PREDICTIVE MODELING PROJECT

DSBA

Contents

Problem 1: Linear Regression

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

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

Dataset for Problem 1: compactiv.xlsx

DATA DICTIONARY:

System measures used:

lread - Reads (transfers per second) between system memory and user memory
lwrite - writes (transfers per second) between system memory and user memory
scall - Number of system calls of all types per second
sread - Number of system read calls per second .
swrite - Number of system write calls per second .
fork - Number of system fork calls per second.
exec - Number of system exec calls per second.
rchar - Number of characters transferred per second by system read calls
wchar - Number of characters 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

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

We have loaded the dataset named compactiv and read the dataset .

- Top 5 rows displayed by head

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap	usr
0	1.0	0.0	2147.0	79.0	68.0	0.2	0.2	40671.000000	53995.0	0.0	...	0.0	0.0	1.6	2.6	16.00	26.40	0.0	4659.125	1730946.0	95.0
1	0.0	0.0	170.0	18.0	21.0	0.2	0.2	448.000000	8385.0	0.0	...	0.0	0.0	0.0	0.0	15.63	16.83	1.0	4659.125	1869002.0	97.0
2	15.0	3.0	2162.0	159.0	119.0	2.0	2.4	197385.728363	31950.0	0.0	...	0.0	1.2	6.0	9.4	150.20	220.20	1.0	702.000	1021237.0	87.0
3	0.0	0.0	160.0	12.0	16.0	0.2	0.2	197385.728363	8670.0	0.0	...	0.0	0.0	0.2	0.2	15.60	16.80	1.0	4659.125	1863704.0	98.0
4	5.0	1.0	330.0	39.0	38.0	0.4	0.4	197385.728363	12185.0	0.0	...	0.0	0.0	1.0	1.2	37.80	47.60	1.0	633.000	1760253.0	90.0

5 rows × 22 columns

- Last 5 rows displayed by tail

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap	usr
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	986647	80
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	1055742	90
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	969106	87
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	1022458	83
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	1756514	94

5 rows × 22 columns

- Information of dataset by 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
```

- Displayed the null values that is present in dataset

  We found that there is null values present in rchar and wchar .

