PM PROJECT

Business report.

Philjoy Dsilva 14-Jan-2023 PGPDSBA

Contents

List of Tables

List of Figures

Problem 1

Comp-activ:

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, you 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.

Data Description :
System measures used:

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.
exec - Number of system exec calls 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
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.

<u>Introduction:</u>
The comp-activ database offers a comprehensive snapshot of computer system activity, sourced from a Sun Sparcstation 20/712 in a multi-user university department. Our aim, as aspiring data scientists, is to establish a predictive linear equation for the 'usr' mode—revealing the percentage of time CPUs operate in user mode. This analysis explores the diverse system attributes influencing this behaviour.

<u>Executive Summary:</u>
This analysis dives into the comp-activ database to decode the intricate relationship between system attributes and the 'usr' mode. Through univariate and multivariate analyses, we aim to construct a predictive model, offering insights crucial for system optimization and resource management. The findings empower data science enthusiasts and system administrators with actionable knowledge.

1.1 : Define the problem and perform exploratory Data Analysis

Ans –

Problem definition

Let's embark on an intriguing journey through the comp-activ database, a fascinating collection of computer system activity from a Sun Sparcstation. Our mission? To demystify the 'usr' mode—understanding how much time CPUs spend in user mode. This adventure is all about connecting the dots between different system features and the 'usr' mode, unraveling the secrets of computer system behavior.

Table 1 - Dataset

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	frees
0	1	0	2147	79	68	0.2	0.20	40671.0	53995.0	0.00	...	0.00	0.0	1.60	2.60	16.00	26.40	CPU_Bound	4670	1730
1	0	0	170	18	21	0.2	0.20	448.0	8385.0	0.00	...	0.00	0.0	0.00	0.00	15.63	16.83	Not_CPU_Bound	7278	1869
2	15	3	2162	159	119	2.0	2.40	NaN	31950.0	0.00	...	0.00	1.2	6.00	9.40	150.20	220.20	Not_CPU_Bound	702	1021
3	0	0	160	12	16	0.2	0.20	NaN	8670.0	0.00	...	0.00	0.0	0.20	0.20	15.60	16.80	Not_CPU_Bound	7248	1863
4	5	1	330	39	38	0.4	0.40	NaN	12185.0	0.00	...	0.00	0.0	1.00	1.20	37.80	47.60	Not_CPU_Bound	633	1760
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
8187	16	12	3009	360	244	1.6	5.81	405250.0	85282.0	8.02	...	55.11	0.6	35.87	47.90	139.28	270.74	CPU_Bound	387	986
8188	4	0	1596	170	146	2.4	1.80	89489.0	41764.0	3.80	...	0.20	0.8	3.80	4.40	122.40	212.60	Not_CPU_Bound	263	1055
8189	16	5	3116	289	190	0.6	0.60	325948.0	52640.0	0.40	...	0.00	0.4	28.40	45.20	60.20	219.80	Not_CPU_Bound	400	969
8190	32	45	5180	254	179	1.2	1.20	62571.0	29505.0	1.40	...	18.04	0.4	23.05	24.25	93.19	202.81	CPU_Bound	141	1022
8191	2	0	985	55	46	1.6	4.80	111111.0	22256.0	0.00	...	0.00	0.2	3.40	6.20	91.80	110.00	CPU_Bound	659	1756

8192 rows × 22 columns

Check shape

Shape – (8192, 22)

