PREDICTIVE MODELLING PROJECT GROUP3

BUSINESS REPORT

PGP DSBA – APR 2023

Submitted by:
Sruthi C

List of Figures

Contents

Problem 1: Linear Regression

The comp-activ database is a collection of 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 cpu runs in user mode) and to find out how each attribute affects the system to be in 'usr' mode using a list of system attributes.

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.

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.

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.

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

Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Problem 2: **Logistic Regression**

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

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.

2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression

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

2.4 Inference: Basis on these predictions, what are the insights and recommendations.
Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Problem Statement – Linear Regression

The comp-activ database is a collection of 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 cpu runs in user mode) and to find out how each attribute affects the system to be in 'usr' mode using a list of system attributes.

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.

- Data frame-

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

8192 rows × 22 columns

freemem	freeswap	usr
4670	1730946	95
7278	1869002	97
702	1021237	87
7248	1863704	98
633	1760253	90
...	...	...
387	986647	80
263	1055742	90
400	969106	87
141	1022458	83
659	1756514	94

Figure 1 : Linear regression Data frame

- Head- Top 5 rows

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

	runqsz	freemem	freeswap	usr
	CPU_Bound	4670	1730946	95
	Not_CPU_Bound	7278	1869002	97
	Not_CPU_Bound	702	1021237	87
	Not_CPU_Bound	7248	1863704	98
	Not_CPU_Bound	633	1760253	90

Figure 2: Head (top 5)

- Tail-Last 5 rows

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt
8187	16	12	3009	360	244	1.6	5.81	405250.0	85282.0	8.02	...	55.11	0.6	35.87	47.90	139.28	270.74
8188	4	0	1596	170	146	2.4	1.80	89489.0	41764.0	3.80	...	0.20	0.8	3.80	4.40	122.40	212.60
8189	16	5	3116	289	190	0.6	0.60	325948.0	52640.0	0.40	...	0.00	0.4	28.40	45.20	60.20	219.80
8190	32	45	5180	254	179	1.2	1.20	62571.0	29505.0	1.40	...	18.04	0.4	23.05	24.25	93.19	202.81
8191	2	0	985	55	46	1.6	4.80	111111.0	22256.0	0.00	...	0.00	0.2	3.40	6.20	91.80	110.00

5 rows × 22 columns

	runqsz	freemem	freeswap	usr
	CPU_Bound	387	986647	80
	Not_CPU_Bound	263	1055742	90
	Not_CPU_Bound	400	969106	87
	CPU_Bound	141	1022458	83
	CPU_Bound	659	1756514	94

Figure 3: Tail (last 5)

- Shape- (8192,22)
  We have 8192 rows and 22 columns.

- Describe-

	count	mean	std	min	25%	50%	75%	max
lread	8192.0	1.955969e+01	53.353799	0.0	2.0	7.0	20.000	1845.00
lwrite	8192.0	1.310620e+01	29.891726	0.0	0.0	1.0	10.000	575.00
scall	8192.0	2.306318e+03	1633.617322	109.0	1012.0	2051.5	3317.250	12493.00
sread	8192.0	2.104800e+02	198.980146	6.0	86.0	166.0	279.000	5318.00
swrite	8192.0	1.500582e+02	160.478980	7.0	63.0	117.0	185.000	5456.00
fork	8192.0	1.884554e+00	2.479493	0.0	0.4	0.8	2.200	20.12
exec	8192.0	2.791998e+00	5.212456	0.0	0.2	1.2	2.800	59.56
rchar	8088.0	1.973857e+05	239837.493526	278.0	34091.5	125473.5	267828.750	2526649.00
wchar	8177.0	9.590299e+04	140841.707911	1498.0	22916.0	46619.0	106101.000	1801623.00
pgout	8192.0	2.285317e+00	5.307038	0.0	0.0	0.0	2.400	81.44
ppgout	8192.0	5.977229e+00	15.214590	0.0	0.0	0.0	4.200	184.20
pgfree	8192.0	1.191971e+01	32.363520	0.0	0.0	0.0	5.000	523.00
pgscan	8192.0	2.152685e+01	71.141340	0.0	0.0	0.0	0.000	1237.00
atch	8192.0	1.127505e+00	5.708347	0.0	0.0	0.0	0.600	211.58
pgin	8192.0	8.277960e+00	13.874978	0.0	0.6	2.8	9.765	141.20
ppgin	8192.0	1.238859e+01	22.281318	0.0	0.6	3.8	13.800	292.61
pflt	8192.0	1.097938e+02	114.419221	0.0	25.0	63.8	159.600	899.80
vflt	8192.0	1.853158e+02	191.000603	0.2	45.4	120.4	251.800	1365.00
freemem	8192.0	1.763456e+03	2482.104511	55.0	231.0	579.0	2002.250	12027.00
freeswap	8192.0	1.328126e+06	422019.426957	2.0	1042623.5	1289289.5	1730379.500	2243187.00
usr	8192.0	8.396887e+01	18.401905	0.0	81.0	89.0	94.000	99.00

Figure 4: Description

- Info-

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

Figure 5: Data Information

There are 13 floats, 8 integers and 1 object.

EXPLORATORY DATA ANALYSIS

Univariate analysis:

A boxplot and histogram are drawn for numerical variables and a countplot in case of categorical variable.

Independent variable: usr

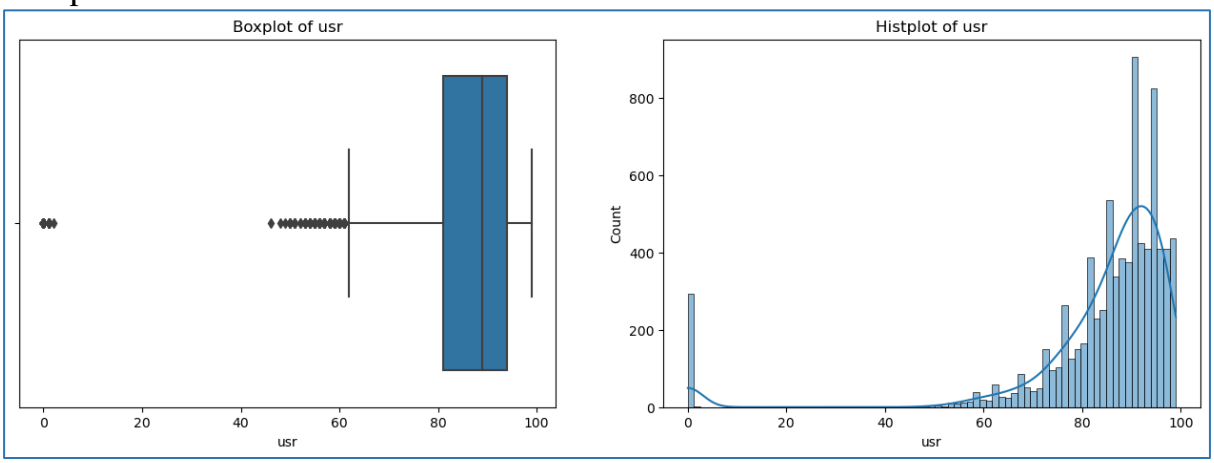


Figure 6: Boxplot and histplot of independent variable

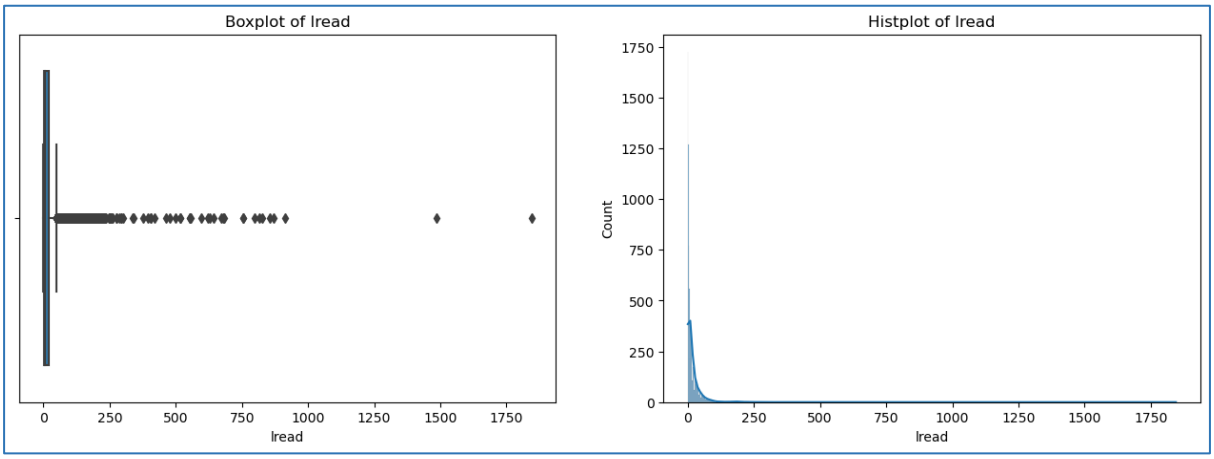
5-point summary -

Minimum usr: 0
First quartile (Q1) of usr: 81.0
Median value: 89.0
Third quartile (Q3) of usr: 94.0

Maximum usr: 99

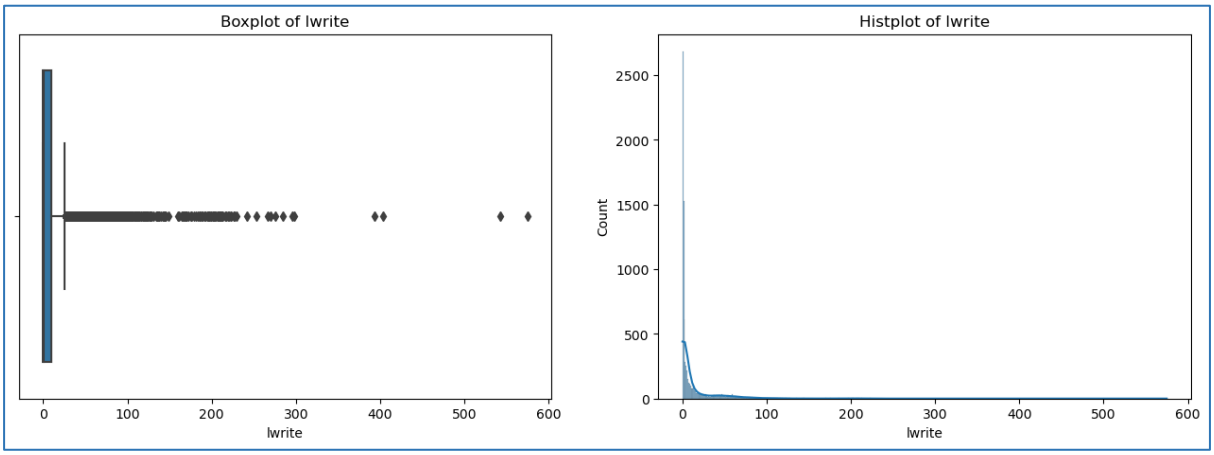
Dependent variables:

- Lread



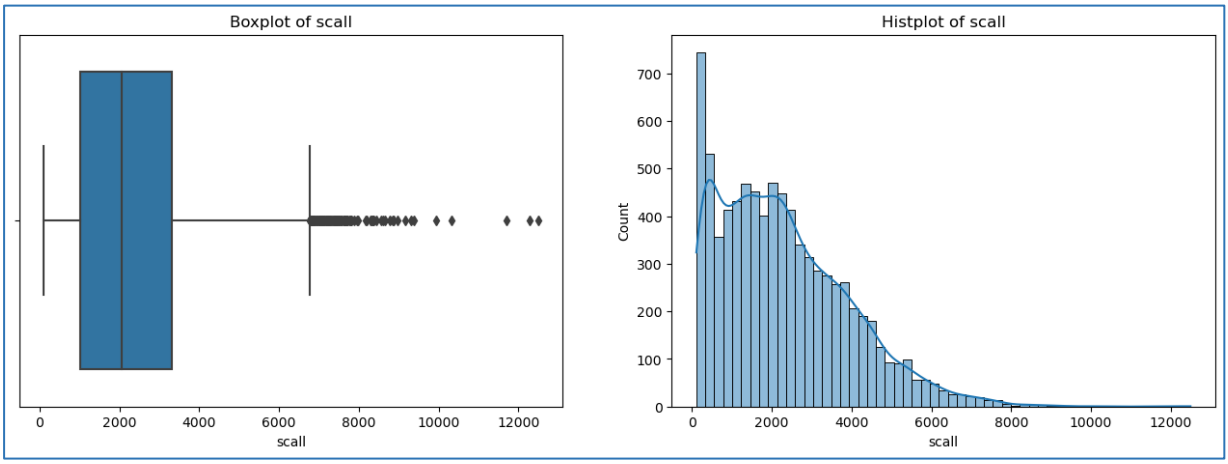
Minimum lread: 0
First quartile (Q1) of lread: 2.0
Median value: 7.0
Third quartile (Q3) of lread: 20.0
Maximum lread: 1845

- Lwrite



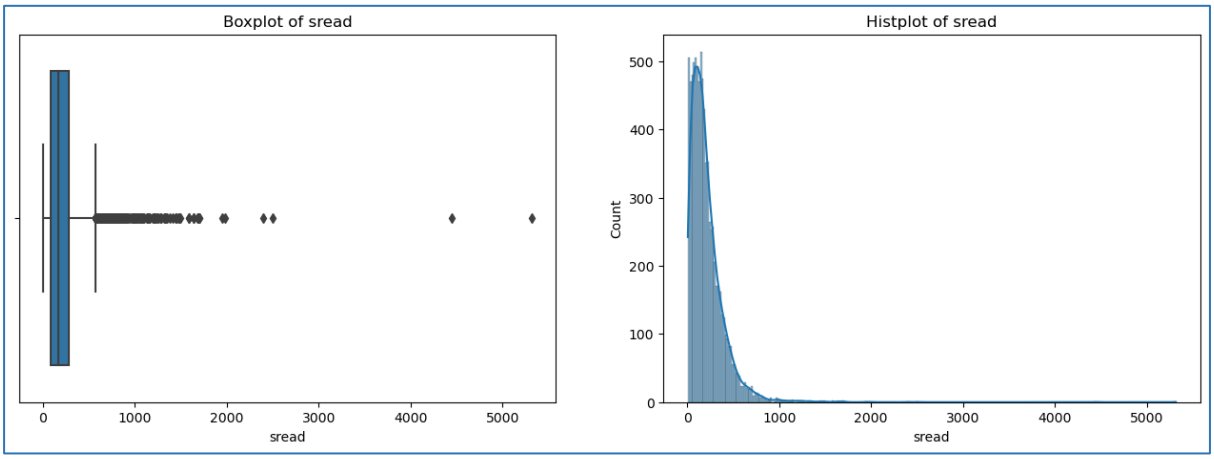
Minimum lwrite: 0
First quartile (Q1) of lwrite: 0.0
Median value: 1.0
Third quartile (Q3) of lwrite: 10.0
Maximum lwrite: 575

- Scall



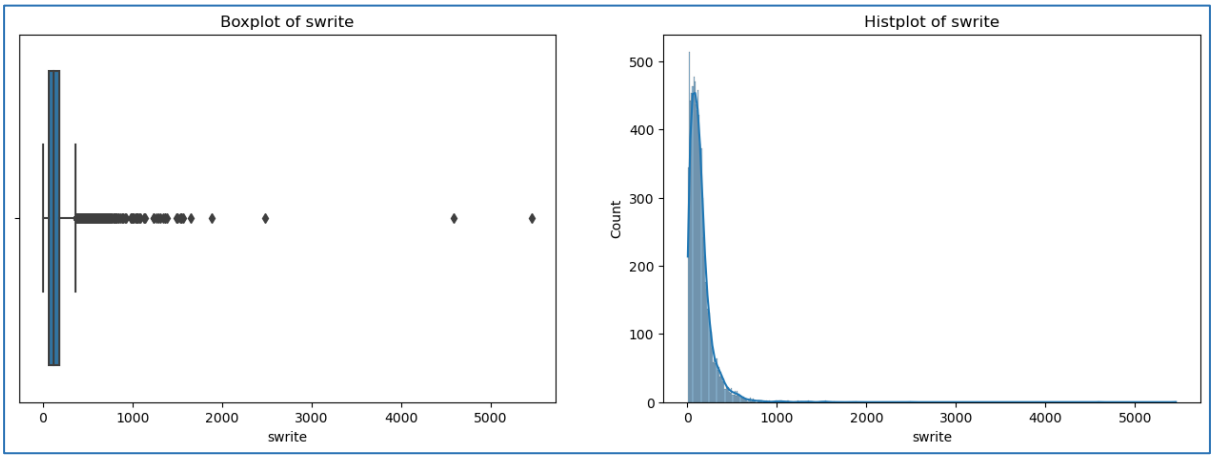
Minimum scall: 109
First quartile (Q1) of scall: 1012.0
Median value: 2051.5
Third quartile (Q3) of scall: 3317.25
Maximum scall: 12493

- Sread



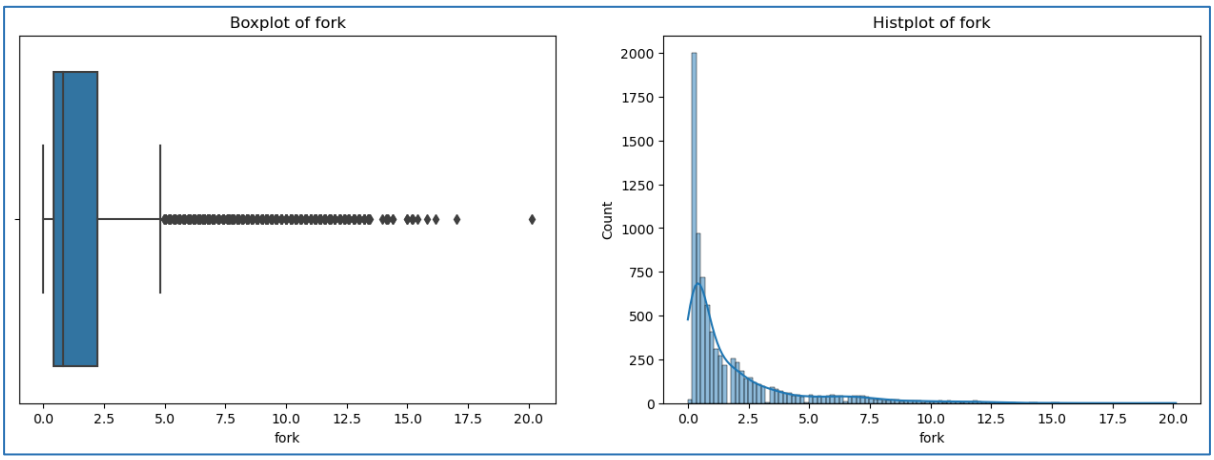
Minimum sread: 6
First quartile (Q1) of sread: 86.0
Median value: 166.0
Third quartile (Q3) of sread: 279.0
Maximum sread: 5318

- Swrite



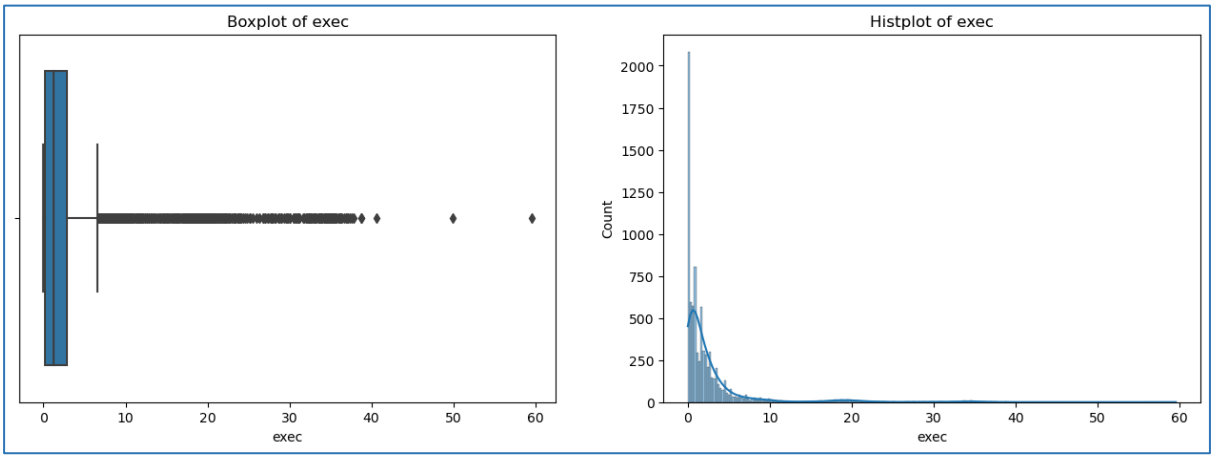
Minimum swrite: 7
First quartile (Q1) of swrite: 63.0
Median value: 117.0
Third quartile (Q3) of swrite: 185.0
Maximum swrite: 5456

- Fork



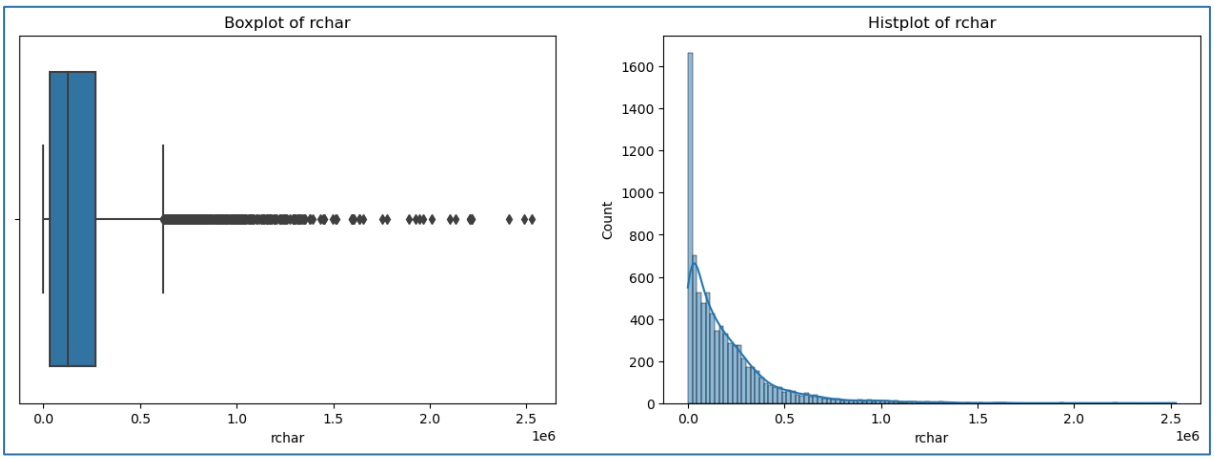
Minimum fork: 0.0
First quartile (Q1) of fork: 0.4
Median value: 0.8
Third quartile (Q3) of fork: 2.2
Maximum fork: 20.12

- Exec



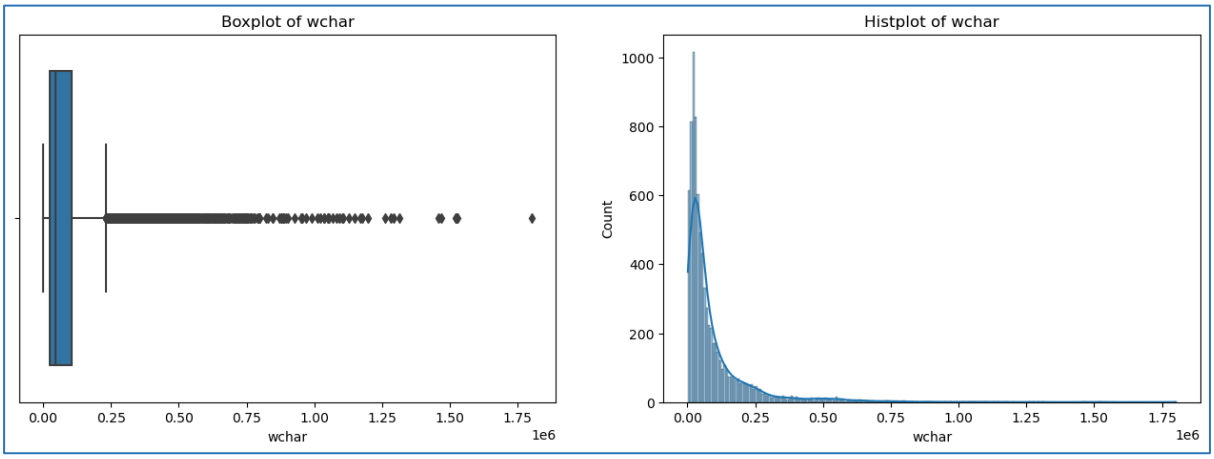
Minimum exec: 0.0
First quartile (Q1) of exec: 0.2
Median value: 1.2
Third quartile (Q3) of exec: 2.8
Maximum exec: 59.56

