Predictive Modelling Project :

-Prapthi Pandian

Table of Contents

Problem 1- Linear Regression

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

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

Dataset for Problem 1: compactiv.xlsx

Data Dictionary:

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 transferred 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.
usr - Portion of time (%) that CPUs run in user mode

1.1 Read the data and do exploratory data analysis. Describe the data briefly. (Check the Data types, shape, EDA, 5 point summary). Perform Univariate, Bivariate Analysis, Multivariate Analysis.

Dataset-

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap	usr	
0	1	0	2147	79	68	0.2	0.2	40671.0	53995.0	0.0	...	0.0	0.0	1.6	2.6	16.00	26.40	CPU_Bound	4670	1730946	95	
1	0	0	170	18	21	0.2	0.2	448.0	8385.0	0.0	...	0.0	0.0	0.0	0.0	15.63	16.83	Not_CPU_Bound	7278	1869002	97	
2	15	3	2162	159	119	2.0	2.4	NaN	31950.0	0.0	...	0.0	1.2	6.0	9.4	150.20	220.20	Not_CPU_Bound	702	1021237	87	
3	0	0	160	12	16	0.2	0.2	NaN	8670.0	0.0	...	0.0	0.0	0.2	0.2	15.60	16.80	Not_CPU_Bound	7248	1863704	98	
4	5	1	330	39	38	0.4	0.4	NaN	12185.0	0.0	...	0.0	0.0	1.0	1.2	37.80	47.60	Not_CPU_Bound	633	1760253	90	

5 rows × 22 columns

Shape - (8192, 22)

```
<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), in
```

Summary Statistics-

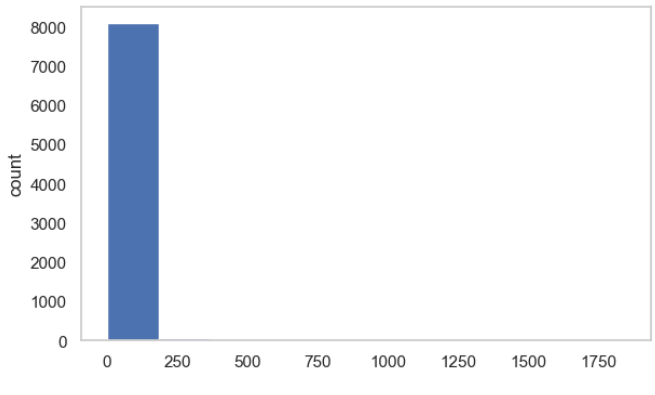
lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgfree	pgscan	atch	pgin	ppgin	pflt	vflt	freemem	freeswap	usr	
count	8192.000000	8192.000000	8192.000000	8192.000000	8192.000000	8192.000000	8192.000000	8.088000e+03	8.177000e+03	8192.000000	...	8192.000000	8192.000000	8192.000000	8192.000000	8192.000000	8192.000000	8192.000000	8192.000000	8.192000e+03	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	21.526849	1.127505	8.277960	12.388586	109.793799	185.315796	1763.456299	1.328126e+06	83.968872
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	71.141340	5.708347	13.874978	22.281318	114.419221	191.000603	2482.104511	4.220194e+05	18.401905
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	0.000000	0.000000	0.000000	0.000000	0.000000	0.200000	55.000000	2.000000e+00	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	0.000000	0.000000	0.600000	0.600000	25.000000	45.400000	231.000000	1.042624e+06	81.000000
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	0.000000	0.000000	2.800000	3.800000	63.800000	120.400000	579.000000	1.289290e+06	89.000000
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	0.000000	0.600000	9.765000	13.800000	159.600000	251.800000	2002.250000	1.730380e+06	94.000000
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	1237.000000	211.580000	141.200000	292.610000	899.800000	1365.000000	12027.000000	2.243187e+06	99.000000

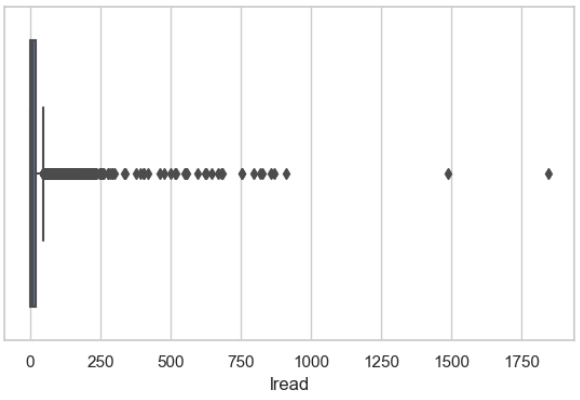
8 rows × 21 columns

- The data has 8192 rows and 22 columns.
- 1 of object data type and 21 of numeric type.
- On an average, 83.9% portion of time, the CPU runs in user mode.
- There are few missing values in rchar and wchar which will be treated later.

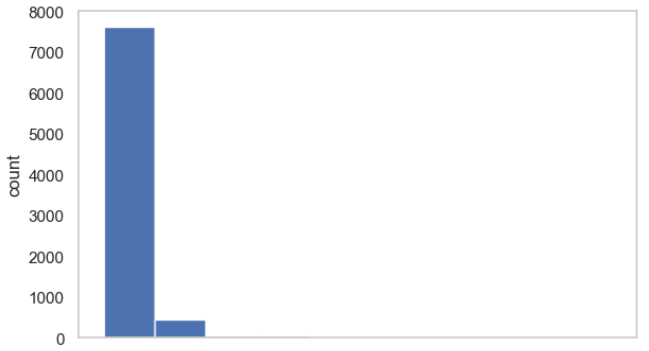
Univariate analysis-

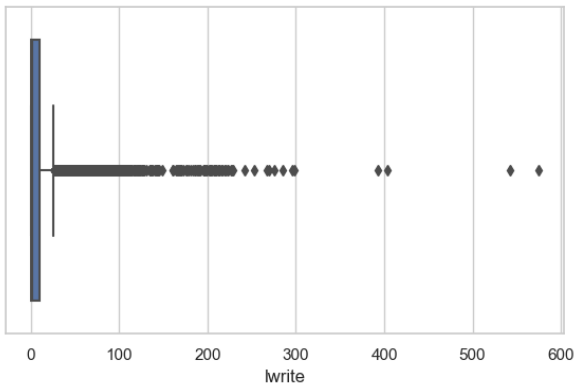
```
Description of lread
--------------------------------------------------------------------------------
count    8192.000000
mean       19.559692
std        53.353799
min         0.000000
25%         2.000000
50%         7.000000
75%        20.000000
max      1845.000000
Name: lread, dtype: float64

Skew : 13.9

Distribution of lread
--------------------------------------------------------------------------------
```



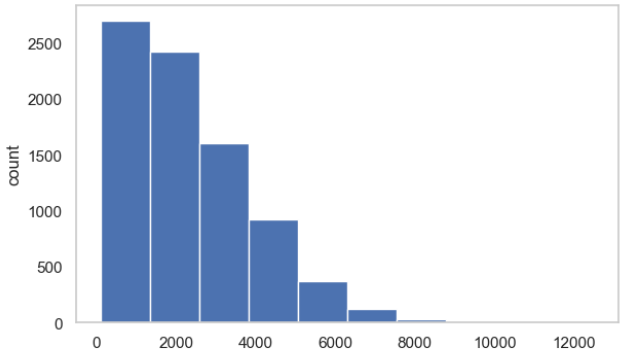


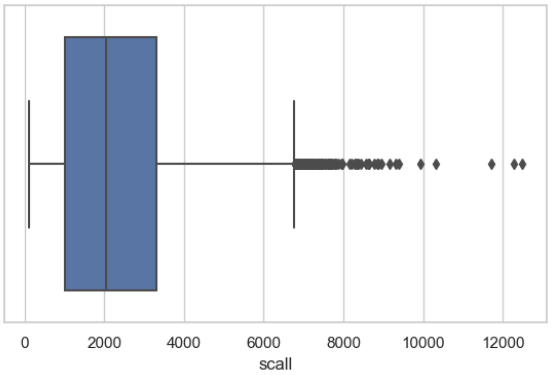
```
Description of lwrite
--------------------------------------------------------------------------------
count    8192.000000
mean       13.106201
std        29.891726
min         0.000000
25%         0.000000
50%         1.000000
75%        10.000000
max       575.000000
Name: lwrite, dtype: float64

Skew : 5.28

Distribution of lwrite
--------------------------------------------------------------------------------
```





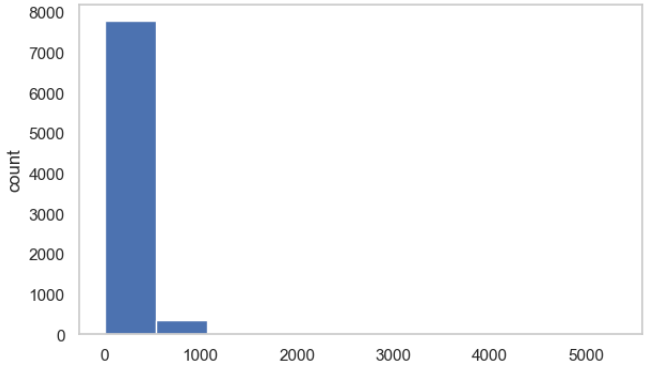

```
Description of scall
--------------------------------------------------------------------------------
count     8192.000000
mean      2306.318237
std       1633.617322
min        109.000000
25%       1012.000000
50%       2051.500000
75%       3317.250000
max      12493.000000
Name: scall, dtype: float64

Skew : 0.9

Distribution of scall
--------------------------------------------------------------------------------
```

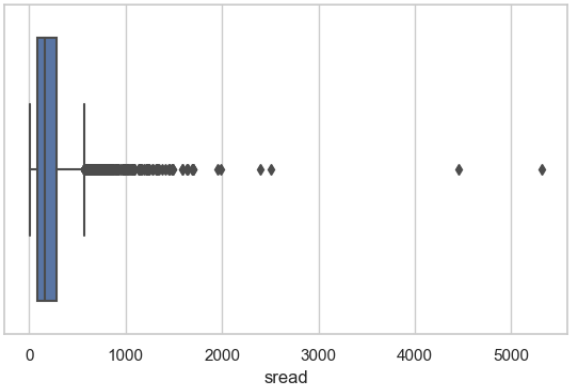




```
Description of sread
--------------------------------------------------------------------------------
count    8192.000000
mean      210.479980
std       198.980146
min         6.000000
25%        86.000000
50%       166.000000
75%       279.000000
max      5318.000000
Name: sread, dtype: float64

Skew : 5.46
```

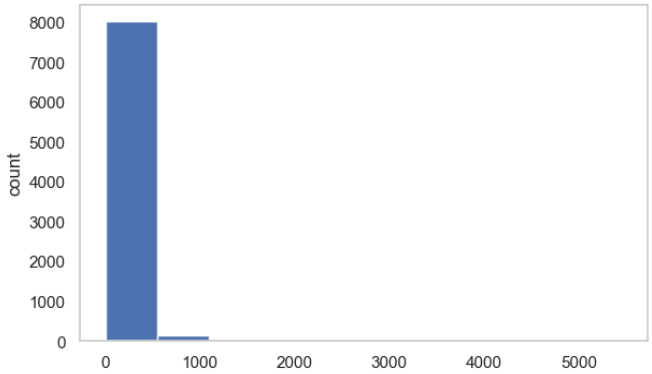
```
Distribution of sread
--------------------------------------------------------------------------------
```

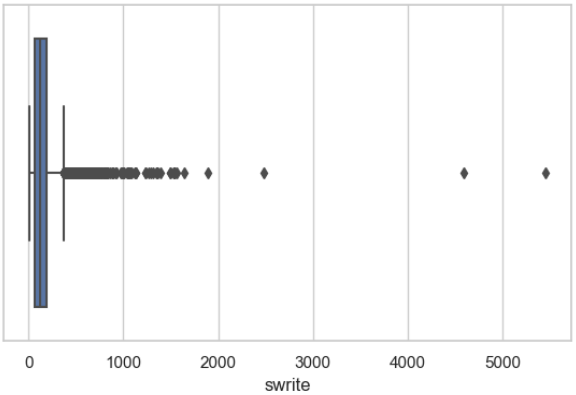




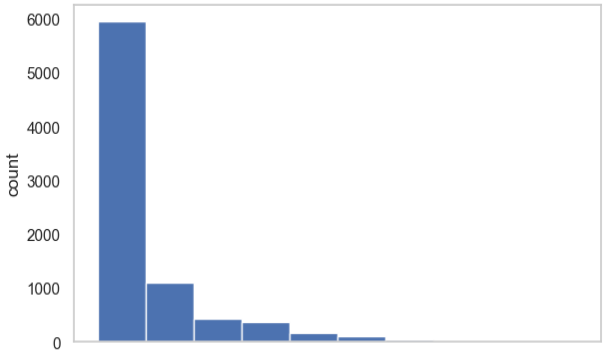

```
Description of swrite
--------------------------------------------------------------------------------
count    8192.000000
mean      150.058228
std       160.478980
min         7.000000
25%        63.000000
50%       117.000000
75%       185.000000
max      5456.000000
Name: swrite, dtype: float64

Skew : 9.61
```

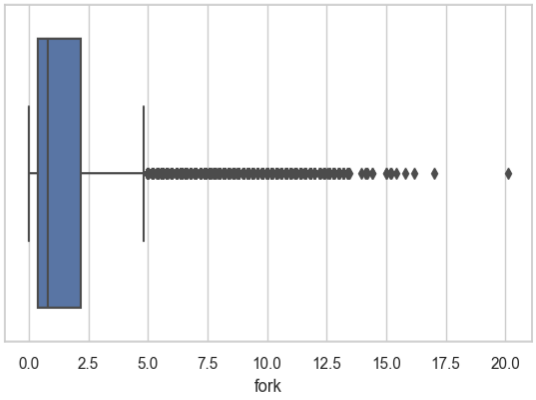
```
Distribution of swrite
--------------------------------------------------------------------------------
```



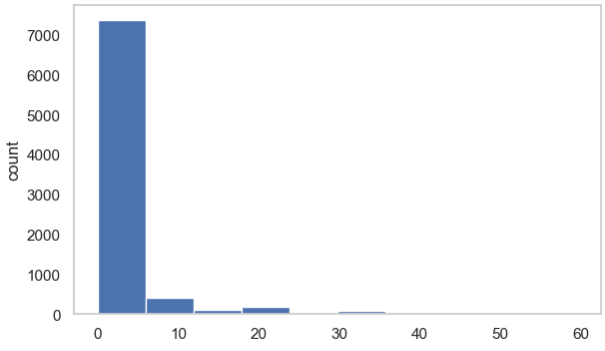


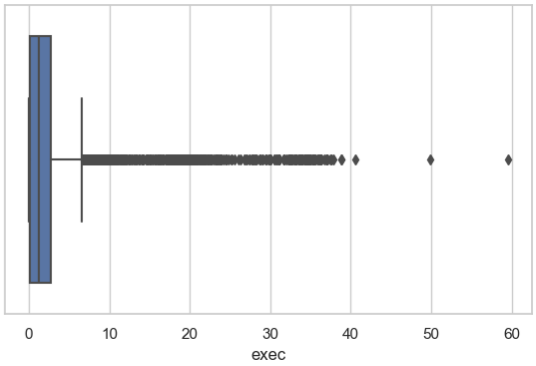
```
Description of fork
--------------------------------------------------------------------------------
count    8192.000000
mean        1.884554
std         2.479493
min         0.000000
25%         0.400000
50%         0.800000
75%         2.200000
max        20.120000
Name: fork, dtype: float64

Skew : 2.25

Distribution of fork
--------------------------------------------------------------------------------
```



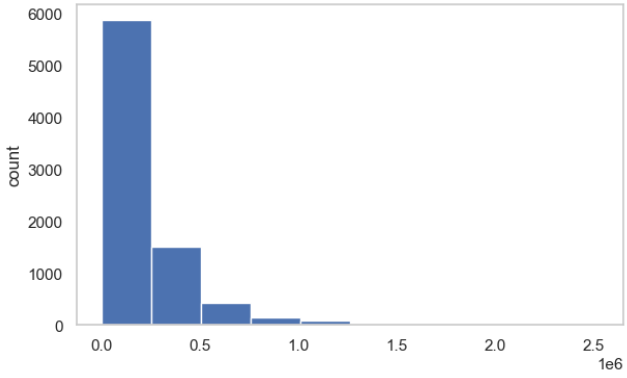


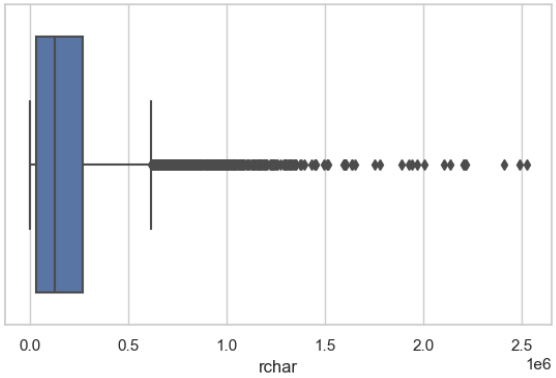

```
Description of exec
--------------------------------------------------------------------------------
count    8192.000000
mean        2.791998
std         5.212456
min         0.000000
25%         0.200000
50%         1.200000
75%         2.800000
max        59.560000
Name: exec, dtype: float64

Skew : 4.07

Distribution of exec
--------------------------------------------------------------------------------
```



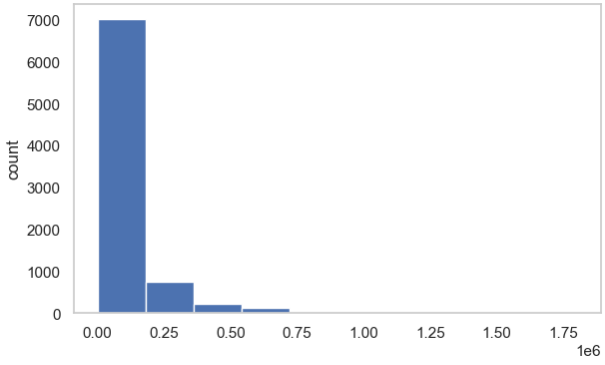


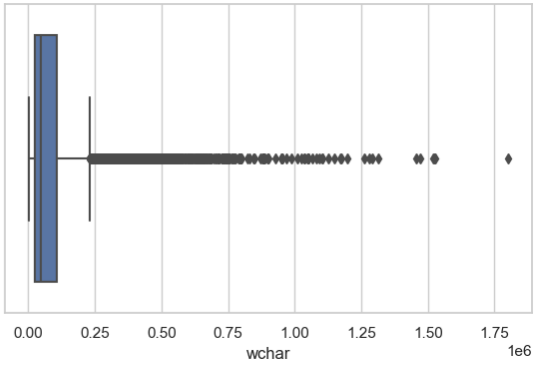
```
Description of rchar
--------------------------------------------------------------------------------
count    8.088000e+03
mean     1.973857e+05
std      2.398375e+05
min      2.780000e+02
25%      3.409150e+04
50%      1.254735e+05
75%      2.678288e+05
max      2.526649e+06
Name: rchar, dtype: float64

Skew : 2.85

Distribution of rchar
--------------------------------------------------------------------------------
```



