Predictive Modeling

Business Report

Jayeesha Chakraborty 6/6/24

Table of Contents

Classification Model Building

Actionable Insights & Recommendations

Data Dictionary of Problem 1

Column Name	Description	Data Type
lread	Reads (transfers per second) between system memory and user memory	int64
lwrite	Writes (transfers per second) between system memory and user memory	int64
scall	Number of system calls of all types per second	int64
sread	Number of system read calls per second	int64
swrite	Number of system write calls per second	float64
fork	Number of system fork calls per second	float64
exec	Number of system exec calls per second	float64
rchar	Number of characters transferred per second by system read calls	float64
wchar	Number of characters transferred per second by system write calls	float64
pgout	Number of page out requests per second	float64
ppgout	Number of pages paged out per second	float64
pgfree	Number of pages per second placed on the free list	float64
pgscan	Number of pages checked if they can be freed per second	float64
atch	Number of page attaches (satisfying a page fault by reclaiming a page in memory) per second	float64
pgin	Number of page-in requests per second	float64
ppgin	Number of pages paged in per second	float64
pflt	Number of page faults caused by protection errors (copy-on-writes)	float64
vflt	Number of page faults caused by address translation	object
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.)	int64
freemem	Number of memory pages available to user processes	int64
freeswap	Number of disk blocks available for page swapping	int64
usr	Portion of time (%) that CPUs run in user mode	int64

Data Dictionary of Problem 2

Column Name	Description	Data Type
Wife_age	Wife Age	float64
Wife_ education	Wife Education 1=uneducated, 2, 3, 4=tertiary	object
Husband_education	Husband Education 1=uneducated, 2, 3, 4=tertiary	object
No_of_children_born	No of children born	float64
Wife_religion	Wife Religion - Non-Scientology, Scientology	object
Wife_Working	Wife's now working? (binary) Yes, No	object
Husband_Occupation	Husband Ocupation 1, 2, 3, 4(random)	int64
Standard_of_living_index	Standard of Living 1=very low, 2, 3, 4=high	object
Media_exposure	Media Exposure – Good, Not Good	object
Contraceptive_method_used	Contraceptive User or not	object

Problem 1

Context

There is an age old computer Sun Sparcstation 20/712 with 128 Mbytes of memory sitting in a multi-user university department where it has been extensively user for internet access, file editing and other CPU intensive programs.

Objective

As an aspiring data scientist, the job is to predict the percentage of time CPUs operating in user mode from the huge dataset of different measures of the computer system given.

Data Overview

- **Shape :** There are 8192 records and 22 columns .
- **Data Types:** There are 13 float64, 8 int64 and 1 object datatypes, overall 21numerical and 1 categorical variables present in the dataset.
- **Independent & Target Variable:** There are 21 independent features and **usr (** the% of time CPUs in user mode) is the target variable.
- **Check Duplicates:** There are no duplicate records in the database.
- **Check Null Values:** There are 104 null values in **rchar** and 15 in **wchar**
- **Statistical Description:**

Metric	Count	Unique	Top	Frequency	Mean	Std	Min	25%	50%	75%	Max
lread	8192.0	NaN	NaN	NaN	19.559692	53.353799	0.0	2.0	7.0	20.0	1845.0
lwrite	8192.0	NaN	NaN	NaN	13.106201	29.891726	0.0	0.0	1.0	10.0	575.0
scall	8192.0	NaN	NaN	NaN	2306.318237	1633.617322	109.0	1012.0	2051.5	3317.25	12493.0
sread	8192.0	NaN	NaN	NaN	210.47998	198.980146	6.0	86.0	166.0	279.0	5318.0
swrite	8192.0	NaN	NaN	NaN	150.058228	160.47898	7.0	63.0	117.0	185.0	5456.0
fork	8192.0	NaN	NaN	NaN	1.884554	2.479493	0.0	0.4	0.8	2.2	20.12
exec	8192.0	NaN	NaN	NaN	2.791998	5.212456	0.0	0.2	1.2	2.8	59.56
rchar	8088.0	NaN	NaN	NaN	197385.73	239837.49	278.0	34091.5	125473.5	267828.75	2526649.0
wchar	8177.0	NaN	NaN	NaN	95902.99	140841.71	1498.0	22916.0	46619.0	106101.0	1801623.0
pgout	8192.0	NaN	NaN	NaN	2.285317	5.307038	0.0	0.0	0.0	2.4	81.44
ppgout	8192.0	NaN	NaN	NaN	5.977229	15.21459	0.0	0.0	0.0	4.2	184.2
pgfree	8192.0	NaN	NaN	NaN	11.919712	32.36352	0.0	0.0	0.0	5.0	523.0

pgscan	8192.0	NaN	NaN	NaN	21.526849	71.14134	0.0	0.0	0.0	0.0	1237.0
atch	8192.0	NaN	NaN	NaN	1.127505	5.708347	0.0	0.0	0.0	0.6	211.58
pgin	8192	NaN	NaN	NaN	8.27796	13.87498	0	0.6	2.8	9.765	141.2
ppgin	8192	NaN	NaN	NaN	12.38859	22.28132	0	0.6	3.8	13.8	292.61
pflt	8192	NaN	NaN	NaN	109.7938	114.4192	0	25	63.8	159.6	899.8
vflt	8192	NaN	NaN	NaN	185.3158	191.0006	0.2	45.4	120.4	251.8	1365
runqsz	8192	2	Not_CPU_Bound	4331	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freemem	8192	NaN	NaN	NaN	1763.456	2482.105	55	231	579	2002.25	12027
freeswap	8192	NaN	NaN	NaN	1328126	422019.4	2	1042624	1289290	1730380	2243187
usr	8192	NaN	NaN	NaN	83.96887	18.40191	0	81	89	94	99

- **Observations:**
  - ✓ All the numerical variables are right or left skewed.
  - ✓ The categorical variable **runqsz** has 2 values CPU- Bound and Not CPU Bound. Out of 8912 records 4331 records are Not CPU Bound which contributes around 50% of the data.
  - ✓ There are records where lread and lwrite are 0 means no read or write calls happened between system and user memory.
  - ✓ The target variable **usr** is 0 in a few records which is not possible in reality.However , for 75% of records **usr** is less than 94%.

Univariate Analysis

- Categorical Variable :

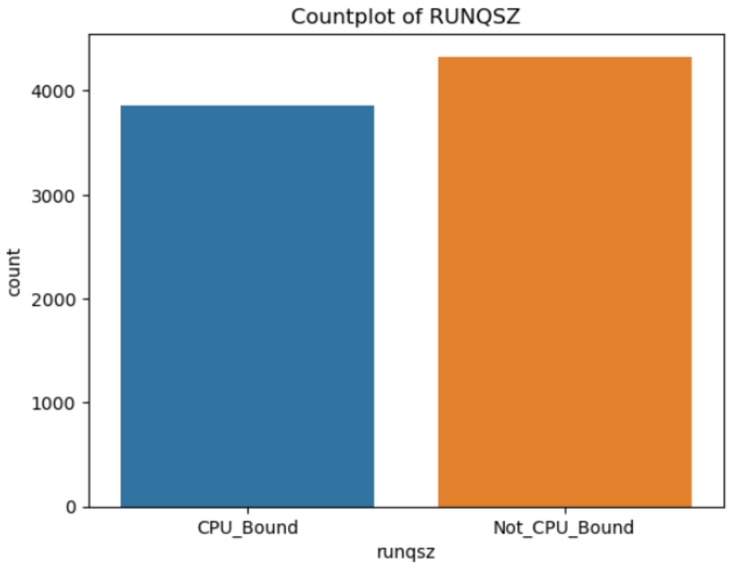


Fig 1 Count Plot of runqsz

There are 4331 operations having Not CPU Bound and 3861 CPU Bounds.

- Numerical Variables:

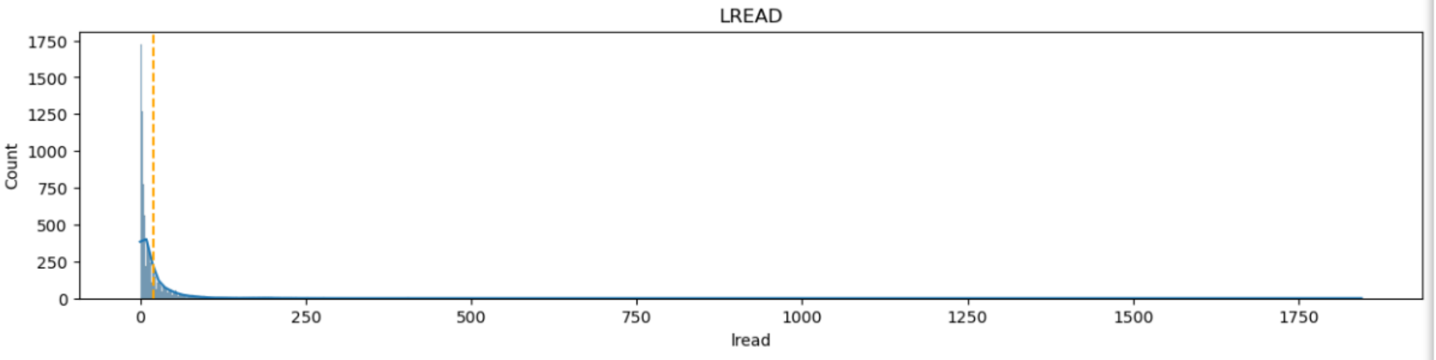


Fig 2 Histogram of lread

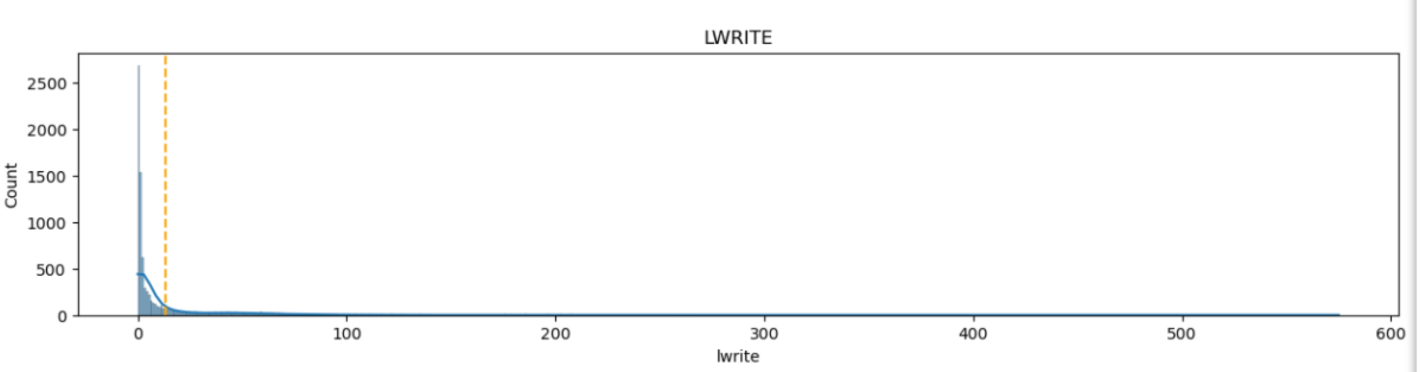


Fig 3 Histogram of lwrite

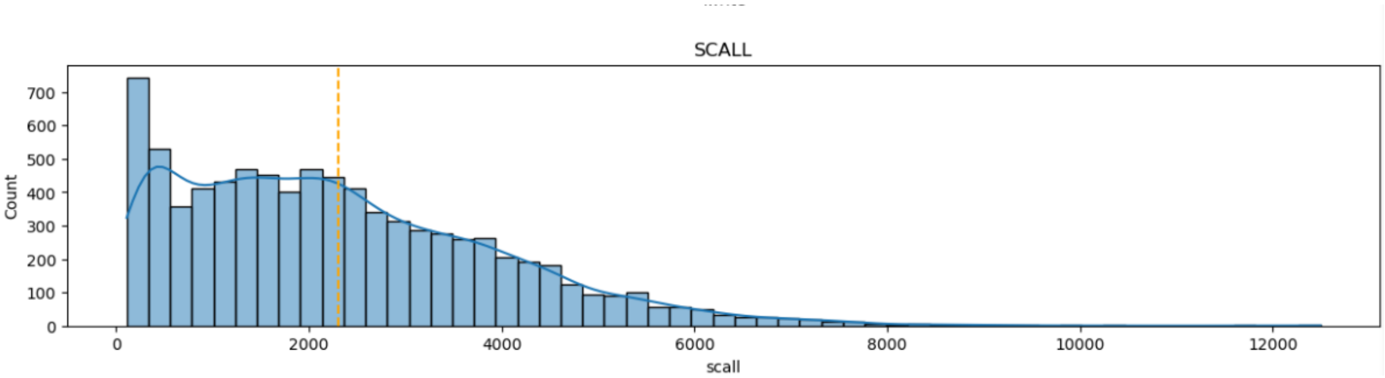


Fig 3 Histogram of scall

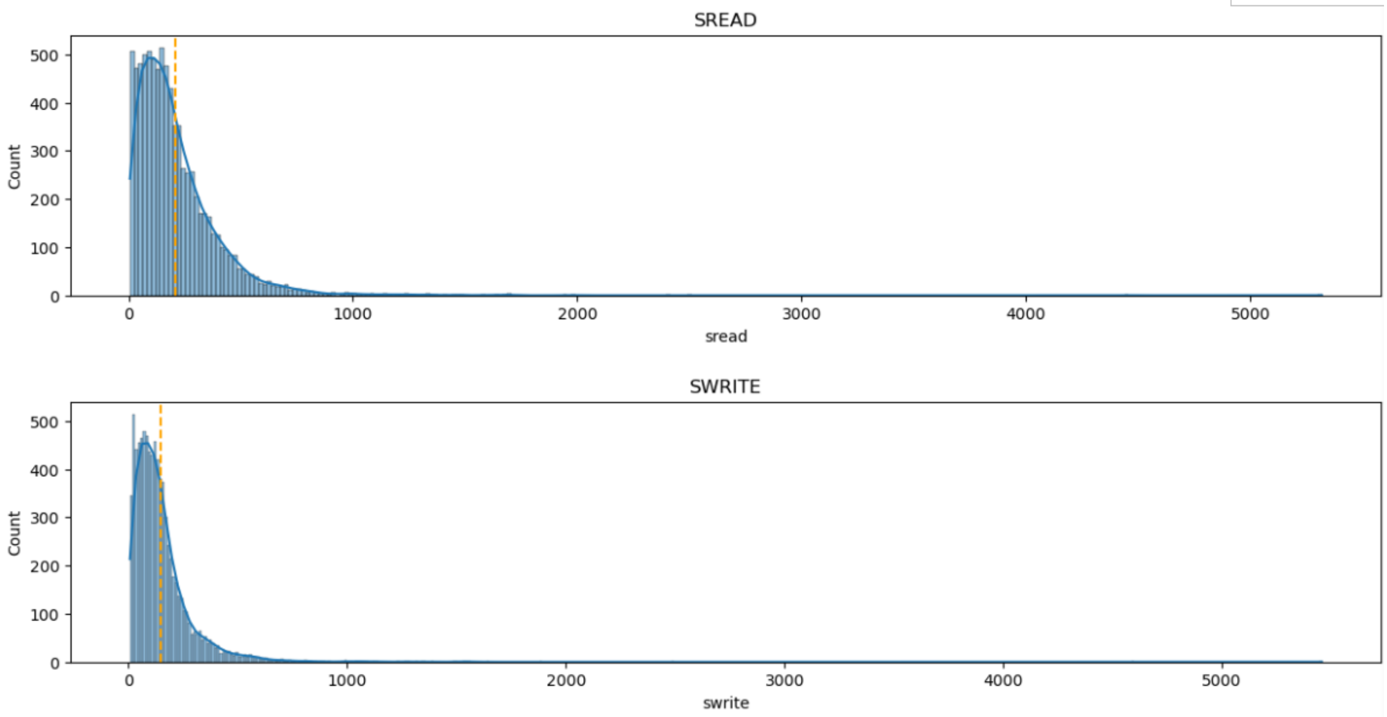


Fig 4 Histogram of sread and swrite

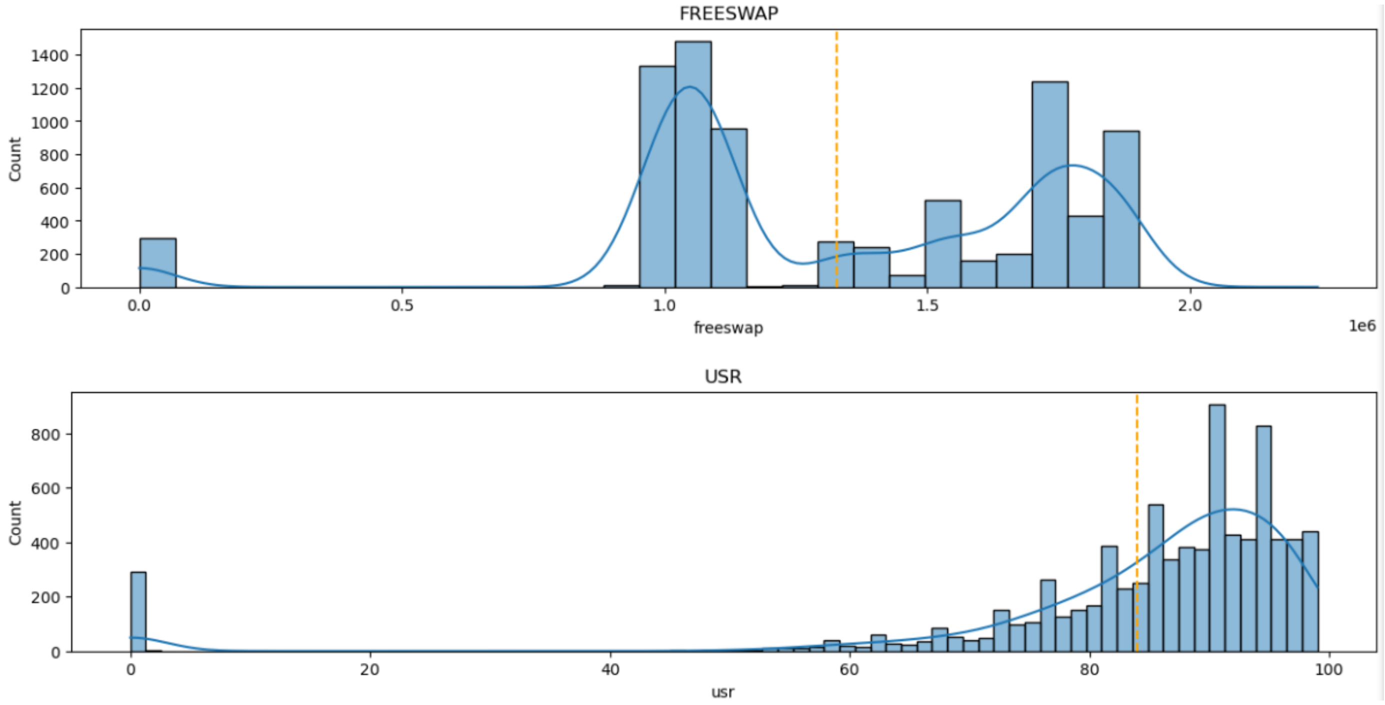


Fig 5 Histogram of freeswap and usr

Observations:

- ✓ As we see from the above figures, most of the numeric features are right skewed except freemem, freeswap and usr.
- ✓ The target variable usr is left skewed.
- ✓ Skewness in values denotes outliers in the dataset which needs to be handled appropriately.
- ✓ Freeswap doesn’t really follow normal distribution

Bivariate Analysis:

Reg plot between dependent and independent variables

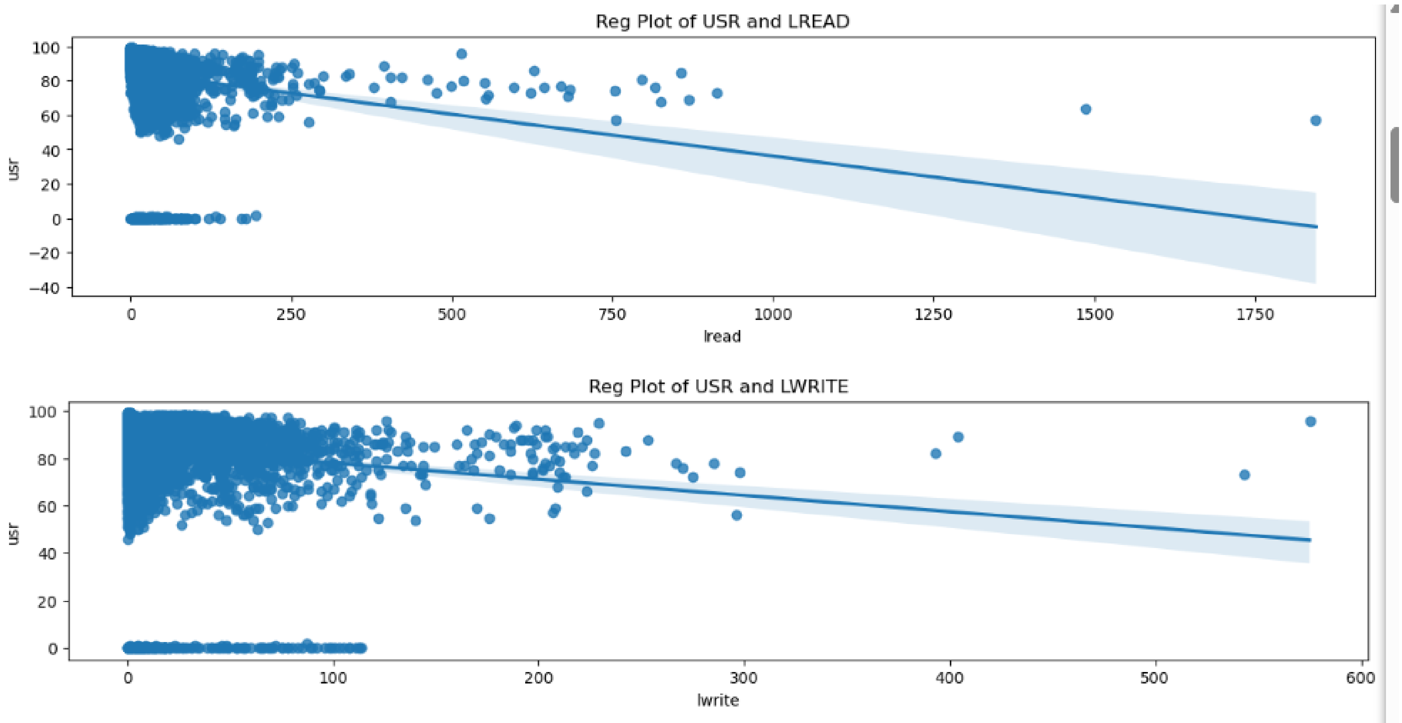


Fig 6 Reg Plot of lread and lwrite with usr

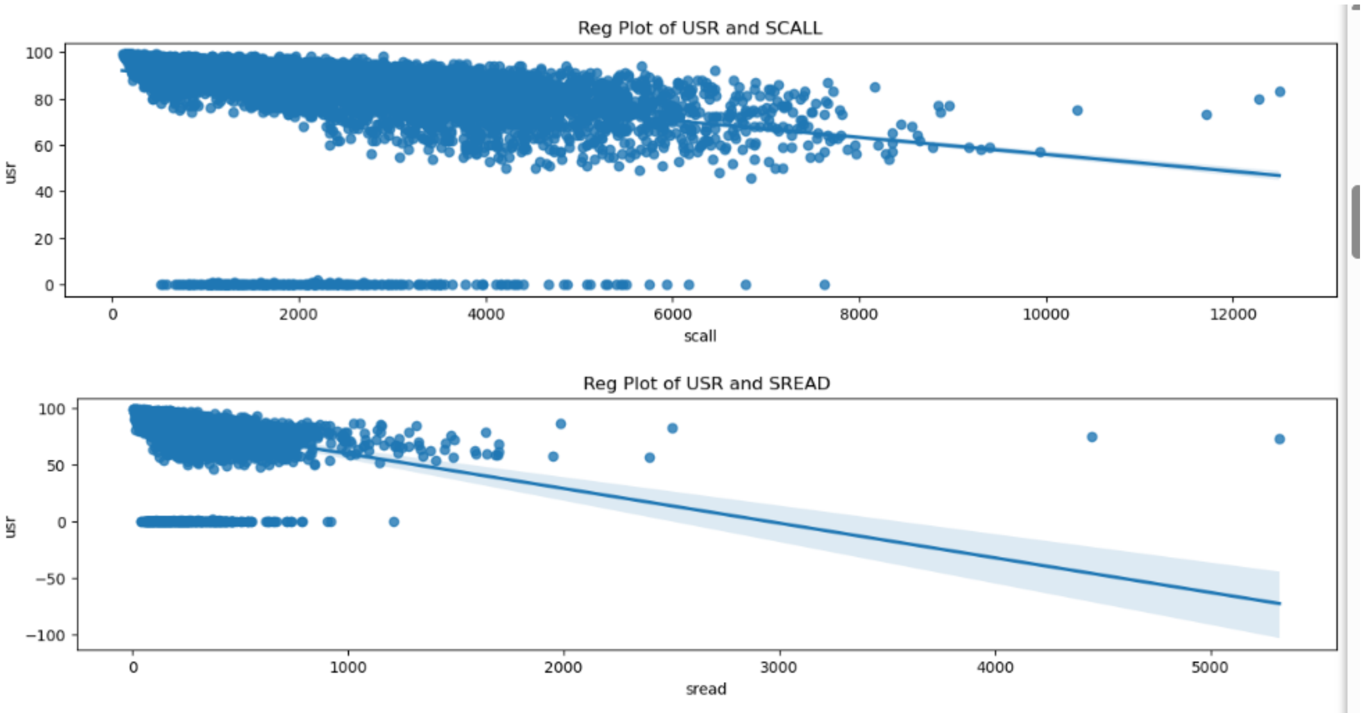


Fig 7 Reg plot scall and sread with usr

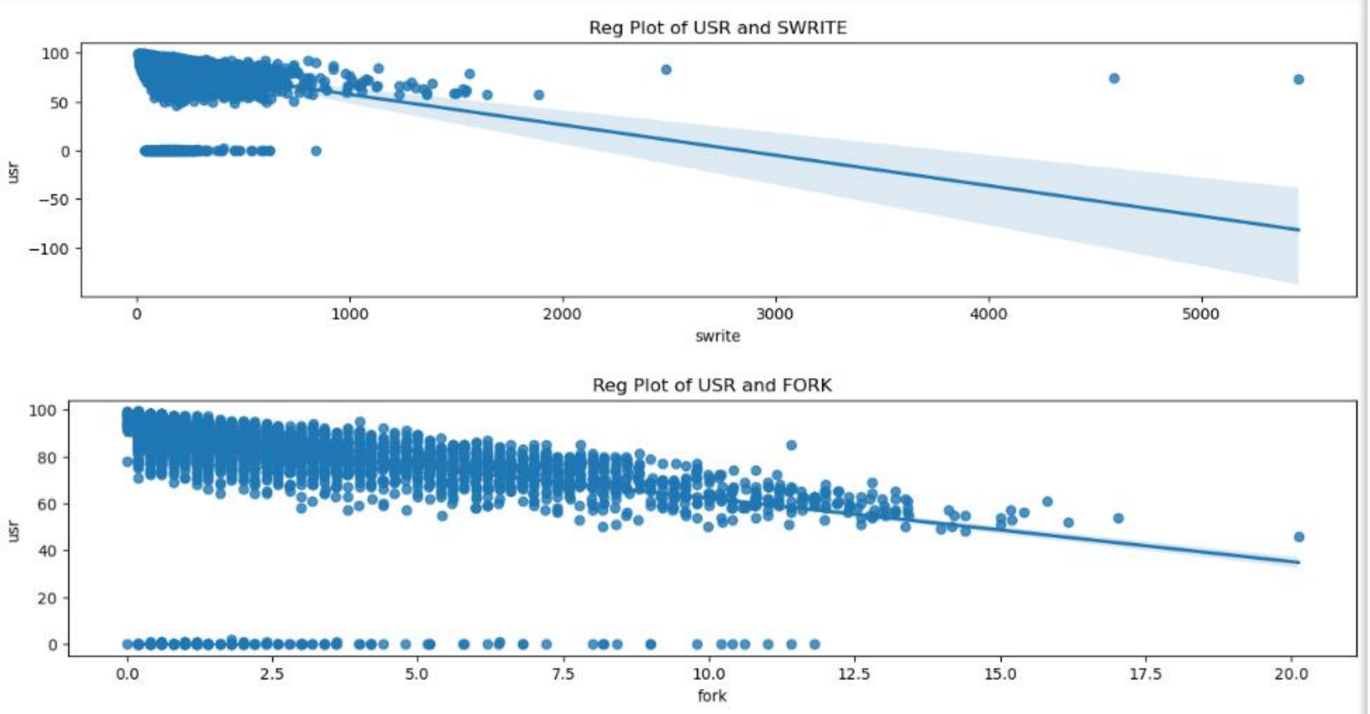


Fig 8 Reg plot swrite and fork with usr

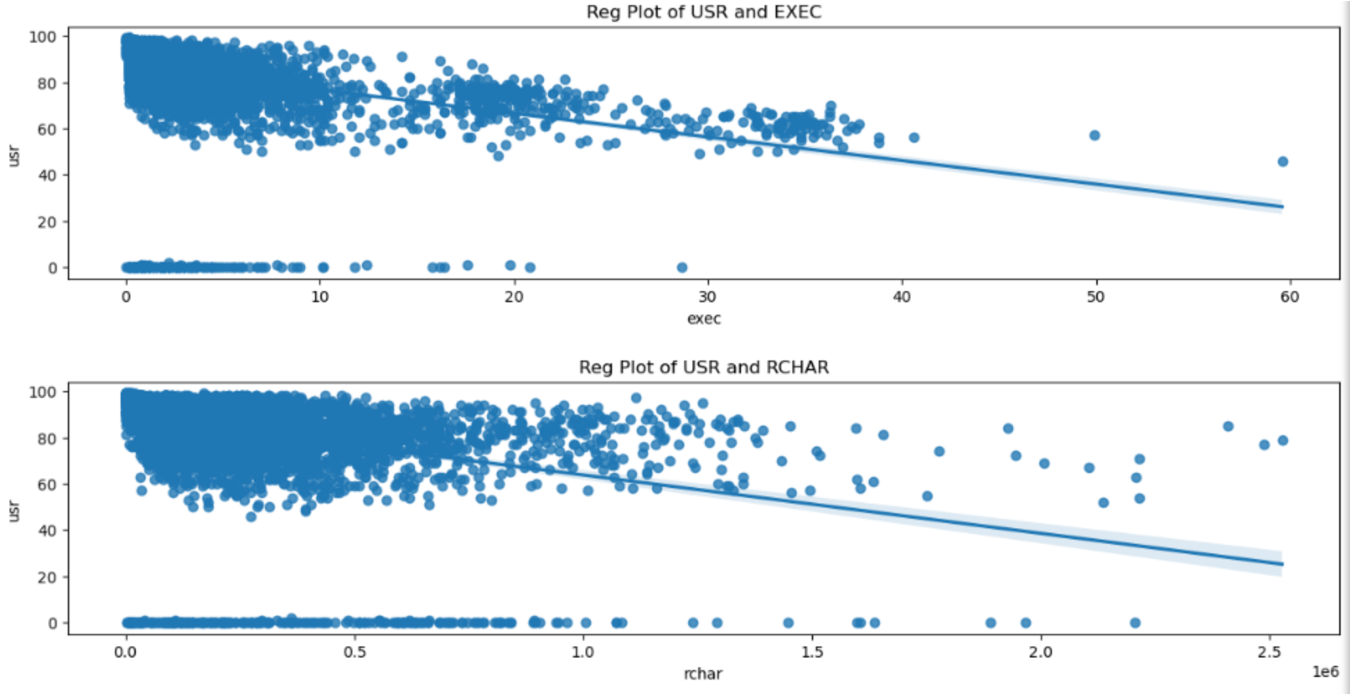


Fig 9 Reg plot of exec and rchar with usr

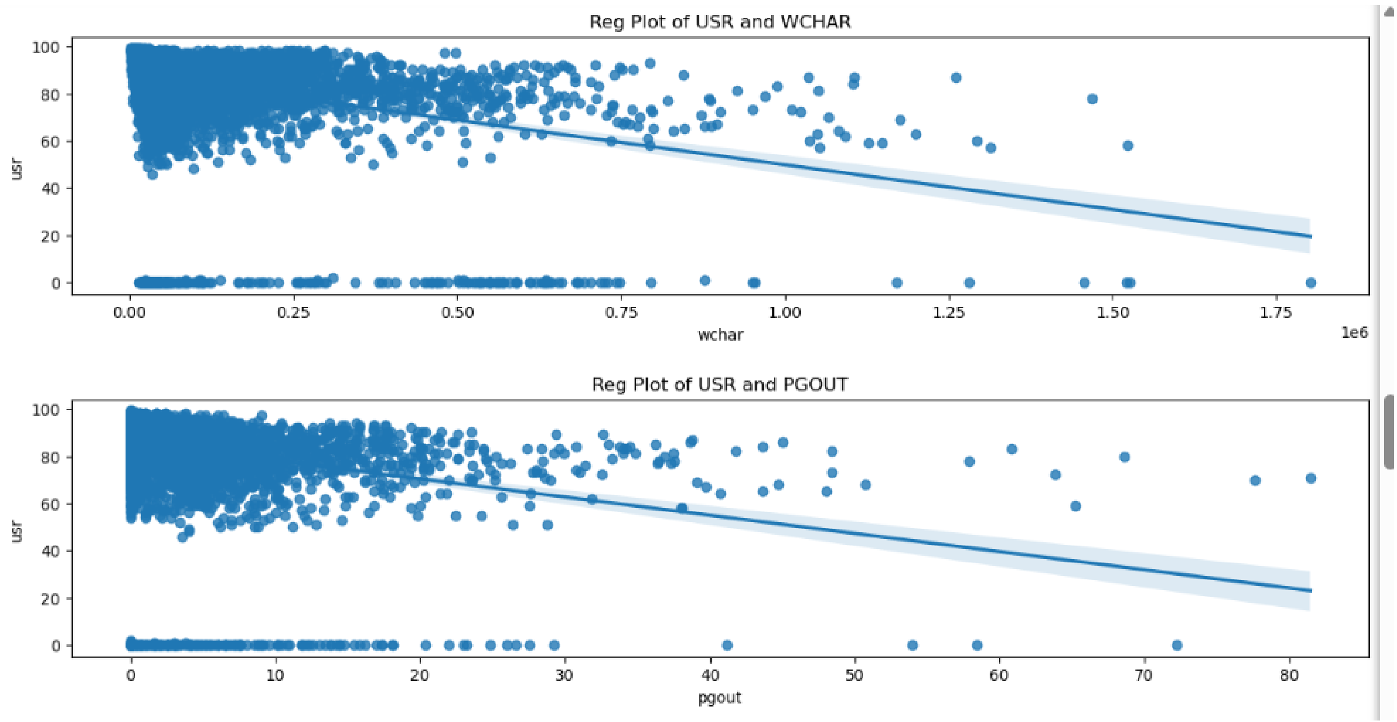


Fig 10 reg plot of wchar and pgout with usr

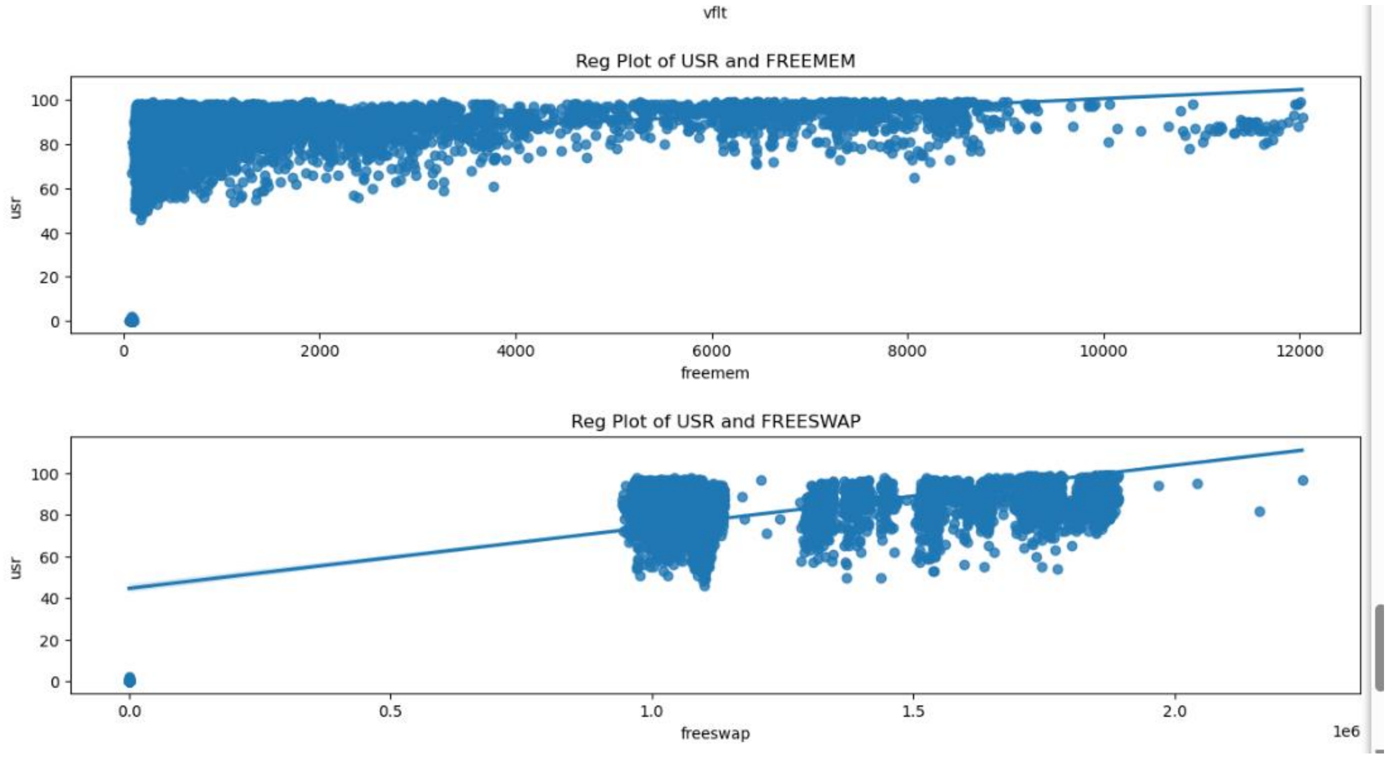


Fig 11 reg plot of freemem and freeswap with usr

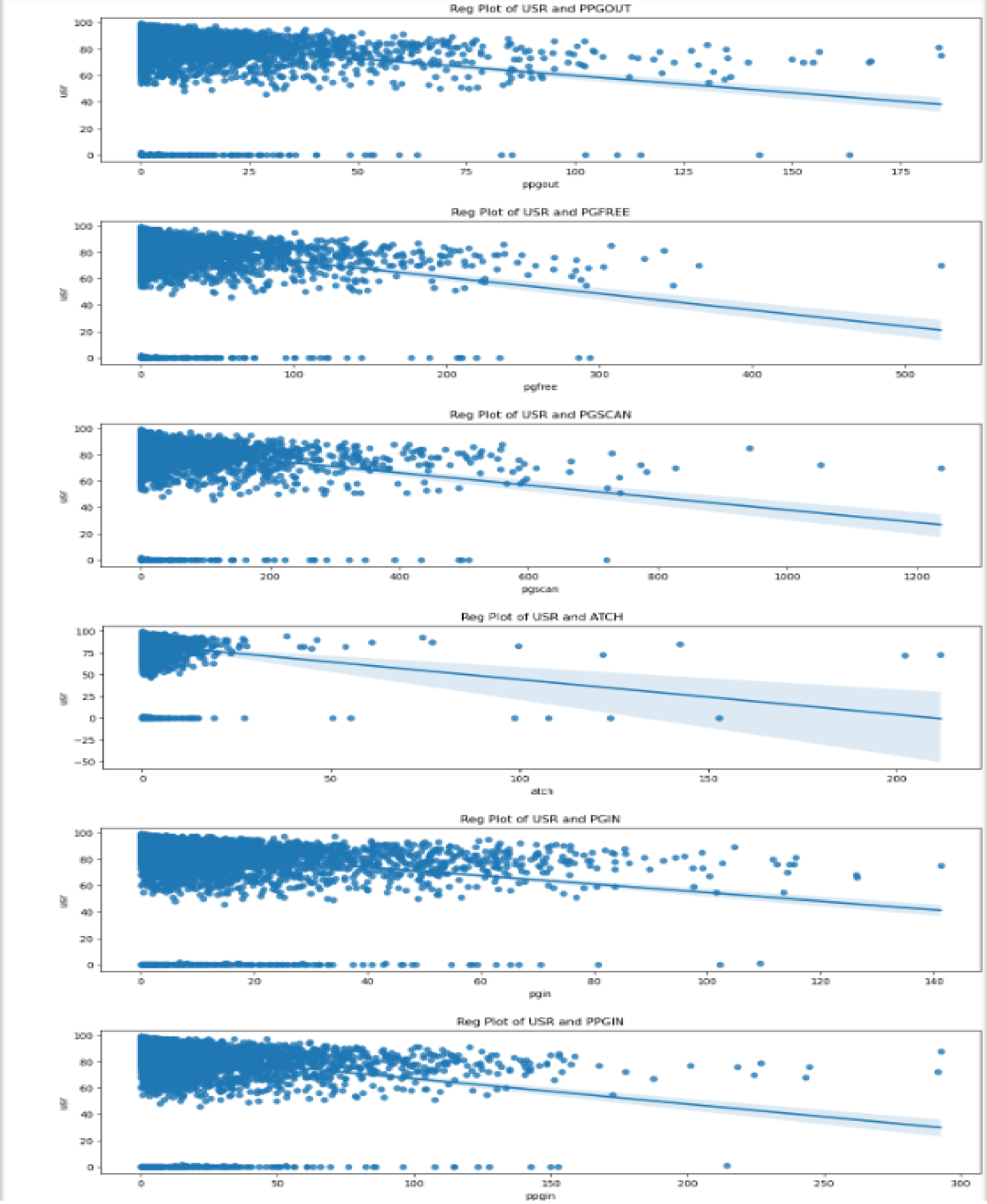


Fig 12 reg plot of rest of features with usr

Observations:

- ✓ Most of the numeric features like lread, lwrite, scalls, sread, swrite, fork, exec, ppgout etc are in Negative linear relationship with usr except freemem and freeswap
- ✓ Freemem and freeswap are in positive linear relationship with usr.

Heatmap:

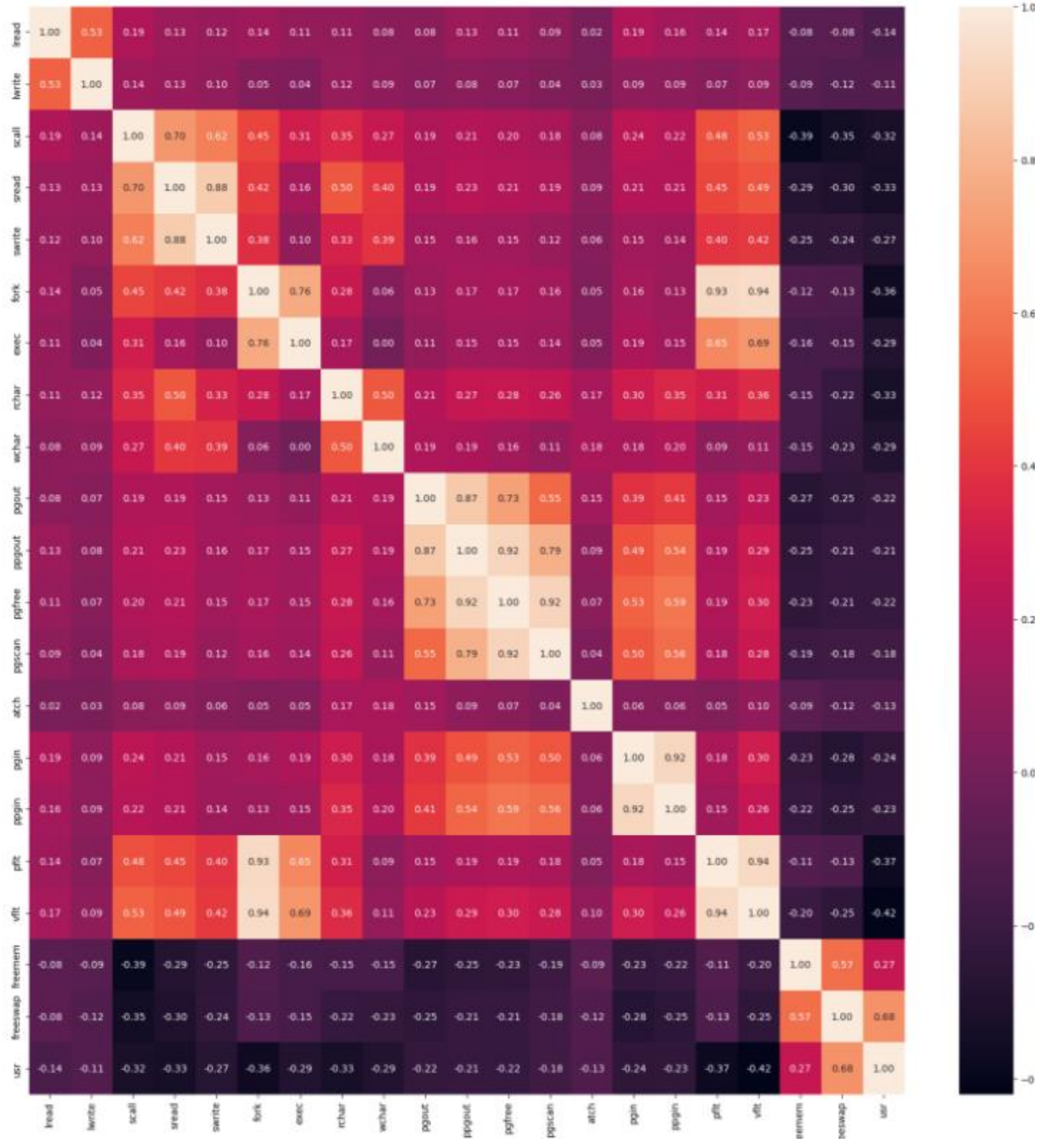


Fig 13 Heatmap correlation between numeric variables

