Predictive Modeling

Coded Project

Anirudh Sardiwal

31 March 2024

Table of Contents

1.0 Problem 1 - Comp-activ

1.1 Basics

lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap	usr
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
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
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
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
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

Fig. 1.1 - Partial Data Head

The data has 8192 rows and 22 columns. Only ‘runqsz’ column is object data type.

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgfree	pgscan	atch	pgin	ppgin	pflt
count	8192.00	8192.00	8192.00	8192.00	8192.00	8192.00	8192.00	8088.00	8177.00	8192.00	...	8192.00	8192.00	8192.00	8192.00	8192.00	8192.00
mean	19.56	13.11	2306.32	210.48	150.06	1.88	2.79	197385.73	95902.99	2.29	...	11.92	21.53	1.13	8.28	12.39	109.79
std	53.35	29.89	1633.62	198.98	160.48	2.48	5.21	239837.49	140841.71	5.31	...	32.36	71.14	5.71	13.87	22.28	114.42
min	0.00	0.00	109.00	6.00	7.00	0.00	0.00	278.00	1498.00	0.00	...	0.00	0.00	0.00	0.00	0.00	0.00
25%	2.00	0.00	1012.00	86.00	63.00	0.40	0.20	34091.50	22916.00	0.00	...	0.00	0.00	0.00	0.60	0.60	25.00
50%	7.00	1.00	2051.50	166.00	117.00	0.80	1.20	125473.50	46619.00	0.00	...	0.00	0.00	0.00	2.80	3.80	63.80
75%	20.00	10.00	3317.25	279.00	185.00	2.20	2.80	267828.75	106101.00	2.40	...	5.00	0.00	0.60	9.76	13.80	159.60
max	1845.00	575.00	12493.00	5318.00	5456.00	20.12	59.56	2526649.00	1801623.00	81.44	...	523.00	1237.00	211.58	141.20	292.61	899.80

Fig. 1.2 - 5 point summaries of variables

1.2 Univariate Analysis

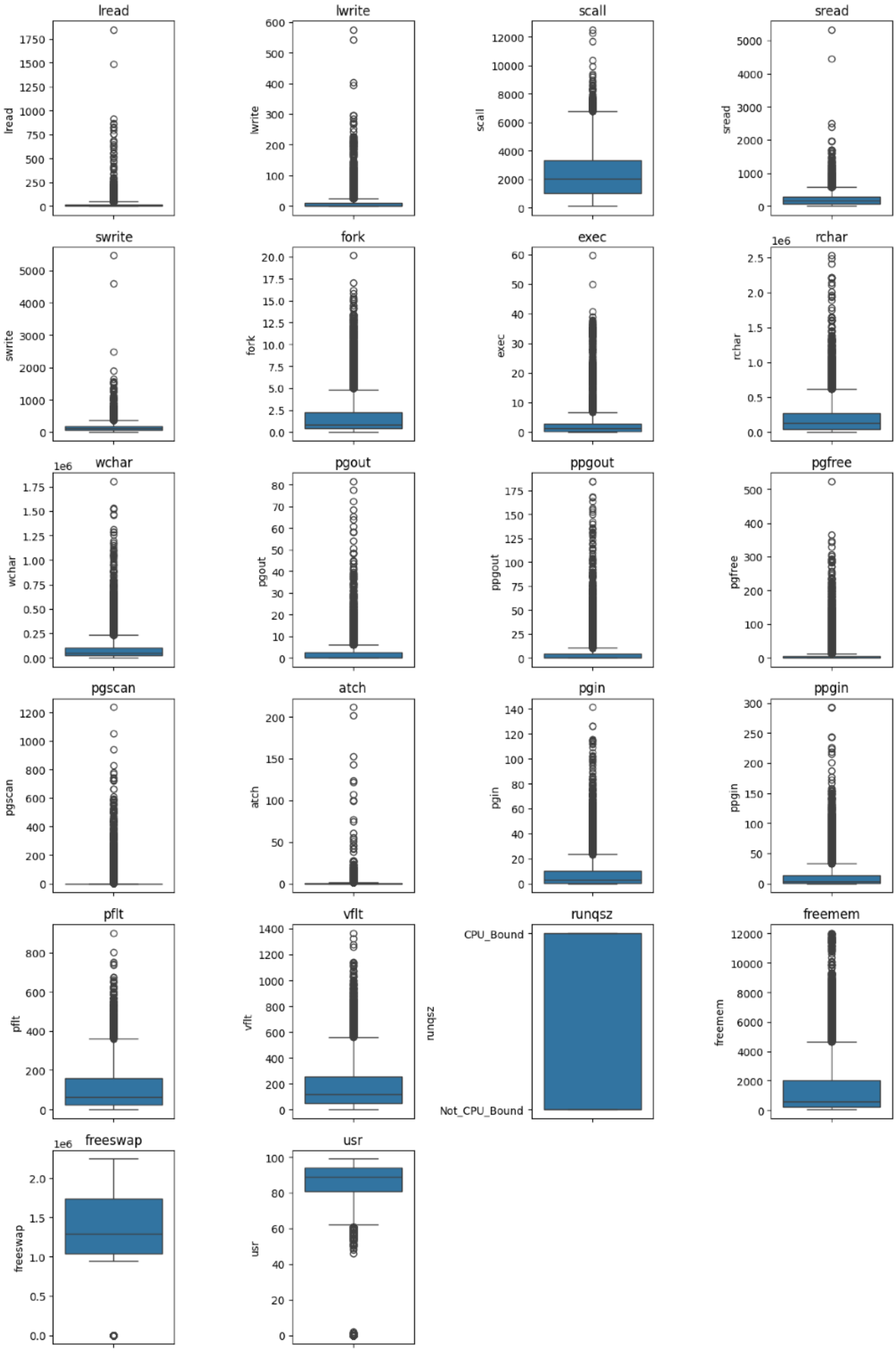


Fig 1.3 - Boxplots of all features

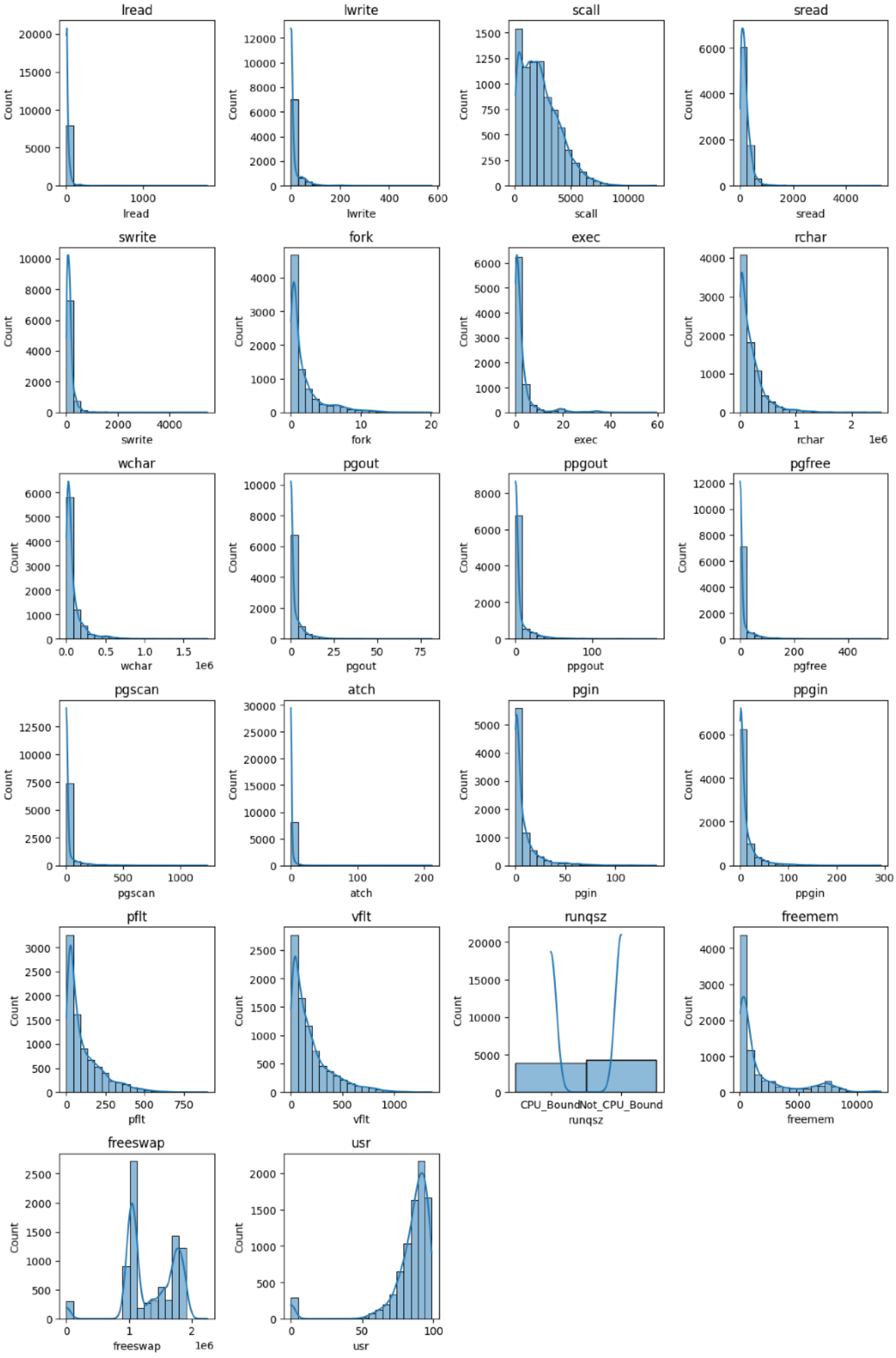


Fig 1.4 - Histogram plots of all features

Features ‘rchar’ and ‘wchar’ contained null values, which were replaced by their respective medians.

