Predective Modeling Project - Business Report

By: Aaryani Kadiyala

PGP-Data Science and Business Analytics

(PGPDSBA.O.JAN24.A)

Table of Contents

1. Problem 1 ---------- 4
2. Problem 2 ---------- 24

List of Tables:

Table 1: Features List dropped ---------- 21
Table 2: CART Model Coefficients ---------- 36
Table 3: Model Comparison Chart ---------- 39

List of Figures:

Figure 1: Dataset Sample ---------- 5
Figure 2: Data types of data frame ---------- 5
Figure 3: Statistical summary ---------- 5
Figure 4: Univariant Chart ---------- 6
Figure 5: Bivariant Chart – runqsz vs usr ---------- 7
Figure 6: Heat Map ---------- 7
Figure 7: Null values of the data set ---------- 8
Figure 8: Zero value check ---------- 8
Figure 9: Box plot for detecting outliers ---------- 9
Figure 10: Linear model before treating outliers ---------- 9
Figure 11: Box plot after treating outliers ---------- 10
Figure 12: X_train ---------- 11
Figure 13: X_test ---------- 11
Figure 14: Linear Regression ---------- 11
Figure 15: VIF Values ---------- 12
Figure 16: Actual & Residual values ---------- 14
Figure 17: Actual Vs Residual values ---------- 15
Figure 18: Pair Plot ---------- 15
Figure 19: Actual Vs Residual values After transformation ---------- 16
Figure 20: Normality of Residuals ---------- 17
Figure 21: Probability Plot ---------- 17
Figure 22: Final Model ---------- 19

Figure 23: Normality of Residuals ---------- 21
Figure 24: QQ Plot of Residuals ---------- 22
Figure 25: Dataset Sample ---------- 25
Figure 26: Data types of the dataset ---------- 25
Figure 27: Statistical summary of the dataset ---------- 25
Figure 28: Null value check before & after of the treatment ---------- 26
Figure 29: Unique values of all the object data types ---------- 26
Figure 30: Univariant Analysis of the dataset ---------- 27
Figure 31: Bivariant Analysis of the dataset ---------- 28
Figure 32: Multivariant Analysis of the dataset ---------- 29
Figure 33: Heatmap of the dataset ---------- 29
Figure 34: Before outlier treatment ---------- 30
Figure 35: After outlier treatment ---------- 30
Figure 36: Statistical summary after encoding ---------- 30
Figure 37: Training dataset ---------- 31
Figure 38: Testing dataset---------- 31
Figure 39: AUC Curve of Training Data Set ---------- 32
Figure 40: AUC Curve of Testing Data Set ---------- 32
Figure 41: Confusion Matrix of training Data Set ---------- 32
Figure 42: Confusion Matrix of testing Data Set ---------- 32
Figure 43: Classification Report of training Data Set ---------- 33
Figure 44: Classification Report of testing Data Set ---------- 33
Figure 45: Confusion Matrix of training & testing Data Set ---------- 34
Figure 46: Classification Report of training & testing Data Set ---------- 35
Figure 47: AUC chart of training & testing Data Set ---------- 35
Figure 48: AUC Curve of Training Data Set ---------- 36
Figure 49: AUC Curve of Testing Data Set ---------- 37
Figure 50: Classification Report of training Data Set ---------- 37
Figure 51: Classification Report of testing Data Set ---------- 37
Figure 52: Confusion Matrix of training Data Set ---------- 38
Figure 53: Confusion Matrix of testing Data Set ---------- 38

Problem 1: Linear Regression

The comp-activ database comprises activity measures of computer systems. Data was gathered from a Sun SPARC station 20/712 with 128 Mbytes of memory, operating in a multi-user university department. Users engaged in diverse tasks, such as internet access, file editing, and CPU-intensive programs.

Being an aspiring data scientist, you aim to establish a linear equation for predicting 'usr' (the percentage of time CPUs operate in user mode). Your goal is to analyse various system attributes to understand their influence on the system's 'usr' mode.

Data Description:

System measures used:

lread - Reads (transfers per second) between system memory and user memory

lwrite - writes (transfers per second) between system memory and user memory

scall - Number of system calls of all types per second

sread - Number of system read calls per second .

swrite - Number of system write calls per second .

fork - Number of system fork calls per second.