Observation:

- ✓ There is 94% collinearity between pflt and vflt, we can keep either of these 2 features to predict usr
- ✓ There is 92% collinearity between ppgin and pgin, we can keep either of these 2 features to predict usr
- ✓ Pgfree is 92% collinear with pgscan and ppgout.
- ✓ There is 87% collinearity between ppgout and pgout.
- ✓ There is 88% collinearity between sread and swrite.

Pairplot

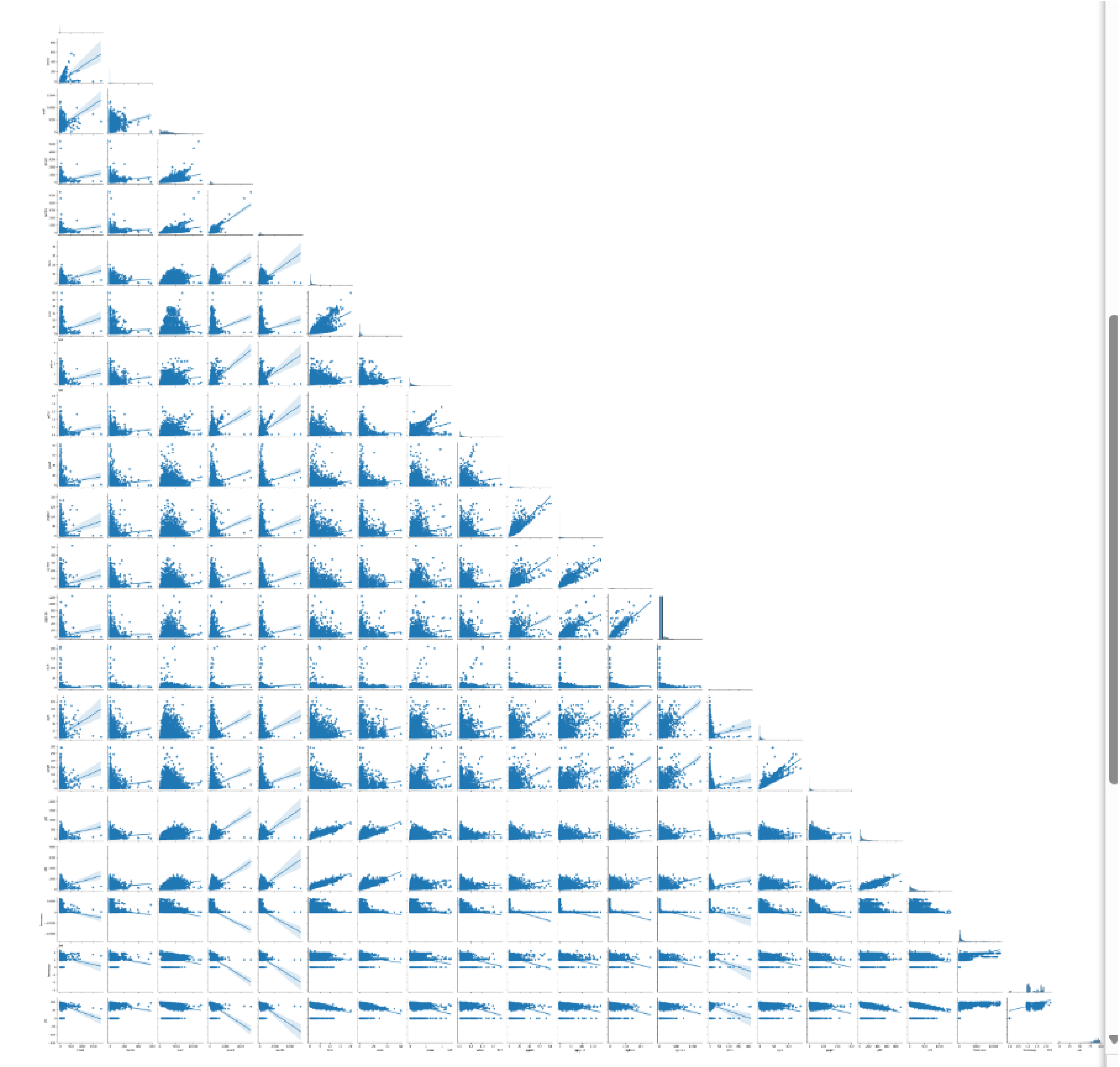


Fig 14 Pairplot of Numeric Variables

Multivariate Analysis :

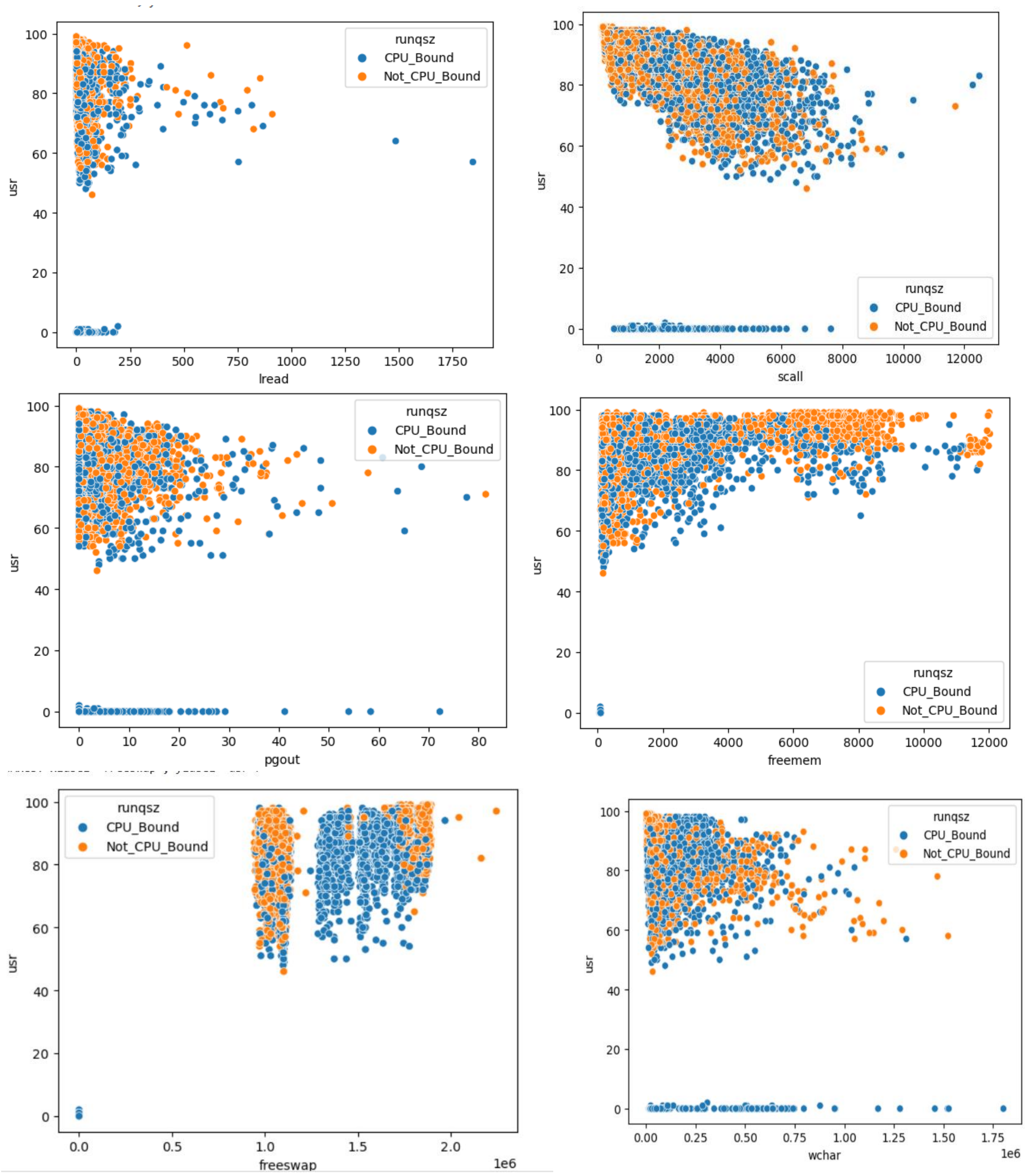


Fig 15 Scatterplots of usr and independent variables with variation of runqsz

Observations:

- ✓ The percentage of time CPU in user mode is high when the read transfer between system and user Memory is low. The trend is same for both CPU bound and CPU Not Bound.

- ✓ With increase of scalls, usr tends to decrease in linear fashion.
- ✓ Usr tends to decrease in the range of 1 to 2 of freeswap. CPU bound operations are predominant Compared to CPU not bound.
- ✓ Wchar and pgout follow the same trend with usr.

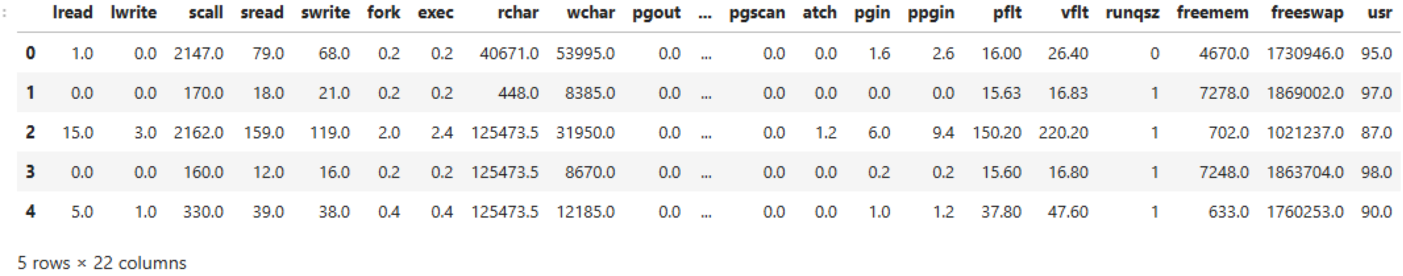
Data Pre-processing:

Missing Value Treatment:

As we saw earlier there are missing values in rchar and wchar columns. There are outliers in both columns. So we are going to impute the null values with median of individual column values.