Total Rows –8192 Total Columns – 22

Table 2 - Shape and info

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8192 entries, 0 to 8191
Data columns (total 22 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   lread     8192 non-null   int64
 1   lwrite    8192 non-null   int64
 2   scall     8192 non-null   int64
 3   sread     8192 non-null   int64
 4   swrite    8192 non-null   int64
 5   fork      8192 non-null   float64
 6   exec      8192 non-null   float64
 7   rchar     8088 non-null   float64
 8   wchar     8177 non-null   float64
 9   pgout     8192 non-null   float64
 10  ppgout    8192 non-null   float64
 11  pgfree    8192 non-null   float64
 12  pgscan    8192 non-null   float64
 13  atch      8192 non-null   float64
 14  pgin      8192 non-null   float64
 15  ppgin     8192 non-null   float64
 16  pflt      8192 non-null   float64
 17  vflt      8192 non-null   float64
 18  runqsz    8192 non-null   object
 19  freemem   8192 non-null   int64
 20  freeswap  8192 non-null   int64
 21  usr       8192 non-null   int64
dtypes: float64(13), int64(8), object(1)
memory usage: 1.4+ MB
```

Data types

Float Datatype – 13

Object Datatype – 1

Int Datatype – 8

Statistical summary

Table 3 - Statistical summary

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgfree
count	8192.000000	8192.000000	8192.000000	8192.000000	8192.000000	8192.000000	8192.000000	8.088000e+03	8.177000e+03	8192.000000	...	8192.000000
mean	19.559692	13.106201	2306.318237	210.479980	150.058228	1.884554	2.791998	1.973857e+05	9.590299e+04	2.285317	...	11.919712
std	53.353799	29.891726	1633.617322	198.980146	160.478980	2.479493	5.212456	2.398375e+05	1.408417e+05	5.307038	...	32.363520
min	0.000000	0.000000	109.000000	6.000000	7.000000	0.000000	0.000000	2.780000e+02	1.498000e+03	0.000000	...	0.000000
25%	2.000000	0.000000	1012.000000	86.000000	63.000000	0.400000	0.200000	3.409150e+04	2.291600e+04	0.000000	...	0.000000
50%	7.000000	1.000000	2051.500000	166.000000	117.000000	0.800000	1.200000	1.254735e+05	4.661900e+04	0.000000	...	0.000000
75%	20.000000	10.000000	3317.250000	279.000000	185.000000	2.200000	2.800000	2.678288e+05	1.061010e+05	2.400000	...	5.000000
max	1845.000000	575.000000	12493.000000	5318.000000	5456.000000	20.120000	59.560000	2.526649e+06	1.801623e+06	81.440000	...	523.000000

Univariate analysis

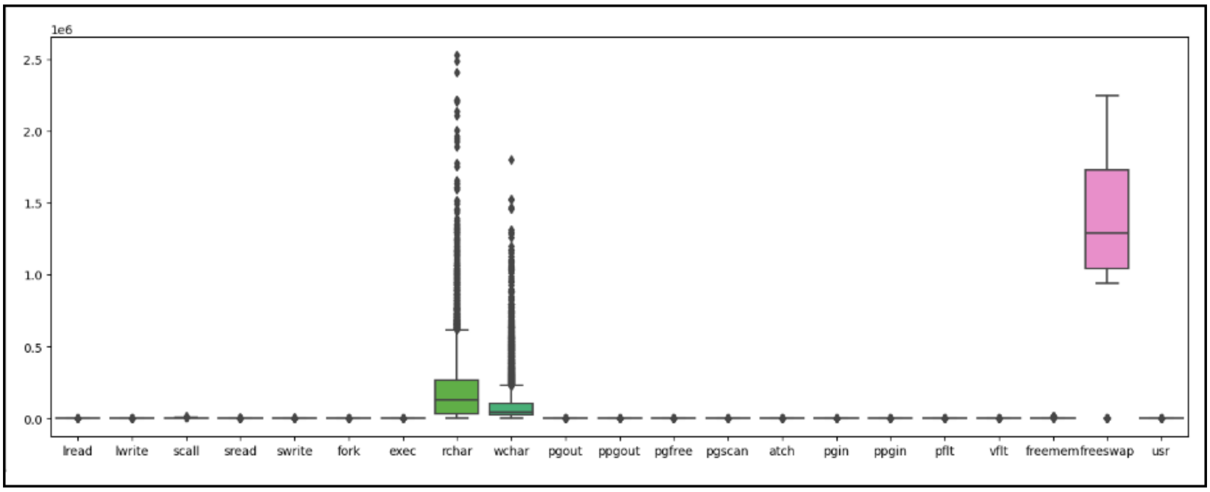

Figure 1 - Box plot of variables

We have outliers for our columns Rchar, wchar

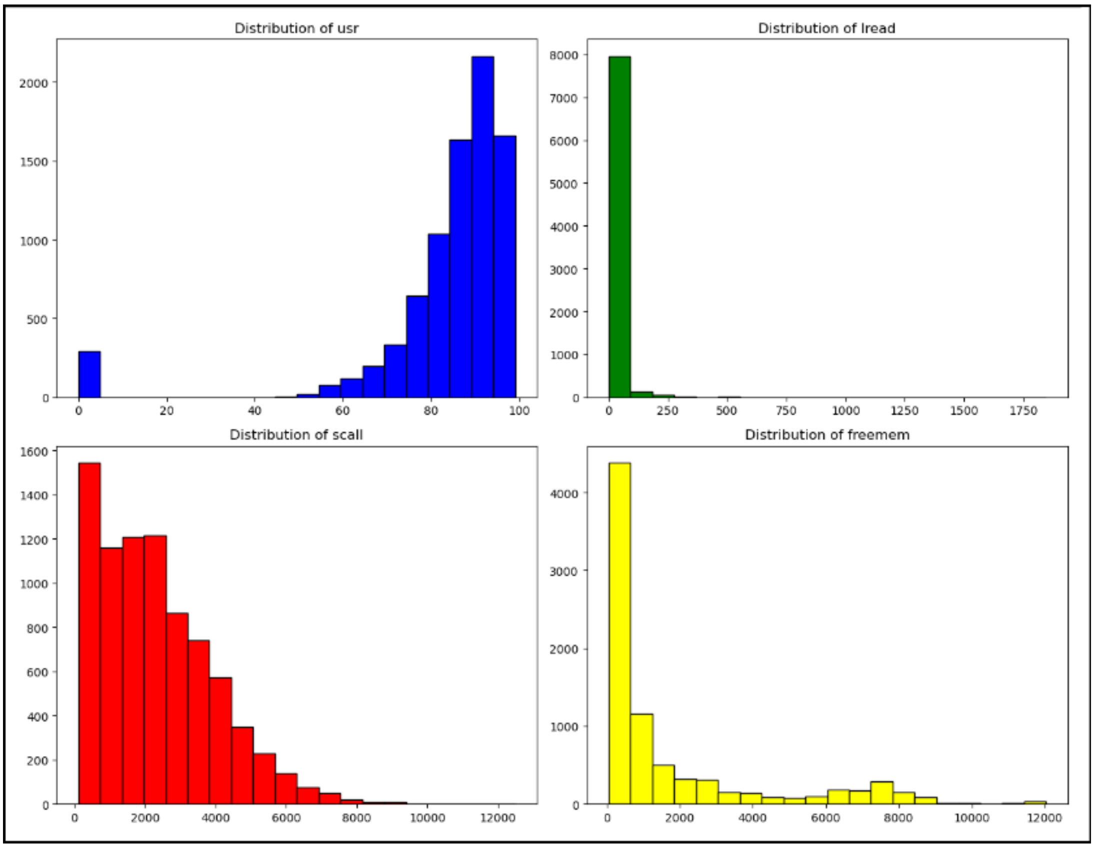

Figure 2 - Histograms for Individual Variables

Scall and Freemem are right skewed

USR is is left skewed

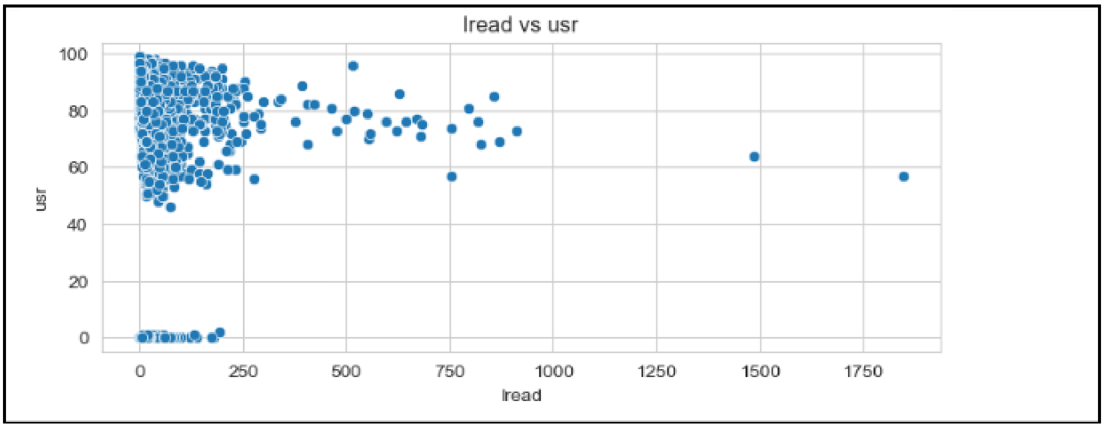

Figure 3 -lread vs usr

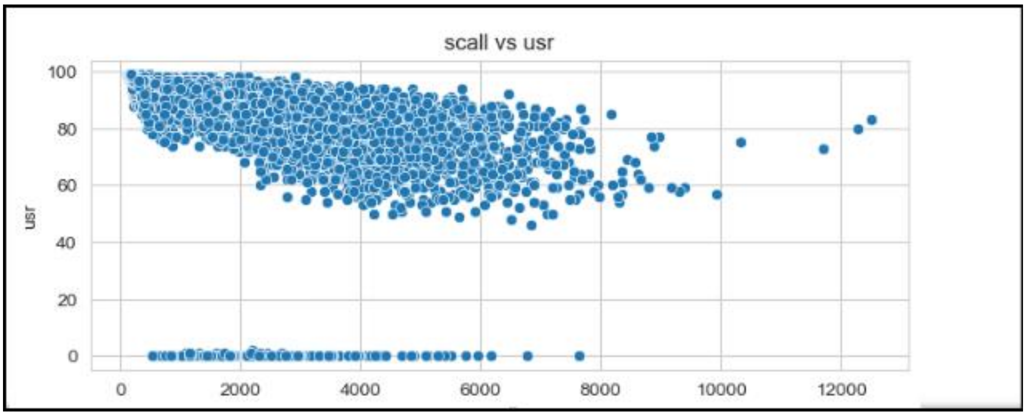

Figure 4 - scall vs usr

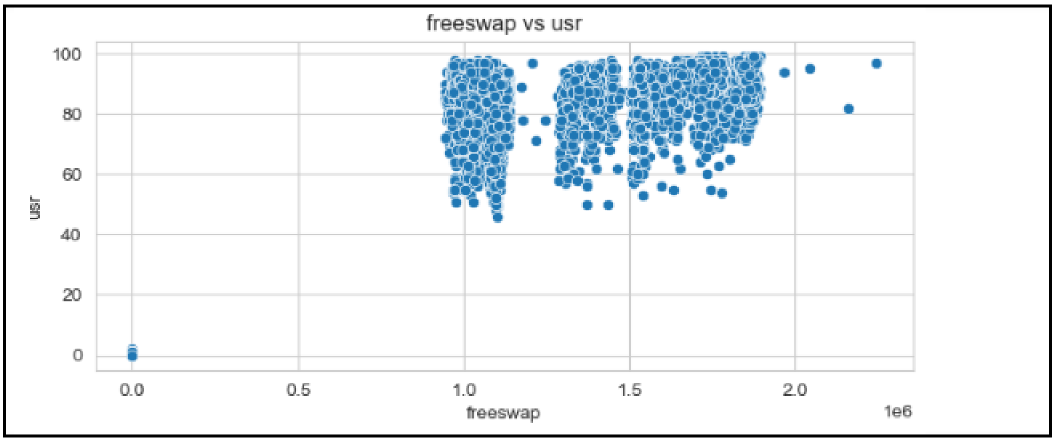

Figure 5 - freeswap vs usr

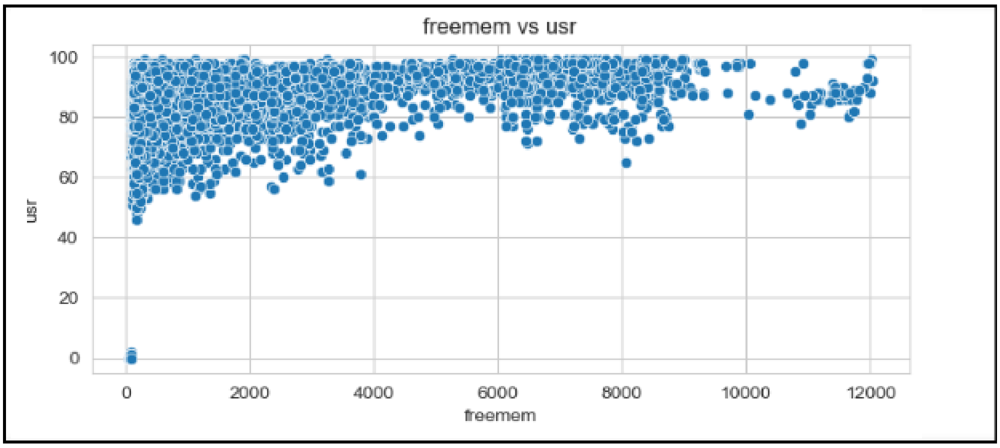

Figure 6 - freemen vs usr

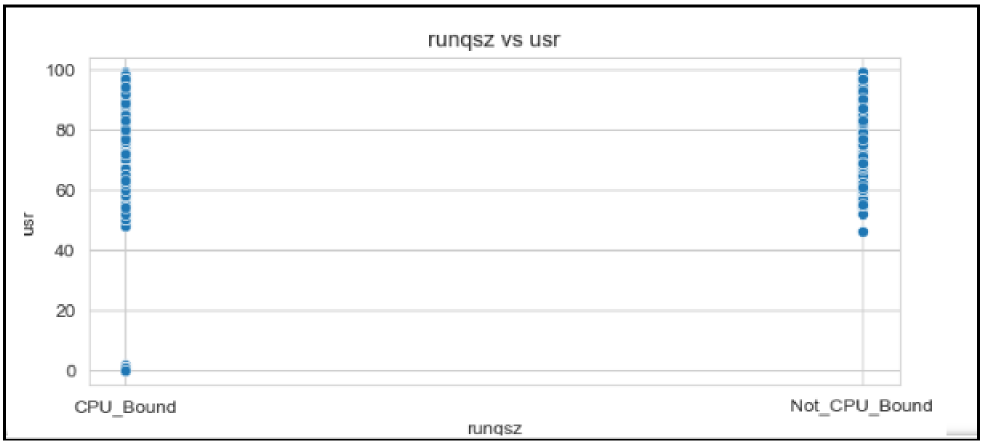

Figure 7 - runqsz vs usr

<u>Findings -</u>

- A noticeable observation is the presence of linear relationships among all variables, with the exception of freeswap and runqsz.

Multivariate analysis

The image can be zoomed for better view.

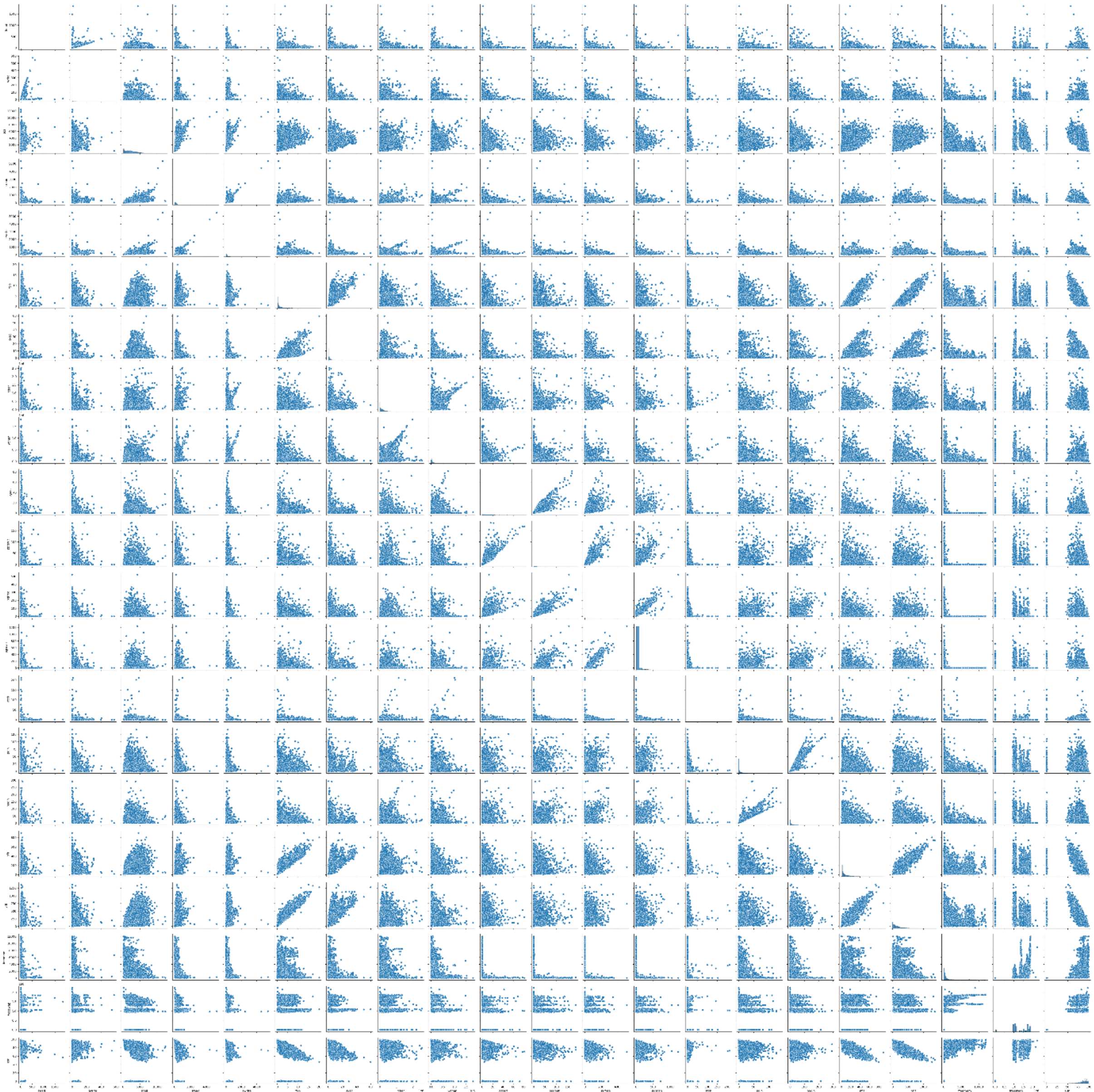

Figure 8 - pair plot of multiple variables

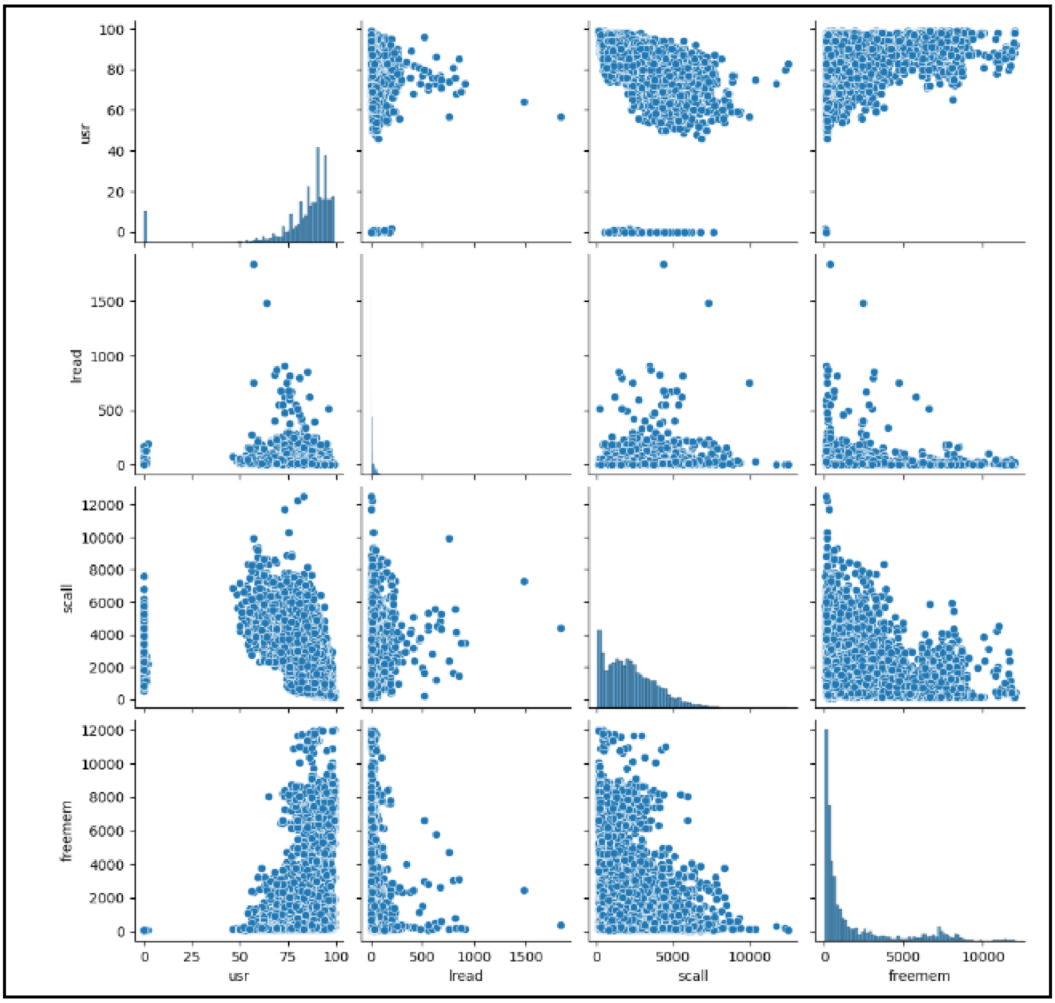

Figure 9 - pairplot 2

- The visual representation in the plots highlights the prevalence of zero values across all variables.
- It is interesting to note that all variables exhibit higher density at elevated rates of usr.

Use appropriate visualizations to identify the patterns and insights

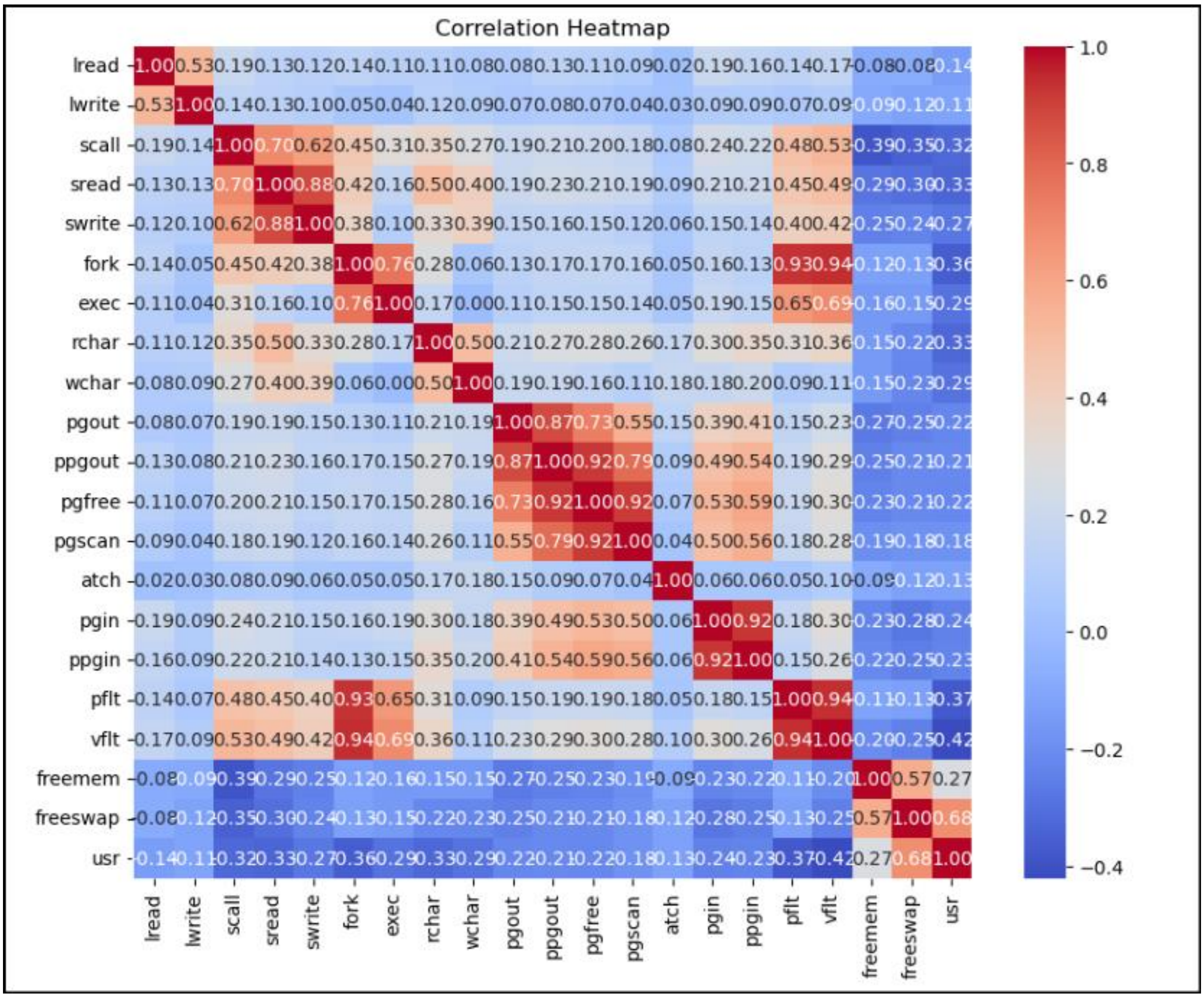

Figure 10 - heatmap of variables

With the analysis we can see the presence of corelation

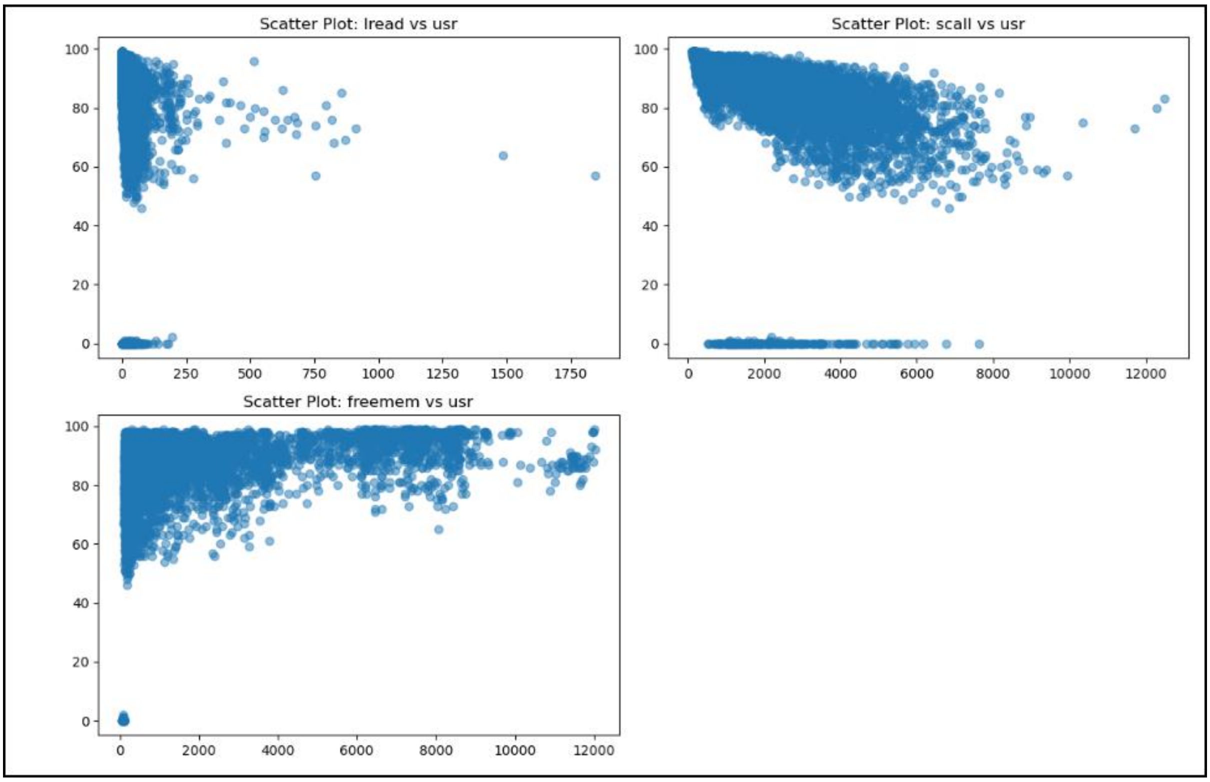

Figure 11 - Scatterplot of variables

As we can see the correlation is week between the variables

Key meaningful observations on individual variables and the relationship between variables

- The dataset encompasses both categorical and numerical values.
- With a total of 8192 rows and 22 columns, only 1 column is of object type, 8 columns are of integer type, and the remaining 13 are of float type.
- Our target variable is 'usr,' while all others serve as predictor variables.
- Upon delving into the univariate analysis, it becomes evident that there are outliers requiring attention.
- Bivariate and multivariate analyses reveal a robust positive correlation between the target variable 'usr' and the predictor variables 'freemem' and 'freeswap.'
- It's worth noting that there are no duplicate records in the provided dataset.

1.2 Data Pre-processing

Prepare the data for modelling: - Missing Value Treatment (if needed)

Table 4 - Null Values