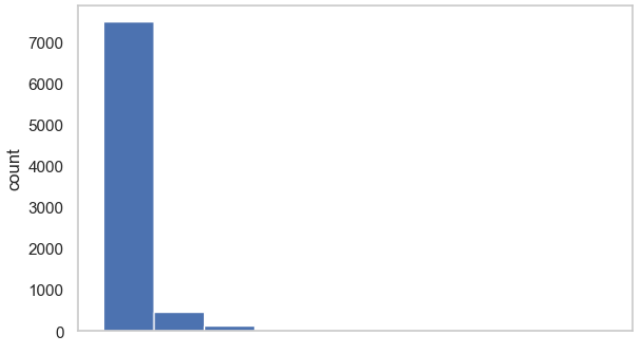


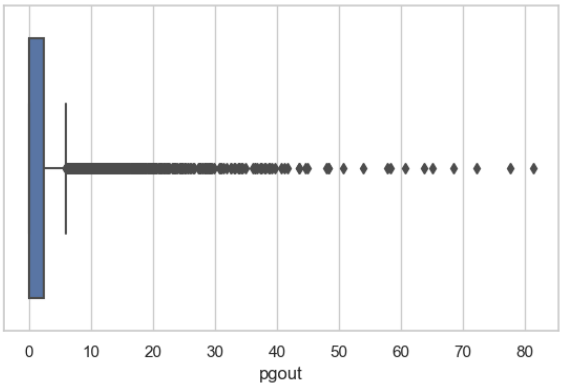

```
Description of wchar
--------------------------------------------------------------------------------
count    8.177000e+03
mean     9.590299e+04
std      1.408417e+05
min      1.498000e+03
25%      2.291600e+04
50%      4.661900e+04
75%      1.061010e+05
max      1.801623e+06
Name: wchar, dtype: float64

Skew : 3.85

Distribution of wchar
--------------------------------------------------------------------------------
```



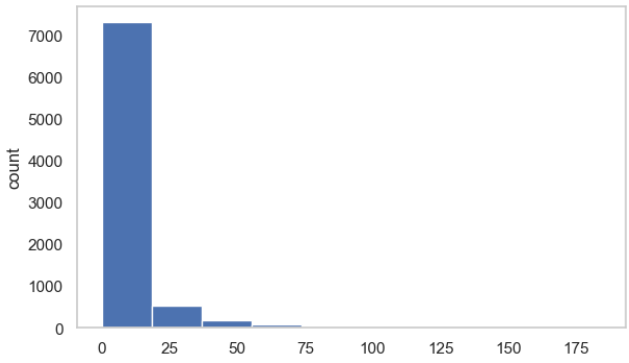


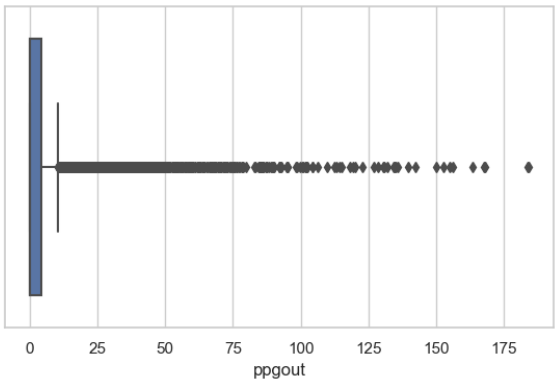
```
Description of pgout
--------------------------------------------------------------------------------
count    8192.000000
mean        2.285317
std         5.307038
min         0.000000
25%         0.000000
50%         0.000000
75%         2.400000
max        81.440000
Name: pgout, dtype: float64

Skew : 5.07

Distribution of pgout
--------------------------------------------------------------------------------
```



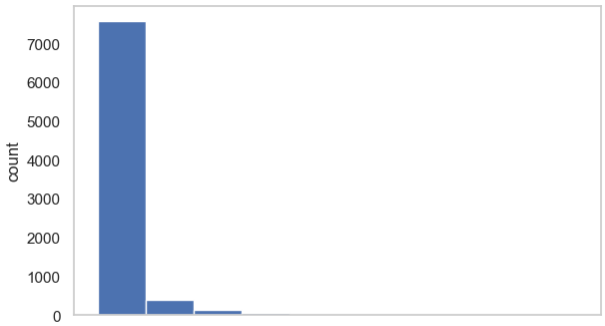


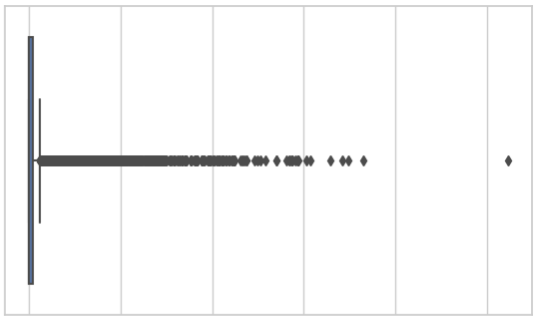

```
Description of ppgout
--------------------------------------------------------------------------------
count    8192.000000
mean        5.977229
std        15.214590
min         0.000000
25%         0.000000
50%         0.000000
75%         4.200000
max       184.200000
Name: ppgout, dtype: float64

Skew : 4.68

Distribution of ppgout
--------------------------------------------------------------------------------
```



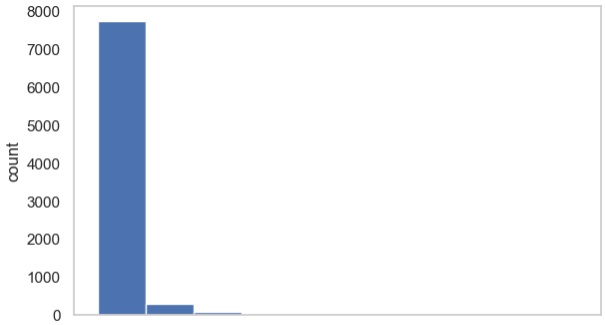


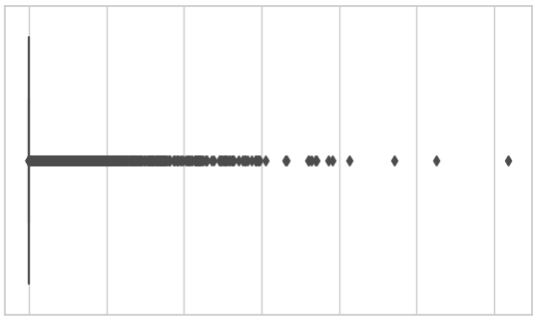
```
Description of pgfree
--------------------------------------------------------------------------------
count    8192.000000
mean       11.919712
std        32.363520
min         0.000000
25%         0.000000
50%         0.000000
75%         5.000000
max       523.000000
Name: pgfree, dtype: float64

Skew : 4.77


Distribution of pgfree
--------------------------------------------------------------------------------
```



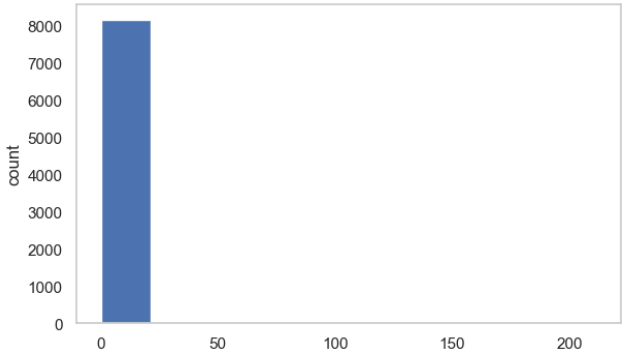


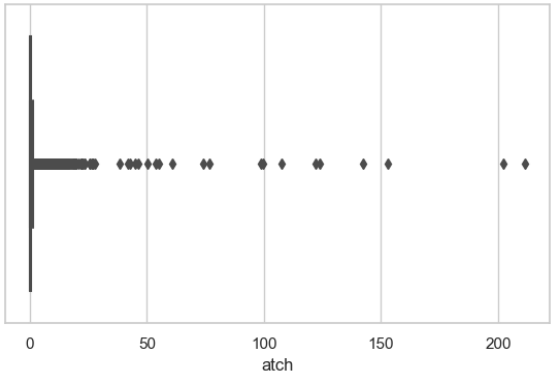

```
Description of pgscan
--------------------------------------------------------------------------------
count    8192.000000
mean       21.526849
std        71.141340
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max      1237.000000
Name: pgscan, dtype: float64

Skew : 5.81


Distribution of pgscan
--------------------------------------------------------------------------------
```



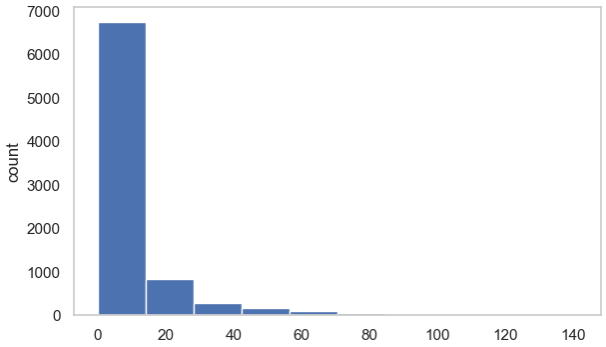


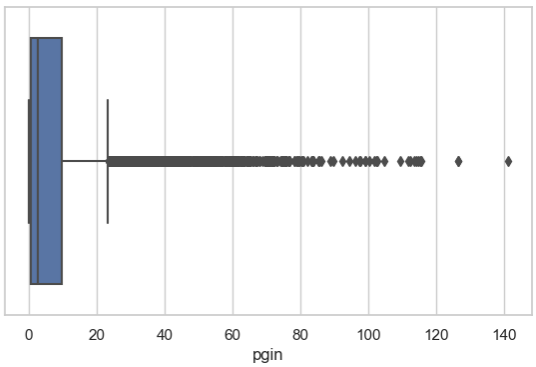
```
Description of atch
--------------------------------------------------------------------------------
count    8192.000000
mean        1.127505
std         5.708347
min         0.000000
25%         0.000000
50%         0.000000
75%         0.600000
max       211.580000
Name: atch, dtype: float64

Skew : 21.54

Distribution of atch
--------------------------------------------------------------------------------
```





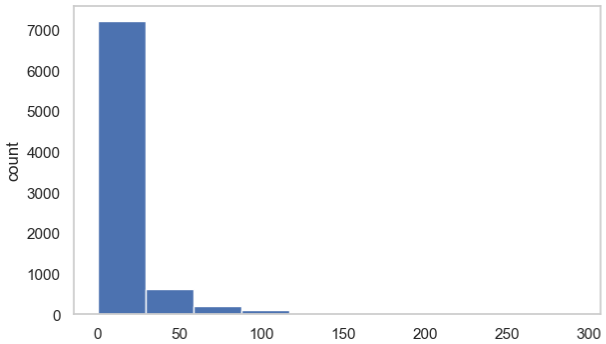

```
Description of pgin
--------------------------------------------------------------------------------
count    8192.000000
mean        8.277960
std        13.874978
min         0.000000
25%         0.600000
50%         2.800000
75%         9.765000
max       141.200000
Name: pgin, dtype: float64

Skew : 3.24

Distribution of pgin
--------------------------------------------------------------------------------
```

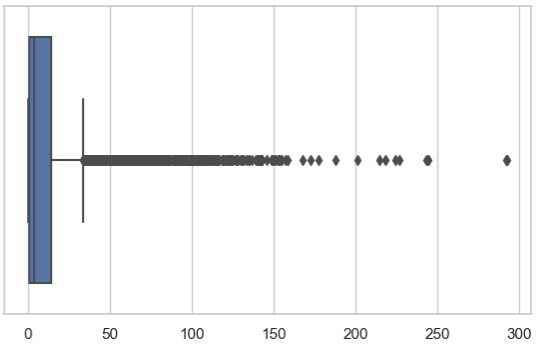




```
Description of ppgin
--------------------------------------------------------------------------------
count    8192.000000
mean       12.388586
std        22.281318
min         0.000000
25%         0.600000
50%         3.800000
75%        13.800000
max       292.610000
Name: ppgin, dtype: float64

Skew : 3.9
```

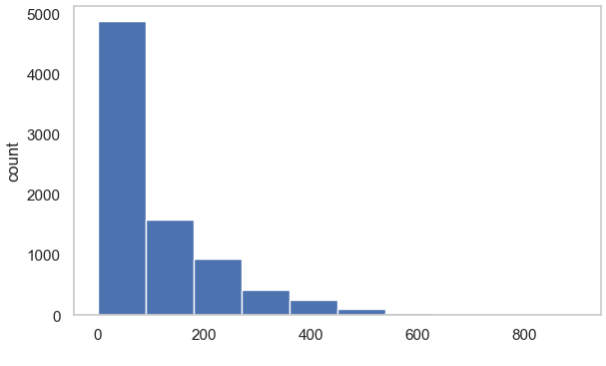
```
Distribution of ppgin
--------------------------------------------------------------------------------
```

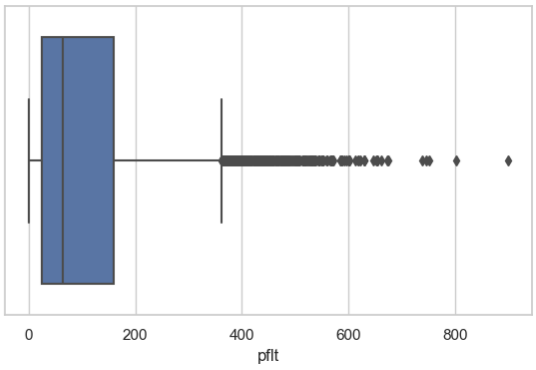




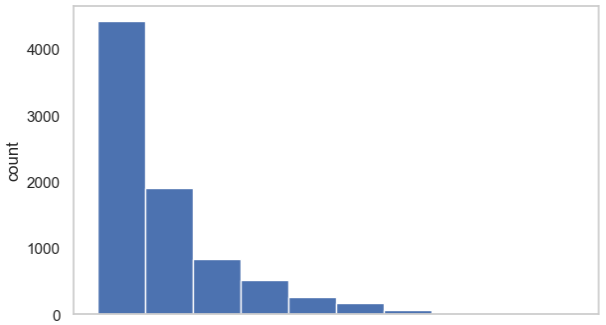

```
Description of pflt
--------------------------------------------------------------------------------
count    8192.000000
mean      109.793799
std       114.419221
min         0.000000
25%        25.000000
50%        63.800000
75%       159.600000
max       899.800000
Name: pflt, dtype: float64

Skew : 1.72
```

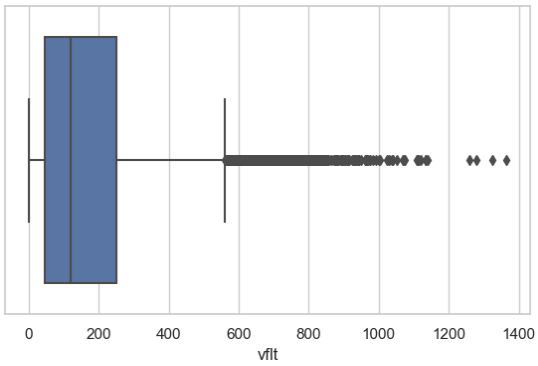
```
Distribution of pflt
--------------------------------------------------------------------------------
```



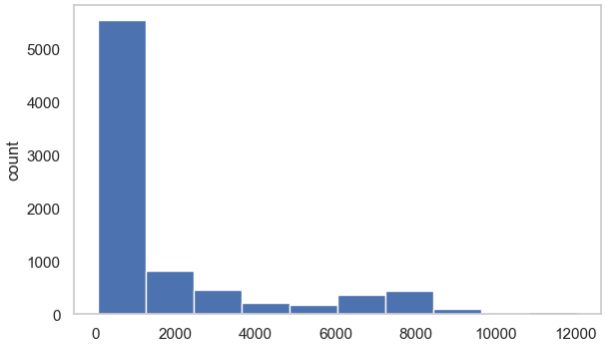


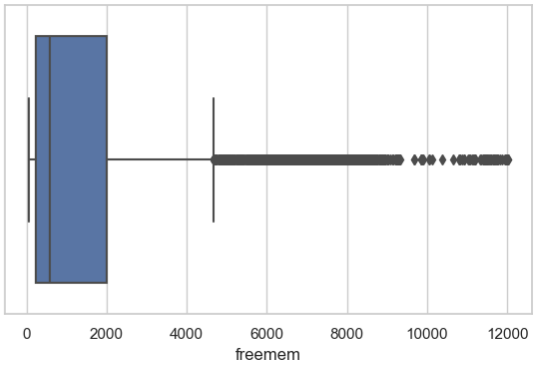

```
Description of vflt
--------------------------------------------------------------------------------
count    8192.000000
mean      185.315796
std       191.000603
min         0.200000
25%        45.400000
50%       120.400000
75%       251.800000
max      1365.000000
Name: vflt, dtype: float64

Skew : 1.74


Distribution of vflt
--------------------------------------------------------------------------------
```



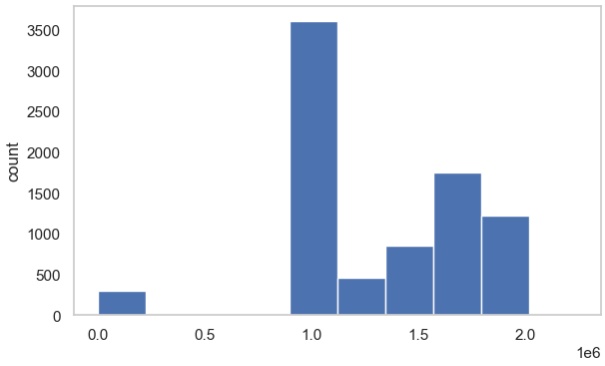


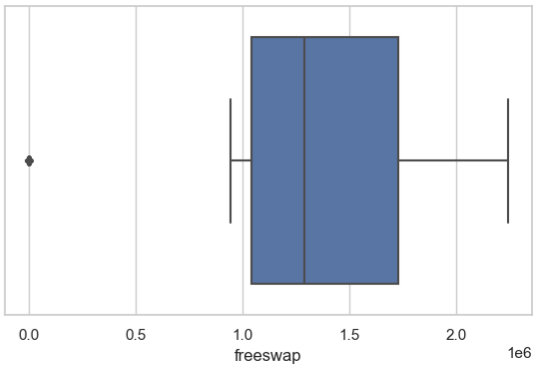

```
Description of freemem
-------------------------------------------------------------------------------
count     8192.000000
mean      1763.456299
std       2482.104511
min         55.000000
25%        231.000000
50%        579.000000
75%       2002.250000
max      12027.000000
Name: freemem, dtype: float64

Skew : 1.81


Distribution of freemem
-------------------------------------------------------------------------------
```





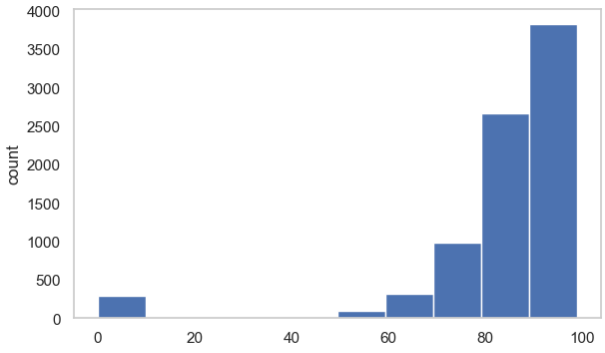

```
Description of freeswap
-------------------------------------------------------------------------------
count    8.192000e+03
mean     1.328126e+06
std      4.220194e+05
min      2.000000e+00
25%      1.042624e+06
50%      1.289290e+06
75%      1.730380e+06
max      2.243187e+06
Name: freeswap, dtype: float64

Skew : -0.79


Distribution of freeswap
-------------------------------------------------------------------------------
```