Encoding:

There is one categorical variable runsqz having 2 values CPU_Bound and Not_ CPU_Bound. We need to encode this variable to build linear regression model. We have used label encoding techniques by which runqsz CPU Bound converted 0 and CPU Not Bound converted to 1.
Also, the data type of runqsz changed to numerical from object type.



	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap	usr
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	1.6	2.6	16.00	26.40	0	4670.0	1730946.0	95.0
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	15.63	16.83	1	7278.0	1869002.0	97.0
2	15.0	3.0	2162.0	159.0	119.0	2.0	2.4	125473.5	31950.0	0.0	...	0.0	1.2	6.0	9.4	150.20	220.20	1	702.0	1021237.0	87.0
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	0.2	0.2	15.60	16.80	1	7248.0	1863704.0	98.0
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	1.0	1.2	37.80	47.60	1	633.0	1760253.0	90.0

5 rows × 22 columns

Fig 16 Data table after encoding

Outlier Treatment

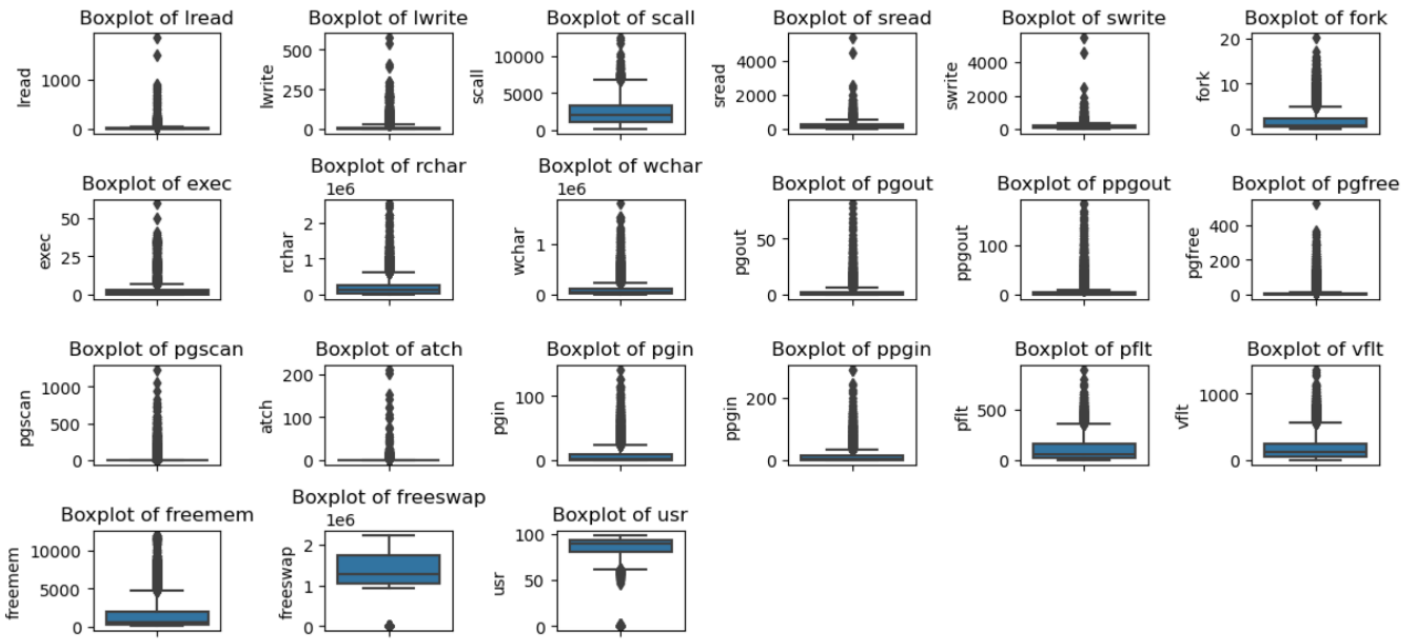


Fig 17 Boxplot of Numerical Features

After imputing the outliers with lower quartile (25%) and upper quartile (75%) values, we get the below boxplot of all numerical variables.

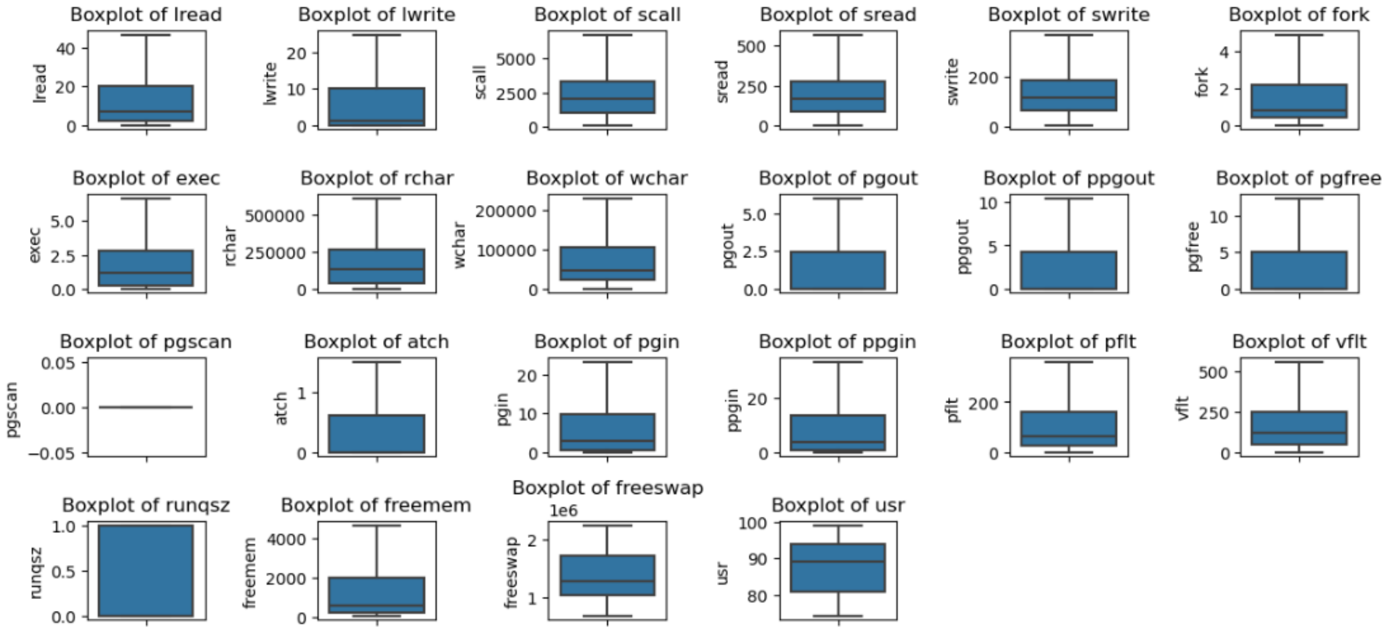


Fig 18 Boxplot of Numerical Features after outlier treatment

Feature Engineering:

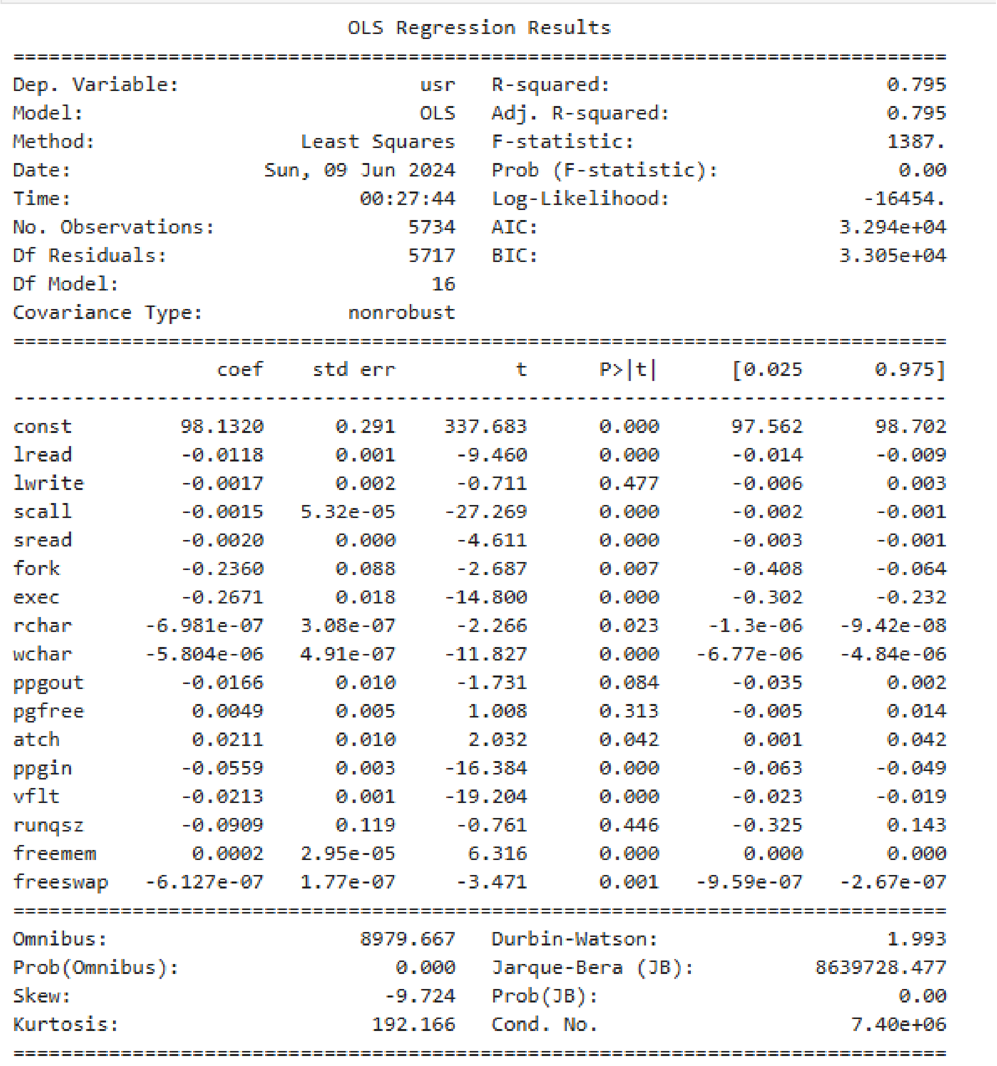
- ✓ As we saw in heatmap there is a strong collinearity between some of the features. We can remove one of those independent variables where collinearity exists. We are removing **pflt, pgin, pgscan, swrite, pgout** because of strong collinearity.
  Hence the number of columns of the dataset got reduced to 17 from 22.
- ✓ We see from the above graphs that there are zeros in usr which is not possible in reality. We are Going to impute the zeros with median as there are outliers in the column.

Train Test Split:

- ✓ We split the whole data set into 2 data frames X & Y. X contains all the independent variables and Y contains the target variable.
- ✓ We further split them into train and test data set with train size of 70% which means 70% data is being Used for model training and 30% for testing.
- ✓ There are 5734 records in train data set and 2458 records in test.

Building Linear Regression Models

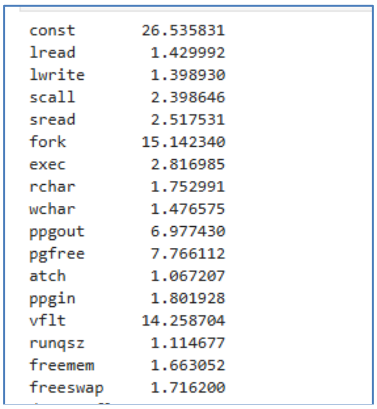
Linear Regression using statsmodels:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.795
Model:                            OLS   Adj. R-squared:                  0.795
Method:                 Least Squares   F-statistic:                     1387.
Date:                Sun, 09 Jun 2024   Prob (F-statistic):               0.00
Time:                        00:27:44   Log-Likelihood:                -16454.
No. Observations:                5734   AIC:                         3.294e+04
Df Residuals:                    5717   BIC:                         3.305e+04
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         98.1320      0.291    337.683      0.000      97.562      98.702
lread         -0.0118      0.001     -9.460      0.000      -0.014      -0.009
lwrite        -0.0017      0.002     -0.711      0.477      -0.006       0.003
scall         -0.0015   5.32e-05    -27.269      0.000      -0.002      -0.001
sread         -0.0020      0.000     -4.611      0.000      -0.003      -0.001
fork          -0.2360      0.088     -2.687      0.007      -0.408      -0.064
exec          -0.2671      0.018    -14.800      0.000      -0.302      -0.232
rchar      -6.981e-07   3.08e-07     -2.266      0.023    -1.3e-06   -9.42e-08
wchar      -5.804e-06   4.91e-07    -11.827      0.000   -6.77e-06   -4.84e-06
ppgout        -0.0166      0.010     -1.731      0.084      -0.035       0.002
pgfree         0.0049      0.005      1.008      0.313      -0.005       0.014
atch           0.0211      0.010      2.032      0.042       0.001       0.042
ppgin         -0.0559      0.003    -16.384      0.000      -0.063      -0.049
vflt          -0.0213      0.001    -19.204      0.000      -0.023      -0.019
runqsz        -0.0909      0.119     -0.761      0.446      -0.325       0.143
freemem        0.0002   2.95e-05      6.316      0.000       0.000       0.000
freeswap   -6.127e-07   1.77e-07     -3.471      0.001   -9.59e-07   -2.67e-07
==============================================================================
Omnibus:                     8979.667   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          8639728.477
Skew:                          -9.724   Prob(JB):                         0.00
Kurtosis:                     192.166   Cond. No.                     7.40e+06
==============================================================================
```

Fig 19 Linear Regression Model using statsmodels

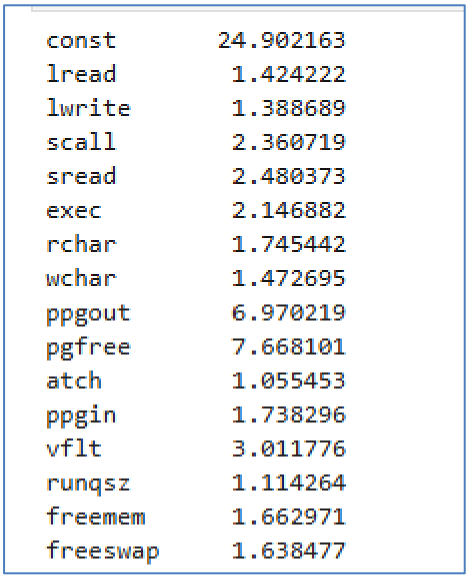
- R-squared and adj R squared values are 79% which is quite good for this model. Let’s check if there is any Multi-collinearity present using variance inflation factor. VIF beyond 5 indicated that there is Multi-collinearity among the predictor variables.



```
const       26.535831
lread        1.429992
lwrite       1.398930
scall        2.398646
sread        2.517531
fork        15.142340
exec         2.816985
rchar        1.752991
wchar        1.476575
ppgout       6.977430
pgfree       7.766112
atch         1.067207
ppgin        1.801928
vflt        14.258704
runqsz       1.114677
freemem      1.663052
freeswap     1.716200
```

Fig 20 VIF of Independent Variables

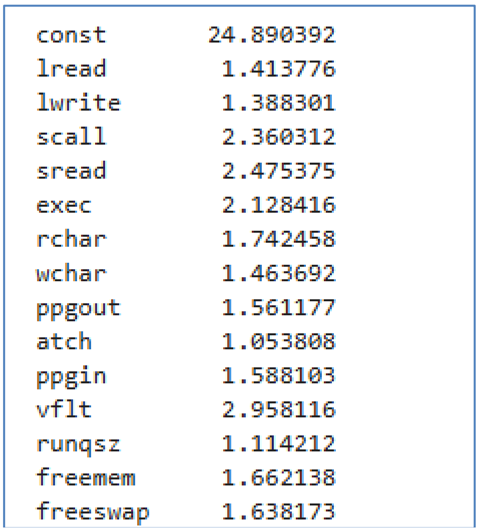
- vif is more than 5 for fork, ppgout, pgfree and vflt. This indicates that these features are related to one or more independent variables. We can drop the features one by one and check R-squared and adj R squared value to ensure the model performance is unaffected.
- After removing fork, we saw that model performance is unchanged, R-squared and adj R squared are 79.4% .
- We measure VIF again for the latest model.



```
const       24.902163
lread        1.424222
lwrite       1.388689
scall        2.360719
sread        2.480373
exec         2.146882
rchar        1.745442
wchar        1.472695
ppgout       6.970219
pgfree       7.668101
atch         1.055453
ppgin        1.738296
vflt         3.011776
runqsz       1.114264
freemem      1.662971
freeswap     1.638477
```

Fig 21 VIF of Independent Variables after removing fork

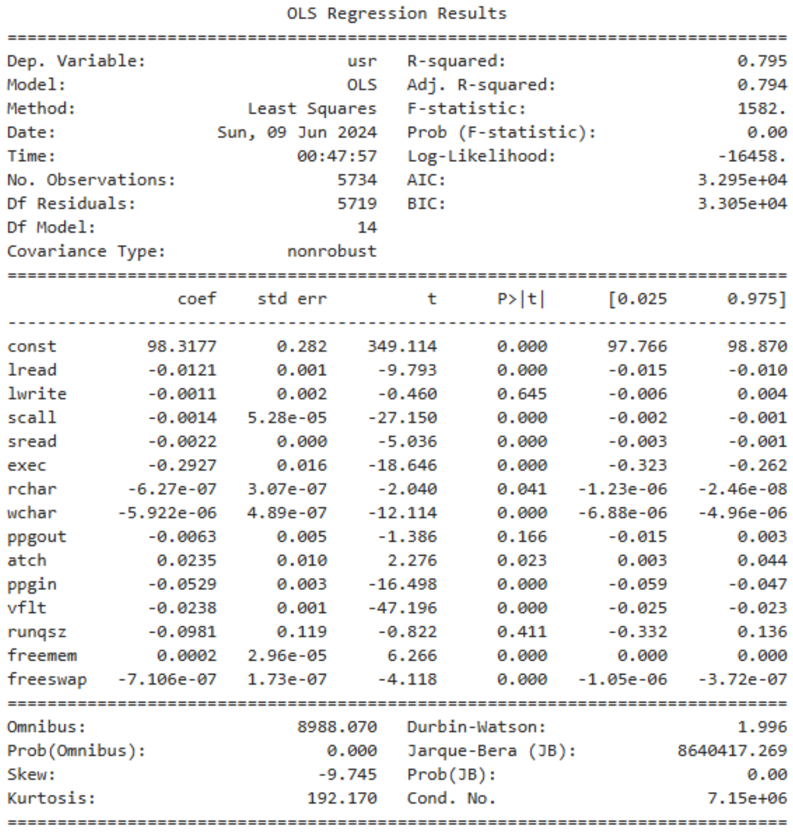
- Now we remove pgfree and again train the model. R-squared and adj R squared are still 79.4% .
- We see that VIF has significantly come down below 5 for all the independent variables. So we Can conclude that there is no more multi collinearity present.