```
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
```

Missing value in Rchar and Wchar

We will treat it using median

Table 5 - After treating null values

```
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)

We can see multiple outliers and we need to remove then for further analysis

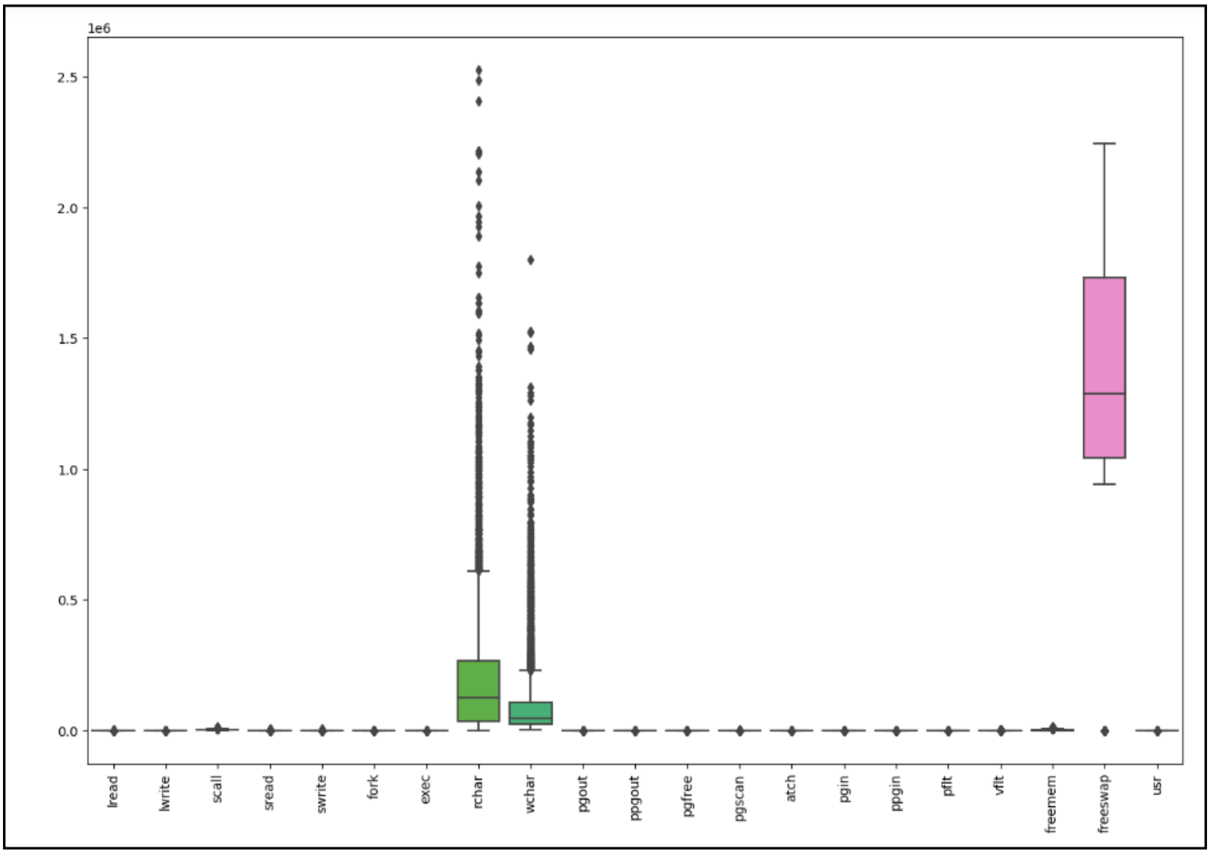

Figure 12 - before outlier treatment

After outlier Treatment

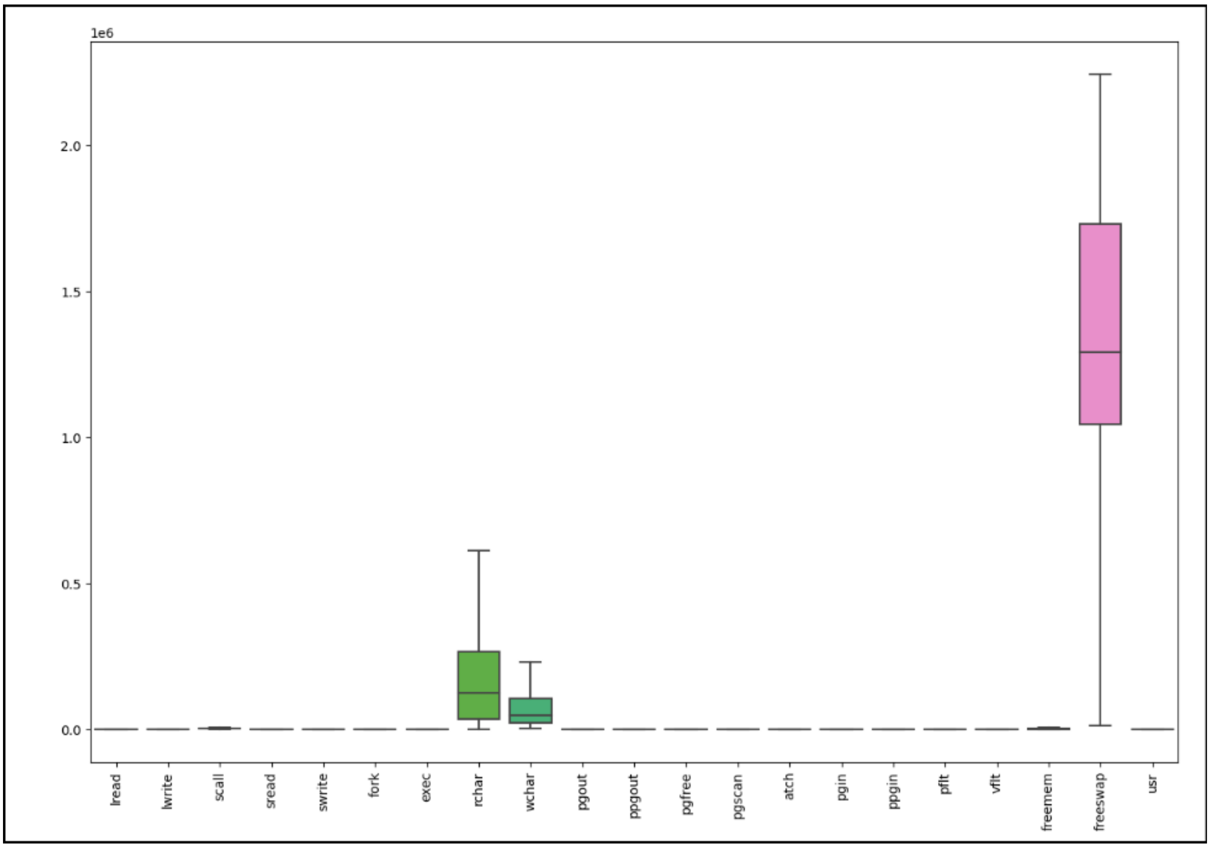

Figure 13 - after outlier treatment

Feature Engineering - Encode the data - Train-test split

Table 6 - encode data

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	ppgout	pgfree	atch	pgin
lread	1.000000	0.834674	0.333572	0.326032	0.306542	0.365676	0.364062	0.255544	0.174410	0.208258	0.219798	0.214400	0.230628	0.283252
lwrite	0.834674	1.000000	0.140213	0.148028	0.132097	0.093166	0.121778	0.117504	0.132595	0.090726	0.089149	0.082519	0.126769	0.113368
scall	0.333572	0.140213	1.000000	0.763001	0.742206	0.474491	0.440010	0.386875	0.331933	0.297487	0.305507	0.300612	0.306359	0.335380
sread	0.326032	0.148028	0.763001	1.000000	0.876652	0.528047	0.369723	0.576127	0.415059	0.301658	0.315648	0.311863	0.292044	0.346654
swrite	0.306542	0.132097	0.742206	0.876652	1.000000	0.519148	0.313323	0.419759	0.430126	0.274708	0.284982	0.280082	0.264283	0.307420
fork	0.365676	0.093166	0.474491	0.528047	0.519148	1.000000	0.774268	0.370995	0.122990	0.197047	0.214192	0.216166	0.201768	0.247715
exec	0.364062	0.121778	0.440010	0.369723	0.313323	0.774268	1.000000	0.324990	0.124145	0.230712	0.248680	0.250468	0.252416	0.299591
rchar	0.255544	0.117504	0.386875	0.576127	0.419759	0.370995	0.324990	1.000000	0.486317	0.250881	0.267983	0.261728	0.265281	0.369470
wchar	0.174410	0.132595	0.331933	0.415059	0.430126	0.122990	0.124145	0.486317	1.000000	0.195926	0.204542	0.187786	0.160142	0.250258
pgout	0.208258	0.090726	0.297487	0.301658	0.274708	0.197047	0.230712	0.250881	0.195926	1.000000	0.950418	0.909151	0.642940	0.437916
ppgout	0.219798	0.089149	0.305507	0.315648	0.284982	0.214192	0.248680	0.267983	0.204542	0.950418	1.000000	0.969091	0.614961	0.464484
pgfree	0.214400	0.082519	0.300612	0.311863	0.280082	0.216166	0.250468	0.261728	0.187786	0.909151	0.969091	1.000000	0.598164	0.464531
atch	0.230628	0.126769	0.306359	0.292044	0.264283	0.201768	0.252416	0.265281	0.160142	0.642940	0.614961	0.598164	1.000000	0.329873
pgin	0.283252	0.113368	0.335380	0.346654	0.307420	0.247715	0.299591	0.369470	0.250258	0.437916	0.464484	0.464531	0.329873	1.000000
ppgin	0.290564	0.118388	0.325487	0.344586	0.302481	0.236216	0.288017	0.390376	0.260535	0.448769	0.482710	0.483354	0.334002	0.961242
pflt	0.375250	0.102631	0.485361	0.529364	0.505560	0.939125	0.758255	0.381489	0.125757	0.206128	0.221593	0.221590	0.208409	0.252192
vflt	0.421079	0.135243	0.548081	0.597892	0.563163	0.932579	0.763429	0.438870	0.157922	0.299740	0.320826	0.322354	0.296459	0.396702
freemem	-0.201369	-0.099558	-0.388969	-0.349887	-0.350318	-0.136045	-0.191919	-0.165552	-0.146838	-0.469831	-0.461413	-0.464708	-0.442062	-0.309647
freeswap	-0.243903	-0.149630	-0.357864	-0.369897	-0.336869	-0.133135	-0.184268	-0.230327	-0.175309	-0.348040	-0.338905	-0.339652	-0.345789	-0.365383
usr	-0.438163	-0.185695	-0.618932	-0.638072	-0.598098	-0.673513	-0.609368	-0.507561	-0.317330	-0.381960	-0.388662	-0.382053	-0.341721	-0.459133