```
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
```

- Shape of the dataset

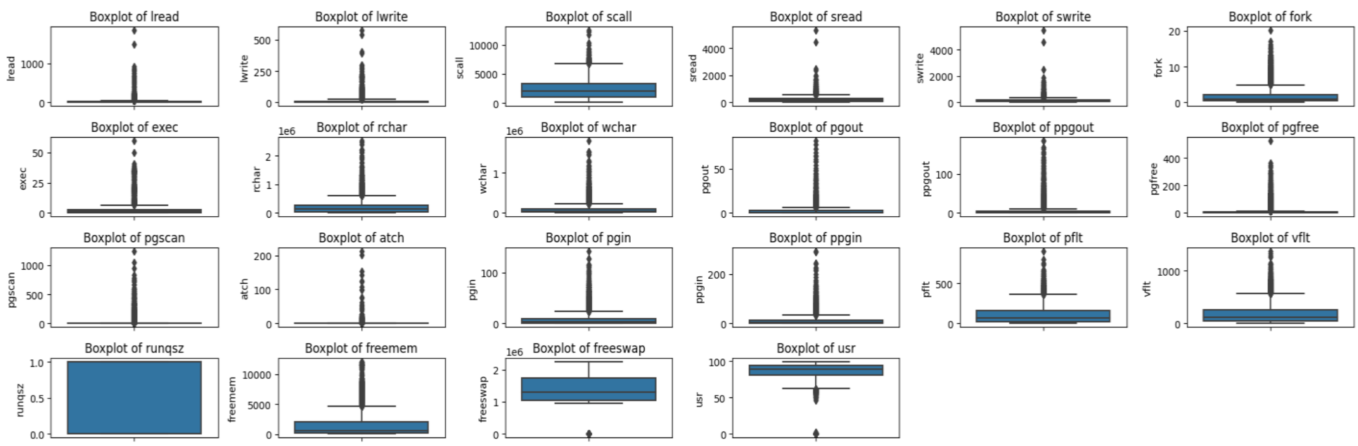
```
(8192, 22)
```

- Describe the dataset shared

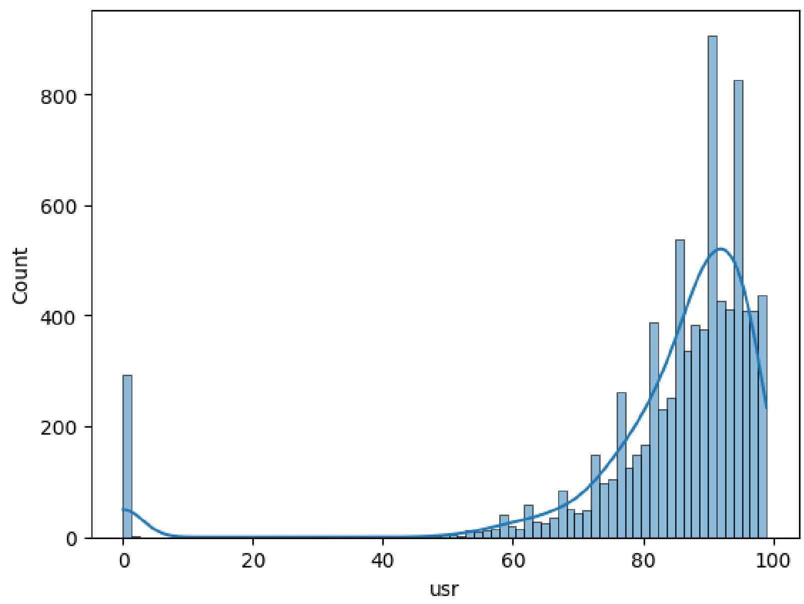
	count	mean	std	min	25%	50%	75%	max
lread	8192.0	1.955969e+01	53.353799	0.0	2.0	7.0	20.000	1845.00
lwrite	8192.0	1.310620e+01	29.891726	0.0	0.0	1.0	10.000	575.00
scall	8192.0	2.306318e+03	1633.617322	109.0	1012.0	2051.5	3317.250	12493.00
sread	8192.0	2.104800e+02	198.980146	6.0	86.0	166.0	279.000	5318.00
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

- UNIVARIATE PLOTS

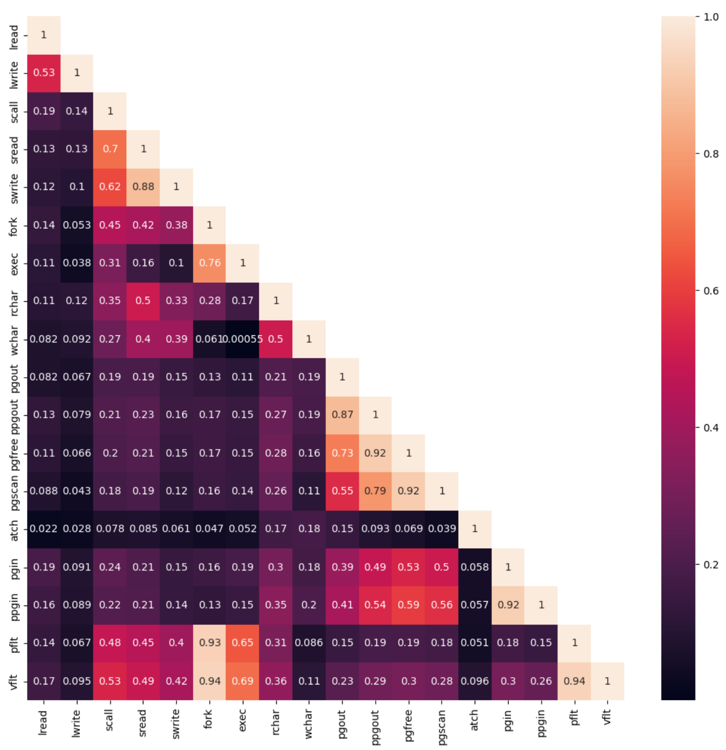
- BOXP PLOT:



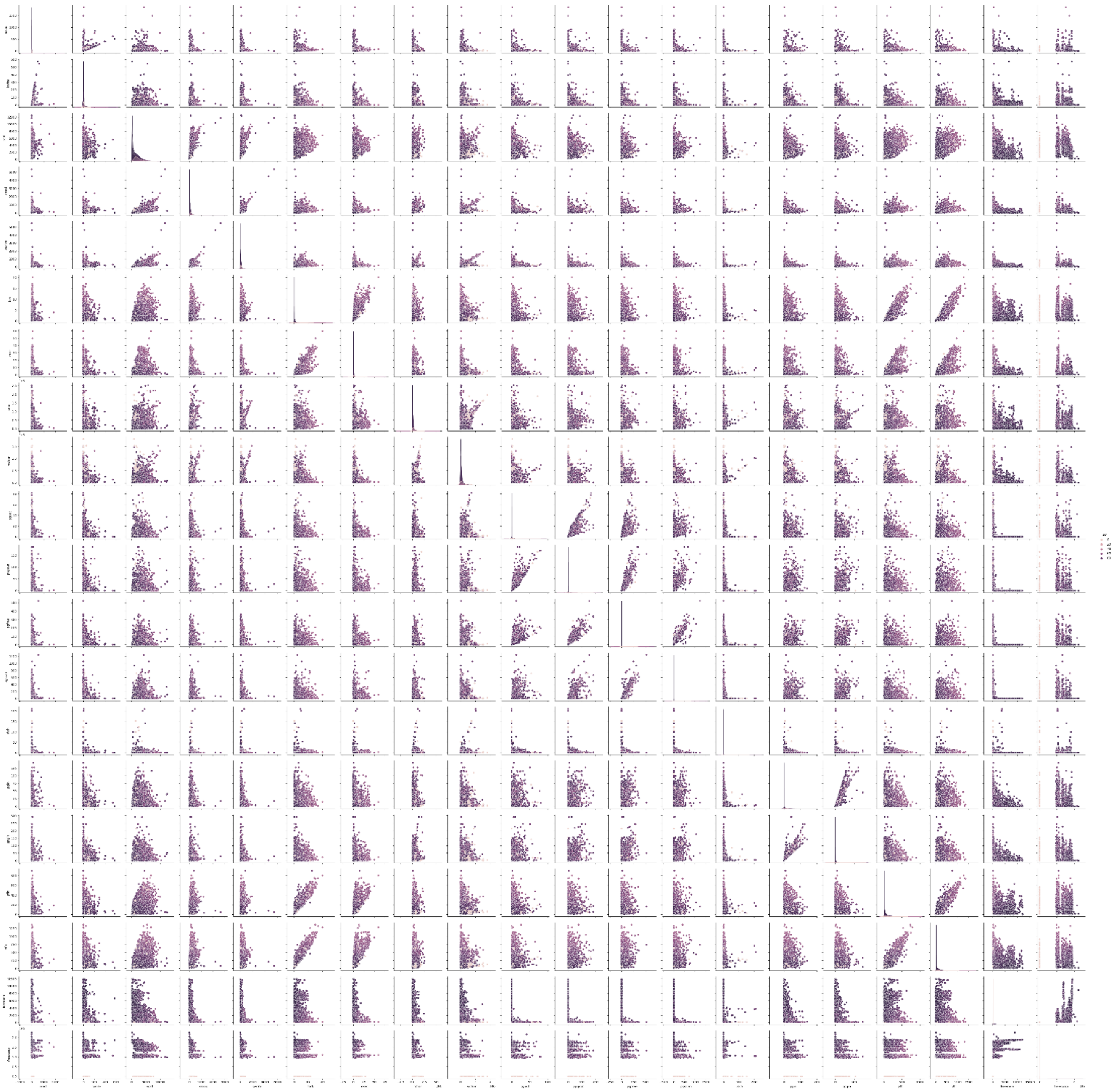
- HISTPLOT FOR UNIVARIATE ANALYSIS



- MULTIVARIATE ANALYSIS
- HEATMAP



- PAIRPLOT



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.

- Check the Duplicates present in dataset

We found that there is no duplicates present in dataset.