- Rchar



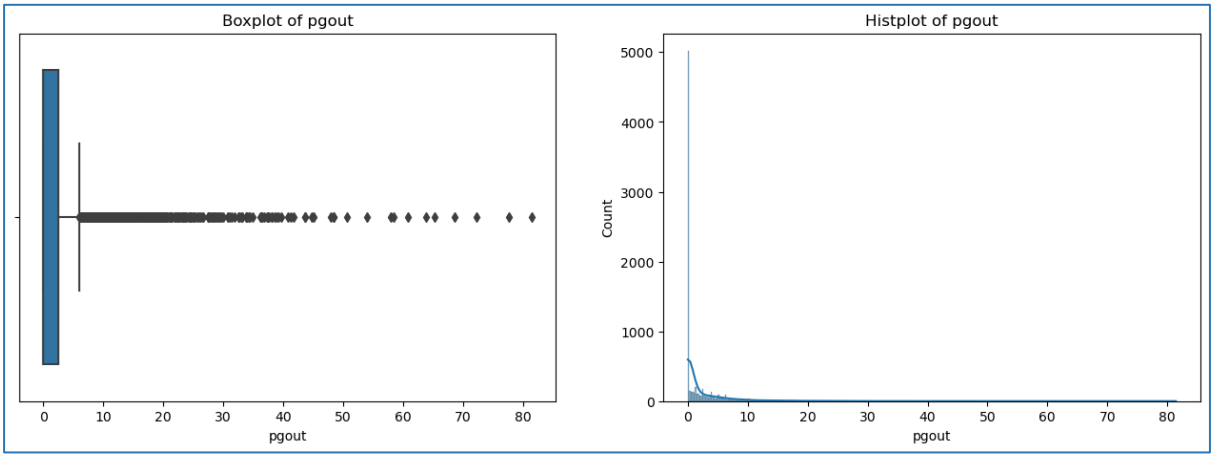
Minimum rchar: 278.0
First quartile (Q1) of rchar: nan
Median value: 125473.5
Third quartile (Q3) of rchar: nan
Maximum rchar: 2526649.0

- Wchar



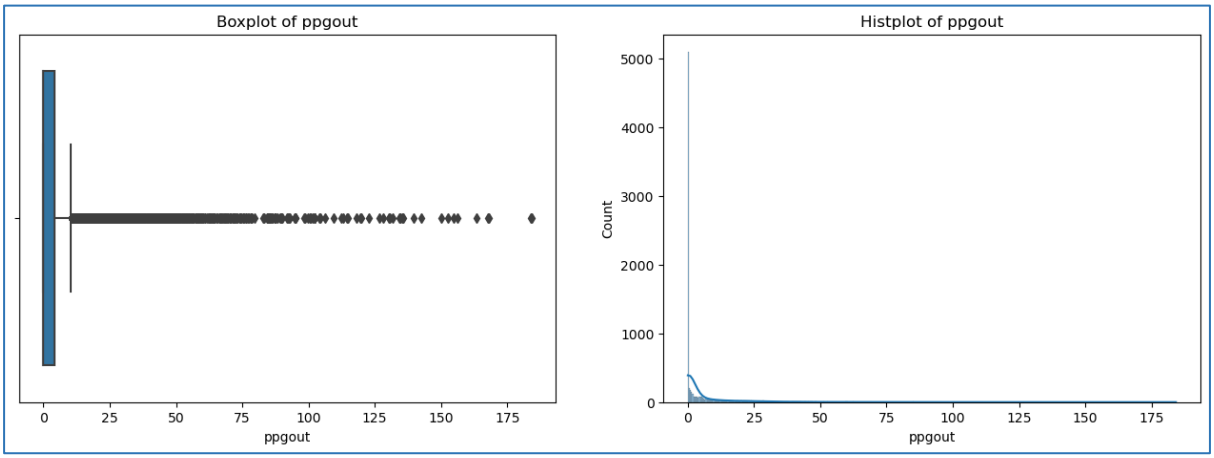
Minimum wchar: 1498.0
First quartile (Q1) of wchar: nan
Median value: 46619.0
Third quartile (Q3) of wchar: nan
Maximum wchar: 1801623.0

- Pgout



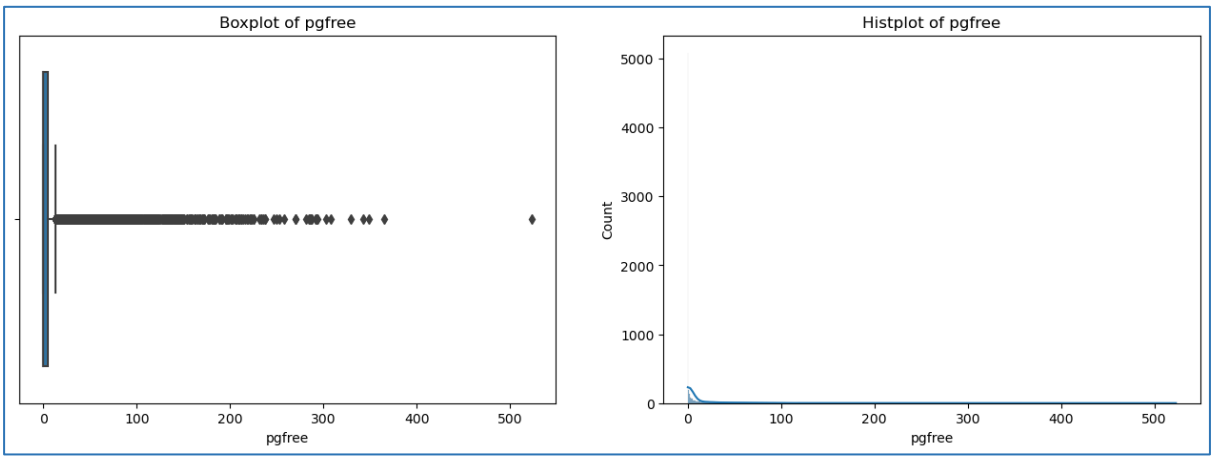
Minimum pgout: 0.0
First quartile (Q1) of pgout: 0.0
Median value: 0.0
Third quartile (Q3) of pgout: 2.4
Maximum pgout: 81.44

- Ppgout



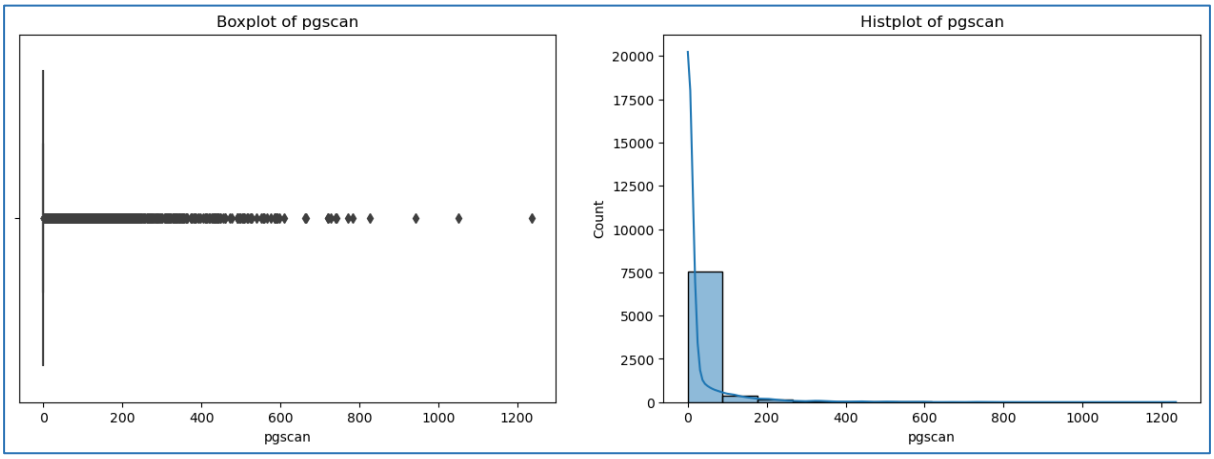
Minimum ppgout: 0.0
First quartile (Q1) of ppgout: 0.0
Median value: 0.0
Third quartile (Q3) of ppgout: 4.2
Maximum ppgout: 184.2

- Pgfree



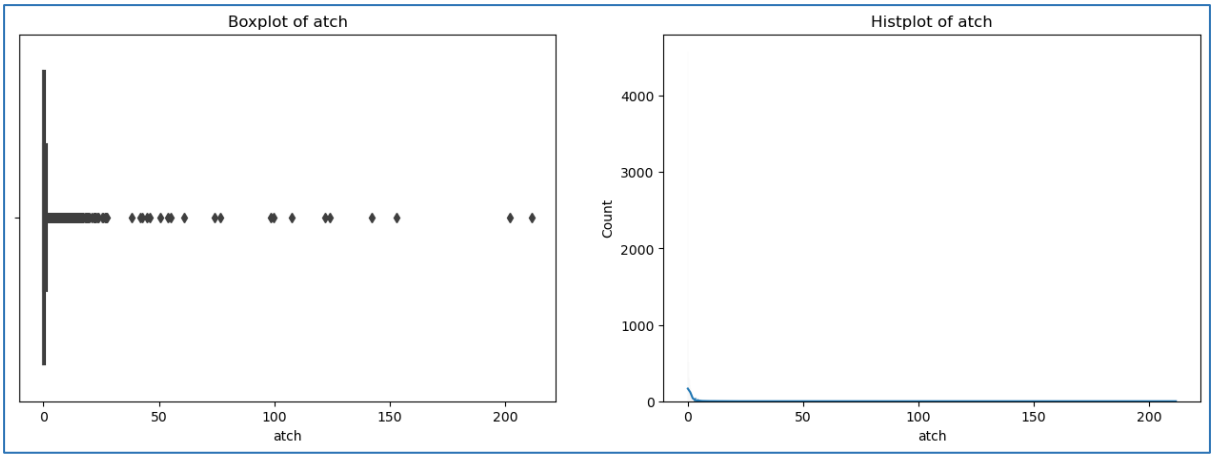
Minimum pgfree: 0.0
First quartile (Q1) of pgfree: 0.0
Median value: 0.0
Third quartile (Q3) of pgfree: 5.0
Maximum pgfree: 523.0

- Pgscan



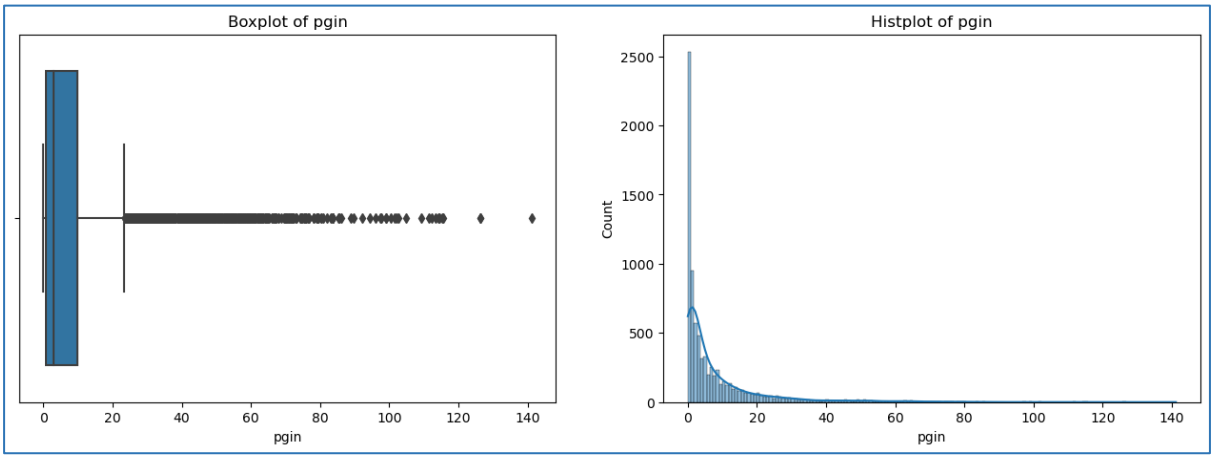
Minimum pgscan: 0.0
First quartile (Q1) of pgscan: 0.0
Median value: 0.0
Third quartile (Q3) of pgscan: 0.0
Maximum pgscan: 1237.0

- Atch



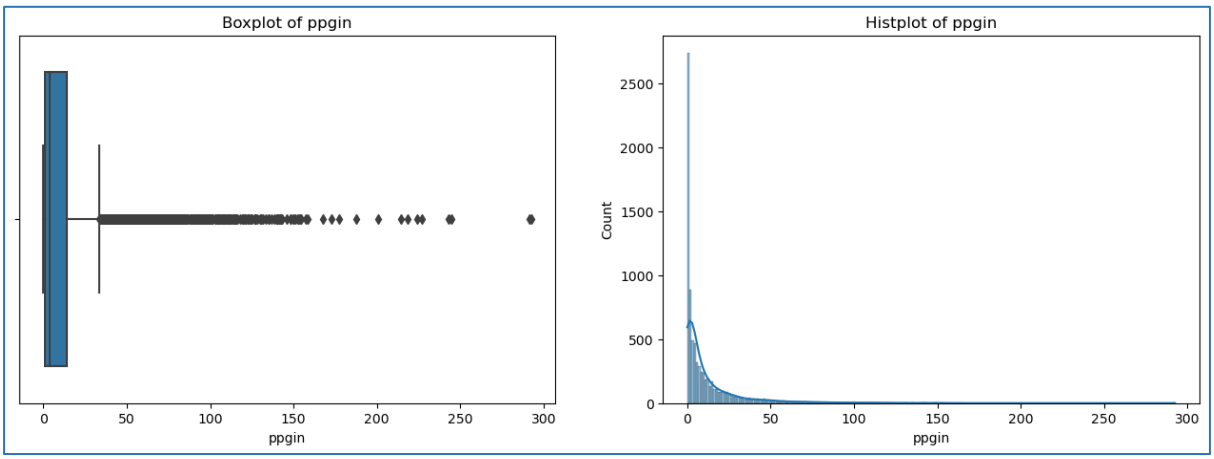
Minimum atch: 0.0
First quartile (Q1) of atch: 0.0
Median value: 0.0
Third quartile (Q3) of atch: 0.6
Maximum atch: 211.58

- Pgin



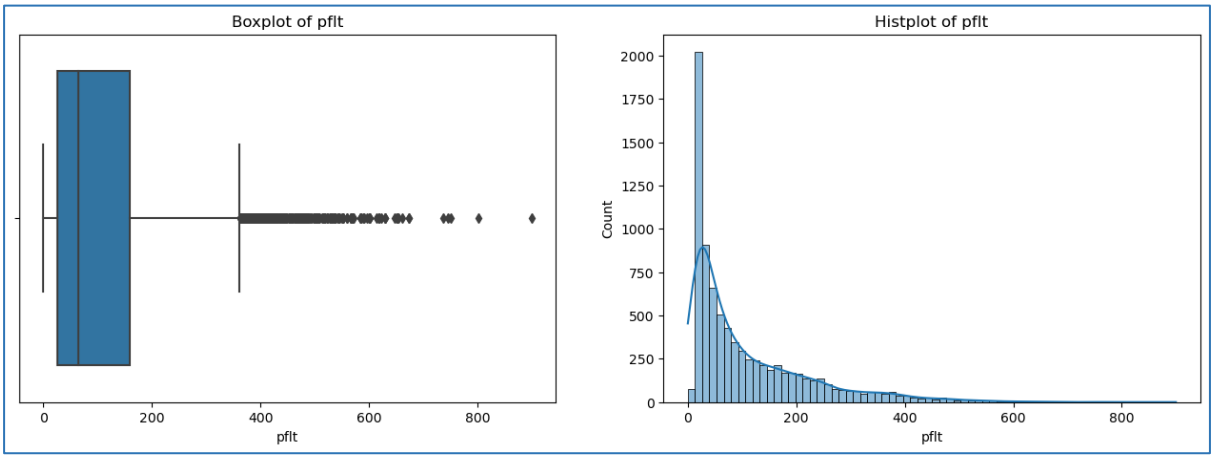
Minimum pgin: 0.0
First quartile (Q1) of pgin: 0.6
Median value: 2.8
Third quartile (Q3) of pgin: 9.765
Maximum pgin: 141.2

- Ppgin



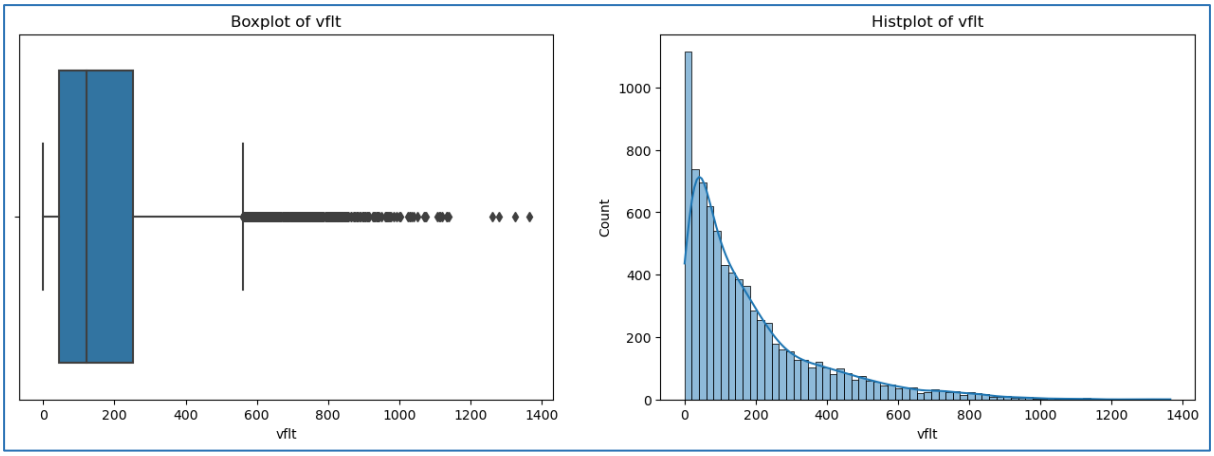
Minimum ppgin: 0.0
First quartile (Q1) of ppgin: 0.6
Median value: 3.8
Third quartile (Q3) of ppgin: 13.8
Maximum ppgin: 292.61

- Pflt



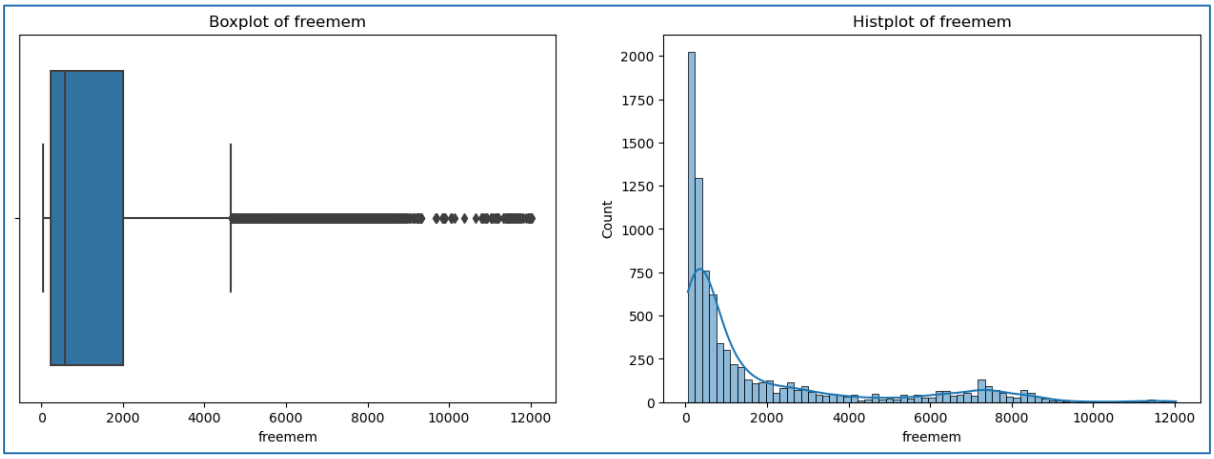
Minimum pflt: 0.0
First quartile (Q1) of pflt: 25.0
Median value: 63.8
Third quartile (Q3) of pflt: 159.6
Maximum pflt: 899.8

- Vflt



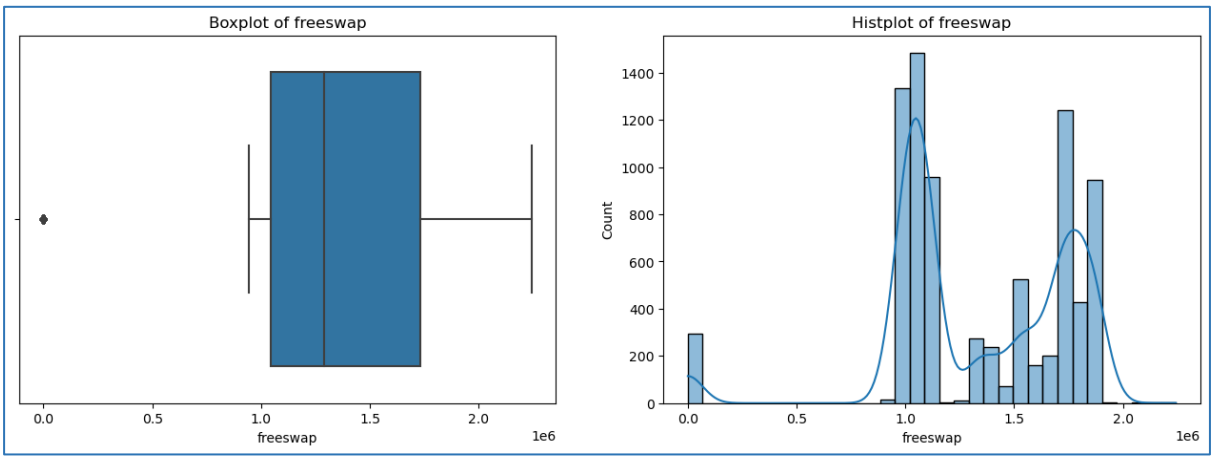
Minimum vflt: 0.2
First quartile (Q1) of vflt: 45.4
Median value: 120.4
Third quartile (Q3) of vflt: 251.8
Maximum vflt: 1365.0

- Freemem



Minimum freemem: 55
First quartile (Q1) of freemem: 231.0
Median value: 579.0
Third quartile (Q3) of freemem: 2002.25
Maximum freemem: 12027

- Freeswap



Minimum freeswap: 2
First quartile (Q1) of freeswap: 1042623.5
Median value: 1289289.5
Third quartile (Q3) of freeswap: 1730379.5
Maximum freeswap: 2243187

- Runqsz

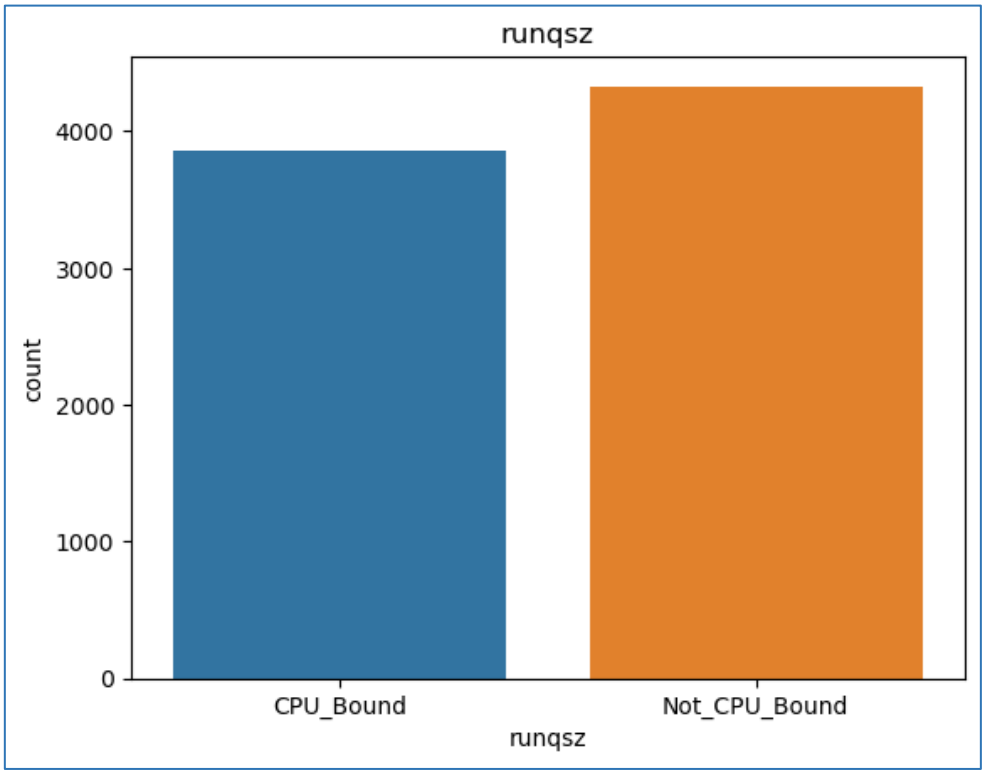


Figure 7: Boxplot, histplot and countplot of dependent variable

Multivariate analysis:

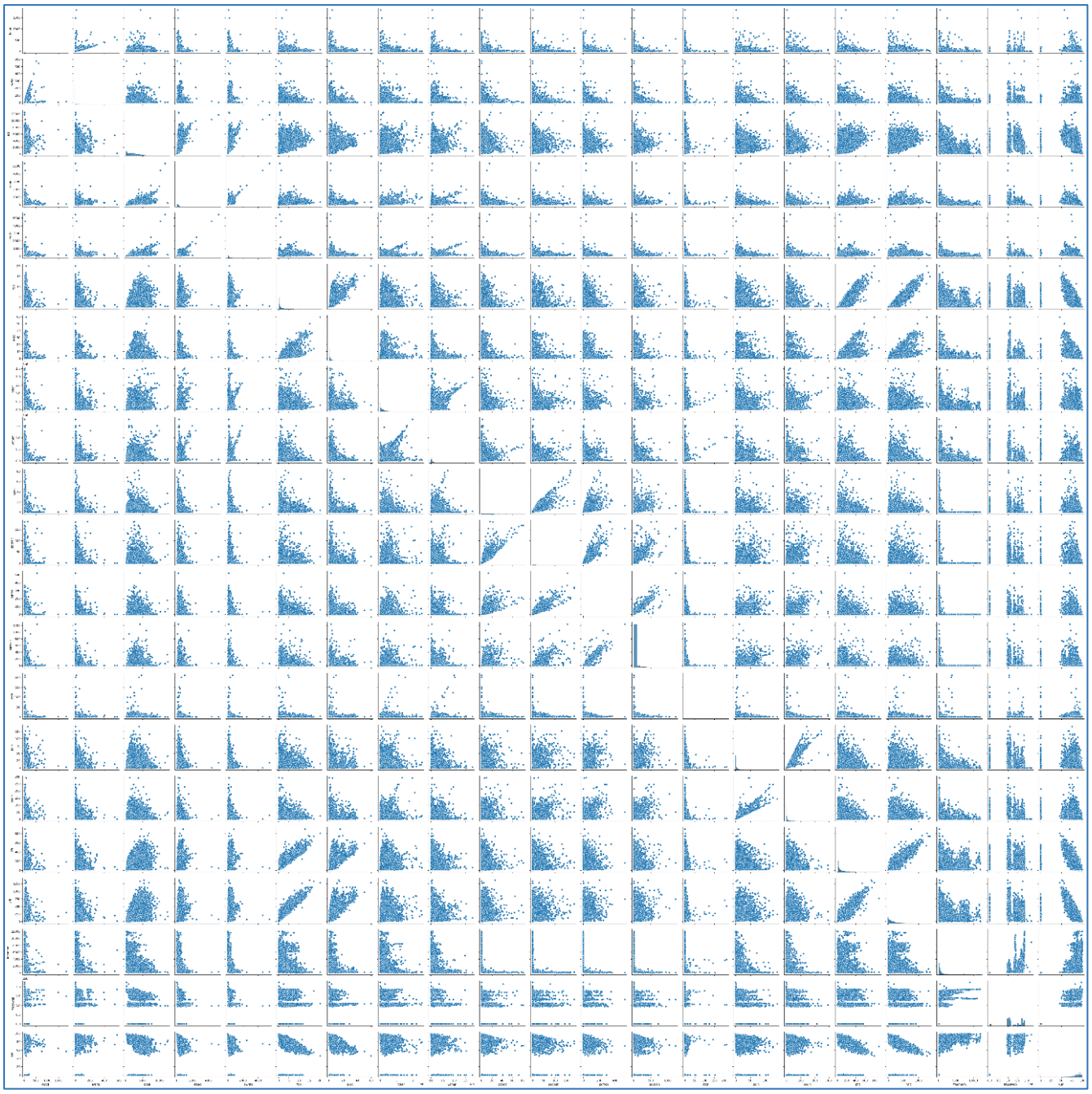


Figure 8: Pairplot

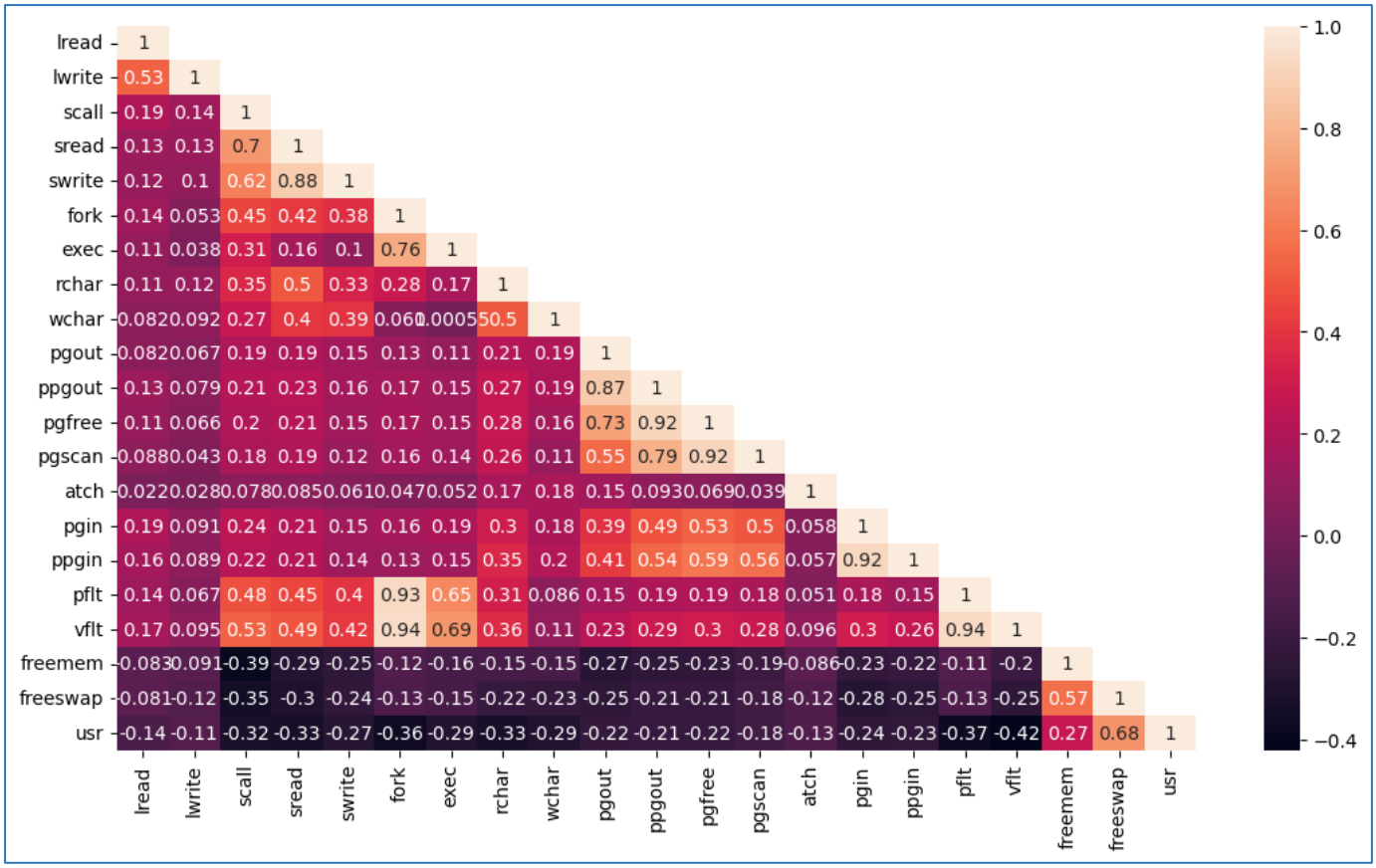


Figure 9: Heatmap

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.

```
lread          0
lwrite         0
scall          0
sread          0
swrite         0
fork           0
exec           0
rchar        104
wchar         15
pgout          0
ppgout         0
pgfree         0
pgscan         0
atch           0
pgin           0
ppgin          0
pflt           0
vflt           0
runqsz         0
freemem        0
freeswap       0
usr            0
dtype: int64
```

Figure 10: Null values

We see that there are null values in rchar and wchar. So we impute the values with median since it's a continuous variable.

```
lread       0
lwrite      0
scall       0
sread       0
swrite      0
fork        0
exec        0
rchar       0
wchar       0
pgout       0
ppgout      0
pgfree      0
pgscan      0
atch        0
pgin        0
ppgin       0
pflt        0
vflt        0
runqsz      0
freemem     0
freeswap    0
usr         0
rcha        0
dtype: int64
```

Figure 11: Null values after imputing

- There are following number of zeros in each variable.

```
Total number of 0s:
 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
runqsz          0
freemem         0
freeswap        0
usr           283
dtype: int64
```

Figure 12: Number of zeros in each variable

Its not necessary to remove these zeros since their number are very high and might cause an impact in our overall model.

- The above boxplots indicate the presence of outliers.
- There are no duplicates in the data.

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

We begin by separating the target and predictors. Then we create dummies.

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt
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
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
2	15	3	2162	159	119	2.0	2.4	125473.5	31950.0	0.0	...	0.0	1.2	6.0	9.4	150.20	220.20
3	0	0	160	12	16	0.2	0.2	125473.5	8670.0	0.0	...	0.0	0.0	0.2	0.2	15.60	16.80
4	5	1	330	39	38	0.4	0.4	125473.5	12185.0	0.0	...	0.0	0.0	1.0	1.2	37.80	47.60

5 rows × 22 columns

freemem	freeswap	constant	runqsz_Not_CPU_Bound
4670	1730946	1	0
7278	1869002	1	1
702	1021237	1	1
7248	1863704	1	1
633	1760253	1	1

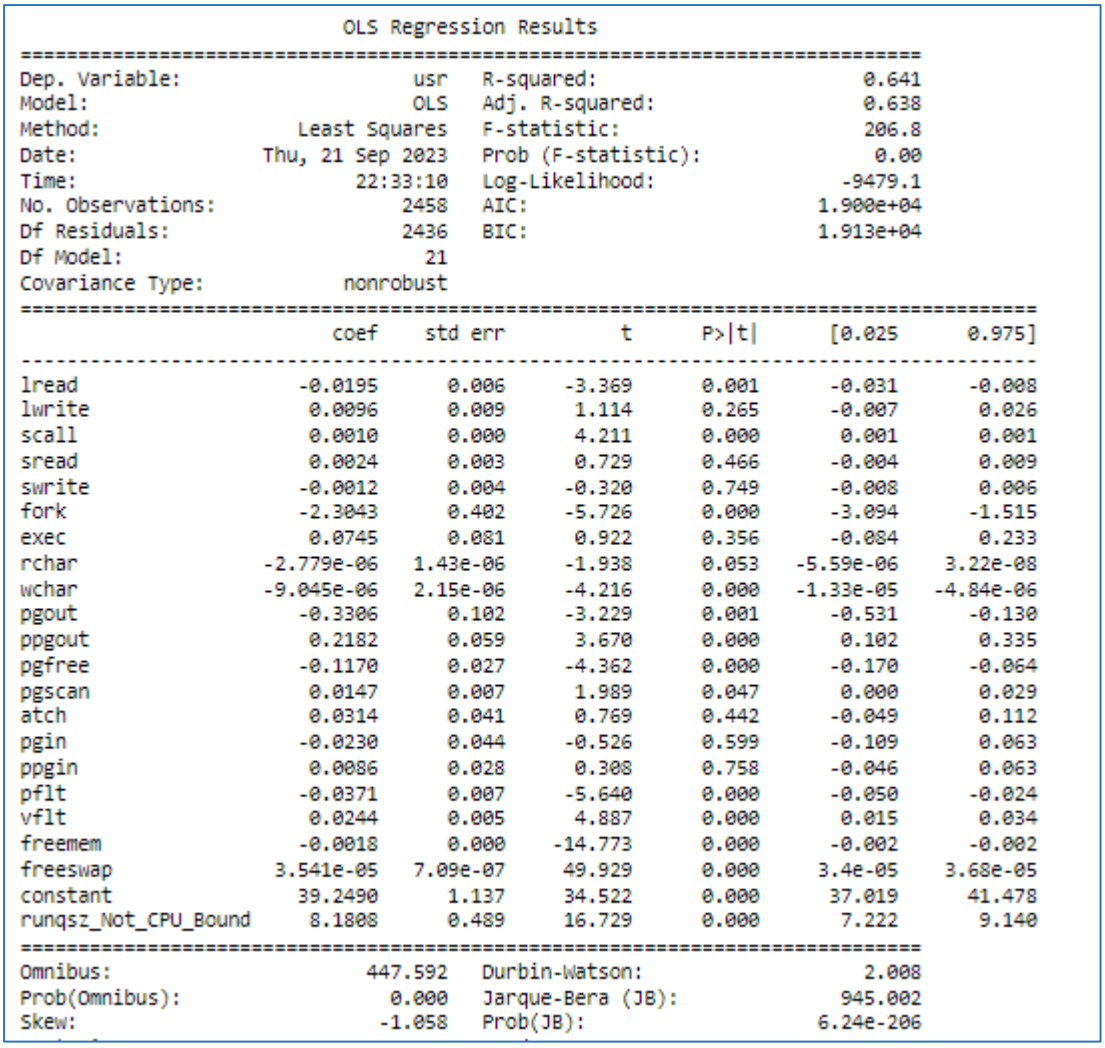
Figure 13: Dataset with dummy variables

We then split the data into train and test. The following is the shape of train an test data.

Shape of Training set : (5734, 22)
Shape of Test set : (2458, 22)

- MODEL BUILDING-:
  I. <u>MODEL-1</u>

  On applying linear regression and building a model, we obtain the following for the test data.



OLS Regression Results

Dep. Variable:	usr	R-squared:	0.641
Model:	OLS	Adj. R-squared:	0.638
Method:	Least Squares	F-statistic:	206.8
Date:	Thu, 21 Sep 2023	Prob (F-statistic):	0.00
Time:	22:33:10	Log-Likelihood:	-9479.1
No. Observations:	2458	AIC:	1.900e+04
Df Residuals:	2436	BIC:	1.913e+04
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
lread	-0.0195	0.006	-3.369	0.001	-0.031	-0.008
lwrite	0.0096	0.009	1.114	0.265	-0.007	0.026
scall	0.0010	0.000	4.211	0.000	0.001	0.001
sread	0.0024	0.003	0.729	0.466	-0.004	0.009
swrite	-0.0012	0.004	-0.320	0.749	-0.008	0.006
fork	-2.3043	0.402	-5.726	0.000	-3.094	-1.515
exec	0.0745	0.081	0.922	0.356	-0.084	0.233
rchar	-2.779e-06	1.43e-06	-1.938	0.053	-5.59e-06	3.22e-08
wchar	-9.045e-06	2.15e-06	-4.216	0.000	-1.33e-05	-4.84e-06
pgout	-0.3306	0.102	-3.229	0.001	-0.531	-0.130
ppgout	0.2182	0.059	3.670	0.000	0.102	0.335
pgfree	-0.1170	0.027	-4.362	0.000	-0.170	-0.064
pgscan	0.0147	0.007	1.989	0.047	0.000	0.029
atch	0.0314	0.041	0.769	0.442	-0.049	0.112
pgin	-0.0230	0.044	-0.526	0.599	-0.109	0.063
ppgin	0.0086	0.028	0.308	0.758	-0.046	0.063
pflt	-0.0371	0.007	-5.640	0.000	-0.050	-0.024
vflt	0.0244	0.005	4.887	0.000	0.015	0.034
freemem	-0.0018	0.000	-14.773	0.000	-0.002	-0.002
freeswap	3.541e-05	7.09e-07	49.929	0.000	3.4e-05	3.68e-05
constant	39.2490	1.137	34.522	0.000	37.019	41.478
runqsz_Not_CPU_Bound	8.1808	0.489	16.729	0.000	7.222	9.140

Omnibus:	447.592	Durbin-Watson:	2.008
Prob(Omnibus):	0.000	Jarque-Bera (JB):	945.002
Skew:	-1.058	Prob(JB):	6.24e-206

Figure 14: MODEL-1

Model performance of:
Training data –

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	10.813214	7.768781	0.64284	0.641464	inf

Test data-

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	11.594014	8.171394	0.631217	0.627885	inf

It shows that there is 63% of observed variance which can be explained by these variables.

Now we check the multicollinearity using VIF (Variance Inflation Factor)

	feature	VIF
0	lread	1.472618
1	lwrite	1.405898
2	scall	2.414301
3	sread	6.836403
4	swrite	5.320692
5	fork	18.210503
6	exec	3.059950
7	rchar	1.974726
8	wchar	1.553348
9	pgout	5.776005
10	ppgout	15.906900
11	pgfree	20.437584
12	pgscan	9.237017
13	atch	1.087328
14	pgin	8.075699
15	ppgin	8.672927
16	pflt	11.834374
17	vflt	20.233207
18	freemem	1.677241
19	freeswap	1.761193
20	constant	27.191591
21	runqsz_Not_CPU_Bound	1.118922

Figure 15: VIF of MODEL-1

Variables which have correlation are not necessary as they inflate the standard errors and correspondingly affect out regression parameters. As a result, we remove such variables.
VIF values less than 5 indiacate no to moderate multicollinearity. So we try to remove variables one by one that are more than 5 and re-run the model to check if there is any improvement in our R-squared value on the test data.

Residuals-

	Actual Values	Fitted Values	Residuals
694	91	83.397386	7.602614
5535	94	83.576344	10.423656
4244	0	42.971134	-42.971134
2472	83	72.705492	10.294508
7052	94	102.942178	-8.942178

Figure 16: Residuals of MODEL-1

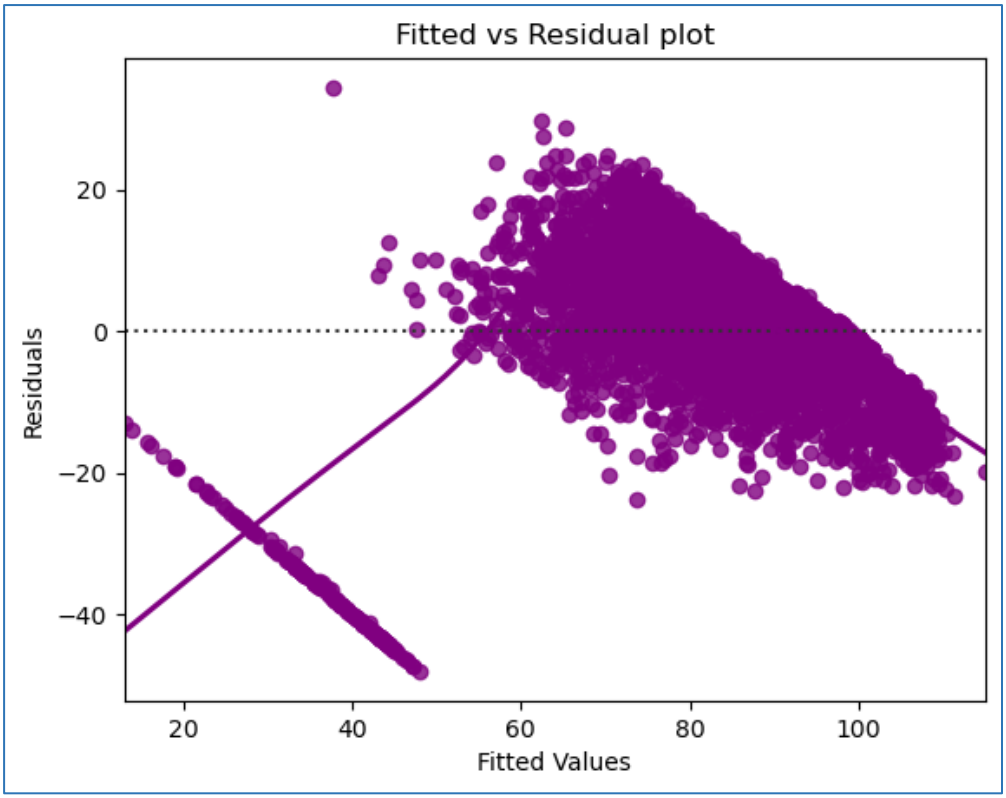


Figure 17: Fitted values with residuals of MODEL-1

Predicted values on test data-

	Actual	Predicted
3904	75	75.000585
2603	78	65.071669
8130	95	84.723177
3100	74	71.089485
2726	97	75.281332
717	89	98.263248
847	90	92.083188
2652	92	96.309702
1198	84	77.439180
6284	86	80.325140

Figure 18: Actual and predicted values of model 1

II. MODEL-2

We treat the outliers and run the model to see if there is any improvement in the R-squared value. IQR method is used to treat outliers.

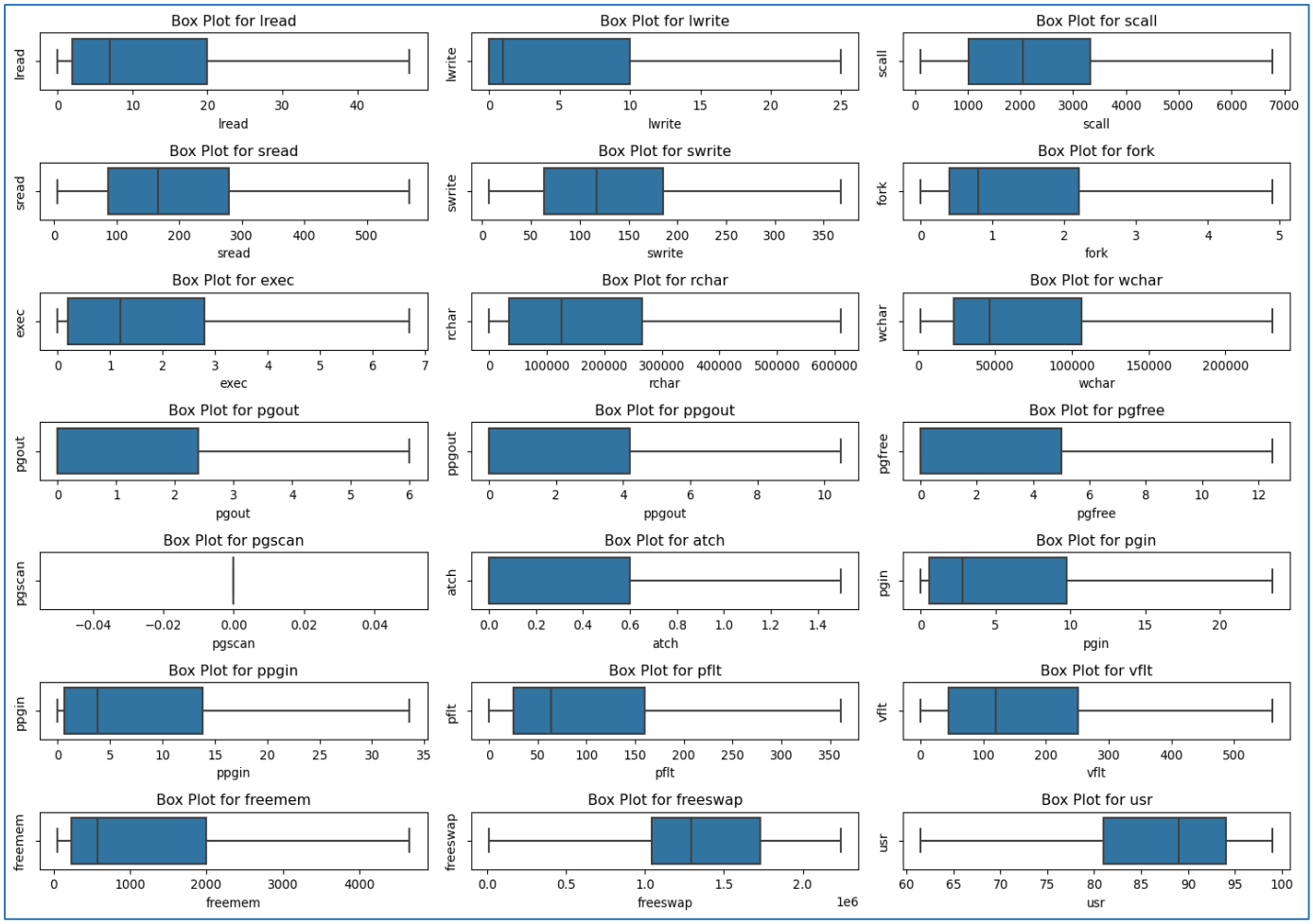
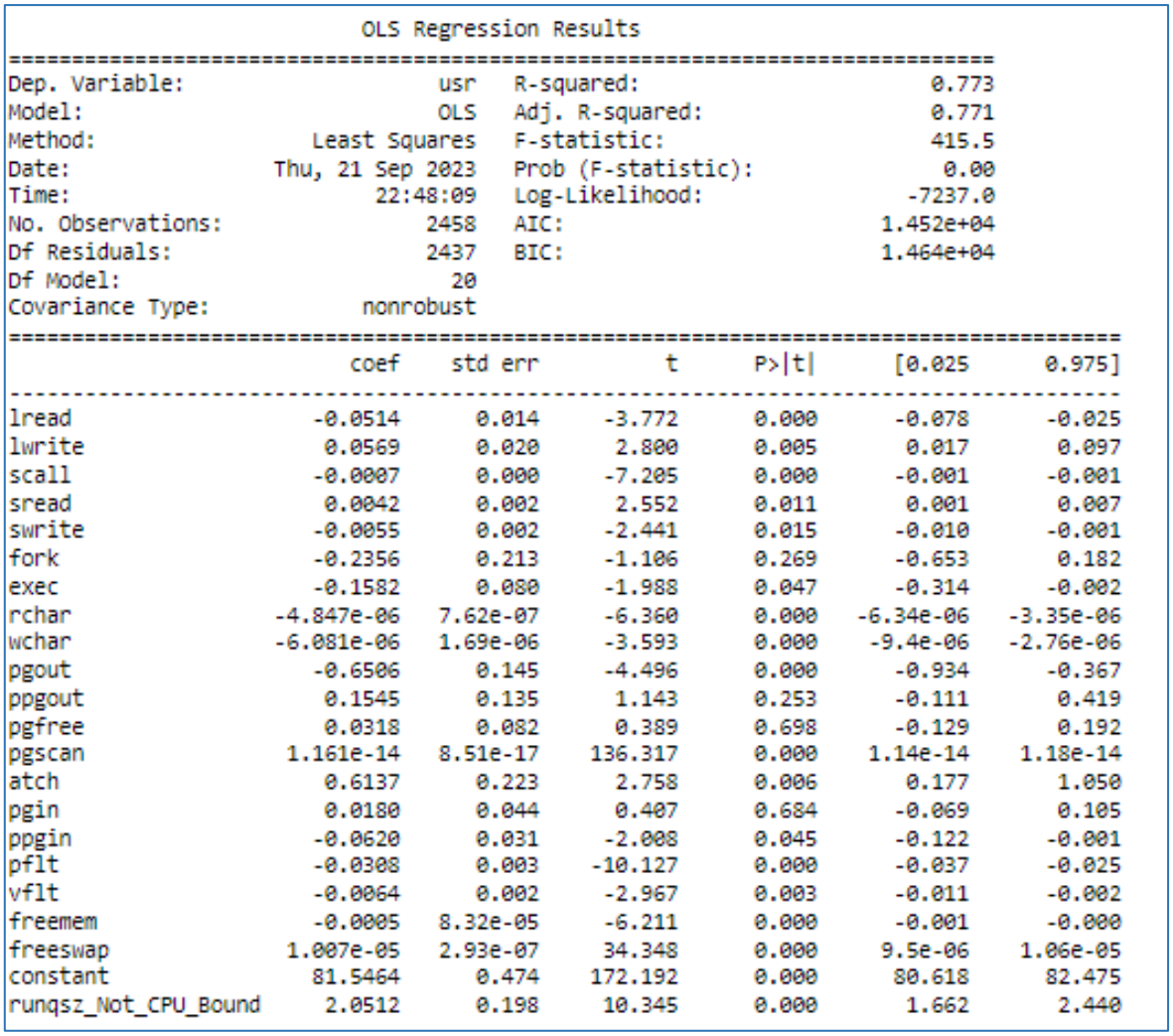