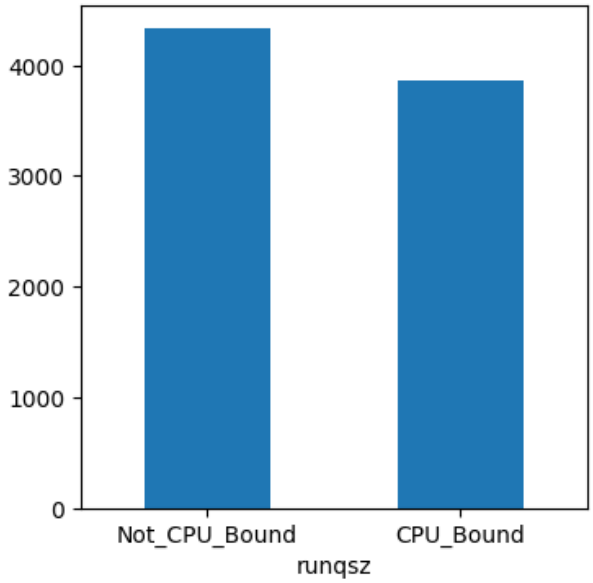


Fig 1.5 - Value counts of feature ‘runqsz’

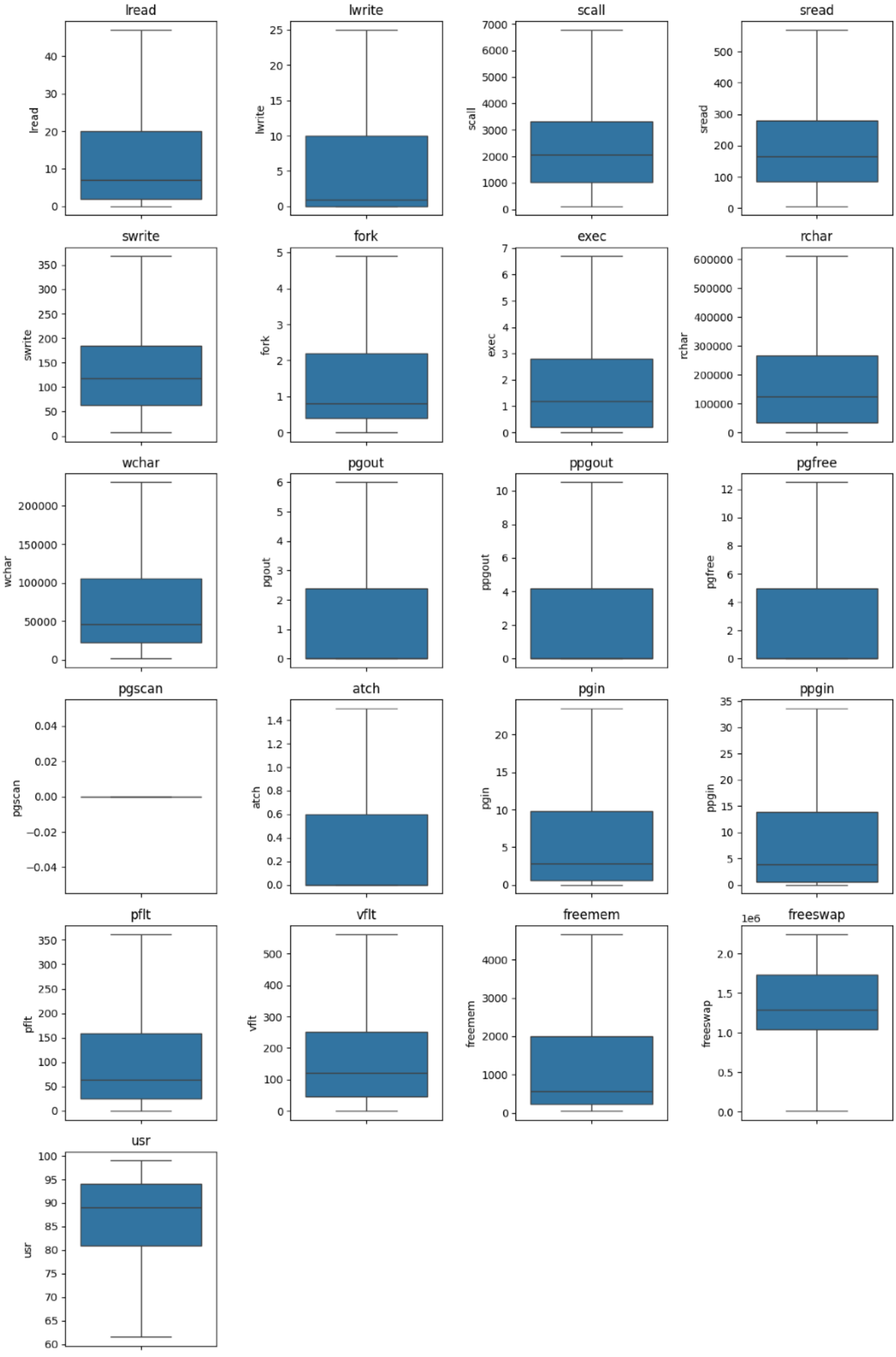


Fig 1.6 - Feature boxplots after outlier treatment

1.3 Bivariate Analysis

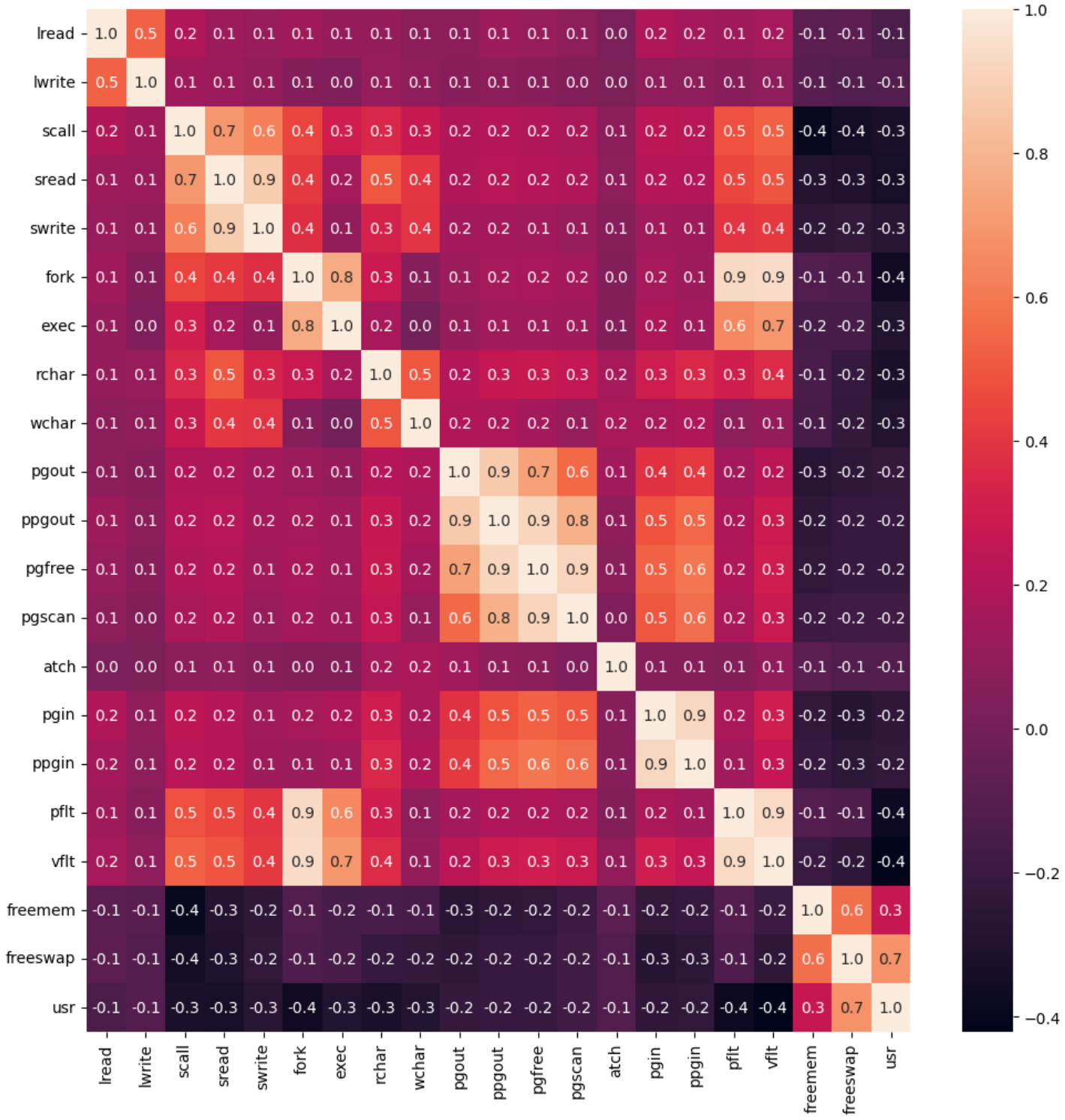


Fig 1.7 - Heatmap of correlation matrix of all features

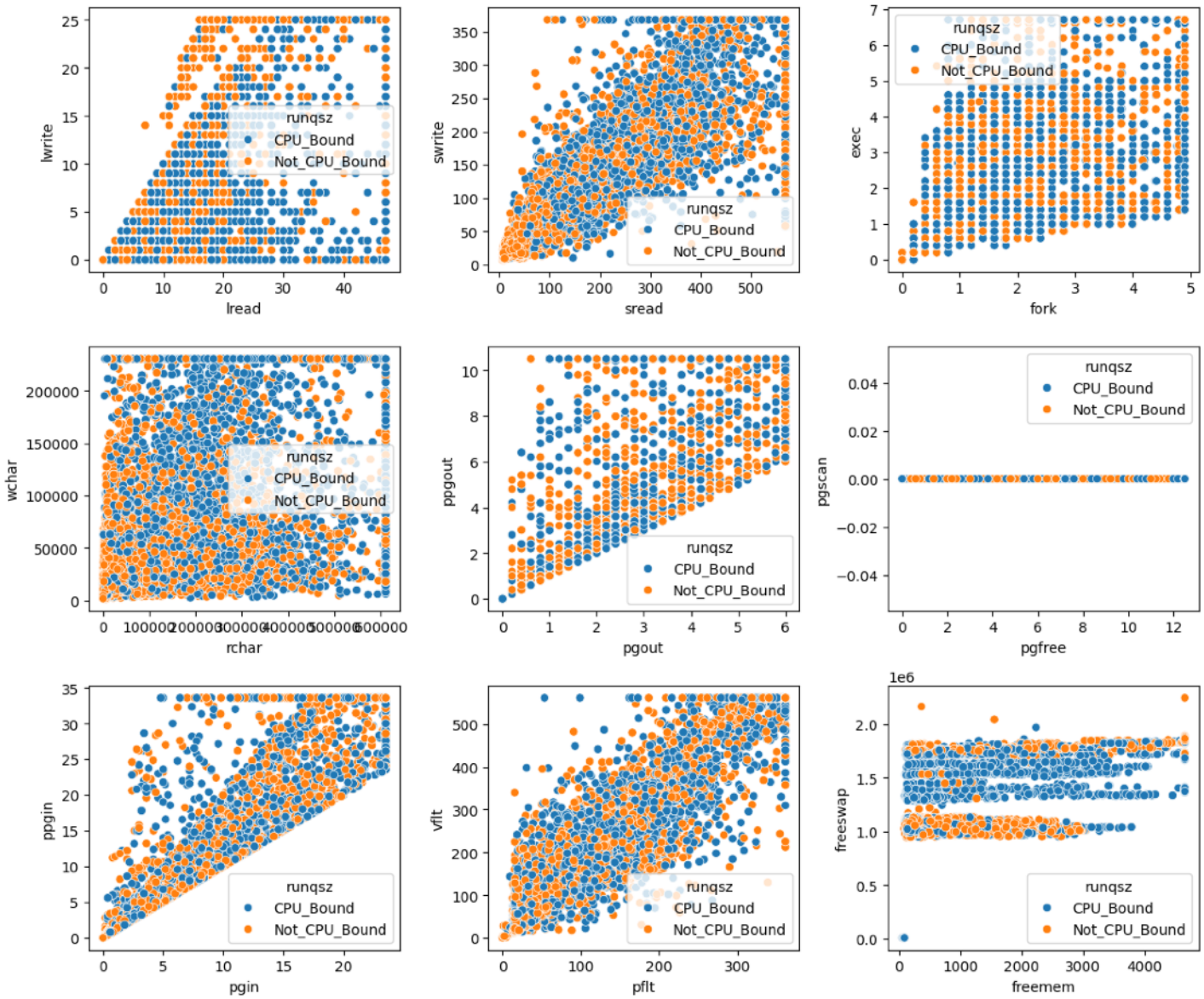


Fig 1.8 - Scatterplots of various features

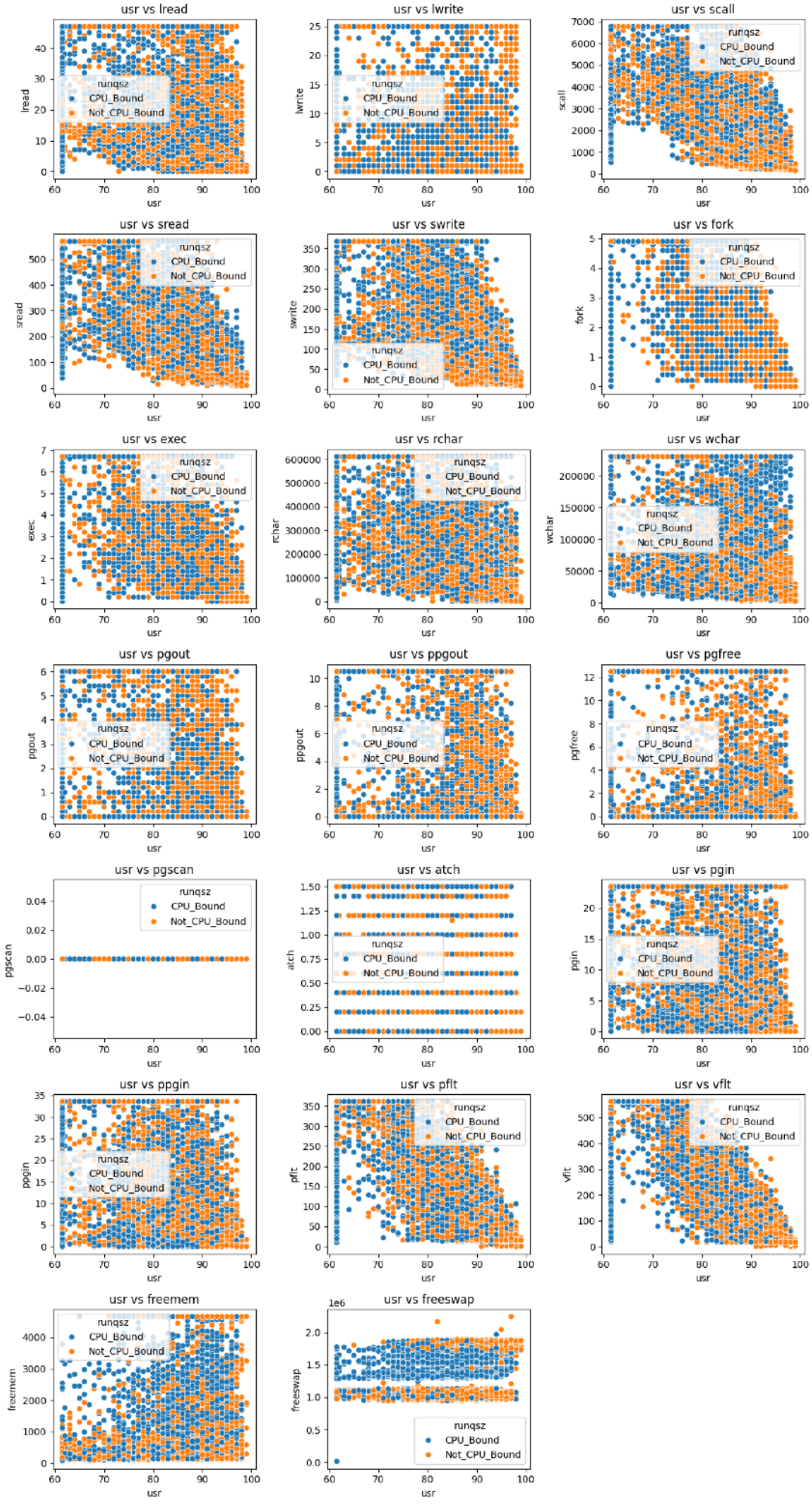


Fig 1.9 - Relation of target variable to various features

1.4 Fit Linear Model

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.796
Model:	OLS	Adj. R-squared:	0.795
Method:	Least Squares	F-statistic:	1115.
Date:	Wed, 27 Mar 2024	Prob (F-statistic):	0.00
Time:	11:47:27	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	-3.003e-14	1.41e-16	-212.676	0.000	-3.03e-14	-2.98e-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.	4.18e+22

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 6.53e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Fig 1.9 - Initial model

1.4.1 Check Multicollinearity

```
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
```

Fig 1.10 - VIF values of initial model

```
VIF values:

const                    27.153614
lwrite                    1.053077
scall                     1.763109
exec                      2.507007
wchar                     1.211316
pgfree                    1.953643
atch                      1.716507
ppgin                     1.480245
vflt                      3.010137
freemem                   1.917788
freeswap                  1.731634
runqsz_Not_CPU_Bound      1.124738
dtype: float64
```

Fig 1.11 - VIF values of final model

1.4.2 Test for Linearity

	Actual Values	Fitted Values	Residuals
0	91.0	89.609360	1.390640
1	94.0	91.614706	2.385294
2	61.5	78.059951	-16.559951
3	83.0	79.634490	3.365510
4	94.0	97.162035	-3.162035

Fig 1.12 - Head of Fitted and Residual values data

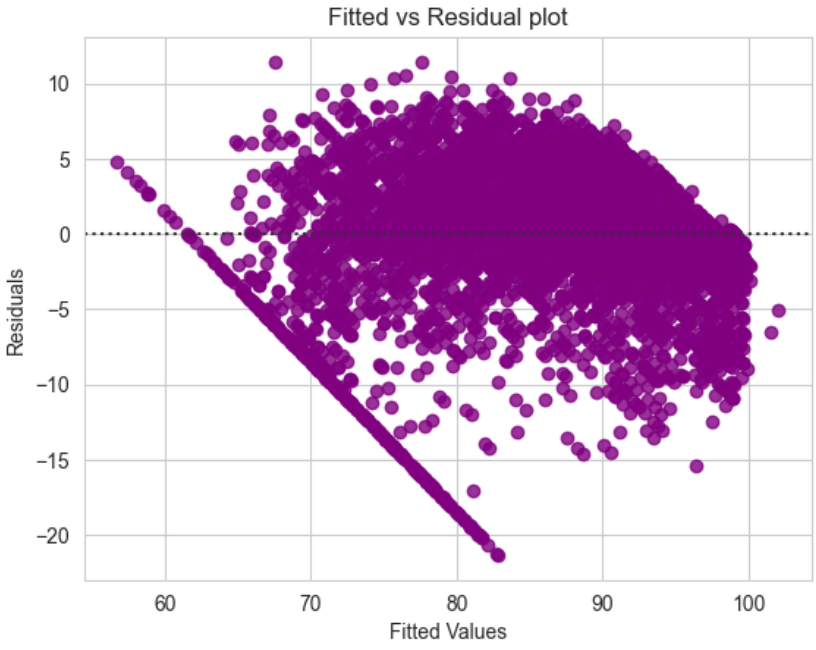
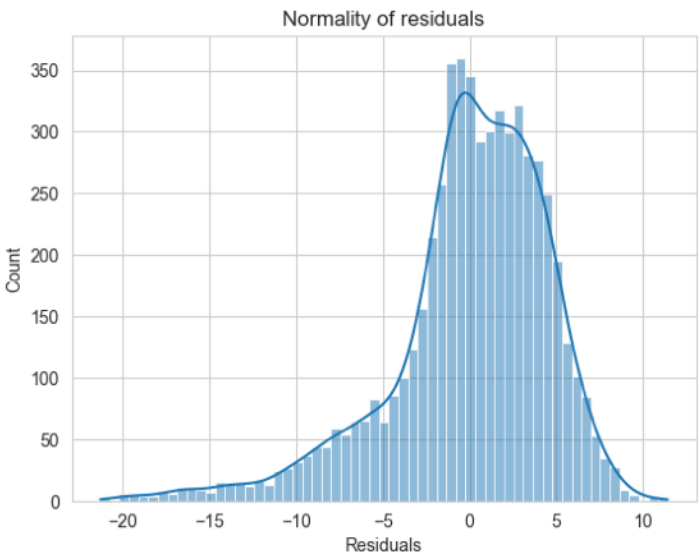
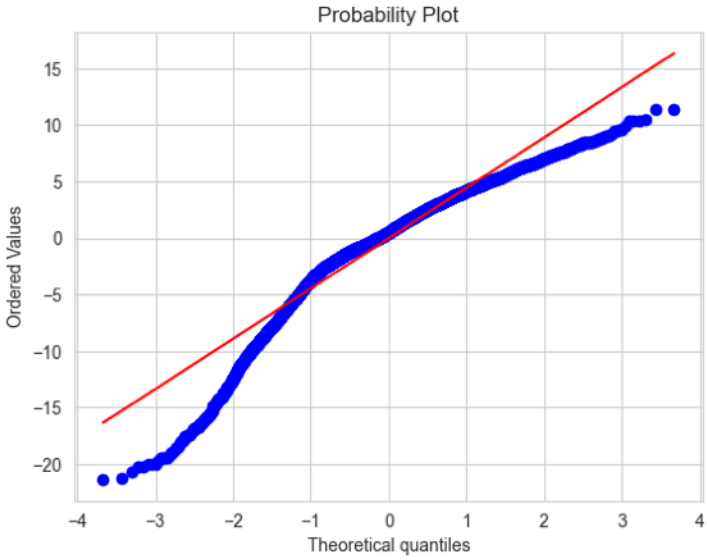


Fig 1.13 - Plot of fitted vs residual values

1.4.3 Test for Normality





Residuals are not Normal.

1.4.4 Test for Homoscedasticity

P-value 0.0015

Residuals are not Homoscedastic

1.5 Final Model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.774
Model:                            OLS   Adj. R-squared:                  0.773