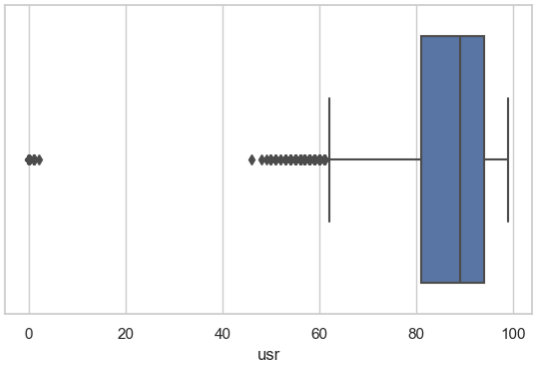




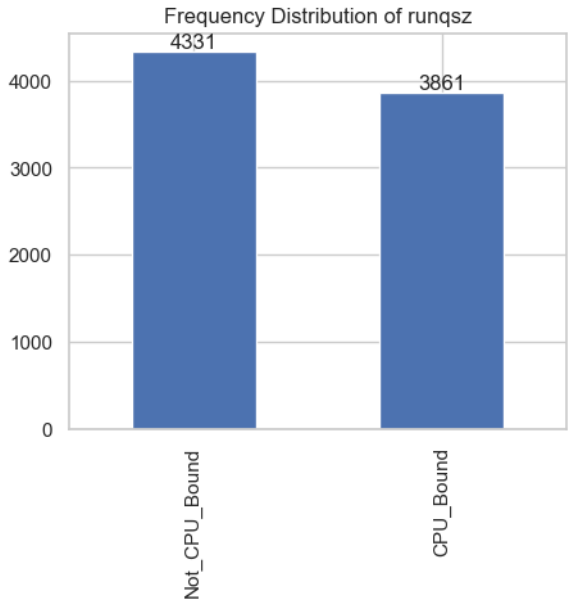

```
Description of usr
-------------------------------------------------------------------------------
count    8192.000000
mean       83.968872
std        18.401905
min         0.000000
25%        81.000000
50%        89.000000
75%        94.000000
max        99.000000
Name: usr, dtype: float64

Skew : -3.42
```

```
Distribution of usr
-------------------------------------------------------------------------------
```



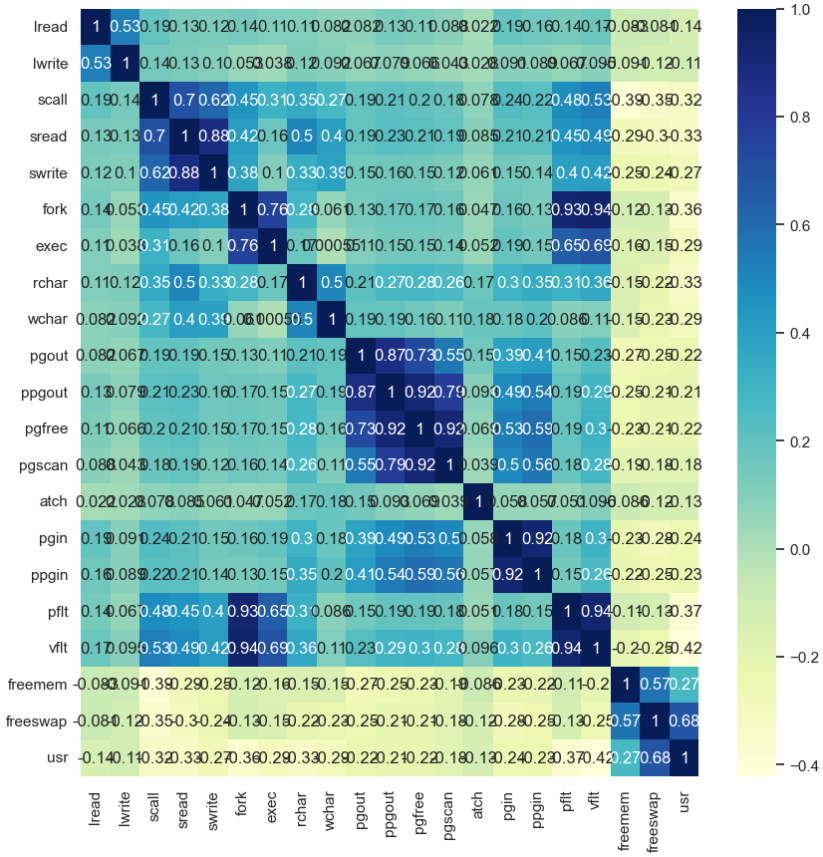



```
Details of runqsz
-------------------------------------------------------------------
Not_CPU_Bound    4331
CPU_Bound        3861
Name: runqsz, dtype: int64
```



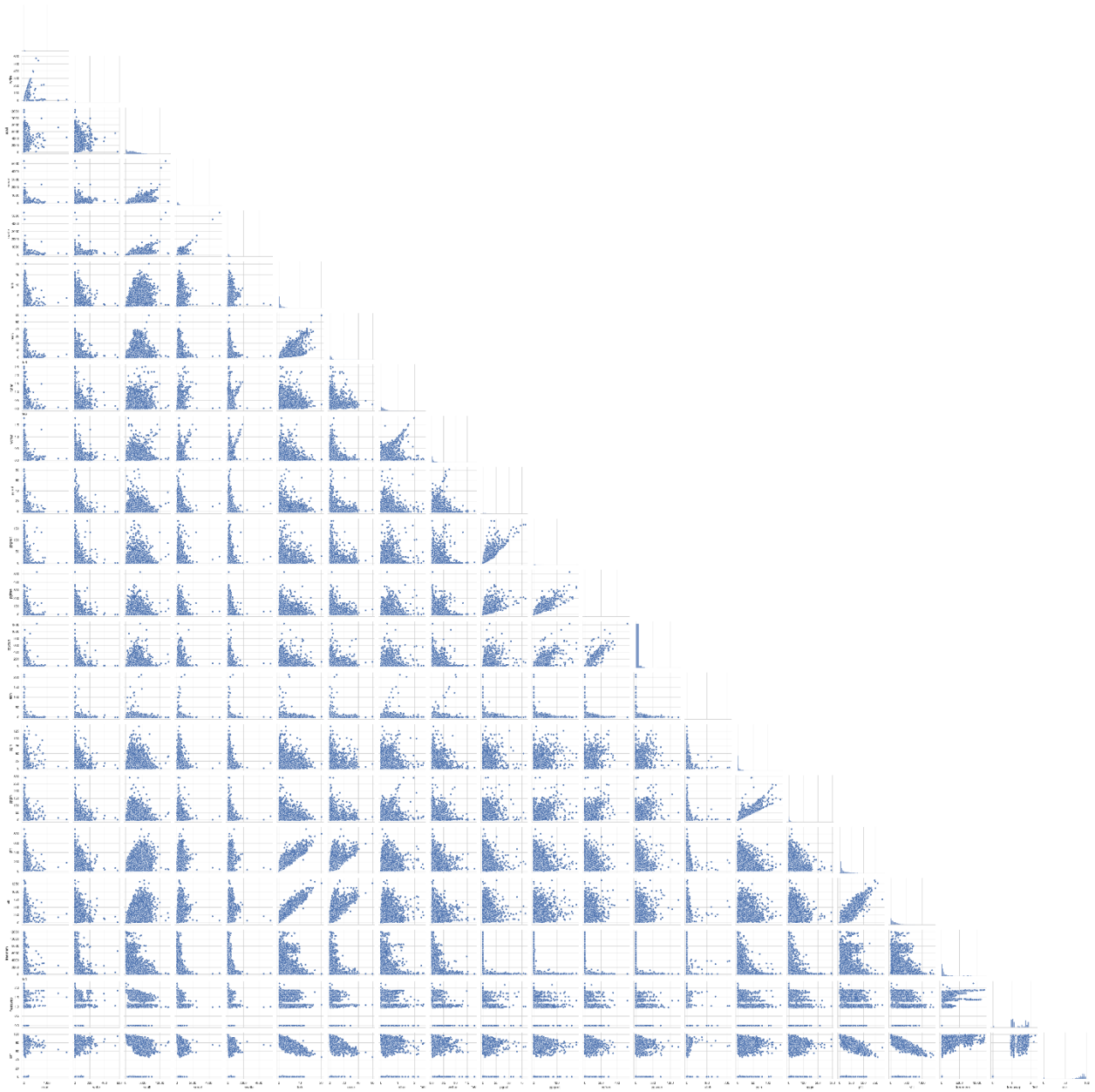
Higher number of systems are Not CPU Bound.

Bivariate Analysis-



We can observe certain amount of correlation between the variables.

Pairwise relationship between the variables-



1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of creating new features if required. Also check for outliers and duplicates if there.

- There are **no duplicates values** present.
- **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
freemem	0
freeswap	0
usr	0
runqsz_Not_CPU_Bound	0
dtype: int64	

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
freemem	0
freeswap	0
usr	0
runqsz_Not_CPU_Bound	0
dtype: int64	

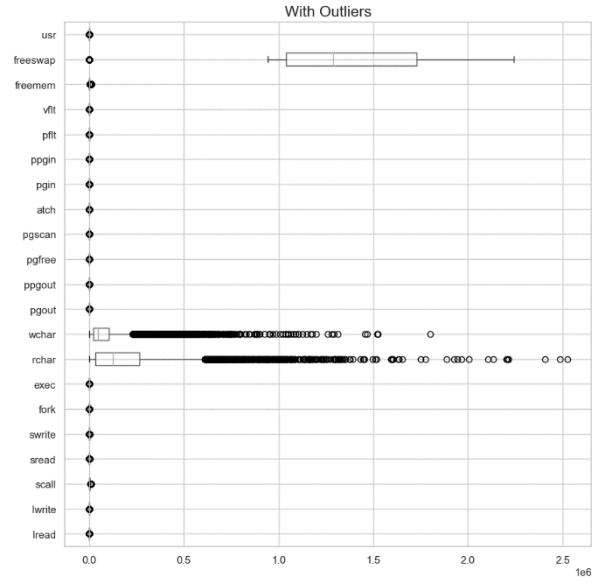
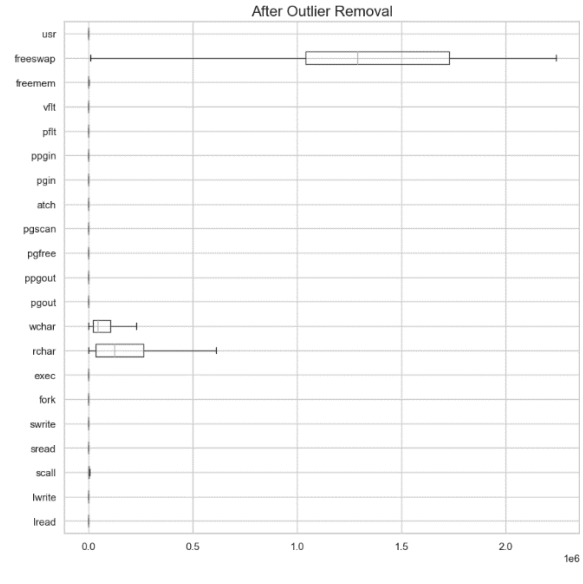
We have used median values for imputing the missing or null values of the continuous numerical variables.

Number of zero values in each column of the Dataframe-

lread	675
lwrite	2684
scall	0
sread	0
swrite	0
fork	21
exec	21
rchar	0
wchar	0
pgout	4878
ppgout	4878
pgfree	4869
pgscan	6448
atch	4575
pgin	1220
ppgin	1220
pflt	3
vflt	0
freemem	0
freeswap	0
usr	283
runqsz_Not_CPU_Bound	3861
dtype: int64	

Since these are valid entries which can contain zero values, we are not treating them.

- **Outliers Treatment-**

Before Treatment	After Treatment
	

We have treated the outlier by adjusting them to the lower and upper bound values by using IQR method.

1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.

For the categorical variable that is nominal (runqsz), we have performed dummy variable encoding. Sample data set post data encoding-

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	freemem	freeswap	usr	runqsz_Not_CPU_Bound	
0	1	0	2147	79	68	0.2	0.2	40671.0	53995.0	0.0	...	0.0	0.0	0.0	1.6	2.6	16.00	26.40	4670	1730946	95	0
1	0	0	170	18	21	0.2	0.2	448.0	8385.0	0.0	...	0.0	0.0	0.0	0.0	0.0	15.63	16.83	7278	1869002	97	1
2	15	3	2162	159	119	2.0	2.4	NaN	31950.0	0.0	...	0.0	1.2	6.0	9.4	150.20	220.20	702	1021237	87	1	
3	0	0	160	12	16	0.2	0.2	NaN	8670.0	0.0	...	0.0	0.0	0.2	0.2	15.60	16.80	7248	1863704	98	1	
4	5	1	330	39	38	0.4	0.4	NaN	12185.0	0.0	...	0.0	0.0	1.0	1.2	37.80	47.60	633	1760253	90	1	

5 rows × 22 columns

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

X dataframe containing the predictor variables-

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgfree	pgscan	atch	pgin	ppgin	pflt	vflt	freemem	freeswap	runqsz_Not_CPU_Bound	
0	1.0	0.0	2147.0	79.0	68.0	0.2	0.2	40671.0	53995.0	0.0	...	0.0	0.0	0.0	1.6	2.6	16.00	26.40	4659.125	1730946.0	0	
1	0.0	0.0	170.0	18.0	21.0	0.2	0.2	448.0	8385.0	0.0	...	0.0	0.0	0.0	0.0	0.0	15.63	16.83	4659.125	1869002.0	1	
2	15.0	3.0	2162.0	159.0	119.0	2.0	2.4	125473.5	31950.0	0.0	...	0.0	0.0	1.2	6.0	9.4	150.20	220.20	702.000	1021237.0	1	
3	0.0	0.0	160.0	12.0	16.0	0.2	0.2	125473.5	8670.0	0.0	...	0.0	0.0	0.0	0.2	0.2	15.60	16.80	4659.125	1863704.0	1	
4	5.0	1.0	330.0	39.0	38.0	0.4	0.4	125473.5	12185.0	0.0	...	0.0	0.0	0.0	1.0	1.2	37.80	47.60	633.000	1760253.0	1	

5 rows × 21 columns

Linear Regression using statsmodel (OLS)-

Regression Summary-

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.796
Model:                            OLS   Adj. R-squared:                  0.795
Method:                 Least Squares   F-statistic:                     1115.
Date:                Sat, 14 Oct 2023   Prob (F-statistic):               0.00
Time:                        17:54:30   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     4.002e-14   1.62e-16    247.538      0.000    3.97e-14    4.03e-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.                     2.95e+22
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.31e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

- **Interpretation of R-squared-**

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

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

  (P > |t|) gives the p-value for each predictor variable to check the null hypothesis.

  Null hypothesis: Predictor variable is not significant
  Alternate hypothesis: Predictor variable is significant

  p-values greater than 0.05 would indicate that the corresponding predictor variables are not significant. However, due to the presence of multicollinearity in our data, the p-values will also change. We need to ensure that there is no multicollinearity in order to interpret the p-values.

- **Multicollinearity check-**

  We are using Variation Inflation Factor (VIF) for testing multicollinearity. It measures the degree of multicollinearity for each variable.
  - If VIF is 1, then there is no correlation among the predictor and the remaining predictor variables.
  - If VIF exceeds 5, there is moderate VIF.
  - And if it is 10 or greater than 10, it shows high multi-collinearity.

  VIF of the predictors-

```
VIF values:

const                 29.229332
lread                  5.350560
lwrite                 4.328397
scall                  2.960609
sread                  6.420172
swrite                 5.597135
fork                  13.035359
exec                   3.241417
rchar                  2.133616
wchar                  1.584381
pgout                 11.360363
ppgout                29.404223
pgfree                16.496748
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
```

  - The VIF values indicate that the features **lread, sread, swrite, fork, pgout, ppgout, pgfree, pgscan, pgin, ppgin, pflt, vflt** are correlated with one or more independent features.

  - To treat multicollinearity, we will have to drop one or more of the correlated features. We will drop the variable that has the least impact on the adjusted R-squared of the model.

Model 1-

Dropping ppgout since it has high VIF value and dropping it does not impact the adj. R squared value

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.796
Model:	OLS	Adj. R-squared:	0.795
Method:	Least Squares	F-statistic:	1174.
Date:	Sat, 14 Oct 2023	Prob (F-statistic):	0.00
Time:	19:58:20	Log-Likelihood:	-16658.
No. Observations:	5734	AIC:	3.336e+04
Df Residuals:	5714	BIC:	3.349e+04
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.1477	0.315	267.138	0.000	83.530	84.765
lread	-0.0635	0.009	-7.077	0.000	-0.081	-0.046
lwrite	0.0482	0.013	3.675	0.000	0.022	0.074
scall	-0.0007	6.28e-05	-10.575	0.000	-0.001	-0.001
sread	0.0003	0.001	0.303	0.762	-0.002	0.002
swrite	-0.0054	0.001	-3.782	0.000	-0.008	-0.003
fork	0.0325	0.132	0.247	0.805	-0.226	0.291
exec	-0.3225	0.052	-6.247	0.000	-0.424	-0.221
rchar	-5.166e-06	4.88e-07	-10.598	0.000	-6.12e-06	-4.21e-06
wchar	-5.45e-06	1.03e-06	-5.283	0.000	-7.47e-06	-3.43e-06
pgout	-0.4264	0.068	-6.286	0.000	-0.559	-0.293
pgfree	0.0477	0.029	1.634	0.102	-0.010	0.105
pgscan	5.609e-14	2.69e-16	208.237	0.000	5.56e-14	5.66e-14
atch	0.6295	0.143	4.407	0.000	0.349	0.909
pgin	0.0212	0.028	0.745	0.456	-0.035	0.077
ppgin	-0.0685	0.020	-3.482	0.001	-0.107	-0.030
pflt	-0.0336	0.002	-16.957	0.000	-0.037	-0.030
vflt	-0.0055	0.001	-3.846	0.000	-0.008	-0.003
freemem	-0.0005	5.07e-05	-9.074	0.000	-0.001	-0.000
freeswap	8.824e-06	1.9e-07	46.472	0.000	8.45e-06	9.2e-06
runqsz_Not_CPU_Bound	1.6130	0.126	12.804	0.000	1.366	1.860

Omnibus:	1102.077	Durbin-Watson:	2.016
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2366.754
Skew:	-1.118	Prob(JB):	0.00
Kurtosis:	5.216	Cond. No.	5.81e+22

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.38e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

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

Model 2-

Dropping ppgout, vflt since it has high VIF value and dropping it does not impact the adj. R squared value

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.796
Model:	OLS	Adj. R-squared:	0.795
Method:	Least Squares	F-statistic:	1235.
Date:	Sat, 14 Oct 2023	Prob (F-statistic):	0.00
Time:	19:58:28	Log-Likelihood:	-16665.
No. Observations:	5734	AIC:	3.337e+04
Df Residuals:	5715	BIC:	3.349e+04
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.0090	0.313	268.139	0.000	83.395	84.623
lread	-0.0654	0.009	-7.281	0.000	-0.083	-0.048
lwrite	0.0491	0.013	3.735	0.000	0.023	0.075
scall	-0.0007	6.28e-05	-10.769	0.000	-0.001	-0.001
sread	-2.068e-05	0.001	-0.021	0.984	-0.002	0.002
swrite	-0.0053	0.001	-3.720	0.000	-0.008	-0.003
fork	-0.2082	0.116	-1.793	0.073	-0.436	0.019
exec	-0.3293	0.052	-6.376	0.000	-0.431	-0.228
rchar	-5.294e-06	4.87e-07	-10.871	0.000	-6.25e-06	-4.34e-06
wchar	-4.982e-06	1.03e-06	-4.858	0.000	-6.99e-06	-2.97e-06
pgout	-0.4205	0.068	-6.194	0.000	-0.554	-0.287
pgfree	0.0408	0.029	1.397	0.162	-0.016	0.098
pgscan	-1.477e-14	6.82e-17	-216.517	0.000	-1.49e-14	-1.46e-14
atch	0.5868	0.143	4.116	0.000	0.307	0.866
pgin	0.0086	0.028	0.305	0.760	-0.047	0.064
ppgin	-0.0685	0.020	-3.476	0.001	-0.107	-0.030
pflt	-0.0373	0.002	-21.570	0.000	-0.041	-0.034
freemem	-0.0005	5.07e-05	-9.165	0.000	-0.001	-0.000
freeswap	8.945e-06	1.87e-07	47.712	0.000	8.58e-06	9.31e-06
runqsz_Not_CPU_Bound	1.6096	0.126	12.761	0.000	1.362	1.857

Omnibus:	1058.324	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2225.362
Skew:	-1.085	Prob(JB):	0.00
Kurtosis:	5.145	Cond. No.	5.21e+23

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 4.2e-32. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

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
pgscan	NaN
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

Model 3-
Dropping ppgout, vflt, ppgin since it has high VIF value and dropping it does not impact the adj. R squared value