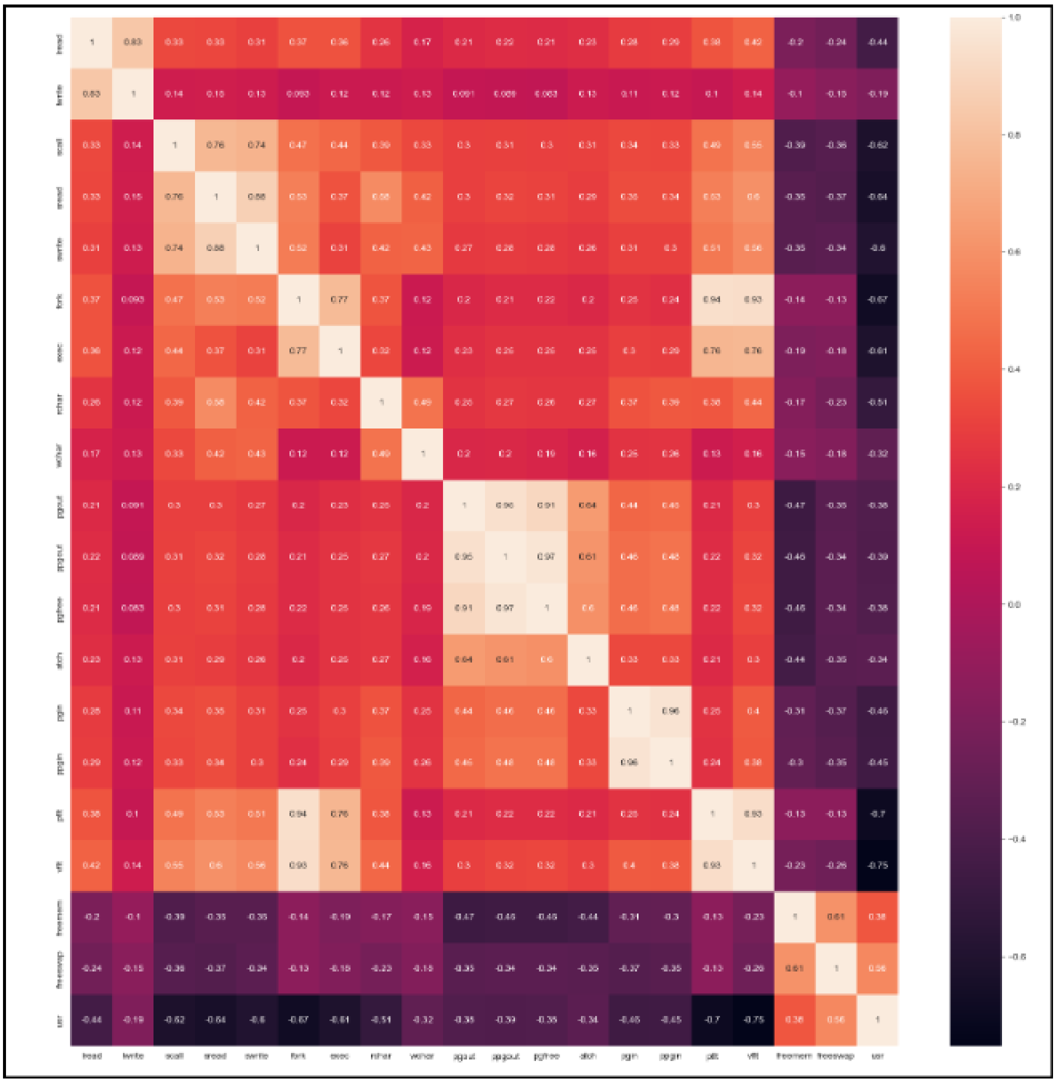

Figure 14 heatmap corelation

The variables have moderate corelation

We will now split the data

X = 'lread', 'lwrite', 'scall', 'sread', 'swrite', 'fork', 'exec', 'rchar',

'wchar', 'pgout', 'ppgout', 'pgfree', 'atch', 'pgin', 'ppgin', 'pflt',

'vflt', 'freemem', 'freeswap'

Y= 'usr'

split X and y into train and test sets in a 70:30 ratio.

Table 7 - Train head

	const	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	ppgout	pgfree	atch	pgin	ppgin	pflt	vflt	freemem	freeswap
694	1.0	1.0	1.0	1345.0	223.0	192.0	0.6	0.6	198703.0	230625.875	0.60	6.20	12.50	1.5	3.80	7.40	28.20	56.60	121.0	1375446.0
5535	1.0	1.0	1.0	1429.0	87.0	67.0	0.2	0.2	7163.0	24842.000	0.00	0.00	0.00	0.0	1.60	1.60	15.77	30.74	1476.0	1021541.0
4244	1.0	47.0	25.0	3273.0	225.0	180.0	0.6	0.4	83246.0	53705.000	5.39	7.19	7.19	1.5	3.99	4.59	59.88	74.05	82.0	10989.5
2472	1.0	13.0	8.0	4349.0	300.0	191.0	2.8	3.0	96009.0	70467.000	0.00	0.00	0.00	0.0	2.80	3.20	129.00	236.80	772.0	993909.0
7052	1.0	17.0	23.0	225.0	13.0	13.0	0.4	1.6	17132.0	12514.000	0.00	0.00	0.00	0.0	0.00	0.00	19.80	23.80	4179.0	1821682.0

Train head

Table 8 - Test head

	const	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	ppgout	pgfree	atch	pgin	ppgin	pflt	vflt	freemem	freeswap
3894	1.0	27.0	25.0	1252.0	53.0	118.0	0.2	0.2	26592.0	54394.000	0.0	0.0	0.0	0.0	0.4	0.6	19.44	20.04	4659.125	1875466.0
4276	1.0	1.0	0.0	996.0	85.0	55.0	0.4	0.4	16667.0	36431.000	0.0	0.0	0.0	0.0	1.0	1.4	35.53	52.10	2979.000	1010114.0
3414	1.0	9.0	7.0	1530.0	247.0	135.0	0.4	0.4	14513.0	61905.000	6.0	10.5	12.5	1.5	14.8	18.4	26.80	186.20	89.000	10989.5
4165	1.0	32.0	4.0	3243.0	182.0	140.0	4.9	5.6	337517.0	94832.000	0.8	1.0	1.0	1.4	4.6	7.0	250.60	420.20	1300.000	1535309.0
7385	1.0	16.0	3.0	5017.0	259.0	249.0	2.8	1.4	73537.0	230625.875	0.0	0.0	0.0	0.0	5.6	5.8	142.80	276.20	2114.000	988600.0

Test Head

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.

Table 9 - OLS model 1

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.790
Model:                            OLS   Adj. R-squared:                  0.790
Method:                 Least Squares   F-statistic:                     1133.
Date:                Sun, 14 Jan 2024   Prob (F-statistic):               0.00
Time:                        03:03:34   Log-Likelihood:                -16738.
No. Observations:                5734   AIC:                         3.352e+04
Df Residuals:                    5714   BIC:                         3.365e+04
Df Model:                          19
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         85.5495      0.300    285.117      0.000      84.961      86.138
lread         -0.0748      0.009     -8.259      0.000      -0.093      -0.057
lwrite         0.0610      0.013      4.598      0.000       0.035       0.087
scall         -0.0007   6.34e-05    -11.772      0.000      -0.001      -0.001
sread          0.0003      0.001      0.266      0.790      -0.002       0.002
swrite        -0.0054      0.001     -3.738      0.000      -0.008      -0.003
fork           0.0576      0.134      0.431      0.667      -0.204       0.320
exec          -0.3323      0.052     -6.347      0.000      -0.435      -0.230
rchar      -5.812e-06   4.92e-07    -11.818      0.000   -6.78e-06   -4.85e-06
wchar      -7.203e-06   1.04e-06     -6.941      0.000   -9.24e-06   -5.17e-06
pgout         -0.3309      0.091     -3.627      0.000      -0.510      -0.152
ppgout        -0.0580      0.080     -0.727      0.467      -0.214       0.098
pgfree         0.0695      0.048      1.435      0.151      -0.025       0.164
atch           0.6753      0.145      4.664      0.000       0.391       0.959
pgin           0.0171      0.029      0.593      0.554      -0.039       0.074
ppgin         -0.0631      0.020     -3.155      0.002      -0.102      -0.024
pflt          -0.0336      0.002    -16.716      0.000      -0.038      -0.030
vflt          -0.0053      0.001     -3.690      0.000      -0.008      -0.003
freemem       -0.0004   5.12e-05     -7.818      0.000      -0.001      -0.000
freeswap    8.627e-06   1.92e-07     44.918      0.000    8.25e-06       9e-06
==============================================================================
Omnibus:                     1325.834   Durbin-Watson:                   2.019
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3434.492
Skew:                          -1.247   Prob(JB):                         0.00
Kurtosis:                       5.856   Cond. No.                     7.17e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.17e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The R-squared value indicates that our model can explain 79.0% of the variance in the training set.

Table 10 - VIF model 1

```
VIF values:

const       25.601099
lread        5.298513
lwrite       4.303179
scall        2.929695
sread        6.420121
swrite       5.597129
fork        13.031712
exec         3.240493
rchar        2.110885
wchar        1.555099
pgout       11.348082
ppgout      29.394226
pgfree      16.486831
atch         1.874624
pgin        13.808466
ppgin       13.947912
pflt        12.001459
vflt        15.970315
freemem      1.945776
freeswap     1.828205
dtype: float64
```

using Variance Inflation Factor (VIF)

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

- On dropping ‘ppgout’ and 'pgfree' adj. R-squared almost remains the same.
- On dropping 'vflt', adj. R-squared decresed by 0.001
- On dropping 'ppgin', adj. R-squared decresed by 0.001
- On dropping 'pgin', adj. R-squared almost remains the same.
- On dropping 'fork', adj. R-squared almost remains the same.
- On dropping 'pflt', adj. R-squared decreased by 0.011
- sharp decline indicates that 'pflt' is an important predictor and should not be removed On dropping
- 'pgout', adj. R-squared decresed by 0.001
- On dropping 'sread', adj. R-squared decresed by 0.001
- In conclusion, observing that the adjusted R-squared remains unaffected upon removing the 'ppgout' column, and considering its substantial influence on the variance, we opt to exclude it from the training set.

Table 11 - OLS model 2

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.790
Model:                            OLS   Adj. R-squared:                  0.790
Method:                 Least Squares   F-statistic:                     1196.
Date:                Sun, 14 Jan 2024   Prob (F-statistic):               0.00
Time:                        03:03:34   Log-Likelihood:                -16739.
No. Observations:                5734   AIC:                         3.352e+04
Df Residuals:                    5715   BIC:                         3.364e+04
Df Model:                          18
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         85.5676      0.299    286.179      0.000      84.981      86.154
lread         -0.0749      0.009     -8.262      0.000      -0.093      -0.057
lwrite         0.0610      0.013      4.600      0.000       0.035       0.087
scall         -0.0007   6.34e-05    -11.777      0.000      -0.001      -0.001
sread          0.0003      0.001      0.264      0.792      -0.002       0.002
swrite        -0.0054      0.001     -3.741      0.000      -0.008      -0.003
fork           0.0600      0.134      0.449      0.654      -0.202       0.322
exec          -0.3333      0.052     -6.368      0.000      -0.436      -0.231
rchar      -5.811e-06   4.92e-07    -11.816      0.000   -6.77e-06   -4.85e-06
wchar      -7.236e-06   1.04e-06     -6.981      0.000   -9.27e-06    -5.2e-06
pgout         -0.3745      0.069     -5.454      0.000      -0.509      -0.240
pgfree         0.0417      0.030      1.406      0.160      -0.016       0.100
atch           0.6767      0.145      4.674      0.000       0.393       0.961
pgin           0.0180      0.029      0.624      0.532      -0.038       0.074
ppgin         -0.0640      0.020     -3.207      0.001      -0.103      -0.025
pflt          -0.0336      0.002    -16.717      0.000      -0.038      -0.030
vflt          -0.0054      0.001     -3.703      0.000      -0.008      -0.003
freemem       -0.0004   5.12e-05     -7.847      0.000      -0.001      -0.000
freeswap    8.621e-06   1.92e-07     44.926      0.000    8.25e-06       9e-06
==============================================================================
Omnibus:                     1324.415   Durbin-Watson:                   2.019
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3427.527
Skew:                          -1.246   Prob(JB):                         0.00
Kurtosis:                       5.852   Cond. No.                     7.14e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.14e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The R-squared value tells us that our model can explain 79.0% of the variance in the training set.

check for multicollinearity

Table 12 - VIF value 2

```
VIF values:

const       25.424423
lread        5.298433
lwrite       4.303148
scall        2.929554
sread        6.420086
swrite       5.597020
fork        13.023854
exec         3.238358
rchar        2.110869
wchar        1.551973
pgout        6.430968
pgfree       6.171219
atch         1.874301
pgin        13.782952
ppgin       13.894353
pflt        12.001459
vflt        15.966065
freemem      1.943526
freeswap     1.825361
dtype: float64
```

- On dropping 'vflt', adj. R-squared decresed by 0.001
- In conclusion, the removal of 'pflt' led to a substantial 0.011 decrease in the adjusted R-squared. This pronounced reduction underscores the significance of 'pflt' as a crucial predictor, strongly advising against its exclusion from the model.

Given the absence of any impact on the adjusted R-squared after removing the 'pgin' column, and considering its high variance influence factor, the decision is made to exclude it from the training set.

Table 13 - OLS model 3

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.790
Model:                            OLS   Adj. R-squared:                  0.790
Method:                 Least Squares   F-statistic:                     1266.
Date:                Sun, 14 Jan 2024   Prob (F-statistic):               0.00
Time:                        03:03:34   Log-Likelihood:                -16739.
No. Observations:                5734   AIC:                         3.351e+04
Df Residuals:                    5716   BIC:                         3.363e+04
Df Model:                          17
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         85.5825      0.298    287.156      0.000      84.998      86.167
lread         -0.0750      0.009     -8.279      0.000      -0.093      -0.057
lwrite         0.0610      0.013      4.602      0.000       0.035       0.087
scall         -0.0007   6.34e-05    -11.766      0.000      -0.001      -0.001
sread          0.0003      0.001      0.255      0.799      -0.002       0.002
swrite        -0.0054      0.001     -3.736      0.000      -0.008      -0.003
fork           0.0571      0.134      0.427      0.669      -0.205       0.319
exec          -0.3324      0.052     -6.353      0.000      -0.435      -0.230
rchar      -5.833e-06    4.9e-07    -11.895      0.000   -6.79e-06   -4.87e-06
wchar       -7.22e-06   1.04e-06     -6.968      0.000   -9.25e-06   -5.19e-06
pgout         -0.3730      0.069     -5.435      0.000      -0.507      -0.238
pgfree         0.0407      0.030      1.377      0.169      -0.017       0.099
atch           0.6771      0.145      4.677      0.000       0.393       0.961
ppgin         -0.0523      0.007     -7.447      0.000      -0.066      -0.039
pflt          -0.0337      0.002    -16.793      0.000      -0.038      -0.030
vflt          -0.0053      0.001     -3.656      0.000      -0.008      -0.002
freemem       -0.0004   5.12e-05     -7.848      0.000      -0.001      -0.000
freeswap    8.614e-06   1.91e-07     44.981      0.000    8.24e-06    8.99e-06
==============================================================================
Omnibus:                     1325.636   Durbin-Watson:                   2.019
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3435.181
Skew:                          -1.247   Prob(JB):                         0.00
Kurtosis:                       5.857   Cond. No.                     7.12e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.12e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The R-squared value indicates that our model can account for 79.0% of the variance within the training set.

Check for multicollinearity

Table 14 - VIF model 3

```
VIF values:

const       25.263079
lread        5.295815
lwrite       4.303110
scall        2.927766
sread        6.418812
swrite       5.596559
fork        13.008265
exec         3.235774
rchar        2.099560
wchar        1.550961
pgout        6.422354
pgfree       6.155198
atch         1.874263
ppgin        1.722709
pflt        11.952041
vflt        15.755601
freemem      1.943522
freeswap     1.817817
dtype: float64
```

Let's eliminate multicollinear columns step by step and assess their impact on our predictive model.