exec - Number of system exec calls per second.

rchar - Number of characters transferred per second by system read calls

wchar - Number of characters transfreed per second by system write calls

pgout - Number of page out requests per second

ppgout - Number of pages, paged out per second

pgfree - Number of pages per second placed on the free list.

pgscan - Number of pages checked if they can be freed per second

atch - Number of page attaches (satisfying a page fault by reclaiming a page in memory) per second

pgin - Number of page-in requests per second

ppgin - Number of pages paged in per second

pflt - Number of page faults caused by protection errors (copy-on-writes).

vflt - Number of page faults caused by address translation .

runqsz - Process run queue size (The number of kernel threads in memory that are waiting for a CPU to run. Typically, this value should be less than 2. Consistently higher values mean that the system might be CPU-bound.)

freemem - Number of memory pages available to user processes

freeswap - Number of disk blocks available for page swapping.

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

Solution:-

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap
0	1	0	2147	79	68	0.2	0.2	40671.0	53995.0	0.0	...	0.0	0.0	1.6	2.6	16.00	26.40	CPU_Bound	4870	1730946
1	0	0	170	18	21	0.2	0.2	448.0	8385.0	0.0	...	0.0	0.0	0.0	0.0	15.63	16.83	Not_CPU_Bound	7278	1869002
2	15	3	2162	159	119	2.0	2.4	NaN	31950.0	0.0	...	0.0	1.2	6.0	9.4	150.20	220.20	Not_CPU_Bound	702	1021237
3	0	0	160	12	16	0.2	0.2	NaN	8670.0	0.0	...	0.0	0.0	0.2	0.2	15.60	16.80	Not_CPU_Bound	7248	1863704
4	5	1	330	39	38	0.4	0.4	NaN	12185.0	0.0	...	0.0	0.0	1.0	1.2	37.80	47.60	Not_CPU_Bound	633	1760253

5 rows × 22 columns

Figure 1: Dataset Sample

Exploratory Data Analysis:

```
<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 2: Data types of data frame

	count	mean	std	min	25%	50%	75%	max
lread	8192.0	19.56	53.35	0.0	2.0	7.0	20.00	1845.00
lwrite	8192.0	13.11	29.89	0.0	0.0	1.0	10.00	575.00
scall	8192.0	2306.32	1633.62	109.0	1012.0	2051.5	3317.25	12493.00
sread	8192.0	210.48	198.98	6.0	86.0	166.0	279.00	5318.00
swrite	8192.0	150.06	160.48	7.0	63.0	117.0	185.00	5456.00
fork	8192.0	1.88	2.48	0.0	0.4	0.8	2.20	20.12
exec	8192.0	2.79	5.21	0.0	0.2	1.2	2.80	59.56
rchar	8088.0	197385.73	239837.49	278.0	34091.5	125473.5	267828.75	2526649.00
wchar	8177.0	95902.99	140841.71	1498.0	22916.0	46619.0	108101.00	1801623.00
pgout	8192.0	2.29	5.31	0.0	0.0	0.0	2.40	81.44
ppgout	8192.0	5.98	15.21	0.0	0.0	0.0	4.20	184.20
pgfree	8192.0	11.92	32.36	0.0	0.0	0.0	5.00	523.00
pgscan	8192.0	21.53	71.14	0.0	0.0	0.0	0.00	1237.00
atch	8192.0	1.13	5.71	0.0	0.0	0.0	0.60	211.58
pgin	8192.0	8.28	13.87	0.0	0.6	2.8	9.76	141.20
ppgin	8192.0	12.39	22.28	0.0	0.6	3.8	13.80	292.61
pflt	8192.0	109.79	114.42	0.0	25.0	63.8	159.60	899.80
vflt	8192.0	185.32	191.00	0.2	45.4	120.4	251.80	1385.00
freemem	8192.0	1763.46	2482.10	55.0	231.0	579.0	2002.25	12027.00
freeswap	8192.0	1328125.96	422019.43	2.0	1042623.5	1289289.5	1730379.50	2243187.00
usr	8192.0	83.97	18.40	0.0	81.0	89.0	94.00	99.00

Figure 3: Statistical summary

- There is total 8192 rows and 22 columns in the dataset.
- Out of 22 columns
    - 1 columns object type
    - 8 columns are integer data type
    - 13 columns are float data type
- rchar and wchar columns has null values

Univariant, Bivariant & Multivariant Analysis:

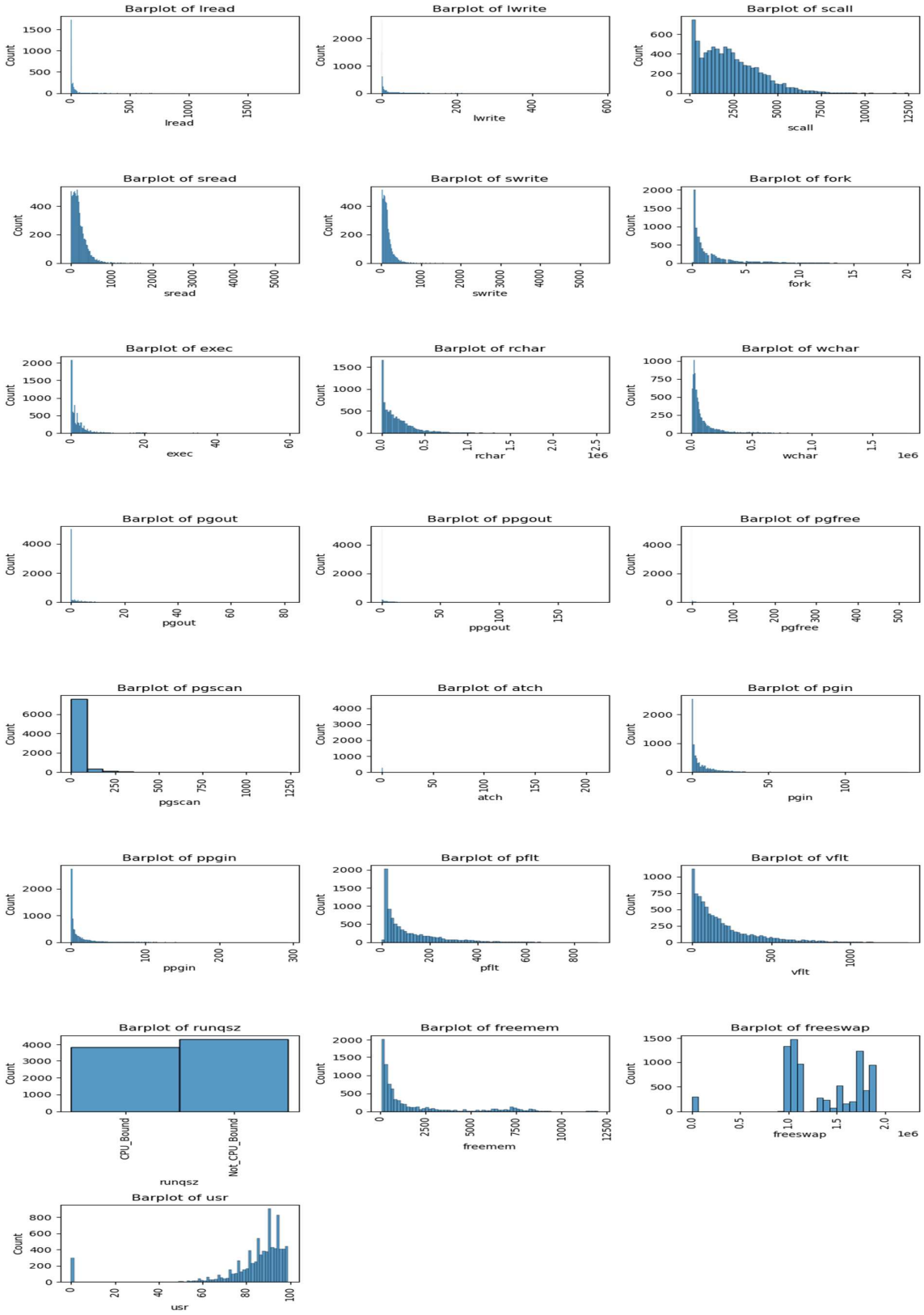

Figure 4: Univariant Chart

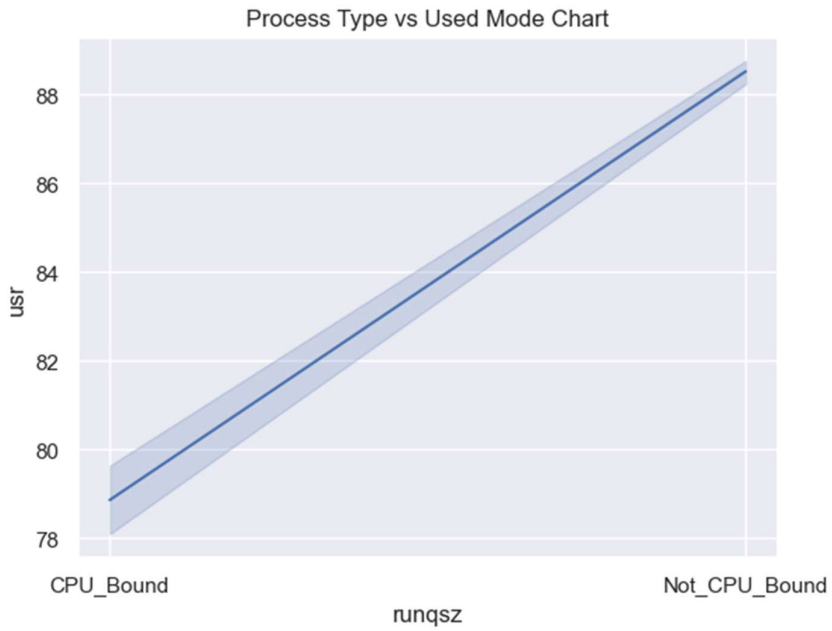

Figure 5: Bivariant Chart – runqsz vs usr

- runqsz has 2 unique values,"Not_CPU_Bound" has more count than CPU_Bound
- usr has less 0 values and more higher values. This shows systems runs more time in user mode

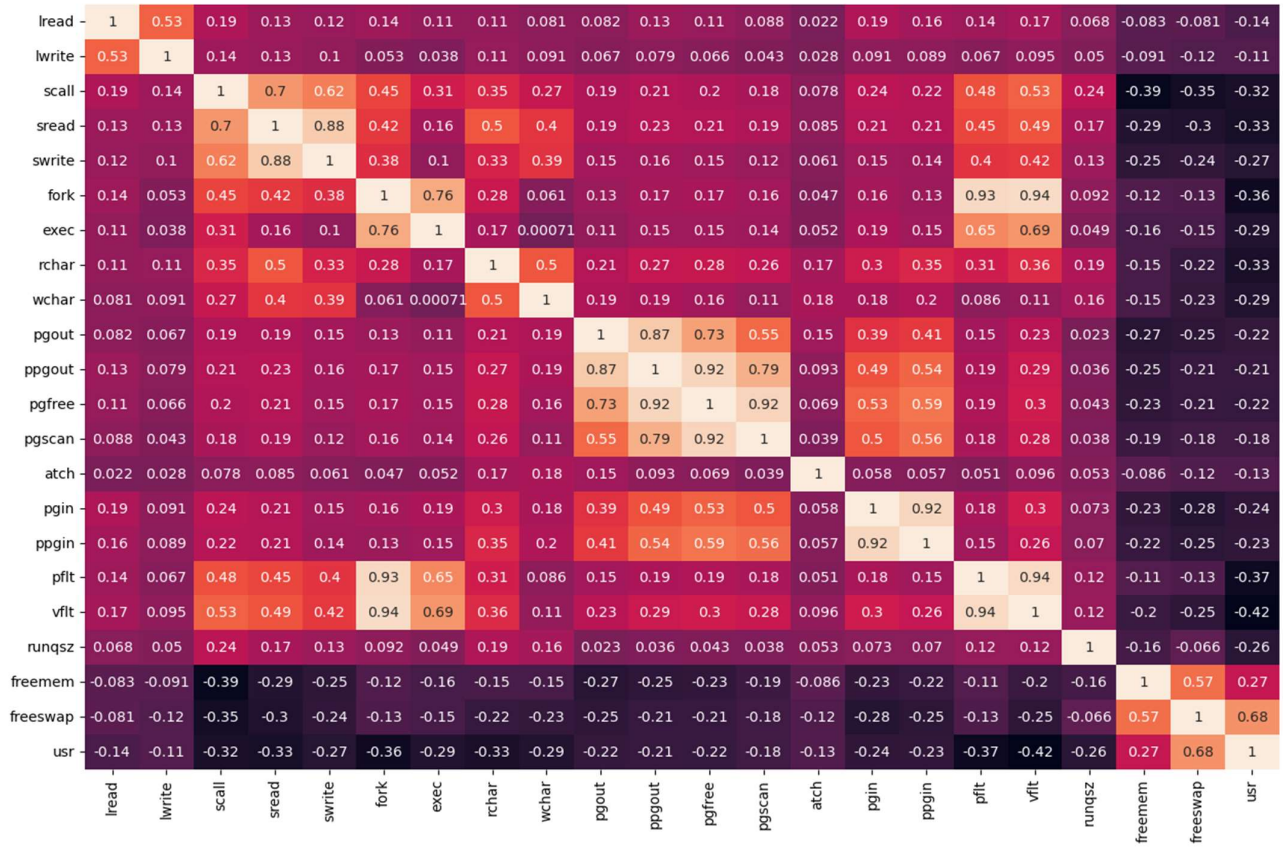

Figure 6: Heat Map

Observations :

Above heat map brings out the correlation between the features.

- There is a high correlation 94% correlation between vflt and fork.
- 93% correlation shown between pflt and fork.
- 92% correlation shown between ppgin and pgin.
- 92% correlation shown between pgfree and ppgout.
- 88% correlation shown between swrite and sread.
- 87% correlation shown between ppgout and pgout.