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.795
Model:	OLS	Adj. R-squared:	0.795
Method:	Least Squares	F-statistic:	1305.
Date:	Sat, 14 Oct 2023	Prob (F-statistic):	0.00
Time:	19:58:34	Log-Likelihood:	-16671.
No. Observations:	5734	AIC:	3.338e+04
Df Residuals:	5716	BIC:	3.350e+04
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.0531	0.313	268.240	0.000	83.439	84.667
lread	-0.0678	0.009	-7.563	0.000	-0.085	-0.050
lwrite	0.0513	0.013	3.909	0.000	0.026	0.077
scall	-0.0007	6.29e-05	-10.692	0.000	-0.001	-0.001
sread	-4.826e-06	0.001	-0.005	0.996	-0.002	0.002
swrite	-0.0054	0.001	-3.732	0.000	-0.008	-0.003
fork	-0.1927	0.116	-1.659	0.097	-0.420	0.035
exec	-0.3290	0.052	-6.364	0.000	-0.430	-0.228
rchar	-5.506e-06	4.84e-07	-11.385	0.000	-6.45e-06	-4.56e-06
wchar	-4.978e-06	1.03e-06	-4.849	0.000	-6.99e-06	-2.97e-06
pgout	-0.4138	0.068	-6.091	0.000	-0.547	-0.281
pgfree	0.0311	0.029	1.070	0.284	-0.026	0.088
pgscan	2.352e-14	1.11e-16	212.653	0.000	2.33e-14	2.37e-14
atch	0.5966	0.143	4.181	0.000	0.317	0.876
pgin	-0.0839	0.009	-8.848	0.000	-0.103	-0.065
pflt	-0.0374	0.002	-21.568	0.000	-0.041	-0.034
freemem	-0.0005	5.08e-05	-9.196	0.000	-0.001	-0.000
freeswap	8.922e-06	1.88e-07	47.572	0.000	8.55e-06	9.29e-06
runqsz_Not_CPU_Bound	1.6017	0.126	12.689	0.000	1.354	1.849

Omnibus:	1052.296	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2207.367
Skew:	-1.081	Prob(JB):	0.00
Kurtosis:	5.137	Cond. No.	2.80e+22

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.45e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

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
pgscan	NaN
atch	1.863536
pgin	1.529142
pflt	9.131545
freemem	1.957713
freeswap	1.785393
runqsz_Not_CPU_Bound	1.155990

dtype: float64

Model 4-

Dropping ppgout, vflt, ppgin, fork since it has high VIF value and dropping it does not impact the adj. R squared value

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.795
Model:	OLS	Adj. R-squared:	0.794
Method:	Least Squares	F-statistic:	1386.
Date:	Sat, 14 Oct 2023	Prob (F-statistic):	0.00
Time:	19:58:42	Log-Likelihood:	-16672.
No. Observations:	5734	AIC:	3.338e+04
Df Residuals:	5717	BIC:	3.349e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.0913	0.313	269.048	0.000	83.479	84.704
lread	-0.0686	0.009	-7.675	0.000	-0.086	-0.051
lwrite	0.0527	0.013	4.025	0.000	0.027	0.078
scall	-0.0007	6.25e-05	-10.572	0.000	-0.001	-0.001
sread	1.712e-05	0.001	0.017	0.986	-0.002	0.002
swrite	-0.0058	0.001	-4.126	0.000	-0.009	-0.003
exec	-0.3583	0.049	-7.375	0.000	-0.454	-0.263
rchar	-5.524e-06	4.84e-07	-11.423	0.000	-6.47e-06	-4.58e-06
wchar	-4.854e-06	1.02e-06	-4.740	0.000	-6.86e-06	-2.85e-06
pgout	-0.4133	0.068	-6.084	0.000	-0.547	-0.280
pgfree	0.0307	0.029	1.054	0.292	-0.026	0.088
pgscan	6.774e-14	3.46e-16	195.552	0.000	6.71e-14	6.84e-14
atch	0.6020	0.143	4.220	0.000	0.322	0.882
pgin	-0.0833	0.009	-8.789	0.000	-0.102	-0.065
pflt	-0.0396	0.001	-37.167	0.000	-0.042	-0.038
freemem	-0.0005	5.08e-05	-9.222	0.000	-0.001	-0.000
freeswap	8.91e-06	1.87e-07	47.536	0.000	8.54e-06	9.28e-06
runqsz_Not_CPU_Bound	1.5972	0.126	12.654	0.000	1.350	1.845

Omnibus:	1044.060	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2198.744
Skew:	-1.071	Prob(JB):	0.00
Kurtosis:	5.148	Cond. No.	4.56e+21

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 5.49e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

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
pgscan	NaN
atch	1.862553
pgin	1.526800
pflt	3.458168
freemem	1.957226
freeswap	1.782829
runqsz_Not_CPU_Bound	1.155448

dtype: float64

Model 5-

Dropping ppgout, vflt, ppgin, fork, sread since it has high VIF value and dropping it does not impact the adj. R squared value

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.795
Model:	OLS	Adj. R-squared:	0.794
Method:	Least Squares	F-statistic:	1478.
Date:	Sat, 14 Oct 2023	Prob (F-statistic):	0.00
Time:	19:58:49	Log-Likelihood:	-16672.
No. Observations:	5734	AIC:	3.338e+04
Df Residuals:	5718	BIC:	3.348e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.0916	0.312	269.421	0.000	83.480	84.703
lread	-0.0686	0.009	-7.682	0.000	-0.086	-0.051
lwrite	0.0528	0.013	4.031	0.000	0.027	0.078
scall	-0.0007	5.97e-05	-11.068	0.000	-0.001	-0.001
swrite	-0.0058	0.001	-5.503	0.000	-0.008	-0.004
exec	-0.3584	0.049	-7.385	0.000	-0.454	-0.263
rchar	-5.52e-06	4.33e-07	-12.758	0.000	-6.37e-06	-4.67e-06
wchar	-4.856e-06	1.02e-06	-4.763	0.000	-6.85e-06	-2.86e-06
pgout	-0.4133	0.068	-6.084	0.000	-0.547	-0.280
pgfree	0.0307	0.029	1.054	0.292	-0.026	0.088
pgscan	-3.693e-14	1.65e-16	-224.373	0.000	-3.73e-14	-3.66e-14
atch	0.6019	0.143	4.221	0.000	0.322	0.881
pgin	-0.0833	0.009	-8.793	0.000	-0.102	-0.065
pflt	-0.0396	0.001	-37.287	0.000	-0.042	-0.038
freemem	-0.0005	5.08e-05	-9.224	0.000	-0.001	-0.000
freeswap	8.91e-06	1.87e-07	47.722	0.000	8.54e-06	9.28e-06
runqsz_Not_CPU_Bound	1.5972	0.126	12.655	0.000	1.350	1.845

Omnibus:	1044.101	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2198.884
Skew:	-1.071	Prob(JB):	0.00
Kurtosis:	5.148	Cond. No.	4.52e+22

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 5.6e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

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
pgscan	NaN
atch	1.861273
pgin	1.525797
pflt	3.436271
freemem	1.956658
freeswap	1.769115
runqsz_Not_CPU_Bound	1.155441

dtype: float64

Model 6-

Dropping ppgout, vflt, ppgin, fork, sread, pgfree since pgfree p-value > 0.05

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.795
Model:	OLS	Adj. R-squared:	0.794
Method:	Least Squares	F-statistic:	1584.
Date:	Sat, 14 Oct 2023	Prob (F-statistic):	0.00
Time:	19:58:51	Log-Likelihood:	-16673.
No. Observations:	5734	AIC:	3.338e+04
Df Residuals:	5719	BIC:	3.348e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.0919	0.312	269.420	0.000	83.480	84.704
lread	-0.0684	0.009	-7.653	0.000	-0.086	-0.051
lwrite	0.0523	0.013	3.995	0.000	0.027	0.078
scall	-0.0007	5.96e-05	-11.111	0.000	-0.001	-0.001
swrite	-0.0058	0.001	-5.481	0.000	-0.008	-0.004
exec	-0.3568	0.049	-7.355	0.000	-0.452	-0.262
rchar	-5.511e-06	4.33e-07	-12.740	0.000	-6.36e-06	-4.66e-06
wchar	-4.872e-06	1.02e-06	-4.779	0.000	-6.87e-06	-2.87e-06
pgout	-0.3540	0.038	-9.287	0.000	-0.429	-0.279
pgscan	-8.181e-14	2.9e-16	-282.464	0.000	-8.24e-14	-8.12e-14
atch	0.6055	0.143	4.247	0.000	0.326	0.885
pgin	-0.0820	0.009	-8.730	0.000	-0.100	-0.064
pflt	-0.0396	0.001	-37.292	0.000	-0.042	-0.038
freemem	-0.0005	5.06e-05	-9.328	0.000	-0.001	-0.000
freeswap	8.915e-06	1.87e-07	47.769	0.000	8.55e-06	9.28e-06
runqsz_Not_CPU_Bound	1.5953	0.126	12.641	0.000	1.348	1.843

Omnibus:	1045.912	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2203.816
Skew:	-1.073	Prob(JB):	0.00
Kurtosis:	5.150	Cond. No.	1.05e+22

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.04e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

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
pgscan	NaN
atch	1.860242
pgin	1.497984
pflt	3.436202
freemem	1.945888
freeswap	1.767780
runqsz_Not_CPU_Bound	1.155214

dtype: float64

Model 7-

Dropping ppgout, vflt, ppgin, fork, sread, pgfree, pgscan since its VIF value is NaN which means it has some strong multi-collinearity

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.795
Model:	OLS	Adj. R-squared:	0.794
Method:	Least Squares	F-statistic:	1584.
Date:	Sat, 14 Oct 2023	Prob (F-statistic):	0.00
Time:	19:58:54	Log-Likelihood:	-16673.
No. Observations:	5734	AIC:	3.338e+04
Df Residuals:	5719	BIC:	3.348e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.0919	0.312	269.420	0.000	83.480	84.704
lread	-0.0684	0.009	-7.653	0.000	-0.086	-0.051
lwrite	0.0523	0.013	3.995	0.000	0.027	0.078
scall	-0.0007	5.96e-05	-11.111	0.000	-0.001	-0.001
swrite	-0.0058	0.001	-5.481	0.000	-0.008	-0.004
exec	-0.3568	0.049	-7.355	0.000	-0.452	-0.262
rchar	-5.511e-06	4.33e-07	-12.740	0.000	-6.36e-06	-4.66e-06
wchar	-4.872e-06	1.02e-06	-4.779	0.000	-6.87e-06	-2.87e-06
pgout	-0.3540	0.038	-9.287	0.000	-0.429	-0.279
atch	0.6055	0.143	4.247	0.000	0.326	0.885
pgin	-0.0820	0.009	-8.730	0.000	-0.100	-0.064
pflt	-0.0396	0.001	-37.292	0.000	-0.042	-0.038
freemem	-0.0005	5.06e-05	-9.328	0.000	-0.001	-0.000
freeswap	8.915e-06	1.87e-07	47.769	0.000	8.55e-06	9.28e-06
runqsz_Not_CPU_Bound	1.5953	0.126	12.641	0.000	1.348	1.843

Omnibus:	1045.912	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2203.816
Skew:	-1.073	Prob(JB):	0.00
Kurtosis:	5.150	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.

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

Final Model-

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.795
Model:	OLS	Adj. R-squared:	0.794
Method:	Least Squares	F-statistic:	1584.
Date:	Sat, 14 Oct 2023	Prob (F-statistic):	0.00
Time:	19:59:29	Log-Likelihood:	-16673.
No. Observations:	5734	AIC:	3.338e+04
Df Residuals:	5719	BIC:	3.348e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.0919	0.312	269.420	0.000	83.480	84.704
lread	-0.0684	0.009	-7.653	0.000	-0.086	-0.051
lwrite	0.0523	0.013	3.995	0.000	0.027	0.078
scall	-0.0007	5.96e-05	-11.111	0.000	-0.001	-0.001
swrite	-0.0058	0.001	-5.481	0.000	-0.008	-0.004
exec	-0.3568	0.049	-7.355	0.000	-0.452	-0.262
rchar	-5.511e-06	4.33e-07	-12.740	0.000	-6.36e-06	-4.66e-06
wchar	-4.872e-06	1.02e-06	-4.779	0.000	-6.87e-06	-2.87e-06
pgout	-0.3540	0.038	-9.287	0.000	-0.429	-0.279
atch	0.6055	0.143	4.247	0.000	0.326	0.885
pgin	-0.0820	0.009	-8.730	0.000	-0.100	-0.064
pflt	-0.0396	0.001	-37.292	0.000	-0.042	-0.038
freemem	-0.0005	5.06e-05	-9.328	0.000	-0.001	-0.000
freeswap	8.915e-06	1.87e-07	47.769	0.000	8.55e-06	9.28e-06
runqsz_Not_CPU_Bound	1.5953	0.126	12.641	0.000	1.348	1.843

Omnibus:	1045.912	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2203.816
Skew:	-1.073	Prob(JB):	0.00
Kurtosis:	5.150	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.

- We have now dropped variables one by one, observed those that do not impact the adj. R-squared value, compared each model and its effect on our predictive model.
- We have dropped the features (ppgout, vflt, ppgin, fork, sread, pgfree, pgscan) causing strong multicollinearity, and yet our model performance hasn't dropped sharply. This shows that these variables did not have much impact on prediction.

Linear Regression Assumptions-

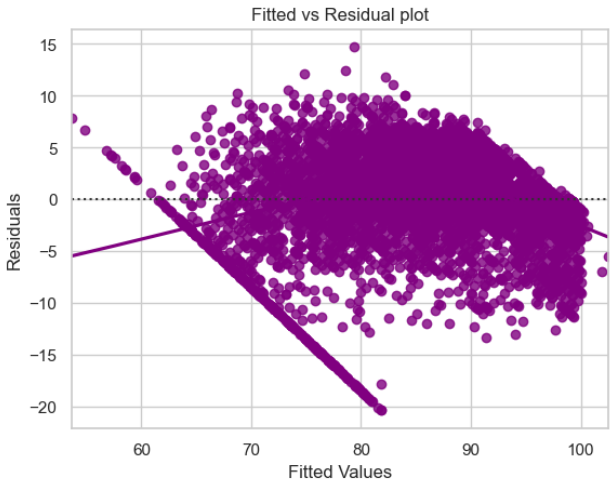
For Linear Regression, we need to check if the following assumptions hold-

1. Linearity
2. Independence
3. Homoscedasticity
4. Normality
5. No strong Multicollinearity

	Actual Values	Fitted Values	Residuals
0	91.0	91.113529	-0.113529
1	94.0	91.759078	2.240922
2	61.5	74.472389	-12.972389
3	83.0	80.847453	2.152547
4	94.0	98.258662	-4.258662

1. Linearity and Independence of predictors-

- Independent and dependent variables must be linearly related
- Residuals are independent

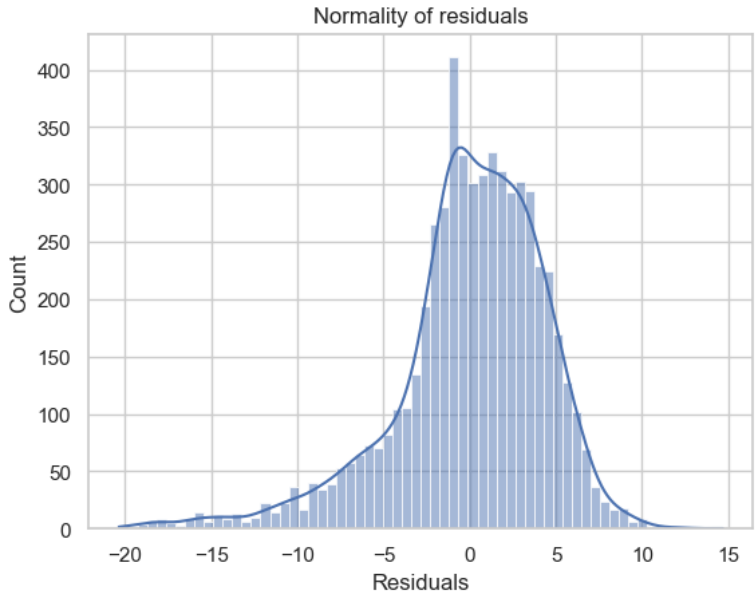


We have plotted the fitted values vs residuals.

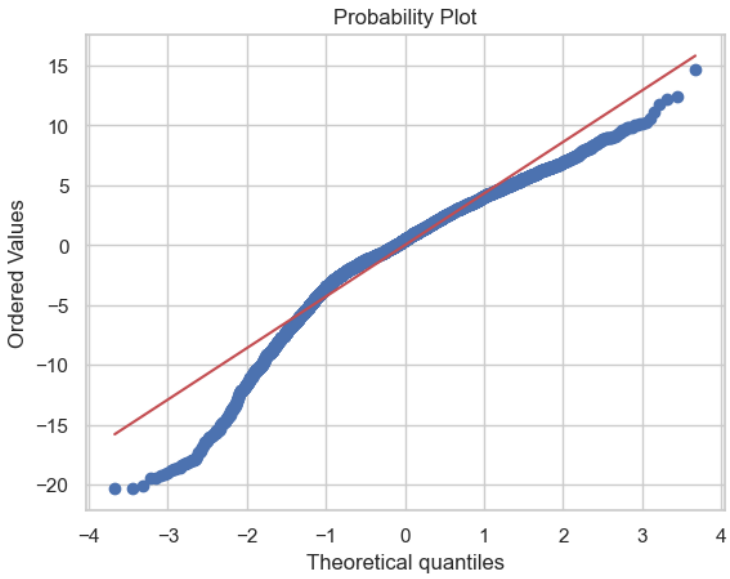
There is no pattern in the data thus the assumption of linearity and independence of predictors are satisfied.

2. Test for Normality-

Error/ Residuals should be normally distributed.



QQ PLOT



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

The non-normality indicates that there are a few unusual data points which must be studied closely to make a better model.

Shapiro-Wilk test-

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

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

ShapiroResult(statistic=0.9426858425140381, pvalue=1.6409205017243608e-42)

Since p-value < 0.05, the residuals are not normal as per Shapiro test.

To make residuals normal, we can apply transformations like log, exponential, etc. as per our data.

3. Test for Homoscedasticity-

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

- If the variance is unequal for the residuals across the regression line, then the data is said to be heteroscedastic.

We are using the **goldfeldquandt** test.
If we get p-value > 0.05 we can say that the residuals are homoscedastic, otherwise they are heteroscedastic.

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

Null hypothesis: Residuals are homoscedastic
Alternate hypothesis: Residuals have heteroscedastic

Value- 0.0013810641992204242

Since p-value < 0.05 we can say that the residuals are heteroscedastic.

How to deal with Heteroscedasticity- It can be fixed by adding other important features or making transformations.

4. No strong Multicollinearity

OLS Regression Results

Dep. Variable:	usr	**R-squared:**	0.795
Model:	OLS	**Adj. R-squared:**	0.794
Method:	Least Squares	**F-statistic:**	1584.
Date:	Sat, 14 Oct 2023	**Prob (F-statistic):**	0.00
Time:	19:59:29	**Log-Likelihood:**	-16673.
No. Observations:	5734	**AIC:**	3.338e+04
Df Residuals:	5719	**BIC:**	3.348e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.0919	0.312	269.420	0.000	83.480	84.704
lread	-0.0684	0.009	-7.653	0.000	-0.086	-0.051
lwrite	0.0523	0.013	3.995	0.000	0.027	0.078
scall	-0.0007	5.96e-05	-11.111	0.000	-0.001	-0.001
swrite	-0.0058	0.001	-5.481	0.000	-0.008	-0.004
exec	-0.3568	0.049	-7.355	0.000	-0.452	-0.262
rchar	-5.511e-06	4.33e-07	-12.740	0.000	-6.36e-06	-4.66e-06
wchar	-4.872e-06	1.02e-06	-4.779	0.000	-6.87e-06	-2.87e-06
pgout	-0.3540	0.038	-9.287	0.000	-0.429	-0.279
atch	0.6055	0.143	4.247	0.000	0.326	0.885
pgin	-0.0820	0.009	-8.730	0.000	-0.100	-0.064
pflt	-0.0396	0.001	-37.292	0.000	-0.042	-0.038
freemem	-0.0005	5.06e-05	-9.328	0.000	-0.001	-0.000
freeswap	8.915e-06	1.87e-07	47.769	0.000	8.55e-06	9.28e-06
runqsz_Not_CPU_Bound	1.5953	0.126	12.641	0.000	1.348	1.843

Omnibus:	1045.912	**Durbin-Watson:**	2.014
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2203.816
Skew:	-1.073	**Prob(JB):**	0.00
Kurtosis:	5.150	**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.