```
0
```

- Displayed the null values that is present in dataset

  We found that there is null values present in rchar and wchar .

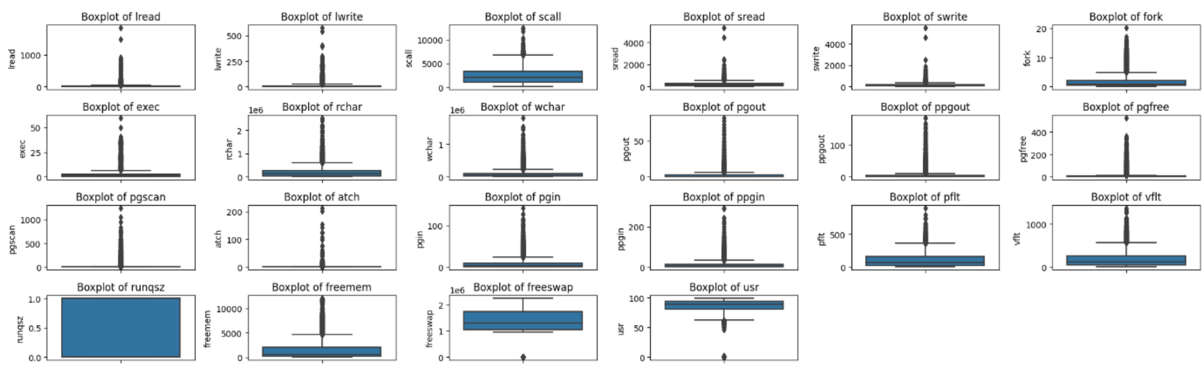
```
lread          0
lwrite         0
scall          0
sread          0
swrite         0
fork           0
exec           0
rchar        104
wchar         15
pgout          0
ppgout         0
pgfree         0
pgscan         0
atch           0
pgin           0
ppgin          0
pflt           0
vflt   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