```
rchar    104
wchar     15
dtype: int64
```

Figure 7: Null values of the data set

```
lread      :  8.24 %
lwrite      :  32.76 %
scall      :  0.0 %
sread      :  0.0 %
swrite      :  0.0 %
fork      :  0.26 %
exec      :  0.26 %
rchar      :  0.0 %
wchar      :  0.0 %
pgout      :  59.55 %
ppgout      :  59.55 %
pgfree      :  59.44 %
pgscan      :  78.71 %
atch      :  55.85 %
pgin      :  14.89 %
ppgin      :  14.89 %
pflt      :  0.04 %
vflt      :  0.0 %
runqsz      :  0.0 %
freemem      :  0.0 %
freeswap      :  0.0 %
usr      :  3.45 %
```

Figure 8: Zero value check

- There are no duplicate values.
- Since, more than 75% of pgscan data are zeroes's column can droped from the data set.
- Other features are with less than 60% of zero values; therefore, those features are not dropped from the data frame.New feature is not needed for the compactiv data set

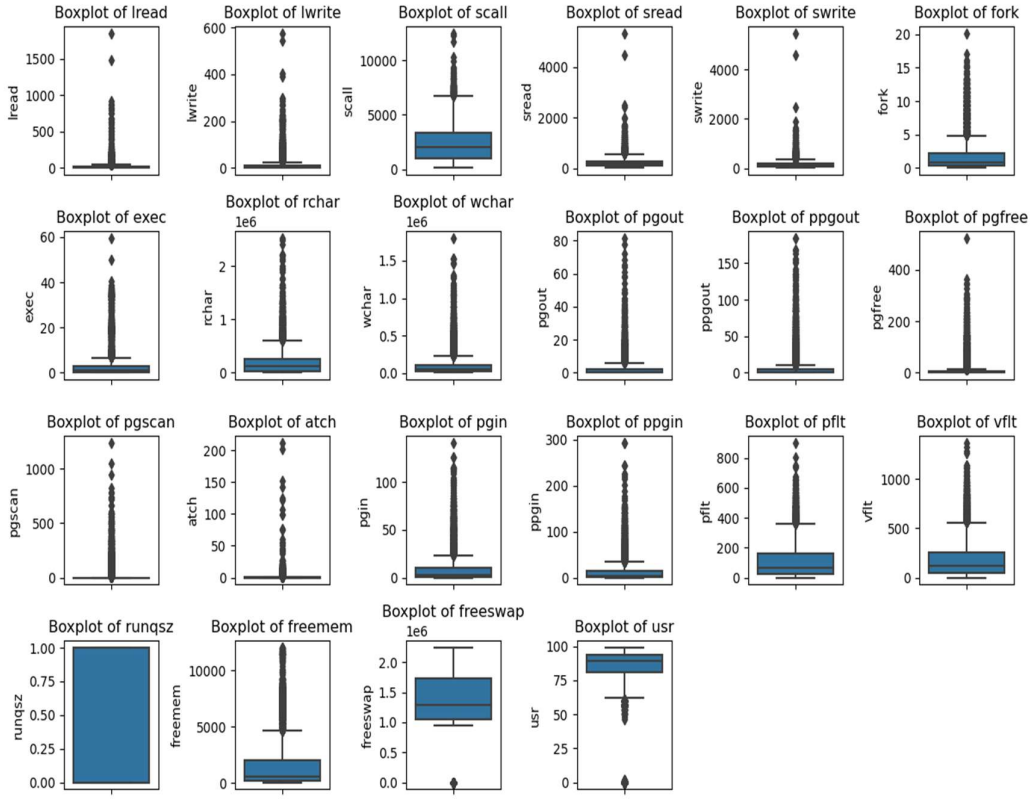

Figure 9: Box plot for detecting outliers

- From the box plot it is clearly visible all the features have outliers except runqsz.

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.640
Model:	OLS	Adj. R-squared:	0.639
Method:	Least Squares	F-statistic:	692.0
Date:	Fri, 14 Jun 2024	Prob (F-statistic):	0.00
Time:	06:16:13	Log-Likelihood:	-31296.
No. Observations:	8192	AIC:	6.264e+04
Df Residuals:	8170	BIC:	6.279e+04
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	50.7881	0.587	86.472	0.000	49.637	51.939
lread	-0.0198	0.003	-7.088	0.000	-0.025	-0.014
lwrite	0.0074	0.005	1.522	0.128	-0.002	0.017
scall	0.0010	0.000	8.759	0.000	0.001	0.001
sread	-9.749e-05	0.002	-0.061	0.952	-0.003	0.003
swrite	-0.0016	0.002	-0.889	0.374	-0.005	0.002
fork	-1.8871	0.209	-9.039	0.000	-2.296	-1.478
exec	-0.0416	0.041	-1.013	0.311	-0.122	0.039
rchar	-3.657e-06	7.19e-07	-5.084	0.000	-5.07e-06	-2.25e-06
wchar	-1.066e-05	1.1e-06	-9.731	0.000	-1.28e-05	-8.51e-06
pgout	-0.2048	0.054	-3.799	0.000	-0.310	-0.099
ppgout	0.1270	0.031	4.109	0.000	0.066	0.188
pgfree	-0.0880	0.016	-5.553	0.000	-0.119	-0.057
pgscan	0.0125	0.005	2.644	0.008	0.003	0.022
atch	-0.0396	0.022	-1.773	0.076	-0.083	0.004
pgin	0.0581	0.024	2.383	0.017	0.010	0.106
ppgin	-0.0392	0.016	-2.507	0.012	-0.070	-0.009
pflt	-0.0401	0.004	-11.182	0.000	-0.047	-0.033
vflt	0.0228	0.003	8.215	0.000	0.017	0.028
runqsz	-7.9475	0.258	-30.828	0.000	-8.453	-7.442
freemem	-0.0017	6.37e-05	-26.141	0.000	-0.002	-0.002
freeswap	3.325e-05	3.82e-07	87.082	0.000	3.25e-05	3.4e-05

Omnibus:	1928.707	Durbin-Watson:	2.026
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5359.778
Skew:	-1.244	Prob(JB):	0.00
Kurtosis:	6.084	Cond. No.	6.79e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.79e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 10: Linear model before treating outliers

Liner regression is sensitive on outliers, R-squared and Adjusted R- squared value are 64% and 63.9% respectively.

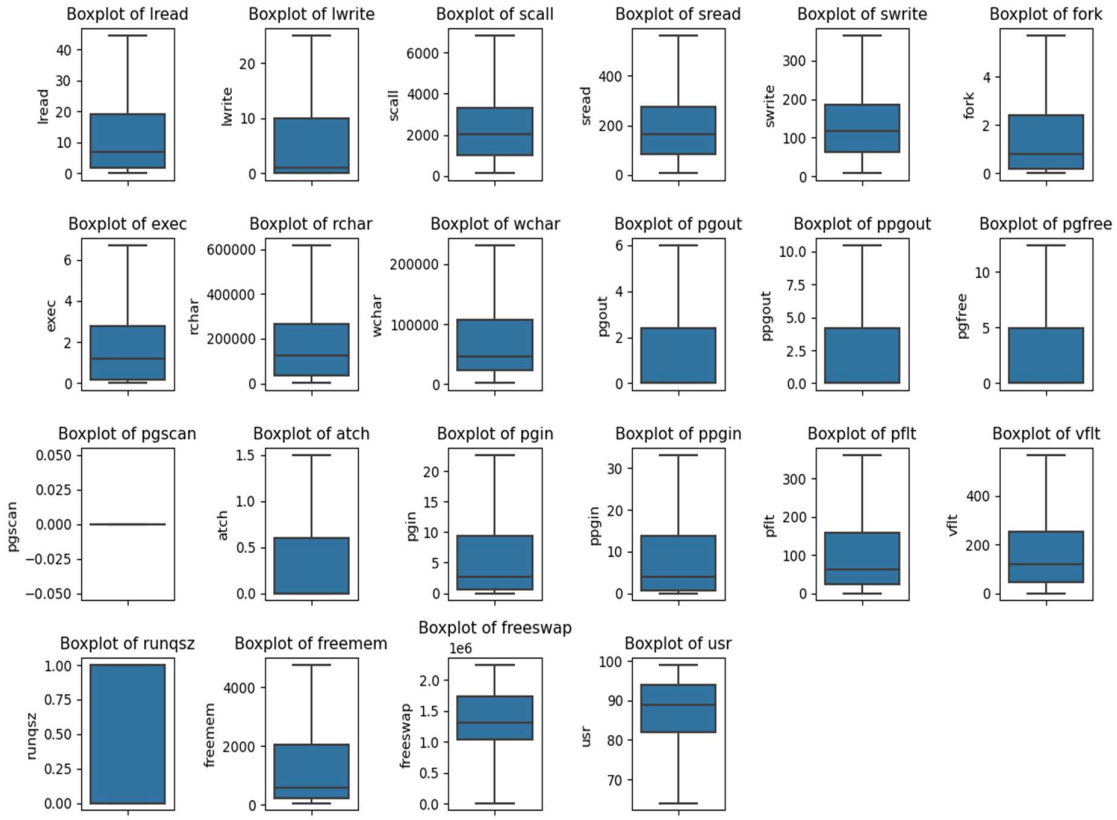

Figure 11: Box plot after treating outliers

Model Building:

Encoding string variable:

In the given data set "runqsz" is the string variable, which is encoded manually. CPU_Bound is encoded as 1 and Not_CPU_Bound is encoded as 0. runqsz variable is type casted from object to integer.

Dummy Encoding is not necessary at this data set since the runqsz has only 2 categories in it.

Split Data

usr variable has taken has a y variable (dependent variable) and all other variables are taken has x variable (independent variable).

The given data set is split into 70:30; 70% data are considered has training data and 30% of data are taken for testing the model.

X_train dataset for training the model; 21 columns with 5734 rows

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgfree	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem
1310	26.0	25.0	5731.0	312.0	224.0	0.8	0.80	155004.0	230625.875	0.0	...	0.00	0.0	1.5	0.2000	0.20	48.8	134.00	1.0	249.0
7365	15.0	3.0	1203.0	61.0	34.0	1.6	1.80	163076.0	33674.000	0.0	...	0.00	0.0	0.0	0.0000	0.00	127.8	199.40	1.0	2744.0
2284	39.0	16.0	5213.0	568.5	368.0	4.9	4.99	435848.0	230625.875	6.0	...	10.18	0.0	1.5	15.7700	17.56	348.1	561.40	1.0	236.0
7076	2.0	0.0	2585.0	203.0	145.0	0.6	0.60	329604.0	126738.000	1.0	...	1.00	0.0	0.8	23.5125	30.46	49.9	194.39	0.0	451.0
3114	2.0	1.0	1827.0	65.0	88.0	0.4	0.20	4487.0	8828.000	0.0	...	0.00	0.0	0.0	0.2000	0.20	17.4	17.00	0.0	689.0

5 rows × 21 columns

Figure 12: X_train

X_test dataset for testing the model; 21 columns with 2458 rows

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgfree	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	fre
5670	14	7	1495	197	169	0.80	1.00	10304.0	24435.0	7.98	...	24.75	38.52	1.0	2.00	2.00	63.07	106.79	1	186	9
5369	10	8	3158	324	172	0.60	2.20	1037534.0	884253.0	0.00	...	0.00	0.00	0.0	26.00	45.80	46.00	79.20	1	510	10
2111	2	0	813	117	113	1.80	0.60	59003.0	24550.0	0.60	...	7.20	14.00	0.0	0.00	0.00	96.00	135.60	0	179	17
6659	48	68	3283	134	125	0.40	0.40	33832.0	23626.0	4.20	...	9.00	0.00	0.6	1.80	2.20	36.40	56.20	0	461	11
5227	12	2	2357	113	96	6.99	20.16	55137.0	36291.0	0.00	...	0.00	0.00	0.0	8.38	12.18	231.14	423.35	0	530	10

5 rows × 21 columns

Figure 13: X_test

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.793
Model:                            OLS   Adj. R-squared:                  0.792
Method:                 Least Squares   F-statistic:                     1093.
Date:                Fri, 14 Jun 2024   Prob (F-statistic):               0.00
Time:                        06:16:50   Log-Likelihood:                -16686.
No. Observations:                5734   AIC:                         3.341e+04
Df Residuals:                    5713   BIC:                         3.355e+04
Df Model:                          20
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     85.0494      0.298    285.844      0.000      84.466      85.633
lread         -0.0633      0.009     -7.147      0.000      -0.081      -0.046
lwrite         0.0456      0.013      3.517      0.000       0.020       0.071
scall         -0.0007   6.35e-05    -11.350      0.000      -0.001      -0.001
sread          0.0026      0.001      2.513      0.012       0.001       0.005
swrite        -0.0064      0.001     -4.423      0.000      -0.009      -0.004
fork          -0.0896      0.133     -0.672      0.502      -0.351       0.172
exec          -0.2494      0.051     -4.861      0.000      -0.350      -0.149
rchar      -4.983e-06   4.87e-07    -10.237      0.000   -5.94e-06   -4.03e-06
wchar      -5.196e-06   1.04e-06     -4.992      0.000   -7.24e-06   -3.16e-06
pgout         -0.4979      0.089     -5.575      0.000      -0.673      -0.323
ppgout        -0.0675      0.081     -0.837      0.403      -0.226       0.091
pgfree         0.1456      0.049      2.966      0.003       0.049       0.242
pgscan     -1.835e-14   7.77e-17   -236.170      0.000   -1.85e-14   -1.82e-14
atch           0.5881      0.143      4.107      0.000       0.307       0.869
pgin           0.0518      0.029      1.778      0.076      -0.005       0.109
ppgin         -0.0846      0.020     -4.178      0.000      -0.124      -0.045
pflt          -0.0324      0.002    -16.409      0.000      -0.036      -0.029
vflt          -0.0062      0.001     -4.399      0.000      -0.009      -0.003
runqsz        -1.7242      0.126    -13.649      0.000      -1.972      -1.477
freemem       -0.0005   5.12e-05     -9.155      0.000      -0.001      -0.000
freeswap   9.223e-06   1.91e-07     48.401      0.000    8.85e-06    9.6e-06
==============================================================================
Omnibus:                      980.750   Durbin-Watson:                   2.022
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1932.140
Skew:                          -1.040   Prob(JB):                         0.00
Kurtosis:                       4.940   Cond. No.                     7.38e+21
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 2.09e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Figure 14: Linear Regression

Observation on initial Linear Regression:

- We have R-squared 0.739 and Adjusted R-squared 0.792
- F-statistic is 1093
- Coefficient of each feature for this initial linear regression model is mostly in negative. The coefficients show how a unit change in X has an effect on the y variable. A positive or negative sign on the coefficient denotes a positive or negative correlation, respectively.

Multicollinearity check:

Multicollinearity occurs when the predictor variables are correlated in the regression model. This correlation is a problem because predictors must be independent. If the variables are highly collinear, we may not be able to rely on the p-value to identify statistically significant independent variables.

Variance Inflation factor technique is used to identify the multicollinearity between the variables.