Predictions-

Linear regression equation-

usr = 84.09190904813205 + -0.06835253681476533 * (lread) + 0.052254429104223274 * (lwrite) + -0.0006624251874706158 * (scall) + -0.005784330295337882 * (swrite) + -0.35675411777005817 * (exec) + -5.511271370539362e-06 * (rchar) + -4.872076935688007e-06 * (wchar) + -0.3540346808159861 * (pgout) + 0.6054715199799747 * (atch) + -0.0819586493033359 * (pgin) + -0.03963012164835577 * (pflt) + -0.0004723752262174181 * (freemem) + 8.91524623463361e-06 * (freeswap) + 1.5953196246012888 * (runqsz_Not_CPU_Bound)

We can now use the model for making predictions on the test data.

- **RMSE** on the train data- 4.431792450848492
  RMSE on the test data- 4.671697953091259

- **MAE** on the train data- 3.2945284938678556
  MAE on the test data- 3.38956487950438

Both RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) are slightly higher on the test data compared to the training data.
MAE indicates that our current model is able to predict usr within a mean error of 3.3 units on the test data.

Linear Regression using sklearn-

- Coefficients for each of the independent attributes-

```
The coefficient for const is 0.0
The coefficient for lread is -0.0683525368148109
The coefficient for lwrite is 0.052254429104294
The coefficient for scall is -0.0006624251874710465
The coefficient for swrite is -0.005784330295333583
The coefficient for exec is -0.3567541177697449
The coefficient for rchar is -5.511271370559487e-06
The coefficient for wchar is -4.8720769356044836e-06
The coefficient for pgout is -0.354034680816096
The coefficient for atch is 0.605471519979
The coefficient for pgin is -0.08195864930327641
The coefficient for pflt is -0.03963012164835959
The coefficient for freemem is -0.0004723752262185039
The coefficient for freeswap is 8.915246234657195e-06
The coefficient for runqsz_Not_CPU_Bound is 1.5953196246013068
```

- **Intercept** for our model is 84.09190904806592

  This represents the estimated value of the response variable (usr) when all predictor variables are zero.

- **R squared** measures the strength of relationship between response and predictor variables in the model.

  R square on training data- 0.7949761695805909

  approximately 79% of the variation in the response variable (usr) is explained by the predictor variables in the model for the training set.

  R square on testing data- 0.7657904864538216
  Around 76.6% of the variation in usr is explained by the predictors for the testing set.

- **RMSE** is a measure of the root of the average squared differences between actual and predicted values on the training data.

  RMSE on Training data- 4.431792450848491
  RMSE on Testing data- 4.671697953091488

- The model has a relatively high R-squared value on both training and testing data. This indicates that the predictors in the model explain a significant proportion of the variation in the response variable.
- The model's performance on the testing data, as indicated by the R-squared is slightly lower than on the training data.

1.4 Inference: Basis on these predictions, what are the business insights and recommendations.

- Higher "lread" values are negatively related to the portion of time the CPU runs in user mode.
- Increase in "lwrite" positively affect the portion of time the CPU runs in user mode.
- Increase in the number of system calls ("scall") is associated with a decrease in "usr." More system calls lead to a lower "usr" value.
- Increase in the number of system write calls ("swrite") is associated with a decrease in "usr." More write calls negatively impact "usr."
- Increase in the number of system exec calls ("exec") is associated with a significant decrease in "usr." More exec calls have a stronger negative impact on "usr."
- Increase in the number of characters transferred by read and write calls have a very slight negative effect on "usr." These variables have a minimal impact.

- Increase in the number of page out requests ("pgout") is associated with a decrease in "usr." More page out requests negatively affect "usr."
- Increase in the number of page attaches ("atch") is associated with a significant increase in "usr." More attaches have a strong positive impact on "usr."
- Increase in the number of page-in requests ("pgin") is associated with a decrease in "usr." More page-in requests negatively affect "usr."
- Increase in page faults caused by protection errors ("pflt") is associated with a decrease in "usr." More protection errors have a negative impact on "usr."
- Increase in the number of available memory pages ("freemem") is associated with a decrease in "usr." More available memory negatively affects "usr."
- Increase in the number of available disk blocks for page swapping ("freeswap") is associated with a slight increase in "usr." More available disk blocks have a minor positive impact on "usr."
- CPU run in user mode is highly influenced by Process run queue size (runqsz). Larger process run queue size ("runqsz_Not_CPU_Bound") is associated with a significant increase in "usr." A larger run queue size has a strong positive impact on "usr."

Problem 2- Logistic Regression, LDA and CART

You are a statistician at the Republic of Indonesia Ministry of Health and you are provided with a data of 1473 females collected from a Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of the survey.

The problem is to predict do/don't they use a contraceptive method of choice based on their demographic and socio-economic characteristics.

Dataset for Problem 2: Contraceptive_method_dataset.xlsx

Data Dictionary:

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=very low, 2, 3, 4=high
9. Media exposure (binary) Good, Not good
10. Contraceptive method used (class attribute) No, Yes

2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, check for duplicates and outliers and write an inference on it. Perform Univariate and Bivariate Analysis and Multivariate Analysis.

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

Shape-

(1473, 10)

The dataset has 1473 rows and 10 variables.

```
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
```

There are few missing values in Wife_age and No_of_children_born which will be treater later.

3 variables are numeric and remaining 7 are categorical.

Husband_Occupation is a categorical variable in encoded format hence, we are converting it to object data type.

```
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   object
 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), object(8)
memory usage: 115.2+ KB
```

Now, we have 2 numeric variables and remaining as categorical.

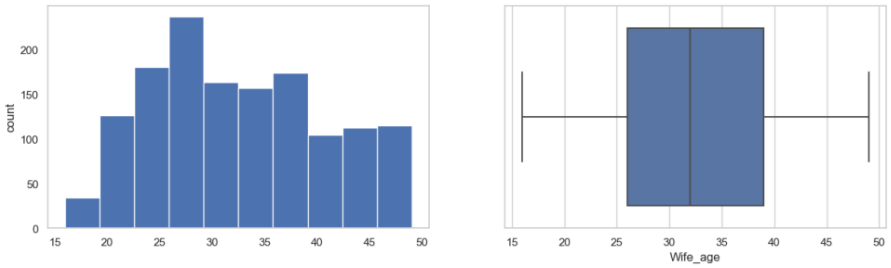
Data Summary-

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Wife_age	1402.0	NaN	NaN	NaN	32.606277	8.274927	16.0	26.0	32.0	39.0	49.0
Wife_ education	1473	4	Tertiary	577	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_education	1473	4	Tertiary	899	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_of_children_born	1452.0	NaN	NaN	NaN	3.254132	2.365212	0.0	1.0	3.0	4.0	16.0
Wife_religion	1473	2	Scientology	1253	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Wife_Working	1473	2	No	1104	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_Occupation	1473.0	4.0	3.0	585.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Standard_of_living_index	1473	4	Very High	684	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Media_exposure	1473	2	Exposed	1364	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Contraceptive_method_used	1473	2	Yes	844	NaN	NaN	NaN	NaN	NaN	NaN	NaN

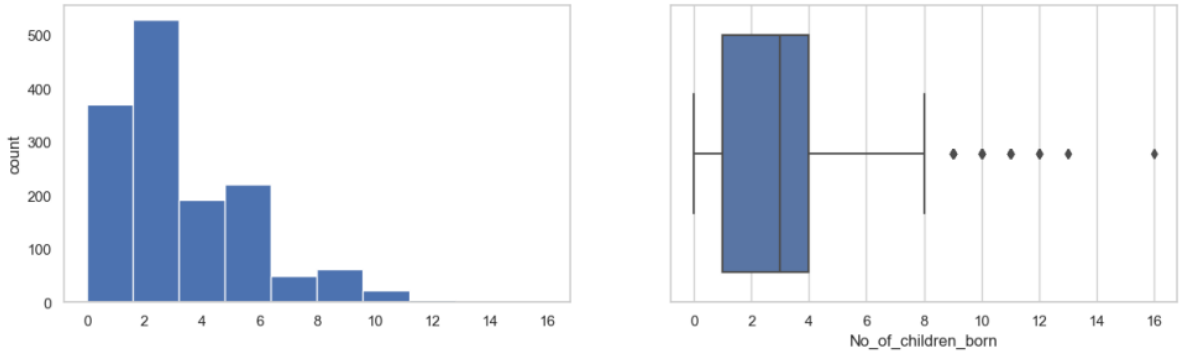
	Wife_age	No_of_children_born
count	1402.000000	1452.000000
mean	32.606277	3.254132
std	8.274927	2.365212
min	16.000000	0.000000
25%	26.000000	1.000000
50%	32.000000	3.000000
75%	39.000000	4.000000
max	49.000000	16.000000

Univariate Analysis-

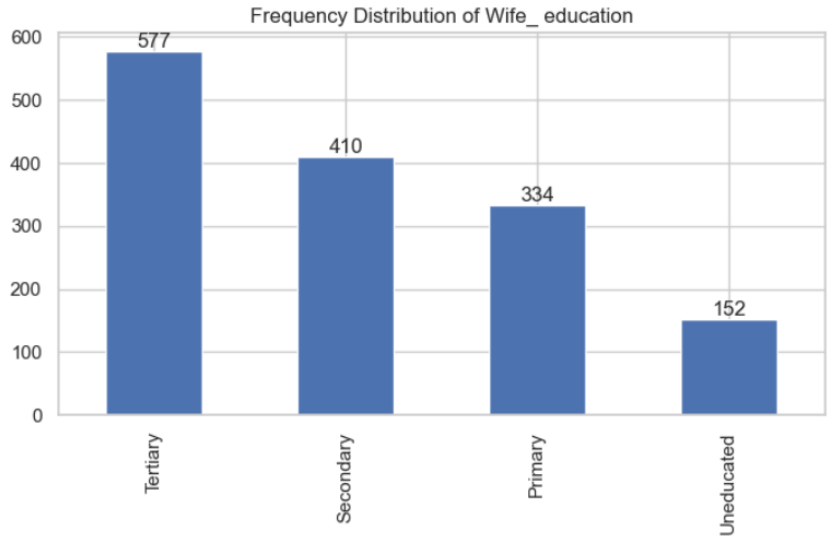
```
Description of Wife_age
--------------------------------------------------------------------------------
count    1402.000000
mean       32.606277
std         8.274927
min        16.000000
25%        26.000000
50%        32.000000
75%        39.000000
max        49.000000
Name: Wife_age, dtype: float64

Skew : 0.25

Distribution of Wife_age
--------------------------------------------------------------------------------
```



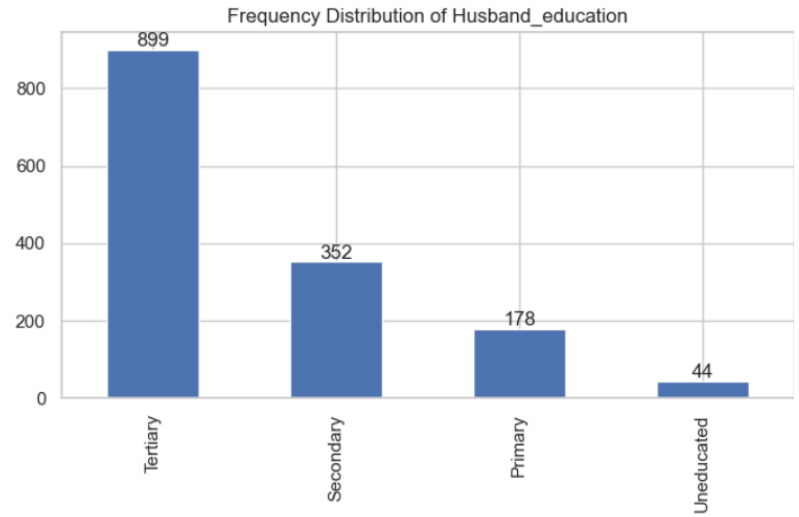

```
Description of No_of_children_born
--------------------------------------------------------------------------------
count    1452.000000
mean        3.254132
std         2.365212
min         0.000000
25%         1.000000
50%         3.000000
75%         4.000000
max        16.000000
Name: No_of_children_born, dtype: float64

Skew : 1.11

Distribution of No_of_children_born
--------------------------------------------------------------------------------
```



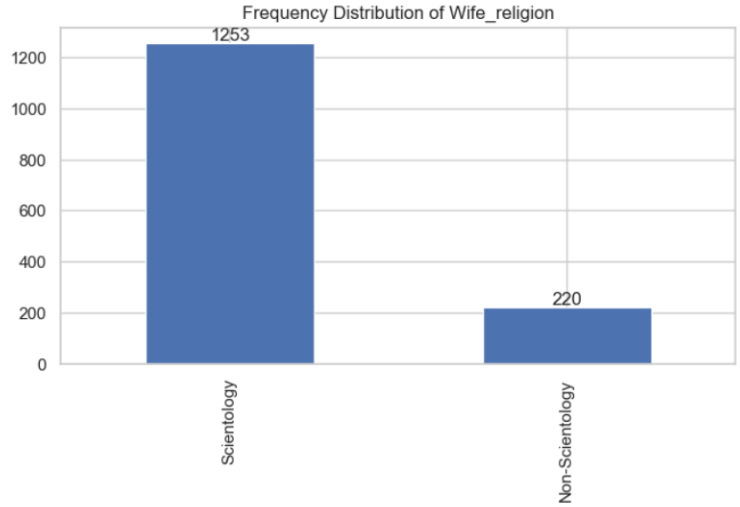
```
Details of Wife_ education
----------------------------------------------------------------
Tertiary      577
Secondary     410
Primary       334
Uneducated    152
Name: Wife_ education, dtype: int64
```



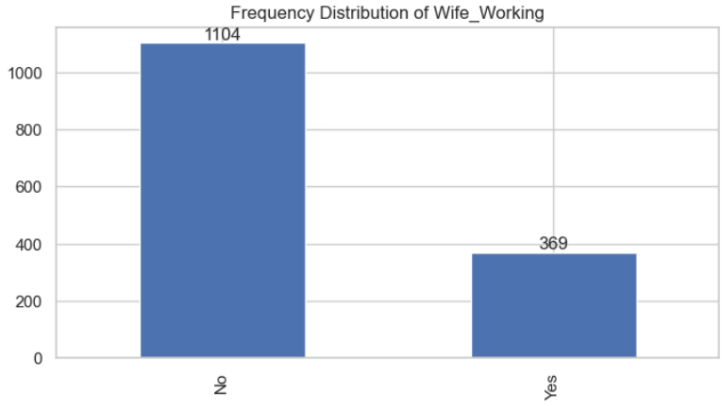

```
Details of Husband_education
----------------------------------------------------------------
Tertiary      899
Secondary     352
Primary       178
Uneducated     44
Name: Husband_education, dtype: int64
```



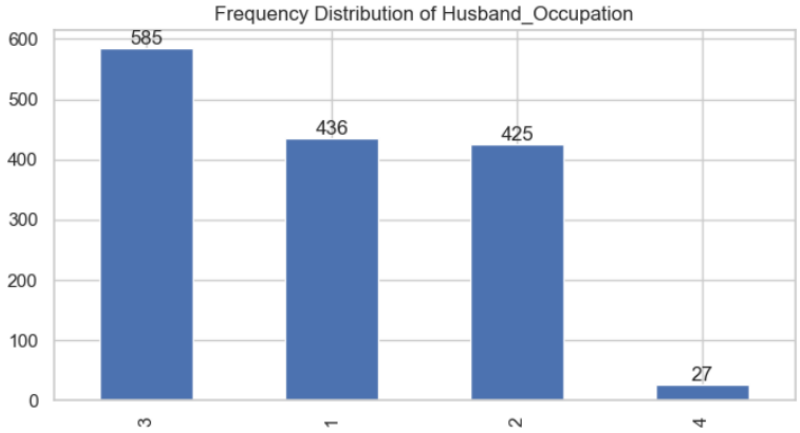
```
Details of Wife_religion
----------------------------------------------------------------
Scientology        1253
Non-Scientology     220
Name: Wife_religion, dtype: int64
```



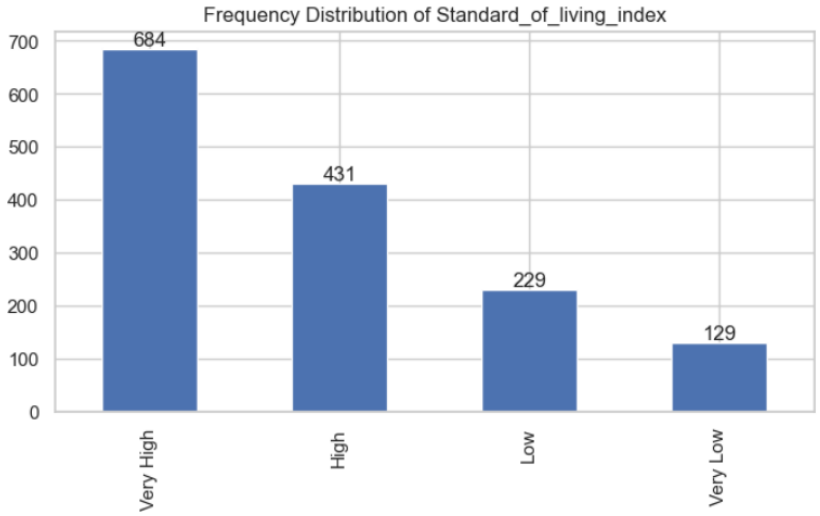

```
Details of Wife_Working
----------------------------------------------------------------
No     1104
Yes     369
Name: Wife_Working, dtype: int64
```



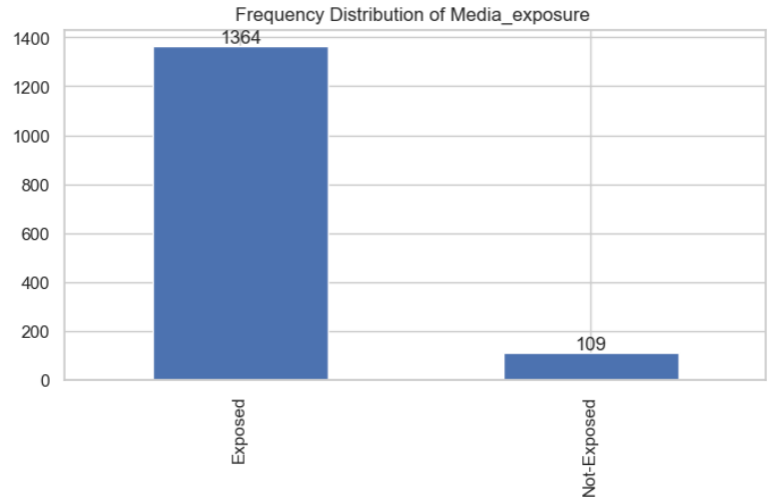
```
Details of Husband_Occupation
-----------------------------------------------------------------
3    585
1    436
2    425
4     27
Name: Husband_Occupation, dtype: int64
```