Method:                 Least Squares   F-statistic:                     1777.
Date:                Fri, 29 Mar 2024   Prob (F-statistic):               0.00
Time:                        11:51:07   Log-Likelihood:                -16958.
No. Observations:                5734   AIC:                         3.394e+04
Df Residuals:                    5722   BIC:                         3.402e+04
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const                     84.2204      0.321    262.470      0.000      83.591      84.849
lwrite                    -0.0275      0.007     -4.028      0.000      -0.041      -0.014
scall                     -0.0009   5.11e-05    -17.249      0.000      -0.001      -0.001
exec                      -0.4397      0.048     -9.194      0.000      -0.533      -0.346
wchar                  -1.209e-05   9.51e-07    -12.714      0.000   -1.4e-05   -1.02e-05
pgfree                    -0.0864      0.017     -4.992      0.000      -0.120      -0.052
atch                       0.3880      0.144      2.697      0.007       0.106       0.670
ppgin                     -0.0427      0.007     -6.317      0.000      -0.056      -0.029
vflt                      -0.0288      0.001    -44.142      0.000      -0.030      -0.028
freemem                   -0.0005   5.28e-05     -9.062      0.000      -0.001      -0.000
freeswap               8.47e-06   1.94e-07     43.639      0.000    8.09e-06    8.85e-06
runqsz_Not_CPU_Bound      1.8228      0.131     13.931      0.000       1.566       2.079
==============================================================================
Omnibus:                     1104.165   Durbin-Watson:                   2.019
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2236.070
Skew:                          -1.147   Prob(JB):                         0.00
Kurtosis:                       5.024   Cond. No.                     7.40e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.4e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

1.5.1 Observations

- The model is able to explain 77% variance in the data.
- A unit increase in ‘runqsz - Not CPU bound’ will result in 1.82 units increase in ‘usr’, all else remaining constant.
- ‘runqsz - Not CPU bound’, ‘exec’, and ‘atch’ will have maximum impact on ‘usr’ if changed.

