Q-Q plot.

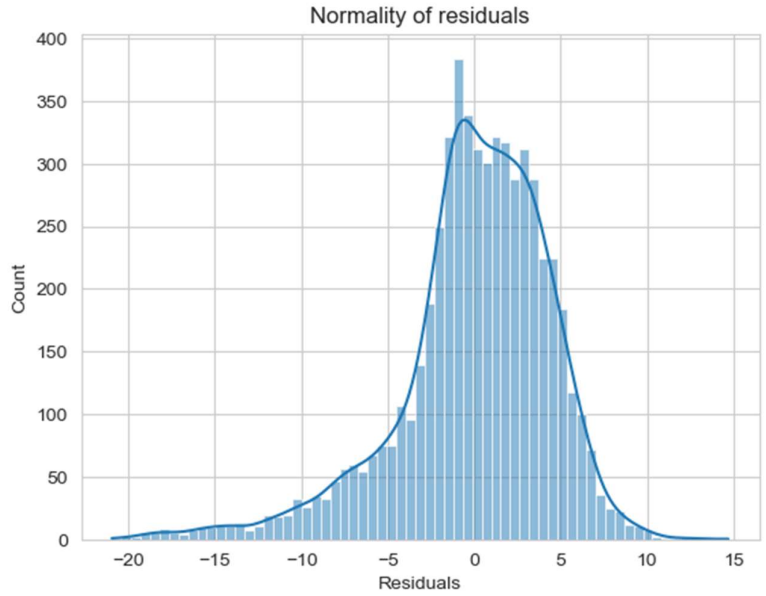

Figure 28

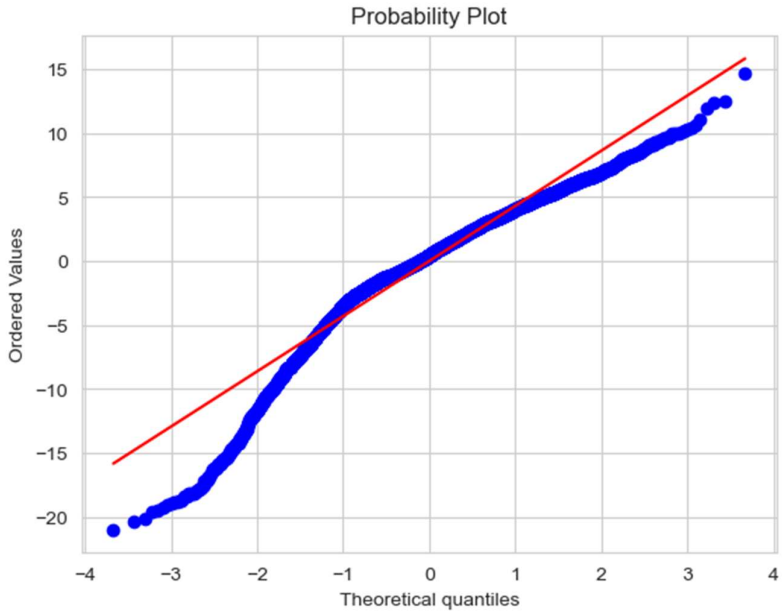

Figure 29

We can see that the datapoints are approximately normally distributed.

3.Homoscedasticity:

Using goldfeldquandt test.

- Null hypothesis: Residuals are homoscedastic.
- Alternate hypothesis: Residuals have heteroscedasticity.
- P values =0.001
- As the P value is less than 5%, we fail we reject the null hypothesis. Thus the variance of the residuals are homoscedastic.

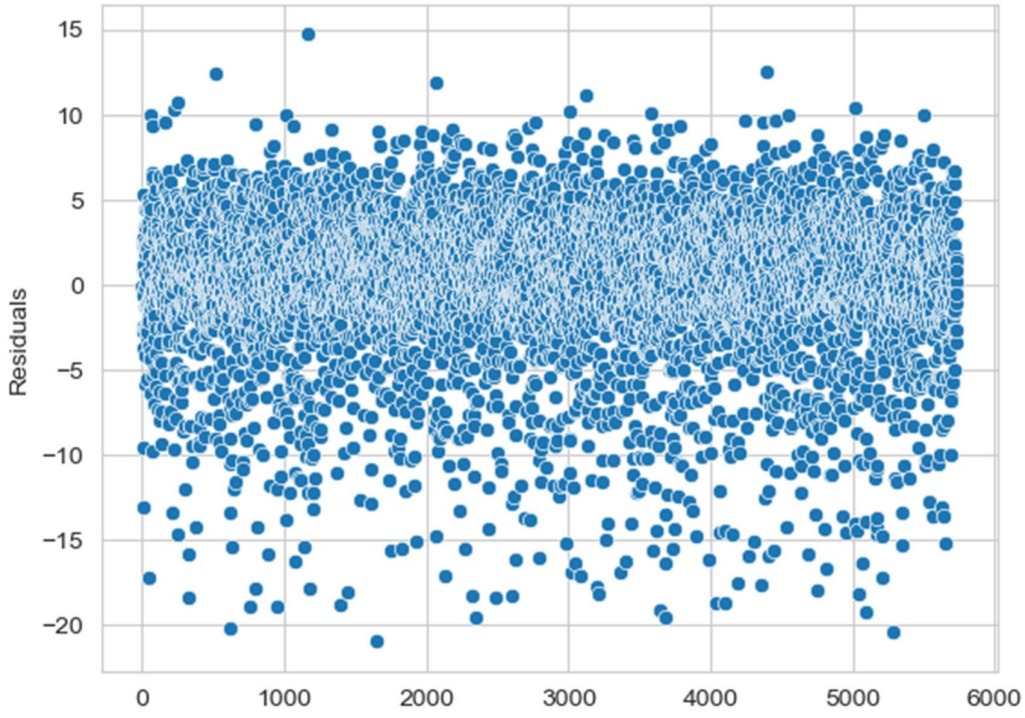

Figure 30

Root Means Squared Error for the Test and Train:

- RMSE Train: 4.437.
- RMSE Test: 4.679.

We can see that RMSE on the train and test sets are comparable. So, our model is not suffering from overfitting.

Mean Absolute Error for the Test and Train:

- MAE Train: 3.295.
- MAE Test: 3.394.

MAE indicates that our current model is able to predict USR within a mean error of 3.39 units on the test data.

Linear Regression using sklearn :

The Coefficient of each column are:

```
array([ 0.00000000e+00, -3.73606013e-02, -6.70419577e-04, -5.82414599e-03,
       -3.69591358e-01, -5.53294065e-06, -4.57199838e-06, -3.57151693e-01,
        6.12726858e-01, -8.72321981e-02, -4.04722425e-02, -4.67682615e-04,
        8.91643215e-06,  1.63296436e+00])
```

Figure 31

Intercept: 84.152

R Square:

- The coefficient of determination R^2 of the prediction on Train set 0.794.
- The coefficient of determination R^2 of the prediction on Test set 0.764.

Root Means Squared Error for the Test and Train:

- RMSE Train: 4.437.
- RMSE Test: 4.679.

We can see that RMSE on the train and test sets are comparable. So, our model is not suffering from overfitting.

Mean Absolute Error for the Test and Train:

- MAE Train: 3.295.
- MAE Test: 3.394.

MAE indicates that our current model is able to predict USR within a mean error of 3.39 units on the test data.

Equation of the Linear Regression Model:

USR = 84.152+ (-0.037) * lread + (-0.0006)*scall + (-0.005)* swrite + (-0.369) * exec + (-5.532940645573015e-06)* rchar + (-4.571998378510463e-06)* wchar + (-0.357) * pgout + 0.612 * (atch) + (-0.087) * pgin + (-0.0404) * pflt + (-0.0004) * freemem + 8.916432149272178e-06 * (freeswap) + (1.632) * runqsz_Not_CPU_Bound .

Business Insights & Recommendations:

- From the equation, we can see that the top 2 variables with the maximum impact on the USR are CPU not bound condition and atch.
- This implies that if the number of processes that are not CPU bound increases by one unit, the USR increases by 1.63 units.
- 1 Unit increase in number of page attaches per second, increases the USR by 0.61 units.
- Number of program execution requests per second has the highest negative relationship with the USR.
- An increase in 1 unit of exec, decreases the USR by -0.36 units.

Problem 2:

Define the problem and perform exploratory Data Analysis:

- 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.
- Your task 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.

Shape and Datatypes:

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

Figure 32

- The dataset consists of 1473 rows and 10 columns
- Among the 10 columns, **7** are of object datatype, 2 are of float datatype, and 1 is of integer datatype.