Let drop the variable "fork" from our train data

Table 15 - OLS model 4

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.790
Model:                            OLS   Adj. R-squared:                  0.790
Method:                 Least Squares   F-statistic:                     1346.
Date:                Sun, 14 Jan 2024   Prob (F-statistic):               0.00
Time:                        03:03:34   Log-Likelihood:                -16739.
No. Observations:                5734   AIC:                         3.351e+04
Df Residuals:                    5717   BIC:                         3.363e+04
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         85.5665      0.296    289.400      0.000      84.987      86.146
lread         -0.0749      0.009     -8.270      0.000      -0.093      -0.057
lwrite         0.0608      0.013      4.587      0.000       0.035       0.087
scall         -0.0007   6.28e-05    -11.916      0.000      -0.001      -0.001
sread          0.0002      0.001      0.234      0.815      -0.002       0.002
swrite        -0.0053      0.001     -3.719      0.000      -0.008      -0.003
exec          -0.3261      0.050     -6.499      0.000      -0.424      -0.228
rchar      -5.835e-06    4.9e-07    -11.899      0.000    -6.8e-06   -4.87e-06
wchar      -7.225e-06   1.04e-06     -6.974      0.000   -9.26e-06   -5.19e-06
pgout         -0.3728      0.069     -5.433      0.000      -0.507      -0.238
pgfree         0.0405      0.030      1.370      0.171      -0.017       0.098
atch           0.6736      0.145      4.660      0.000       0.390       0.957
ppgin         -0.0529      0.007     -7.662      0.000      -0.066      -0.039
pflt          -0.0334      0.002    -17.949      0.000      -0.037      -0.030
vflt          -0.0050      0.001     -3.923      0.000      -0.007      -0.002
freemem       -0.0004   5.12e-05     -7.848      0.000      -0.001      -0.000
freeswap    8.623e-06    1.9e-07     45.331      0.000    8.25e-06       9e-06
==============================================================================
Omnibus:                     1325.032   Durbin-Watson:                   2.019
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3427.344
Skew:                          -1.247   Prob(JB):                         0.00
Kurtosis:                       5.851   Cond. No.                     7.05e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.05e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Following the removal of features responsible for significant multicollinearity and those deemed statistically insignificant, our model's performance has not experienced a substantial decline. This suggests that these variables lacked considerable predictive power in our model.

Table 16 - OLS model 5

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.787
Model:                            OLS   Adj. R-squared:                  0.786
Method:                 Least Squares   F-statistic:                     1625.
Date:                Sun, 14 Jan 2024   Prob (F-statistic):               0.00
Time:                        03:03:35   Log-Likelihood:                -16783.
No. Observations:                5734   AIC:                         3.359e+04
Df Residuals:                    5720   BIC:                         3.369e+04
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         85.4441      0.296    288.345      0.000      84.863      86.025
lwrite        -0.0354      0.007     -5.361      0.000      -0.048      -0.022
scall         -0.0008   6.03e-05    -12.916      0.000      -0.001      -0.001
swrite        -0.0059      0.001     -5.518      0.000      -0.008      -0.004
exec          -0.4084      0.049     -8.304      0.000      -0.505      -0.312
rchar           -6e-06   4.41e-07    -13.607      0.000   -6.86e-06   -5.14e-06
wchar       -6.376e-06   1.03e-06     -6.200      0.000   -8.39e-06   -4.36e-06
pgout         -0.3521      0.069     -5.097      0.000      -0.488      -0.217
pgfree         0.0275      0.030      0.925      0.355      -0.031       0.086
atch           0.6217      0.145      4.278      0.000       0.337       0.907
ppgin         -0.0699      0.007    -10.556      0.000      -0.083      -0.057
pflt          -0.0414      0.001    -39.277      0.000      -0.043      -0.039
freemem       -0.0004   5.15e-05     -7.760      0.000      -0.001      -0.000
freeswap    8.812e-06   1.89e-07     46.673      0.000    8.44e-06    9.18e-06
==============================================================================
Omnibus:                     1268.997   Durbin-Watson:                   2.014
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3195.014
Skew:                          -1.206   Prob(JB):                         0.00
Kurtosis:                       5.748   Cond. No.                     7.02e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.02e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Table 17 - VIF values 4

```
VIF values:

const       24.318585
lwrite       1.051894
scall        2.601101
swrite       2.521669
exec         1.368012
rchar        1.623734
wchar        1.438804
pgout        2.027035
atch         1.858638
ppgin        1.463216
freemem      1.911325
freeswap     1.727889
dtype: float64
```

VIF for all features is less than 3

Table 18 - VIF for all features

	Actual Values	Fitted Values	Residuals
0	91.0	91.765169	-0.765169
1	94.0	91.261125	2.738875
2	61.5	76.430666	-14.930666
3	83.0	81.272261	1.727739
4	94.0	97.101512	-3.101512

Checking linearity and independence

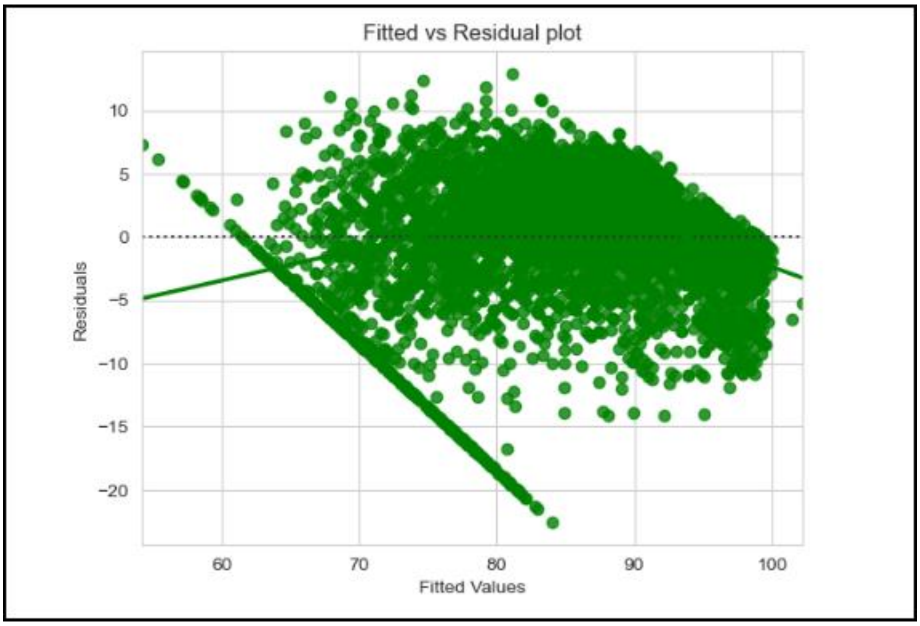

Figure 17 - Fitted vs residual

It's noted that the pattern has slightly diminished, and the data points now appear to be more randomly distributed.

Lets check the distribution of the data

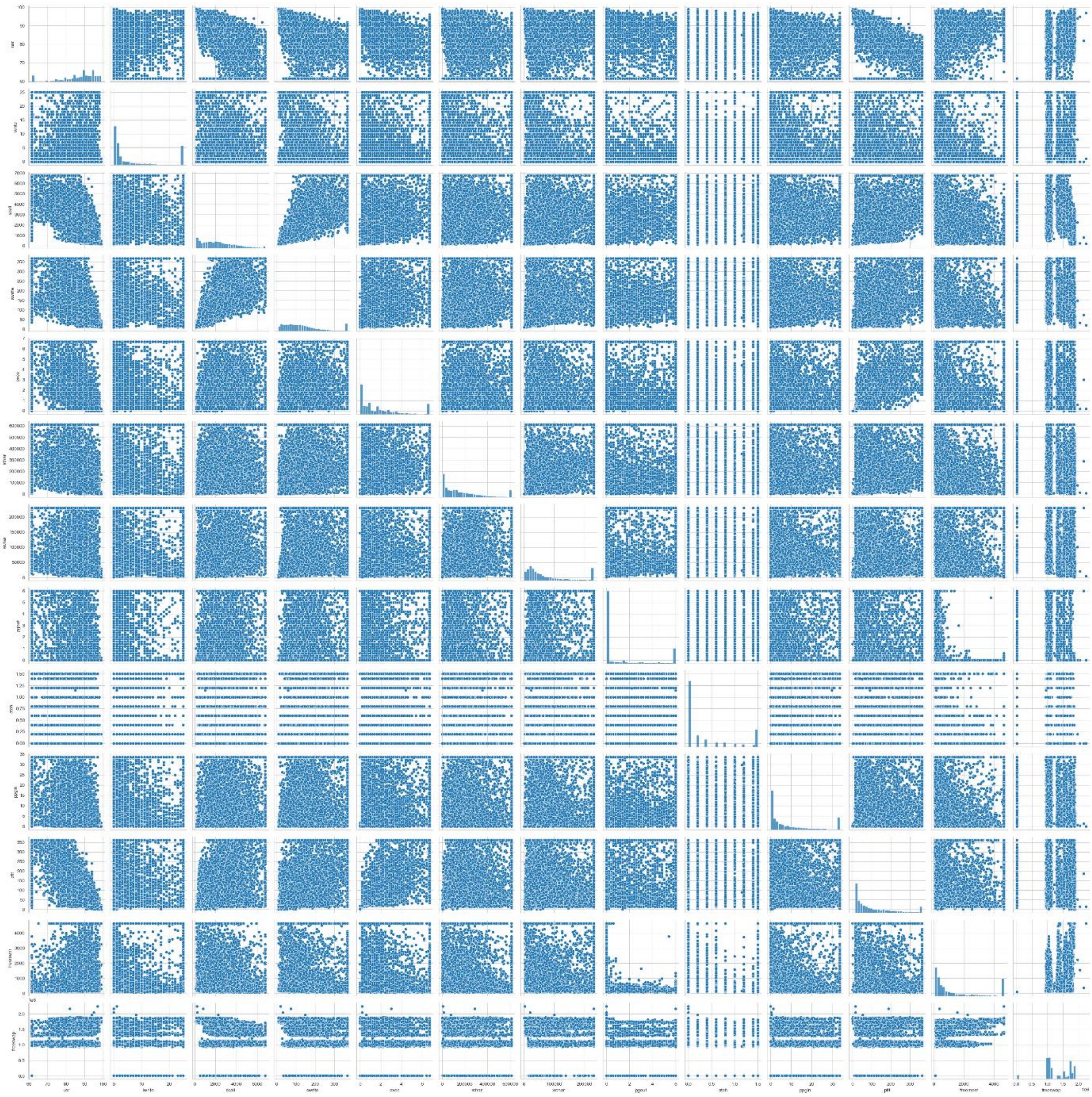

Figure 18 – Pairplot distribution of data

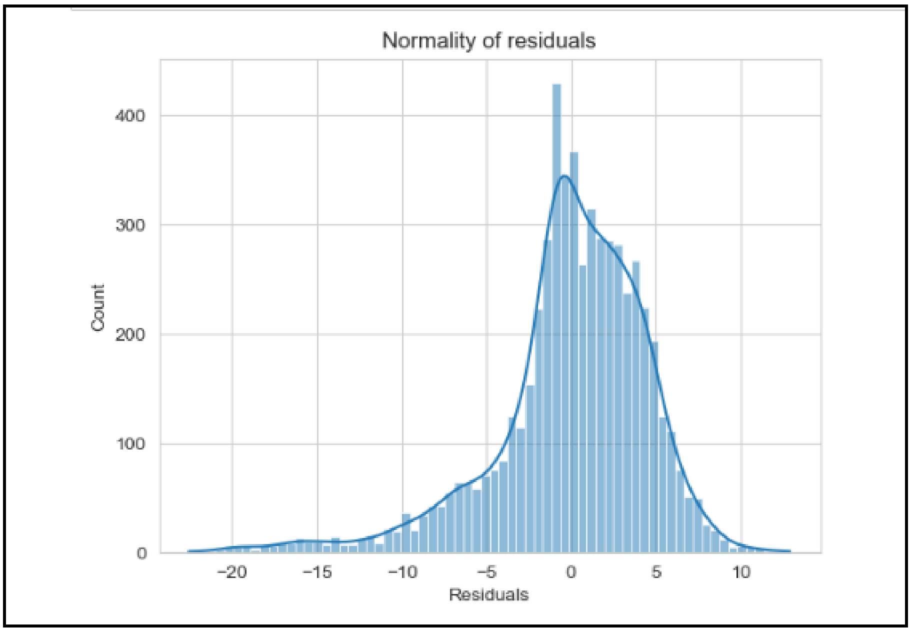

Figure 19 - Normality of residuals

The residual terms demonstrate a normal distribution. A visual examination of the QQ plot of residuals can validate the normality assumption, where the normal probability plot should closely align with a straight line.

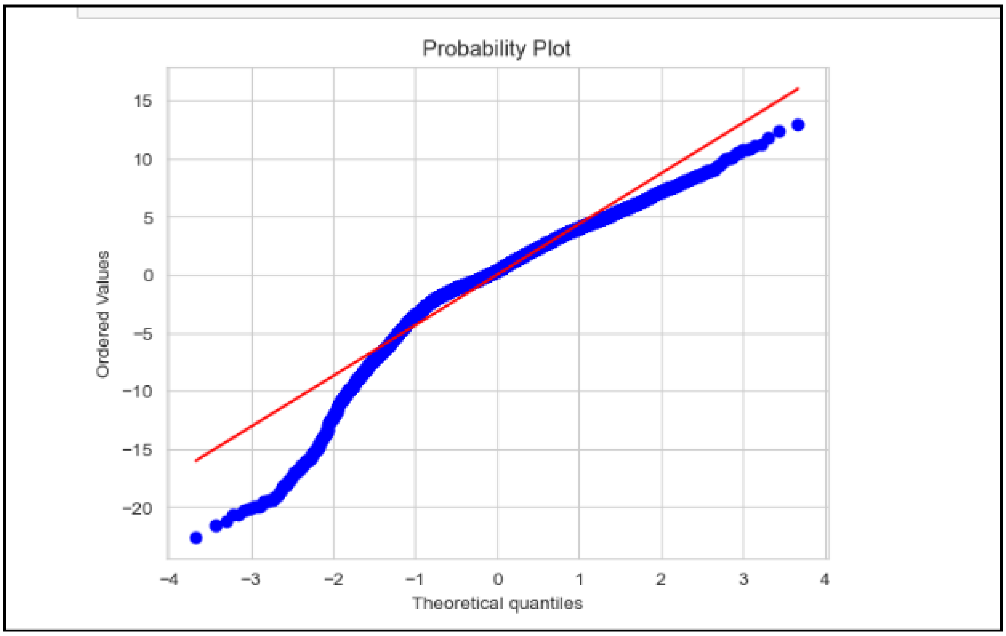

Figure 20 - Probability plot

Partially the points are lying on the straight line in QQ plot

[('F statistic', 1.1299923459713648), ('p-value', 0.0005576651863864631)]

assumptions of linear regression are now satisfied.