```
VIF values:

const       25.629887
lread        5.222496
lwrite       4.230336
scall        2.987824
sread        6.555928
swrite       5.666273
fork        13.195160
exec         3.216047
rchar        2.088006
wchar        1.583353
pgout       11.215199
ppgout      30.947431
pgfree      17.468614
pgscan            NaN
atch         1.848297
pgin        14.475734
ppgin       14.673035
pflt        11.703237
vflt        15.370510
runqsz       1.151214
freemem      1.974160
freeswap     1.847552
dtype: float64
```

Figure 15: VIF Values

The VIF values are to uniquely identify the top variables with high collinearity between variables. pgscan is called out has more than the 75% of zeroes. Pgscan will be dropped from the training dataset and new regression model will be created.

Model 1:

R-squared - 0.793
Adj. R-squared – 0.792
pgout has highest VIF value, it will be dropped to build Model 2.

Model 2:

Based on the VIF value pgout is dropped and new model is created.
R-squared - 0.793
Adj. R-squared – 0.792
vflt has highest VIF value, it will be dropped to build Model 3.

Model 3:

Based on the VIF value vflt is dropped and new model is created.
R-squared - 0.792
Adj. R-squared – 0.791
ppgin has highest VIF value, it will be dropped to build Model 4.

Model 4:

Based on the VIF value ppgin is dropped and new model is created.
R-squared - 0.791
Adj. R-squared – 0.791
fork has highest VIF value, it will be dropped to build Model 5.

Model 5:

Based on the VIF value fork is dropped and new model is created.
R-squared - 0.791
Adj. R-squared – 0.790
sread has highest VIF value, it will be dropped to build Model 6.

Model 6:

Based on the VIF value sread is dropped and new model is created.
R-squared - 0.791
Adj. R-squared – 0.790
pgout has highest VIF value, it will be dropped to build Model 7.

Model 7:

Based on the VIF value pgout is dropped and new model is created.
R-squared - 0.789
Adj. R-squared – 0.788
lread has highest VIF value, it will be dropped to build Model 8.

Model 8:

Based on the VIF value lread is dropped and new model is created.
R-squared - 0.793
Adj. R-squared – 0.792
lread has highest VIF value, but R-squared value is lesser while comparing with pflt therefore pflt will be dropped to build Model 9.

Model 9:

Based on the VIF value pflt is dropped and new model is created.
R-squared - 0.736
Adj. R-squared – 0.735
lread has highest VIF value, it will be dropped to build Model 10.

Model 10:

Based on the VIF value lread is dropped and new model is created.
R-squared - 0.725
Adj. R-squared – 0.725

scall has highest VIF value, it will be dropped to build Model 11.

Model 11:

Based on the VIF value scall is dropped and new model is created.
R-squared - 0.721
Adj. R-squared – 0.721
wchar has p value > 0,05, it will be dropped to build Model 12.

Model 12:

Based on the VIF value wchar is dropped and new model is created.
R-squared - 0.721
Adj. R-squared – 0.720
atch has p value > 0,05, it will be dropped to build Model 13.

Model 13:

Based on the VIF value atch is dropped and new model is created.
R-squared - 0.721
Adj. R-squared – 0.720

The below predicted & fitted data frame has been derived.

	Actual Values	Fitted Values	Residuals
0	81.0	86.560085	-5.560085
1	93.0	90.534799	2.465201
2	64.0	67.462342	-3.462342
3	86.0	85.458306	0.541694
4	94.0	98.038729	-4.038729

Figure 16: Actual & Residual values

Assumptions of Linear Regression:
These assumptions are essential conditions that should be met before we draw inferences regarding the model estimates or use the model to make a prediction.
For Linear Regression, we need to check if the following assumptions hold:-

- Linearity
- Independence
- Homoscedasticity
- Normality of error terms
- No strong Multicollinearity

TEST FOR LINEARITY AND INDEPENDENCE
Why the test?

- Linearity describes a straight-line relationship between two variables, predictor variables must have a linear relation with the dependent variable.

How to check linearity?

- Make a plot of fitted values vs residuals. If they don't follow any pattern (the curve is a straight line), then we say the model is linear otherwise model is showing signs of non-linearity.

How to fix if this assumption is not followed?

- We can try different transformations. Below plot shows the fitted and residual values of the regression model.

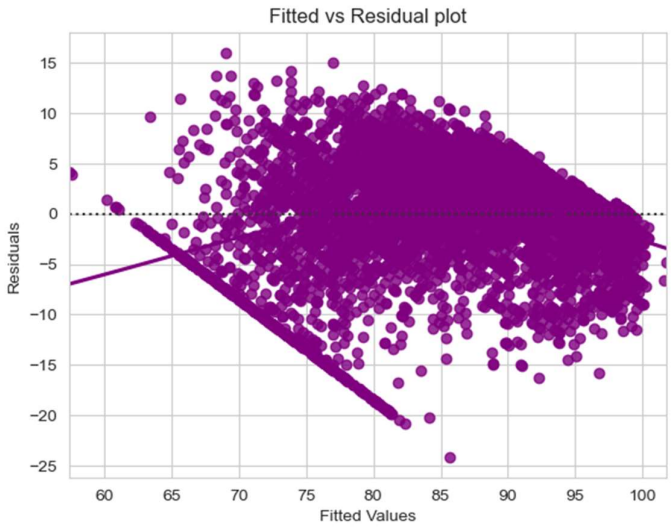

Figure 17: Actual Vs Residual values

- We can observe a pattern in the residual vs fitted values, hence we will try to transform the continuous variables in the data.

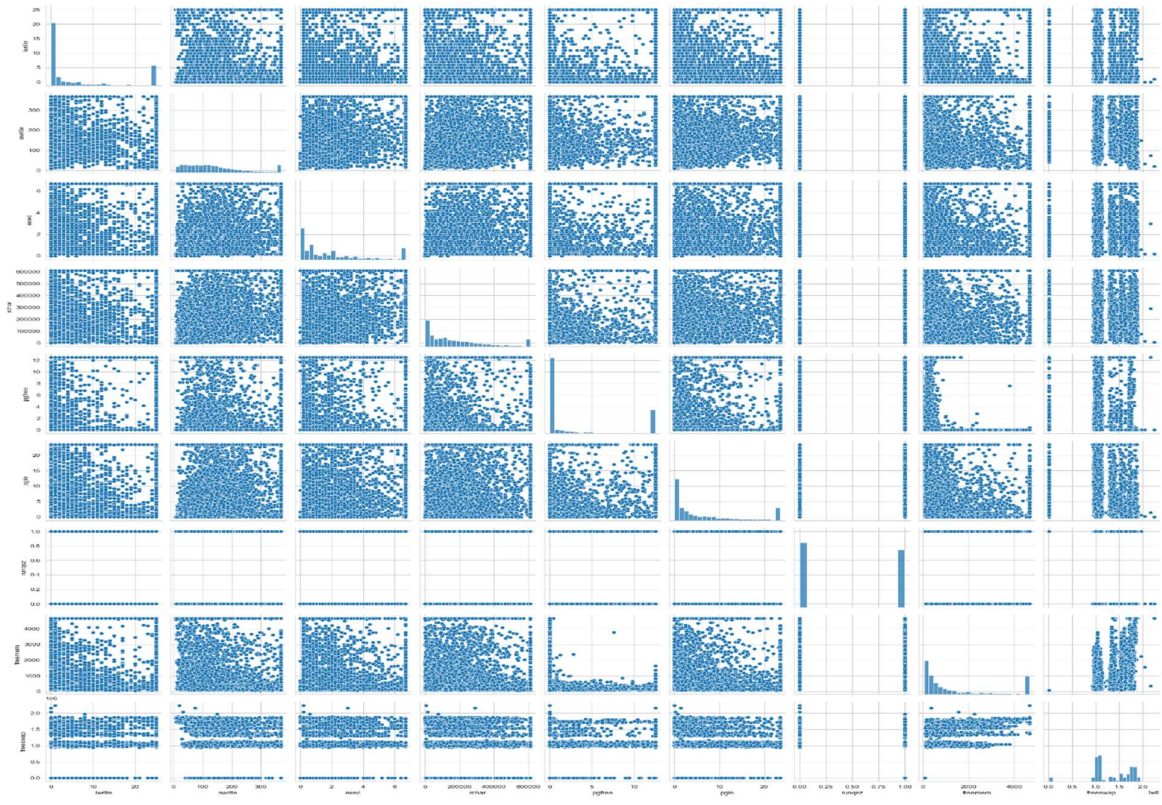

Figure 18: Pair Plot

From the above Pair plot we can see 'swrite and freeswap ' column has a slight nonlinear relationship with 'usr'. We can transform the swrite and freeswap' variables by square the values and 2 new columns will be introduced to the dataset swrite_sq and freeswap_sq respectively.

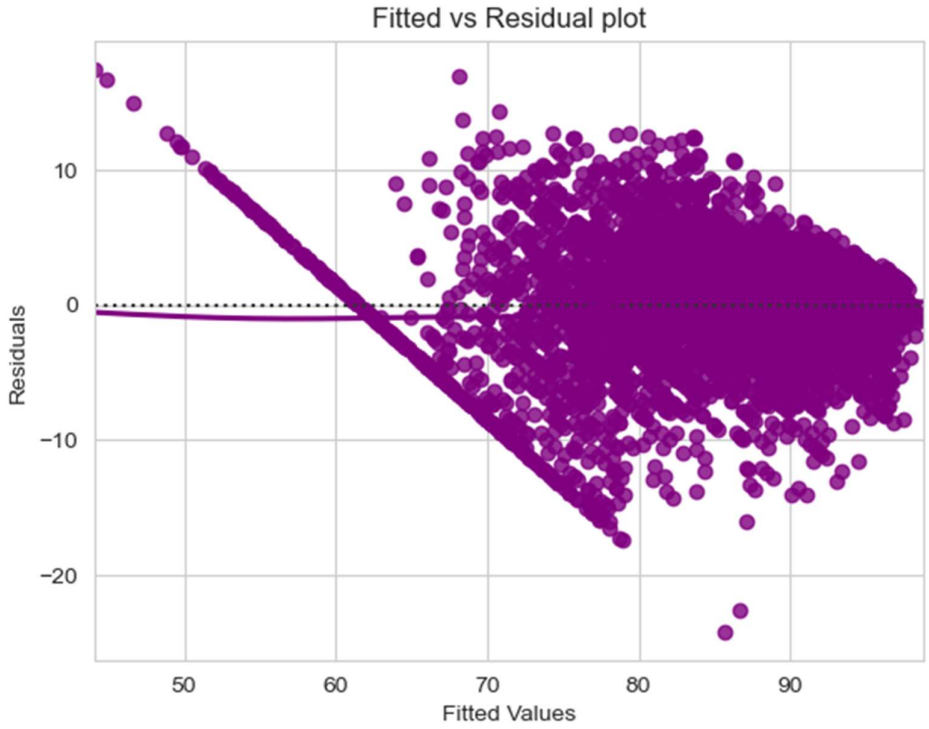

Figure 19: Actual Vs Residual values After transformation

This transformation makes the model more effective which can be seen through R-squared: 0.817 and Adj R-squared 0.817.

TEST FOR NORMALITY

What is the test?

- Error terms/residuals should be normally distributed.
- If the error terms are not normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares.

What does non-normality indicate?

- It suggests that there are a few unusual data points which must be studied closely to make a better model.

How to check the Normality?

- It can be checked via QQ Plot - residuals following normal distribution will make a straight line plot, otherwise not.
- Another test to check for normality is the Shapiro-Wilk test.

How to Make residuals normal?

- We can apply transformations like log, exponential, arcsinh, etc as per our data.

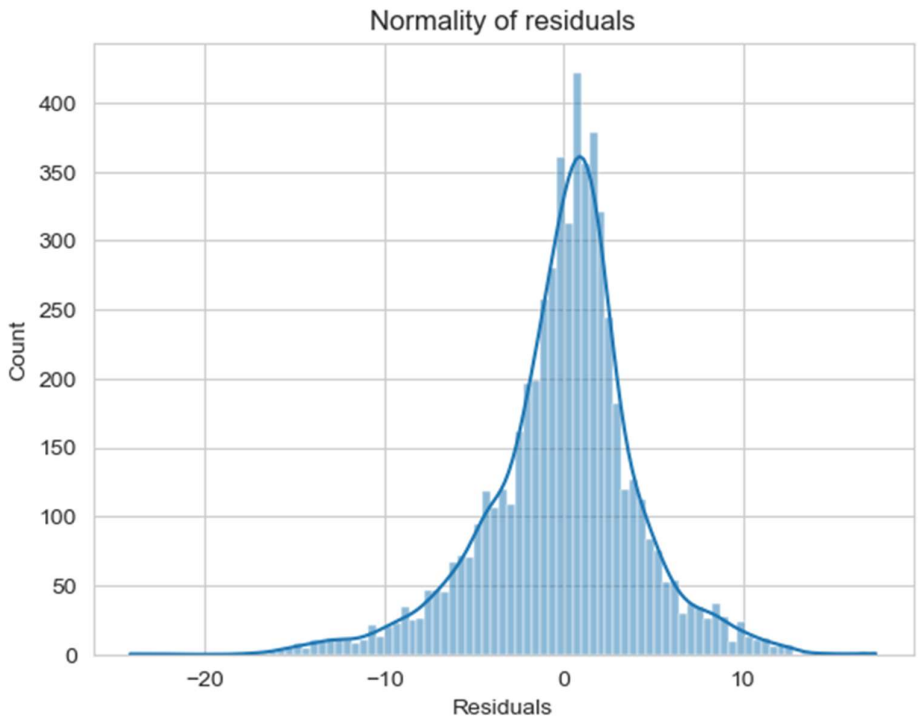

Figure 20: Normality of Residuals

As we can see in the above plot, the residuals are nearly normally distributed

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

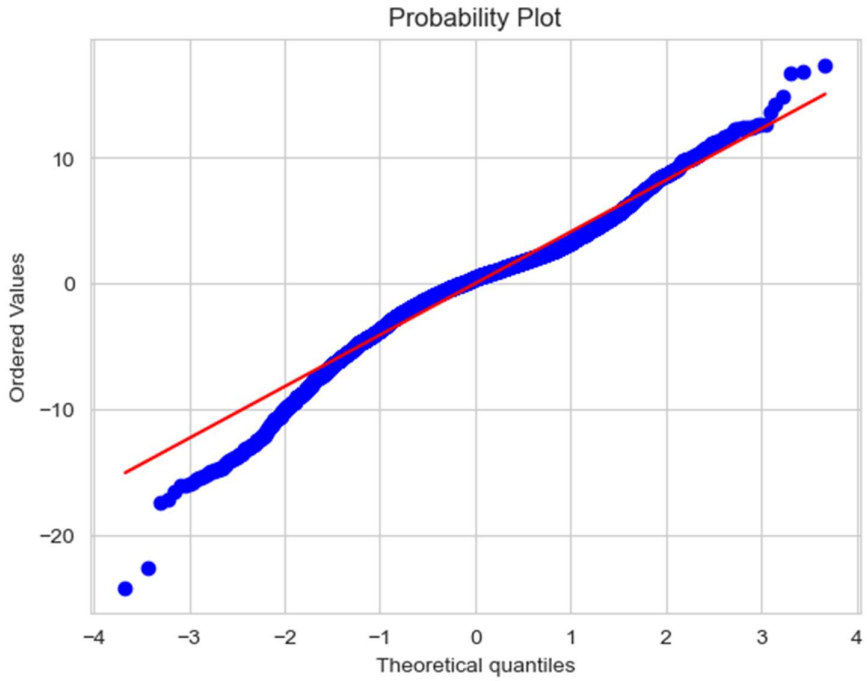

Figure 21: Probability Plot

- Major points are lying on the straight line in QQ plot

The Shapiro-Wilk test can also be used for checking the normality. The null and alternate hypotheses of the test are as follows:

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

Shapiro-Wilk test result :

Statistic = 0.9659520387649536
Pvalue = 9.229837001750153e-35

- Since p-value < 0.05, the residuals are not normal as per shapiro test.
- Strictly speaking - the residuals are not normal. However, as an approximation, we might be willing to accept this distribution as close to being normal

Test For Homoscedasticity

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

- Heteroscedacity- If the variance is unequal for the residuals across the regression line, then the data is said to be heteroscedastic. In this case the residuals can form an arrow shape or any other non-symmetrical shape.

Why the test?

- The presence of non-constant variance in the error terms results in heteroscedasticity. Generally, non-constant variance arises in presence of outliers.

How to check if model has Heteroscedasticity?

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

How to deal with Heteroscedasticity?

- Can be fixed via adding other important features or making transformations.

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

- Null hypothesis: Residuals are homoscedastic
- Alternate hypothesis: Residuals have hetroscedasticity

HOMOSCEDASTICITY test result
F statistic = 1.0110951626451194
p-value = 0.3840883958116449

- Since p-value > 0.05 we can say that the residuals are homoscedastic.

Final model:

All the assumptions of linear regression are now satisfied. Let's check the summary of our final model.

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.817
Model:	OLS	Adj. R-squared:	0.817
Method:	Least Squares	F-statistic:	2321.
Date:	Fri, 14 Jun 2024	Prob (F-statistic):	0.00
Time:	06:28:45	Log-Likelihood:	-16331.
No. Observations:	5734	AIC:	3.269e+04
Df Residuals:	5722	BIC:	3.277e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	74.8472	0.365	204.817	0.000	74.131	75.564
lwrite	-0.0311	0.006	-5.155	0.000	-0.043	-0.019
swrite	-0.0300	0.003	-11.488	0.000	-0.035	-0.025
exec	-2.0179	0.030	-67.131	0.000	-2.077	-1.959
rchar	-6.113e-06	3.76e-07	-16.264	0.000	-6.85e-06	-5.38e-06
pgfree	-0.0735	0.014	-5.371	0.000	-0.100	-0.047