Data Dictionary:

Description
1. Wife's age (numerical)-Current age of Wife
2. Wife's education (categorical) -Uneducated, primary, secondary, tertiary.
3. Husband's education (categorical) - Uneducated, primary, secondary, tertiary.
4. Number of children ever born (numerical)
5. Wife's religion (binary) Non-Scientology, Scientology
6. Wife's now working? (binary) Yes, No
7. Husband's occupation (categorical)
8. Standard-of-living index (categorical)
9. Media exposure (binary) Good, Not good
10. Contraceptive method used (class attribute) No,Yes

Table 2

Statistical Summary:

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

Figure 33

- 50% if the females have age less than 32.
- The average number of children is around 3.
- The youngest female participant in the survey is 16 years old.

Count of different values for categorical variables:

```
Wife_ education
Tertiary      577
Secondary     410
Primary       334
Uneducated    152
Name: count, dtype: int64


Husband_education
Tertiary      899
Secondary     352
Primary       178
Uneducated     44
Name: count, dtype: int64


Wife_religion
Scientology        1253
Non-Scientology     220
Name: count, dtype: int64


Wife_Working
No     1104
Yes     369
Name: count, dtype: int64


Standard_of_living_index
Very High    684
High         431
Low          229
Very Low     129
Name: count, dtype: int64


Media_exposure
Exposed        1364
Not-Exposed     109
Name: count, dtype: int64


Contraceptive_method_used
Yes    844
No     629
Name: count, dtype: int64
```

Figure 34

Univariate Analysis:

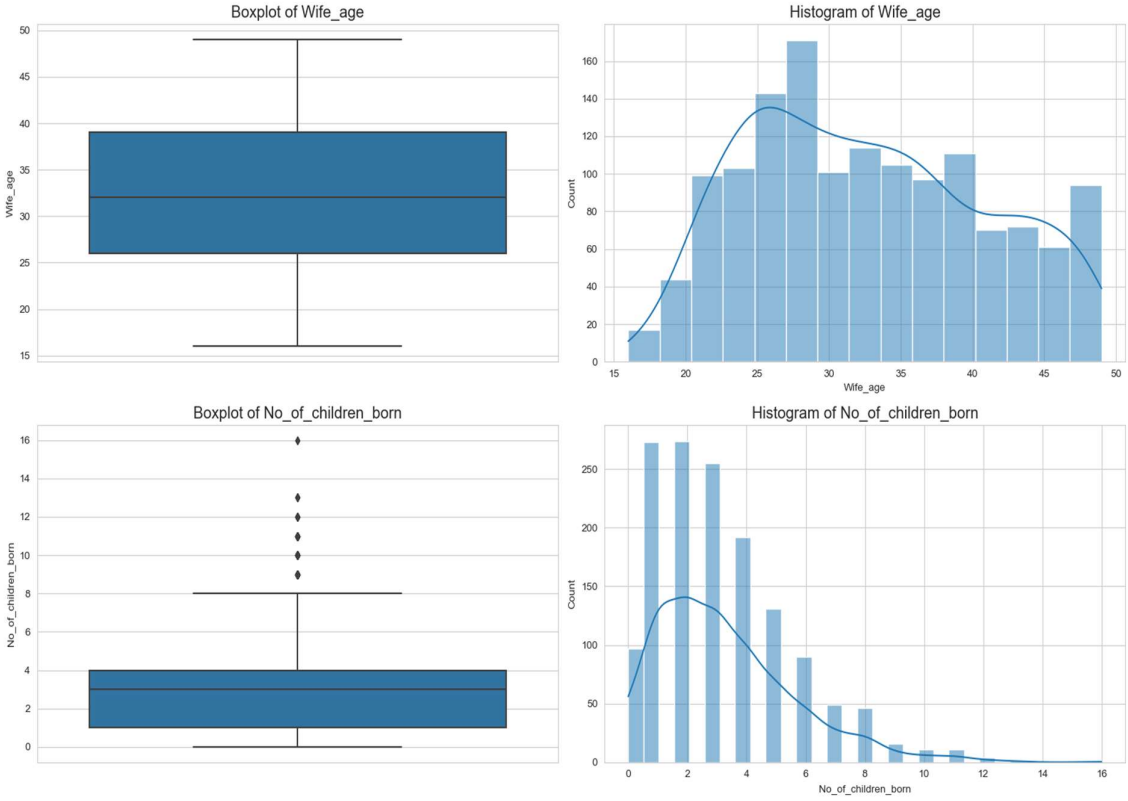

Figure 35

- The box plot confirms the statistical summary of the numerical variables.
- There are outliers present in No of children column.
- The wife age is slightly right skewed and the skewness for No of children born is comparatively significant.

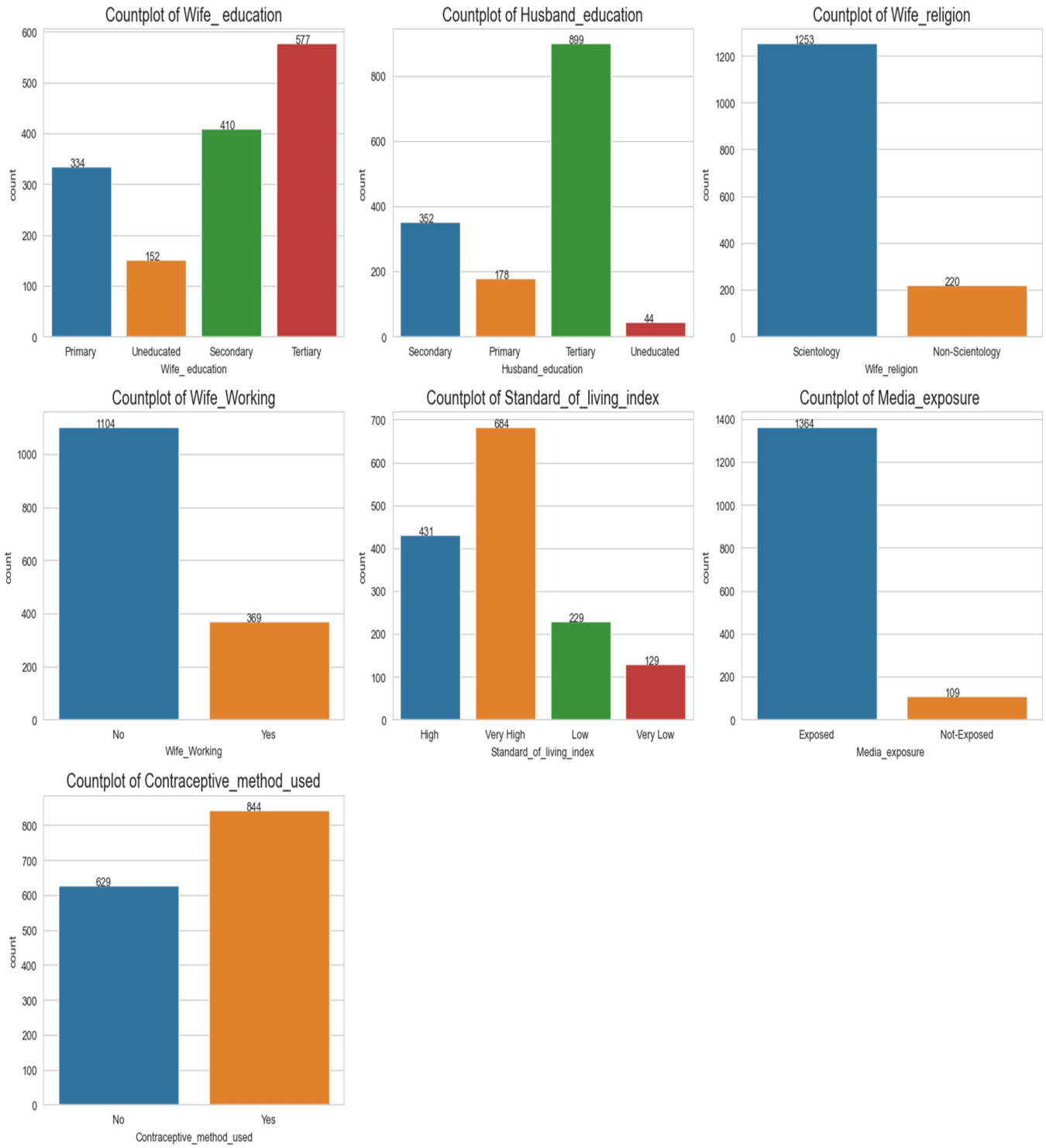
Countplot of Wife_ education
count
Wife_ education
Primary
Uneducated
Secondary
Tertiary
334
152
410
577
Countplot of Husband_education
count
Husband_education
Secondary
Primary
Tertiary
Uneducated
352
178
899
44
Countplot of Wife_religion
count
Wife_religion
Scientology
Non-Scientology
1253
220
Countplot of Wife_Working
count
Wife_Working
No
Yes
1104
369
Countplot of Standard_of_living_index
count
Standard_of_living_index
High
Very High
Low
Very Low
431
684
229
129
Countplot of Media_exposure
count
Media_exposure
Exposed
Not-Exposed
1364
109
Countplot of Contraceptive_method_used
count
Contraceptive_method_used
No
Yes
629
844

Figure 36

Education:

- The number of females who have completed the tertiary education is significantly higher as compared to other categories.
- The count is around 577 for tertiary followed by secondary with 410.
- This situation also applies to the husband's education.