1.5.2 Equation

```
usr = 84.22 + -0.027 * ( lwrite ) +  -0.001 * ( scall ) +  -0.44 * ( exec ) +  -0.0 * ( wchar ) +  -0.086 * ( pgfree ) +  0.388 * ( atch ) +
-0.043 * ( ppgin ) +  -0.029 * ( vflt ) +  -0.0 * ( freemem ) +  0.0 * ( freeswap ) +  1.823 * ( runqsz_Not_CPU_Bound )
```

1.6 Errors

RMSE Train Data - 4.607

RMSE Test Data - 4.852

MAE Train Data - 3.415

MAE Test Data - 3.572

- RMSE on the Train and Test data are comparable. This means our model is not suffering from overfitting.
- MAE indicates that our model is able to predict ‘usr’ with a mean error of 3.57 units on Test data, which is acceptable.
- Therefore the model is moderately good for prediction and inference.

1.7 Actionable Insights and Recommendations

- To optimise the portion of time that cpus run in user mode (usr), it would help to focus on the Process run queue size (runqsz), Number of system exec calls per second (exec), and Number of page attaches per second (atch).

2.0 Problem 2 - Indonesia MOH Survey

2.1 Basics

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

Fig 2.1 - Data Head

```
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)
```

Fig 2.2 - Data types of various features

	Wife_age	No_of_children_born
count	1402.00	1452.00
mean	32.61	3.25
std	8.27	2.37
min	16.00	0.00
25%	26.00	1.00
50%	32.00	3.00
75%	39.00	4.00
max	49.00	16.00

Fig 2.3 - 5 point summaries of numerical columns

2.2 Univariate Analysis

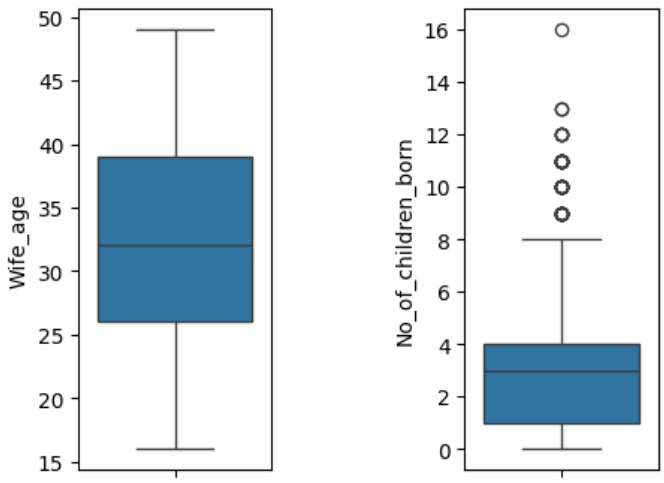


Fig 2.4 - Boxplots of Numerical features

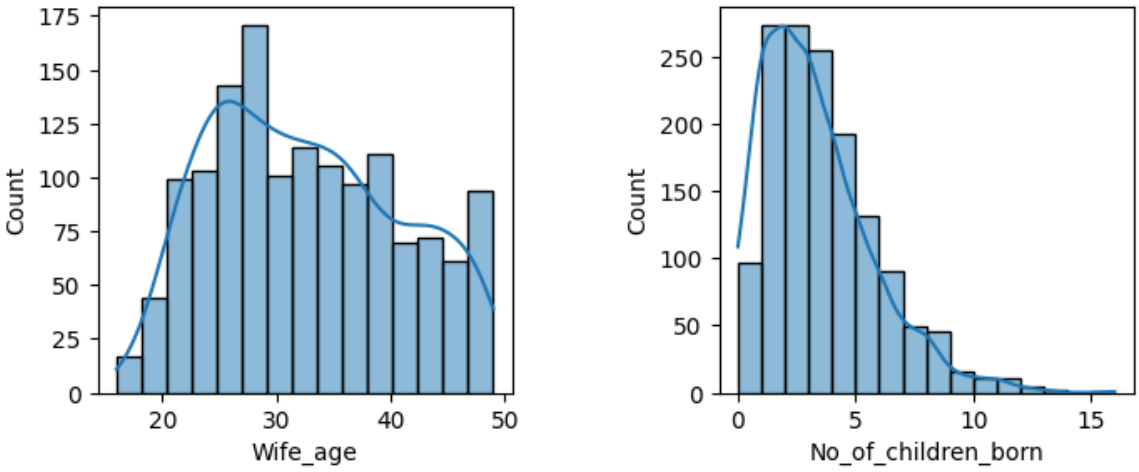


Fig 2.5 - Histograms of Numerical features

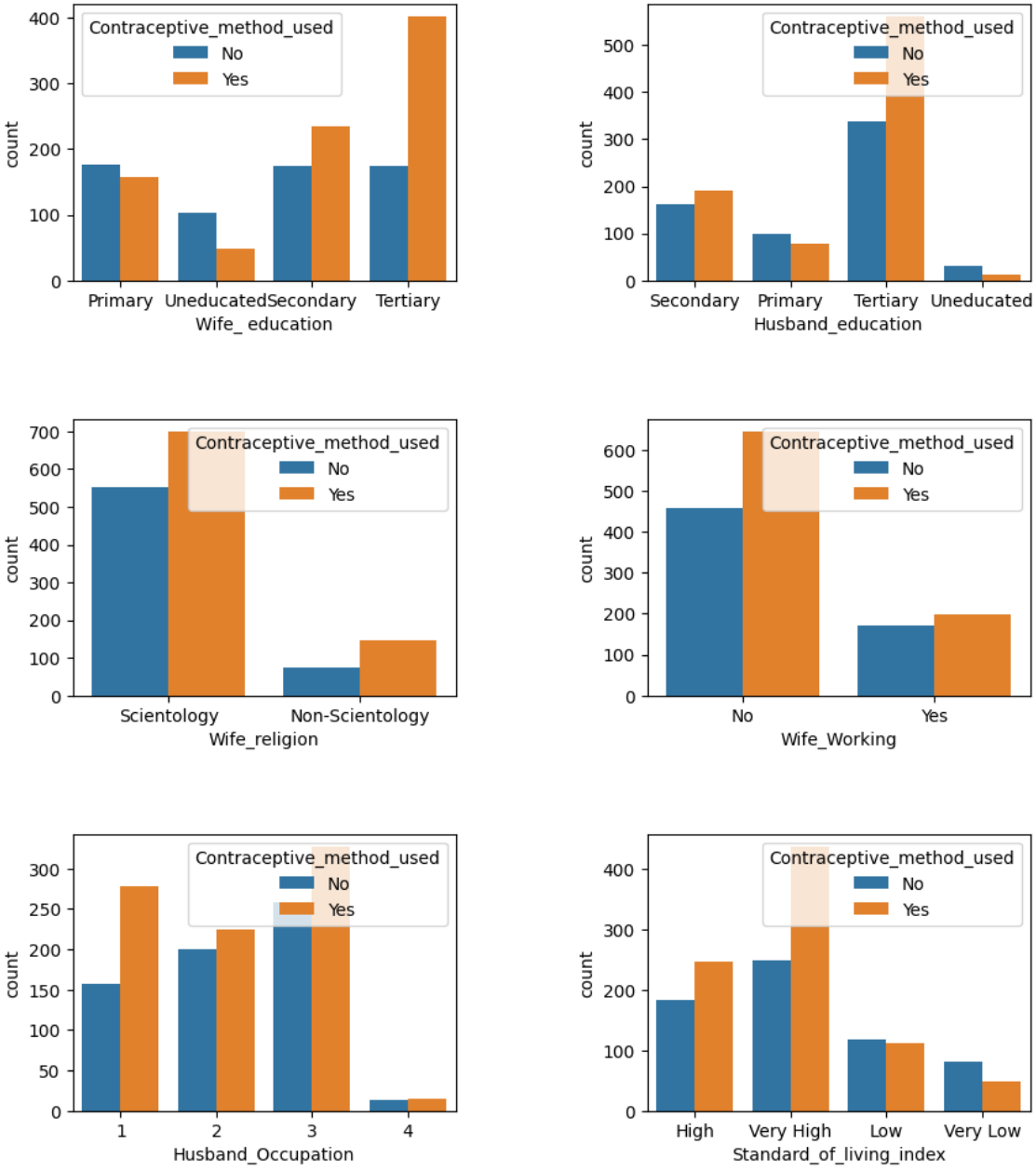


Fig 2.6 - Countplots of categorical features differentiated by target variable

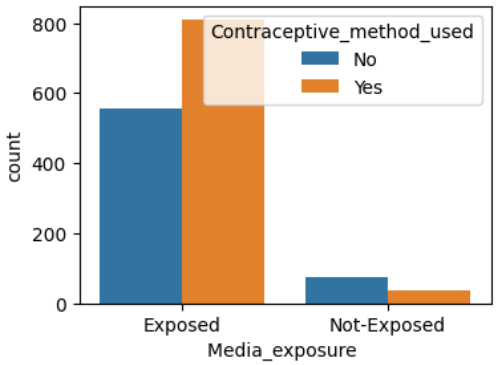


Fig 2.6 - Countplot of Media Exposure

2.2.1 Observations

- Median age of wife is 32 and number of children born are 3.
- Contraceptive use is maximum in Husbands and Wives with Tertiary education.
- Majority wives and husbands have either secondary or tertiary education.
- Contraceptive use is more with wives following Scientology. Majority wives follow Scientology.
- Contraceptive use is proportionately more when wives are not working. Majority of women are not working.
- Majority people have high to very high Standard of living, and contraceptive use is proportionately more.
- Majority have media exposure, and majority of those with media exposure have contraceptive use.