pgin	-0.1588	0.009	-17.693	0.000	-0.176	-0.141
runqsz	-0.4667	0.122	-3.840	0.000	-0.705	-0.228
freemem	0.0004	5.07e-05	8.571	0.000	0.000	0.001
freeswap	3.487e-05	5.19e-07	67.251	0.000	3.39e-05	3.59e-05
freeswap_sq	-1.256e-11	2.39e-13	-52.492	0.000	-1.3e-11	-1.21e-11
swrite_sq	-1.382e-05	6.3e-06	-2.193	0.028	-2.62e-05	-1.46e-06

Omnibus:	457.699	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1074.531
Skew:	-0.491	Prob(JB):	4.66e-234
Kurtosis:	4.880	Cond. No.	1.46e+13

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.46e+13. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 22: Final Model

Observations:

- R-squared of the model is 0.817 and adjusted R-squared is 0.817, which shows that the model is able to explain ~81% variance in the data. This is quite good.

Predictions:

Model Parameters:

const	74.847189
lwrite	-0.031144
swrite	-0.02995
exec	-2.017936
rchar	-0.000006
pgfree	-0.073518
pgin	-0.158754
runqsz	-0.466721

freemem	0.000434
freeswap	0.000035
freeswap_sq	-0.000000
swrite_sq	-0.000014

Equation of Linear Regression:

usr = 74.8471891993446 + -0.031144299605293357 * (lwrite) + -0.029950302359855907 * (swrite) + -2.017935849804573 * (exec) + -6.112959549440973e-06 * (rchar) + -0.073517857785101 * (pgfree) + -0.15875359921697435 * (pgin) + -0.46672076582632216 * (runqsz) + 0.00043445364032663585 * (freemem) + 3.487109155021237e-05 * (freeswap) + -1.2556835025774306e-11 * (freeswap_sq) + -1.3823165912853517e-05 * (swrite_sq)

Observations:

- Freemem is the only positive feature which has a positive tendency towards usr (CPU runs in user mode). When Freemem unit increases chances of CPU runs in user mode increases.
- A unit increase in the freemem will result in a 0.0004 unit increase in the usr, all other variables remaining constant.
- The usr of a process of CPU_Bound will be -0.4667 units lesser than a process of Not_CPU_Bound, all other variables remaining constant.

Predictions of test dataset:

RMSE OF TRAIN DATA = 4.175
RMSE OF TEST DATA = 4.13

MAE OF TRAIN DATA = 3.866
MAE OF TEST DATA = 2.978

Observations:

- We can see that RMSE on the train and test sets are comparable. So, our model is not suffering from overfitting.
- MAE indicates that our current model is able to predict usr within a mean error of 2.9 units on the test data.
- Hence, we can conclude the final model is good for prediction as well as inference purposes.

Inference

We constructed a number of models by removing variables one at a time in order to produce an effective model. By taking into account several aspects like R-squared, Adj R-squared, P value, and creating VIF, the variables are eliminated. On beforehand we have to clean up the data by handling the outliers and impute the missing values before moving on to the linear regression model. We have tried to build a Linear Regression without treating the outliers which gave us a very low R-squared value which shows the model is not efficient.

Linear Regression before Outlier treatment: R-squared and Adjusted R-squared value are 73.9% and 79.2% respectively. This value is considered has a very low score therefore we have moved over to build an effective liner regression model.

Below is the iteration we have gone to bring the linear regression model.

Variables	R-Squared	Adj R-Squared
pgscan	0.793	0.792
pgout	0.793	0.792
vflt	0.792	0.791
ppgin	0.791	0.791
sread	0.791	0.790
pgout	0.789	0.788
pflt	0.736	0.35
lread	0.725	0.725
scall	0.721	0.721
wchar	0.721	0.720
atch	0.721	0.720

Table 1: Features List dropped

Variable column has the variables that we have dropped one by one, corresponding changes in R- squared and Adj R-squared are filled up beside to it. It shows we have started with 0.793 and ended with 0.721 even though it shows negative improvement in R-squared value, the reason we have choosen to drop the variables is they have Multicollinearity within the independent variables which effects the effective model therefore we have removed the variables which has Multicollinearity.

To improve the model, we have transformed the swrite and freeswap variables by square the values and 2 new columns will be introduced to the dataset swrite_sq and freeswap_sq respectively. Which makes the model more effective and it can be measured by the R-squared: 0.817 and Adj R- squared 0.817 values.

For linear regression the residuals have to be in normal distributed, in our model the residuals are build up close to normal distribution form which make the model very effective. Shapiro test which helps to identify if the residuals are in normal distribution, p value (Pvalue = 9.229837001750153e-35) on the Shapiro test is lesser than 0.05 therefore it is proved the residual is not normally distributed.

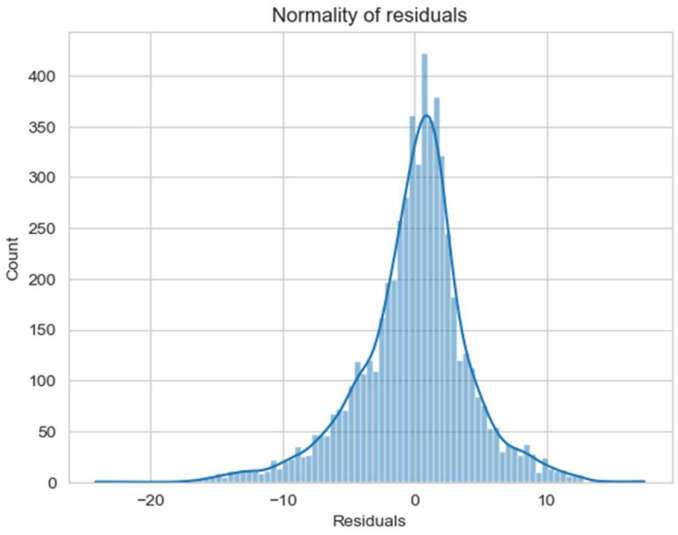

Figure 23: Normality of Residuals

Homoscedasticity test is performed to check if the presence of non-constant variance in the error terms results in heteroscedasticity. The P-value on Homoscedasticity test is 0.3840883958116449 therefore the null hypothesis is rejected so that we can say that the residuals are homoscedastic.

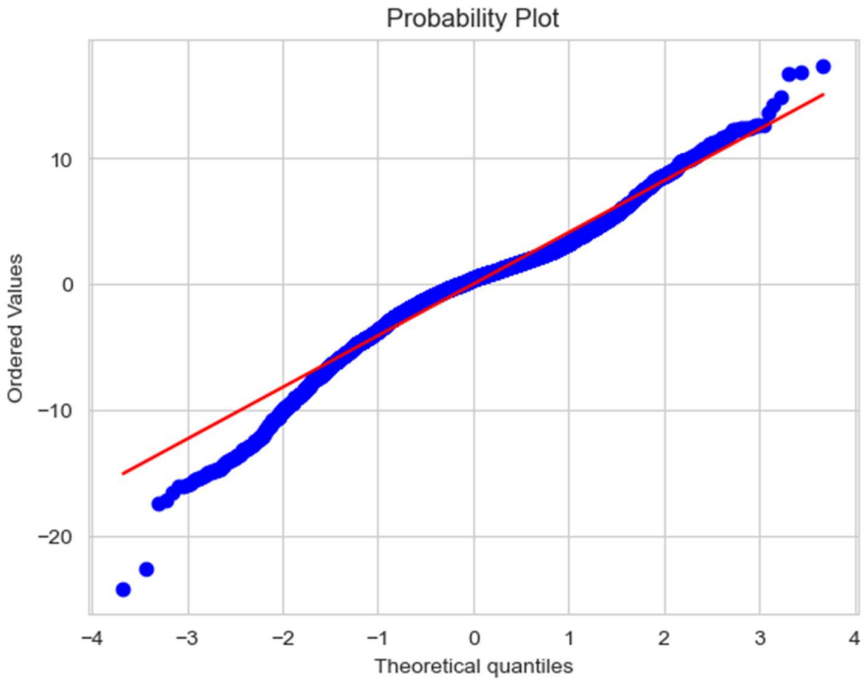

Figure 24: QQ Plot of Residuals

QQ plot shows the majority of residuals are on the linear line. This is an evidence of an effective linear regression model.

Last but not the least, our model worked very well in both training and test data. This is tested using RMSE and MAE. RMSE on the train and test sets are comparable (Train data: 4.175 & Test data:4.13). Therefore, our model is not suffering from overfitting. MAE indicates that our current model is able to predict usr within a mean error of 2.9 units on the test data.

Recommendations:

usr -Portion of time (%) that CPUs run in user mode can be predicted using the below linear regression equation.

usr = 74.8471891993446 + -0.031144299605293357 * (lwrite) +
-0.029950302359855907 * (swrite) + -2.017935849804573 * (exec) +
-6.112959549440973e-06 * (rchar) + -0.073517857785101 * (pgfree) +
-0.15875359921697435 * (pgin)
+ -0.46672076582632216 * (runqsz) +
0.00043445364032663585 * (freemem)
+ 3.487109155021237e-05 * (freeswap) +
-1.2556835025774306e-11 * (freeswap_sq)
+ -1.3823165912853517e-05 * (swrite_sq)

The dependent variable urs -Portion of time CPUs run in user mode rises as the following variables' units fall.

lwrite - writes (transfers per second) between system memory and user memory

swrite - Number of system write calls per second.

exec - Number of system exec calls per second.

rchar - Number of characters transferred per second by system read calls

pgfree - Number of pages per second placed on the free list.

pgin - Number of page-in requests per second

ppgin - Number of pages paged in per second

runqsz - Process run queue size (

freemem - Number of memory pages available to user processes

freeswap - Number of disk blocks available for page swapping.

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

Through this model, we advise that there is a greater likelihood of an increase in the amount of time CPUs are used in user mode when the aforementioned factors are used sparingly.

Problem – 2 : Logistic Regression, LDA & CART

Objective

In your role as a statistician at the Republic of Indonesia Ministry of Health, you have been entrusted with a dataset containing information from a Contraceptive Prevalence Survey. This dataset encompasses data from 1473 married females who were either not pregnant or were uncertain of their pregnancy status during the survey.

Your task involves predicting whether these women opt for a contraceptive method of choice. This prediction will be based on a comprehensive analysis of their demographic and socio-economic attributes.

Data Description

Wife's age (numerical)

Wife's education (categorical) 1=uneducated, 2, 3, 4=tertiary

Husband's education (categorical) 1=uneducated, 2, 3, 4=tertiary

Number of children ever born (numerical)

Wife's religion (binary) Non-Scientology, Scientology

Wife's now working? (binary) Yes, No

Husband's occupation (categorical) 1, 2, 3, 4(random)

Standard-of-living index (categorical) 1=verlow, 2, 3, 4=high

Media exposure (binary) Good, Not good

Contraceptive method used (class attribute) No,Yes

Solution :

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

Figure 25: Dataset Sample

Exploratory Data Analysis (EDA):