Contraceptive Methods:

Most of the females have opted for contraceptive methods.

Employment Status:

- Majority of the wives, to be precise around 1104 are unemployed.

Living Standard:

- Very High: 684.
- High:431.
- Low:229.
- Very low: 129.

Media Exposure:

- The number of females who are exposed to media are 12.5 times higher than those who are not exposed to any form of media.

Bivariate Analysis:

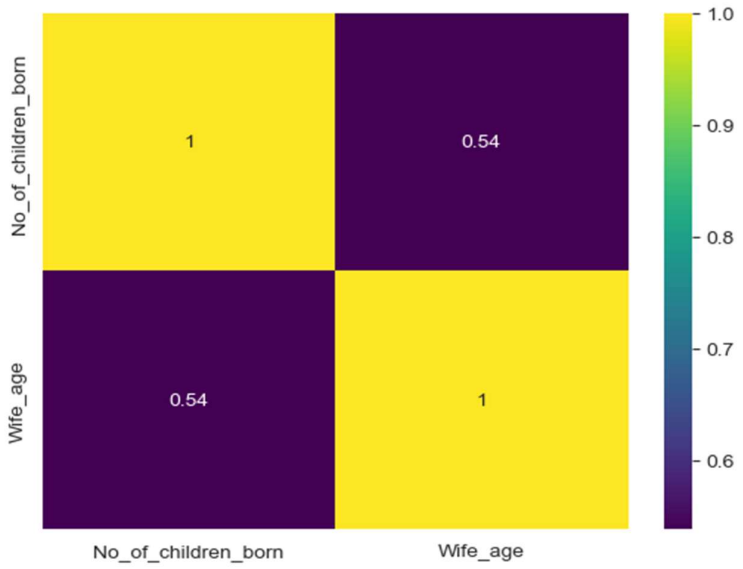

Figure 37

There is a slight correlation between wife age and No of children born but not that significant.

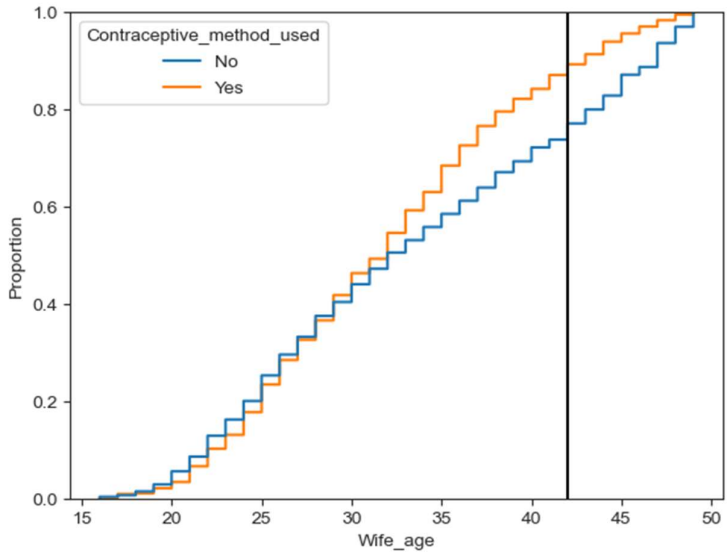

Figure 38

90% of the wives who opt for contraceptive measures are below the age of 42.

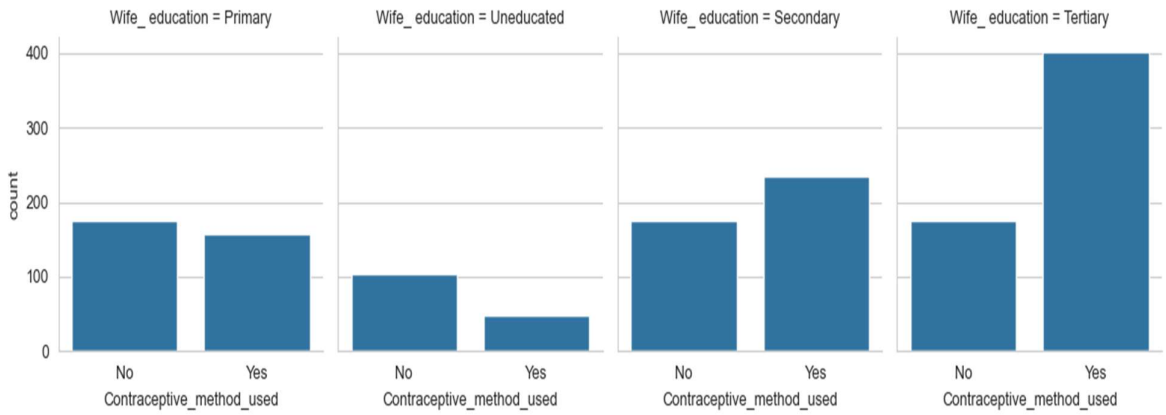

Figure 39

We can see that people with higher education tend to choose contraceptive methods as compared to those with primary education.

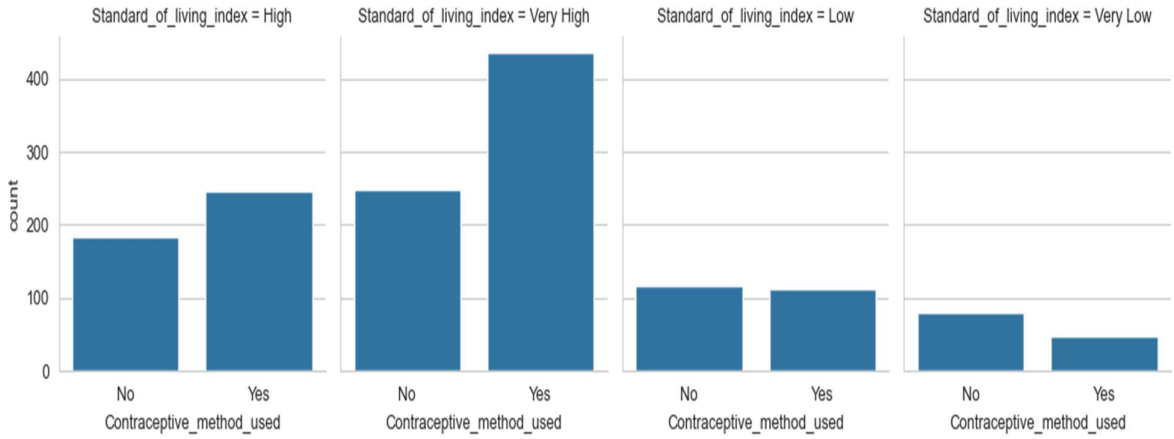

Figure 40

- The standard of living also plays a significant role in deciding whether to choose the contraceptive method or not.
- For individuals with a very high standard of living, there is a significant preference for using contraceptive methods.
- This could be due to various financial factors such a cost of the method which might be affordable to certain groups.

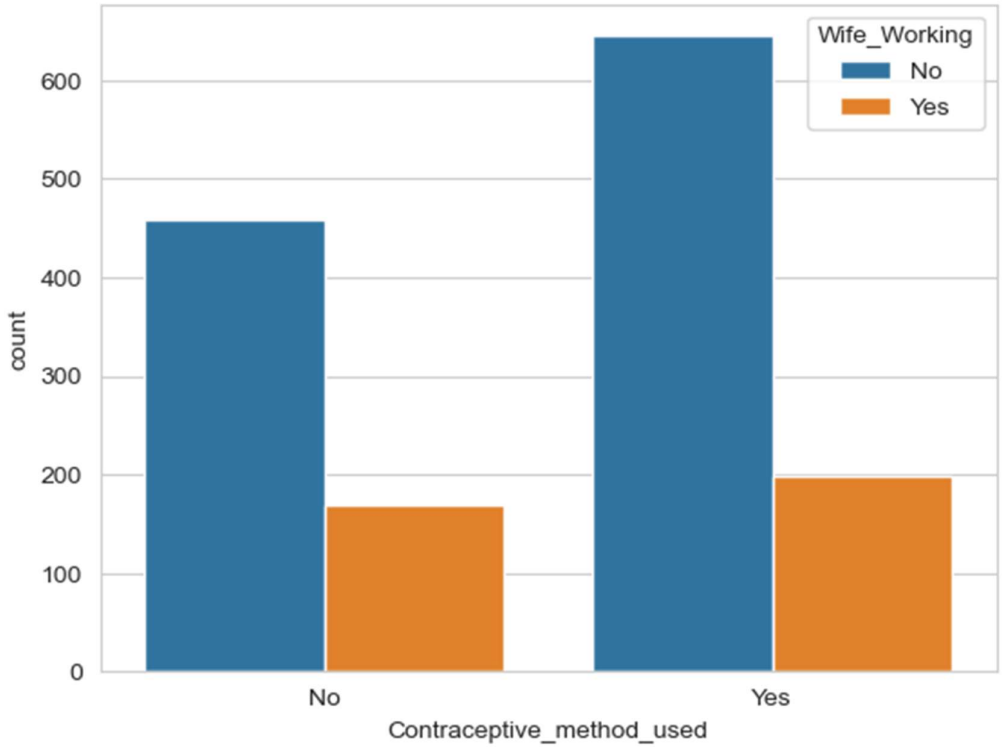

Figure 41

Wives Who Are Not Working:

- Approximately 450 wives who are not working do not use contraceptive methods.
- Around 640 wives who are not working use contraceptive methods.