Best Model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.787
Model:                            OLS   Adj. R-squared:                  0.786
Method:                 Least Squares   F-statistic:                     1625.
Date:                Sun, 14 Jan 2024   Prob (F-statistic):               0.00
Time:                        03:04:30   Log-Likelihood:                -16783.
No. Observations:                5734   AIC:                         3.359e+04
Df Residuals:                    5720   BIC:                         3.369e+04
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         85.4441      0.296    288.345      0.000      84.863      86.025
lwrite        -0.0354      0.007     -5.361      0.000      -0.048      -0.022
scall         -0.0008   6.03e-05    -12.916      0.000      -0.001      -0.001
swrite        -0.0059      0.001     -5.518      0.000      -0.008      -0.004
exec          -0.4084      0.049     -8.304      0.000      -0.505      -0.312
rchar           -6e-06   4.41e-07    -13.607      0.000   -6.86e-06   -5.14e-06
wchar      -6.376e-06   1.03e-06     -6.200      0.000   -8.39e-06   -4.36e-06
pgout         -0.3521      0.069     -5.097      0.000      -0.488      -0.217
pgfree         0.0275      0.030      0.925      0.355      -0.031       0.086
atch           0.6217      0.145      4.278      0.000       0.337       0.907
ppgin         -0.0699      0.007    -10.556      0.000      -0.083      -0.057
pflt          -0.0414      0.001    -39.277      0.000      -0.043      -0.039
freemem       -0.0004   5.15e-05     -7.760      0.000      -0.001      -0.000
freeswap    8.812e-06   1.89e-07     46.673      0.000    8.44e-06    9.18e-06
==============================================================================
Omnibus:                     1268.997   Durbin-Watson:                   2.014
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3195.014
Skew:                          -1.206   Prob(JB):                         0.00
Kurtosis:                       5.748   Cond. No.                     7.02e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.02e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 21 - Best model

```
usr = 85.44405142467511 + -0.035431681356878006 * ( lwrite ) +  -0.0007
789166996726056 * ( scall ) +  -0.0059364238127426 * ( swrite ) +  -0.4
08362100281623 * ( exec ) +  -6.000393832825783e-06 * ( rchar ) +  -6.3
76410814972524e-06 * ( wchar ) +  -0.35214422917288685 * ( pgout ) +  0
.027503334107575 66 * ( pgfree ) +  0.6216576279548014 * ( atch ) +  -0.
06988532485141626 * ( ppgin ) +  -0.04141258260531031 * ( pflt ) +  -0.
0004000185573245753 * ( freemem ) +  8.81187709338839e-06 * ( freeswap
)
```

RMSE of train data 4.517865444540217

RMSE of test data 4.789314038671895

Observation: The comparable RMSE values on both the train and test sets suggest that our model isn't encountering overfitting issues. The MAE indicates that our current model can predict 'mpg' within a mean error of the test data. Therefore, we can confidently conclude that the "fitres_42" model is suitable for both prediction and inference purposes.

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

```
const       85.444051
lwrite      -0.035432
scall       -0.000779
swrite      -0.005936
exec        -0.408362
rchar       -0.000006
wchar       -0.000006
pgout       -0.352144
pgfree       0.027503
atch         0.621658
ppgin       -0.069885
pflt        -0.041413
freemem     -0.000400
freeswap     0.000009
dtype: float64
```

Figure 22 - linear regression equation

```
usr = 85.44405142467511 + -0.035431681356878006 * ( lwrite ) +  -0.0007
789166996726056 * ( scall ) +  -0.0059364238127426 * ( swrite ) +  -0.4
08362100281623 * ( exec ) +  -6.000393832825783e-06 * ( rchar ) +  -6.3
76410814972524e-06 * ( wchar ) +  -0.35214422917288685 * ( pgout ) +  0
.02750333410757566 * ( pgfree ) +  0.6216576279548014 * ( atch ) +  -0.
06988532485141626 * ( ppgin ) +  -0.04141258260531031 * ( pflt ) +  -0.
0004000185573245753 * ( freemem ) +  8.81187709338839e-06 * ( freeswap
)
```

Positive Impact:

pgfree: An increase in pgfree is associated with an increase in 'usr.'

atch: An increase in atch is linked to an increase in 'usr.'

Negative Impact:

exec: An increase in exec is associated with a decrease in 'usr.'

pflt: An increase in pflt is linked to a decrease in 'usr.'

Conclude with the key takeaways (actionable insights and recommendations) for the business

1. Efficient Resource Management

- Focus on optimizing activities related to pgfree and text atch for improved 'usr' mode.
- Allocate resources strategically to enhance system performance.

2. Execution Call Monitoring

- Keep a close eye on exec calls, as they negatively impact 'usr.'
- Streamline processes with frequent execution calls to improve overall efficiency.

3. Continuous Monitoring and Improvement

- Regularly monitor system attributes for emerging patterns or issues.
- Implement ongoing improvements based on data analysis to adapt to changing system requirements.

4. Protection Fault Mitigation:

- Address protection faults Pflt to minimize their adverse effect on 'usr.'
- Implement measures to reduce protection errors and ensure system stability.

These insights aim to guide efficient resource allocation, address potential performance issues, and ensure the stability of the computer systems' 'usr' mode.

Problem 2

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.

Data Description

1. Wife's age (numerical)
2. Wife's education (categorical) 1=uneducated, 2, 3, 4=tertiary
3. Husband's education (categorical) 1=uneducated, 2, 3, 4=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) 1, 2, 3, 4(random)
8. Standard-of-living index (categorical) 1=verlow, 2, 3, 4=high
9. Media exposure (binary) Good, Not good
10. Contraceptive method used (class attribute) No,Yes

Introduction:

As a statistician at the Republic of Indonesia Ministry of Health, I am tasked with analyzing a dataset from a Contraceptive Prevalence Survey. The dataset includes information from 1473 married females, focusing on predicting contraceptive choices based on their demographic and socio-economic attributes.

Executive Summary:

The dataset encompasses key variables such as age, education, religion, employment status, and more. Our goal is to build a predictive model to understand the factors influencing contraceptive decisions among married females. This analysis is crucial for guiding public health initiatives and interventions related to family planning in Indonesia. The process involves exploring data patterns, identifying correlations, and utilizing statistical models to provide actionable insights for effective policy recommendations.

2.1 Define the problem and perform exploratory Data Analysis

Problem Definition

Predicting contraceptive choices among married females based on demographic and socio-economic attributes for informed public health strategies in Indonesia.

Table 19 - Dataset

	Wife_age	Wife_education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure
0	24.0	Primary	Secondary	3.0	Scientology	No	2	High	Exposed
1	45.0	Uneducated	Secondary	10.0	Scientology	No	3	Very High	Exposed
2	43.0	Primary	Secondary	7.0	Scientology	No	3	Very High	Exposed
3	42.0	Secondary	Primary	9.0	Scientology	No	3	High	Exposed
4	36.0	Secondary	Secondary	8.0	Scientology	No	3	Low	Exposed

Check shape, Data types, statistical summary

Shape – (1473, 10)

Total Rows – 1473, Total Columns – 10

Object Datatype – 7

Int Datatype – 1

float64 - 2

Table 20 - Info

```
<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
```

Table 21 - Statistical data

	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

Univariate analysis - Multivariate analysis - Use appropriate visualizations to identify the patterns and insights

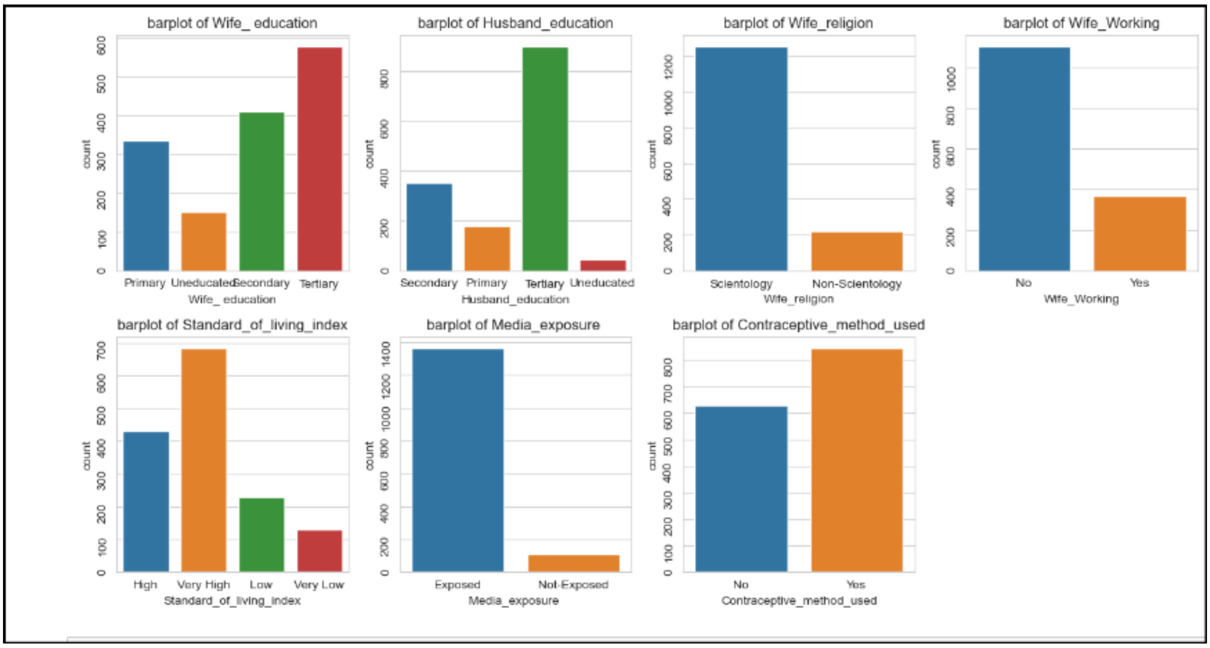

Figure 23 - coutplot of multiple variable

The dataset reveals a higher proportion of educated women and men, with a notable disparity showing more uneducated women than men.

Additionally, a significant number of women are not actively working.

Media exposure is widespread among the surveyed population.

The majority of individuals in the dataset are using contraceptive methods, and a preference for a very high standard of living is evident.

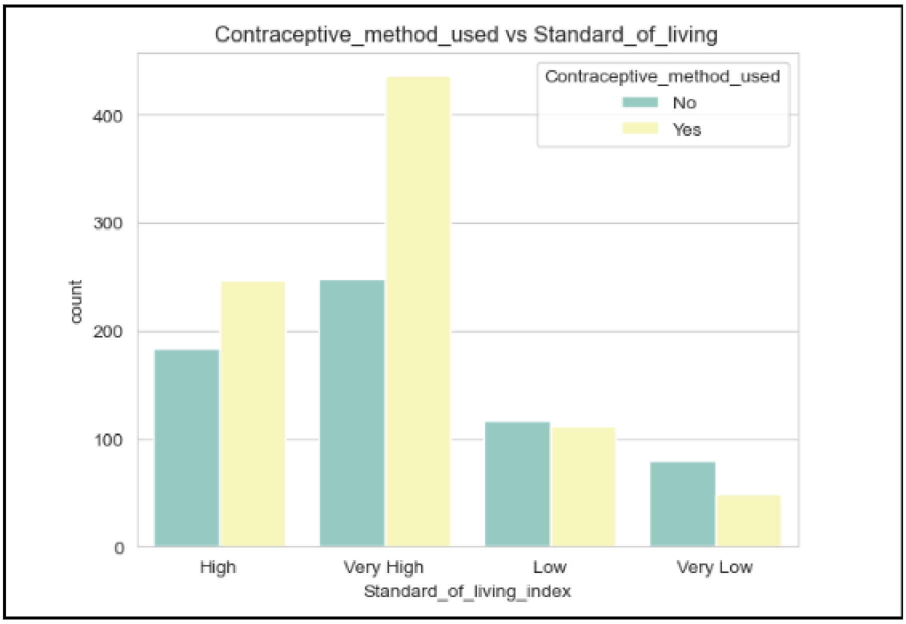

Figure 24 - standard of living index

People with high standard of living prefer contraceptive method

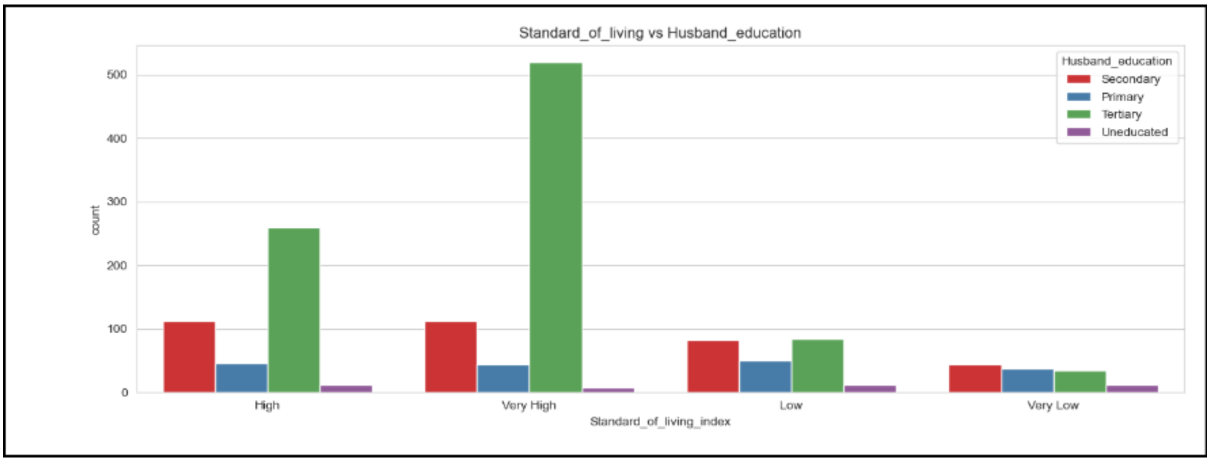

Figure 25 - Husband edcuation

Those husbands having tertiary education have very high standard of living

Key meaningful observations on individual variables and the relationship between variables

Wife's Age:

The age distribution of married females in the dataset varies, providing a diverse range of age groups for analysis.

Education Levels:

There is a significant number of educated women, and the dataset reflects varying education levels for both wives and husbands.

Number of Children Ever Born:

The dataset includes information on the number of children ever born, indicating the family size of the surveyed population.

Religion and Occupation:

The dataset captures information on the wife's religion and the husband's occupation, providing insights into the religious and occupational diversity of the respondents.

Working Status:

A substantial number of women in the dataset are not currently working, highlighting potential implications for family planning and decision-making.

Standard-of-Living Index:

The standard-of-living index varies, with a notable preference for a very high standard of living among the surveyed population.

Media Exposure:

Media exposure is prevalent among the respondents, indicating a potential influence on their awareness and decision-making processes.

Contraceptive Usage:

The majority of individuals in the dataset are currently using contraceptive methods, showcasing the prevalence of family planning practices.