Figure 19: Boxplots after treating outliers



OLS Regression Results

Dep. Variable:	usr	R-squared:	0.773
Model:	OLS	Adj. R-squared:	0.771
Method:	Least Squares	F-statistic:	415.5
Date:	Thu, 21 Sep 2023	Prob (F-statistic):	0.00
Time:	22:48:09	Log-Likelihood:	-7237.0
No. Observations:	2458	AIC:	1.452e+04
Df Residuals:	2437	BIC:	1.464e+04
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
lread	-0.0514	0.014	-3.772	0.000	-0.078	-0.025
lwrite	0.0569	0.020	2.800	0.005	0.017	0.097
scall	-0.0007	0.000	-7.205	0.000	-0.001	-0.001
sread	0.0042	0.002	2.552	0.011	0.001	0.007
swrite	-0.0055	0.002	-2.441	0.015	-0.010	-0.001
fork	-0.2356	0.213	-1.106	0.269	-0.653	0.182
exec	-0.1582	0.080	-1.988	0.047	-0.314	-0.002
rchar	-4.847e-06	7.62e-07	-6.360	0.000	-6.34e-06	-3.35e-06
wchar	-6.081e-06	1.69e-06	-3.593	0.000	-9.4e-06	-2.76e-06
pgout	-0.6506	0.145	-4.496	0.000	-0.934	-0.367
ppgout	0.1545	0.135	1.143	0.253	-0.111	0.419
pgfree	0.0318	0.082	0.389	0.698	-0.129	0.192
pgscan	1.161e-14	8.51e-17	136.317	0.000	1.14e-14	1.18e-14
atch	0.6137	0.223	2.758	0.006	0.177	1.050
pgin	0.0180	0.044	0.407	0.684	-0.069	0.105
ppgin	-0.0620	0.031	-2.008	0.045	-0.122	-0.001
pflt	-0.0308	0.003	-10.127	0.000	-0.037	-0.025
vflt	-0.0064	0.002	-2.967	0.003	-0.011	-0.002
freemem	-0.0005	8.32e-05	-6.211	0.000	-0.001	-0.000
freeswap	1.007e-05	2.93e-07	34.348	0.000	9.5e-06	1.06e-05
constant	81.5464	0.474	172.192	0.000	80.618	82.475
runqsz_Not_CPU_Bound	2.0512	0.198	10.345	0.000	1.662	2.440

Figure 20: MODEL-2

We notice that on treating outliers, the R-square value has drastically improved to 77.3% which is a good sign.

Model performance of:
Training data –

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.419536	3.285268	0.796109	0.795323	4.07469

Test data-

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.596708	3.447267	0.773249	0.771201	4.270487

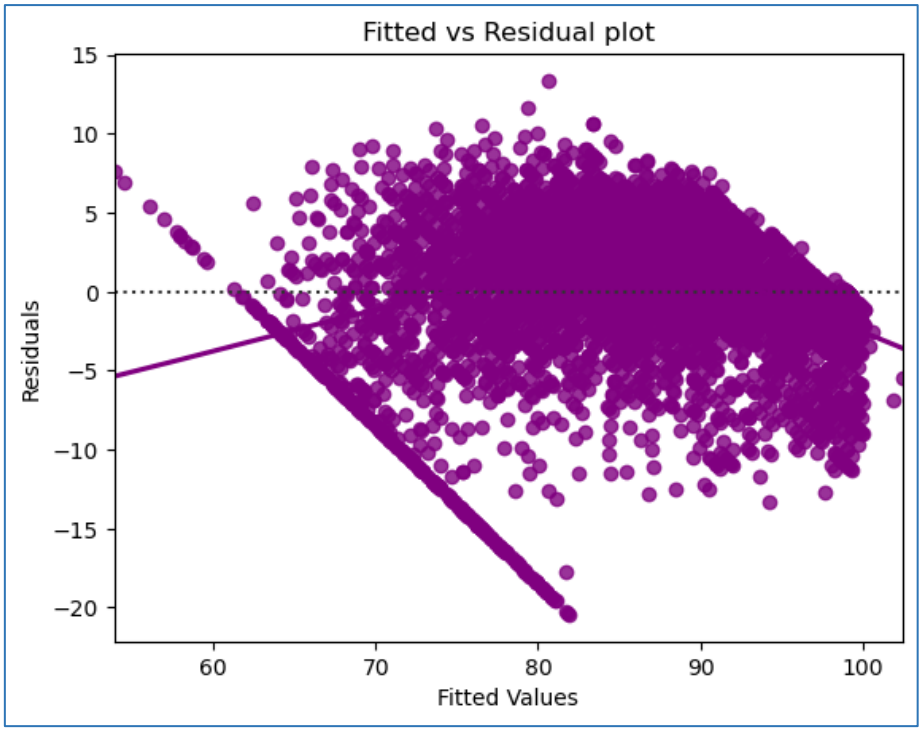
VIF values-

	feature	VIF
0	lread	5.350560
1	lwrite	4.328397
2	scall	2.960609
3	sread	6.420172
4	swrite	5.597135
5	fork	13.035359
6	exec	3.241417
7	rchar	2.133616
8	wchar	1.584381
9	pgout	11.360363
10	ppgout	29.404223
11	pgfree	16.496748
12	pgscan	NaN
13	atch	1.875901
14	pgin	13.809339
15	ppgin	13.951855
16	pflt	12.001460
17	vflt	15.971049
18	freemem	1.961304
19	freeswap	1.841239
20	constant	29.229332
21	runqsz_Not_CPU_Bound	1.156815

Figure 20: VIF of MODEL-2

Residuals-

	Actual Values	Fitted Values	Residuals
694	91.0	91.507801	-0.507801
5535	94.0	91.778831	2.221169
4244	61.5	74.855263	-13.355263
2472	83.0	80.729524	2.270476
7052	94.0	98.268952	-4.268952



Predicted values on test data-

	Actual	Predicted
3904	75.0	78.226652
2603	78.0	76.059766
8130	95.0	91.685942
3100	74.0	72.067431
2726	97.0	90.564449
717	89.0	92.847531
847	90.0	91.087303
2652	92.0	91.736531
1198	84.0	82.296006
6284	86.0	81.256856

We now systematically drop the numerical columns with VIF>5. We will ignore the VIF values for dummy variables and the constant(intercept).

III. MODEL-3

We drop the column 'pgfree' whose VIF=16.4

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.773
Model:	OLS	Adj. R-squared:	0.771
Method:	Least Squares	F-statistic:	437.5
Date:	Thu, 21 Sep 2023	Prob (F-statistic):	0.00
Time:	23:14:43	Log-Likelihood:	-7237.1
No. Observations:	2458	AIC:	1.451e+04
Df Residuals:	2438	BIC:	1.463e+04
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	81.5493	0.473	172.251	0.000	80.621	82.478
lread	-0.0517	0.014	-3.797	0.000	-0.078	-0.025
lwrite	0.0571	0.020	2.811	0.005	0.017	0.097
scall	-0.0007	0.000	-7.202	0.000	-0.001	-0.001
sread	0.0043	0.002	2.568	0.010	0.001	0.008
swrite	-0.0056	0.002	-2.452	0.014	-0.010	-0.001
fork	-0.2359	0.213	-1.108	0.268	-0.654	0.182
exec	-0.1571	0.080	-1.976	0.048	-0.313	-0.001
rchar	-4.852e-06	7.62e-07	-6.367	0.000	-6.35e-06	-3.36e-06
wchar	-6.127e-06	1.69e-06	-3.630	0.000	-9.44e-06	-2.82e-06
pgout	-0.6629	0.141	-4.696	0.000	-0.940	-0.386
ppgout	0.1978	0.076	2.592	0.010	0.048	0.347
pgscan	-5.479e-15	1.66e-16	-33.070	0.000	-5.8e-15	-5.15e-15
atch	0.6175	0.222	2.778	0.006	0.182	1.053
pgin	0.0174	0.044	0.393	0.695	-0.069	0.104
ppgin	-0.0611	0.031	-1.985	0.047	-0.121	-0.001
pflt	-0.0308	0.003	-10.132	0.000	-0.037	-0.025
vflt	-0.0064	0.002	-2.964	0.003	-0.011	-0.002
freemem	-0.0005	8.29e-05	-6.257	0.000	-0.001	-0.000
freeswap	1.007e-05	2.93e-07	34.373	0.000	9.5e-06	1.06e-05
runqsz_Not_CPU_Bound	2.0524	0.198	10.354	0.000	1.664	2.441

Figure 21: MODEL-3

The R-square value remains same i.e- 77.3% indicating that dropping this column does not affect our column much.

Model performance of:
Training data –

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.420747	3.286002	0.795997	0.795247	4.075769

Test data-

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.59685	3.447211	0.773235	0.77128	4.270533

VIF values-

	feature	VIF
0	const	29.104453
1	lread	5.350249
2	lwrite	4.327383
3	scall	2.958105
4	sread	6.420154
5	swrite	5.597127
6	fork	13.034330
7	exec	3.241322
8	rchar	2.133610
9	wchar	1.581770
10	pgout	11.140354
11	ppgout	11.002640
12	pgscan	NaN
13	atch	1.874864
14	pgin	13.809241
15	ppgin	13.950674
16	pflt	11.994236
17	vflt	15.961041
18	freemem	1.951034
19	freeswap	1.840999
20	runqsz_Not_CPU_Bound	1.156119

Figure 22: VIF of MODEL-3

IV. MODEL-4

We drop the column 'vflt' whose VIF=15.9

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.772
Model:	OLS	Adj. R-squared:	0.771
Method:	Least Squares	F-statistic:	459.9
Date:	Thu, 21 Sep 2023	Prob (F-statistic):	0.00
Time:	23:31:41	Log-Likelihood:	-7241.5
No. Observations:	2458	AIC:	1.452e+04
Df Residuals:	2439	BIC:	1.463e+04
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	81.3481	0.469	173.344	0.000	80.428	82.268
lread	-0.0536	0.014	-3.939	0.000	-0.080	-0.027
lwrite	0.0580	0.020	2.852	0.004	0.018	0.098
scall	-0.0008	0.000	-7.393	0.000	-0.001	-0.001
sread	0.0039	0.002	2.371	0.018	0.001	0.007
swrite	-0.0055	0.002	-2.423	0.015	-0.010	-0.001
fork	-0.5452	0.186	-2.932	0.003	-0.910	-0.181
exec	-0.1647	0.080	-2.069	0.039	-0.321	-0.009
rchar	-4.949e-06	7.62e-07	-6.490	0.000	-6.44e-06	-3.45e-06
wchar	-5.654e-06	1.68e-06	-3.360	0.001	-8.95e-06	-2.35e-06
pgout	-0.6475	0.141	-4.583	0.000	-0.925	-0.370
ppgout	0.1861	0.076	2.438	0.015	0.036	0.336
pgscan	-3.381e-15	1.11e-16	-30.487	0.000	-3.6e-15	-3.16e-15
atch	0.5723	0.222	2.577	0.010	0.137	1.008
pgin	0.0058	0.044	0.132	0.895	-0.081	0.092
ppgin	-0.0642	0.031	-2.085	0.037	-0.125	-0.004
pflt	-0.0347	0.003	-12.599	0.000	-0.040	-0.029
freemem	-0.0005	8.3e-05	-6.351	0.000	-0.001	-0.000
freeswap	1.025e-05	2.87e-07	35.692	0.000	9.69e-06	1.08e-05
runqsz_Not_CPU_Bound	2.0430	0.199	10.291	0.000	1.654	2.432

Figure 23: VIF of MODEL-4

The R-square value is now 77.2% showing no much affect on the model by removing vflt as well.

Model performance of:
Training data –

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.426288	3.291979	0.795485	0.794769	4.08224

Test data-

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.605125	3.454343	0.772418	0.77055	4.278755

VIF values-

	feature	VIF
0	const	28.703900
1	lread	5.335255
2	lwrite	4.326141
3	scall	2.950927
4	sread	6.374687
5	swrite	5.595871
6	fork	10.087914
7	exec	3.237609
8	rchar	2.123783
9	wchar	1.559124
10	pgout	11.129401
11	ppgout	10.964532
12	pgscan	NaN
13	atch	1.863323
14	pgin	13.624884
15	ppgin	13.950655
16	pflt	9.130725
17	freemem	1.949957
18	freeswap	1.789929
19	runqsz_Not_CPU_Bound	1.156075

Figure 24: VIF of MODEL-4

V. MODEL-5

We drop the column 'pgin' whose VIF=13.6

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.772
Model:	OLS	Adj. R-squared:	0.771
Method:	Least Squares	F-statistic:	487.1
Date:	Thu, 21 Sep 2023	Prob (F-statistic):	0.00
Time:	23:40:39	Log-Likelihood:	-7241.5
No. Observations:	2458	AIC:	1.452e+04
Df Residuals:	2440	BIC:	1.462e+04
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	81.3549	0.466	174.471	0.000	80.441	82.269
lread	-0.0537	0.014	-3.946	0.000	-0.080	-0.027
lwrite	0.0581	0.020	2.856	0.004	0.018	0.098
scall	-0.0008	0.000	-7.394	0.000	-0.001	-0.001
sread	0.0039	0.002	2.378	0.017	0.001	0.007
swrite	-0.0055	0.002	-2.430	0.015	-0.010	-0.001
fork	-0.5439	0.186	-2.929	0.003	-0.908	-0.180
exec	-0.1643	0.080	-2.065	0.039	-0.320	-0.008
rchar	-4.952e-06	7.62e-07	-6.498	0.000	-6.45e-06	-3.46e-06
wchar	-5.657e-06	1.68e-06	-3.362	0.001	-8.96e-06	-2.36e-06
pgout	-0.6461	0.141	-4.588	0.000	-0.922	-0.370
ppgout	0.1852	0.076	2.437	0.015	0.036	0.334
pgscan	-5.699e-15	4.79e-17	-118.890	0.000	-5.79e-15	-5.61e-15
atch	0.5714	0.222	2.574	0.010	0.136	1.007
ppgin	-0.0604	0.010	-5.768	0.000	-0.081	-0.040
pflt	-0.0347	0.003	-12.620	0.000	-0.040	-0.029
freemem	-0.0005	8.3e-05	-6.355	0.000	-0.001	-0.000
freeswap	1.025e-05	2.86e-07	35.845	0.000	9.69e-06	1.08e-05
runqsz_Not_CPU_Bound	2.0431	0.198	10.294	0.000	1.654	2.432

Figure 25: MODEL-5

The R-square value remains 77.2% after removing 'pgin'.

Model performance of:

Training data –

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.426316	3.291941	0.795483	0.794802	4.082119

Test data-

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.605142	3.454195	0.772416	0.770643	4.278554

VIF values-

	feature	VIF
0	const	28.477684
1	lread	5.333765
2	lwrite	4.328131
3	scall	2.948289
4	sread	6.374589
5	swrite	5.595550
6	fork	10.082898
7	exec	3.234015
8	rchar	2.114410
9	wchar	1.558886
10	pgout	11.098156
11	ppgout	10.924686
12	pgscan	NaN
13	atch	1.882989
14	ppgin	1.588837
15	pflt	9.129997
16	freemem	1.949936
17	freeswap	1.777319
18	runqsz_Not_CPU_Bound	1.156013

Figure 26: VIF of MODEL-5

VI. MODEL-6

We drop the column 'pgout' whose VIF=11

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.770
Model:	OLS	Adj. R-squared:	0.769
Method:	Least Squares	F-statistic:	512.1
Date:	Thu, 21 Sep 2023	Prob (F-statistic):	0.00
Time:	23:41:38	Log-Likelihood:	-7252.1
No. Observations:	2458	AIC:	1.454e+04
Df Residuals:	2441	BIC:	1.464e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	81.2260	0.467	173.800	0.000	80.310	82.142
lread	-0.0500	0.014	-3.666	0.000	-0.077	-0.023
lwrite	0.0531	0.020	2.606	0.009	0.013	0.093
scall	-0.0008	0.000	-7.368	0.000	-0.001	-0.001
sread	0.0044	0.002	2.622	0.009	0.001	0.008
swrite	-0.0058	0.002	-2.558	0.011	-0.010	-0.001
fork	-0.5065	0.186	-2.719	0.007	-0.872	-0.141
exec	-0.1583	0.080	-1.982	0.048	-0.315	-0.002
rchar	-4.956e-06	7.65e-07	-6.478	0.000	-6.46e-06	-3.46e-06
wchar	-5.813e-06	1.69e-06	-3.442	0.001	-9.13e-06	-2.5e-06
ppgout	-0.1306	0.032	-4.037	0.000	-0.194	-0.067
pgscan	-2.627e-14	2.26e-16	-116.339	0.000	-2.67e-14	-2.58e-14
atch	0.3034	0.215	1.411	0.158	-0.118	0.725
ppgin	-0.0589	0.011	-5.605	0.000	-0.080	-0.038
pflt	-0.0354	0.003	-12.849	0.000	-0.041	-0.030
freemem	-0.0005	8.33e-05	-6.171	0.000	-0.001	-0.000
freeswap	1.03e-05	2.87e-07	35.903	0.000	9.74e-06	1.09e-05
runqsz_Not_CPU_Bound	2.0212	0.199	10.145	0.000	1.630	2.412

Figure 27: MODEL-6

The R-square value has now become 77% after removing 'pgout' which is not much difference.

Model performance of:

Training data –

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.433326	3.299085	0.794834	0.794188	4.091236

Test data-

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.624959	3.465708	0.770453	0.768759	4.29466

VIF values-

	feature	VIF
0	const	28.380737
1	lread	5.331247
2	lwrite	4.324757
3	scall	2.947447
4	sread	6.374433
5	swrite	5.594045
6	fork	10.082410
7	exec	3.229571
8	rchar	2.113240
9	wchar	1.558201
10	ppgout	2.015936
11	pgscan	NaN
12	atch	1.784355
13	ppgin	1.564118
14	pflt	9.129145
15	freemem	1.944308
16	freeswap	1.775425
17	runqsz_Not_CPU_Bound	1.154502

Figure 28: VIF of MODEL-6

VII. MODEL-7

We drop the column 'fork' whose VIF=10

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.770
Model:	OLS	Adj. R-squared:	0.768
Method:	Least Squares	F-statistic:	544.3
Date:	Thu, 21 Sep 2023	Prob (F-statistic):	0.00
Time:	23:46:46	Log-Likelihood:	-7255.8
No. Observations:	2458	AIC:	1.454e+04
Df Residuals:	2442	BIC:	1.464e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	81.2986	0.467	174.012	0.000	80.382	82.215
lread	-0.0527	0.014	-3.874	0.000	-0.079	-0.026
lwrite	0.0578	0.020	2.840	0.005	0.018	0.098
scall	-0.0007	0.000	-7.040	0.000	-0.001	-0.001
sread	0.0043	0.002	2.590	0.010	0.001	0.008
swrite	-0.0072	0.002	-3.236	0.001	-0.012	-0.003
exec	-0.2306	0.075	-3.060	0.002	-0.378	-0.083
rchar	-5.031e-06	7.66e-07	-6.571	0.000	-6.53e-06	-3.53e-06
wchar	-5.3e-06	1.68e-06	-3.153	0.002	-8.6e-06	-2e-06
ppgout	-0.1283	0.032	-3.963	0.000	-0.192	-0.065
pgscan	1.47e-13	8.39e-16	175.170	0.000	1.45e-13	1.49e-13
atch	0.3123	0.215	1.451	0.147	-0.110	0.734
ppgin	-0.0580	0.011	-5.512	0.000	-0.079	-0.037
pflt	-0.0413	0.002	-24.494	0.000	-0.045	-0.038
freemem	-0.0005	8.33e-05	-6.269	0.000	-0.001	-0.000
freeswap	1.029e-05	2.87e-07	35.817	0.000	9.72e-06	1.09e-05
runqsz_Not_CPU_Bound	2.0292	0.199	10.173	0.000	1.638	2.420

Figure 29: MODEL-7

The R-square value remains to be 77% after removing 'fork'

Model performance of:

Training data –

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.434575	3.298056	0.794719	0.794108	4.08879

Test data-

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.631958	3.466697	0.769758	0.768154	4.292293

VIF values-

	feature	VIF
0	const	28.229904
1	lread	5.310527
2	lwrite	4.305161
3	scall	2.910216
4	sread	6.373324
5	swrite	5.388096
6	exec	2.849598
7	rchar	2.111886
8	wchar	1.550023
9	ppgout	2.015794
10	pgscan	NaN
11	atch	1.783484
12	ppgin	1.560581
13	pflt	3.461241
14	freemem	1.943876
15	freeswap	1.772908
16	runqsz_Not_CPU_Bound	1.153933

Figure 30: VIF of MODEL-7

VIII. MODEL8-

We drop the column 'pgscan'

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.770
Model:	OLS	Adj. R-squared:	0.768
Method:	Least Squares	F-statistic:	544.3
Date:	Thu, 21 Sep 2023	Prob (F-statistic):	0.00
Time:	23:51:00	Log-Likelihood:	-7255.8
No. Observations:	2458	AIC:	1.454e+04
Df Residuals:	2442	BIC:	1.464e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	81.2986	0.467	174.012	0.000	80.382	82.215
lread	-0.0527	0.014	-3.874	0.000	-0.079	-0.026
lwrite	0.0578	0.020	2.840	0.005	0.018	0.098
scall	-0.0007	0.000	-7.040	0.000	-0.001	-0.001
sread	0.0043	0.002	2.590	0.010	0.001	0.008
swrite	-0.0072	0.002	-3.236	0.001	-0.012	-0.003
exec	-0.2306	0.075	-3.060	0.002	-0.378	-0.083
rchar	-5.031e-06	7.66e-07	-6.571	0.000	-6.53e-06	-3.53e-06
wchar	-5.3e-06	1.68e-06	-3.153	0.002	-8.6e-06	-2e-06
ppgout	-0.1283	0.032	-3.963	0.000	-0.192	-0.065
atch	0.3123	0.215	1.451	0.147	-0.110	0.734
ppgin	-0.0580	0.011	-5.512	0.000	-0.079	-0.037
pflt	-0.0413	0.002	-24.494	0.000	-0.045	-0.038
freemem	-0.0005	8.33e-05	-6.269	0.000	-0.001	-0.000
freeswap	1.029e-05	2.87e-07	35.817	0.000	9.72e-06	1.09e-05
runqsz_Not_CPU_Bound	2.0292	0.199	10.173	0.000	1.638	2.420