Wives Who Are Working:

- Nearly 170 wives who are working do not use contraceptive methods, while about 190 opt for contraceptive methods.

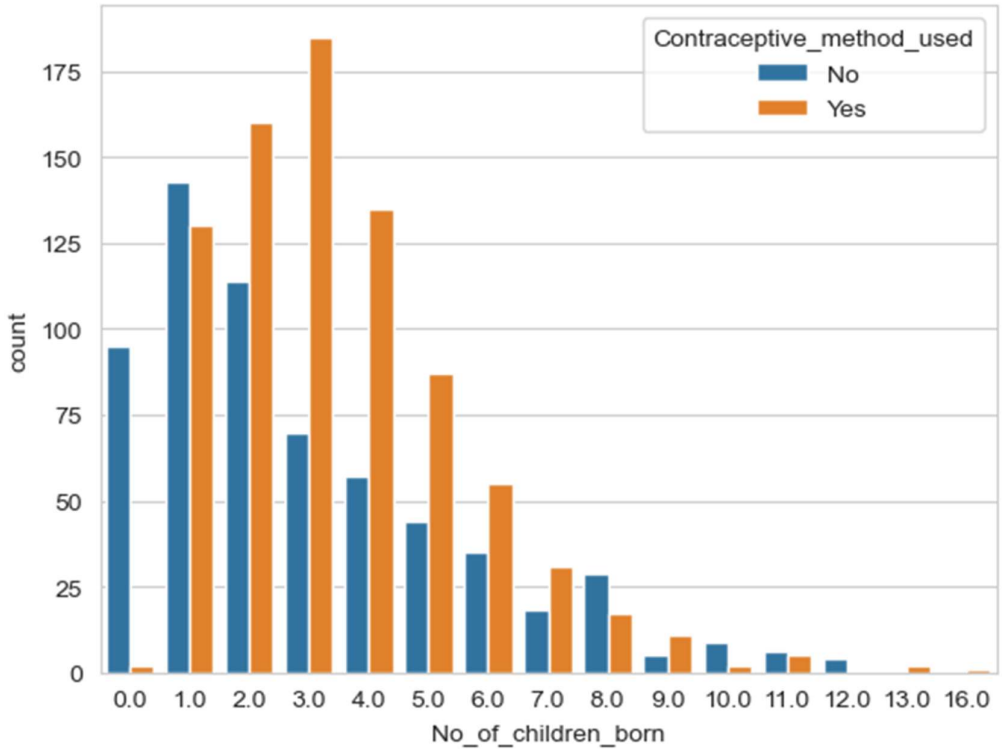

Figure 42

- Wives with 1 or no children seem to show hesitation while choosing the contraceptive methods.
- While females with 2-7 children prefer contraceptive methods.

Data Pre-processing:

There are null values present in wife age and No of children. Let's impute those values using the median of the respective columns.

Before Imputing

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

After Imputing

```
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
```

Figure 43

Converting Categorical Variables into Integers:

Wife's & Husband's Education:

- Uneducated-1
- Primary-2
- Secondary-3
- Tertiary-4

Wife's Religion:

- Non-Scientology-0
- Scientology-1

Wife Working & Contraceptive Method:

- Yes -1
- No- 0

Media Exposure:

- Exposed-1
- Not Exposed -0

Standard of living:

- Very low-1
- Low-2
- High-3
- Very high-4

After Imputation:

Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_method_used
1	0	2	3	1	0
1	0	3	4	1	0
1	0	3	4	1	0
1	0	3	3	1	0
1	0	3	2	1	0

Figure 44

Outliers:

An observation is considered to be an outlier if that particular observation has been mistakenly captured in the data set. Treating outliers sometimes results in the models having better performance but the models lose out on generalization.

We are proceeding without treating the outliers, as this information seems legit. Further the same should be conveyed to the stakeholders before proceeding.

Train-Test Split:

Let us split the dataset with predictor variable and dependent variable as a separate dataset.

- Shape of X Train: (975,9)
- Shape of X Test: (418,9)

Model Using Logistic Regression:

Classification Report for the Train data:

Classification Report

	precision	recall	f1-score	support
0	0.68	0.54	0.60	422
1	0.70	0.81	0.75	553
accuracy			0.69	975
macro avg	0.69	0.67	0.67	975
weighted avg	0.69	0.69	0.68	975

Figure 45

Classification Report for the Test data:

Classification Report	precision	recall	f1-score	support
0	0.64	0.45	0.53	192
1	0.63	0.78	0.69	226
accuracy			0.63	418
macro avg	0.63	0.62	0.61	418
weighted avg	0.63	0.63	0.62	418

Figure 46

- The accuracy score is around 69% for train data and 63% for test data.
- Our class of interest 1 (i.e.) people choosing contraceptive methods, has a recall of 0.78 and f1-score of 0.69 which is good.

Confusion Matrix:

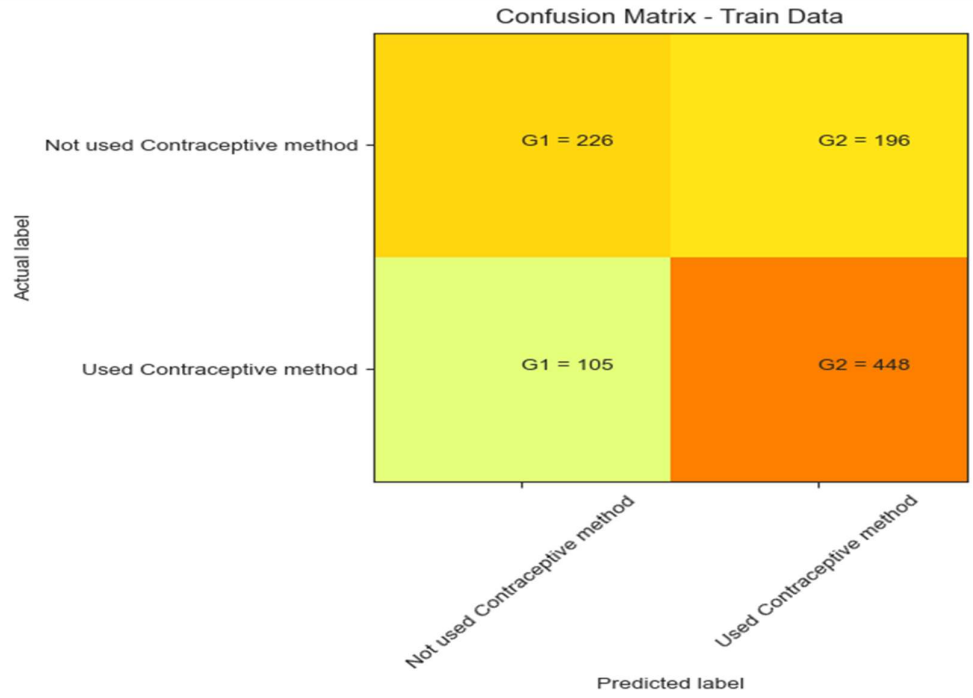

Figure 47

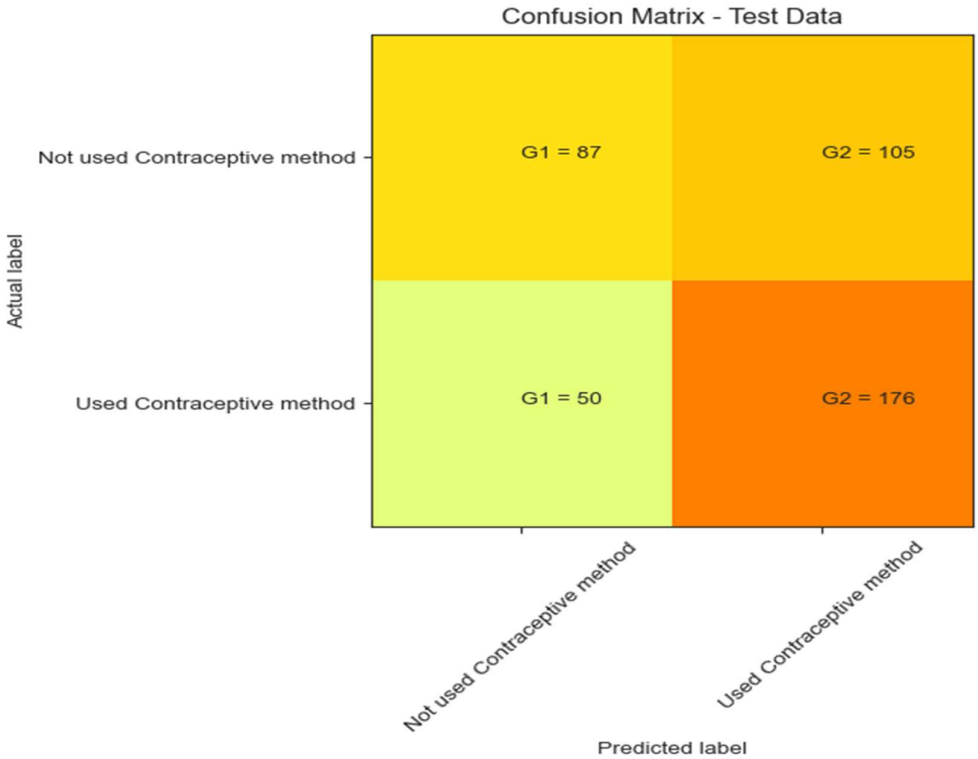

Figure 48

AUC-ROC Score:

For Train Data:

AUC score:0.725

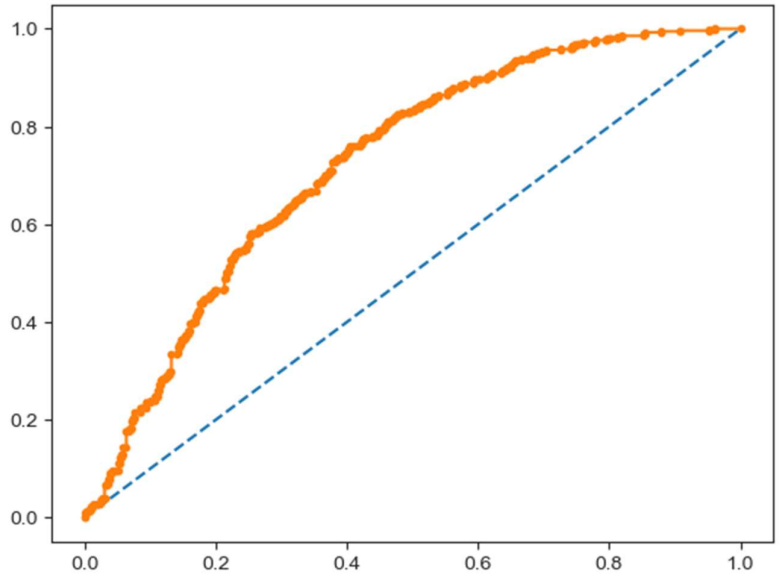

Figure 49

For Test Data:

AUC Score :0.657

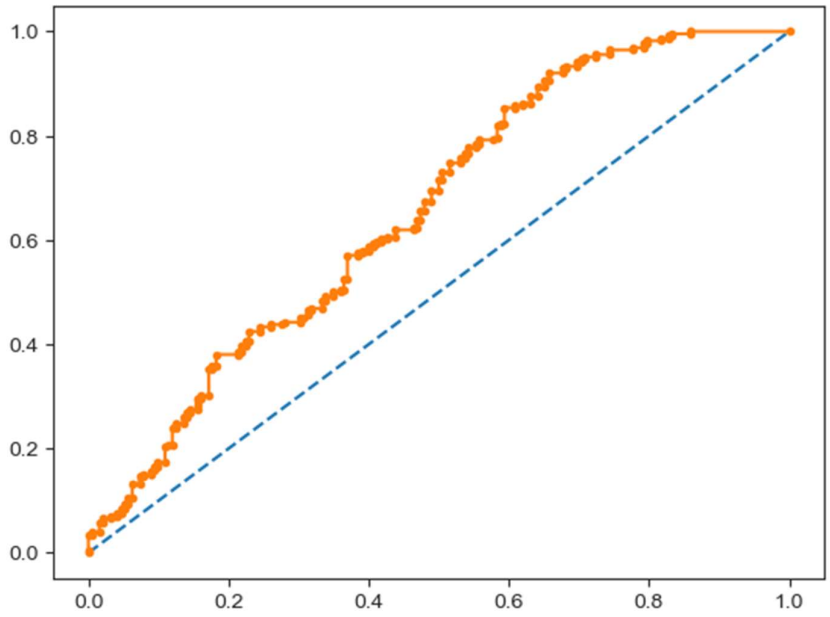

Figure 50

Model Using LDA:

Classification Report for the Train data:

Classification Report	precision	recall	f1-score	support
0	0.70	0.51	0.59	422
1	0.69	0.83	0.75	553
accuracy			0.69	975
macro avg	0.69	0.67	0.67	975
weighted avg	0.69	0.69	0.68	975

Figure 51

Classification Report for the Test data:

Classification Report	precision	recall	f1-score	support
0	0.63	0.43	0.51	192
1	0.62	0.79	0.69	226
accuracy			0.62	418
macro avg	0.62	0.61	0.60	418
weighted avg	0.62	0.62	0.61	418

Figure 52

- The model accuracy has been dropped from 69% for the train data to 62% for the testing data.
- The recall score for the train data for the class of interest is around 0.83 and it is 0.79 for test data.
- The f1 score is identical to the logistic regression model.
- LDF=-0.906+ X1*(-0.083) + X2*(0.489) + X3*(0.075) + X4*(0.354) + X5*(-0.426) + X6*(-0.022) + X7*(0.173) + X8*(0.299) + X9*(0.182)

Confusion Matrix:

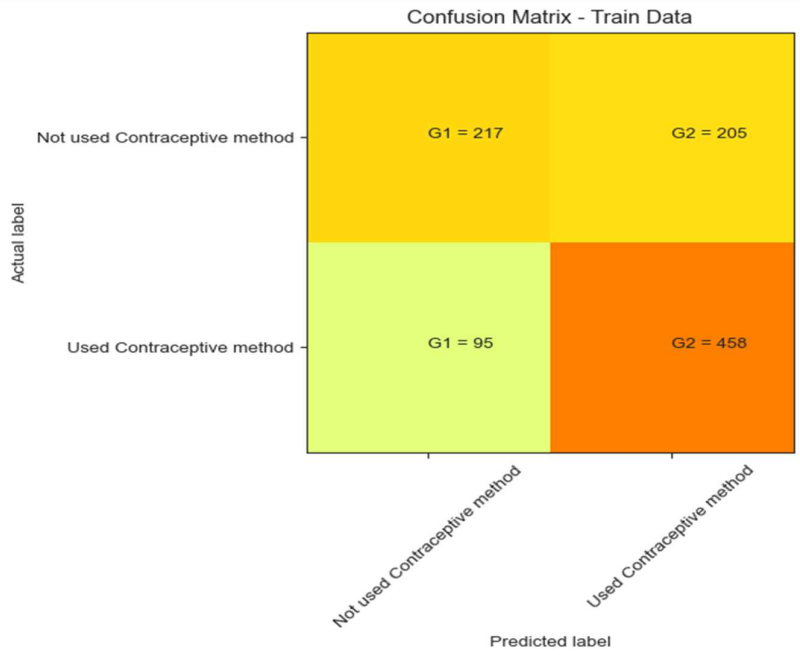

Figure 53

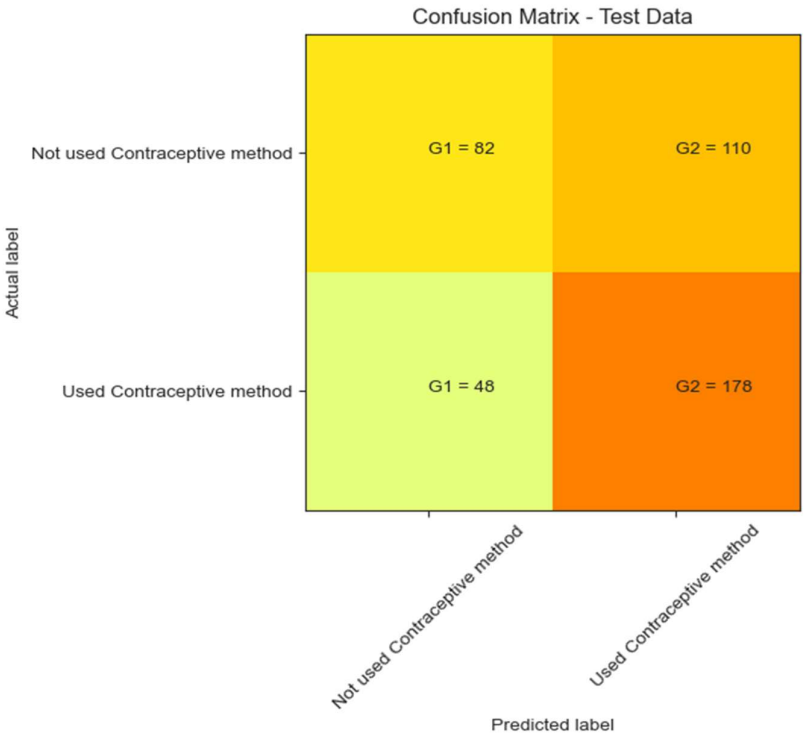

Figure 54

AUC-ROC Score:

For Train Data:

AUC score:0.724

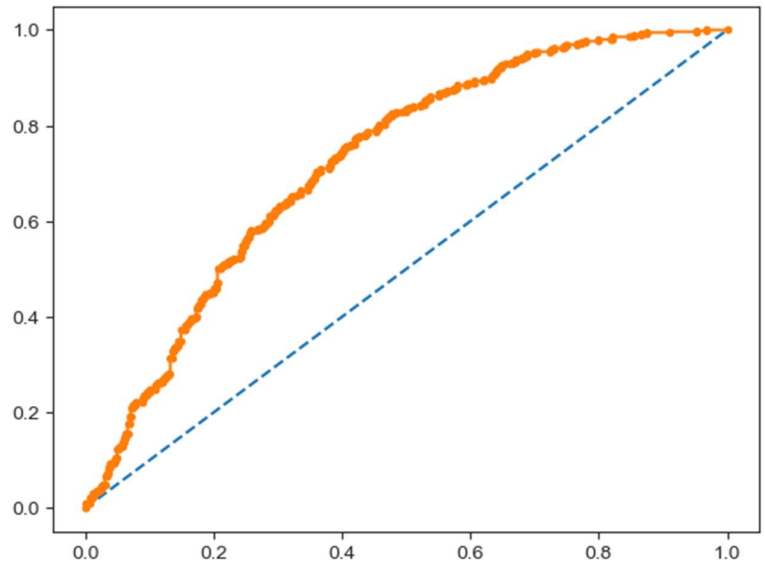

Figure 55

For Test Data:

AUC Score :0.656

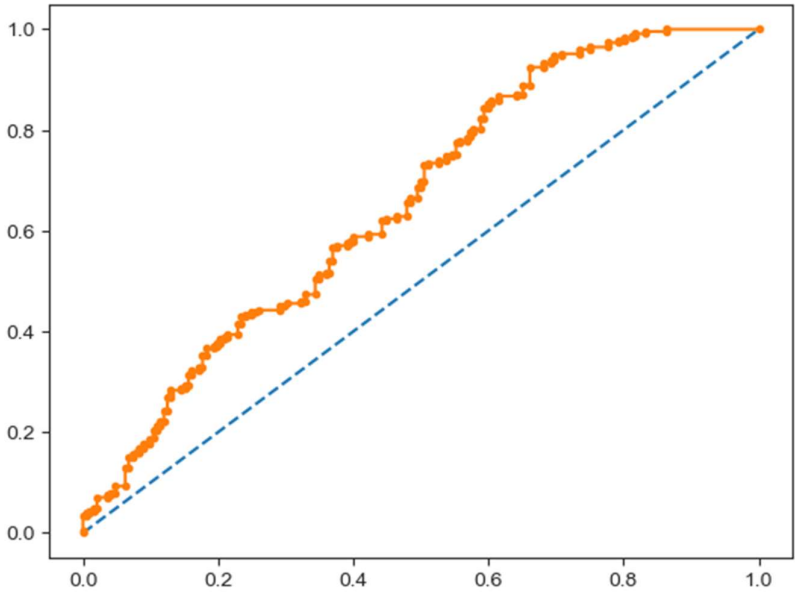

Figure 56

Model Using CART:

Classification Report for the Train data:

	precision	recall	f1-score	support
0	0.96	1.00	0.98	422
1	1.00	0.97	0.99	553
accuracy			0.98	975
macro avg	0.98	0.99	0.98	975
weighted avg	0.98	0.98	0.98	975

Figure 57

Classification Report for the Test data:

	precision	recall	f1-score	support
0	0.57	0.55	0.56	192
1	0.63	0.65	0.64	226
accuracy			0.60	418
macro avg	0.60	0.60	0.60	418
weighted avg	0.60	0.60	0.60	418

Figure 58

- The accuracy has significantly dropped from 98% to 60%.
- This could be an indication of overfitting.
- Overfitting occurs when the model learns the noise and random fluctuations in the training data rather than the underlying pattern. This means it doesn't generalize well to new, unseen data.

Confusion Matrix:

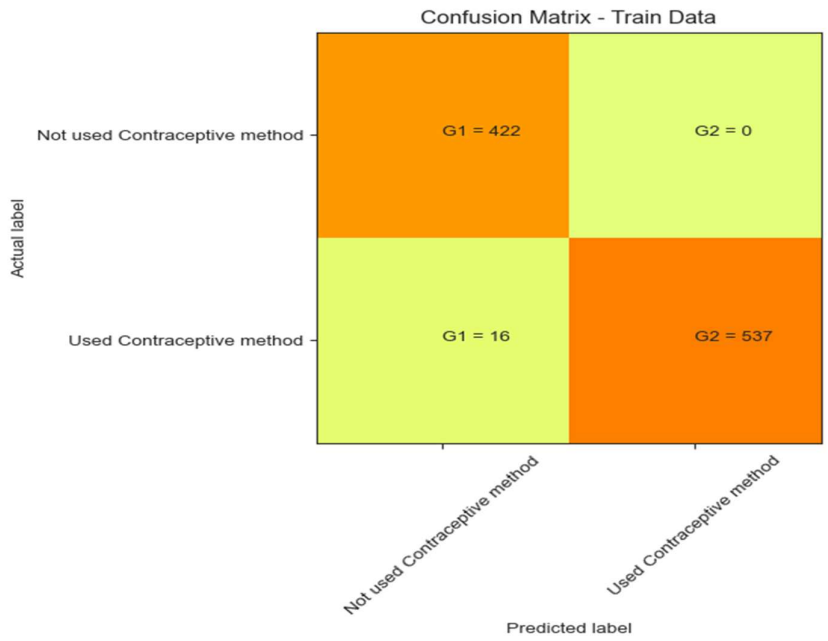

Figure 59

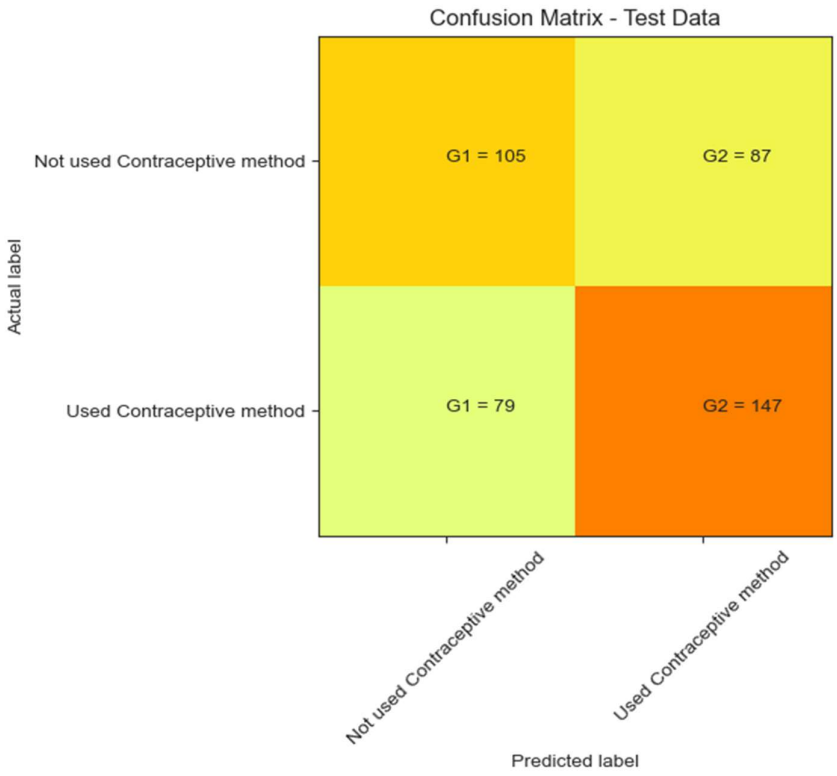

Figure 60

AUC-ROC Score:

For Train Data:

AUC score:0.99.