These observations provide a foundation for exploring relationships between variables

2.2 Data Pre-processing

Prepare the data for modelling: - Missing value Treatment (if needed) - Outlier Detection(treat, if needed)

Table 22 - Missing values

```
Wife_age                     71
Wife_ education               0
Husband_education             0
No_of_children_born          21
Wife_religion                 0
Wife_Working                  0
Husband_Occupation            0
Standard_of_living_index      0
Media_exposure                0
Contraceptive_method_used     0
dtype: int64
```

Missing values are present for No of children born and Wife age

We will treat them with median

Table 23 - Post treatment of missing values

```
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
```

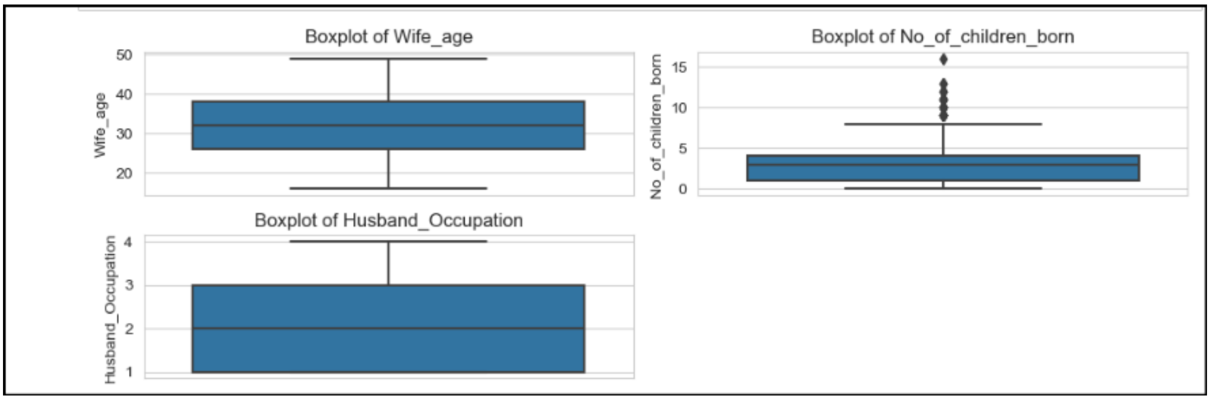

Figure 26 - Outliers

There our outliers in children born and we need to remove them.

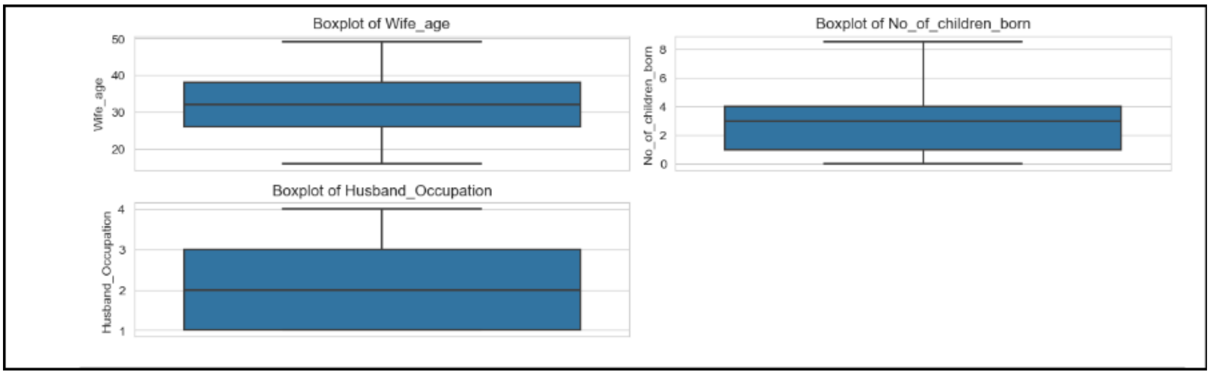

Figure 27 - Post treatment of outlier

Post treatment of outlier

Feature Engineering (if needed) - Encode the data - Train-test split

create dummy variables for categorical columns in the DataFrame df2.

Table 24 - Data encoding

	Wife_age	No_of_children_born	Husband_Occupation	Wife_ education_Secondary	Wife_ education_Tertiary	Wife_ education_Uneducated	Husband_education_Secondary
1468	33.0	3.0	2.0	0	1	0	0
1469	33.0	3.0	1.0	0	1	0	0
1470	39.0	3.0	1.0	1	0	0	1
1471	33.0	3.0	2.0	1	0	0	1
1472	17.0	1.0	2.0	1	0	0	1

We'll separate the data into independent (x) and dependent (y) variables. Subsequently, we'll split the data into a 70:30 ratio for training and testing purposes. This implies that 70% of the data will be used to train the model, while the remaining 30% will be reserved for testing the model.

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 hyperparameters using GridSearch - 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

Logistic Regression

Table 25 – Classification for training data

```
0.6493212669683258
[[218 214]
 [103 496]]
              precision    recall  f1-score   support

         0.0       0.68      0.50      0.58       432
         1.0       0.70      0.83      0.76       599

    accuracy                           0.69      1031
   macro avg       0.69      0.67      0.67      1031
weighted avg       0.69      0.69      0.68      1031
```

Classification report for training data

Table 26 – Classification for testing data

```
0.6493212669683258
[[ 89 108]
 [ 47 198]]
              precision    recall  f1-score   support

         0.0       0.65      0.45      0.53       197
         1.0       0.65      0.81      0.72       245

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

Classification report for testing data

Confusion matrix for train Data

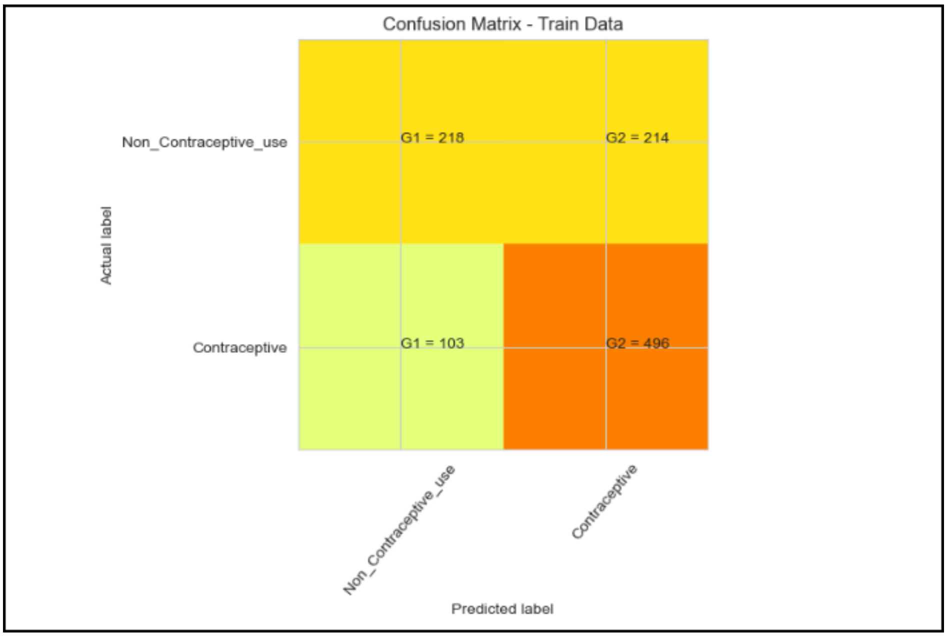

Figure 28 - Confusin for train data

Confusion matrix for test data

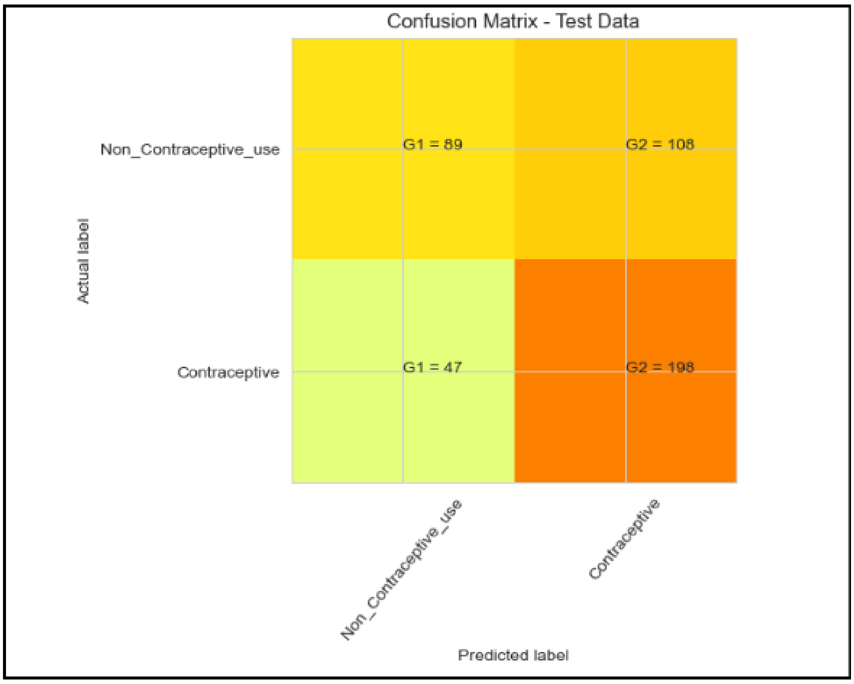

Figure 29 - confusion for test data

The analysis of the confusion matrix on the training data reveals the following:

- True Negative (TN): 198 instances were correctly identified as negative.
- True Positive (TP): 218 instances were correctly identified as positive.
- False Negative (FN): 214 instances were actually positive but predicted as negative.
- False Positive (FP): 103 instances were actually negative but predicted as positive.

Similarly, the confusion matrix analysis on the testing data shows:

- True Negative (TN): 496 instances were correctly identified as negative.
- True Positive (TP): 89 instances were correctly identified as positive.
- False Negative (FN): 108 instances were actually positive but predicted as negative.
- False Positive (FP): 47 instances were actually negative but predicted as positive.

Linear Discriminant Analysis

Coefficient of LDA is a follows

Table 27 - LDA ARRAY

```
array([[-0.08935853,  0.38082614,  0.09030706,  0.6349701 ,  1.54470159,
        -0.3754929 ,  0.17097053, -0.14116475, -0.38729069, -0.46933499,
        -0.12916726, -0.03764121,  0.29537192, -0.39828833, -0.25029414]])
```

Table 28 - Coefficent data