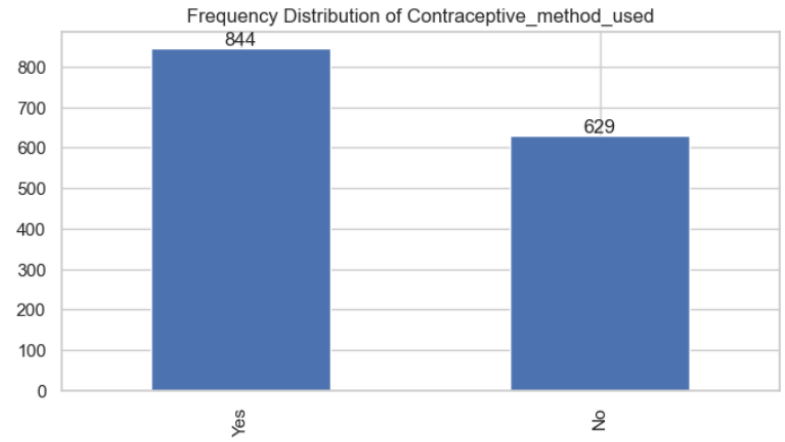

```
Details of Standard_of_living_index
-----------------------------------------------------------------
Very High    684
High         431
Low          229
Very Low     129
Name: Standard_of_living_index, dtype: int64
```



```
Details of Media_exposure
-----------------------------------------------------------------
Exposed        1364
Not-Exposed     109
Name: Media_exposure , dtype: int64
```



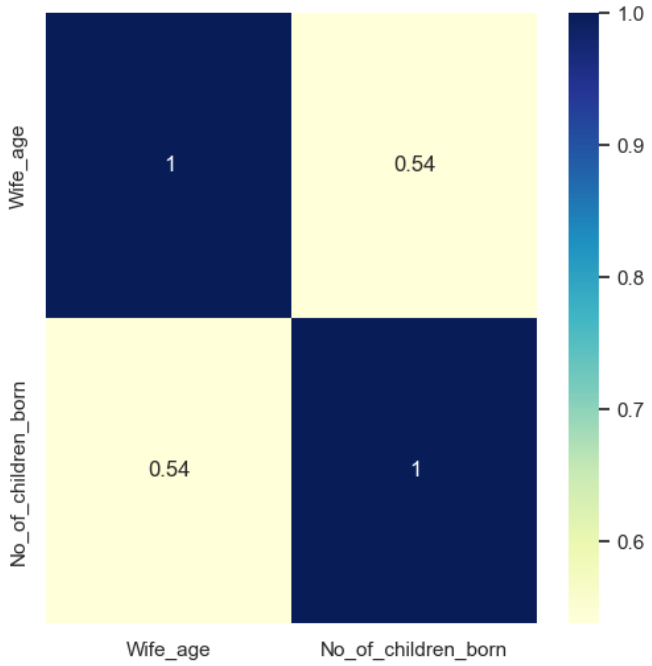

```
Details of Contraceptive_method_used
-----------------------------------------------------------------
Yes    844
No     629
Name: Contraceptive_method_used, dtype: int64
```



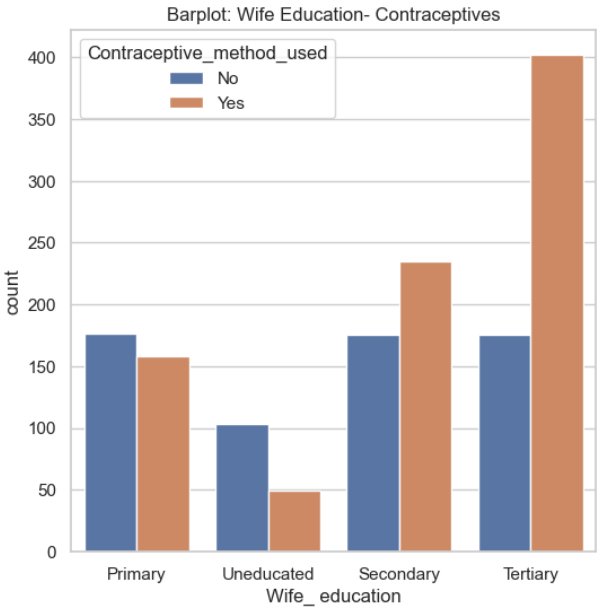
- Majority of the age of the wives range from late 20's- early 40's.
- Majority of people have 1- 4 children, but a few also have more than 15.

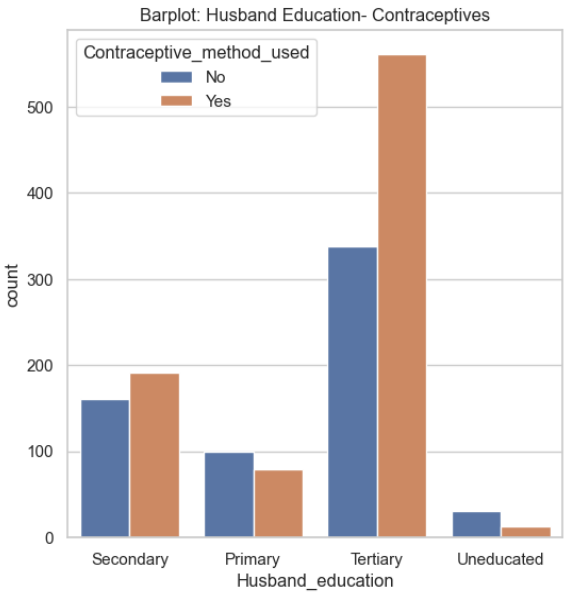
- In terms of education, for both men and women, Tertiary is the highest education level which is obtained by most of them.
- While comparatively, uneducated men/ women are of least numbers.
- Also, there are more number of educated men than women.
- More number of women belong to Scientology religion.
- Most of the women are unemployed.
- There is more frequency at level 3 Occupation of the Husbands followed by level 1 and level 4 being the least number.
- More number of people have a very high standard of living.
- High number of people are exposed to media.
- More number of people go in for the option of using contraceptive methods.

Bivariate Analysis

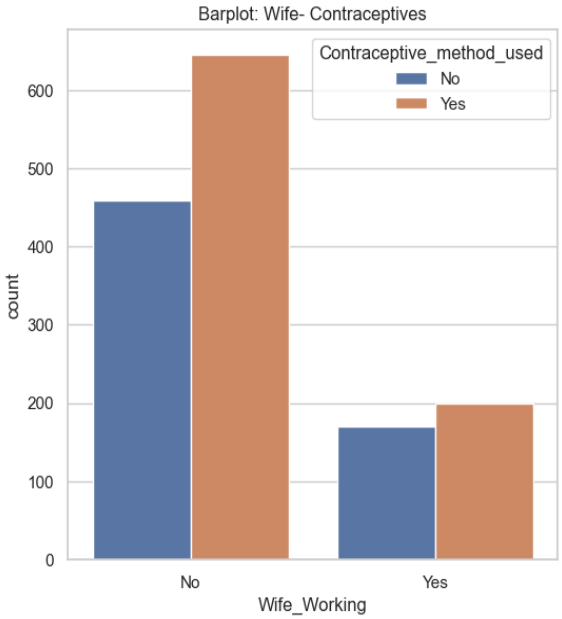


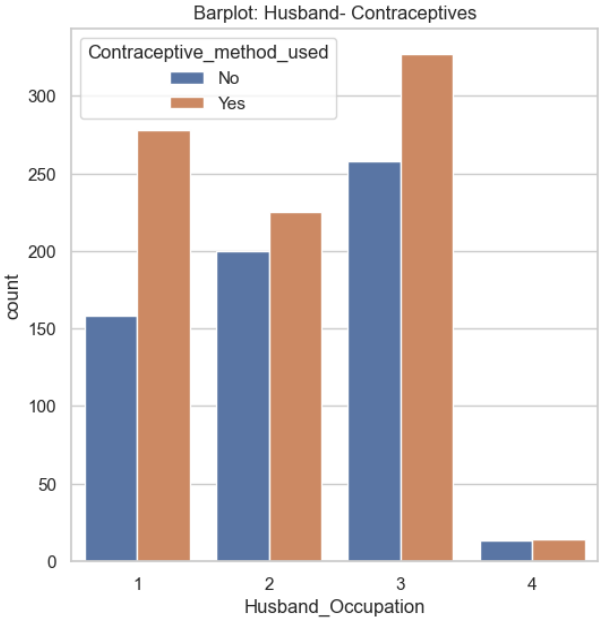
- Based on the correlation coefficient of 0.54, we can conclude that there is a moderate positive linear relationship between wife age and the number of children born.
- Older wives tend to have more children, but it's not a perfect relationship.

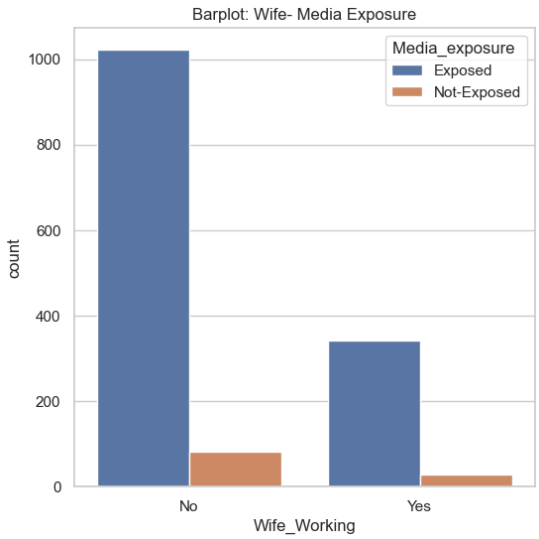




The more the people are educated, the more they opt for contraceptive methods.

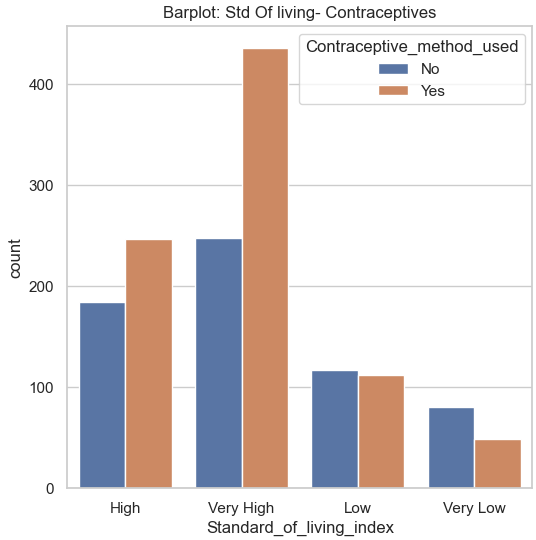




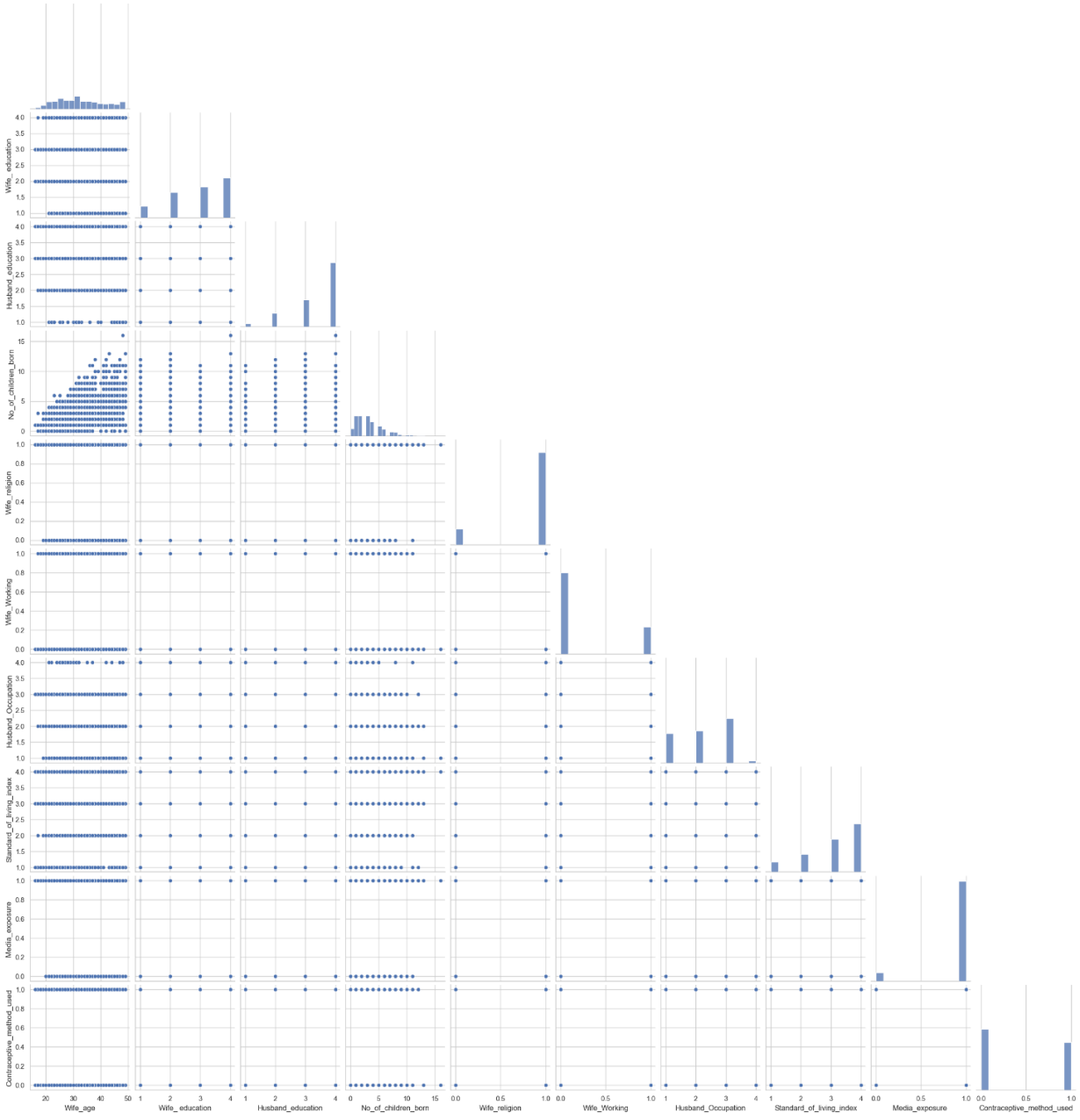


Non- working women have more exposure towards media which may also be the reason for their high proportion of using contraceptives.

Also, highest with husband occupation at level 3, followed by level 1 and level 4 at the least.



Also, people falling under very high and high standard of living have higher volume of people using contraceptive methods than those under low / very low standards.



We have **80 duplicate rows** and have dropped them.

Printing all the unique values-

```
WIFE_ EDUCATION :  4
Uneducated    150
Primary       330
Secondary     398
Tertiary      515
Name: Wife_ education, dtype: int64


HUSBAND_EDUCATION :  4
Uneducated     44
Primary       175
Secondary     347
Tertiary      827
Name: Husband_education, dtype: int64


WIFE_RELIGION :  2
Non-Scientology     207
Scientology        1186
Name: Wife_religion, dtype: int64


WIFE_WORKING :  2
Yes     350
No     1043
Name: Wife_Working, dtype: int64


HUSBAND_OCCUPATION :  4
4     27
1    381
2    415
3    570
Name: Husband_Occupation, dtype: int64


STANDARD_OF_LIVING_INDEX :  4
Very Low     129
Low          227
High         419
Very High    618
Name: Standard_of_living_index, dtype: int64


MEDIA_EXPOSURE  :  2
Not-Exposed     109
Exposed        1284
Name: Media_exposure , dtype: int64


CONTRACEPTIVE_METHOD_USED :  2
No     614
Yes    779
Name: Contraceptive_method_used, dtype: int64
```

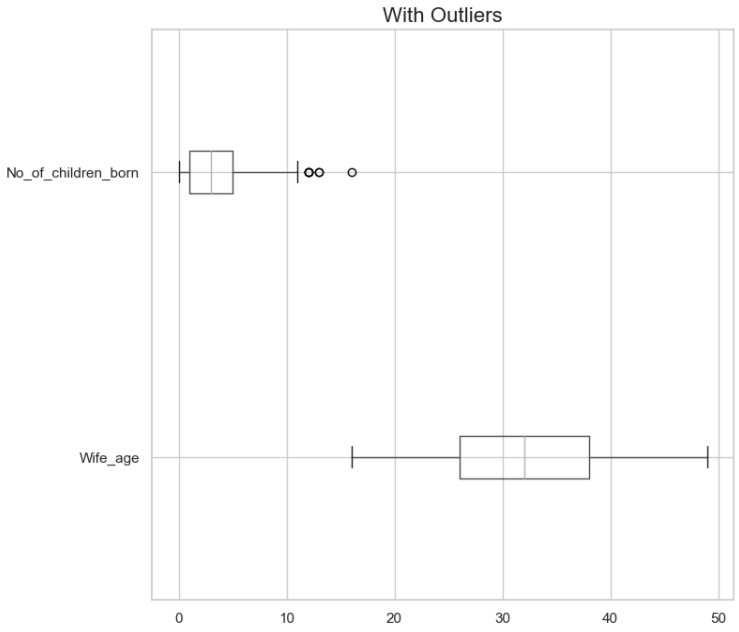
Printing null values-

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

Post Null value treatment-

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

After replacing all null values with median, null values have tended to o.



Number of children born has outliers which is valid. Hence, we are not treating it.

2.2 Do not scale the data. Encode the data (having string values) for Modelling.

Since LDA works with Numeric independent variables, we have encoded all the object feature types as below-

'Wife_ education' – Uneducated- 1, Primary- 2, Secondary – 3, Tertiary – 4

'Husband_education'- Uneducated- 1, Primary- 2, Secondary – 3, Tertiary – 4

'Wife_religion'- Scientology- 1, Non-Scientology- 0

'Wife_Working'- Yes- 1, No -0

'Standard_of_living_index'- Very Low – 1, Low -2, High- 3, Very High - 4

'Media_exposure' – Exposed- 1, Not-Exposed- 0

'Contraceptive_method_used'- Yes- 0, No- 1

	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_method_used	
0	24.0	2	3	3.0	1	0	2	3	1	1	
1	45.0	1	3	10.0	1	0	3	4	1	1	
2	43.0	2	3	7.0	1	0	3	4	1	1	
3	42.0	3	2	9.0	1	0	3	3	1	1	
4	36.0	3	3	8.0	1	0	3	2	1	1	

Converting them all to numeric type-

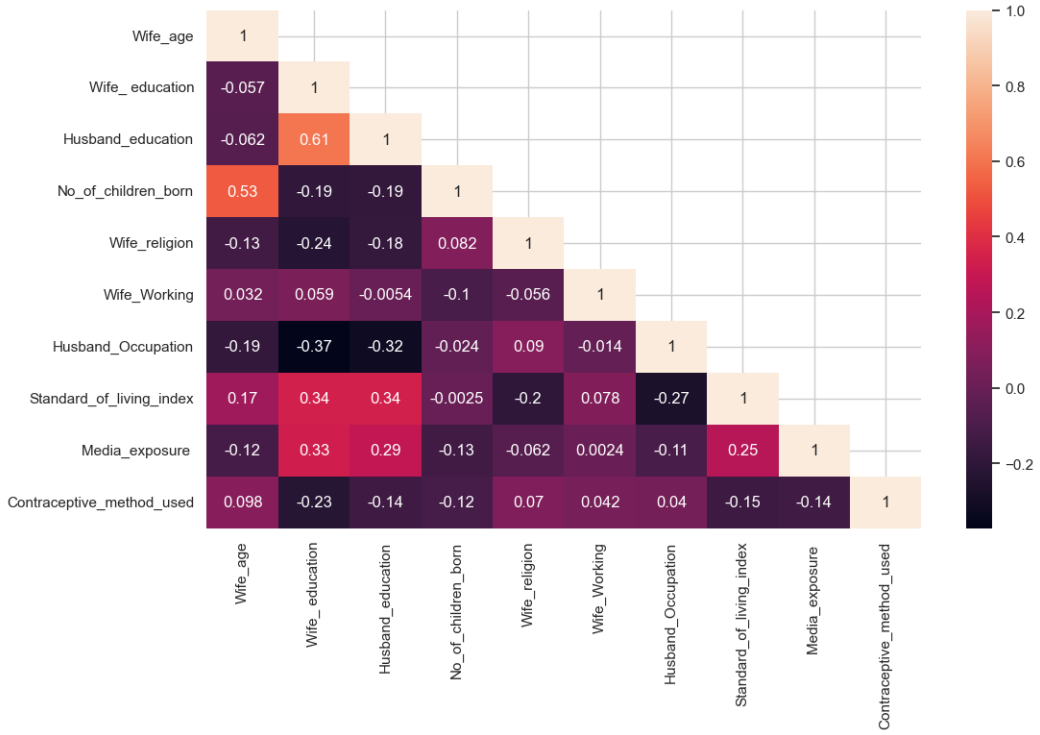
```
Int64Index: 1393 entries, 0 to 1472
Data columns (total 10 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   Wife_age                   1393 non-null   float64
 1   Wife_ education            1393 non-null   int64  
 2   Husband_education          1393 non-null   int64  
 3   No_of_children_born        1393 non-null   float64
 4   Wife_religion              1393 non-null   int64  
 5   Wife_Working               1393 non-null   int64  
 6   Husband_Occupation         1393 non-null   int64  
 7   Standard_of_living_index   1393 non-null   int64  
 8   Media_exposure             1393 non-null   int64  
 9   Contraceptive_method_used  1393 non-null   int64  
dtypes: float64(2), int64(8)
memory usage: 119.7 KB
```