2.3 Bivariate and Multivariate Analysis

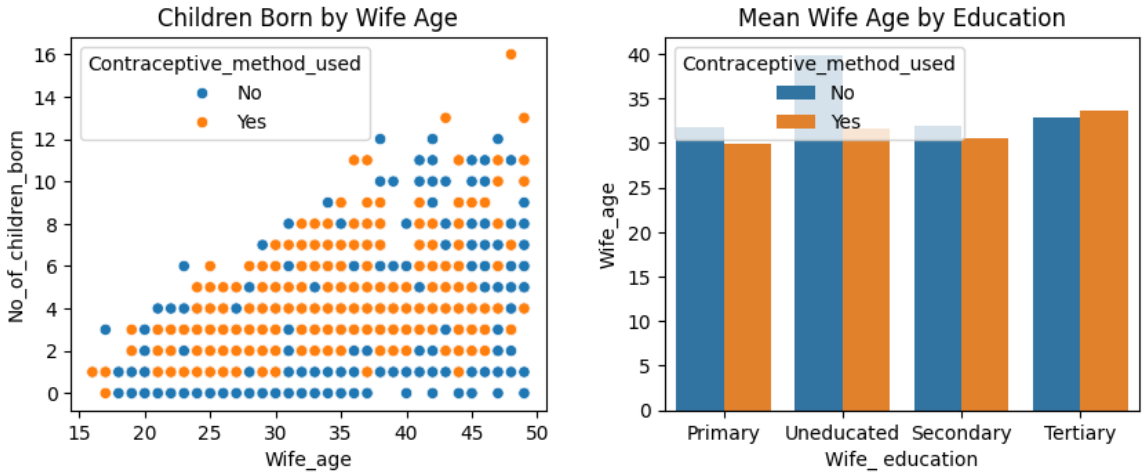


Fig 2.7 - Children born by Wife age & Mean wife age by Education

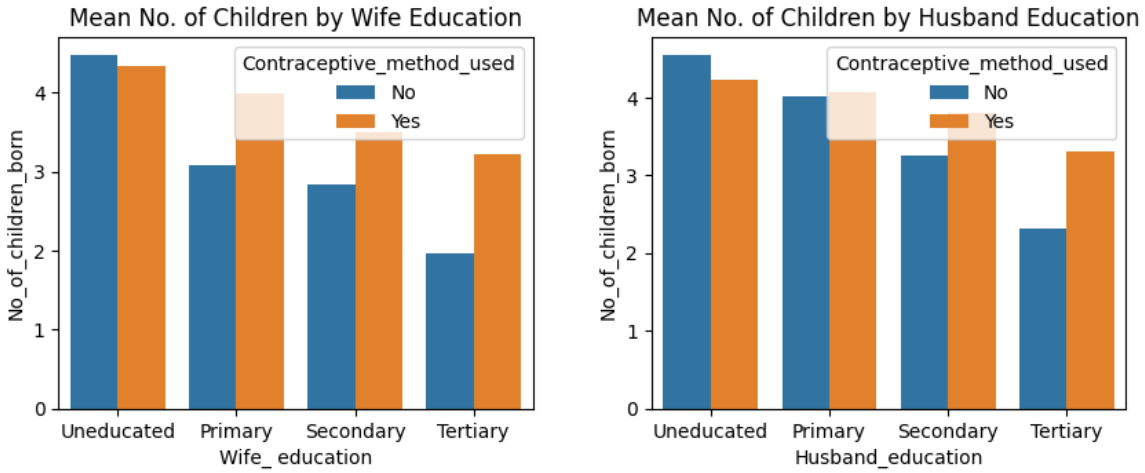


Fig 2.8 - Mean number of children by Wife and Husband Education

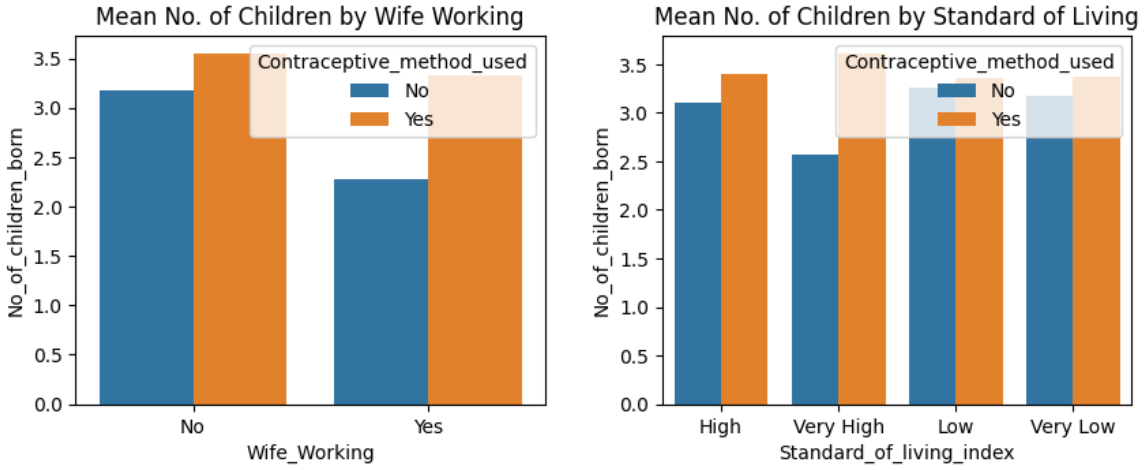


Fig 2.9 - Mean number of children by Wife working and Standard of Living

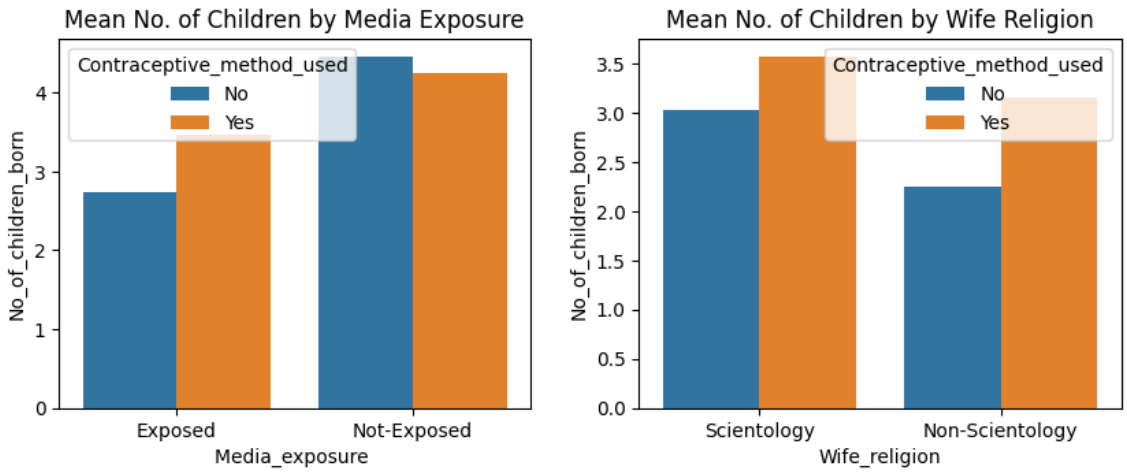


Fig 2.10 - Mean number of children by Media Exposure and Wife religion

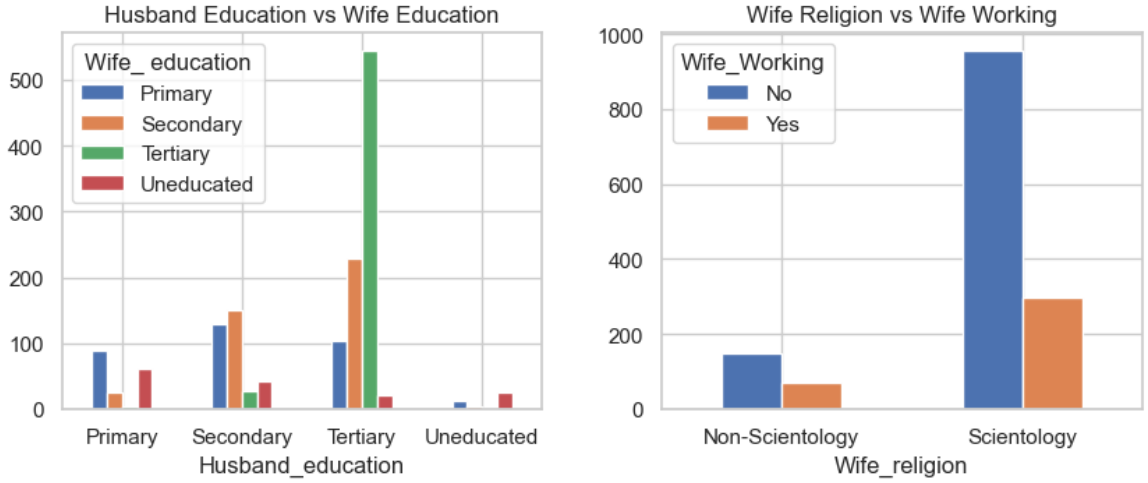


Fig 2.11 - Husband vs Wife education & Wife religion by working

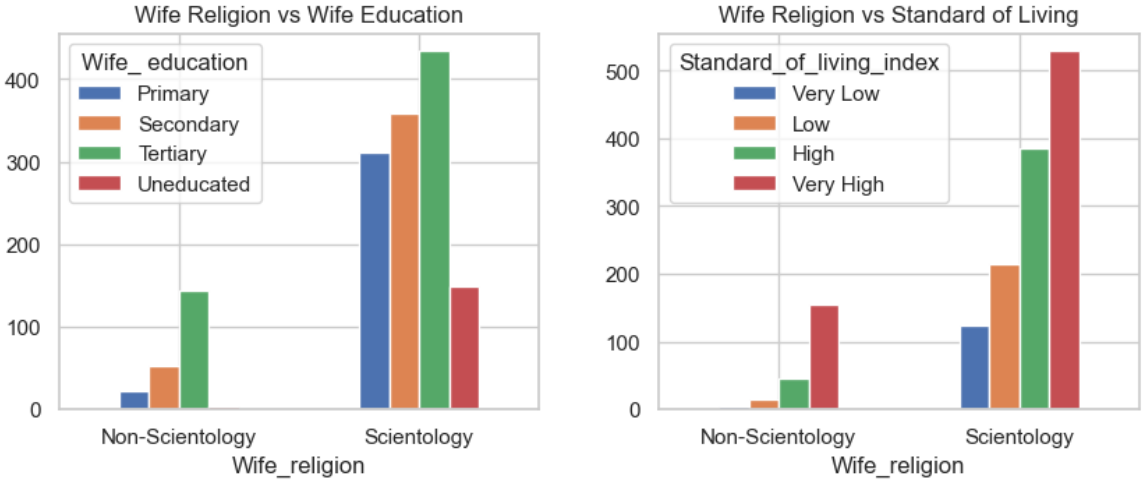


Fig 2.12 - Wife religion vs Wife education & Standard of living

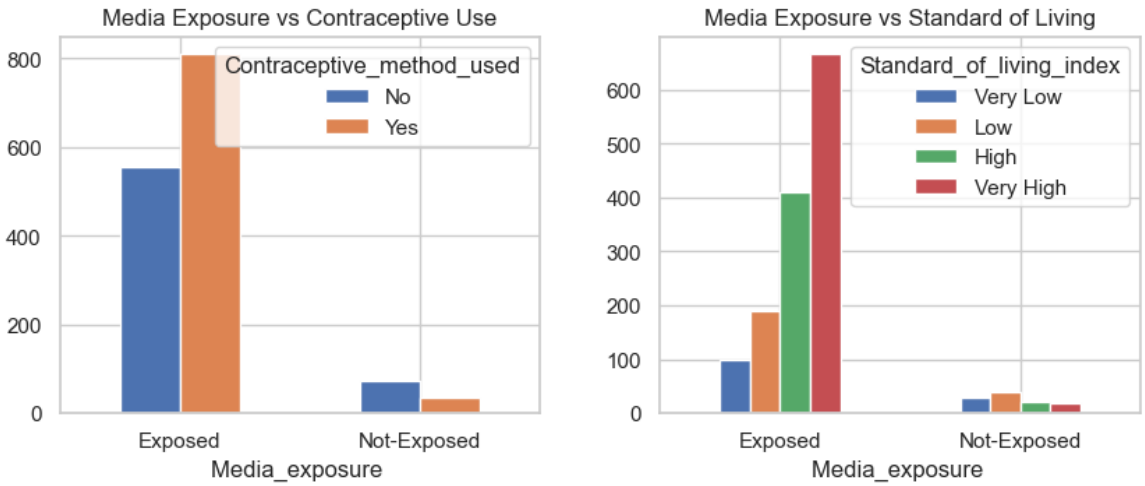


Fig 2.13 - Media exposure vs contraceptive use & Standard of living

2.3.1 Observations

- Older women have more children.
- Uneducated people tend to have the most children while those with Tertiary education tend to have the least.
- Non-working wives tend to have more children. Whether the wife is working or not the contraceptive use is higher in both cases.
- Those with very high standard of living tend to have slightly less children and their contraceptive use is higher.
- Those exposed to Media have lower number of children and their contraceptive use is higher.
- Those following the religion of Scientology have more children while more number of people are using contraceptive.
- Maximum number of husbands and wives have tertiary education, that means majority people are well educated.
- Majority wives following the religion of Scientology are not working.
- Majority wives following the religion of Scientology are well educated.
- Majority wives following the religion of Scientology have very high to high standard of living.
- Those exposed to media have high to very high standard of living.

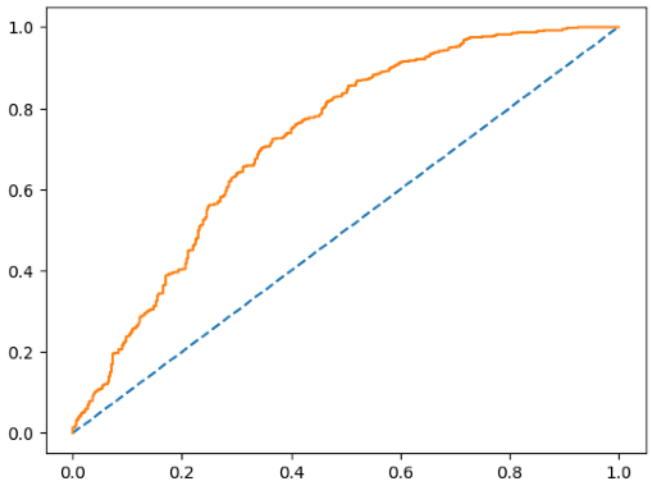
2.4 Logistic Regression