Figure 31: MODEL-8

The R-square value remains to be 77% after removing 'pgscan'

Model performance of:

Training data –

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.434575	3.298056	0.794719	0.794144	4.08879

Test data-

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.631958	3.466697	0.769758	0.768249	4.292293

VIF values-

	feature	VIF
0	const	28.229904
1	lread	5.310527
2	lwrite	4.305161
3	scall	2.910216
4	sread	6.373324
5	swrite	5.388096
6	exec	2.849598
7	rchar	2.111886
8	wchar	1.550023
9	ppgout	2.015794
10	atch	1.783484
11	ppgin	1.560581
12	pflt	3.461241
13	freemem	1.943876
14	freeswap	1.772908
15	runqsz_Not_CPU_Bound	1.153933

Figure 32: VIF of MODEL-8

IX. MODEL9-

We drop the column 'sread' with VIF=6.37

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.769
Model:	OLS	Adj. R-squared:	0.768
Method:	Least Squares	F-statistic:	581.3
Date:	Fri, 22 Sep 2023	Prob (F-statistic):	0.00
Time:	00:22:33	Log-Likelihood:	-7259.2
No. Observations:	2458	AIC:	1.455e+04
Df Residuals:	2443	BIC:	1.464e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	81.3561	0.467	174.129	0.000	80.440	82.272
lread	-0.0547	0.014	-4.023	0.000	-0.081	-0.028
lwrite	0.0607	0.020	2.987	0.003	0.021	0.101
scall	-0.0006	9.76e-05	-6.544	0.000	-0.001	-0.000
swrite	-0.0032	0.002	-2.002	0.045	-0.006	-6.66e-05
exec	-0.2369	0.075	-3.141	0.002	-0.385	-0.089
rchar	-4.207e-06	6.97e-07	-6.035	0.000	-5.57e-06	-2.84e-06
wchar	-5.682e-06	1.68e-06	-3.390	0.001	-8.97e-06	-2.4e-06
ppgout	-0.1238	0.032	-3.824	0.000	-0.187	-0.060
atch	0.2880	0.215	1.338	0.181	-0.134	0.710
ppgin	-0.0586	0.011	-5.566	0.000	-0.079	-0.038
pflt	-0.0410	0.002	-24.334	0.000	-0.044	-0.038
freemem	-0.0005	8.34e-05	-6.269	0.000	-0.001	-0.000
freeswap	1.022e-05	2.86e-07	35.692	0.000	9.65e-06	1.08e-05
runqsz_Not_CPU_Bound	2.0520	0.200	10.285	0.000	1.661	2.443

Figure 33: MODEL-9

The R-square value has slight reduced to 76.9%

Model performance of:

Training data –

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.434575	3.298059	0.794719	0.79418	4.088793

Test data-

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.638314	3.467131	0.789128	0.787708	4.29552

VIF values-

	feature	VIF
0	const	28.155508
1	lread	5.303116
2	lwrite	4.295939
3	scall	2.653012
4	swrite	3.012529
5	exec	2.842704
6	rchar	1.694305
7	wchar	1.536667
8	ppgout	2.013184
9	atch	1.782293
10	ppgin	1.559541
11	pflt	3.439371
12	freemem	1.943347
13	freeswap	1.759247
14	runqsz_Not_CPU_Bound	1.153926

Figure 34: VIF of MODEL-9

Since now all VIF values are below 5, now lets look at p value. P value must be below 0.05.

Here we notice that p-value for 'atch' is 0.181. So lets now remove 'atch' variable.

X. MODEL10-

We drop the column 'atch'.

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.769
Model:	OLS	Adj. R-squared:	0.768
Method:	Least Squares	F-statistic:	625.7
Date:	Fri, 22 Sep 2023	Prob (F-statistic):	0.00
Time:	00:29:12	Log-Likelihood:	-7260.1
No. Observations:	2458	AIC:	1.455e+04
Df Residuals:	2444	BIC:	1.463e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	81.4295	0.464	175.472	0.000	80.520	82.339
lread	-0.0546	0.014	-4.012	0.000	-0.081	-0.028
lwrite	0.0614	0.020	3.022	0.003	0.022	0.101
scall	-0.0006	9.75e-05	-6.495	0.000	-0.001	-0.000
swrite	-0.0032	0.002	-1.985	0.047	-0.006	-3.91e-05
exec	-0.2298	0.075	-3.053	0.002	-0.377	-0.082
rchar	-4.133e-06	6.95e-07	-5.946	0.000	-5.5e-06	-2.77e-06
wchar	-5.726e-06	1.68e-06	-3.416	0.001	-9.01e-06	-2.44e-06
ppgout	-0.1044	0.029	-3.607	0.000	-0.161	-0.048
ppgin	-0.0592	0.011	-5.629	0.000	-0.080	-0.039
pflt	-0.0411	0.002	-24.444	0.000	-0.044	-0.038
freemem	-0.0005	8.23e-05	-6.573	0.000	-0.001	-0.000
freeswap	1.021e-05	2.86e-07	35.662	0.000	9.64e-06	1.08e-05
runqsz_Not_CPU_Bound	2.0570	0.200	10.311	0.000	1.666	2.448

Figure 35: MODEL-10

The R-square value remains to be 76.9%.

Model performance of:

Training data –

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.439013	3.30166	0.794308	0.793804	4.093329

Test data-

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	4.640013	3.467199	0.768957	0.767633	4.296027

Residuals-

	Actual Values	Fitted Values	Residuals
694	91.0	89.530544	1.469456
5535	94.0	91.804461	2.195539
4244	61.5	74.717158	-13.217158
2472	83.0	80.954011	2.045989
7052	94.0	98.274704	-4.274704

Figure 36: Residuals of MODEL-10

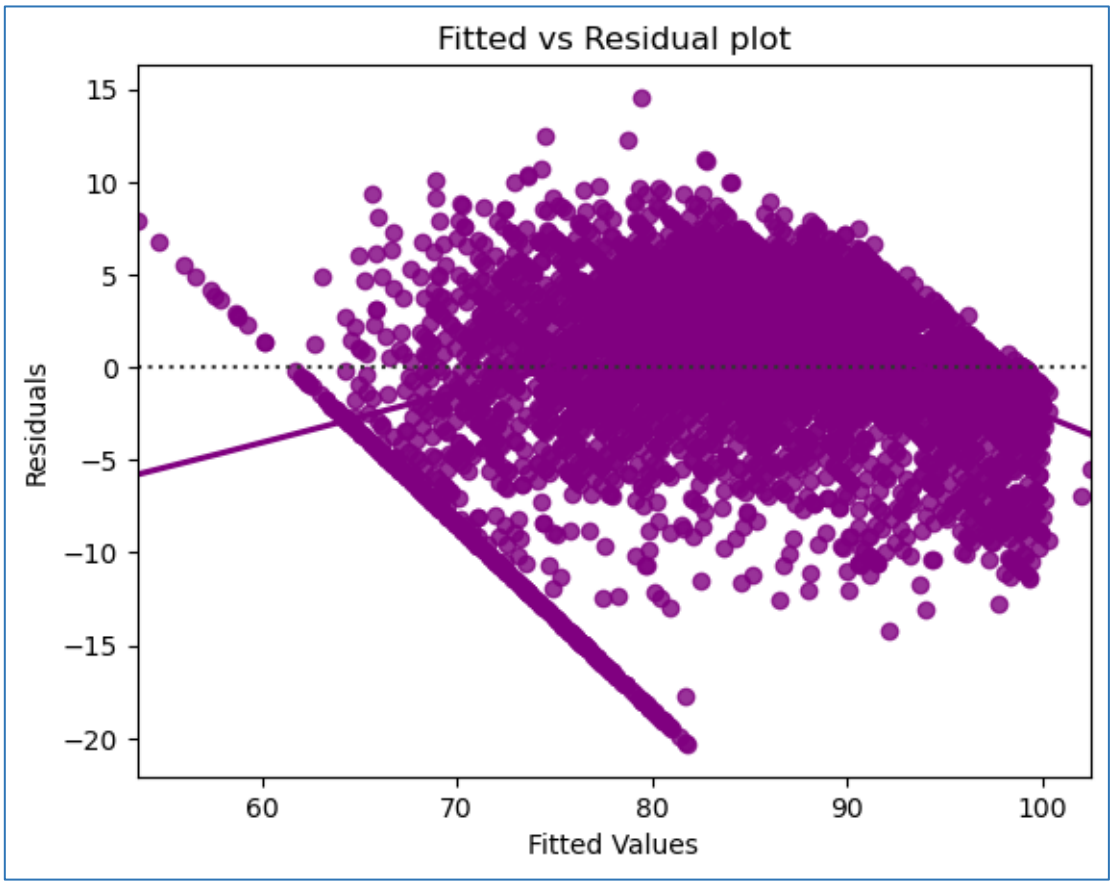


Figure 37: Fitted values vs Residuals of MODEL-10

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

A. Model Insights:

We conducted a linear regression analysis to predict the 'usr' variable based on various performance-related factors.

The model exhibited good performance, with an R-squared value of approximately 0.769 on the test data.

The key performance factors affecting 'usr' include 'lread,' 'lwrite,' 'scall,' 'swrite,' 'exec,' 'rchar,' 'wchar,' 'ppgout,' 'ppgin,' 'pflt,' 'freemem,' 'freeswap,' and 'runqsz_Not_CPU_Bound.'

B. Performance Factors:

Memory Usage: Factors related to memory usage ('lread,' 'lwrite,' 'ppgout,' 'ppgin,' 'pflt,' 'freemem,' 'freeswap') significantly impact 'usr.' Managing memory efficiently can improve system performance.

System Calls: 'scall' has a negative impact on 'usr,' suggesting that an excessive number of system calls may lead to increased resource usage.

'swrite' is negatively associated with 'usr,' indicating that excessive write operations may negatively affect system performance.

Processor Execution: 'exec' has a negative impact on 'usr,' suggesting that optimizing program execution is crucial for performance.

Character Operations: Both 'rchar' and 'wchar' have a significant negative impact on 'usr.' Minimizing character operations can improve system performance.

C. Recommendations:

Memory Optimization: Focus on memory optimization strategies to reduce memory-related resource usage.

System Call Reduction: Optimize code to minimize the number of system calls.

I/O Optimization: Efficiently manage write operations to reduce resource usage.

Processor Execution: Optimize program execution to reduce resource consumption.

Character Operations: Minimize character operations to improve system efficiency.

Performance Monitoring: Continuously monitor these factors to maintain optimal system performance.

D. Project Summary:

We conducted a comprehensive analysis to understand the factors influencing system resource usage.

Data preprocessing included data cleaning and splitting the dataset into training and testing sets.

Linear regression models were trained and evaluated for predictive performance.

We presented detailed insights into the performance factors affecting 'usr' and their impact.

Actionable recommendations were provided to optimize system performance.

E. Business Impact:

Implementing the recommendations can lead to improved system performance and resource utilization.

Efficient resource management can result in cost savings and better user experiences.

Monitoring these performance factors can help organizations proactively address resource issues and ensure system reliability.

Problem Statement – Predictive Modelling

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

Perform the following in given order:

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.

- Data frame-

	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
...	...	...	...	...	...	...	...	...	...	...
1468	33.0	Tertiary	Tertiary	NaN	Scientology	Yes	2	Very High	Exposed	Yes
1469	33.0	Tertiary	Tertiary	NaN	Scientology	No	1	Very High	Exposed	Yes
1470	39.0	Secondary	Secondary	NaN	Scientology	Yes	1	Very High	Exposed	Yes
1471	33.0	Secondary	Secondary	NaN	Scientology	Yes	2	Low	Exposed	Yes
1472	17.0	Secondary	Secondary	1.0	Scientology	No	2	Very High	Exposed	Yes

Figure 38: Logistic regression dataset

- Head – Top 5 rows

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

Media_exposure	Contraceptive_method_used
Exposed	No
Exposed	No
Exposed	No
Exposed	No
Exposed	No

Figure 39: Head (top 5)

- Tail – Last 5 rows

	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index
1468	33.0	Tertiary	Tertiary	NaN	Scientology	Yes	2	Very High
1469	33.0	Tertiary	Tertiary	NaN	Scientology	No	1	Very High
1470	39.0	Secondary	Secondary	NaN	Scientology	Yes	1	Very High
1471	33.0	Secondary	Secondary	NaN	Scientology	Yes	2	Low
1472	17.0	Secondary	Secondary	1.0	Scientology	No	2	Very High

Media_exposure	Contraceptive_method_used
Exposed	Yes
Exposed	Yes
Exposed	Yes
Exposed	Yes
Exposed	Yes

Figure 40: Tail (Last 5)

- Shape – (1473,10)
  The dataset has 1473 rows and 10 columns.

- Info –

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

Figure 41: Data information

There are 2 floats, 1 integer and 7 objects in the data

- Describe –

	Wife_age	No_of_children_born	Husband_Occupation
count	1402.000000	1452.000000	1473.000000
mean	32.606277	3.254132	2.137814
std	8.274927	2.365212	0.864857
min	16.000000	0.000000	1.000000
25%	26.000000	1.000000	1.000000
50%	32.000000	3.000000	2.000000
75%	39.000000	4.000000	3.000000
max	49.000000	16.000000	4.000000

Figure 42: Data description

- Null values –

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

Figure 43: Null values

There are 71 missing values in Wife_age and 21 in No_of_children_born.

- Duplicate values – There are total of 82 duplicated values in Dataset.

Below is the sample data of Duplicated Values

	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index
79	38.000000	Tertiary	Tertiary	1.0	Scientology	Yes	1	Very High
167	26.000000	Tertiary	Tertiary	1.0	Scientology	No	1	Very High
224	47.000000	Tertiary	Tertiary	4.0	Scientology	No	1	Very High
270	30.000000	Tertiary	Tertiary	2.0	Scientology	No	1	Very High
299	26.000000	Tertiary	Tertiary	1.0	Scientology	No	1	Very High
394	29.000000	Tertiary	Tertiary	0.0	Scientology	Yes	2	Very High
414	20.000000	Primary	Secondary	3.0	Scientology	No	3	Very High
462	36.000000	Tertiary	Tertiary	3.0	Scientology	No	1	Very High
492	37.000000	Tertiary	Tertiary	3.0	Scientology	No	1	Very High
528	29.000000	Tertiary	Tertiary	2.0	Scientology	Yes	1	High
576	41.000000	Tertiary	Tertiary	4.0	Non-Scientology	Yes	2	Very High
585	39.000000	Tertiary	Tertiary	3.0	Scientology	No	1	Very High
586	24.000000	Tertiary	Tertiary	1.0	Scientology	No	1	Very High
622	46.000000	Tertiary	Tertiary	4.0	Scientology	No	1	Very High
627	44.000000	Tertiary	Tertiary	4.0	Scientology	No	1	Very High
646	24.000000	Tertiary	Tertiary	1.0	Scientology	Yes	1	Very High
655	29.000000	Tertiary	Tertiary	2.0	Scientology	No	3	Very High

Figure 44: Duplicated values

- The missing values were treated by writing a user defined function. The wife age was treated by using mean of the dataset and No of children born was treated by using median of the dataset

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

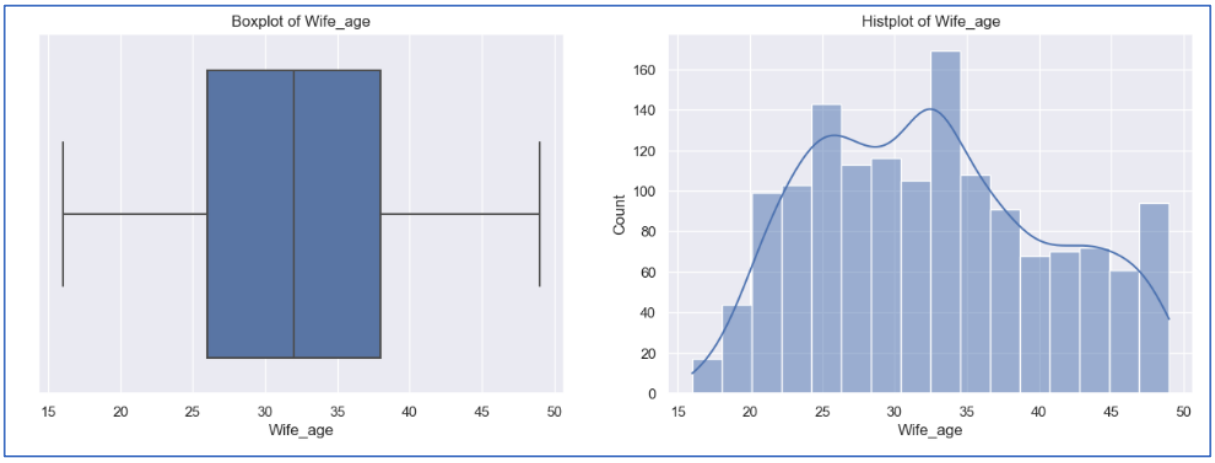
Figure 45: After imputing null values

EXPLORATORY DATA ANALYSIS

Univariate analysis:

- Numeric Variables-:

We draw boxplot and histplot for the numercal variables namely Wife age and No of children born Wife Age



- ✓ No of children Born

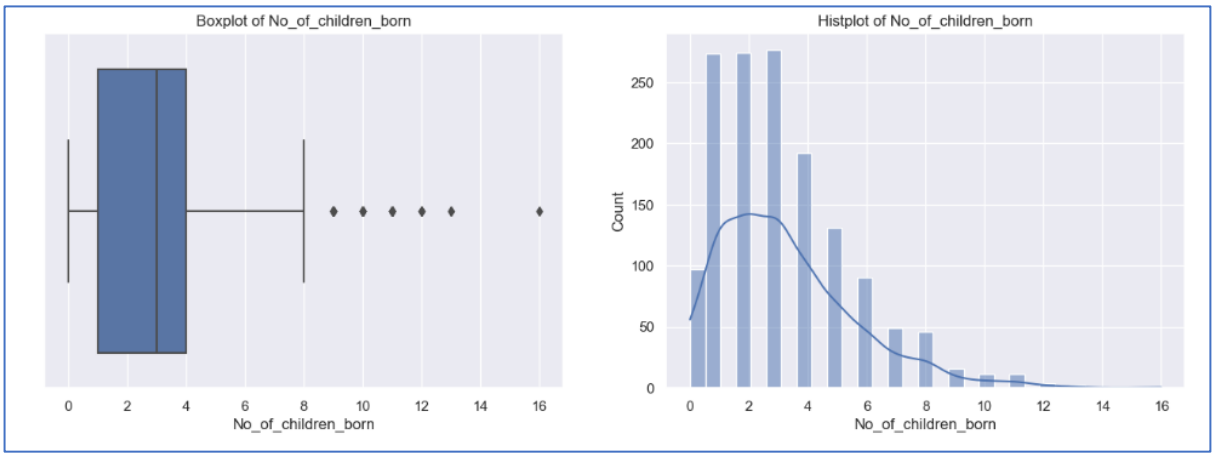
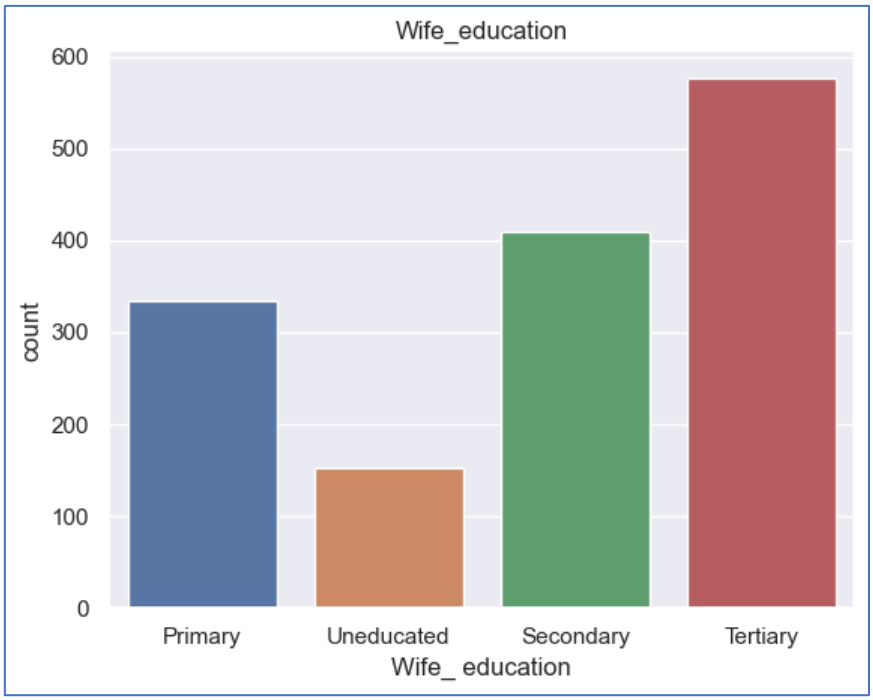


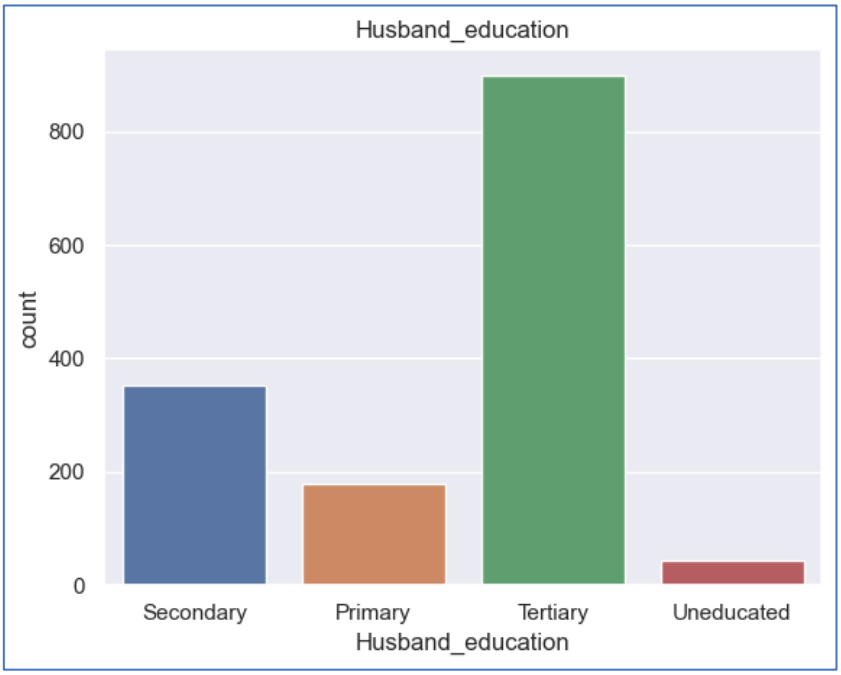
Figure 46: Box and Hist plots

- Categorical Variables-: We draw countplot for the following categorical variables