```
<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 26: Data types of the dataset

- There are 80 duplicates in the data set and all the the duplicates are dropped.

Statistical Summary after dropping duplicates is shown below,

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Wife_age	1326.0	NaN	NaN	NaN	32.557315	8.289259	16.0	26.0	32.0	39.0	49.0
Wife_ education	1393	4	Tertiary	515	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_education	1393	4	Tertiary	827	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_of_children_born	1372.0	NaN	NaN	NaN	3.290816	2.399697	0.0	1.0	3.0	5.0	16.0
Wife_religion	1393	2	Scientology	1186	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Wife_Working	1393	2	No	1043	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_Occupation	1393.0	NaN	NaN	NaN	2.174444	0.85459	1.0	1.0	2.0	3.0	4.0
Standard_of_living_index	1393	4	Very High	618	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Media_exposure	1393	2	Exposed	1284	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Contraceptive_method_used	1393	2	Yes	779	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 27: Statistical summary of the dataset

- All the numerical columns have numerical values alone.

- From the above 5 point summary we can observe that count for each feature is shown has 1393 after removing the duplicates.

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

```
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 28: Null value check before & after of the treatment

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


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


Wife_religion
Scientology        1186
Non-Scientology     207
Name: count, dtype: int64


Wife_Working
No     1043
Yes     350
Name: count, dtype: int64


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


Media_exposure
Exposed        1284
Not-Exposed     109
Name: count, dtype: int64


Contraceptive_method_used
Yes    779
No     614
Name: count, dtype: int64
```

Figure 29: Unique values of all the object data types

Univariant, Bivariant & Multivariant Analysis:

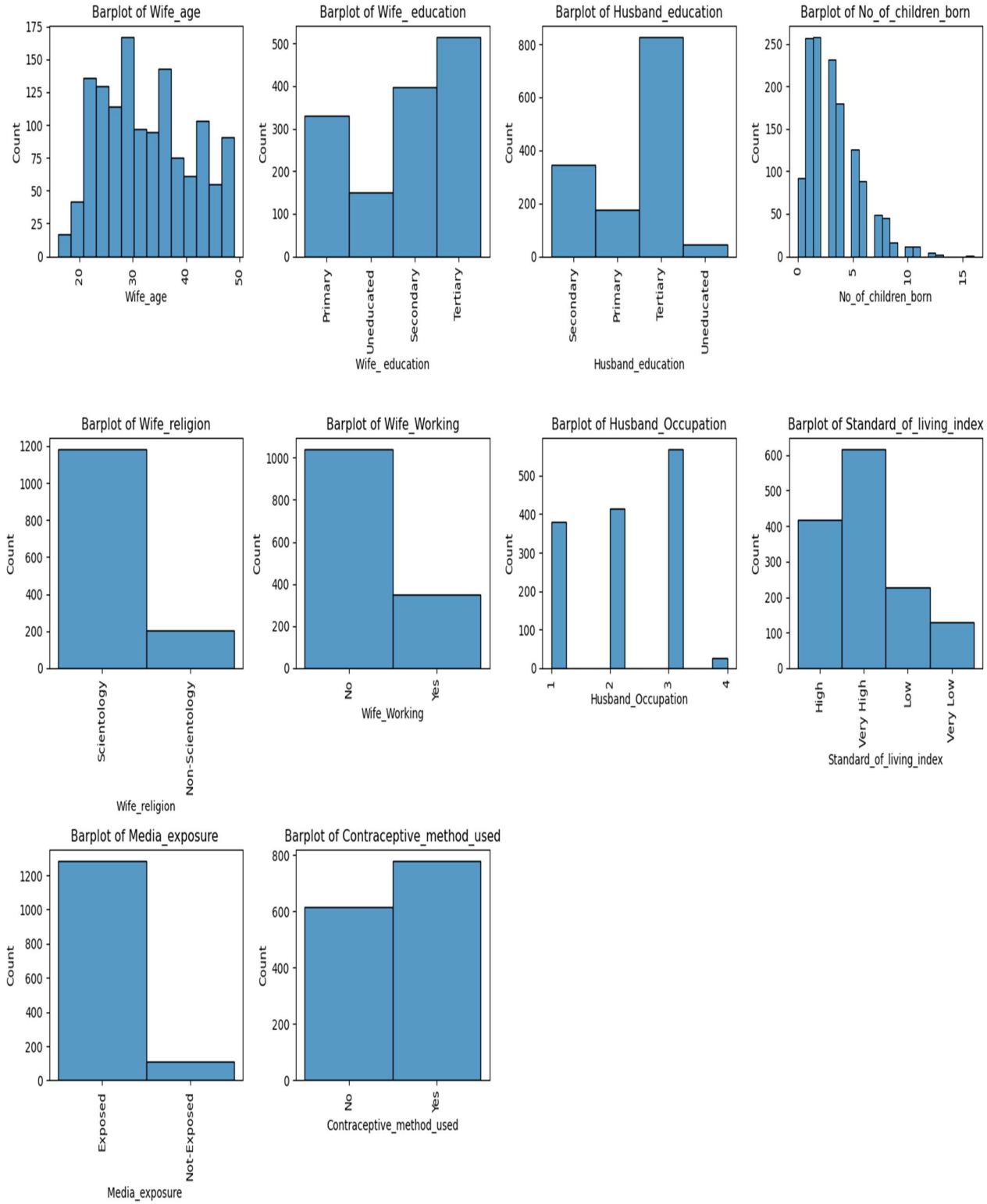

Figure 30: Univariant Analysis of the dataset

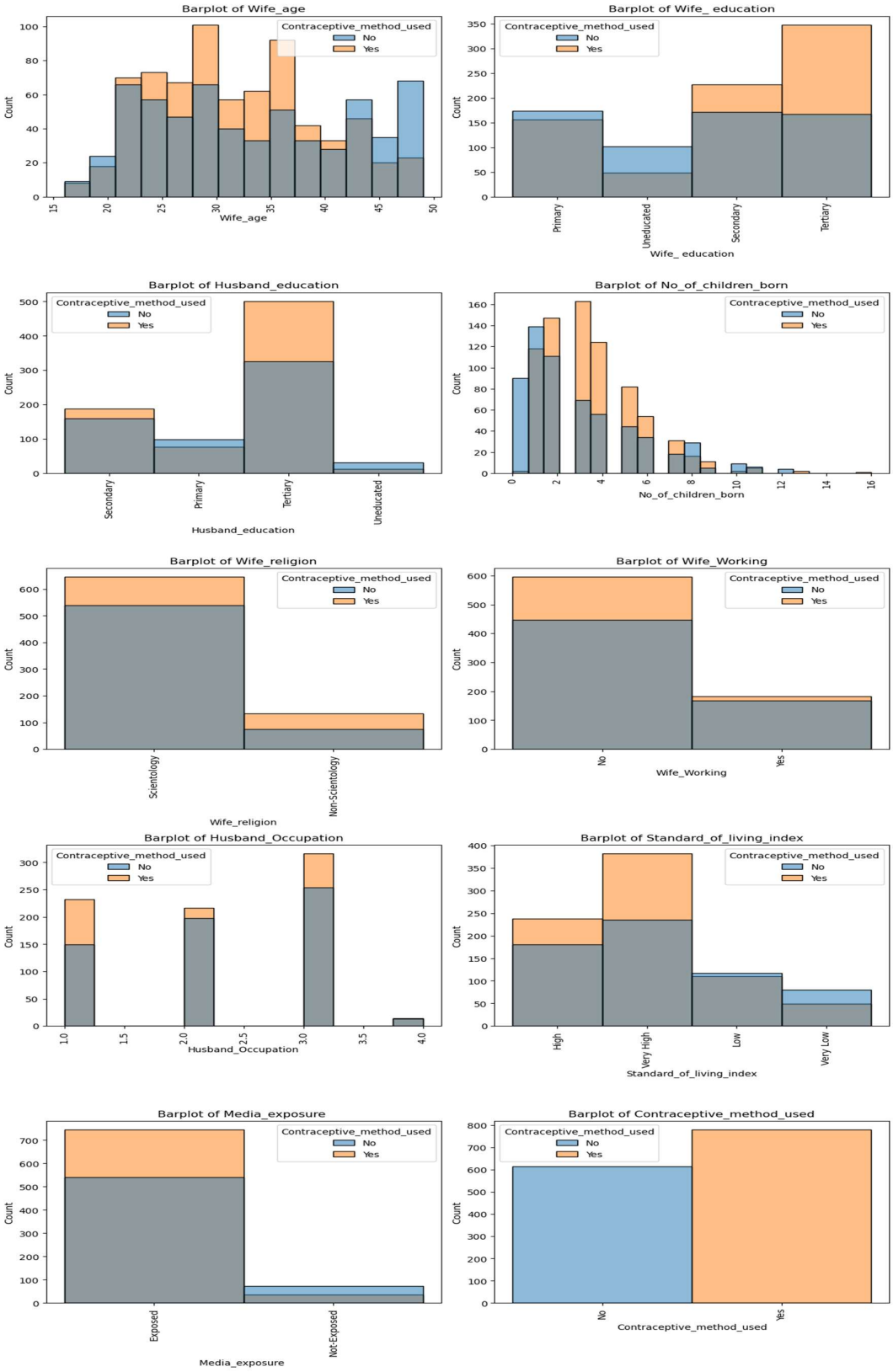

Figure 31: Bivariant Analysis of the dataset

- All variables are neatly distributed.

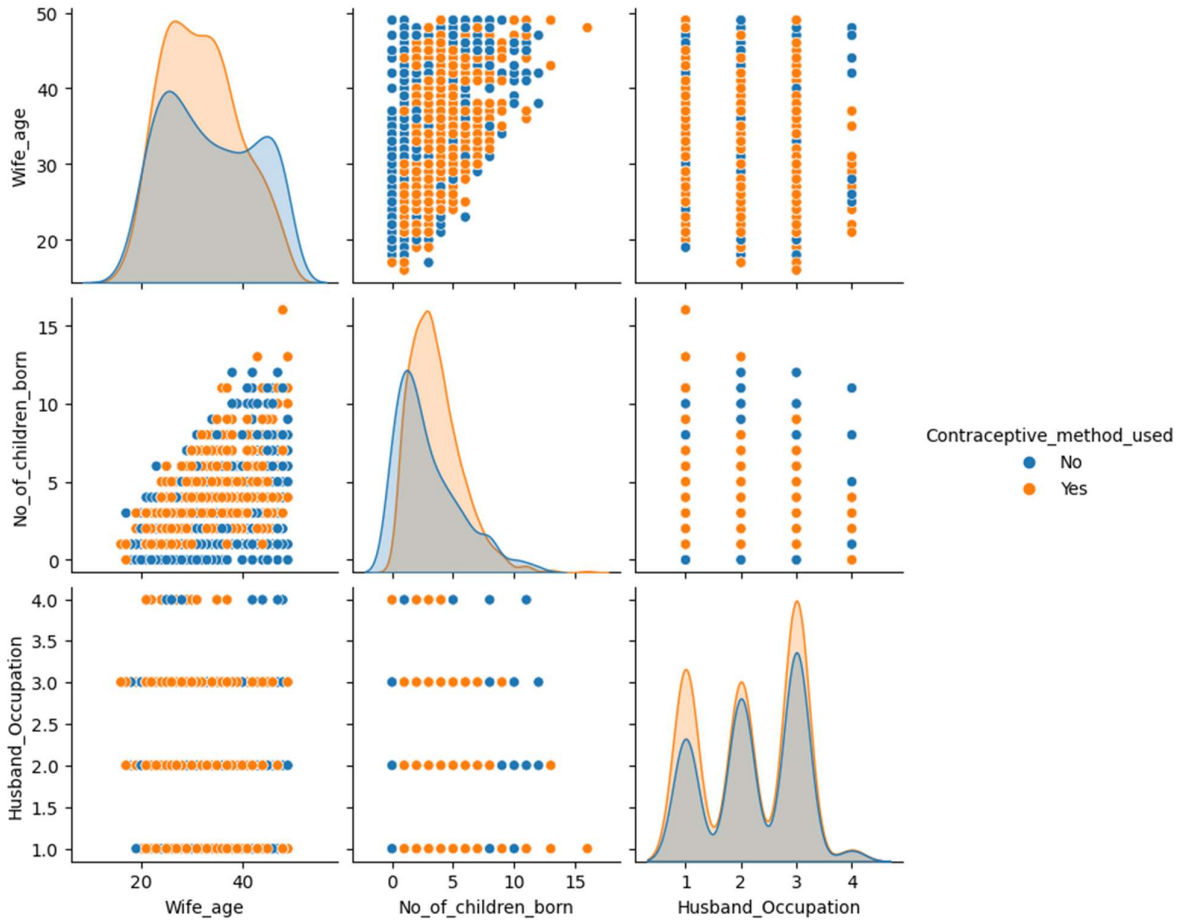

Figure 32: Multivariant Analysis of the dataset

- There is no variance in the depth of variables, scattered data will help models to perform well
- Each variable has equivalent contribution of Contraceptive_method_used dependent variable

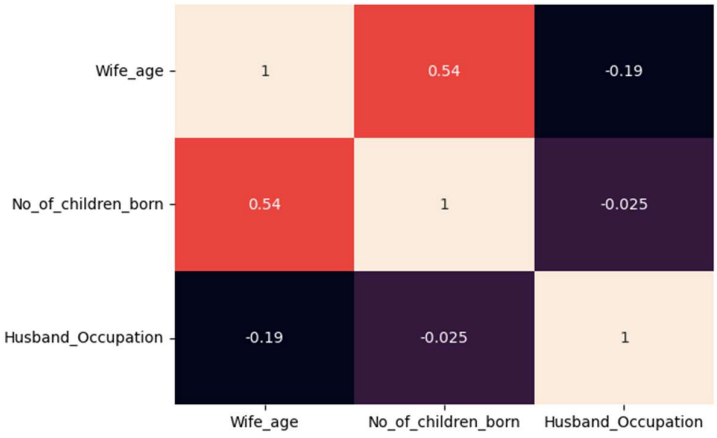

Figure 33: Heatmap of the dataset

- Wife_age vs No_of_children_born has correlation of 54%

Outlier Treatment:

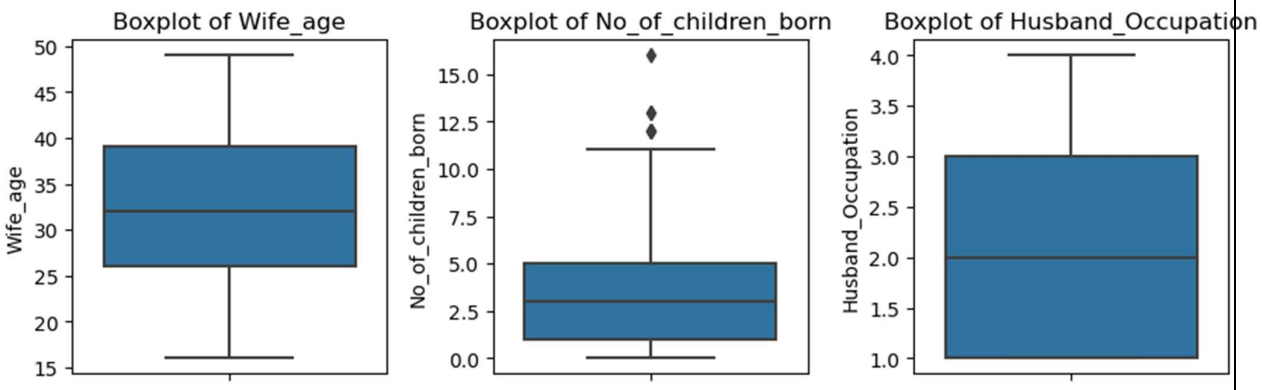

Figure 34: Before outlier treatment

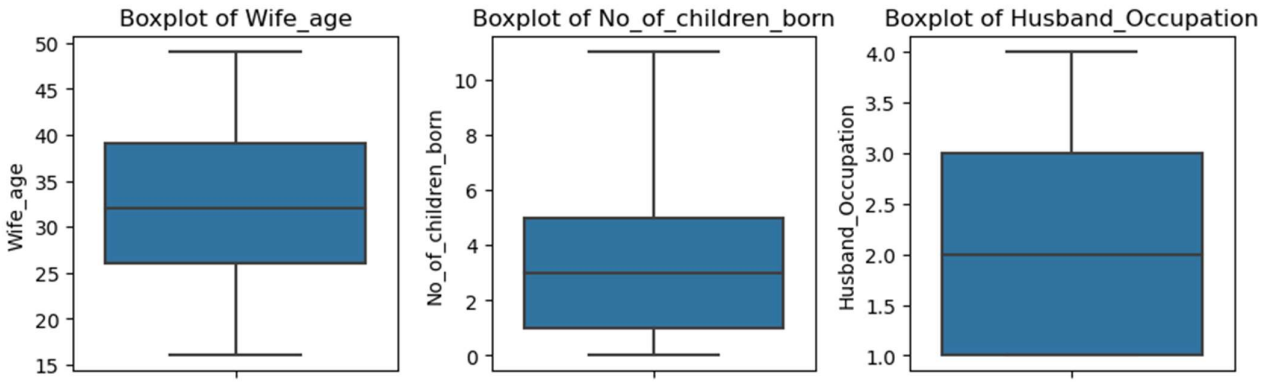

Figure 35: After outlier treatment

	count	mean	std	min	25%	50%	75%	max
Wife_age	1393.0	32.56	8.09	16.0	26.0	32.0	38.0	49.0
Wife_ education	1393.0	1.35	0.96	0.0	1.0	1.0	2.0	3.0
Husband_education	1393.0	1.53	0.75	0.0	1.0	2.0	2.0	3.0
No_of_children_born	1393.0	3.29	2.38	0.0	1.0	3.0	5.0	16.0
Wife_religion	1393.0	0.85	0.36	0.0	1.0	1.0	1.0	1.0
Wife_Working	1393.0	0.25	0.43	0.0	0.0	0.0	1.0	1.0
Husband_Occupation	1393.0	2.17	0.85	1.0	1.0	2.0	3.0	4.0
Standard_of_living_index	1393.0	1.33	1.00	0.0	0.0	2.0	2.0	3.0
Media_exposure	1393.0	0.08	0.27	0.0	0.0	0.0	0.0	1.0
Contraceptive_method_used	1393.0	0.56	0.50	0.0	0.0	1.0	1.0	1.0

Figure 36: Statistical summary after encoding

Split Data:

Contraceptive_method_used variable has taken has a y variable (dependent variable) and all other variables are taken has x variable (independent variable).

The given data set is split into 70:30; 70% data is taken as training data and 30% of data are taken for testing the model.

	Wife_age	Wife_education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposı
1265	35.0	0	2	4.0	0	0	3	1	
850	33.0	0	0	2.0	1	0	3	2	
791	31.0	1	2	5.0	1	0	3	3	
294	29.0	2	2	1.0	1	0	1	0	
289	35.0	3	1	0.0	1	0	2	1	

Figure 37: Training dataset

	Wife_age	Wife_education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposı
327	35.0	2	2	2.0	0	0	3	2	
1154	25.0	1	0	1.0	1	1	3	0	
1023	23.0	2	2	1.0	1	1	3	0	
536	28.0	2	2	2.0	1	1	1	0	
652	30.0	0	0	4.0	1	0	2	2	

Figure 38: Testing dataset

Models:

Logistic Regression:

Using logistic regression we are trying to predict the dependent variable; logistic regression is used in predicting the categorical dependent variable. To perform the regression, model the data set has to be all numeric, to achieve this we have encoded all the object data in the dataset to numeric.

Logistic Regression Model Score: 0.6564102564102564

As shown above we have obtained 65.6 as a Logistic Regression Model Score

AUC on the training: 0.665
AUC on the test: 0.665

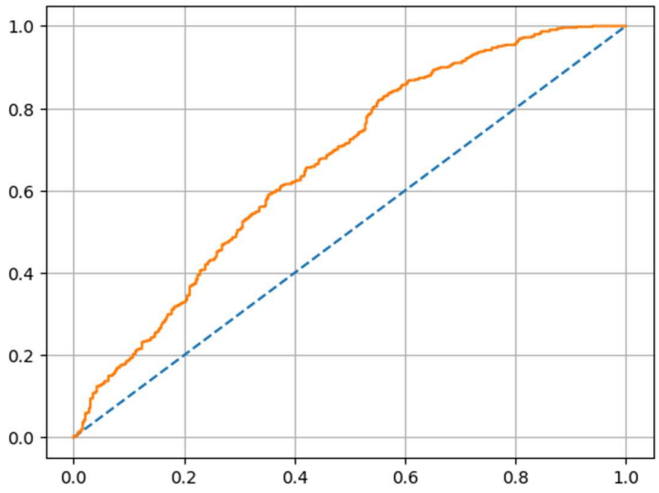

Figure 39: AUC Curve of Training Data Set

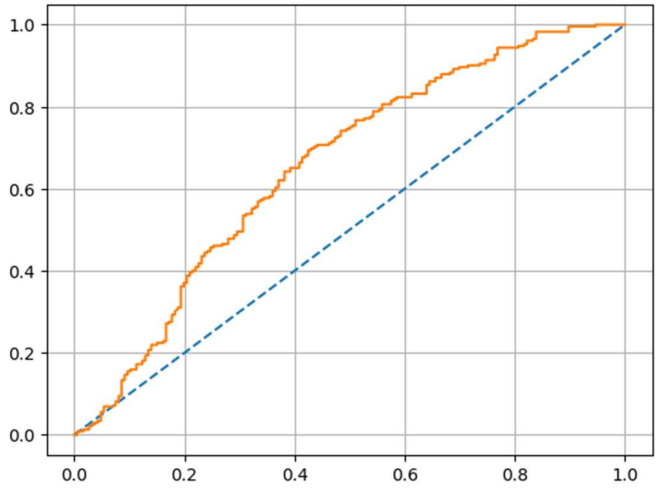

Figure 40 : AUC Curve of Testing Data Set

From the Figure 35 & 36 we could clearly visualize Logistic regression model is performed well in both the Train and Test data.

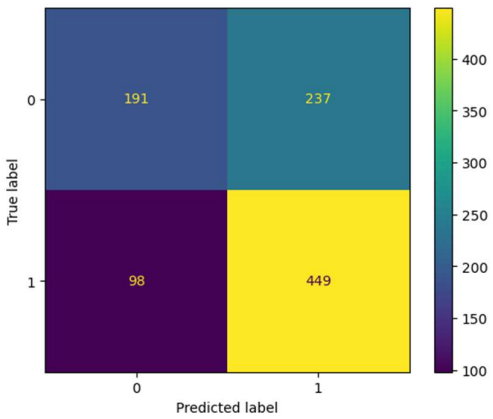

Figure 41: Confusion Matrix of training Data Set

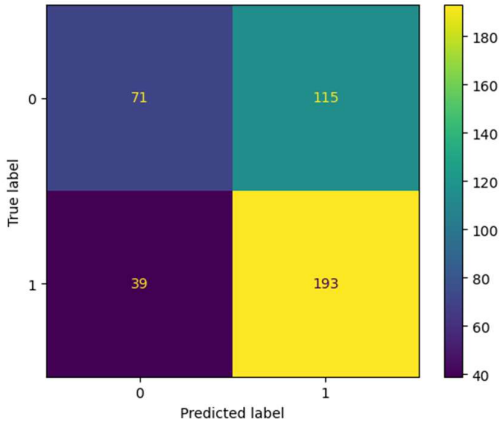

Figure 42: Confusion Matrix of testing Data Set

Inference from Train data:

- 449 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Contraceptive method used No is predicted as No
- 237 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Contraceptive method used Yes is predicted as Yes
- 191 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 98 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Inference from Test data:

- 193 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Contraceptive method used No is predicted as No
- 115 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Contraceptive method used Yes is predicted as Yes
- 71 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 39 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Classification Report:

	precision	recall	f1-score	support
0	0.66	0.45	0.53	428
1	0.65	0.82	0.73	547
accuracy			0.66	975
macro avg	0.66	0.63	0.63	975
weighted avg	0.66	0.66	0.64	975

Figure 43: Classification Report of training Data Set

	precision	recall	f1-score	support
0	0.65	0.38	0.48	186
1	0.63	0.83	0.71	232
accuracy			0.63	418
macro avg	0.64	0.61	0.60	418
weighted avg	0.64	0.63	0.61	418

Figure 44: Classification Report of testing Data Set

LDA (Linear Discriminant Analysis):

LDA = 2.35 + (-0.0787 * Wife_age) + (0.187 * Wife_ education)
+ (0.177 * Husband_education) + (0.243 * No_of_children_born)
+ (-0.673 * Wife_religion) + (-0.114 * Wife_Working)
+ (-0.043 * Husband_Occupation) + (-0.024 * Standard_of_living_index)
+ (-1.216 * Media_exposure)

With LDA model we came up with the above equation; equation starts with constant and all the variable have their coefficient. Based on the value of coefficient the variable contributes on prediction of y (dependent) variable.

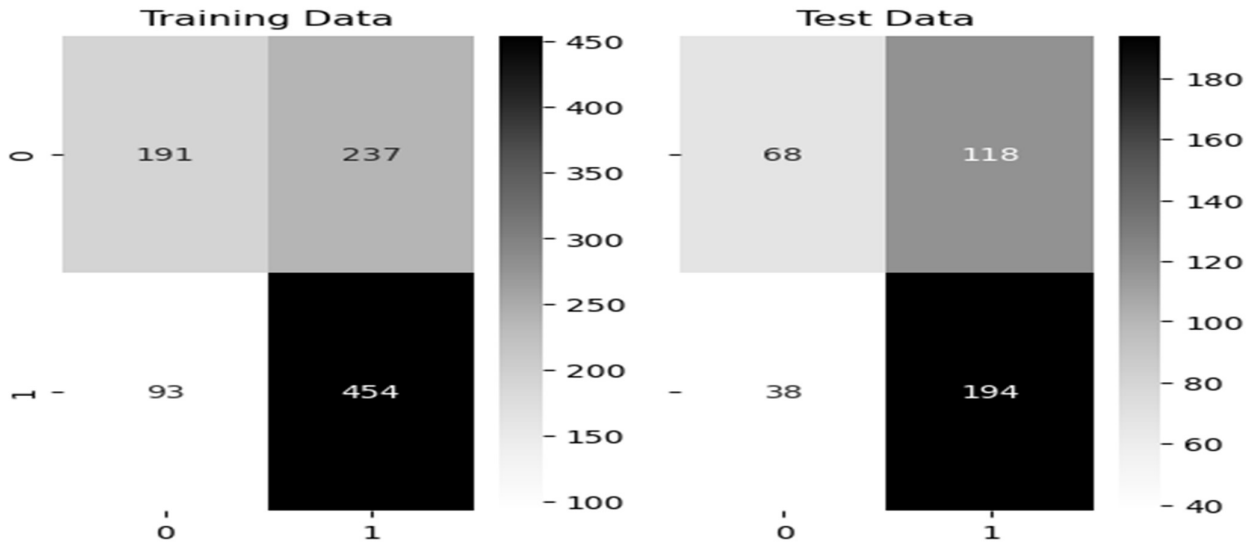

Figure 45: Confusion Matrix of training & testing Data Set

Inference from Train data:

- 454 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Contraceptive method used No is predicted as No
- 237 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Contraceptive method used Yes is predicted as Yes
- 191 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 93 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Inference from Test data:

- 194 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Contraceptive method used No is predicted as No

- 118 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Contraceptive method used Yes is predicted as Yes
- 68 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 38 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.67	0.45	0.54	428
1	0.66	0.83	0.73	547
accuracy			0.66	975
macro avg	0.66	0.64	0.63	975
weighted avg	0.66	0.66	0.65	975