Accuracy score for Train data = 0.693

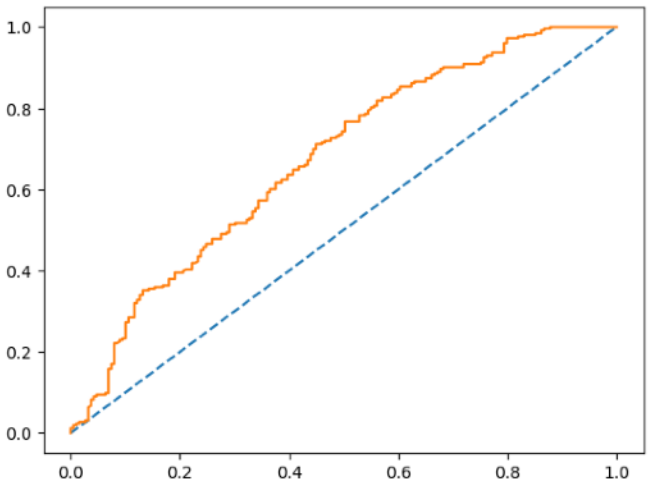
Accuracy score for Test data = 0.652

AUC score for Train data = 0.723

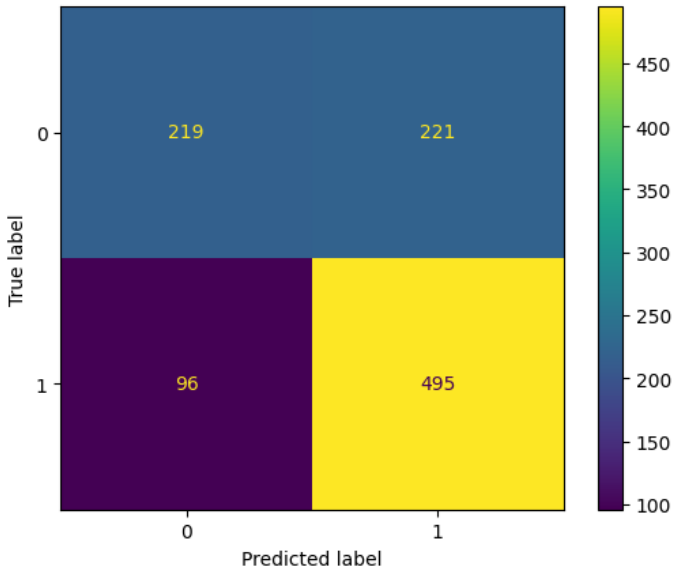
AUC score for Test data = 0.674



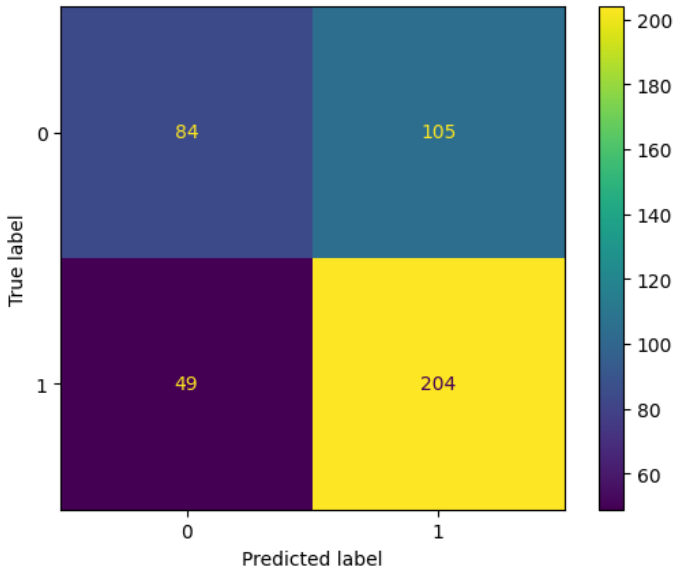
2.4.1 - Curve for Train data



2.4.2 - Curve for Test data



2.4.3 - Confusion matrix for Train data



2.4.4 - Confusion matrix for Test data

	precision	recall	f1-score	support
0	0.70	0.50	0.58	440
1	0.69	0.84	0.76	591
accuracy			0.69	1031
macro avg	0.69	0.67	0.67	1031
weighted avg	0.69	0.69	0.68	1031

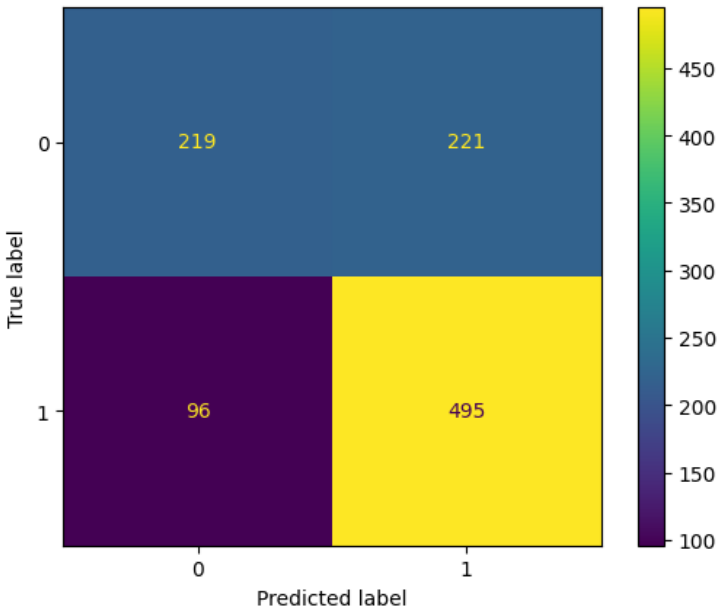
2.4.5 - Classification report for Train data

	precision	recall	f1-score	support
0	0.63	0.44	0.52	189
1	0.66	0.81	0.73	253
accuracy			0.65	442
macro avg	0.65	0.63	0.62	442
weighted avg	0.65	0.65	0.64	442

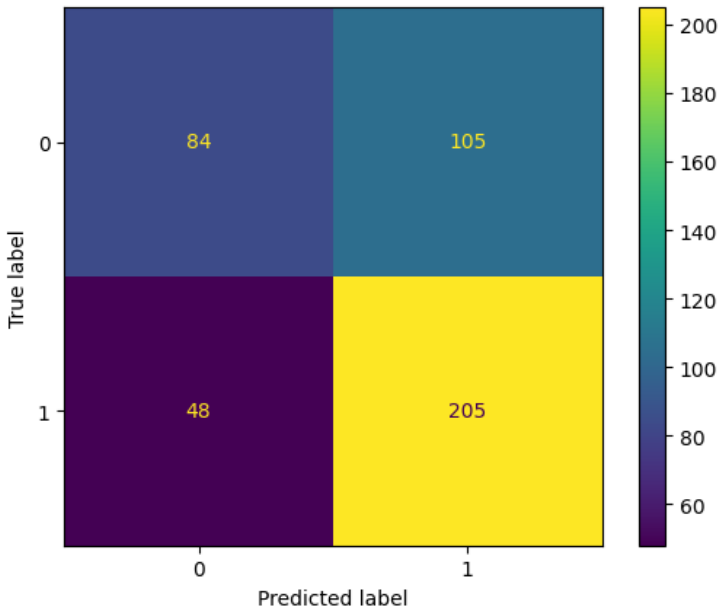
2.4.6 - Classification report for Test data

2.4.1 Applying GridSearch

Best Parameters - penalty: l2, solver: sag, tol: 0.0001



2.4.7 - Confusion matrix for Train data



2.4.8 - Confusion matrix for Test data

	precision	recall	f1-score	support
0	0.70	0.50	0.58	440
1	0.69	0.84	0.76	591
accuracy			0.69	1031
macro avg	0.69	0.67	0.67	1031
weighted avg	0.69	0.69	0.68	1031

2.4.9 - Classification report for Train data

	precision	recall	f1-score	support
0	0.64	0.44	0.52	189
1	0.66	0.81	0.73	253
accuracy			0.65	442
macro avg	0.65	0.63	0.63	442
weighted avg	0.65	0.65	0.64	442

2.4.10 - Classification report for Test data

2.4.2 Inferences

Overall Accuracy = 65% - is Average

For Contraceptive Use = No (Label 0)

Precision (64%) : 64% subjects are correctly predicted to not using contraceptive out of all subjects predicted to not be using contraceptive - Average performance

Recall (44%) : Out of all subjects actually not using contraceptive, only 44% have been correctly predicted. - Poor performance as it is worse than random selection.