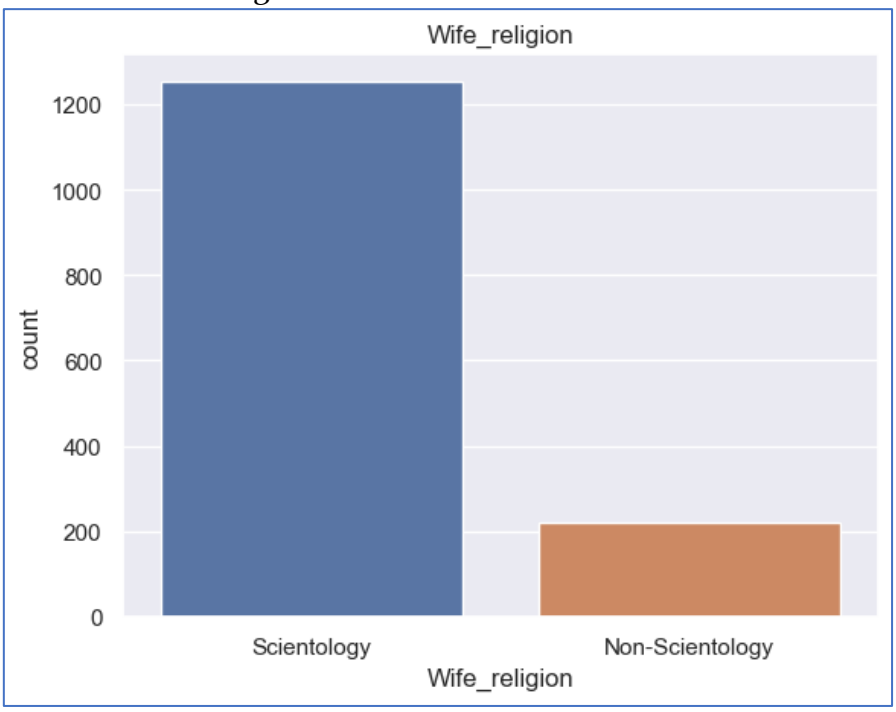
  ✓ Wife Education



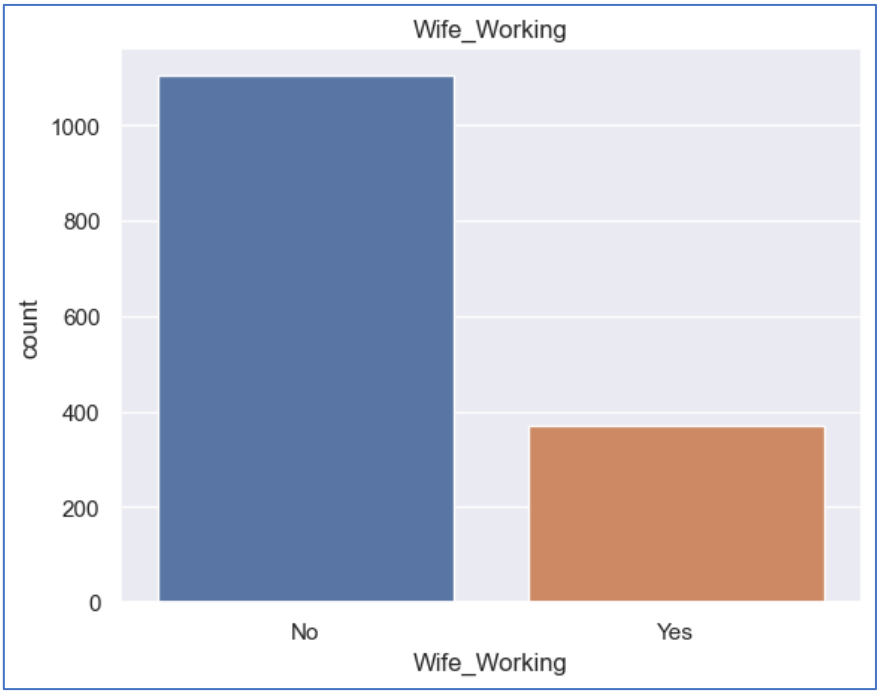
  ✓ Husband Education



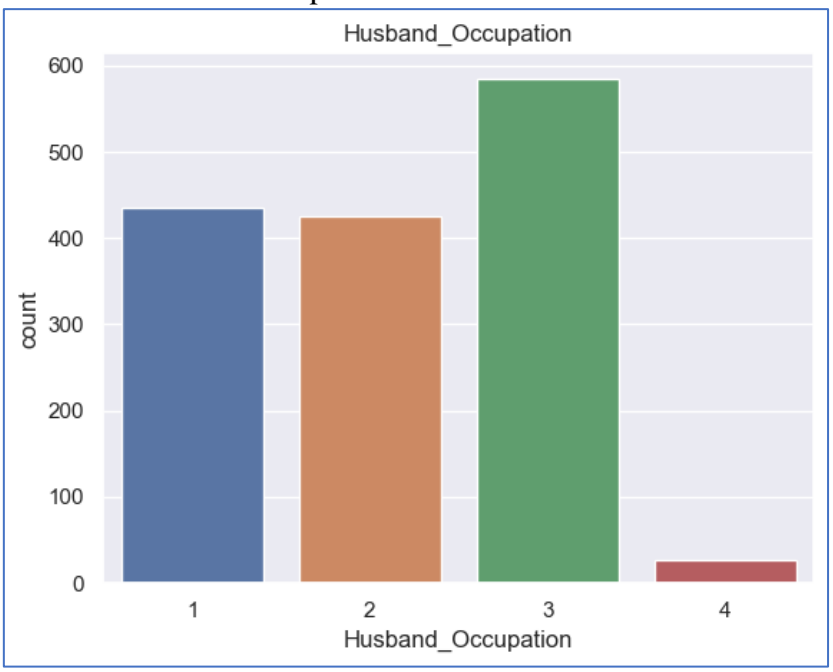
✓ Wife religion



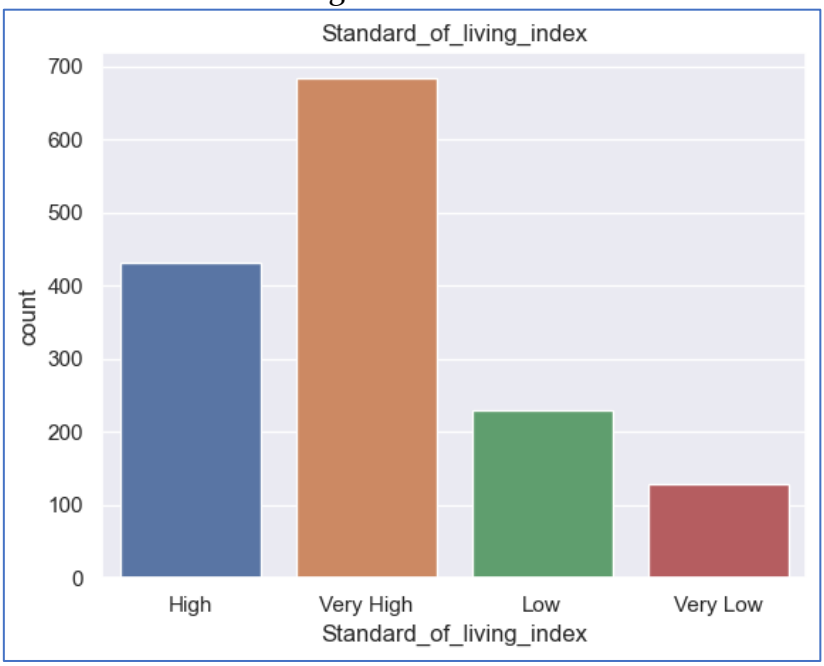
✓ Wife working



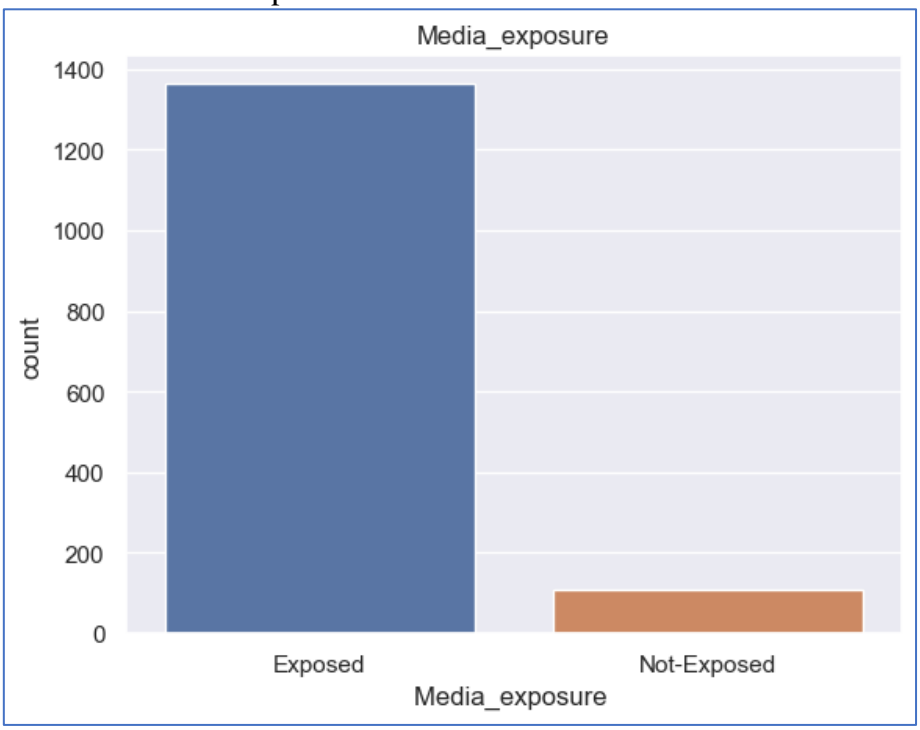
✓ Husband Occupation



✓ Standard of living index



✓ Media exposure



✓ Contraceptive method used

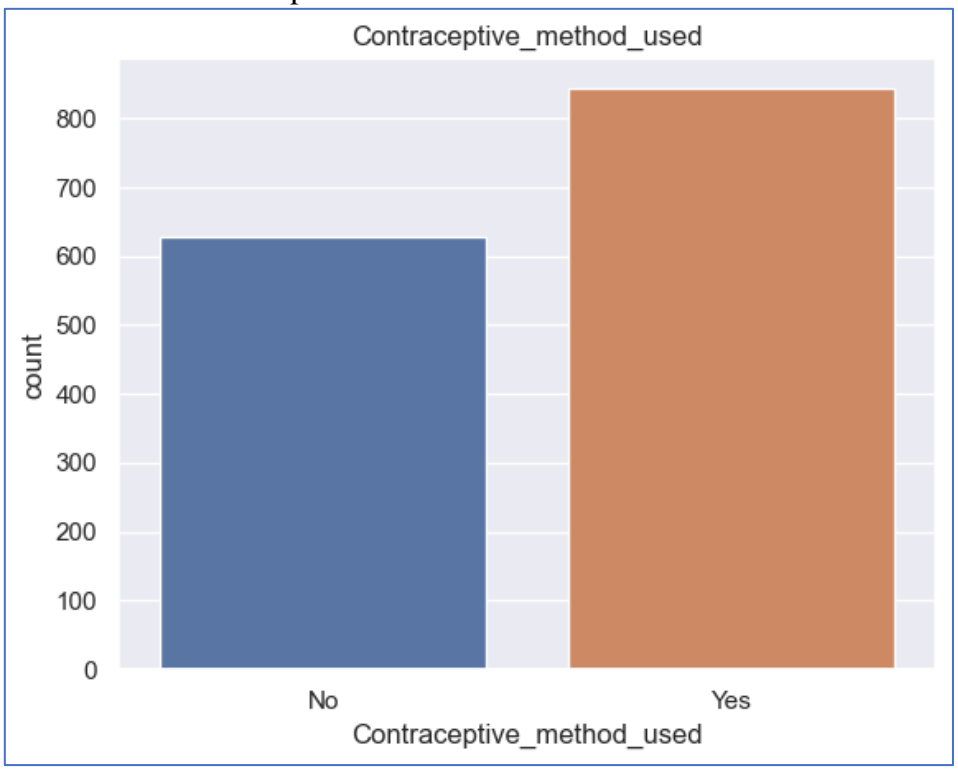


Figure 47: Count plots of Categorical Variables

Bivariate analysis:

- ✓ Pairplot

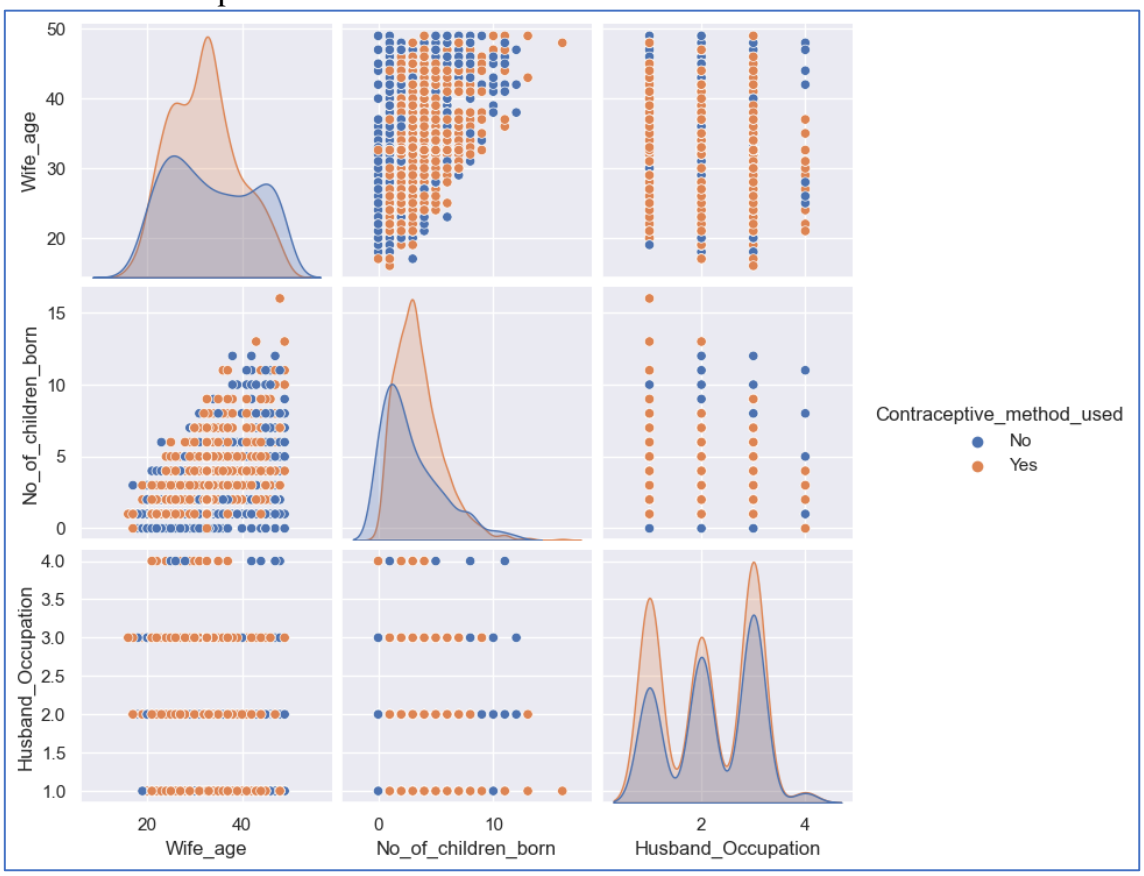


Figure 48: Pairplot

✓ Heatmap

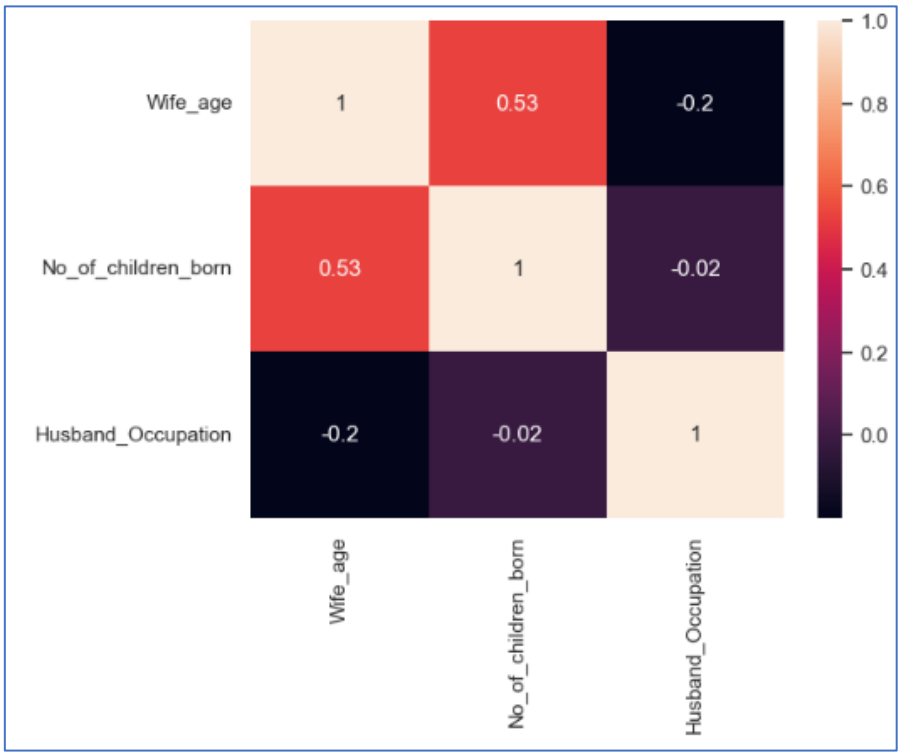
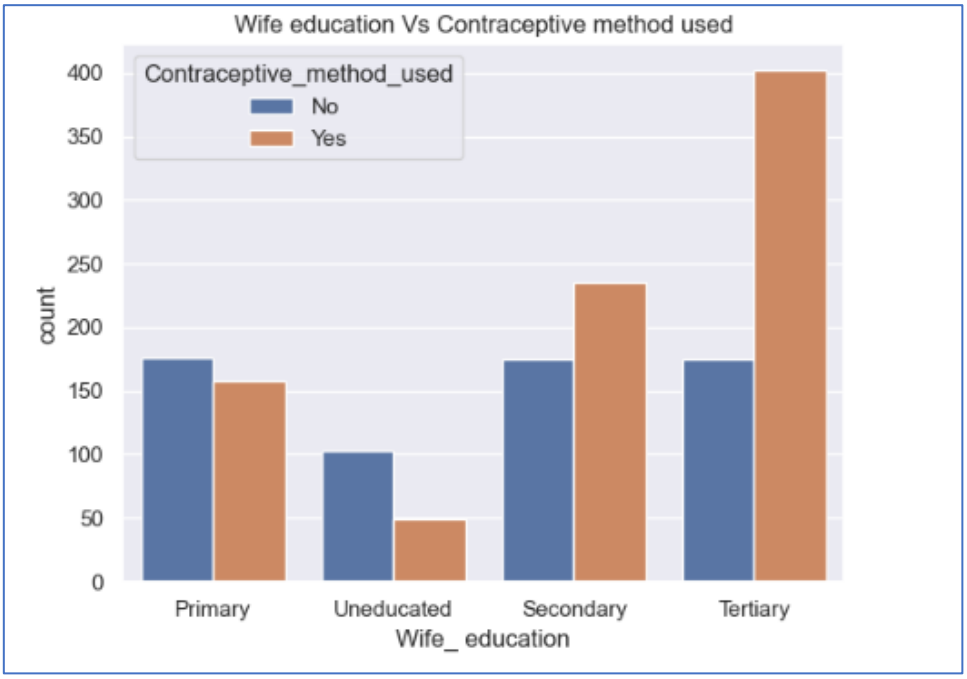


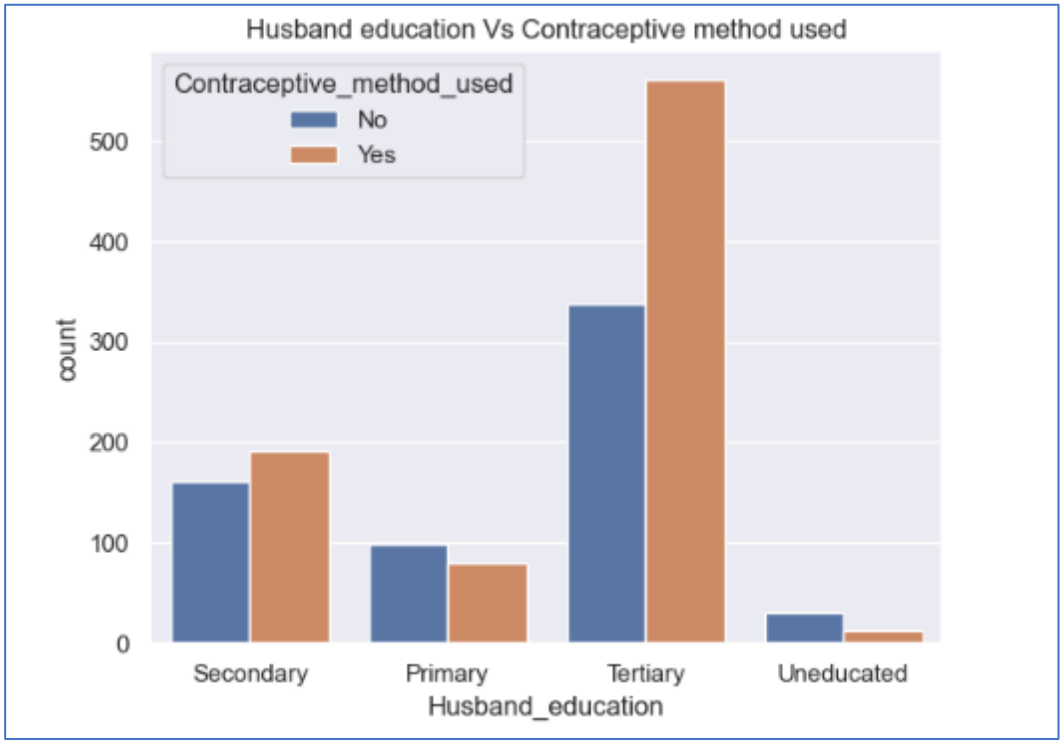
Figure 49: Heatmap

The values of the axis are not close to 1. Hence it is clear that the values are not strongly correlated.

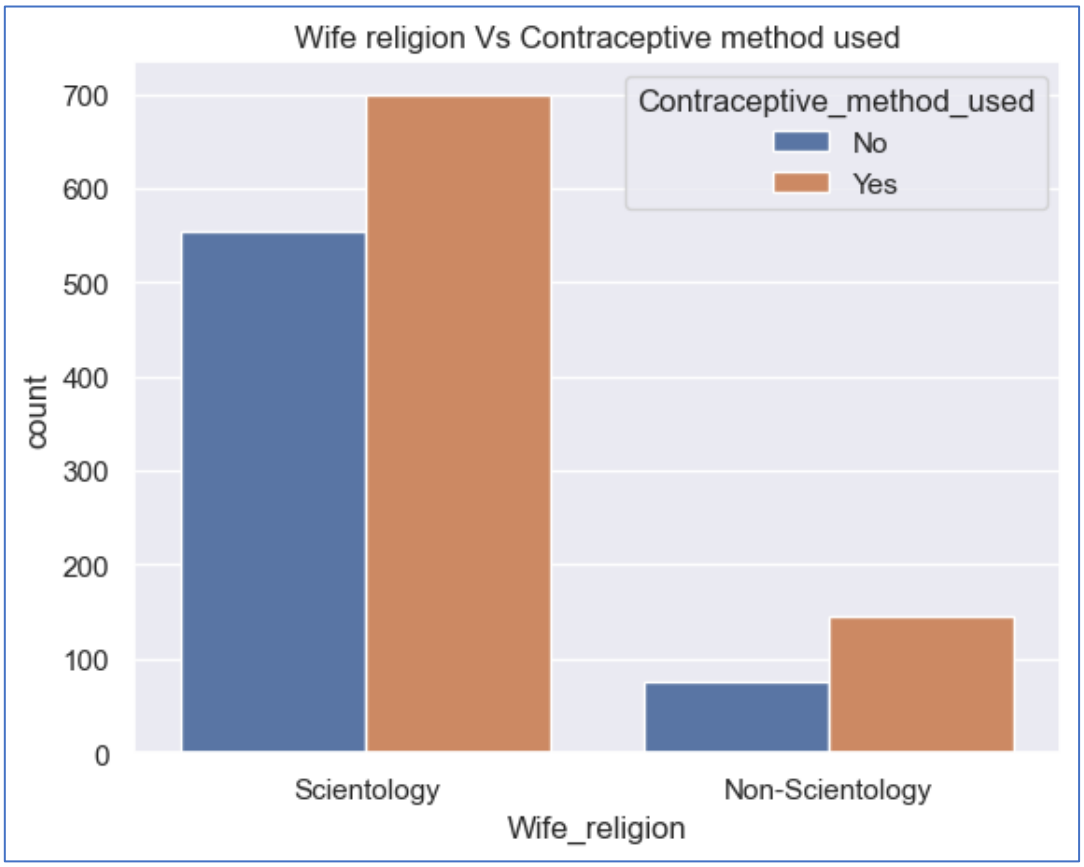
✓ Wife Education vs Contraceptive method used