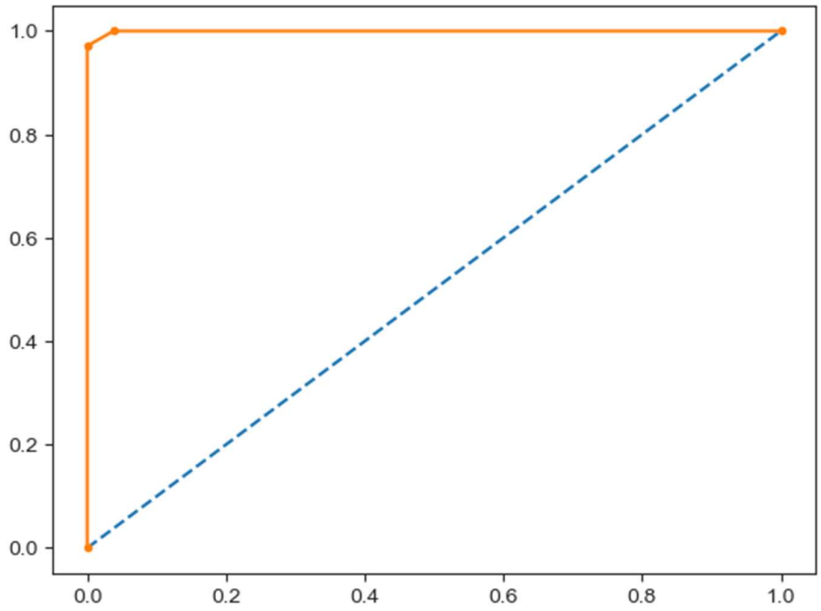

Figure 61

For Test Data:

AUC score:0.60

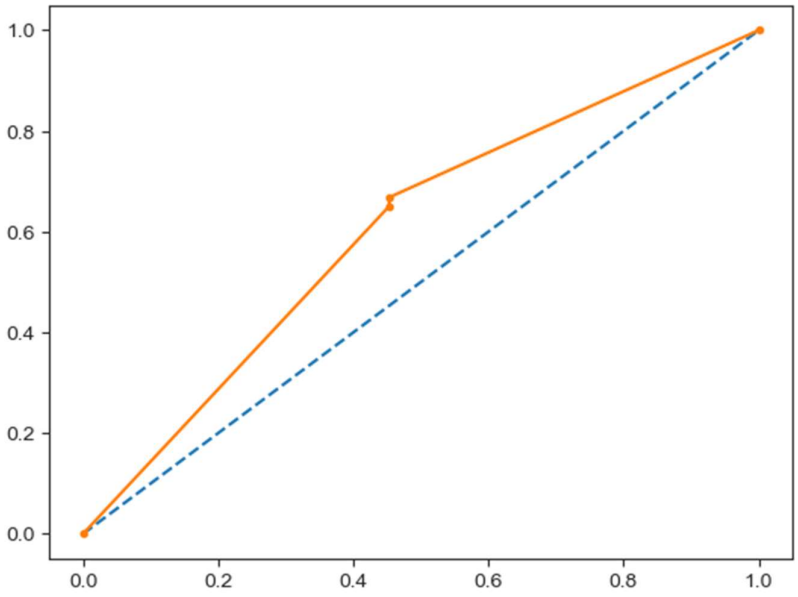

Figure 62

Tree Before Pruning:

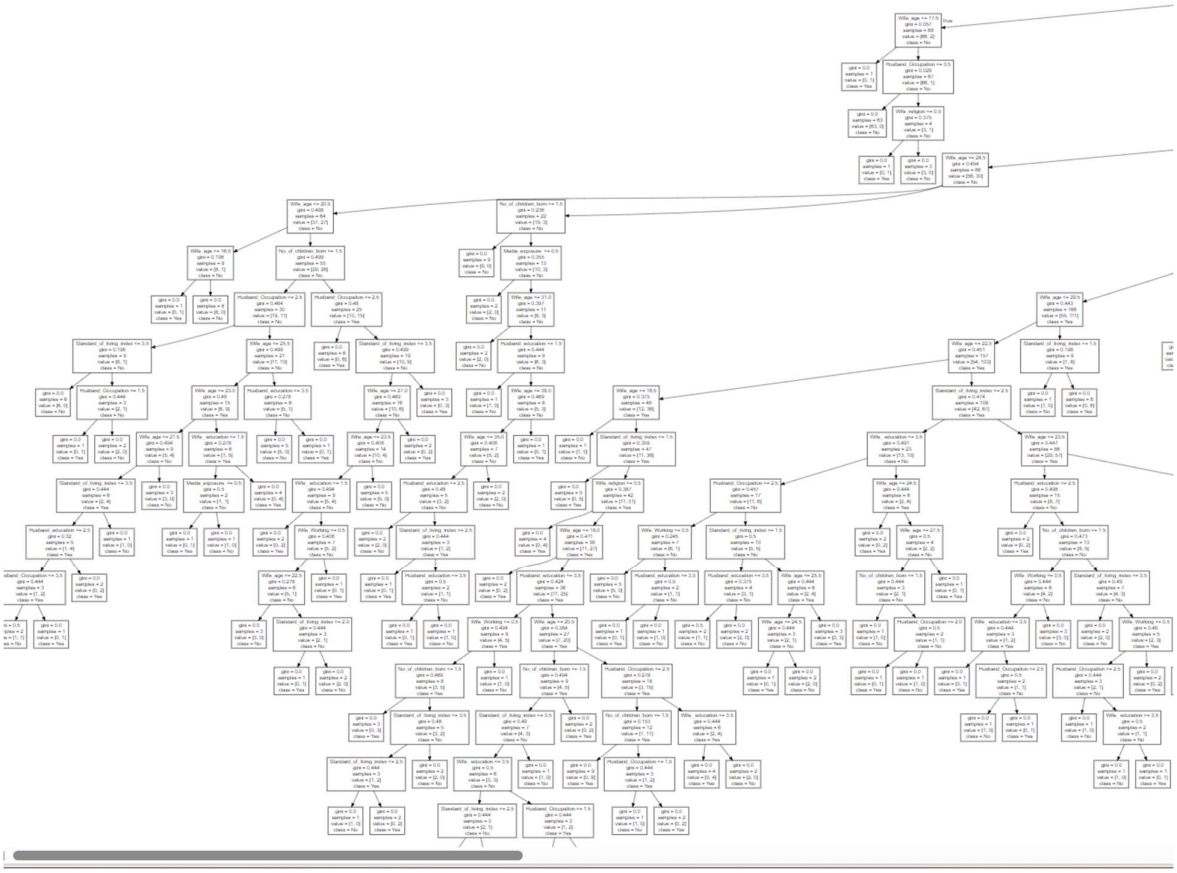

Figure 63

It is evident that the tree has overgrown, which leads to overfitting of the model.

Feature Importance before pruning:

	Imp
Wife_age	0.292197
No_of_children_born	0.240709
Wife_ education	0.106970
Husband_Occupation	0.106614
Standard_of_living_index	0.103483
Husband_education	0.051409
Wife_Working	0.048122
Wife_religion	0.033211
Media_exposure	0.017285

Figure 64

Regularising the Decision Tree:

Tuning the hyperparameter using GridSearchCV.

The best hyperparameters for the given model are:

```
DecisionTreeClassifier(ccp_alpha=0.001, max_depth=10, max_features='sqrt',
                       min_samples_leaf=15, min_samples_split=5,
                       random_state=1)
```

Figure 65

Classification Report for the train data after regularisation:

	precision	recall	f1-score	support
0	0.75	0.62	0.68	422
1	0.74	0.84	0.79	553
accuracy			0.74	975
macro avg	0.75	0.73	0.73	975
weighted avg	0.75	0.74	0.74	975

Classification Report for the test data after regularisation:

	precision	recall	f1-score	support
0	0.69	0.54	0.60	192
1	0.67	0.79	0.72	226
accuracy			0.67	418
macro avg	0.68	0.66	0.66	418
weighted avg	0.68	0.67	0.67	418

Figure 66

For Test Data:

AUC score:0.744

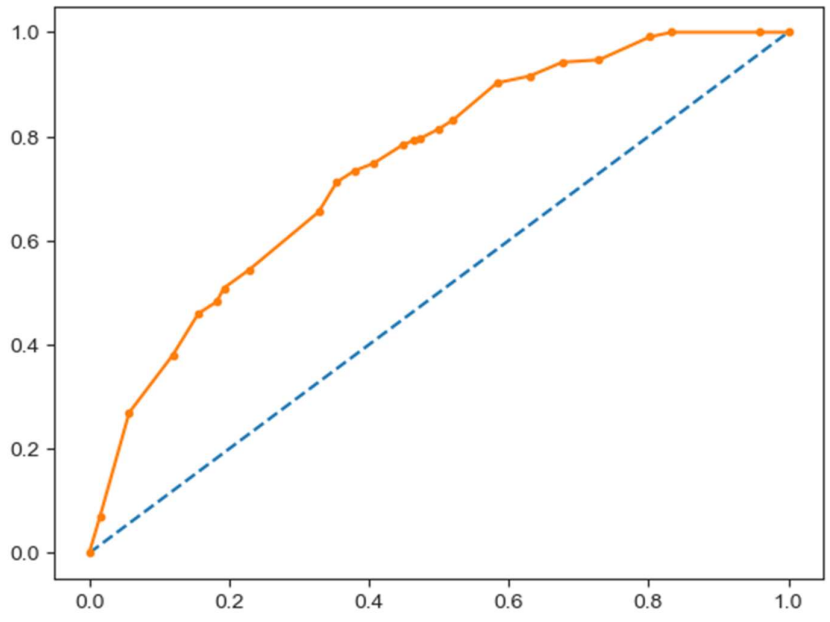

Figure 67

Confusion Matrix:

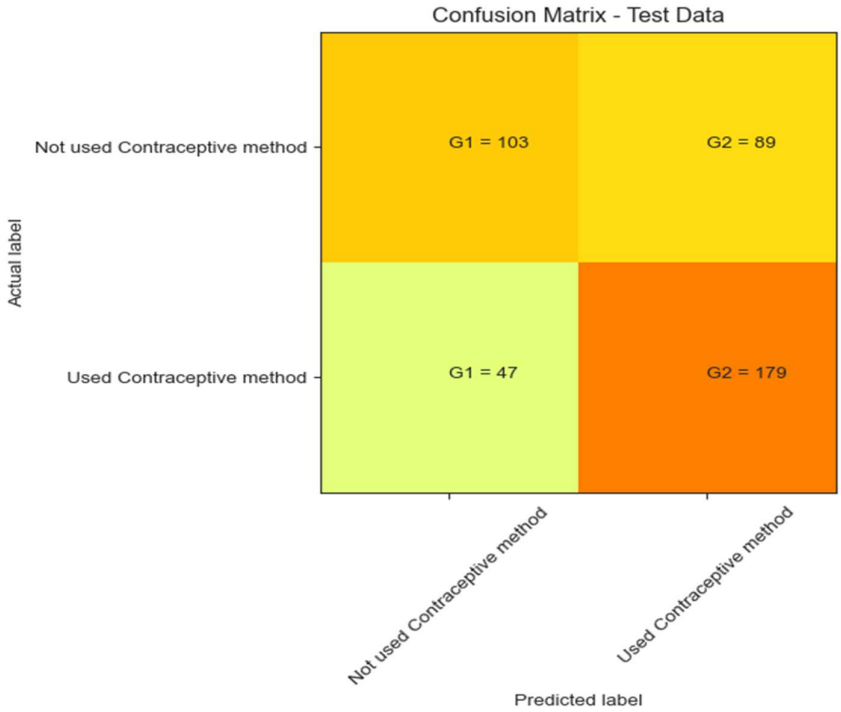

Figure 68

Tree After Regularisation:

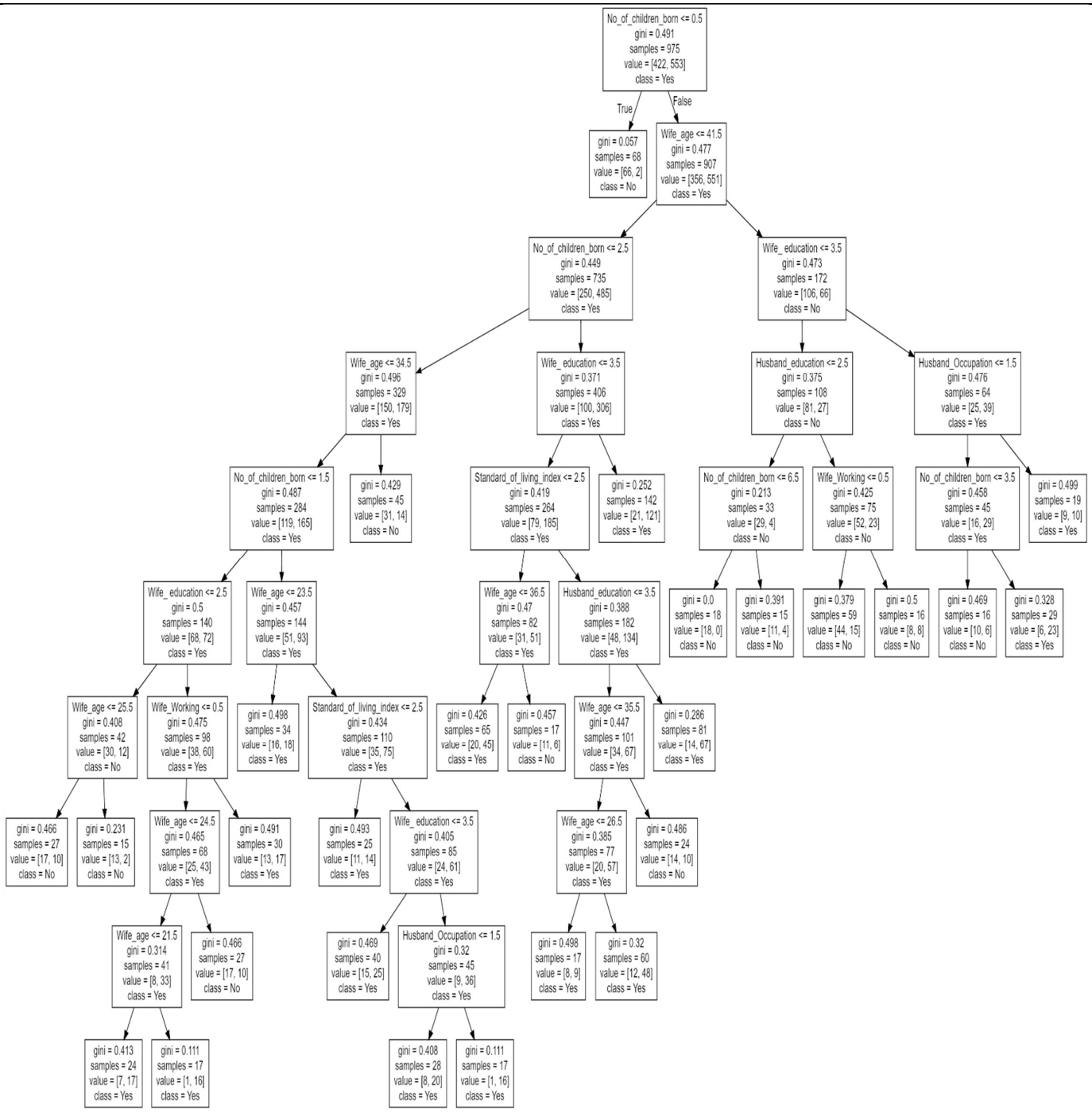

Figure 69

Comparison Between Models:

Model Metrics	Scores
logistic Regression:	
Accuracy	0.63
Recall	0.78
precision	0.63
F1-Score	0.69
AUC Score	0.65
LDA:	
Accuracy	0.62
Recall	0.79
precision	0.62
F1-Score	0.69
AUC Score	0.65
CART:	
Accuracy	0.67
Recall	0.79
precision	0.67
F1-Score	0.72
AUC Score	0.74

Table 3

- Among the three models, the CART model has a better metrics.
- Further as compared to the other models, CART is easy to interpret.
- CART model can automatically detect the most important variable as compared to LDA & logistic Regression.
- They are not sensitive to outliers.
- Multicollinearity has less effect on CART model comparatively.

For these reasons, we are proceeding with the CART model.

Feature Importance:

	Imp
No_of_children_born	0.459324
Wife_age	0.317840
Wife_ education	0.155576
Husband_education	0.027978
Standard_of_living_index	0.017100
Wife_Working	0.011926
Husband_Occupation	0.010256
Wife_religion	0.000000
Media_exposure	0.000000

Figure 70

Business Insights & Recommendations:

- The most important variables in predicting the class of interest are the No of children born followed by the age of the wife.
- Wives with no children seem to show hesitation while choosing the contraceptive methods. While females with 2-7 children prefer contraceptive methods. So, our target audience is families with children.
- 90% of the wives who opt for contraceptive measures are below the age of 42.
- It is evident that the education level of the wives and their respective partners play a significant role. Partners with higher education tend to opt for contraceptive methods, especially those with tertiary education. Raising awareness among people with lower education could bring a significant change in the numbers.
- People with higher standard of living prefer contraceptive measures. In order to attract people with low income, the treatment cost can be made affordable so that more people will prefer contraceptive measures.

- Wife religion and media exposure has no effect.
- Working professionals tend to prefer contraceptive measure as compared with those who are unemployed.