For Contraceptive Use = Yes (Label 1)

Precision (64%) : 64% subjects are correctly predicted to using contraceptive out of all subjects predicted to be using contraceptive - Average performance.

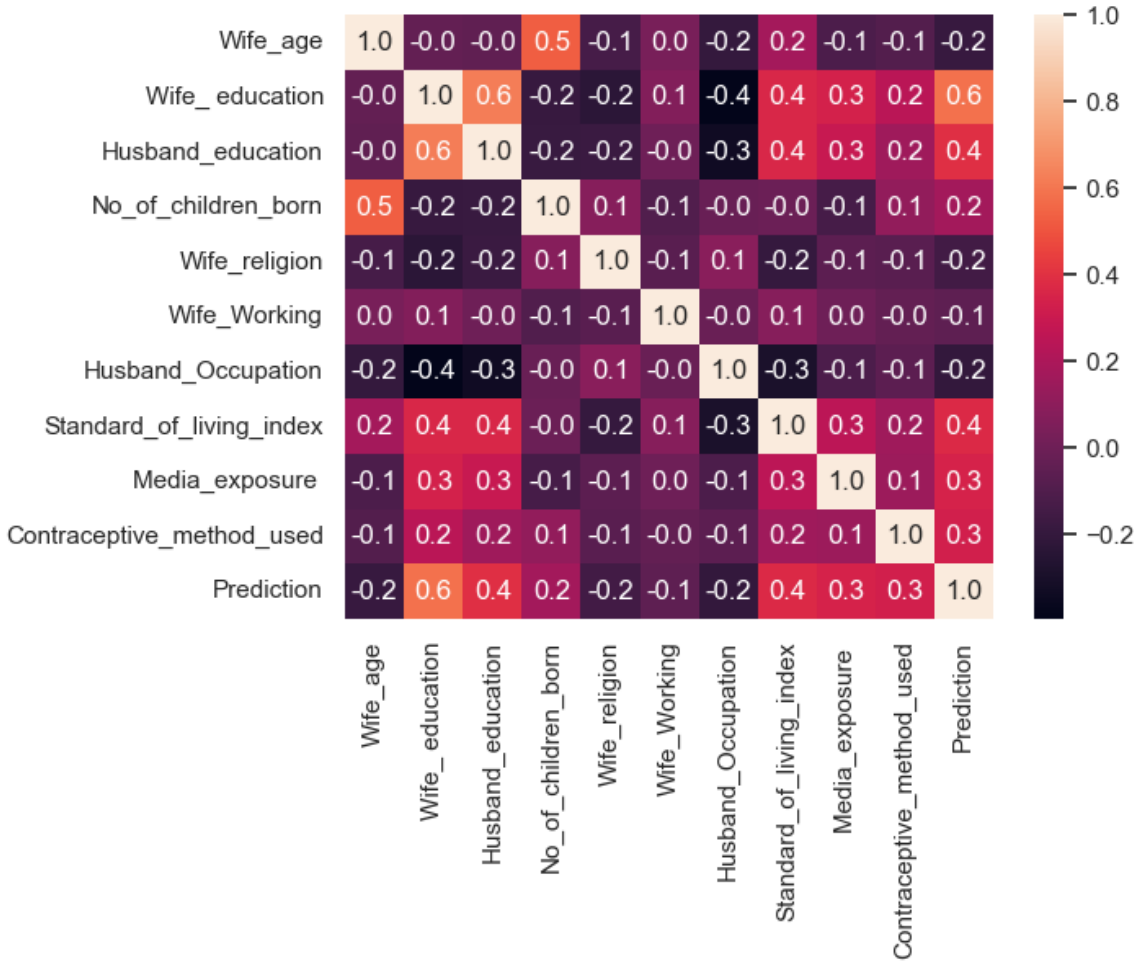
Recall (81%) : Out of all subjects actually using contraceptive, 81% have been correctly predicted. - Good performance.

2.5 Linear Discriminant Analysis

Contraceptive method used actual count:

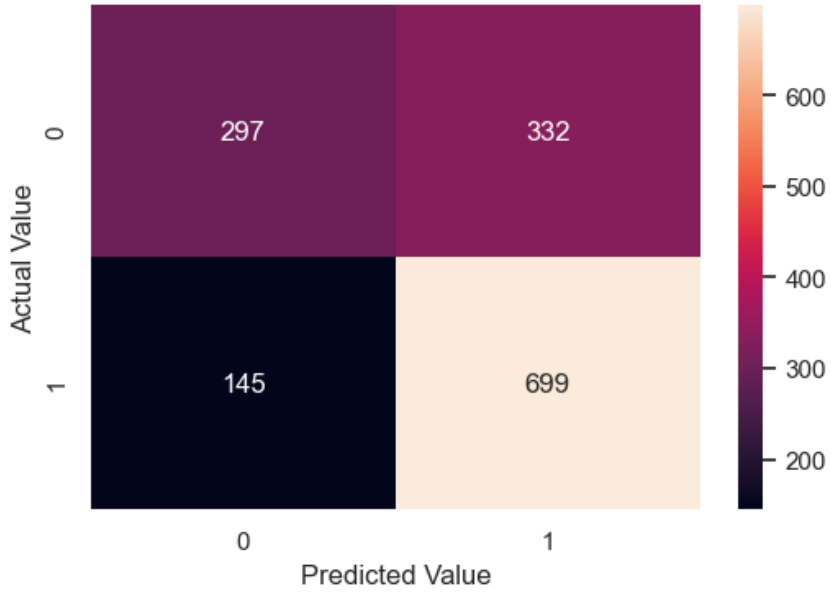
1 : 844

0 : 629



2.5.1 - Heatmap of correlation of various features

2.5.1 Confusion Matrix and Classification Report



2.5.2 - Confusion matrix of model

	precision	recall	f1-score	support
0	0.67	0.47	0.55	629
1	0.68	0.83	0.75	844
accuracy			0.68	1473
macro avg	0.67	0.65	0.65	1473
weighted avg	0.68	0.68	0.66	1473

2.5.3 - Classification report of model

Overall accuracy = 68% - Average

For Contraceptive Use = No (Label 0)

Precision (67%) : 67% subjects are correctly predicted to not using contraceptive out of all subjects predicted to not be using contraceptive - Average performance

Recall (47%) : Out of all subjects actually not using contraceptive, only 47% have been correctly predicted. - Poor performance as it is worse than random selection.

For Contraceptive Use = Yes (Label 1)

Precision (68%) : 68% subjects are correctly predicted to using contraceptive out of all subjects predicted to be using contraceptive - Average performance.

Recall (83%) : Out of all subjects actually using contraceptive, 83% have been correctly predicted. - Good performance.

2.5.2 Building the Equation

```
LDF = 0.344 + -0.608 * ( Wife_age ) +  0.56 * ( Wife_ education ) +  0.005 * ( Husband_education ) +
0.725 * ( No_of_children_born ) +  -0.13 * ( Wife_religion ) +  -0.062 * ( Wife_Working ) +  0.067 *
( Husband_Occupation ) +  0.256 * ( Standard_of_living_index ) +  0.114 * ( Media_exposure  )
```

Wife Age, Wife education and Number of children born have the most weight in predicting contraceptive use, out of which Wife Age is negatively correlated. That means as Wife's Age increases, the tendency to use Contraceptive decreases.

2.5.3 Classification by Discriminant Score & Probability

```
FOR Row: 0

--> { DS:  0.352 >=0 , Classify as 1}
--------------------------------------
FOR Row: 1

--> { DS:  0.457 >=0 , Classify as 1}
--------------------------------------
FOR Row: 2

--> { DS:  0.233 >=0 , Classify as 1}
--------------------------------------
FOR Row: 3

--> { DS:  1.472 >=0 , Classify as 1}
--------------------------------------
FOR Row: 4

--> { DS:  1.621 >=0 , Classify as 1}
--------------------------------------
FOR Row: 5

--> { DS:  0.989 >=0 , Classify as 1}
--------------------------------------
FOR Row: 6

--> { DS:  0.301 >=0 , Classify as 1}
--------------------------------------
```

2.5.4 - Discriminant score of each row

```
FOR Row: 0

--> { prob(Y=1|X) = 0.587 >=0.5 , Classify as 1 }
-------------------------------------------------
FOR Row: 1

--> { prob(Y=1|X) = 0.672 >=0.5 , Classify as 1 }
-------------------------------------------------
FOR Row: 2

--> { prob(Y=1|X) = 0.621 >=0.5 , Classify as 1 }
-------------------------------------------------
FOR Row: 3

--> { prob(Y=1|X) = 0.813 >=0.5 , Classify as 1 }
-------------------------------------------------
FOR Row: 4

--> { prob(Y=1|X) = 0.795 >=0.5 , Classify as 1 }
-------------------------------------------------
FOR Row: 5

--> { prob(Y=1|X) = 0.728 >=0.5 , Classify as 1 }
-------------------------------------------------
FOR Row: 6

--> { prob(Y=1|X) = 0.509 >=0.5 , Classify as 1 }
-------------------------------------------------
FOR Row: 7
```

2.5.5 - 2.5.4 - Probability of each row

Discriminant Score

1055 Rows classified as 1 (actually 844)

418 rows classified as 0 (actually 629)