```
const       24.890392
lread        1.413776
lwrite       1.388301
scall        2.360312
sread        2.475375
exec         2.128416
rchar        1.742458
wchar        1.463692
ppgout       1.561177
atch         1.053808
ppgin        1.588103
vflt         2.958116
runqsz       1.114212
freemem      1.662138
freeswap     1.638173
```

Fig 22 VIF of Independent Variables after removing pgfree

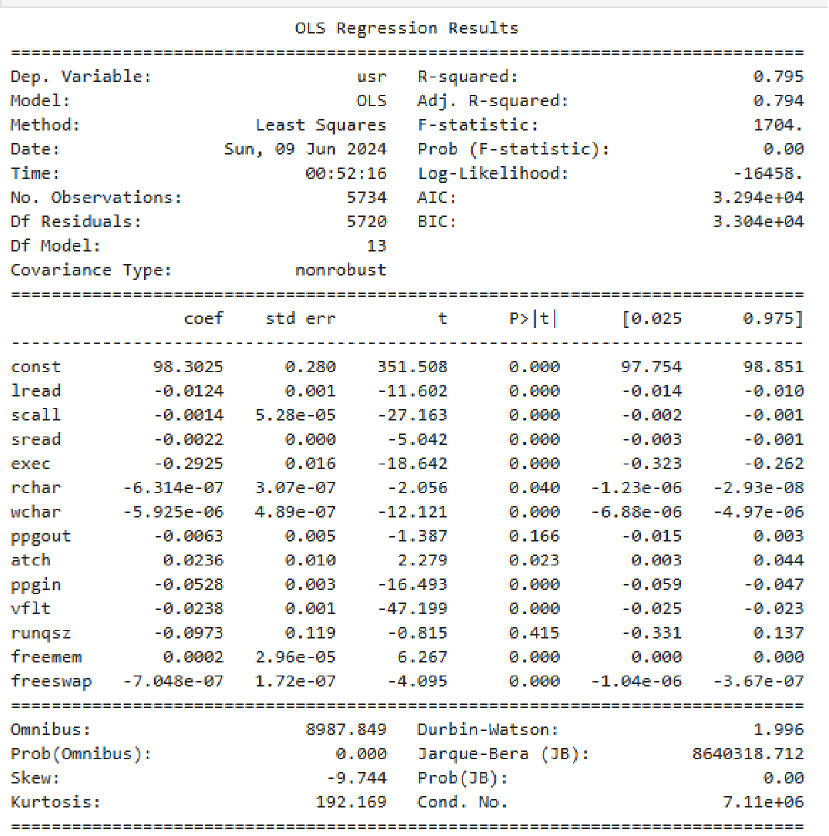
- Let’s check the model performance summary.



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.795
Model:                            OLS   Adj. R-squared:                  0.794
Method:                 Least Squares   F-statistic:                     1582.
Date:                Sun, 09 Jun 2024   Prob (F-statistic):               0.00
Time:                        00:47:57   Log-Likelihood:                -16458.
No. Observations:                5734   AIC:                         3.295e+04
Df Residuals:                    5719   BIC:                         3.305e+04
Df Model:                          14                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         98.3177      0.282    349.114      0.000      97.766      98.870
lread         -0.0121      0.001     -9.793      0.000      -0.015      -0.010
lwrite        -0.0011      0.002     -0.460      0.645      -0.006       0.004
scall         -0.0014   5.28e-05    -27.150      0.000      -0.002      -0.001
sread         -0.0022      0.000     -5.036      0.000      -0.003      -0.001
exec          -0.2927      0.016    -18.646      0.000      -0.323      -0.262
rchar       -6.27e-07   3.07e-07     -2.040      0.041   -1.23e-06   -2.46e-08
wchar      -5.922e-06   4.89e-07    -12.114      0.000   -6.88e-06   -4.96e-06
ppgout        -0.0063      0.005     -1.386      0.166      -0.015       0.003
atch           0.0235      0.010      2.276      0.023       0.003       0.044
ppgin         -0.0529      0.003    -16.498      0.000      -0.059      -0.047
vflt          -0.0238      0.001    -47.196      0.000      -0.025      -0.023
runqsz        -0.0981      0.119     -0.822      0.411      -0.332       0.136
freemem        0.0002   2.96e-05      6.266      0.000       0.000       0.000
freeswap   -7.106e-07   1.73e-07     -4.118      0.000   -1.05e-06   -3.72e-07
==============================================================================
Omnibus:                     8988.070   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          8640417.269
Skew:                          -9.745   Prob(JB):                         0.00
Kurtosis:                     192.170   Cond. No.                     7.15e+06
==============================================================================
```

Fig 23 Linear Regression Model using statsmodels after removing multicollinearity

- From the summary we see that p-value for lwrite coefficient is 0.64 i.e which is greater than 0.5. Hence we fail to reject the Null hypothesis(predictor variable is not significant). So we are going to remove lwrite as it is not significant variable.



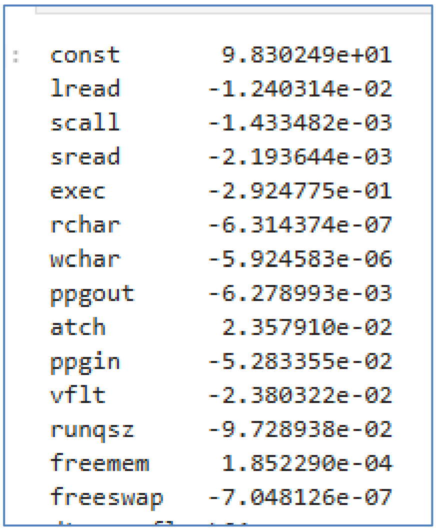
```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.795
Model:                            OLS   Adj. R-squared:                  0.794
Method:                 Least Squares   F-statistic:                     1704.
Date:                Sun, 09 Jun 2024   Prob (F-statistic):               0.00
Time:                        00:52:16   Log-Likelihood:                -16458.
No. Observations:                5734   AIC:                         3.294e+04
Df Residuals:                    5720   BIC:                         3.304e+04
Df Model:                          13                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         98.3025      0.280    351.508      0.000      97.754      98.851
lread         -0.0124      0.001    -11.602      0.000      -0.014      -0.010
scall         -0.0014   5.28e-05    -27.163      0.000      -0.002      -0.001
sread         -0.0022      0.000     -5.042      0.000      -0.003      -0.001
exec          -0.2925      0.016    -18.642      0.000      -0.323      -0.262
rchar      -6.314e-07   3.07e-07     -2.056      0.040   -1.23e-06   -2.93e-08
wchar      -5.925e-06   4.89e-07    -12.121      0.000   -6.88e-06   -4.97e-06
ppgout        -0.0063      0.005     -1.387      0.166      -0.015       0.003
atch           0.0236      0.010      2.279      0.023       0.003       0.044
ppgin         -0.0528      0.003    -16.493      0.000      -0.059      -0.047
vflt          -0.0238      0.001    -47.199      0.000      -0.025      -0.023
runqsz        -0.0973      0.119     -0.815      0.415      -0.331       0.137
freemem        0.0002   2.96e-05      6.267      0.000       0.000       0.000
freeswap   -7.048e-07   1.72e-07     -4.095      0.000   -1.04e-06   -3.67e-07
==============================================================================
Omnibus:                     8987.849   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          8640318.712
Skew:                          -9.744   Prob(JB):                         0.00
Kurtosis:                     192.169   Cond. No.                     7.11e+06
==============================================================================
```

Fig 24 Final Linear Regression Model

- We see that the relationship between atch,freemem features individually with usr is positive as the coefficients are positive, for the rest of the features the relationship is negative.

- R-squared and adjusted R-squared of the model are 0.79, which shows that the model is able to explain ~79% variance in the data. This is quite good.

- **The significant predictors variable of this linear regression models are lread, scall, sread, exec, rchar, Wchar, ppgout, atch, ppgin, vflt, runqsz, freemem and freeswap**

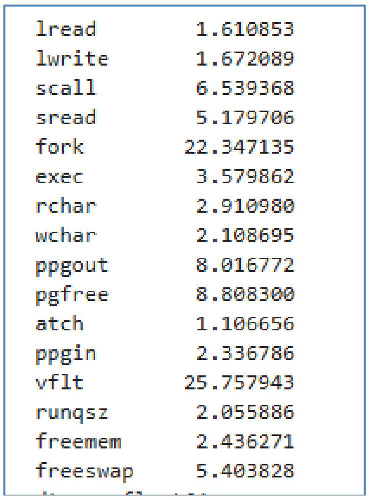


```
:  const        9.830249e+01
   lread       -1.240314e-02
   scall       -1.433482e-03
   sread       -2.193644e-03
   exec        -2.924775e-01
   rchar       -6.314374e-07
   wchar       -5.924583e-06
   ppgout      -6.278993e-03
   atch         2.357910e-02
   ppgin       -5.283355e-02
   vflt        -2.380322e-02
   runqsz      -9.728938e-02
   freemem      1.852290e-04
   freeswap    -7.048126e-07
```

Fig 25 Important Features with the coefficients

Build Linear Regression Model using skitlearn

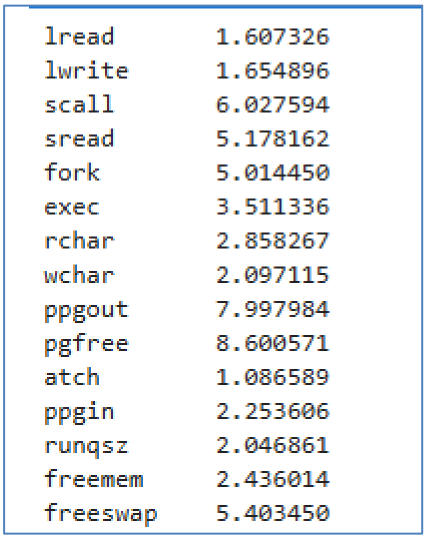
- After we split the data set into train and test set, we build the linear regression model using skitlearn And fit the train data. R- squared value of train data and test are 79% and 77% respectively which is Quite good.
- We follow the same steps as followed in the previous model i.e calculate VIF and check multicollinearity in the model.



```
lread        1.610853
lwrite       1.672089
scall        6.539368
sread        5.179706
fork        22.347135
exec         3.579862
rchar        2.910980
wchar        2.108695
ppgout       8.016772
pgfree       8.808300
atch         1.106656
ppgin        2.336786
vflt        25.757943
runqsz       2.055886
freemem      2.436271
freeswap     5.403828
```

Fig 26 VIF of independent variables using skitlearn linear regression model

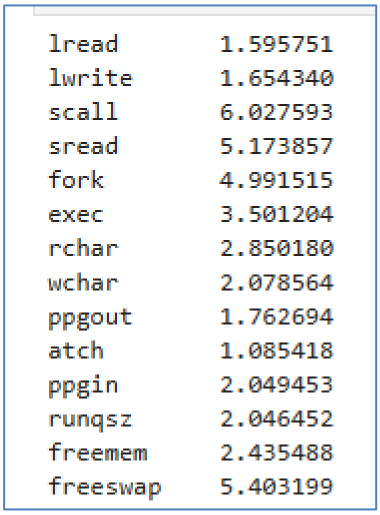
- We remove vflt column and rebuild the model. R- squared value of train data and test for the revised model are 78% and 76% respectively. Lets reverify VIF for the new model.



```
lread       1.607326
lwrite      1.654896
scall       6.027594
sread       5.178162
fork        5.014450
exec        3.511336
rchar       2.858267
wchar       2.097115
ppgout      7.997984
pgfree      8.600571
atch        1.086589
ppgin       2.253606
runqsz      2.046861
freemem     2.436014
freeswap    5.403450
```

Fig 27 VIF of independent variables after removing vflt

- We remove pgfree column and rebuild the model. R- squared value of train data and test for the revised model are still 78% and 76% respectively. Lets reverify VIF for the new model.

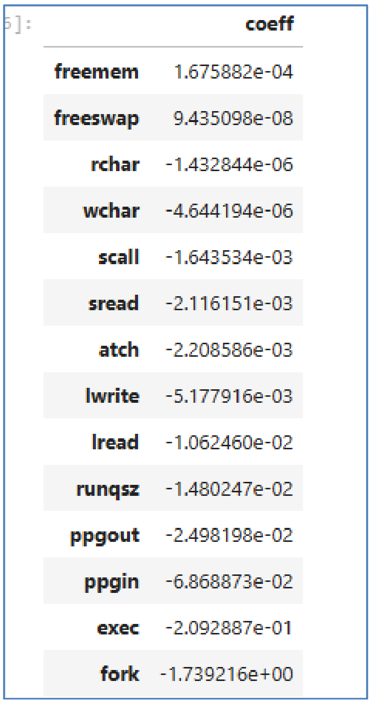


```
lread       1.595751
lwrite      1.654340
scall       6.027593
sread       5.173857
fork        4.991515
exec        3.501204
rchar       2.850180
wchar       2.078564
ppgout      1.762694
atch        1.085418
ppgin       2.049453
runqsz      2.046452
freemem     2.435488
freeswap    5.403199
```

Fig 28 VIF of independent variables after removing pgfree

- We remove scall column and rebuild the model. R- squared value of train data and test for the revised model get down to 74% and 73% respectively which is not reducing the model performance. Hence We will not drop scall, it must have been holding useful information.

- **So the final columns of the linear regression model are 'lread', 'lwrite', 'scall', 'sread', 'fork', 'exec', 'rchar', 'wchar', 'ppgout', 'atch', 'ppgin', 'runqsz', 'freemem', 'freeswap'**



	coeff
freemem	1.675882e-04
freeswap	9.435098e-08
rchar	-1.432844e-06
wchar	-4.644194e-06
scall	-1.643534e-03
sread	-2.116151e-03
atch	-2.208586e-03
lwrite	-5.177916e-03
lread	-1.062460e-02
runqsz	-1.480247e-02
ppgout	-2.498198e-02
ppgin	-6.868873e-02
exec	-2.092887e-01
fork	-1.739216e+00

Fig 29 Important Features with the coefficients

Linear Regression Model after scaling data:

- ✓ We scaled the data using MinMaxScaler. All data ranges between 0 to 1
- ✓ We split the scaled data to train and test set and build the model using skitlearn.
- ✓ We find R-sqaured value and check VIF to ensure that multicollinearity is removed. The steps are same as above model.
- ✓ We build the final model with scaled data and check the performance.

Decision Tree Regression Model

- ✓ Like any other model, we split the whole data set into train and test part and instantiate model With best hyperparameters.
- ✓ Fit the model to train data set and check the model performance.

Performance Comparison of Various Regression models:

Model	R-Squared on Train Data	R-Squared on Test Data	RMSE on Train Data	RMSE on Test Data
Linear Regression using statsmodel	79%		4.26	4.24
Linear Regression using skitlearn	78%	76%	4.4	4.34
Decision Tree Regressor	81%	73%	4.03	4.64
Linear Regression using skitlearn on scaled data	78%	76.7%	0.045(Scaled)	0.044(scaled)

Observations:

- ✓ We see from the above table that Linear Regression model built using statsmodels provides stable R-squared value which is 79%. **This indicates, predictor variables are able to explain 79% of variance In the target variable which is the percentage of time CPU in user mode.**
- ✓ RMSE value of this model for both train and test data is less compared to the other models.

Linear Regression Model Equation:

Usr = 98.3 + -0.012 * (lread) + -0.001 * (scall) + -0.002 * (sread) + -0.292 * (exec) + -6.314e-07 * (rchar) + -5.924e-06 * (wchar) + -0.006 * (ppgout) + 0.0235 * (atch) + -0.053 * (ppgin) + -0.0238 * (vflt) + -0.097 * (runqsz) + 0.0001 * (freemem) + -7.045e-07 * (freeswap)

Key Takeaways:

- ✓ Out of **21** independent variables (excluding the target variable) present in the data set, we observed that there are a few features interrelated to each other. Through various techniques we could reduce the number of features to **13**. Hence the dataset became much simpler to process.

- ✓ We see from the above equation, most of the predictor features are in **negative linear relationship** with The target variable.

- ✓ Here the coefficient values indicate the importance of each feature to predict usr. Number of systems Execution calls **exec** has the highest coefficient value which is **0.292**. **With 1 unit decrease of exec, The percentage of time CPU in user mode increase by 0.292 unit**. The least important feature to predict Usr is freeswap.

- ✓ RMSE (root mean square error) of the final model for train and test data are **4.26 and 4.24** respectively which is quite low. Lower RMSE indicates better prediction of usr, hence better performance of the model.

- ✓ We see from the below plot that except a few cases when actual usr is very low (close to 0), the model Is able to predict usr quite correctly.