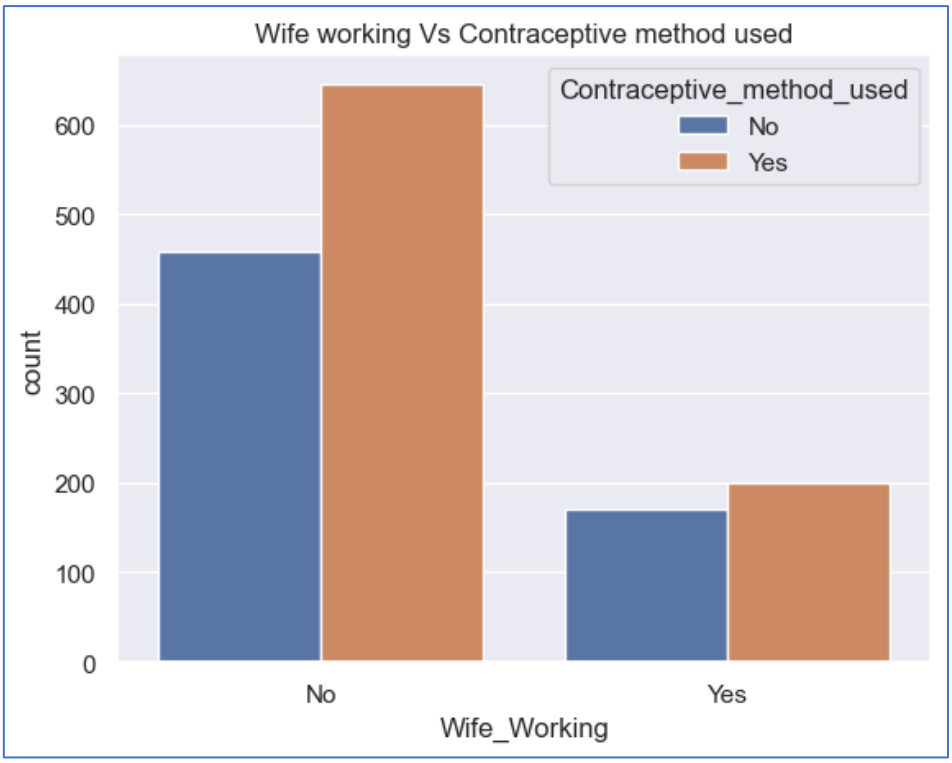
✓ Husband Education vs Contraceptive method used



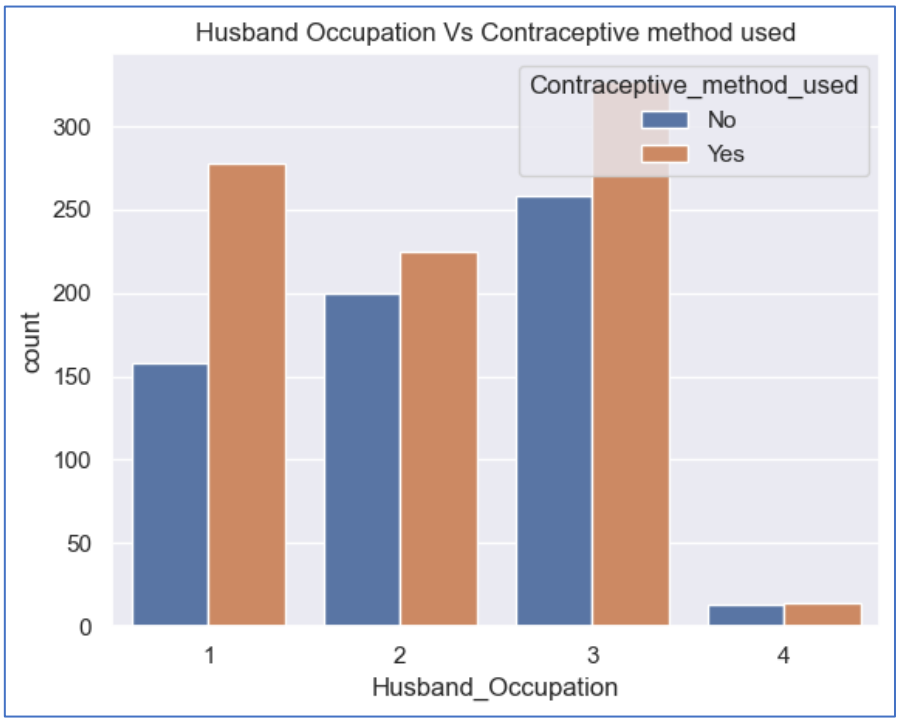
✓ Wife religion vs Contraceptive method used



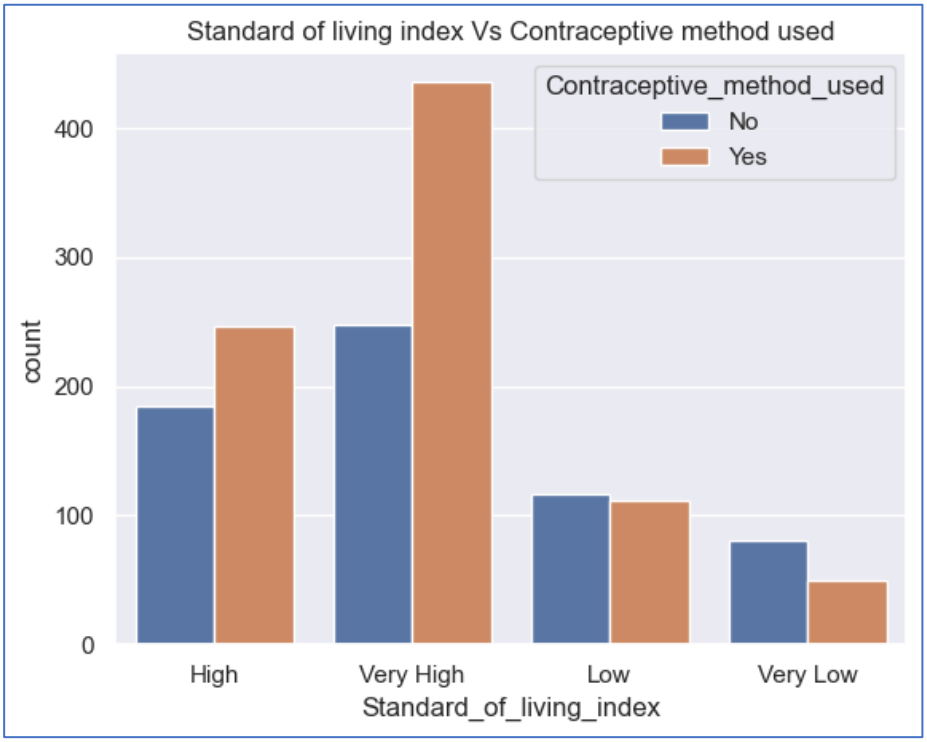
✓ Wife working vs Contraceptive method used



✓ Husband Occupation vs Contraceptive method used



✓ Standard of living index vs Contraceptive method used



✓ Media Exposure vs Contraceptive metho

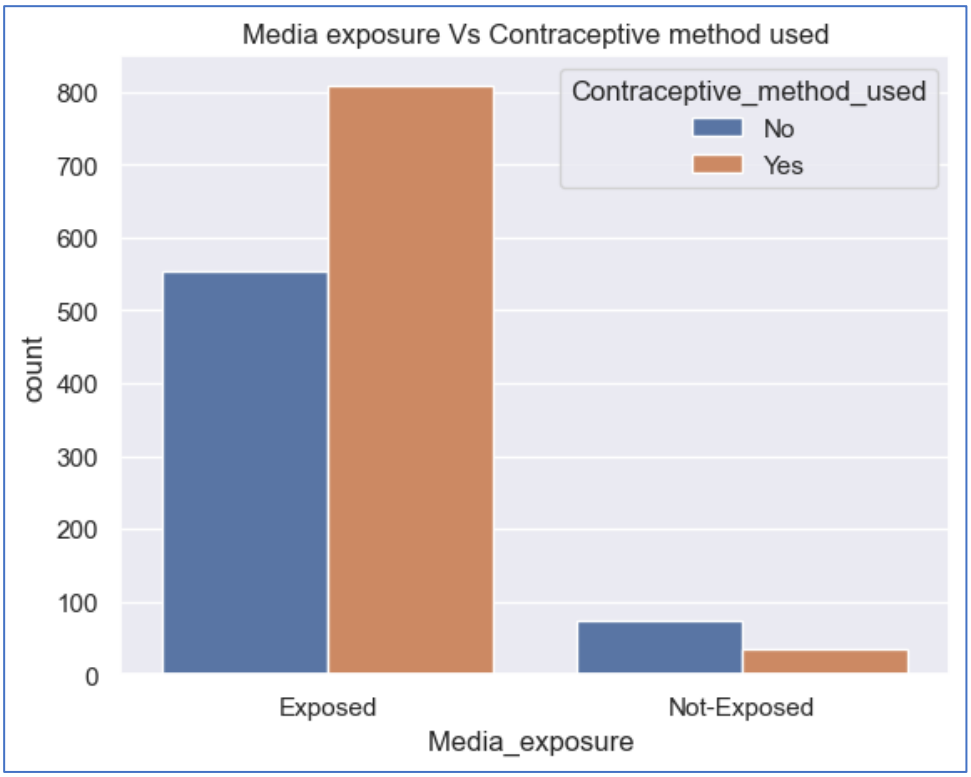


Figure 50: Bivariate analysis

Multivariate analysis:

- ✓ Standard of living index vs No of children born vs contraceptive method used

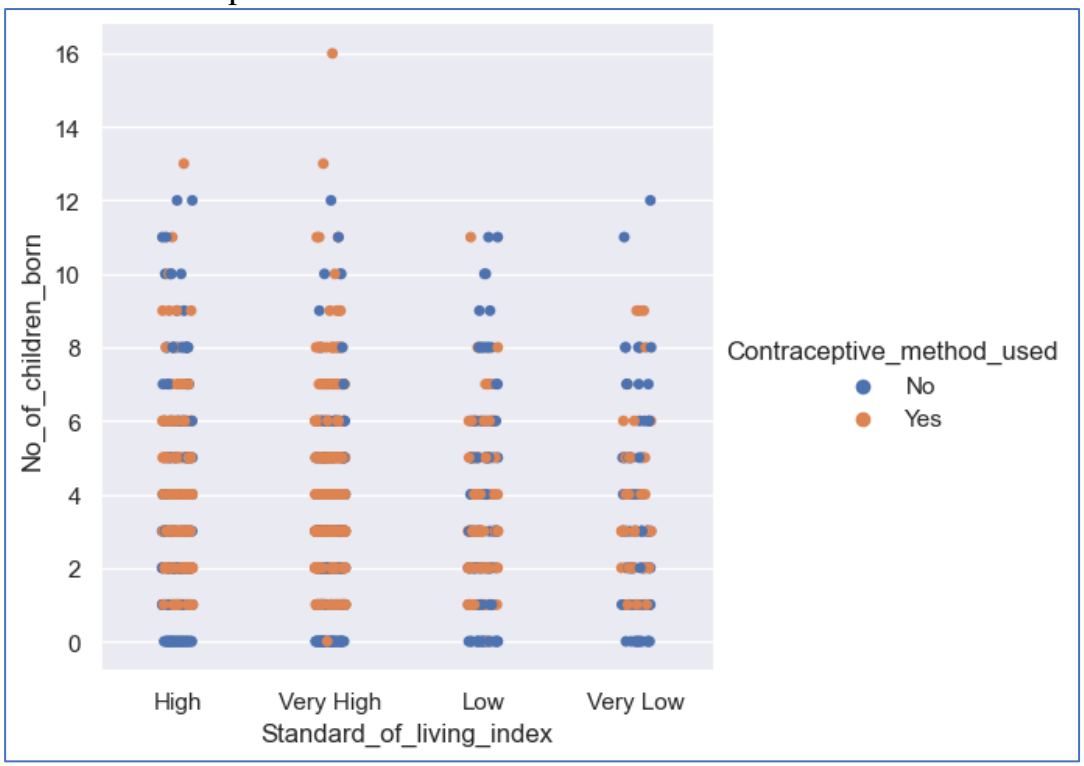


Figure 51: Cat plot

2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression

In this section, we present the results of applying logistic regression to the dataset for predicting contraceptive method usage based on demographic and socio-economic characteristics.

Data Pre-processing and Splitting:

- The dataset was pre-processed to prepare it for logistic regression modelling. The steps involved in data pre-processing include:
  - Separating the target variable (**Contraceptive_method_used**) from the predictor variables (**X** and **Y**).
  - Encoding the target variable into numerical format, where "Yes" is represented as 1 and "No" as 0.
  - Adding a constant term to the predictor variables, which is a common practice

- Creating dummy variables for categorical predictors using one-hot encoding.
- Splitting the data into a training set (70%) and a test set (30%) using a random seed for reproducibility.

Data Split Summary:

- The shape of the training set is (1031, 16), indicating that it consists of 1031 samples and 16 predictor variables.
- The shape of the test set is (442, 16), indicating that it contains 442 samples with the same set of predictor variables as the training set.

Logistic Regression Model:

- A logistic regression model was fitted to the training data to predict contraceptive method usage.
- The logistic regression model summary provides valuable insights into the model's performance and the significance of predictor variables.

Logit Regression Results

Dep. Variable:	Contraceptive_method_used	No. Observations:	1031
Model:	Logit	Df Residuals:	1016
Method:	MLE	Df Model:	14
Date:	Sun, 24 Sep 2023	Pseudo R-squ.:	0.1105
Time:	14:24:09	Log-Likelihood:	-624.70
converged:	True	LL-Null:	-702.33
Covariance Type:	nonrobust	LLR p-value:	6.407e-26

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.6639	0.447	3.723	0.000	0.788	2.540
Wife_age	-0.0715	0.011	-6.362	0.000	-0.093	-0.049
No_of_children_born	0.3130	0.041	7.718	0.000	0.234	0.393
Wife_ education_Secondary	0.4651	0.198	2.347	0.019	0.077	0.853
Wife_ education_Tertiary	1.1031	0.221	5.001	0.000	0.671	1.535
Wife_ education_Uneducated	-0.5912	0.292	-2.026	0.043	-1.163	-0.019
Husband_education_Secondary	0.0940	0.253	0.371	0.710	-0.402	0.590
Husband_education_Tertiary	-0.1344	0.262	-0.513	0.608	-0.648	0.379
Husband_education_Uneducated	-0.2970	0.468	-0.635	0.526	-1.214	0.620
Wife_religion_Scientology	-0.5085	0.205	-2.483	0.013	-0.910	-0.107
Wife_Working_Yes	-0.0219	0.159	-0.137	0.891	-0.334	0.291
Standard_of_living_index_Low	-0.0494	0.219	-0.226	0.821	-0.478	0.379
Standard_of_living_index_Very High	0.1870	0.169	1.109	0.267	-0.143	0.517
Standard_of_living_index_Very Low	-0.6550	0.272	-2.411	0.016	-1.188	-0.122
Media_exposure _Not-Exposed	-0.5122	0.303	-1.689	0.091	-1.107	0.082

Figure 52: MODEL-1

- Pseudo R-squared: 0.1110, indicating the goodness of fit of the model.
- Convergence: The model successfully converged, indicating that the optimization algorithm reached a solution.

- LLR p-value: 1.597e-25, suggesting the overall statistical significance of the model.

Interpreting Coefficients:

- The logit regression results show that the following factors are statistically significantly associated with the use of contraception:
- Wife's age (negative coefficient)
- Number of children born (positive coefficient)
- Wife's education level (positive coefficient)
- Wife's religion (negative coefficient for Scientology)
- Standard of living index (negative coefficient for Very Low)

Coefficients

1. **Constant (1.4118):**
    - The constant represents the log-odds when all other predictor variables are zero. In this case, it's 1.4118.
2. **Wife Age (-0.0701):**
    - For each one-unit increase in wife's age, the log-odds of using a contraceptive method decrease by 0.0701 units.
3. **Number of Children Born (0.3139):**
    - For each additional child born, the log-odds of using a contraceptive method increase by 0.3139 units.
4. **Husband Occupation (0.0758):**
    - Husband's occupation does not appear to have a statistically significant effect on contraceptive method usage (p-value > 0.05).
5. **Wife Education Secondary (0.4783):**
    - Women with secondary education are associated with an increase of 0.4783 units in the log-odds of using a contraceptive method compared to those with lower levels of education.
6. **Wife Education Tertiary (1.1500):**
    - Women with tertiary education are associated with a substantial increase of 1.1500 units in the log-odds of using a contraceptive method compared to those with lower levels of education.

7. **Wife Education Uneducated (-0.6058):**
    - Women with no formal education are associated with a decrease of 0.6058 units in the log-odds of using a contraceptive method compared to those with lower levels of education.
8. **Husband Education Secondary (0.0945):**
    - Husbands with secondary education do not appear to have a statistically significant effect on contraceptive method usage (p-value > 0.05).
9. **Husband Education Tertiary (-0.1211):**
    - Husbands with tertiary education do not appear to have a statistically significant effect on contraceptive method usage (p-value > 0.05).
10. **Husband Education Uneducated (-0.3043):**
    - Husbands with no formal education do not appear to have a statistically significant effect on contraceptive method usage (p-value > 0.05).
11. **Wife Religion Scientology (-0.4950):**
    - Women belonging to the Scientology religion are associated with a decrease of 0.4950 units in the log-odds of using a contraceptive method compared to those of other religions.
12. **Wife Working Yes (-0.0281):**
    - Women who are working are associated with a slight decrease of 0.0281 units in the log-odds of using a contraceptive method compared to non-working women, but this difference is not statistically significant (p-value > 0.05).
13. **Standard of Living Index Low (-0.0585):**
    - Living in areas with a low standard of living is associated with a small decrease of 0.0585 units in the log-odds of using a contraceptive method, but this effect is not statistically significant (p-value > 0.05).
14. **Standard of Living Index Very High (0.1963):**
    - Living in areas with a very high standard of living is associated with an increase of 0.1963 units in the log-odds of using a contraceptive method, but this effect is not statistically significant (p-value > 0.05).
15. **Standard of Living Index Very Low (-0.6643):**

- Living in areas with a very low standard of living is associated with a substantial decrease of 0.6643 units in the log-odds of using a contraceptive method compared to other areas.

16. **Media Exposure Not-Exposed (-0.5150):**

- Individuals who have not been exposed to media are associated with a decrease of 0.5150 units in the log-odds of using a contraceptive method compared to those who have been exposed. However, this difference is not statistically significant (p-value > 0.05).

CONFUSION MATRIX

- **TRAIN DATA**

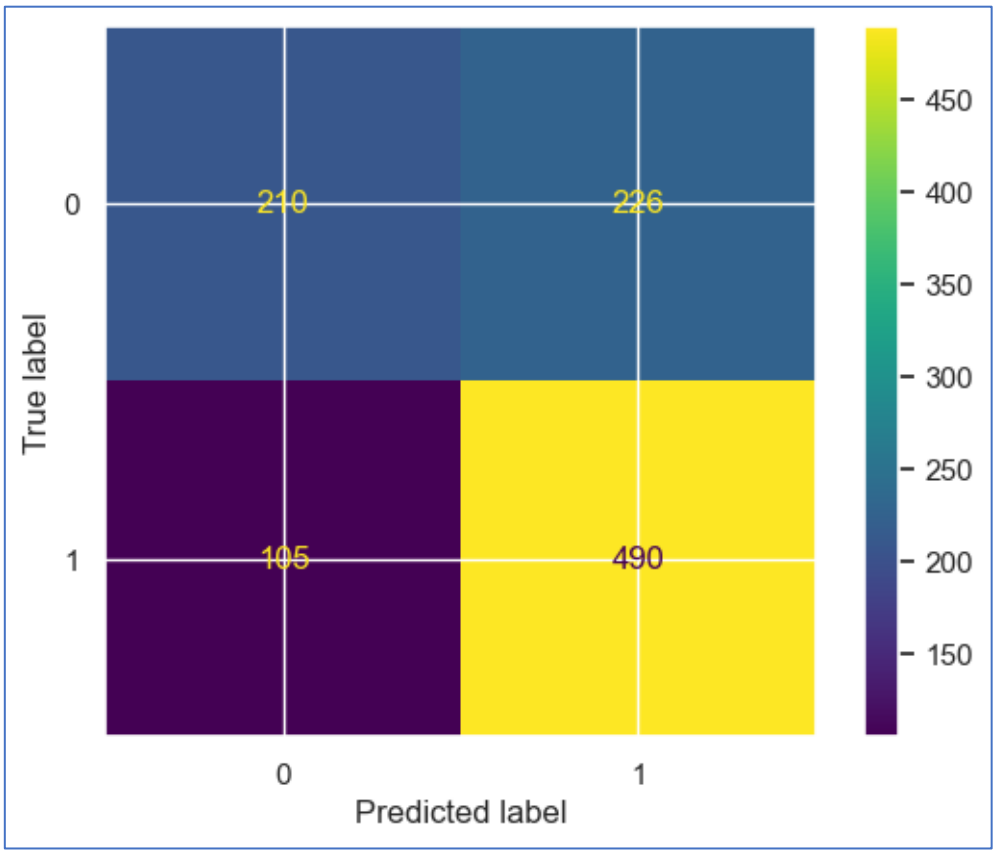


Figure 53: Confusion matrix of training data for MODEL-1

- True Negatives (TN): 210
- False Positives (FP): 226
- False Negatives (FN): 105
- True Positives (TP): 490

	Accuracy	Recall	Precision	F1
0	0.678952	0.823529	0.684358	0.747521

- **TEST DATA**

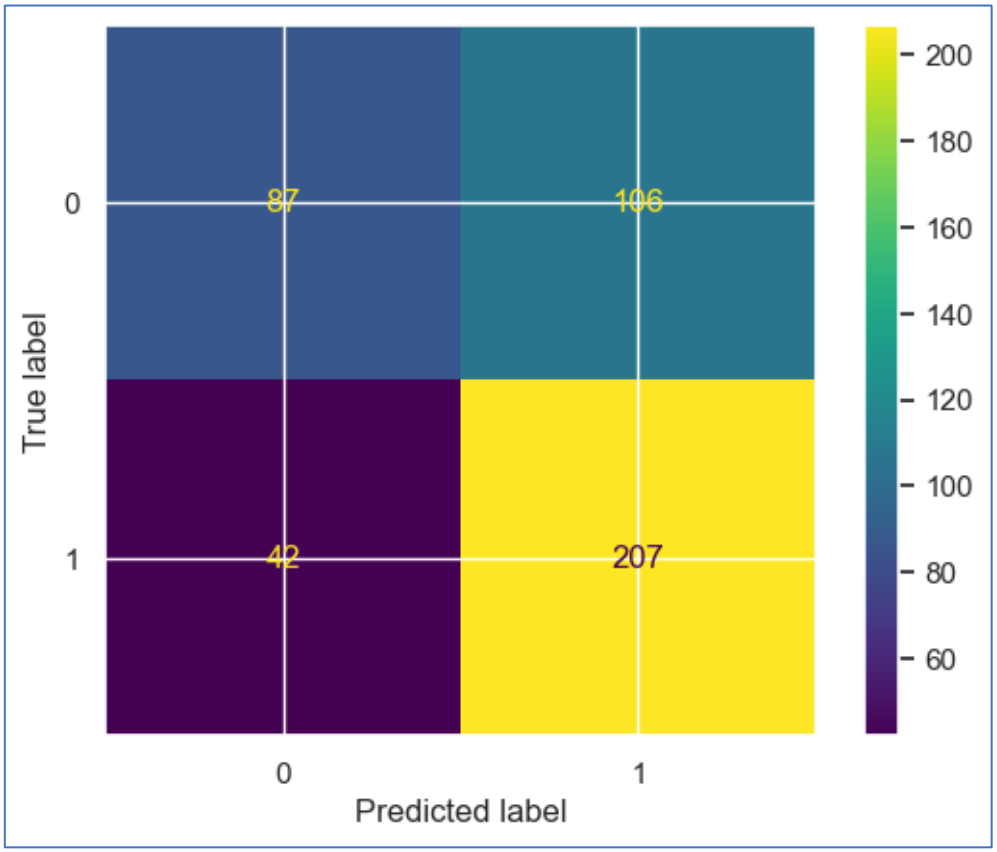


Figure 54: Confusion matrix of test data for MODEL-1

- True Negatives (TN): 87
- False Positives (FP): 106
- False Negatives (FN): 42
- True Positives (TP): 207

	Accuracy	Recall	Precision	F1
0	0.665158	0.831325	0.661342	0.736655

- **VIF**

```
Series before feature selection:

const                                62.475809
Wife_age                              1.644567
No_of_children_born                   1.544618
Husband_Occupation                    1.315302
Wife_ education_Secondary             1.762549
Wife_ education_Tertiary              2.654850
Wife_ education_Uneducated            1.548559
Husband_education_Secondary           2.491186
Husband_education_Tertiary            3.550438
Husband_education_Uneducated          1.285343
Wife_religion_Scientology             1.123917
Wife_Working_Yes                      1.032792
Standard_of_living_index_Low          1.408206
Standard_of_living_index_Very High    1.526686
Standard_of_living_index_Very Low     1.307327
Media_exposure _Not-Exposed           1.289283
dtype: float64
```

Figure 55: VIF for MODEL-1

Since all their VIF are below 5, lets look at p-values.

We notice that the p value for few columns is more than 0.05, so lets remove them one by one.