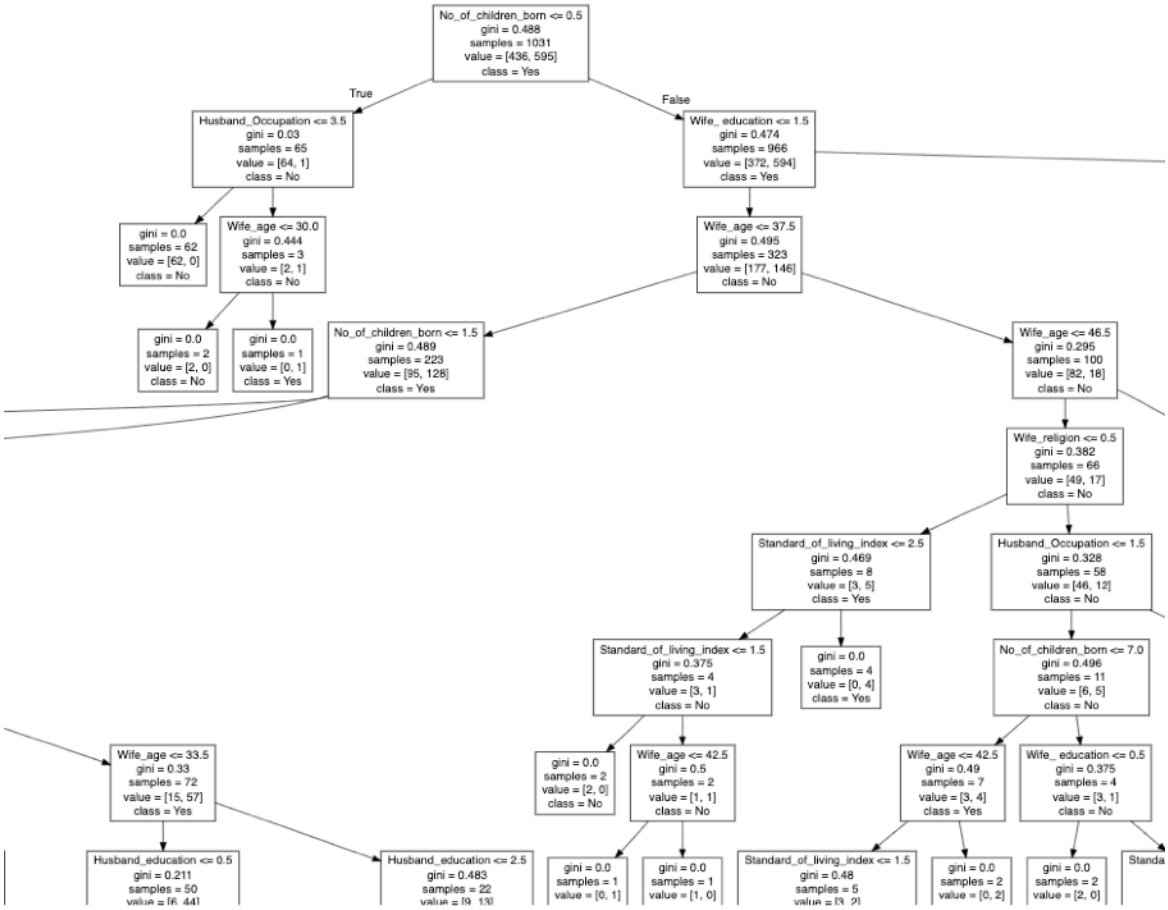
Probability

1031 Rows classified as 1 (actually 844)

442 rows classified as 0 (actually 629)

2.6 CART

2.6.1 Basic Tree



2.6.1 - Part of Decision Tree

	Imp
Wife_age	0.347416
No_of_children_born	0.222045
Wife_ education	0.106002
Husband_Occupation	0.087502
Standard_of_living_index	0.087465
Husband_education	0.058855
Wife_Working	0.046789
Wife_religion	0.031956
Media_exposure	0.011970

2.6.2 - Variable Importance

2.6.2 Regularising the Model

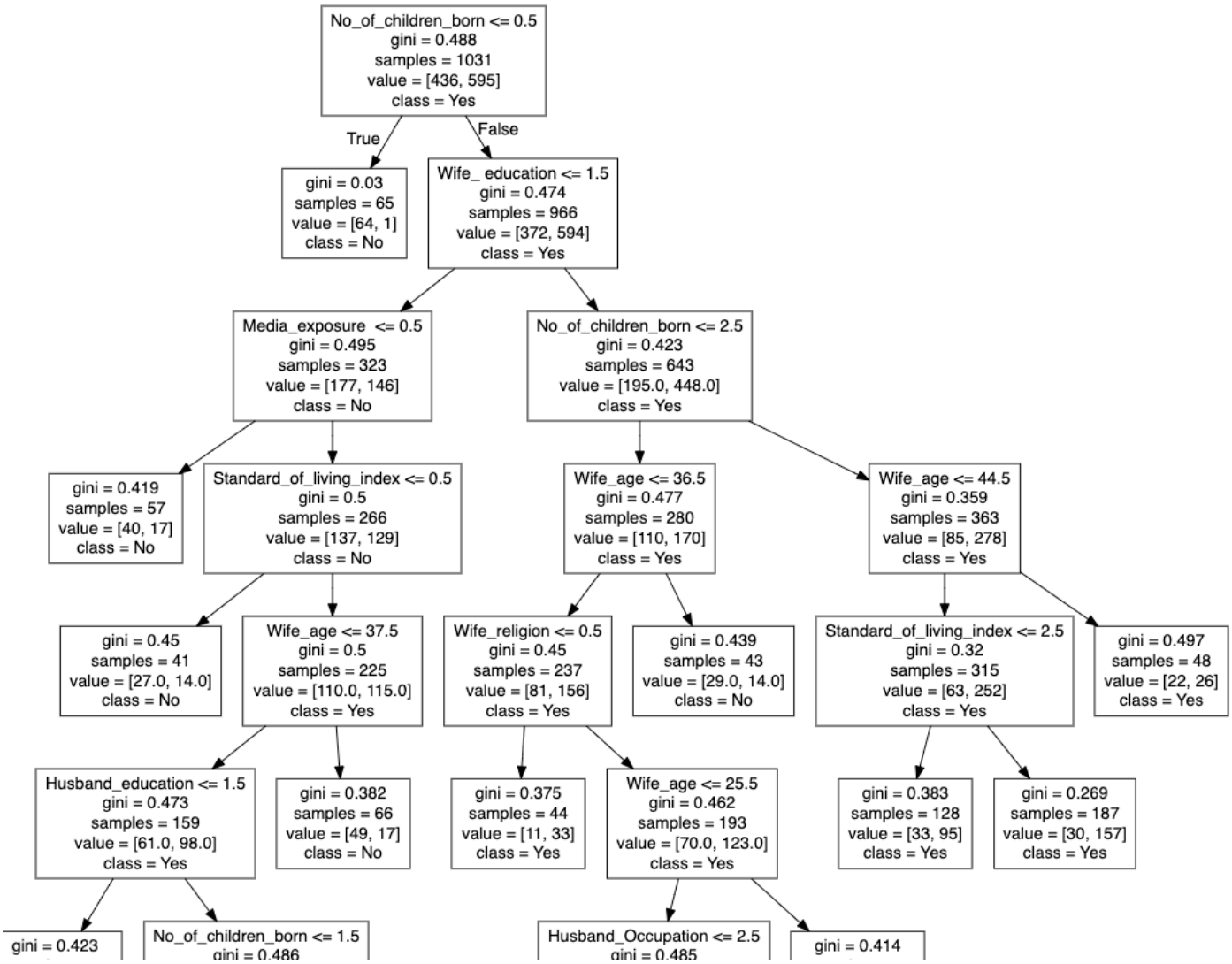
Best Parameters given by GridSearch

```
{'ccp_alpha': 0.001,
 'criterion': 'gini',
 'max_depth': 10,
 'max_features': 'sqrt',
 'min_samples_leaf': 20}
```

2.6.3 - Best parameters given by GridSearch

	Imp
No_of_children_born	0.492656
Wife_age	0.217457
Wife_ education	0.207260
Standard_of_living_index	0.027666
Media_exposure	0.026342
Husband_Occupation	0.014741
Wife_religion	0.007325
Husband_education	0.006552
Wife_Working	0.000000

2.6.4 - Variable importances of Regularised model



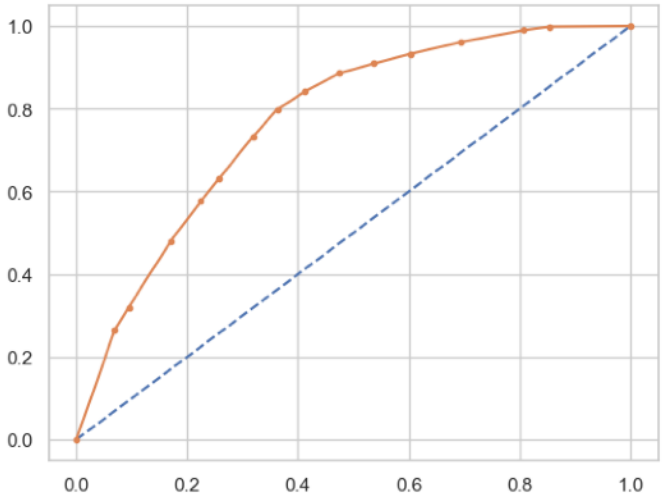
2.6.5 - Part of Decision tree of Regularised model

- Number of children born moves at first place and its importance changes from 0.34 to 0.49
- Wife age moves down to second place but importance chances little from 0.22 to 0.21
- Wife education remains at the same spot but its importance changes from 0.11 to 0.21

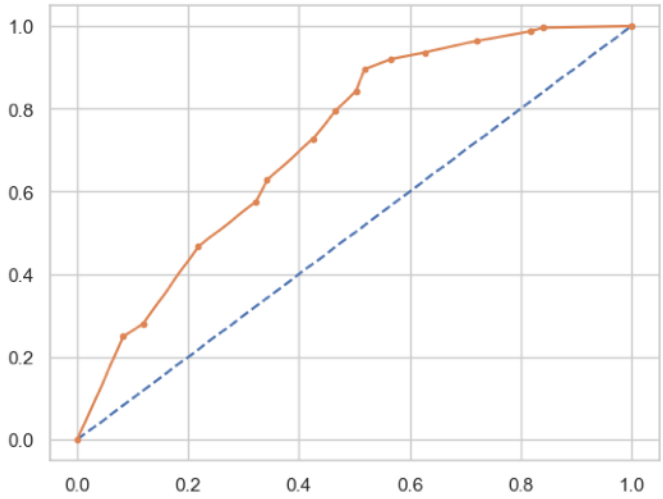
2.6.3 AUC - ROC

AUC for Training data = 0.771

AUC for Test data = 0.719



2.6.6 - ROC curve for Train data



2.6.7 - ROC curve for Test data

2.6.4 Confusion Matrix

	precision	recall	f1-score	support
0	0.73	0.59	0.65	436
1	0.74	0.84	0.79	595
accuracy			0.73	1031
macro avg	0.73	0.71	0.72	1031
weighted avg	0.73	0.73	0.73	1031

2.6.8 - Confusion matrix for Train data

	precision	recall	f1-score	support
0	0.71	0.50	0.59	193
1	0.68	0.84	0.76	249
accuracy			0.69	442
macro avg	0.70	0.67	0.67	442
weighted avg	0.70	0.69	0.68	442

2.6.9 - Confusion matrix for Test data

Overall Accuracy = 69% - is Average

For Contraceptive Use = No (Label 0)

Precision (71%) : 71% subjects are correctly predicted to not using contraceptive out of all subjects predicted to not be using contraceptive - Average performance

Recall (50%) : Out of all subjects actually not using contraceptive, only 50% have been correctly predicted. - Poor performance as it is as good as random selection.

For Contraceptive Use = Yes (Label 1)

Precision (68%) : 68% subjects are correctly predicted to using contraceptive out of all subjects predicted to be using contraceptive - Average performance.

Recall (84%) : Out of all subjects actually using contraceptive, 84% have been correctly predicted. - Good performance.

2.6.5 Actionable Insights and Recommendations

- To optimize contraceptive use, focus on Wife Age, Wife Education, and Number of children born.
- Advertisements or other efforts to pursuade women to use contraceptives should focus on older, less educated women with more children, as this demographic is least likely to use contraceptives.
- Recall of 0s and Precision of 1s will be the imporant metric here. Recall of subjects not using contraceptives is poor, which means those who are not using contraceptives have been marked as using. For this more data is required to Train the model regarding subjects who are not using contraceptives. This will improve the Precision of those using contraceptives as well.