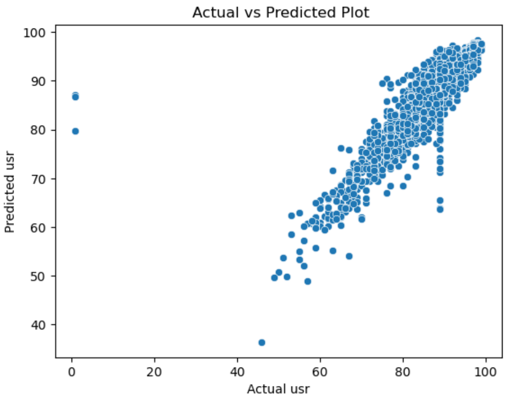


Fig 30 Actual USR vs Predicted USR

_Problem 2

Context

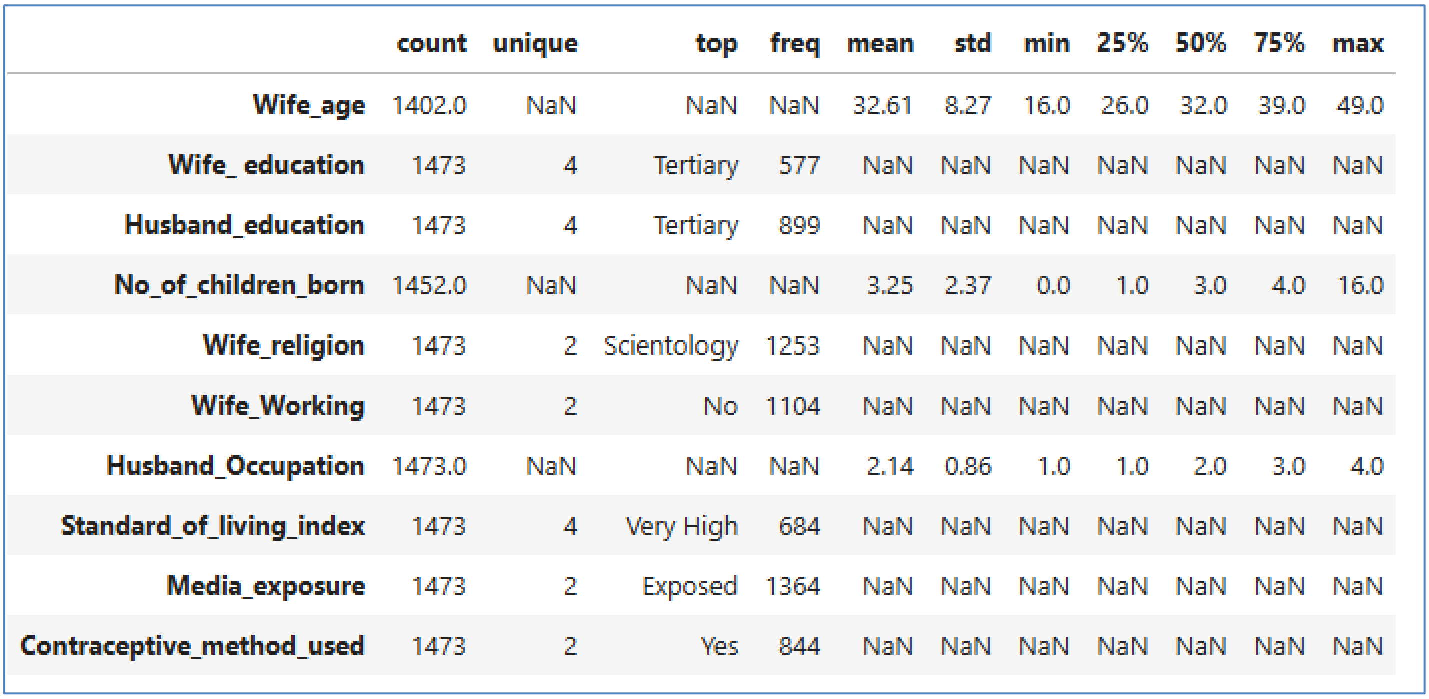
There is dataset from Republic of Indonesia Ministry of Health which encompasses data from 1473 married females who were either not pregnant or were uncertain of their pregnancy status during the survey.

Objective

As an aspiring data scientist, the job is to predict whether these women opt for a contraceptive method of choice.

Data Overview

- **Shape :** There are 1473 records and 10 columns .
- **Data Types:** There are 2 float64, 1 int64 and 7 object datatypes, overall 31numerical and 7 categorical variables present in the dataset.
- **Independent & Target Variable:** There are 9 independent features and **contraceptive_method_used** is the target variable.
- **Check Duplicates:** There are 80 duplicate records in the database.
- **Check Null Values:** There are 71 null values in **Wife_age** and 21 in **No_of_children_born**
- **Statistical Description:**

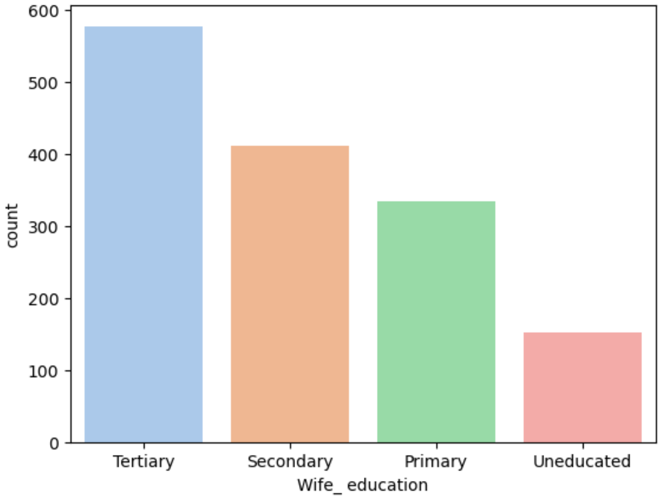


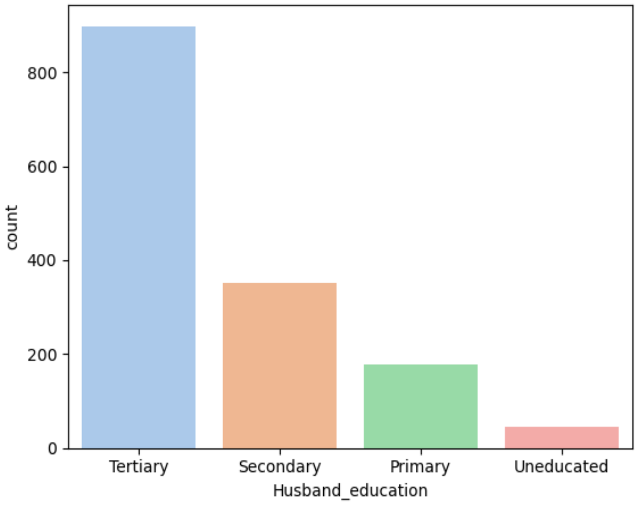
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Wife_age	1402.0	NaN	NaN	NaN	32.61	8.27	16.0	26.0	32.0	39.0	49.0
Wife_ education	1473	4	Tertiary	577	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_education	1473	4	Tertiary	899	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_of_children_born	1452.0	NaN	NaN	NaN	3.25	2.37	0.0	1.0	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	NaN	NaN	NaN	2.14	0.86	1.0	1.0	2.0	3.0	4.0
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

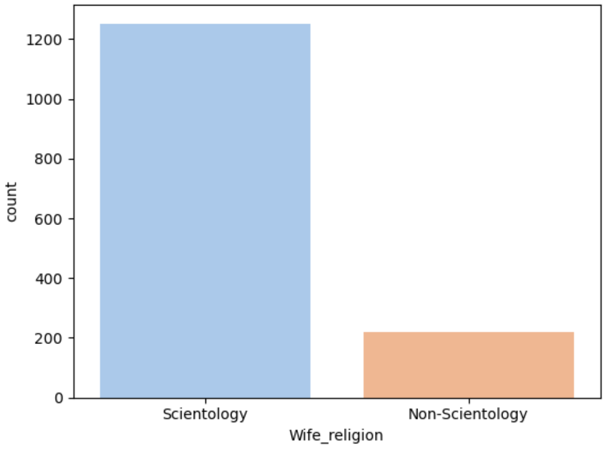
Observations:

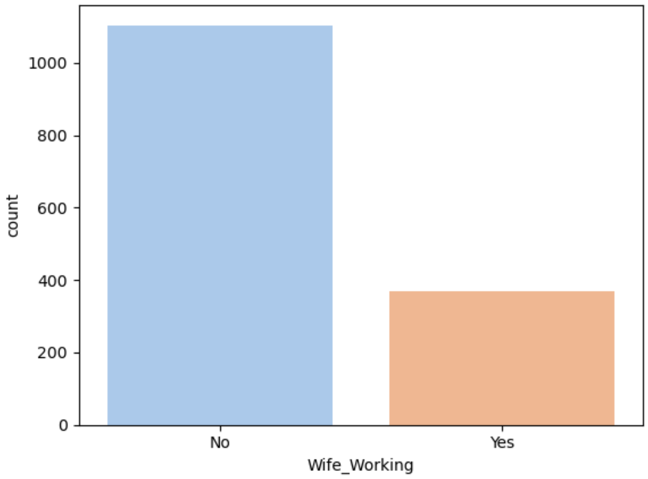
- ✓ Most of the features are categorical in nature.
- ✓ There are women from min age of 16 to max 49. 75% of them are under 39.
- ✓ Around 40% women and 61% of the husbands are coming from tertiary education background.
- ✓ Most of the women believe in Scientology.
- ✓ The standard of living is very high for 46% households.
- ✓ Around 57% women seems to have used contraceptive method.

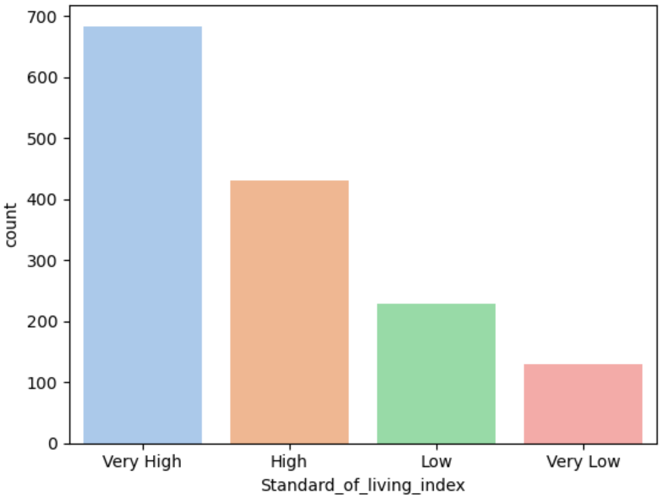
Univariate Analysis:

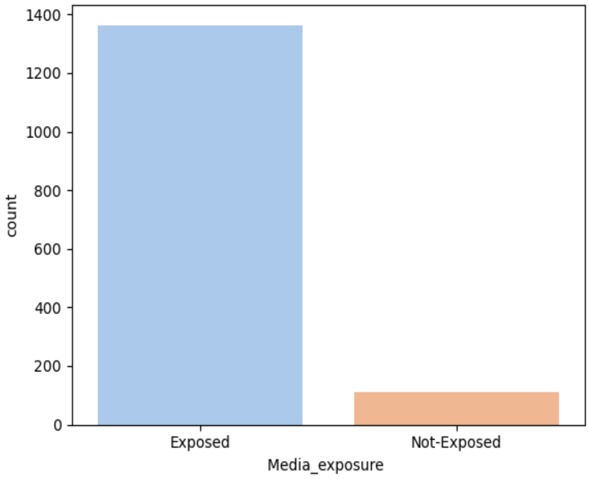












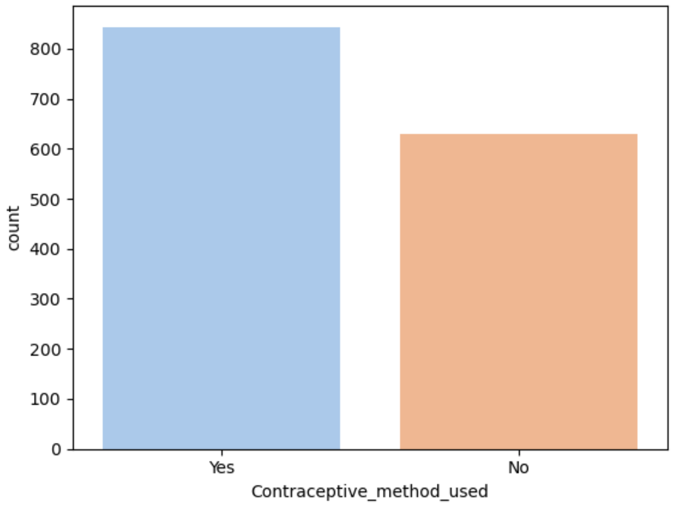


Fig 31 Count Plot of Categorical Variables

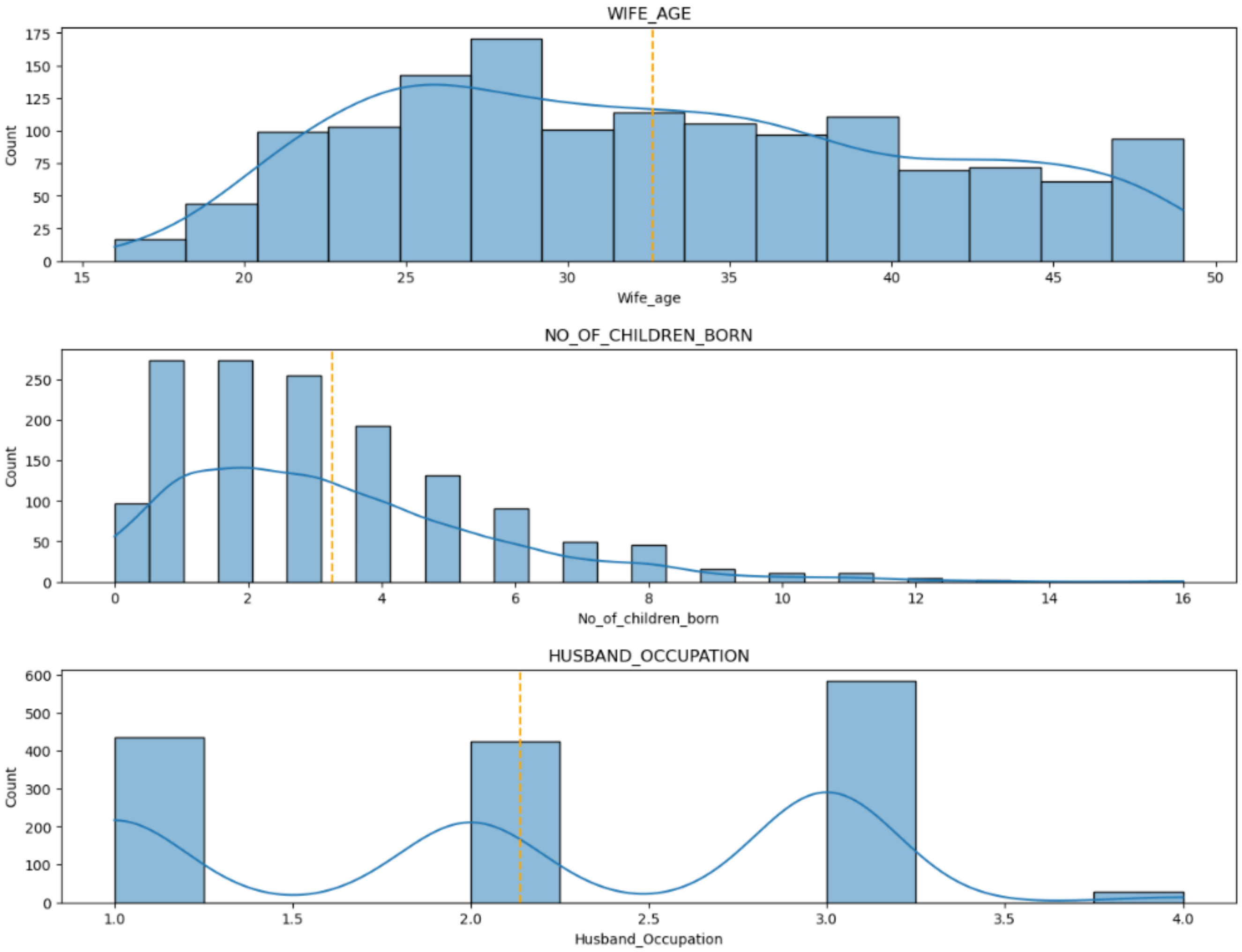


Fig 32 Histplot of Numericl Variables

Observations:

- ✓ No of children born feature is right skewed which indicates the presence of outliers.
- ✓ Wife age somewhat follows normal distribution.
- ✓ Most women are exposed to media.

Bivariate Analysis:

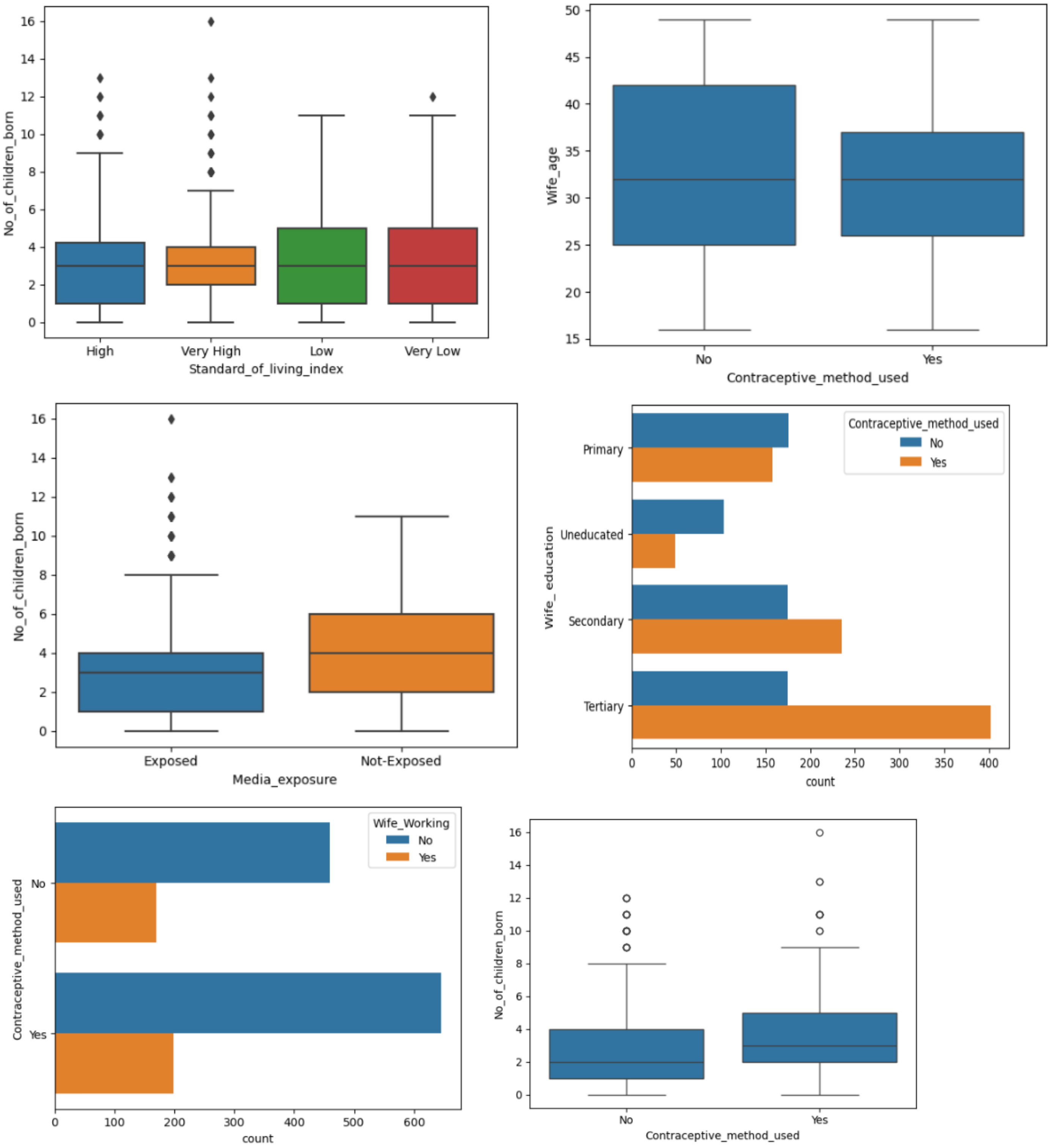


Fig 33 Bivariate Analysis

Observations:

- ✓ The max no of children born in a very high living standard family is below 8, however there are a few outliers. On the other side, in a very low living standard family around 12 children are born at max.
- ✓ The average age of women used contraceptive method or didn’t is same.
- ✓ Average no of children born is higher in the family where women are not exposed to media.
- ✓ The women having education till tertiary naturally are more conscious of using contraceptive methods compared to less educated women.
- ✓ The women working or not doesn’t seem to have any impact on the decision of using contraceptive method or not.