Printing value counts of Contraceptive methods

```
0    779
1    614
Name: Contraceptive_method_used, dtype: int64
```

Checking correlation-

	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_method_used
Wife_age	1.000000	-0.057025	-0.062177	0.528428	-0.134091	0.032000	-0.187070	0.171420	-0.119737	0.098228
Wife_ education	-0.057025	1.000000	0.606868	-0.193076	-0.236765	0.058523	-0.370799	0.341412	0.334415	-0.228341
Husband_education	-0.062177	0.606868	1.000000	-0.186437	-0.181347	-0.005429	-0.317932	0.342141	0.285320	-0.144646
No_of_children_born	0.528428	-0.193076	-0.186437	1.000000	0.081623	-0.103751	-0.024213	-0.002481	-0.132228	-0.118343
Wife_religion	-0.134091	-0.236765	-0.181347	0.081623	1.000000	-0.055791	0.090034	-0.201524	-0.061603	0.070084
Wife_Working	0.032000	0.058523	-0.005429	-0.103751	-0.055791	1.000000	-0.013669	0.078367	0.002385	0.042433
Husband_Occupation	-0.187070	-0.370799	-0.317932	-0.024213	0.090034	-0.013669	1.000000	-0.270934	-0.106340	0.040438
Standard_of_living_index	0.171420	0.341412	0.342141	-0.002481	-0.201524	0.078367	-0.270934	1.000000	0.245634	-0.146416
Media_exposure	-0.119737	0.334415	0.285320	-0.132228	-0.061603	0.002385	-0.106340	0.245634	1.000000	-0.139744
Contraceptive_method_used	0.098228	-0.228341	-0.144646	-0.118343	0.070084	0.042433	0.040438	-0.146416	-0.139744	1.000000



- There is a moderate positive correlation between a wife's age and the number of children born. As the wife's age increases, the number of children born tends to increase.
- There is moderately positive correlation between the education level of wives and husbands. This suggests that couples tend to have similar levels of education.
- There is a moderate positive correlation between a wife/ husbands education and the standard of living index. This indicates that higher education levels are associated with a higher standard of living.
- There is a moderate positive correlation between wife's education and media exposure. More educated wives tend to have higher media exposure.
- There is a moderate negative correlation between husband's education and the contraceptive method used. This suggests that the choice of contraceptive method may vary based on the husband's education.
- There is a weak negative correlation between the standard of living index and the contraceptive method used. This indicates that the standard of living may have some influence on the choice of contraceptive method.

We have split the data into train and test set (70:30 ratio).

To ensure balance of 0's and 1's in both the test and train data set, we have used function stratify.

Train values-

```
0    0.558974
1    0.441026
Name: Contraceptive_method_used, dtype: float64
```

Test values-

```
0    0.559809
1    0.440191
Name: Contraceptive_method_used, dtype: float64
```

```
Number of rows and columns of the training set for the independent variables: (975, 9)
Number of rows and columns of the training set for the dependent variable: (975,)
Number of rows and columns of the test set for the independent variables: (418, 9)
Number of rows and columns of the test set for the dependent variable: (418,)
```

Logistic Regression model-

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

Getting the predicted class and probabilities-

Probability prediction on training data-

```
array([1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0,
       1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 1,
       0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0,
       1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0,
       1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0,
       0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1,
       0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0,
       1, 0, 1, 0, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0,
       0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0,
       0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1,
       1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1,
       1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1,
       1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0,
       0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,
       1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0,
       0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0,
       0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0,
       0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
       1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0,
       0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0,
       1, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1,
       0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
       1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0,
       0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0,
       1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1,
       0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1,
       0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0,
       1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0,
       1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0,
       0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0,
       1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 0,
       0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0,
       0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1,
       0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0,
       0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0,
       0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1,
       0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
       0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0,
       1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0,
       1, 0, 0, 1, 1, 1, 0], dtype=int64)
```

Probability prediction on test data-

```
array([0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1,
       0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0,
       1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0,
       1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0,
       0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1,
       0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
       0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0,
       1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1,
       1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1,
       1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0,
       1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0,
       0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0,
       0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0,
       0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0,
       0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0,
       1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1,
       1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0,
       0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0],
      dtype=int64)
```

Printing Predicted class probabilities for training data-

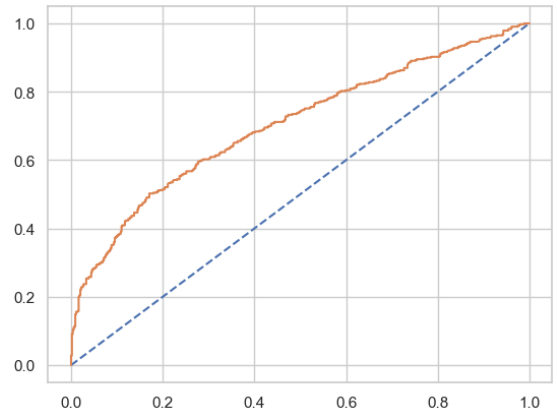
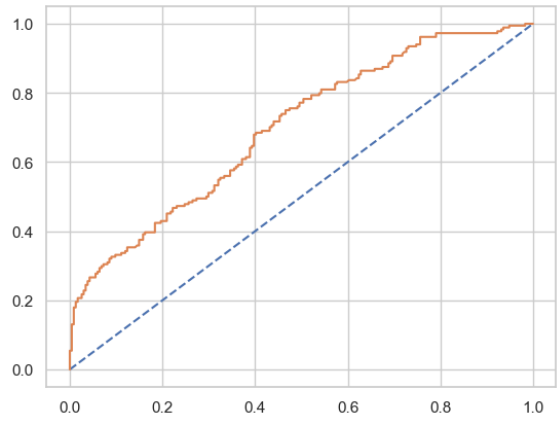
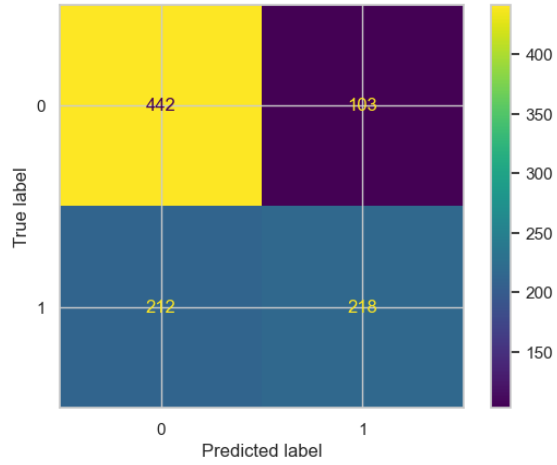
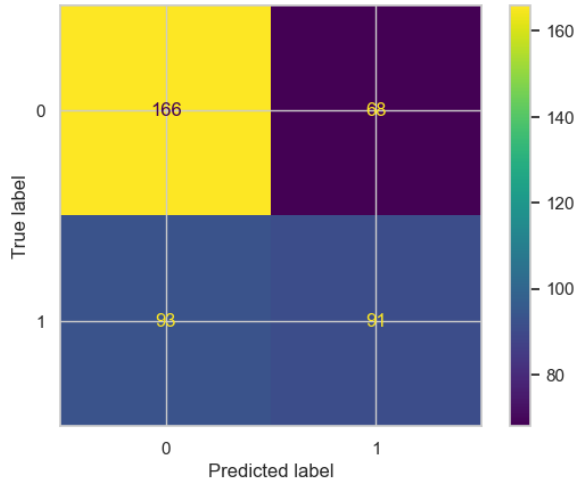
	0	1
0	0.491259	0.508741
1	0.794138	0.205862
2	0.813545	0.186455
3	0.835770	0.164230
4	0.412272	0.587728
...	...	...
970	0.810950	0.189050
971	0.493644	0.506356
972	0.144435	0.855565
973	0.220199	0.779801
974	0.791254	0.208746

975 rows × 2 columns

Printing Predicted class probabilities for test data-

	0	1
0	0.819851	0.180149
1	0.634207	0.365793
2	0.659119	0.340881
3	0.476370	0.523630
4	0.263909	0.736091
...	...	...
413	0.114172	0.885828
414	0.721743	0.278257
415	0.756333	0.243667
416	0.431756	0.568244
417	0.542762	0.457238

418 rows × 2 columns

	Train Data	Test Data
AUC	0.703	0.703
ROC		
Accuracy	0.676923076923077	0.6148325358851675
Confusion Matrix	array([[442, 103], [212, 218]], dtype=int64)	array([[166, 68], [93, 91]], dtype=int64)
		
Classification report	precision recall f1-score support 0 0.68 0.81 0.74 545 1 0.68 0.51 0.58 430 accuracy 0.68 975 macro avg 0.68 0.66 0.66 975 weighted avg 0.68 0.68 0.67 975	precision recall f1-score support 0 0.64 0.71 0.67 234 1 0.57 0.49 0.53 184 accuracy 0.61 418 macro avg 0.61 0.60 0.60 418 weighted avg 0.61 0.61 0.61 418

Applying GridSearchCV for Logistic Regression-

Best params and best estimators-

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

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

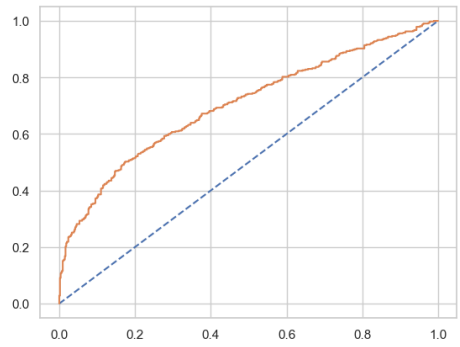
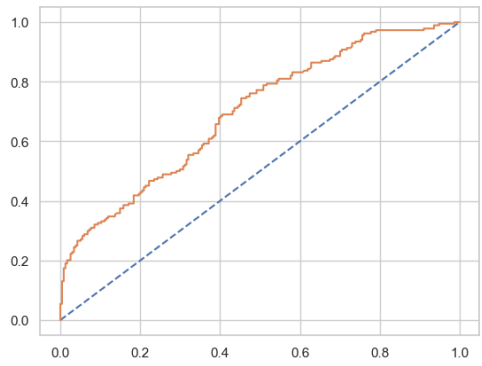
Post building model using these params-

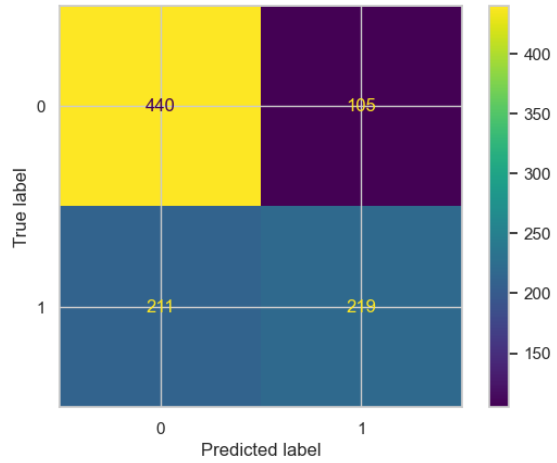
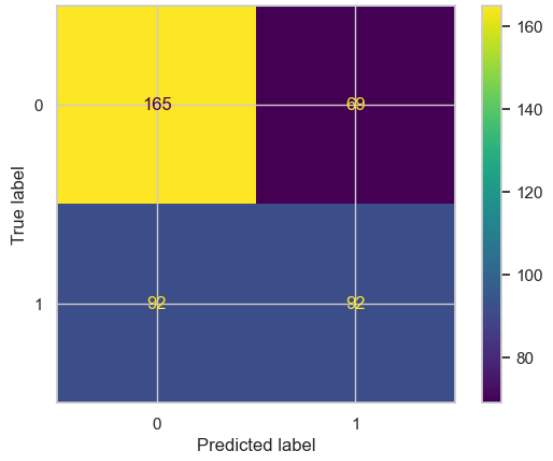
Probabilities on train set-

	0	1
0	0.491261	0.508739
1	0.793561	0.206439
2	0.813929	0.186071
3	0.833828	0.166172
4	0.421639	0.578361

Probabilities on test data set-

	0	1
0	0.817926	0.182074
1	0.634511	0.365489
2	0.657651	0.342349
3	0.475477	0.524523
4	0.272231	0.727769

	Train Data	Test Data
AUC	0.703	0.703
ROC		
Accuracy	0.6758974358974359	0.6148325358851675
Confusion Matrix	`array([[440, 105],` `[211, 219]], dtype=int64)`	`array([[165, 69],` `[ 92, 92]], dtype=int64)`

		
Classification report	precision recall f1-score support 0 0.68 0.81 0.74 545 1 0.68 0.51 0.58 430 accuracy 0.68 975 macro avg 0.68 0.66 0.66 975 weighted avg 0.68 0.68 0.67 975	precision recall f1-score support 0 0.64 0.71 0.67 234 1 0.57 0.50 0.53 184 accuracy 0.61 418 macro avg 0.61 0.60 0.60 418 weighted avg 0.61 0.61 0.61 418

We do not observe much variation on the model post optimization.

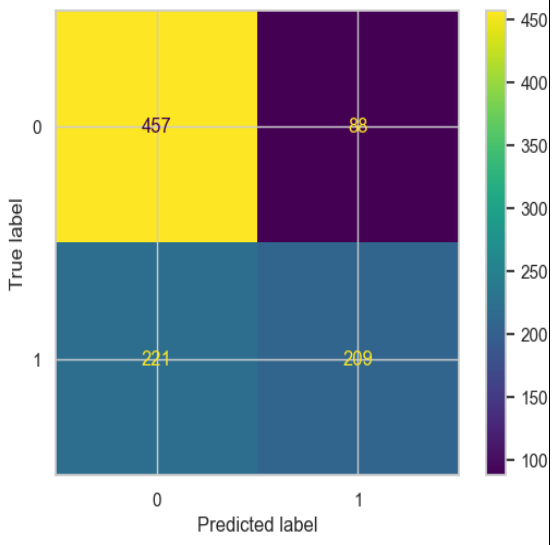
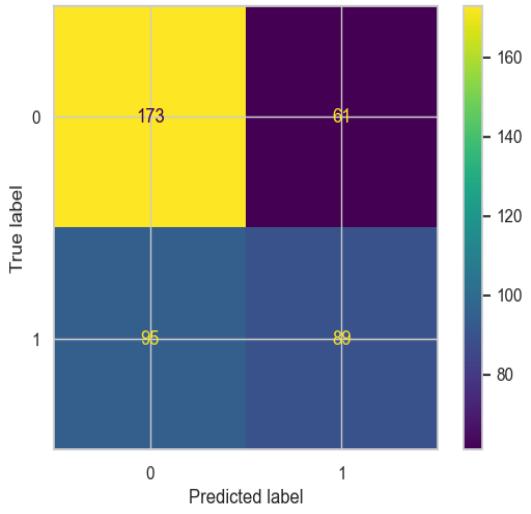
LDA (Linear Discriminant Analysis)-

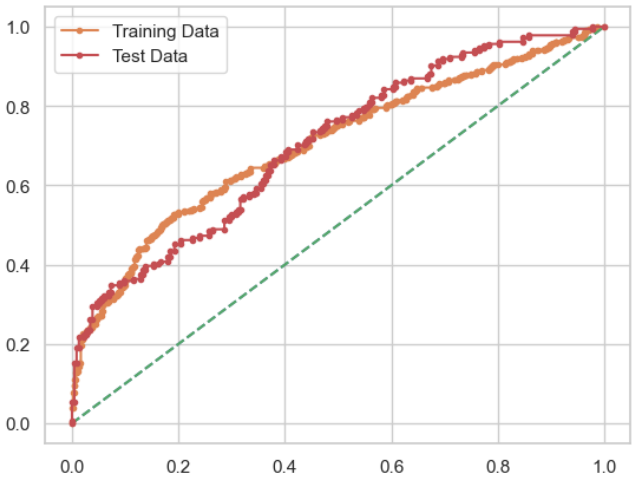
Probability prediction on training data-

```
array([0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0,
       1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 1,
       0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0,
       1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0,
       1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0,
       0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1,
       0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0,
       1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0,
       0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0,
       0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
       1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1,
       1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1,
       1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
       0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,
       1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0,
       0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0,
       0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
       1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0,
       0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0,
       1, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0,
       0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
       1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0,
       0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0,
       1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1,
       0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1,
       0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0,
       1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0,
       1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0,
       1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 0,
       0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
       0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1,
       0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0,
       0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0,
       0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1,
       0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
       0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0,
       1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0,
       1, 0, 0, 1, 1, 1, 0], dtype=int64)
```

Probability prediction on test data-

```
array([0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1,
       0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0,
       1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0,
       1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0,
       0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1,
       0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
       1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0,
       1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1,
       1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1,
       1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1,
       1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0,
       0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0,
       0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0,
       0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0,
       1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1,
       1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0],
      dtype=int64)
```

	Train Dataset	Test Dataset
Confusion Matrix	`array([[457, 88], [221, 209]], dtype=int64)`	`array([[173, 61], [ 95, 89]], dtype=int64)`
		
Classification report	Classification Report of the training data: precision recall f1-score support 0 0.67 0.84 0.75 545 1 0.70 0.49 0.57 430 accuracy 0.68 975 macro avg 0.69 0.66 0.66 975 weighted avg 0.69 0.68 0.67 975	Classification Report of the test data: precision recall f1-score support 0 0.65 0.74 0.69 234 1 0.59 0.48 0.53 184 accuracy 0.63 418 macro avg 0.62 0.61 0.61 418 weighted avg 0.62 0.63 0.62 418
AUC	0.703	0.705



Printing 1st 20 data points in a predicted probability class-

Train dataset-

```
array([[0.50831811, 0.49168189],
       [0.81160197, 0.18839803],
       [0.79329158, 0.20670842],
       [0.84796879, 0.15203121],
       [0.40985458, 0.59014542],
       [0.59835143, 0.40164857],
       [0.74714196, 0.25285804],
       [0.46492291, 0.53507709],
       [0.48757061, 0.51242939],
       [0.87340548, 0.12659452],
       [0.34457659, 0.65542341],
       [0.55260921, 0.44739079],
       [0.52557647, 0.47442353],
       [0.64245896, 0.35754104],
       [0.66027055, 0.33972945],
       [0.41212167, 0.58787833],
       [0.41458502, 0.58541498],
       [0.76658553, 0.23341447],
       [0.33400859, 0.66599141],
       [0.77971172, 0.22028828]])
```

Test dataset-

```
array([[0.81548004, 0.18451996],
       [0.61375202, 0.38624798],
       [0.64486414, 0.35513586],
       [0.49675781, 0.50324219],
       [0.27804999, 0.72195001],
       [0.49945663, 0.50054337],
       [0.54479929, 0.45520071],
       [0.6246769 , 0.3753231 ],
       [0.82097008, 0.17902992],
       [0.24062221, 0.75937779],
       [0.53554255, 0.46445745],
       [0.183702  , 0.816298  ],
       [0.64801241, 0.35198759],
       [0.80477943, 0.19522057],
       [0.73054665, 0.26945335],
       [0.61058935, 0.38941065],
       [0.81355584, 0.18644416],
       [0.79014939, 0.20985061],
       [0.78699925, 0.21300075],
       [0.73020957, 0.26979043]])
```

Intercept value-

array([0.84034378])

Coefficients for the Linear Discriminant Function-

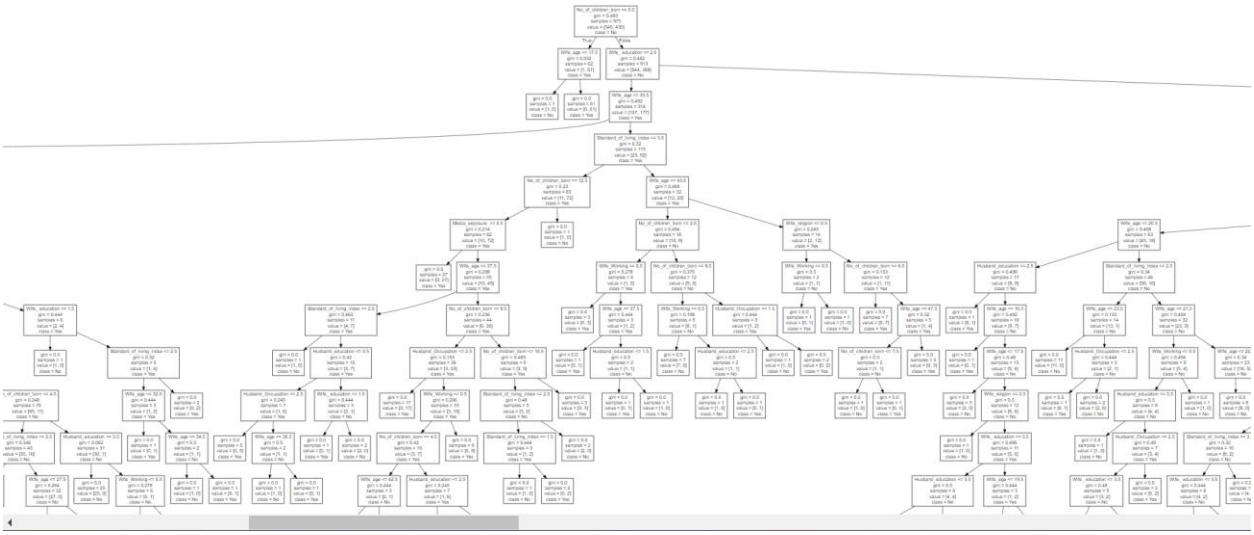
```
array([[ 0.07511072, -0.50898147, -0.01209484, -0.3032077 ,  0.38161852,
         0.13955146, -0.10950659, -0.24636313, -0.43083527]])
```

Rounded up coefficients-

```
array([[ 0.08, -0.51, -0.01, -0.3 ,  0.38,  0.14, -0.11, -0.25, -0.43]])
```

Coefficients represent each independent variables weight in Linear Discriminant Function.