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.64	0.37	0.47	186
1	0.62	0.84	0.71	232
accuracy			0.63	418
macro avg	0.63	0.60	0.59	418
weighted avg	0.63	0.63	0.60	418

Figure 46: Classification Report of training & testing Data Set

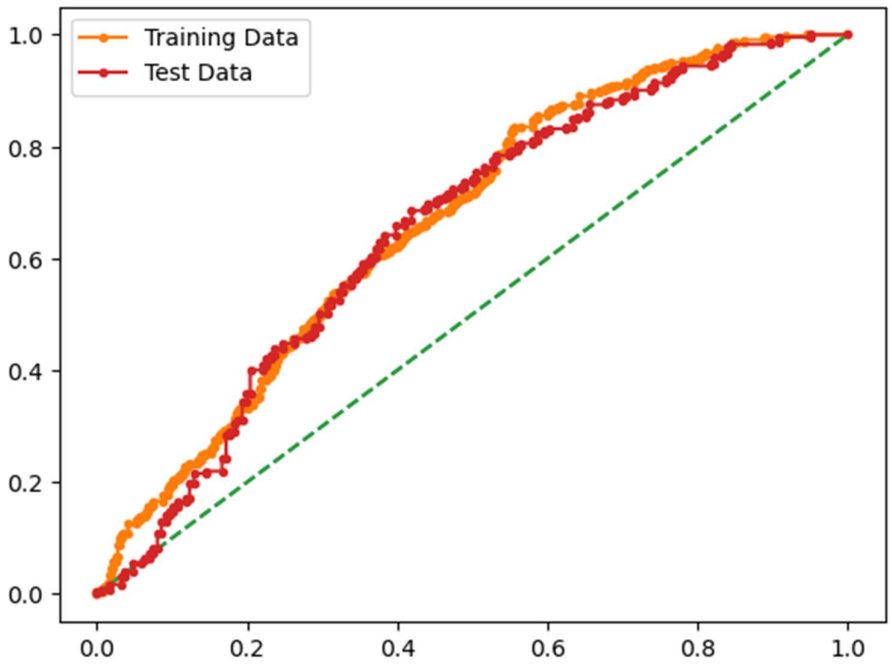

Figure 47: AUC chart of training & testing Data Set

From the Figure 42 we could clearly visualize LDA model is performed well in both the Train and Test data

CART Model:

Performed CART model to predict the dependent variable, in our data set "Contraceptive method used" is the dependent variable, where other variables are used to predict "Contraceptive method used"

Feature	Coefficient
Wife_age	0.29922
No_of_children_born	0.255744
Wife_ education	0.100791
Husband_education	0.084069
Husband_Occupation	0.07782
Standard_of_living_index	0.074705
Wife_religion	0.041152
Wife_Workin	0.04026
Media_exposure	0.026238

Table 2: CART Model Coefficients

In CART model we could clearly see the entire coefficients are positive. The unit increase in the independent variable likely turns to be a positive impact to dependent variable.

AUC on the training: 0.859
AUC on the test: 0.859

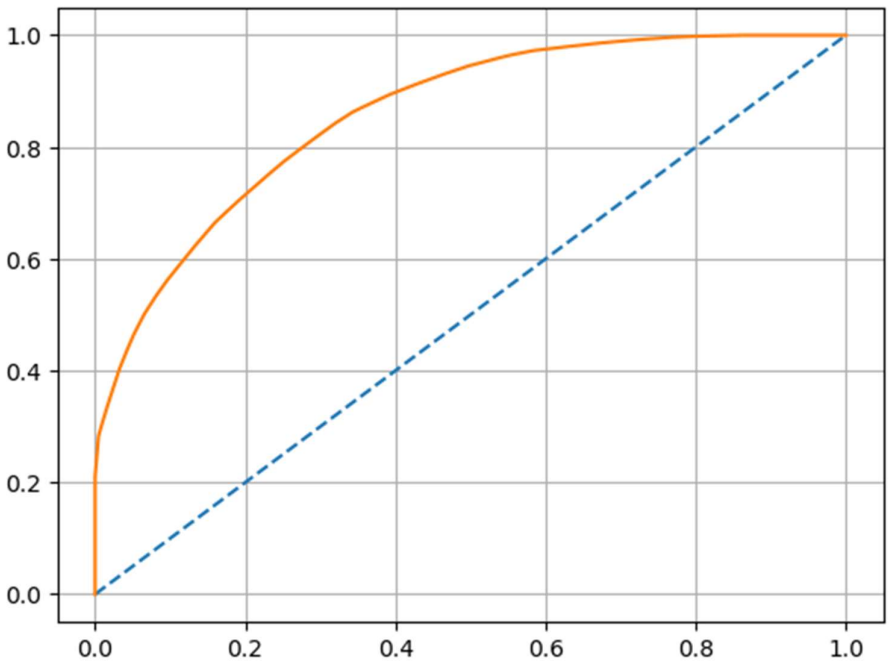

Figure 48: AUC Curve of Training Data Set

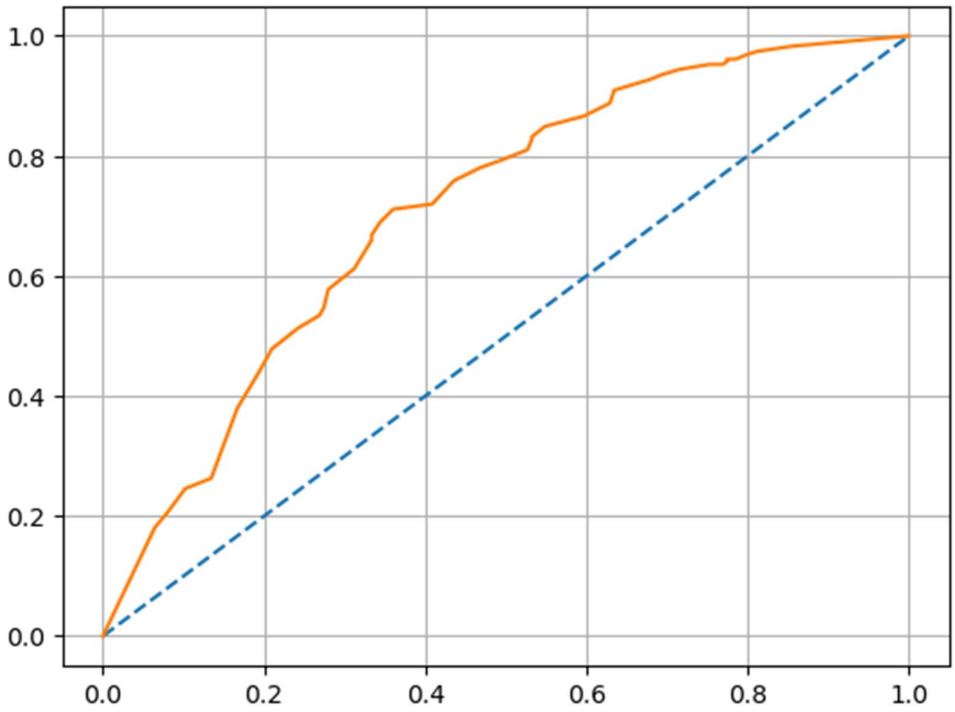

Figure 49: AUC Curve of Testing Data Set

From the Figure 44 we could clearly visualize CART model is performed well in both the Train and Test data

Classification Report:

	precision	recall	f1-score	support
0	0.79	0.66	0.72	428
1	0.76	0.86	0.81	547
accuracy			0.77	975
macro avg	0.78	0.76	0.76	975
weighted avg	0.77	0.77	0.77	975

Figure 50: Classification Report of training Data Set

	precision	recall	f1-score	support
0	0.69	0.47	0.56	186
1	0.66	0.83	0.73	232
accuracy			0.67	418
macro avg	0.67	0.65	0.65	418
weighted avg	0.67	0.67	0.65	418

Figure 51: Classification Report of testing Data Set

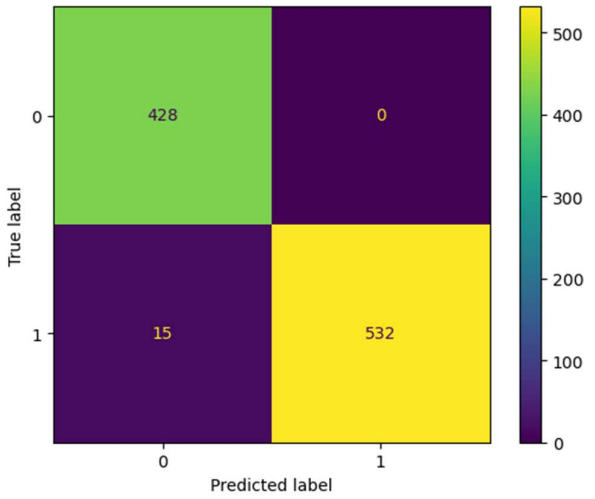

Figure 52: Confusion Matrix of training Data Set

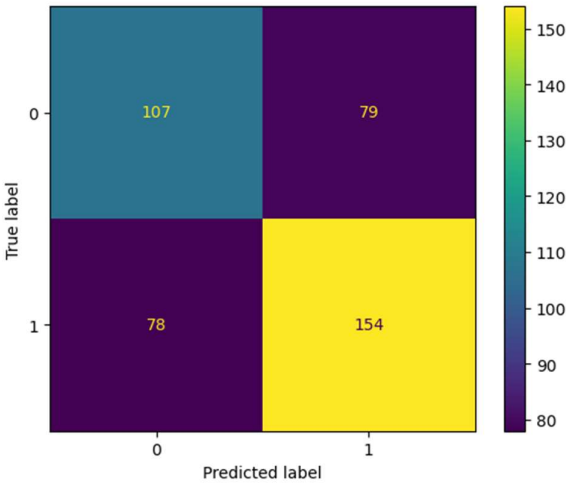

Figure 53: Confusion Matrix of testing Data Set

Inference from Train data:

- 532 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Contraceptive method used No is predicted as No
- 428 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Contraceptive method used Yes is predicted as Yes
- 15 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 0 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Inference from Test data:

- 154 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Contraceptive method used No is predicted as No
- 107 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Contraceptive method used Yes is predicted as Yes
- 79 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 78 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Model Comparison:

Logistic regression, LDA and CART models are thoroughly explained in the before sections. We are here to compare the all 3 models and identify which make more sense with respect you predicting dependent variable (Contraceptive method used).

	Logistic Regression		LDA		CART	
	Train	Test	Train	Test	Train	Test
AUC	0.665	0.665	0.665	0.653	0.859	0.859
Accuracy	0.66	0.64	0.66	0.63	0.77	0.67
Precision 0	0.66	0.65	0.67	0.64	0.79	0.69
Precision 1	0.65	0.63	0.66	0.62	0.76	0.66
Recall 0	0.45	0.38	0.45	0.37	0.66	0.47
Recall 1	0.82	0.83	0.83	0.84	0.86	0.83
F1-Score 0	0.53	0.48	0.54	0.47	0.72	0.56
F1-Score 1	0.73	0.71	0.73	0.71	0.81	0.73

Table 3: Model Comparison Chart

Table 2 helps us to understand how each models came out with the important component like AUC, Accuracy, precision, recall,f1-score. Logistic regression performed well on predicting the dependent variable, but when it is compared with LDA and CART model it shows lesser performance in both train and test data. LDA performed well than Logistic Regression, in precision 0 both Logistic and LDA model outcome are same. Accuracy of LDA is much better than Logistic Regression.

CART performed well than other models.

- CART has highest values in most of the criteria
- Highest Accuracy score 0.85
- Top score 0.79 in precision 0
- Top score 0.86 in recall 1
- Top score 0.81 in f1-score 1
- Performed well in both train and test data

Inference:

We constructed three different models' Logistic regression, LDA and CART models to predict Contraceptive method used dependent variable. By taking into account several aspects like coefficient, AUC, Accuracy, precision, recall, f1-score we were able to compare models between them. On beforehand we did the encoding so make sure the data are ready to build the Logistic regression, LDA and CART models. Outliers are treated and object variables are encoded to convert it to numeric variable.

As explained in the model comparison CART model performed well than the other models. This is evident by reviewing the Table 2.

Below is the coefficient values from CART mode

Feature	Coefficient
Wife_age	0.29922
No_of_children_born	0.255744
Wife_ education	0.100791
Husband_education	0.084069
Husband_Occupation	0.07782

Standard_of_living_index	0.074705
Wife_religion	0.041152
Wife_Workin	0.04026
Media_exposure	0.026238

Where we have highest Coefficient that variable is the main contributor in predicting dependent variable. In our case Contraceptive method used is the dependent variable all other variables are independent variable. All the variable has positive Coefficient, this shows where there is a unit increase in the independent variable, dependent variable has the impact of Coefficient times. For an example, Wife age unit increase impact the Contraceptive method used by 0.29 times, No of children born age unit increase impact the Contraceptive method used by 0.25 times.