Multivariate Analysis:

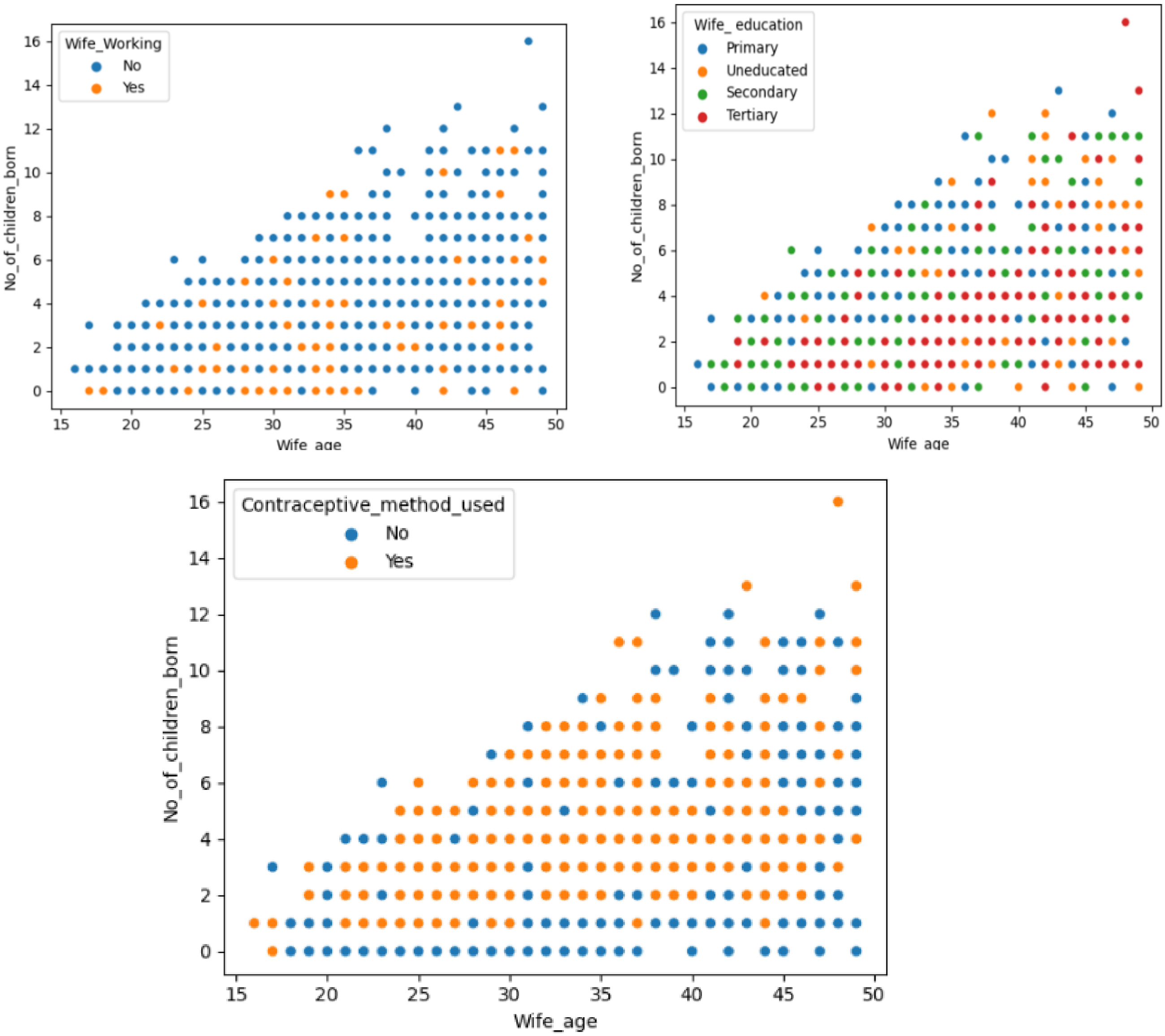
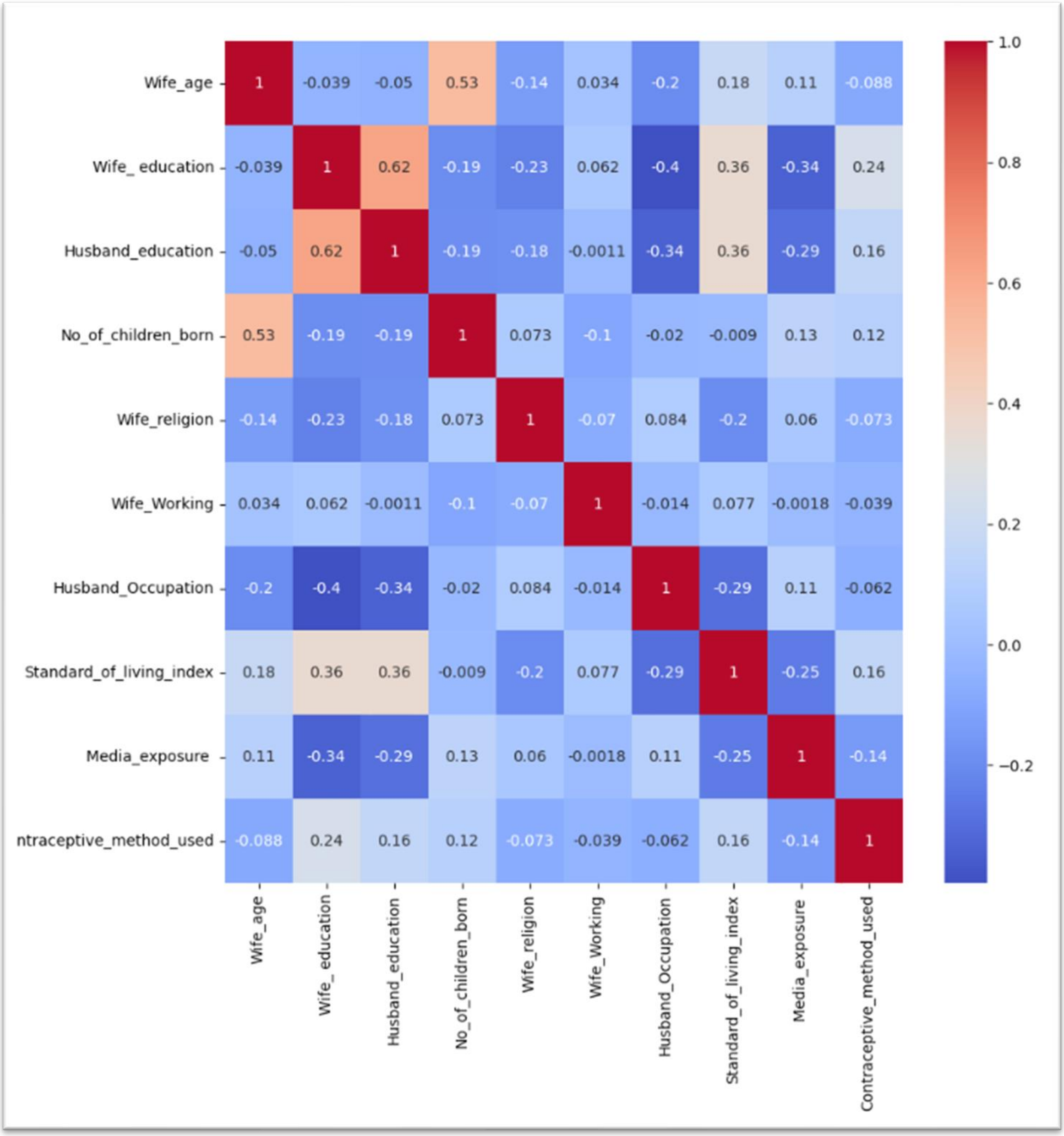
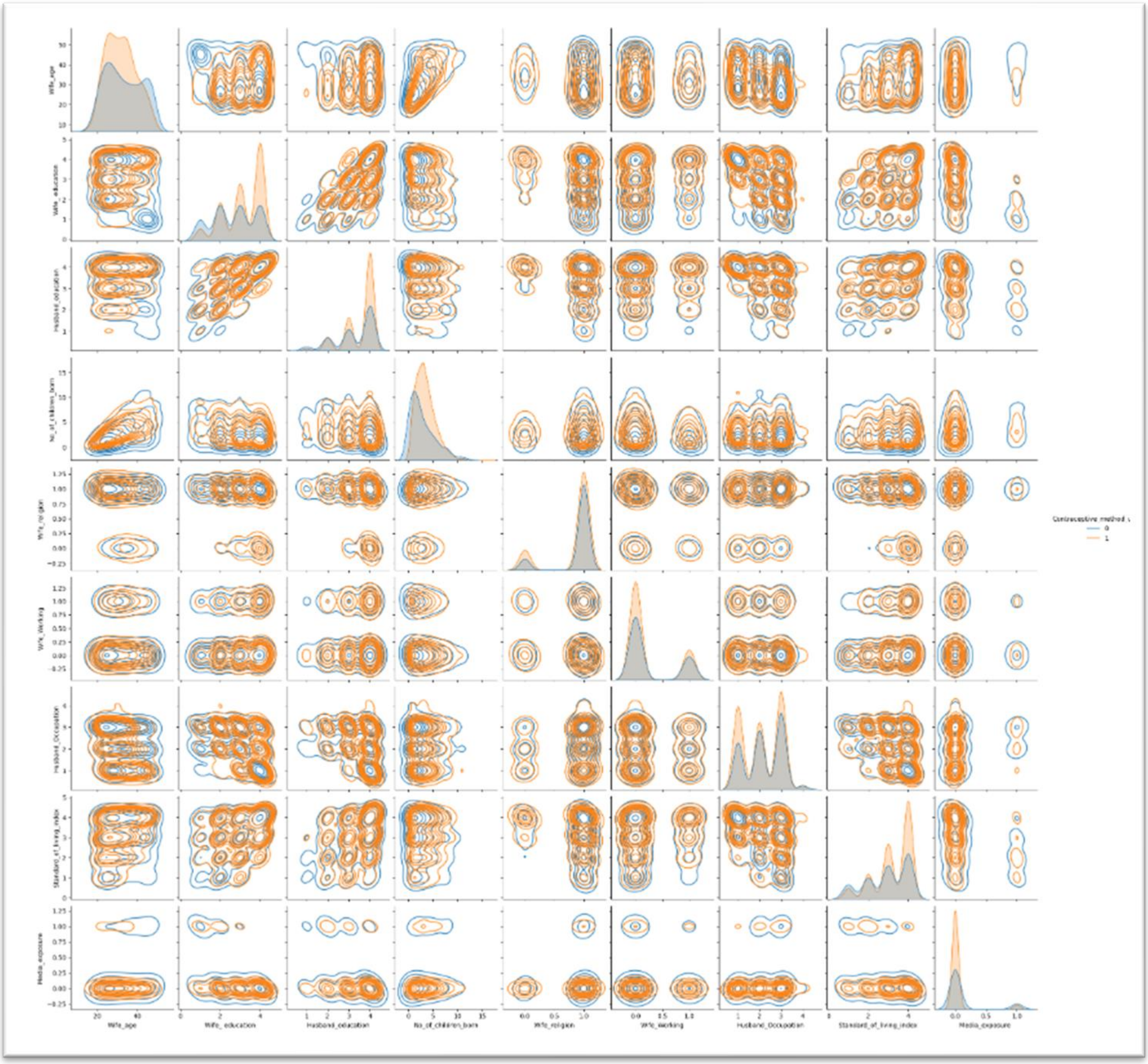


Fig 34 Multivariate Analysis

Heatmap


Wife_age 1 -0.039 -0.05 0.53 -0.14 0.034 -0.2 0.18 0.11 -0.088
Wife_ education -0.039 1 0.62 -0.19 -0.23 0.062 -0.4 0.36 -0.34 0.24
Husband_education -0.05 0.62 1 -0.19 -0.18 -0.0011 -0.34 0.36 -0.29 0.16
No_of_children_born 0.53 -0.19 -0.19 1 0.073 -0.1 -0.02 -0.009 0.13 0.12
Wife_religion -0.14 -0.23 -0.18 0.073 1 -0.07 0.084 -0.2 0.06 -0.073
Wife_Working 0.034 0.062 -0.0011 -0.1 -0.07 1 -0.014 0.077 -0.0018 -0.039
Husband_Occupation -0.2 -0.4 -0.34 -0.02 0.084 -0.014 1 -0.29 0.11 -0.062
Standard_of_living_index 0.18 0.36 0.36 -0.009 -0.2 0.077 -0.29 1 -0.25 0.16
Media_exposure 0.11 -0.34 -0.29 0.13 0.06 -0.0018 0.11 -0.25 1 -0.14
ntraceptive_method_used -0.088 0.24 0.16 0.12 -0.073 -0.039 -0.062 0.16 -0.14 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
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Pairplot:



Observations:

- ✓ We see from the pair plot that wife religion and wife working or not don't have much impact on the decision making whether the women used contraceptive method or not.
- ✓ There is 62% collinearity between wife education and husband education which is not strong enough to remove either of two.
- ✓ There is 53% collinearity between wife age and no of children born.

Data Pre-processing:

Missing Value Treatment:

There are missing values in Wife age and No of children born column. We impute the age with mean and the no of children born with median value because of presence of the outliers. Now there are no more missing Values.

Outlier Treatment:

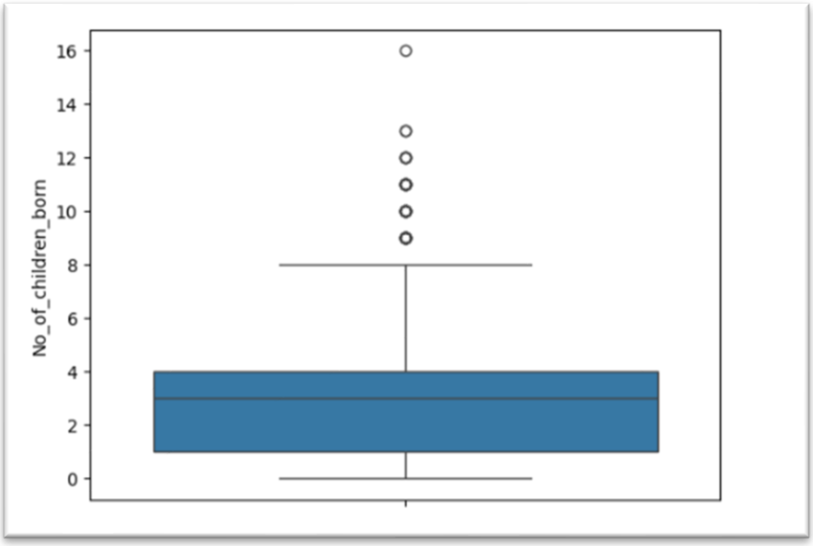


Fig 34 Check Outliers

As we see in the above plot, there are only a very few outliers. We can keep it in the dataset.

Encoding**:**

- ✓ There are 2 types of categorical variables present in the dataset, ordinal variables – wife education, Husband education and standard of living index. Another nominal variable – Wife religion, wife media exposure, wife working, and contraceptive method used or not.
- ✓ Ordinal variables are encoded with proper ordering from 0 to 4 lowest to highest.
- ✓ Nominal variables are encoded using label encoding technique.
- ✓ Encoding changes the data type of categorical variables to numeric one.

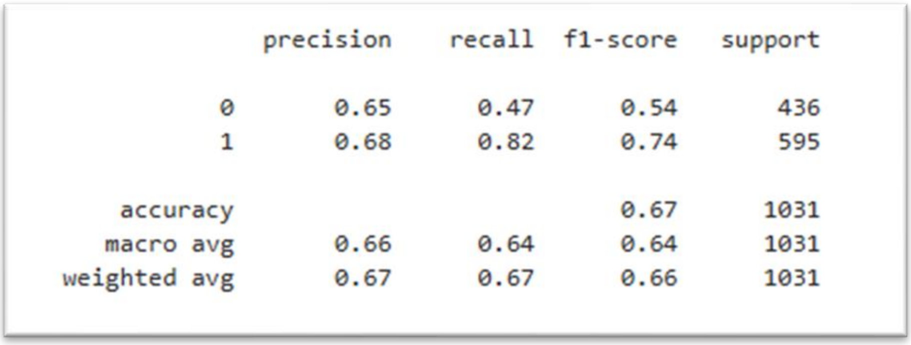
Split Train and Test Data:

- ✓ We split the whole data set into 2 data frames X & Y. X contains all the independent variables and Y contains the target variable.
- ✓ We further split them into train and test data set with train size of 70% which means 70% data is being Used for model training and 30% for testing.
- ✓ There are 1031 records in the train data set and 442 records in the test.