```
The coefficient for Wife_age is -0.08935853494596956
The coefficient for No_of_children_born is 0.38082614077700017
The coefficient for Husband_Occupation is 0.09030705841213124
The coefficient for Wife_ education_Secondary is 0.6349700985239196
The coefficient for Wife_ education_Tertiary is 1.5447015885975135
The coefficient for Wife_ education_Uneducated is -0.37549289911044414
The coefficient for Husband_education_Secondary is 0.17097052569898152
The coefficient for Husband_education_Tertiary is -0.14116474811655483
The coefficient for Husband_education_Uneducated is -0.3872906896794378
The coefficient for Wife_religion_Scientology is -0.4693349895134461
The coefficient for Wife_Working_Yes is -0.12916726253578836
The coefficient for Standard_of_living_index_Low is -0.03764120954991867
The coefficient for Standard_of_living_index_Very High is 0.2953719197526194
The coefficient for Standard_of_living_index_Very Low is -0.39828832728498603
The coefficient for Media_exposure _Not-Exposed is -0.25029413609573725
```

-0.09 x Wife_age + 0.38 x No_of_children_born + 0.09 x Husband_Occupation + 0.63 x Wife_ education_Secondary + 1.54 x Wife_ education_Tertiary + -0.38 x Wife_ education_Uneducated + 0.17 x Husband_education_Secondary + -0.14 x Husband_education_Tertiary + -0.39 x Husband_education_Uneducated + -0.47 x Wife_religion_Scientology + -0.13 x Wife_Working_Yes + -0.04 x Standard_of_living_index_Low + 0.3 x Standard_of_living_index_Very High + -0.4 x Standard_of_living_index_Very Low + -0.25 x Media_exposure _Not-Exposed +

From the provided equation and coefficients, it's evident that the predictor "Wife_education_Tertiary" holds the highest magnitude, suggesting a significant impact on classification. On the other hand, the predictor "Wife_religion_Scientology" possesses the smallest magnitude, indicating a relatively lesser influence on classification, helping discern the least.

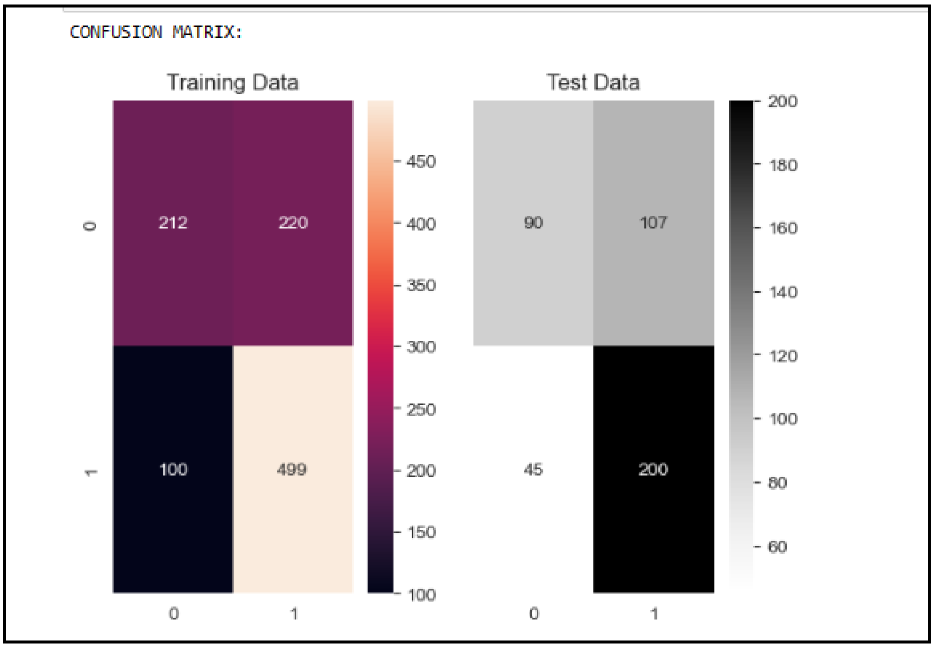

Figure 30 - Confusion Matrix LDA

Table 29 - Classification training data LDA

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.68	0.49	0.57	432
1	0.69	0.83	0.76	599
accuracy			0.69	1031
macro avg	0.69	0.66	0.66	1031
weighted avg	0.69	0.69	0.68	1031

Table 30 - Classification test data LDA

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.67	0.46	0.54	197
1	0.65	0.82	0.72	245
accuracy			0.66	442
macro avg	0.66	0.64	0.63	442
weighted avg	0.66	0.66	0.64	442

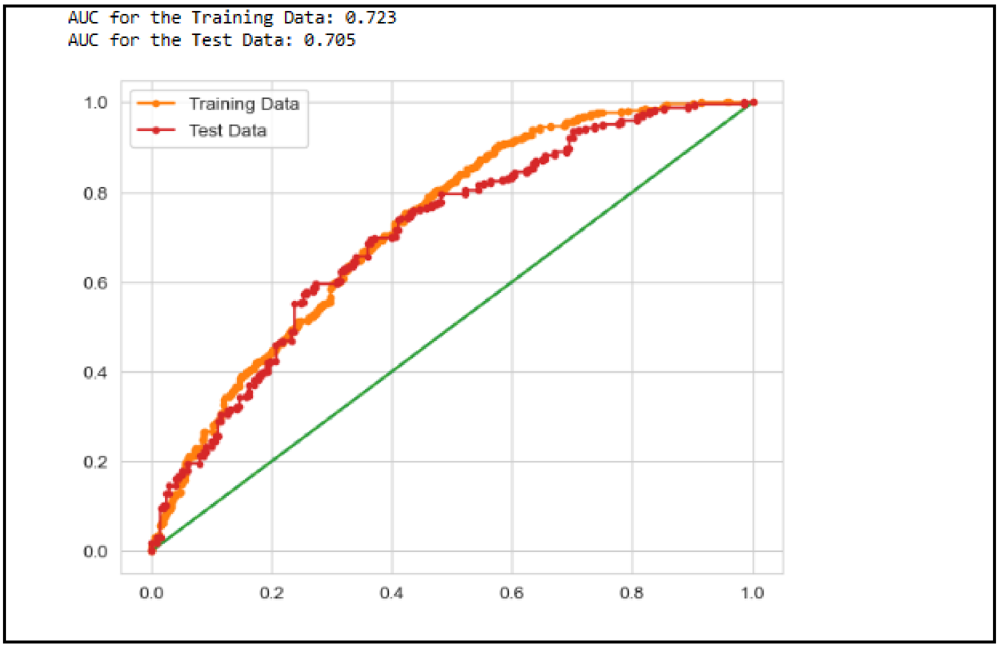

Figure 31 - AUC data

AUC for the Training Data: 0.723

AUC for the Test Data: 0.705

The ROC (Receiver Operating Characteristic) curve is a graphical representation of a model's performance, depicting the true positive rate against the false positive rate. AUC (Area Under the Curve) quantifies the entire area under the ROC curve. In this case, the AUC for both the training and testing data is 72% and 70%, respectively. This indicates that the model covers approximately 70% of the data, demonstrating good performance.

CART - Classification and Regression Trees

We divide the data into independent variables 'X' and dependent variable 'y'. The dataset is split into 70% for training the model and 30% for testing. To implement the CART Algorithm, necessary packages are imported. The data is then fitted to the algorithm using the Gini Index as the classifier for classification at each stage.

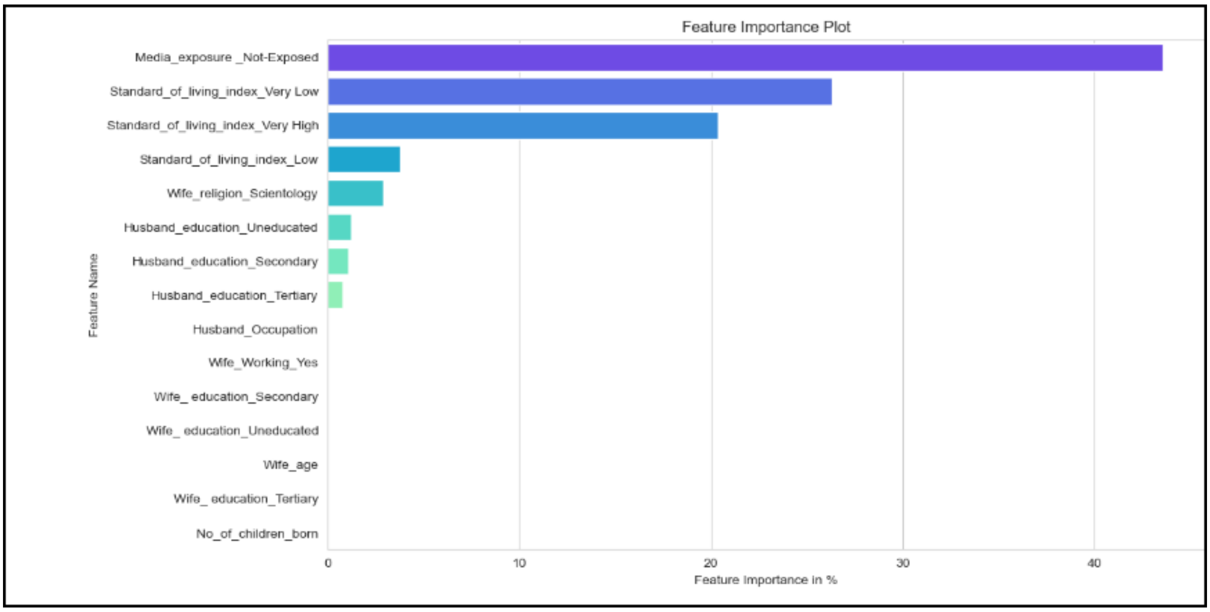

Figure 32 - Feature importance plot

AUC of training data is 78%

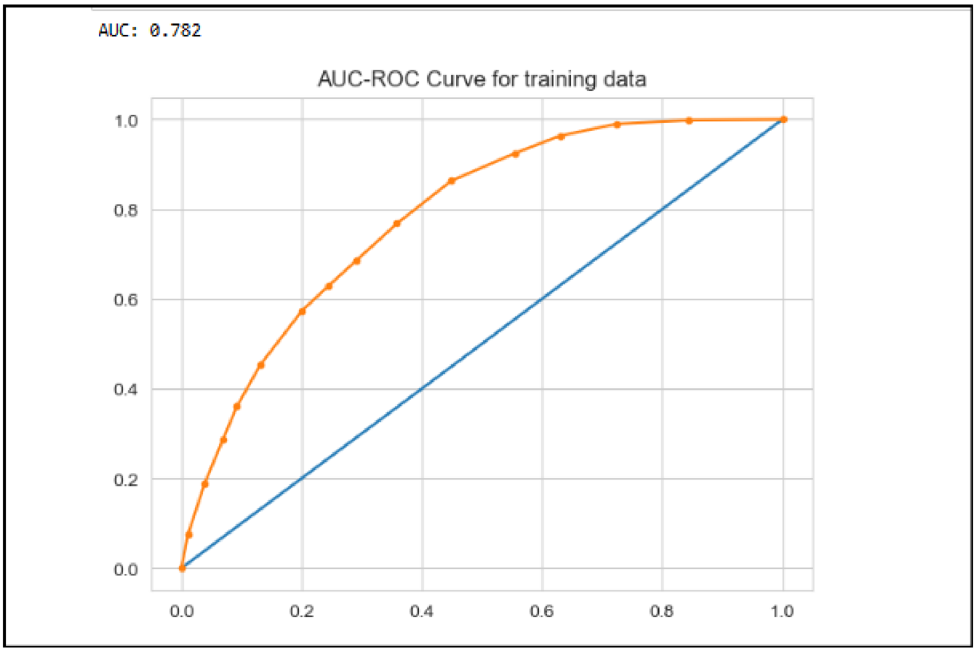

Figure 33 - AUC cart training data

AUC of testing data is 73%

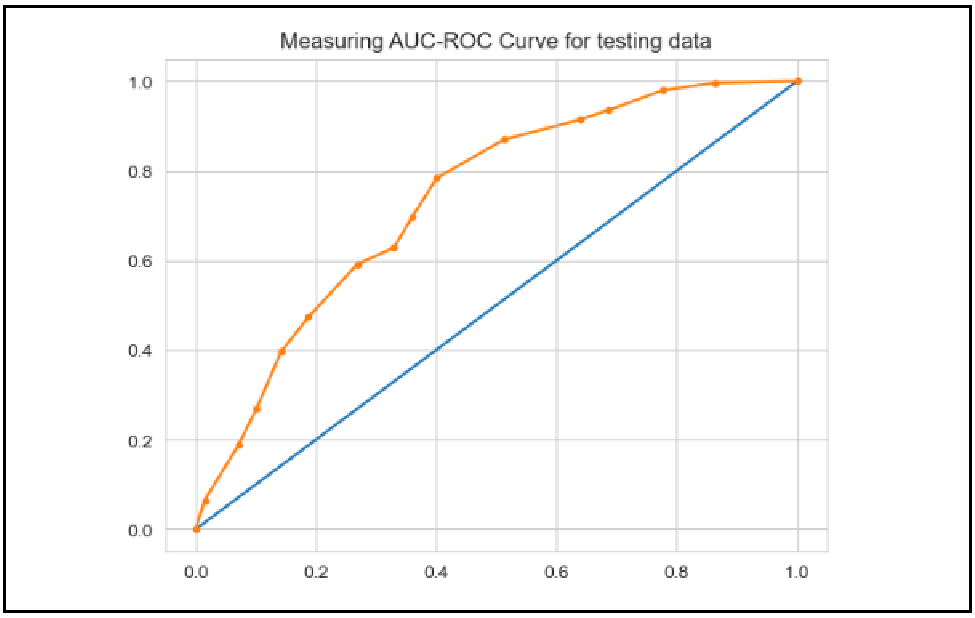

Figure 34 - AUC test data

The analysis of the confusion matrix on the training data reveals 517 True Negatives, 238 True Positives, 194 False Negatives, and 84 False Positives. In the testing data, there are 213 True Negatives, 96 True Positives, 101 False Negatives, and 32 False Positives.

Confusion matrix for training data

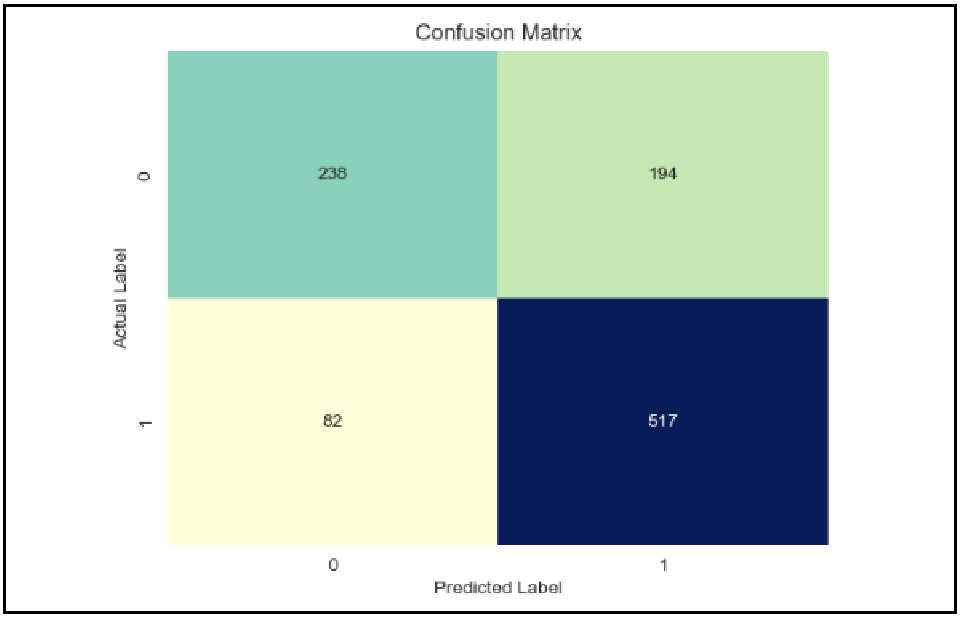

Figure 35 - Confusion matrix for training data

Confusion matrix for testing data

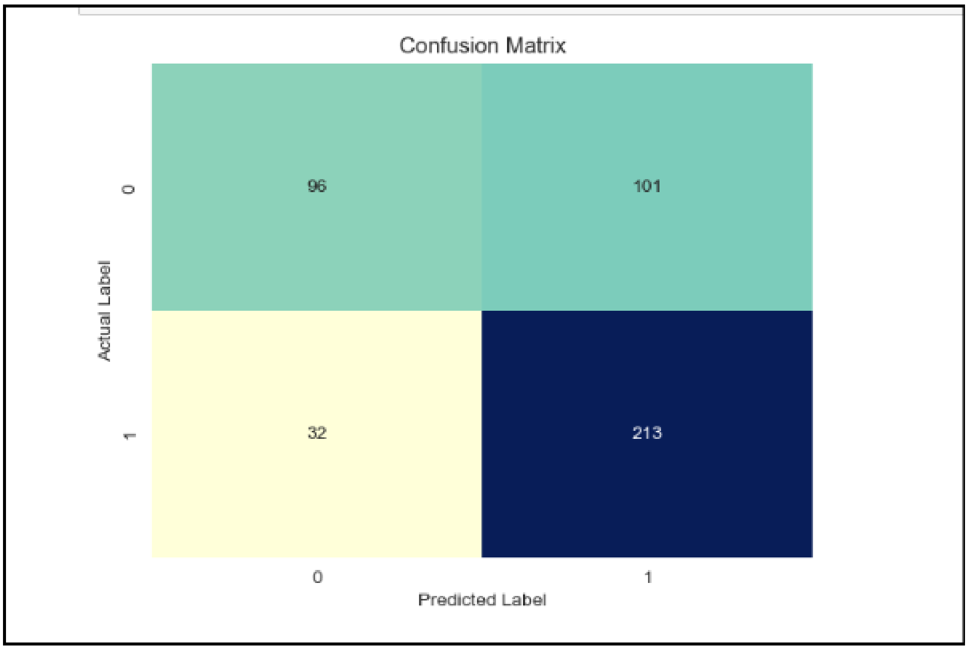

Figure 36 - Confusion matrix for testing data

Among the models employed, Logistic Regression achieved an accuracy of 65%, LDA achieved 69%, and the CART algorithm led with an accuracy of 70%. Additionally, the AUC for LDA on the training data was 72%, on the test data was 70%, and for the CART algorithm, it was 78% on the training data and 73% on the test data. In comparison, the CART algorithm outperforms Logistic Regression and LDA, demonstrating higher accuracy.

2.4 Business Insights & Recommendations

Comment on the importance of features based on the best model - Conclude with the key takeaways (actionable insights and recommendations) for the business

After employing various models to predict contraceptive choices, it was observed that certain features played a crucial role in determining the outcomes. The best-performing model, the Classification and Regression Trees (CART) algorithm, provided insights into feature importance.

Key features contributing significantly to the prediction of contraceptive choices include wife's age, education levels of both wife and husband, the number of children ever born, wife's religion, and media exposure. These features exhibited higher importance in influencing the model's decision-making process.

Key Insights:

1. Age Matters: A woman's age is crucial in predicting contraceptive choices.
2. Education is Significant: Both the wife's and husband's education levels strongly influence decisions.
3. Religion Plays a Role: The wife's religion impacts contraceptive choices, emphasizing the need for cultural sensitivity.
4. Media Exposure Counts: Media exposure is influential, suggesting targeted campaigns for better outcomes.
5. Family Planning Programs: Consider the number of children born for more comprehensive family planning strategies.
6. Support for Working Women: Tailored interventions for working women could enhance family planning success.
7. Socio-economic Factors: Husband's occupation and living standards contribute, albeit less significantly.

Recommendations:

1. Age-Based Programs: Tailor family planning initiatives based on age groups.
2. Educational Campaigns: Focus on educational programs for informed decision-making.

3. Cultural Sensitivity: Consider religious perspectives in family planning strategies.
4. Media-Centric Approaches: Leverage media for targeted campaigns and education.
5. Comprehensive Strategies: Integrate factors like the number of children into family planning programs.
6. Support for Working Women: Implement policies supporting working women in family planning.
7. Socio-economic Tailoring: Tailor interventions based on occupation and living standards.
8. In summary, adopting a nuanced approach considering these insights is recommended for effective family planning programs, promoting better reproductive health outcomes for married women.

The end