CART (Classification & Regression Tree)-



Variable Importance-

	Imp
Wife_age	0.294870
No_of_children_born	0.253824
Standard_of_living_index	0.096836
Wife_ education	0.089735
Husband_Occupation	0.081690
Husband_education	0.080626
Wife_Working	0.063336
Wife_religion	0.026510
Media_exposure	0.012574

"Wife_age" and "No_of_children_born" are the most important features, while "Media_exposure" has the lowest importance in making predictions whether contraceptive methods are being used or not.

Predicting on Training and Test dataset

Shape-

ytrain_predict (975,)
ytest_predict (418,)

Predicted classes on Test -

```
array([1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1,
       0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1,
       1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0,
       1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0,
       0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0,
       1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1,
       0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0,
       0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1,
       0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1,
       1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1,
       1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0,
       0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0,
       0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0,
       0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0,
       1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1,
       1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0,
       1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1,
       0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1],
      dtype=int64)
```

Predicted classes on Train -

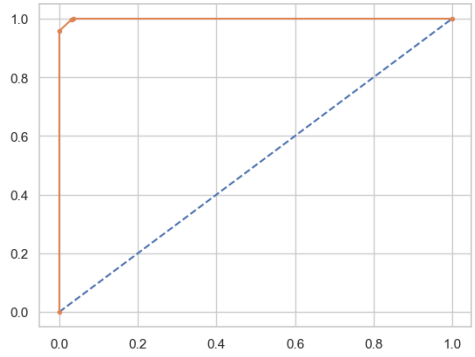
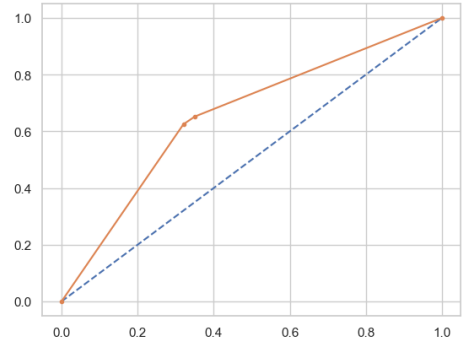
```
array([1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1,
       1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 1,
       1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0,
       1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0,
       0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0,
       1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0,
       0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
       0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0,
       0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 0,
       0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0,
       1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0,
       1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1,
       1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
       0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0,
       0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0,
       1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0,
       1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,
       0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0,
       1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0,
       1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1,
       1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
       0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 1, 0,
       1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0,
       1, 1, 1, 0, 1, 1, 0], dtype=int64)
```

Printing 1st 20 rows of Predicted class probabilities for the test data-

```
array([[0., 1.],
       [0., 1.],
       [1., 0.],
       [1., 0.],
       [0., 1.],
       [1., 0.],
       [1., 0.],
       [1., 0.],
       [1., 0.],
       [0., 1.],
       [1., 0.],
       [1., 0.],
       [1., 0.],
       [1., 0.],
       [1., 0.],
       [1., 0.],
       [1., 0.],
       [1., 0.],
       [1., 0.]])
```

Printing 1st 20 rows of Predicted class probabilities for the train data-

```
array([1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1],
      dtype=int64)
```

	Train Dataset	**Test Dataset**
AUC	0.999	0.648
		
Confusion Matrix	`array([[545, 0], [ 18, 412]], dtype=int64)`	`array([[157, 77], [ 69, 115]], dtype=int64)`
Accuracy	0.9815384615384616	0.6507177033492823
Classification report	precision recall f1-score support 0 0.97 1.00 0.98 545 1 1.00 0.96 0.98 430 accuracy 0.98 975 macro avg 0.98 0.98 0.98 975 weighted avg 0.98 0.98 0.98 975	precision recall f1-score support 0 0.69 0.67 0.68 234 1 0.60 0.62 0.61 184 accuracy 0.65 418 macro avg 0.65 0.65 0.65 418 weighted avg 0.65 0.65 0.65 418

Looks like Decision Tree Classifier is over-fitting since the model's training accuracy is significantly higher than its test accuracy. This means the model performs well on the training data and its ability to make accurate predictions on new, unseen data is not as strong. We will make use of Grid Search to get the best parameters and prune the tree.

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

Comparing the models-

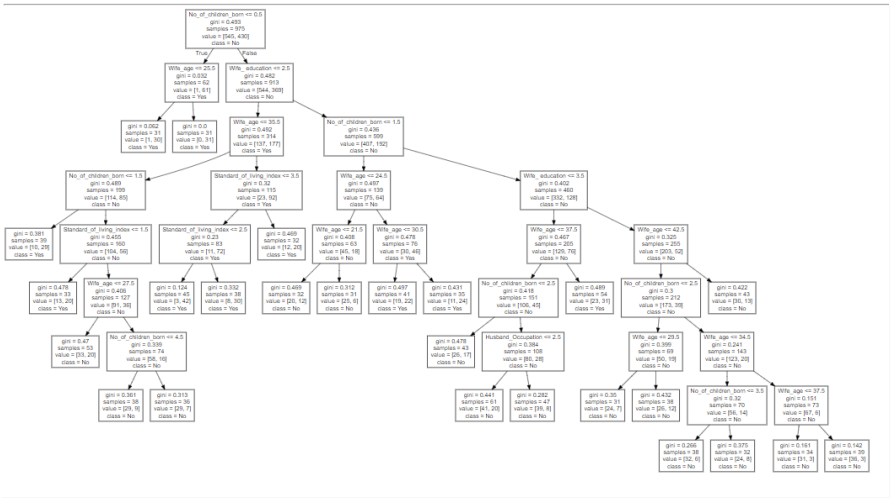
	Train Accuracy	Test Accuracy
Decision Tree Classifier	0.981538	0.643541
LDA	0.682051	0.622010
Logistic Regression	0.676923	0.617225

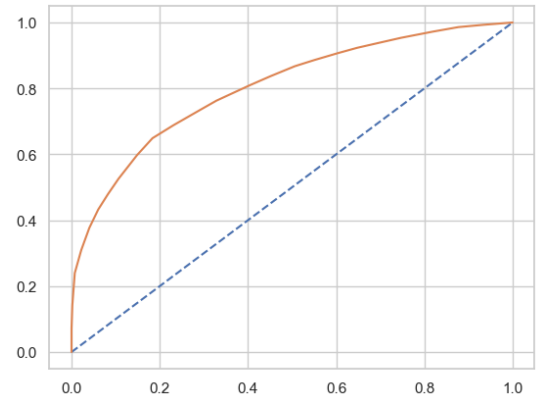
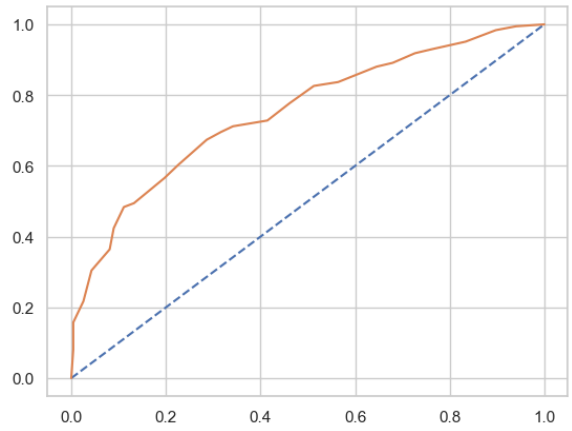
The Decision Tree Classifier has high training accuracy but a lower test accuracy, indicating that it may be overfitting the training data.

Best params-

```
{'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 30, 'min_samples_split': 35}
```

Building model using best params-

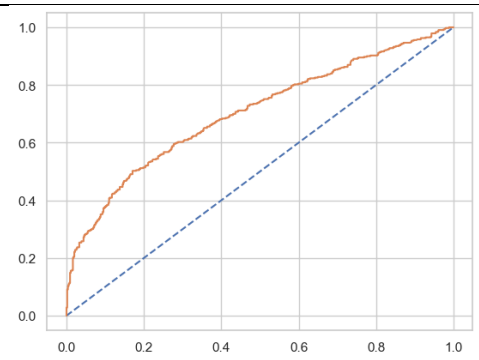
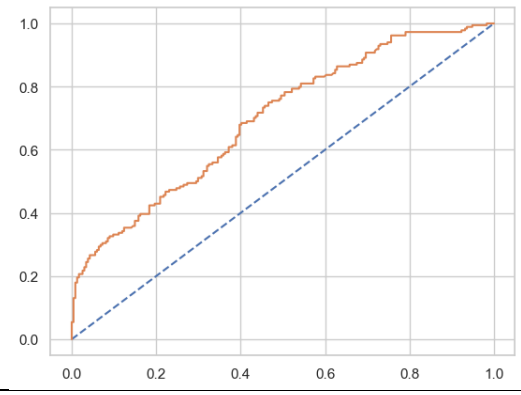
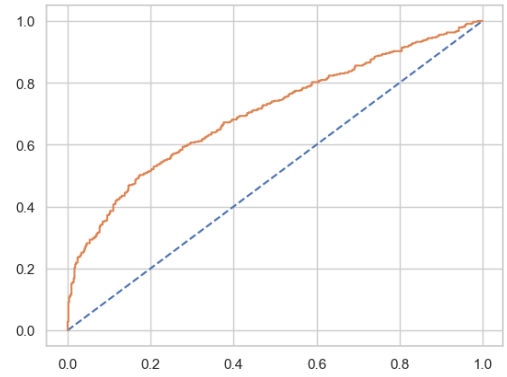
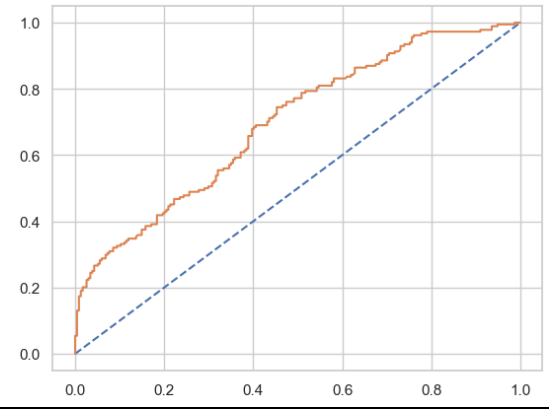


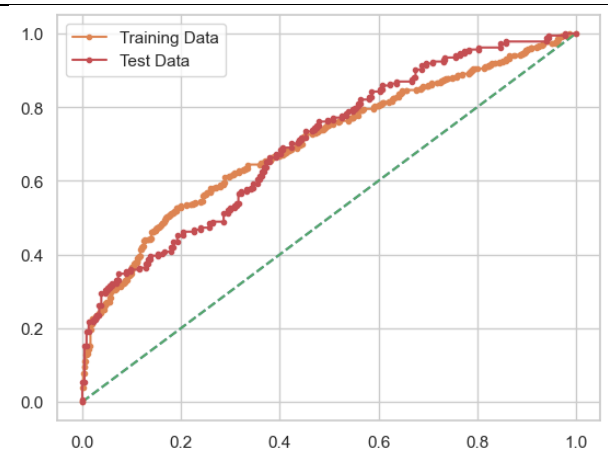
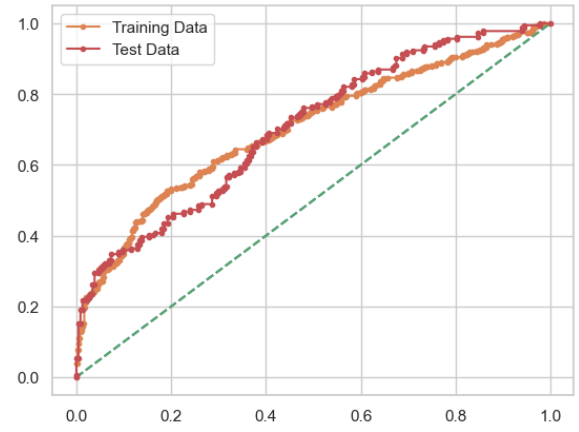
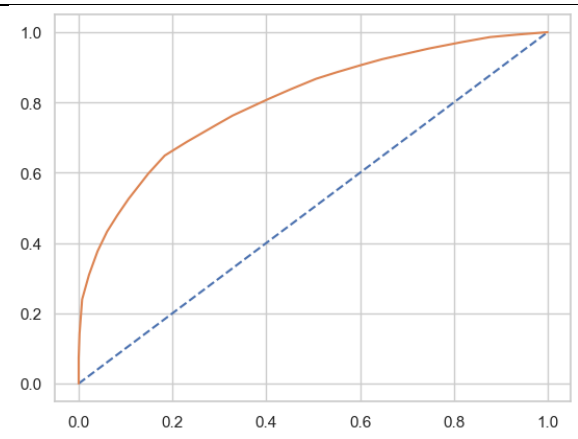
	Train Dataset	Test Dataset
AUC	0.803	0.803
		
Classification report	precision recall f1-score support 0 0.75 0.82 0.78 545 1 0.74 0.65 0.69 430 accuracy 0.74 975 macro avg 0.74 0.73 0.73 975 weighted avg 0.74 0.74 0.74 975	precision recall f1-score support 0 0.71 0.77 0.74 234 1 0.68 0.60 0.64 184 accuracy 0.70 418 macro avg 0.69 0.69 0.69 418 weighted avg 0.70 0.70 0.70 418

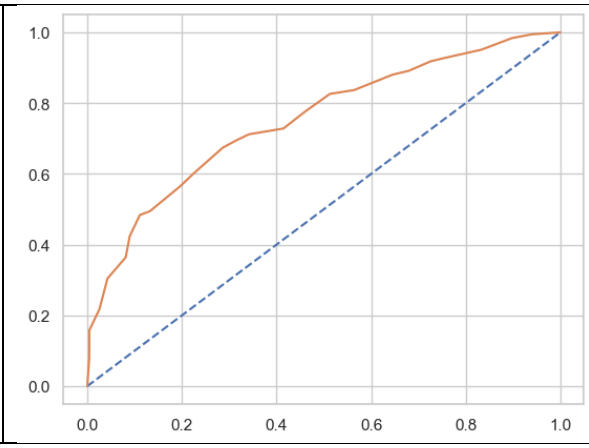
Comparison of models post pruning-

	Train Accuracy	Test Accuracy
Decision Tree Classifier	0.742564	0.698565
LDA	0.682051	0.622010
Logistic Regression	0.676923	0.617225

	LR Train	LR Test	Tuned LR Train	Tuned LR Test	LDA Train	LDA Test	CART Train	CART Test
Accuracy	0.68	0.61	0.68	0.61	0.68	0.63	0.74	0.70
AUC	0.703	0.703	0.703	0.703	0.703	0.705	0.803	0.803
Recall	0.51	0.49	0.51	0.50	0.49	0.48	0.65	0.60
Precision	0.68	0.57	0.68	0.57	0.70	0.59	0.74	0.68
F1 Score	0.58	0.53	0.58	0.53	0.57	0.53	0.69	0.64

	ROC
LR Train	
LR Test	
Tuned LR Train	
Tuned LR Test	

LDA Train	
LDA Test	
CART Train	

CART Test	

Accuracy- How accurately the model classifies the data point.
More the accuracy, lesser the false predictions.

Sensitivity/ Recall- How many of actual True data points are identified as True data points by the model.

Precision- Among the points identified as positive by the mode, how many are actual positives.

AUC score represents degree/ measure of separability. i.e. how much the model is capable of distinguish ing between classes.
Value closer to 1 tells that there is good separability between the predicted classes and thus the model is good for prediction.

ROC Curve – For visualizing the classifier performance.
Steeper the ROC curve, stronger the model.

F1 score helps to know if Type 1 / Type 2 error is high/ low on average.

- Tuning the logistic regression model does not seem to have a significant impact on its performance. The accuracy, AUC, and other metrics remain quite similar to the previous version.
- Comparing the metrices, we can observe that the CART model has performed better than other models with an overall accuracy of 70-74 % for both our classes of interest.
- Considering the Recall value, CART is able to identify 65% of true positives accurately
- AUC captured is 80% for both test and train data.
- No overfitting or underfitting is observed and overall the CART model is a good model for classification.

2.4 Inference: Basis on these predictions, what are the insights and recommendations.

- From the models, we can conclude that the Number of children born and the wife's age are two important factors that can influence the decision if the women will use contraceptive methods or not.
- In terms of education, for both men and women, Tertiary is the highest education level which is obtained by most of them. While comparatively, uneducated men/ women are of least numbers.
- The more the people are educated, the more they opt for contraceptive methods.
- Also, there are a greater number of educated men than women.
- Majority of people have 1- 4 children, but a few also have more than 15.
- There is a moderate positive linear relationship between wife age and the number of children born. Older wives tend to have more children.
- Non- working women have more exposure towards media which may also be the reason for their high proportion of using contraceptives.
- There is more frequency at level 3 Occupation of the Husbands followed by level 1 and level 4 being the least number which is directly proportional to the usage of contraceptives.
- Also, people falling under very high and high standard of living have higher volume of people using contraceptive methods than those under low / very low standards.
- Exposure towards media, education and occupation plays a vital role in contraceptive methods usage.
- It is advised to reach out to women who do not use contraceptives to understand the reason and educate them about its needs and usage.

THE END.