Model Building:

Logistic Regression Model:

- We instantiate the model and fit it to the train data set.
- The model score on test data is **68.7%** and train data **66.9%**
- Classification report of the model on train data as follows



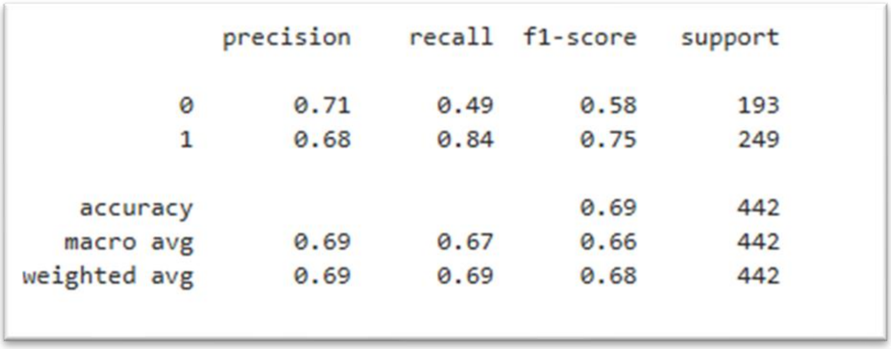
```
              precision    recall  f1-score   support

           0       0.65      0.47      0.54       436
           1       0.68      0.82      0.74       595

    accuracy                           0.67      1031
   macro avg       0.66      0.64      0.64      1031
weighted avg       0.67      0.67      0.66      1031
```

Fig 35

- Classification report of the model on test data as follows:



```
              precision    recall  f1-score   support

           0       0.71      0.49      0.58       193
           1       0.68      0.84      0.75       249

    accuracy                           0.69       442
   macro avg       0.69      0.67      0.66       442
weighted avg       0.69      0.69      0.68       442
```

Fig 36

- As per test data, Precision of class 1 is 68% means that out of total contraceptive method used case prediction, only 68% times the prediction is correct, other times the model is predicting as contraceptive method not used.
- Out of total contraceptive method not used case prediction, 71% times the prediction is correct. When the model predicts contraceptive method not used case as contraceptive method used one (incorrect prediction – False Positive) , the model fails to decide if there is any chance of pregnancy or not.
- Recall of class 1 is 84% means that 84% times the model is predicting a woman has used Contraceptive method out of total Contraceptive method used cases.
- Recall of class 0 is 49% means that 49% times the model is predicting a woman has not used Contraceptive method out of total Contraceptive method not used cases.
- Confusion matrix is as follows:

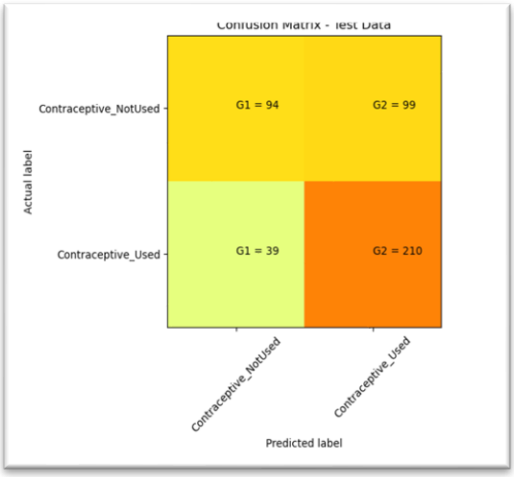
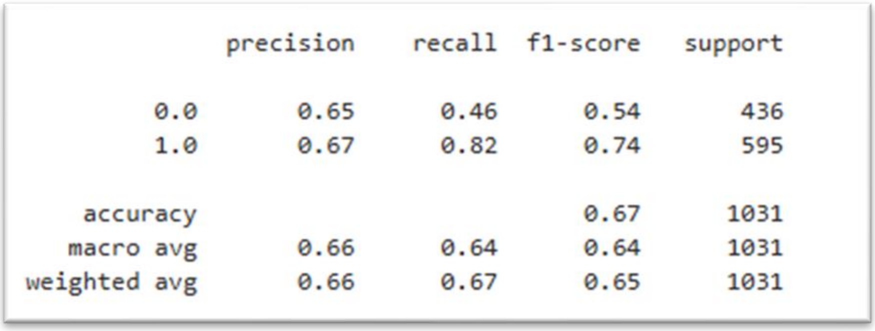


Fig 37

- ✓ As we see from the above matrix out of actual 193 contraceptive method not used cases, the model Could predict 94 correctly.
- ✓ Out of actual 249 contraceptive method used cases; the model predicted 210 records correctly.

Linear Discriminant Analysis:

- ✓ Data needs to be scaled before fitting it to LDA model. Here we perform MinMaxScaling technique as most of the column values range between 0 to 1 after encoding.
- ✓ Once the data is scaled and split, we instantiate the model and fit the train data to the model.
- ✓ Model score on the train data is **66.6%** and test data is **68.5%**
- ✓ Classification report of the model on train data as follows



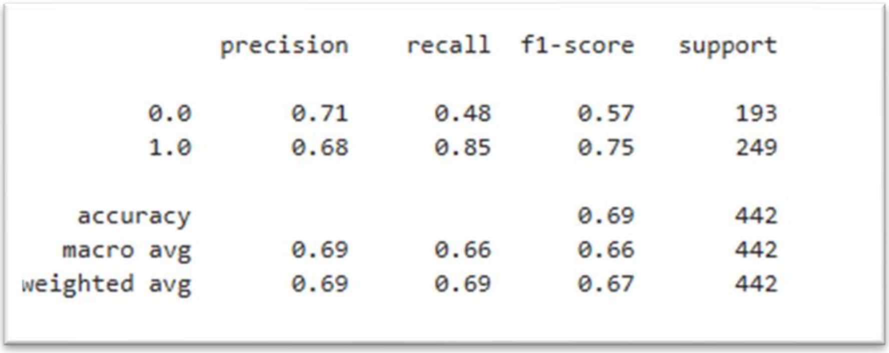
```
              precision    recall  f1-score   support

         0.0       0.65      0.46      0.54       436
         1.0       0.67      0.82      0.74       595

    accuracy                           0.67      1031
   macro avg       0.66      0.64      0.64      1031
weighted avg       0.66      0.67      0.65      1031
```

Fig 38

- ✓ Classification report of the model on test data as follows:



```
              precision    recall  f1-score   support

         0.0       0.71      0.48      0.57       193
         1.0       0.68      0.85      0.75       249

    accuracy                           0.69       442
   macro avg       0.69      0.66      0.66       442
weighted avg       0.69      0.69      0.67       442
```

Fig 39

- ✓ As per test data, Precision of class 1 is 68% means that out of total contraceptive method used case prediction, only 68% times the prediction is correct, other times the model is predicting as contraceptive method not used.
- ✓ Out of total contraceptive method not used case prediction, 71% times the prediction is correct. When the model predicts contraceptive method not used case as contraceptive method used one (incorrect prediction – False Positive) , the model fails to decide if there is any chance of pregnancy or not.
- ✓ Recall of class 1 is 85% means that 85% times the model is predicting a woman has used Contraceptive method out of total Contraceptive method used cases.
- ✓ Recall of class 0 is 48% means that 48% times the model is predicting a woman has not used Contraceptive method out of total Contraceptive method not used cases.
- ✓ Confusion matrix on the test data is as follows:

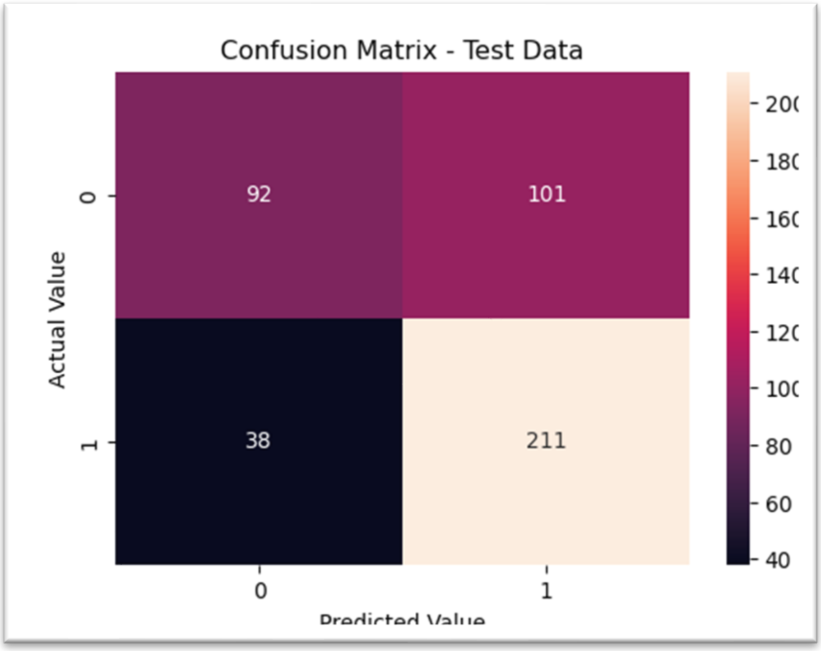
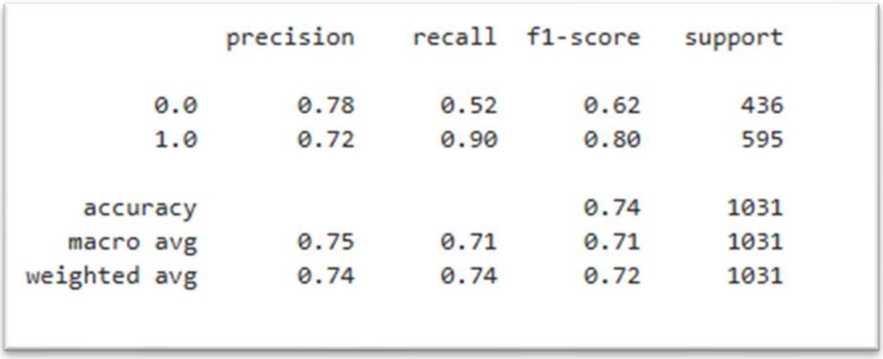


Fig 40

- ✓ As we see from the above matrix out of actual 193 contraceptive method not used cases, the model Could predict 92 correctly.
- ✓ Out of actual 249 contraceptive method used cases; the model predicted 211 records correctly.

Decision Tree Classifier:

- ✓ We instantiate the model with Gini criteria and fit the train data to it.
- ✓ The model score on train data is **98.3%** and test data is **67%** which indicates overfitting of the model.
- ✓ We do prune the model by tuning the hyperparameters to deal with overfitting. As we find the best Parameters we build the CART model again and find out the scores.
- ✓ The regularized model score on the train data is **73.5%** and test data **71.7%.**
- ✓ Classification report on train data as follows:



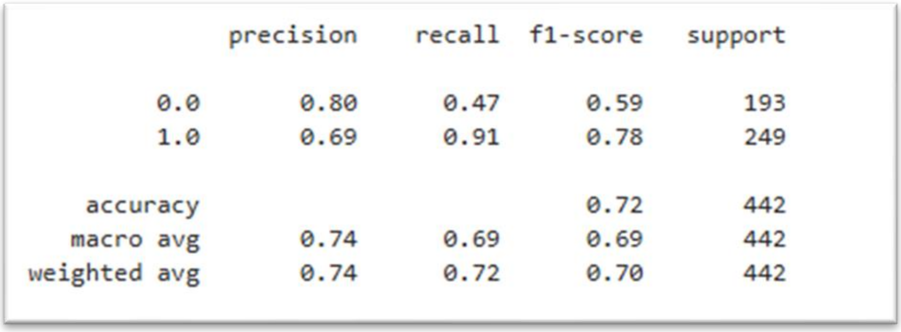
```
              precision    recall  f1-score   support

         0.0       0.78      0.52      0.62       436
         1.0       0.72      0.90      0.80       595

    accuracy                           0.74      1031
   macro avg       0.75      0.71      0.71      1031
weighted avg       0.74      0.74      0.72      1031
```

Fig 41

✓ Classification report on test data as follows:



```
              precision    recall  f1-score   support

         0.0       0.80      0.47      0.59       193
         1.0       0.69      0.91      0.78       249

    accuracy                           0.72       442
   macro avg       0.74      0.69      0.69       442
weighted avg       0.74      0.72      0.70       442
```

Fig 42

✓ As per test data, Precision of class 1 is 69% means that out of total contraceptive method used case prediction, only 69% times the prediction is correct, other times the model is predicting as contraceptive method not used.

✓ Out of total contraceptive method not used case prediction, 80% times the prediction is correct. When the model predicts contraceptive method not used case as contraceptive method used one (incorrect prediction – False Positive) , the model fails to decide if there is any chance of pregnancy or not.

✓ Recall of class 1 is 91% means that 91% times the model is predicting a woman has used Contraceptive method out of total Contraceptive method used cases.

✓ Recall of class 0 is 47% means that 47% times the model is predicting a woman has not used Contraceptive method out of total Contraceptive method not used cases.

✓ Confusion matrix on the test data is as follows:

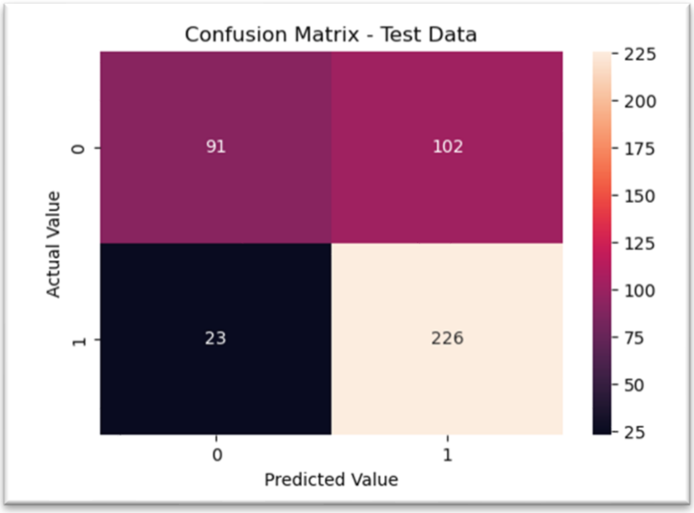


Fig 43

- ✓ AUC value on train data is 0.79 and test data 0.754. Higher the value of AUC, better is model performance.

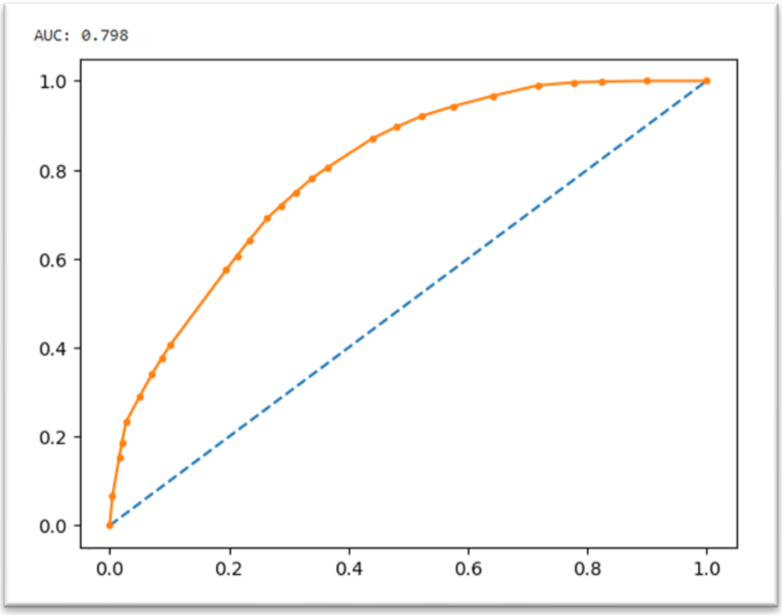


Fig 44 AUC on Train Data

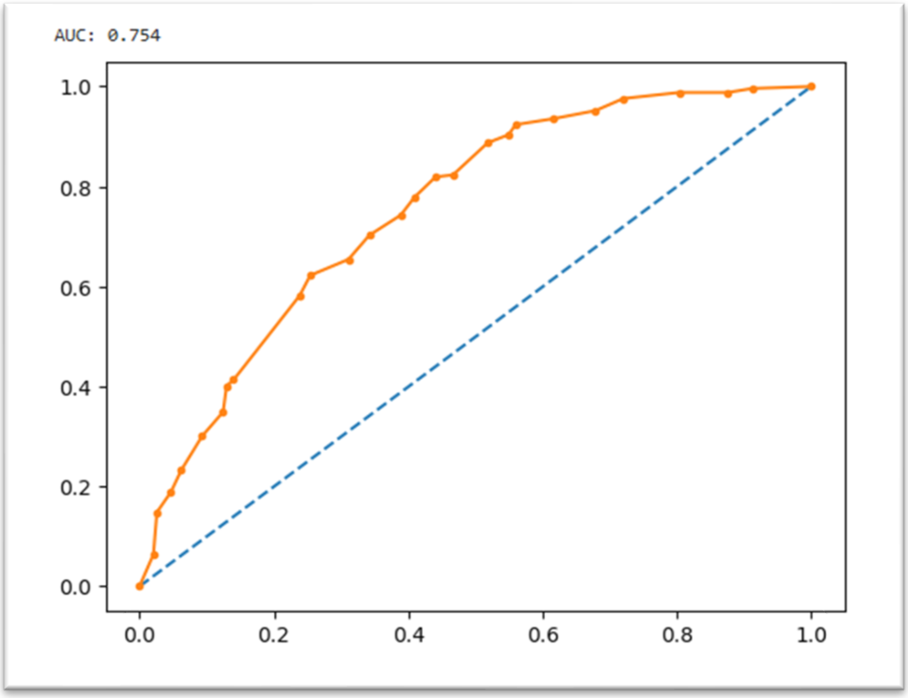


Fig 45 AUC on Test Data

Model Performance Comparison:

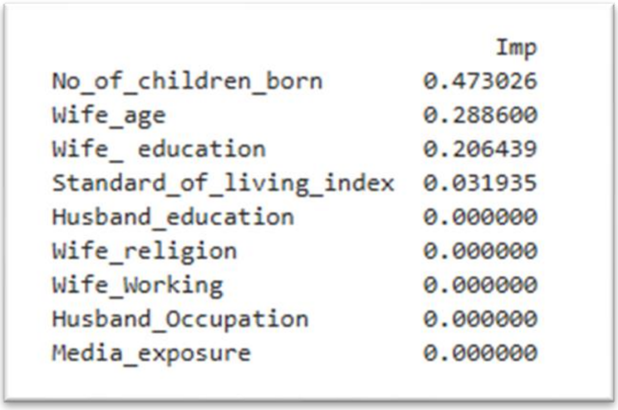
Model	Score on Train data (%)	Score on Test data(%)	Precision of Class 1 on Test Data(%)	Recall of Class 1 on Test Data (%)	Precision of Class 0 on Test Data(%)	Recall of Class 0 on Test Data(%)
Logistic Regression	68.7	66.9	68	84	71	49
LDA	66.6	68	67	85	71	48
CART	73.5	71.5	69	91	80	47

Observation:

- As we observe from the table that score of CART model on train and test is comparatively better and stable.
- Here we need to check Recall of class 0 (contraceptive method not used) to understand the model performance. If a woman didn't use contraceptive method, being pregnant is a possibility. But if the model predicts it incorrectly, then woman will not be able to take appropriate course of action.
- We see Recall of class 0 for all three model is not going to atleast 50% . However, Recall of Class 1 is Significantly good which means the models are quite good predicting a women who has used contraceptive method.

Check Feature Importance:

Feature importance of CART model :



```
                                  Imp
No_of_children_born          0.473026
Wife_age                     0.288600
Wife_ education              0.206439
Standard_of_living_index     0.031935
Husband_education            0.000000
Wife_religion                0.000000
Wife_Working                 0.000000
Husband_Occupation           0.000000
Media_exposure               0.000000
```

As we see from the above fig, No of children born feature mostly contributing to the prediction of contraceptive method used or not, then wife age, wife education and standard of living index follows.

Key Takeaways :

CART model is the best model to predict if a woman used contraceptive method or not compared to Logistic Regression and LDA. However, overall performance is not quite good to avoid the risk of identifying someone having possibility of being pregnant as non-pregnant.