MODEL-2

We remove husband occupation.

```
                           Logit Regression Results
==============================================================================
Dep. Variable:     Contraceptive_method_used   No. Observations:                 1031
Model:                                Logit   Df Residuals:                     1021
Method:                                 MLE   Df Model:                            9
Date:                      Sun, 24 Sep 2023   Pseudo R-squ.:                  0.1069
Time:                              14:29:42   Log-Likelihood:                -627.25
converged:                             True   LL-Null:                       -702.33
Covariance Type:                 nonrobust   LLR p-value:                 8.171e-28
=========================================================================================================
                                            coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                     1.6481      0.406      4.062      0.000       0.853       2.443
Wife_age                                 -0.0729      0.011     -6.532      0.000      -0.095      -0.051
No_of_children_born                       0.3122      0.040      7.798      0.000       0.234       0.391
Wife_ education_Secondary                 0.4409      0.188      2.344      0.019       0.072       0.810
Wife_ education_Tertiary                  1.0169      0.191      5.316      0.000       0.642       1.392
Wife_ education_Uneducated               -0.7562      0.276     -2.739      0.006      -1.297      -0.215
Wife_religion_Scientology                -0.4876      0.204     -2.388      0.017      -0.888      -0.087
Standard_of_living_index_Low             -0.0862      0.211     -0.408      0.684      -0.501       0.328
Standard_of_living_index_Very High        0.1919      0.168      1.142      0.253      -0.137       0.521
Standard_of_living_index_Very Low        -0.7138      0.265     -2.697      0.007      -1.232      -0.195
=========================================================================================================
```

Figure 56: MODEL-2

- **TRAIN DATA**

	Accuracy	Recall	Precision	F1
0	0.679922	0.828571	0.683773	0.74924

- **TEST DATA**

	Accuracy	Recall	Precision	F1
0	0.665158	0.831325	0.661342	0.736655

The F1 score remains the same.

MODEL-3

We now remove all column of husband education since all their p-values are more than 0.05

Logit Regression Results

Dep. Variable:	Contraceptive_method_used	No. Observations:	1031
Model:	Logit	Df Residuals:	1019
Method:	MLE	Df Model:	11
Date:	Sun, 24 Sep 2023	Pseudo R-squ.:	0.1092
Time:	14:25:38	Log-Likelihood:	-625.64
converged:	True	LL-Null:	-702.33
Covariance Type:	nonrobust	LLR p-value:	3.037e-27

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.6444	0.408	4.034	0.000	0.845	2.443
Wife_age	-0.0719	0.011	-6.423	0.000	-0.094	-0.050
No_of_children_born	0.3141	0.040	7.806	0.000	0.235	0.393
Wife_ education_Secondary	0.4290	0.188	2.276	0.023	0.060	0.798
Wife_ education_Tertiary	0.9994	0.192	5.209	0.000	0.623	1.375
Wife_ education_Uneducated	-0.6195	0.287	-2.160	0.031	-1.182	-0.057
Wife_religion_Scientology	-0.4990	0.204	-2.441	0.015	-0.900	-0.098
Wife_Working_Yes	-0.0222	0.159	-0.140	0.889	-0.334	0.289
Standard_of_living_index_Low	-0.0219	0.215	-0.102	0.919	-0.444	0.400
Standard_of_living_index_Very High	0.1868	0.168	1.109	0.267	-0.143	0.517
Standard_of_living_index_Very Low	-0.6410	0.268	-2.390	0.017	-1.167	-0.115
Media_exposure _Not-Exposed	-0.5317	0.301	-1.767	0.077	-1.121	0.058

Figure 57: MODEL-3

- **TRAIN DATA**

	Accuracy	Recall	Precision	F1
0	0.677013	0.828571	0.680939	0.747536

- **TRAIN DATA**

	Accuracy	Recall	Precision	F1
0	0.676471	0.835341	0.670968	0.744186

MODEL-4

We now remove wife working column.

Logit Regression Results

Dep. Variable:	Contraceptive_method_used	No. Observations:	1031
Model:	Logit	Df Residuals:	1020
Method:	MLE	Df Model:	10
Date:	Sun, 24 Sep 2023	Pseudo R-squ.:	0.1092
Time:	14:29:11	Log-Likelihood:	-625.65
converged:	True	LL-Null:	-702.33
Covariance Type:	nonrobust	LLR p-value:	7.582e-28

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.6399	0.406	4.035	0.000	0.843	2.436
Wife_age	-0.0719	0.011	-6.433	0.000	-0.094	-0.050
No_of_children_born	0.3146	0.040	7.845	0.000	0.236	0.393
Wife_ education_Secondary	0.4296	0.188	2.280	0.023	0.060	0.799
Wife_ education_Tertiary	0.9984	0.192	5.208	0.000	0.623	1.374
Wife_ education_Uneducated	-0.6201	0.287	-2.162	0.031	-1.182	-0.058
Wife_religion_Scientology	-0.4988	0.204	-2.440	0.015	-0.899	-0.098
Standard_of_living_index_Low	-0.0223	0.215	-0.104	0.918	-0.444	0.400
Standard_of_living_index_Very High	0.1855	0.168	1.103	0.270	-0.144	0.515
Standard_of_living_index_Very Low	-0.6405	0.268	-2.388	0.017	-1.166	-0.115
Media_exposure _Not-Exposed	-0.5333	0.301	-1.773	0.076	-1.123	0.056

Figure 58: MODEL-4

- **TRAIN DATA**

	Accuracy	Recall	Precision	F1
0	0.675073	0.82521	0.680055	0.745634

- **TEST DATA**

	Accuracy	Recall	Precision	F1
0	0.674208	0.831325	0.669903	0.741935

MODEL-5

We now remove media exposure column

Logit Regression Results

Dep. Variable:	Contraceptive_method_used	No. Observations:	1031
Model:	Logit	Df Residuals:	1021
Method:	MLE	Df Model:	9
Date:	Sun, 24 Sep 2023	Pseudo R-squ.:	0.1069
Time:	14:29:42	Log-Likelihood:	-627.25
converged:	True	LL-Null:	-702.33
Covariance Type:	nonrobust	LLR p-value:	8.171e-28

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.6481	0.406	4.062	0.000	0.853	2.443
Wife_age	-0.0729	0.011	-6.532	0.000	-0.095	-0.051
No_of_children_born	0.3122	0.040	7.798	0.000	0.234	0.391
Wife_ education_Secondary	0.4409	0.188	2.344	0.019	0.072	0.810
Wife_ education_Tertiary	1.0169	0.191	5.316	0.000	0.642	1.392
Wife_ education_Uneducated	-0.7562	0.276	-2.739	0.006	-1.297	-0.215
Wife_religion_Scientology	-0.4876	0.204	-2.388	0.017	-0.888	-0.087
Standard_of_living_index_Low	-0.0862	0.211	-0.408	0.684	-0.501	0.328
Standard_of_living_index_Very High	0.1919	0.168	1.142	0.253	-0.137	0.521
Standard_of_living_index_Very Low	-0.7138	0.265	-2.697	0.007	-1.232	-0.195

Figure 59: MODEL-5

- **TRAIN DATA**

	Accuracy	Recall	Precision	F1
0	0.675073	0.820168	0.681564	0.74447

- **TEST DATA**

	Accuracy	Recall	Precision	F1
0	0.674208	0.839357	0.667732	0.743772

Here though the p-values of Standard of living index column is high, we notice that one of the column, i.e- standard of living index very low has pvalue of 0.007 which is less that 0.05 we retain the all three column of standard of living index

CONFUSION MATRIX

TRAIN DATA-

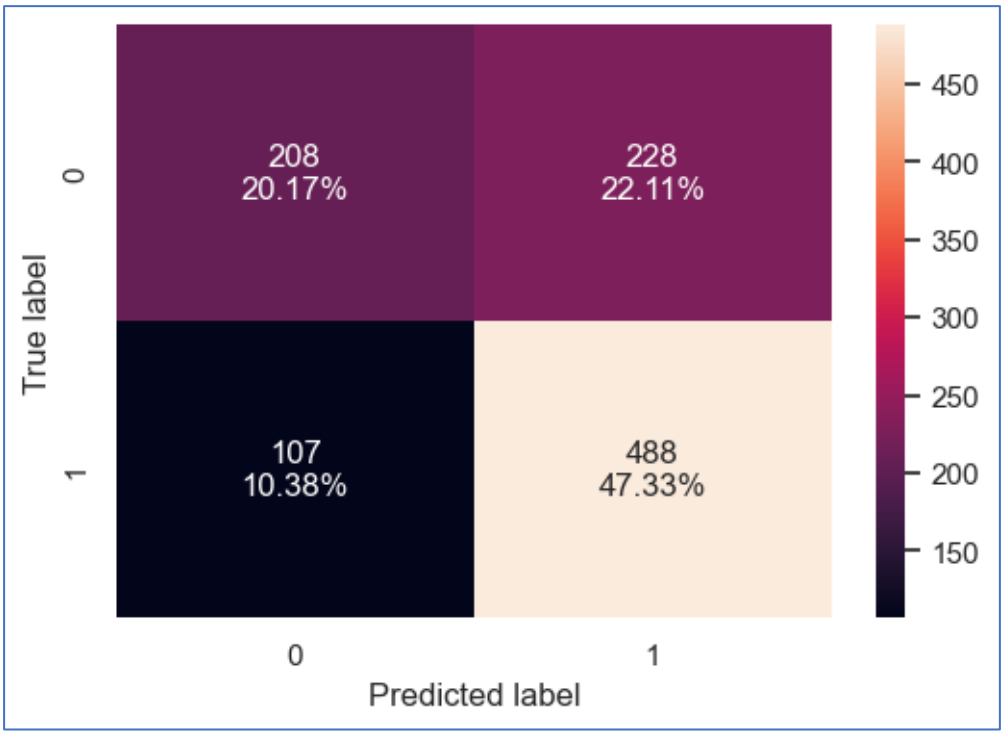


Figure 60: Confusion matrix of training data for MODEL-5

- True Positives (TP): 208
- False Positives (FP): 228
- True Negatives (TN): 488
- False Negatives (FN): 107

TEST DATA

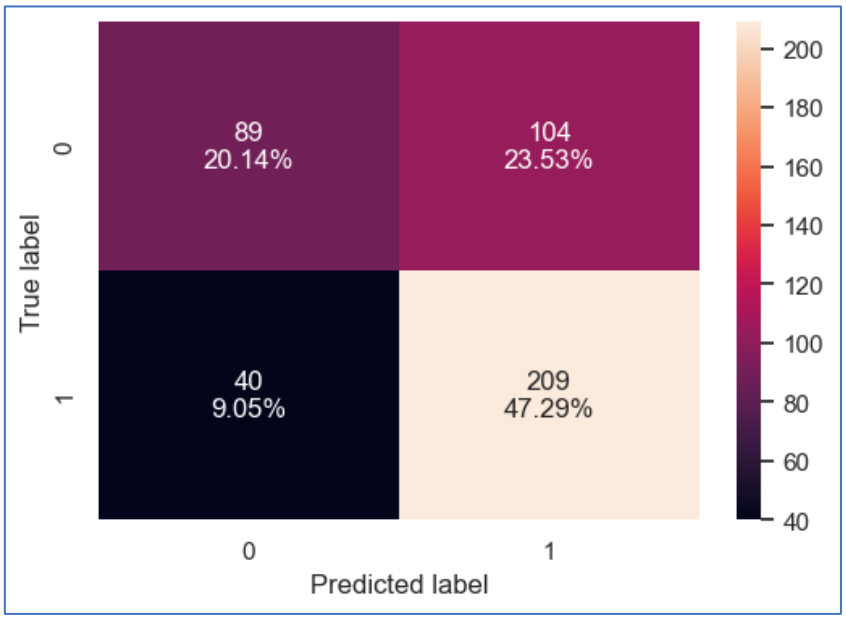


Figure 61: Confusion matrix of test data for MODEL-5

- True Positives (TP): 89
- False Positives (FP): 104
- True Negatives (TN): 209
- False Negatives (FN): 40

ROC AUC Score on Training Data:

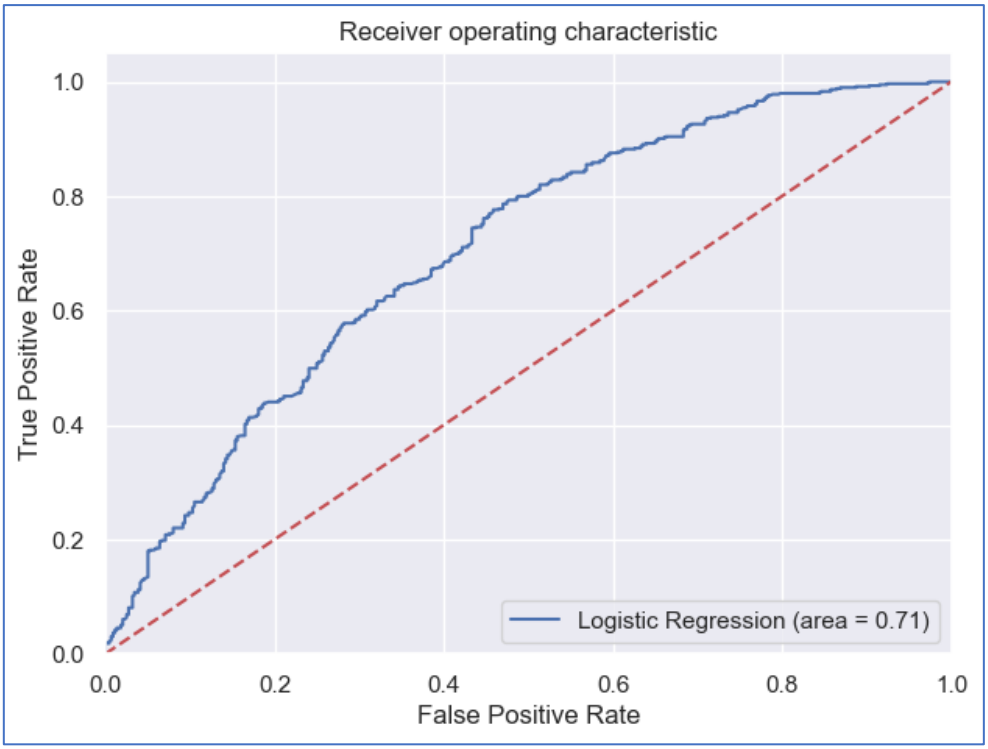


Figure 62: ROC-AUC on training data

- ROC AUC Score: 0.706

Interpretation:

- The ROC AUC score measures the model's ability to discriminate between the positive and negative classes.
- An ROC AUC score of 0.706 indicates that our logistic regression model has a fair ability to make this distinction on the training data.

Confusion Matrix with Optimal Threshold (Training Data):

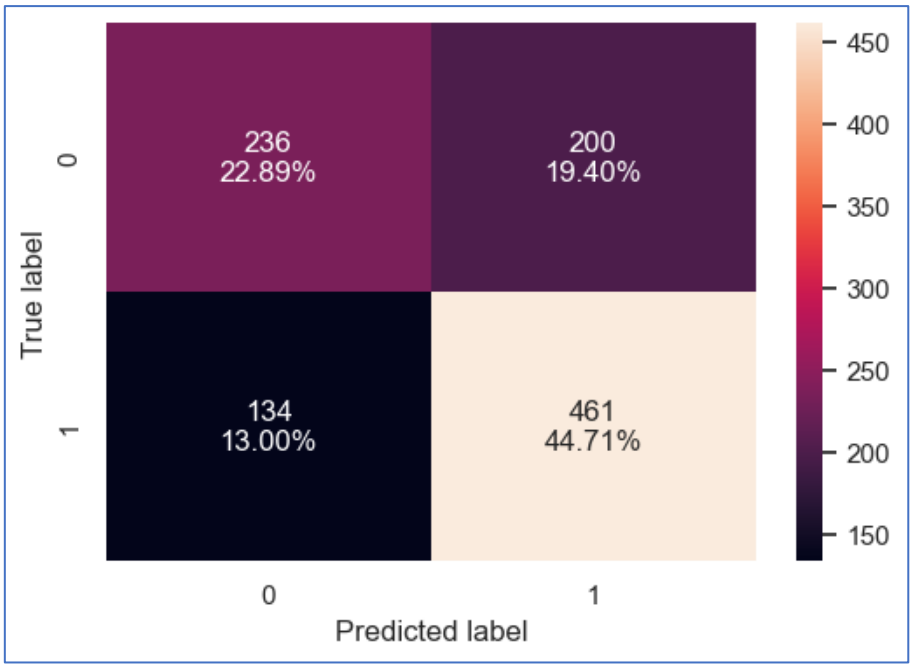


Figure 63: Confusion matrix with optimal threshold on training data

Optimal Threshold Calculation:

- Optimal Threshold (AUC-ROC): 0.537

Interpretation:

- The optimal threshold represents the point on the ROC curve that balances true positives and false positives.
- An optimal threshold of approximately 0.537 is chosen based on the ROC curve analysis.

Model Performance Metrics with Optimal Threshold (Training Data):

	Accuracy	Recall	Precision	F1
0	0.676043	0.77479	0.697428	0.734076

Conclusion:

- The logistic regression model demonstrates moderate discriminatory power with an ROC AUC score of 0.706.
- The optimal threshold of 0.537 is selected to balance true positives and false positives.
- Model performance at this threshold yields an accuracy of 0.676, indicating that 67.6% of instances are correctly classified.

- The model's ability to identify positive instances (recall) is 77.5%, and it maintains a reasonably high F1 score of 0.734.

In summary, the ROC AUC analysis and associated metrics provide insights into the logistic regression model's performance in classifying contraceptive method usage. While the model shows promise, further evaluation on test data is recommended to assess its generalization capability.

ROC AUC Score on Test Data:

- ROC AUC Score: 0.705

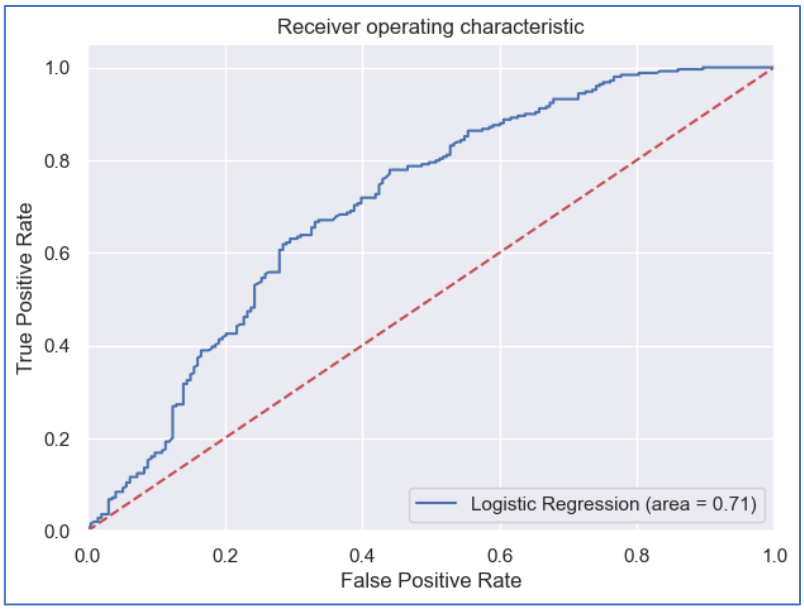


Figure 64: ROC-AUC on test data

Interpretation:

- The ROC AUC score on the test data measures the model's ability to discriminate between positive and negative classes.
- An ROC AUC score of 0.705 indicates that our logistic regression model maintains a fair ability to distinguish between classes on the test data.

Confusion Matrix with Optimal Threshold (Test Data):

- True Positives: 99
- False Positives: 94

- True Negatives: 196
- False Negatives: 53

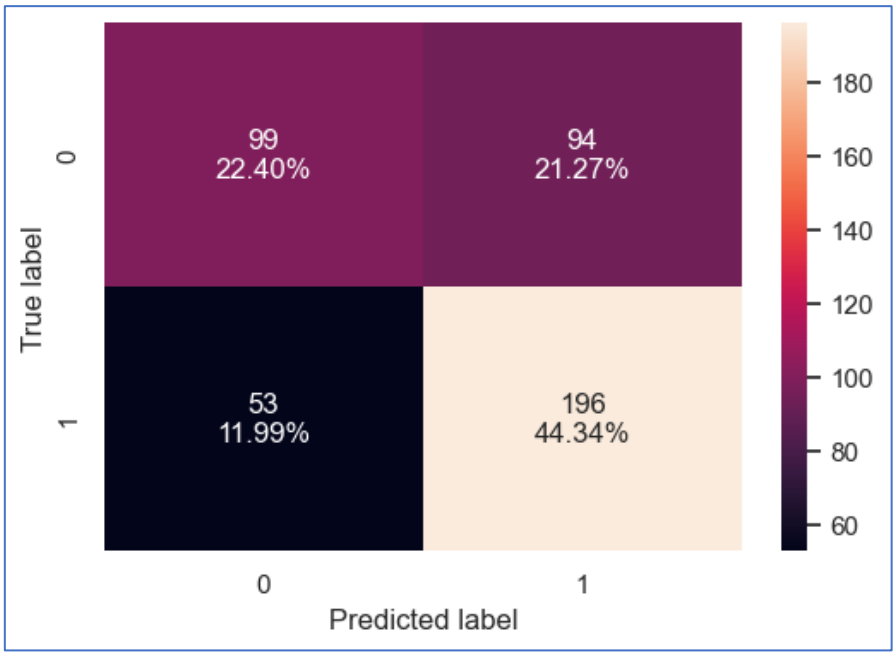


Figure 65: Confusion matrix with optimal threshold on test data

Interpretation:

- The confusion matrix provides detailed insights into model performance at the optimal threshold on the test data.
- It helps us understand the number of correctly and incorrectly classified instances on the test dataset.

Model Performance Metrics with Optimal Threshold (Test Data):

	Accuracy	Recall	Precision	F1
0	0.667421	0.787149	0.675862	0.727273

Interpretation:

- Accuracy: The proportion of correctly classified instances on the test data.
- Recall: The ability to correctly identify positive instances (sensitivity) on the test dataset.
- Precision: The ability to avoid false positives on the test dataset.

- F1 Score: A combined metric that balances precision and recall on the test data.

Conclusion:

- The logistic regression model's ROC AUC score on the test data is 0.705, indicating its continued ability to distinguish between the positive and negative classes.
- The confusion matrix at the optimal threshold on the test dataset shows that 99 instances were correctly identified as positive, with 196 instances correctly classified as negative.
- Model performance metrics at this threshold yield an accuracy of 0.667, indicating that 66.7% of instances on the test data are correctly classified.
- The model maintains a relatively high recall of 78.7% on the test dataset, meaning it effectively identifies positive instances.
- The precision is 67.6%, reflecting the model's ability to avoid false positives on the test data.
- The F1 score of 0.727 demonstrates a balanced trade-off between precision and recall on the test data.

In summary, the ROC AUC analysis on the test dataset reaffirms the logistic regression model's ability to classify contraceptive method usage. The model exhibits promising performance, with strengths in recall and a balanced F1 score, making it suitable for real-world applications.

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

A. **Model Insights:**

- We performed a logistic regression analysis to predict contraceptive method usage among women based on various socio-economic and demographic factors.
- The model exhibited moderate to good performance, with ROC AUC scores ranging from 0.705 on the test data.
- The key factors influencing contraceptive method usage include wife's age, the number of children born, wife's education level (especially tertiary education), wife's religion, and the standard of living index.

B. **Socio-Economic Factors:**

- **Wife's Education:** Women with tertiary education are more likely to use contraceptive methods compared to those with lower education levels.
- **Standard of Living:** Living in areas with very low standards of living is associated with lower contraceptive method usage. Conversely, very high standard of living areas show a slight increase in usage.
- **Religion:** Membership in the Scientology religion is associated with lower contraceptive usage.

C. **Demographic Factors:**

- **Wife's Age:** As women get older, they tend to use contraceptive methods less.
- **Number of Children Born:** An increase in the number of children born is associated with higher contraceptive method usage.

D. **Recommendations:**

- **Promote Education:** Encourage women to pursue higher education, as it is strongly correlated with increased contraceptive usage. Education empowers women to make informed family planning choices.
- **Awareness Programs:** Implement awareness programs in areas with lower standards of living to educate women about the importance of family planning and contraceptive methods.
- **Religious Sensitivity:** Sensitivity to religious beliefs is crucial. Tailor family planning campaigns to address the concerns of different religious groups.
- **Age-Related Counseling:** Provide family planning counseling that considers the age of women. Younger women may need more support and information.
- **Healthcare Access:** Ensure access to healthcare facilities that provide contraceptive services.

E. **Project Summary:**

- We conducted a comprehensive analysis to understand the factors influencing contraceptive method usage.

- Data preprocessing included data cleaning, encoding categorical variables, and splitting the dataset into training and testing sets.
- Logistic regression models were trained and evaluated for predictive performance, with ROC AUC scores indicating model strength.
- The optimal threshold was determined using ROC AUC analysis to balance precision and recall.
- We presented detailed insights into the factors affecting contraceptive usage, including socio-economic and demographic variables.
- Actionable recommendations were provided to promote family planning and improve contraceptive usage among women.

F. **Business Impact:**

- Implementing the recommendations can lead to improved family planning and reproductive health outcomes.
- Increased contraceptive usage can help families make informed choices about family size, leading to better maternal and child health outcomes.
- Improved access to education and healthcare services can empower women and promote gender equality.
- These efforts align with broader societal goals of health and well-being.

In conclusion, this analysis not only predicts contraceptive method usage but also provides valuable insights and recommendations for policymakers and organizations working in the field of reproductive health and family planning. The results can guide targeted interventions to promote family planning and improve the overall well-being of women and families.