- As we found that there is null values present in rchar and wchar so we have to impute these values to proceed further. We have to impute the same using mean values
- After imputing we have to check whether there is null values present or not after imputation

```
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
dtype: int64
```

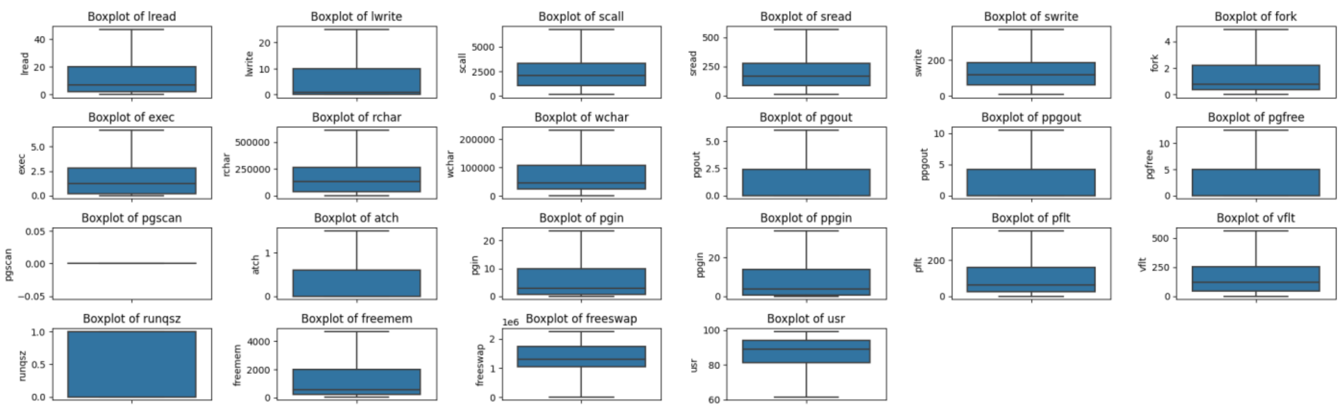
We found that after imputation there is no null values present now.

- Need to check if there is any outliers present in dataset.



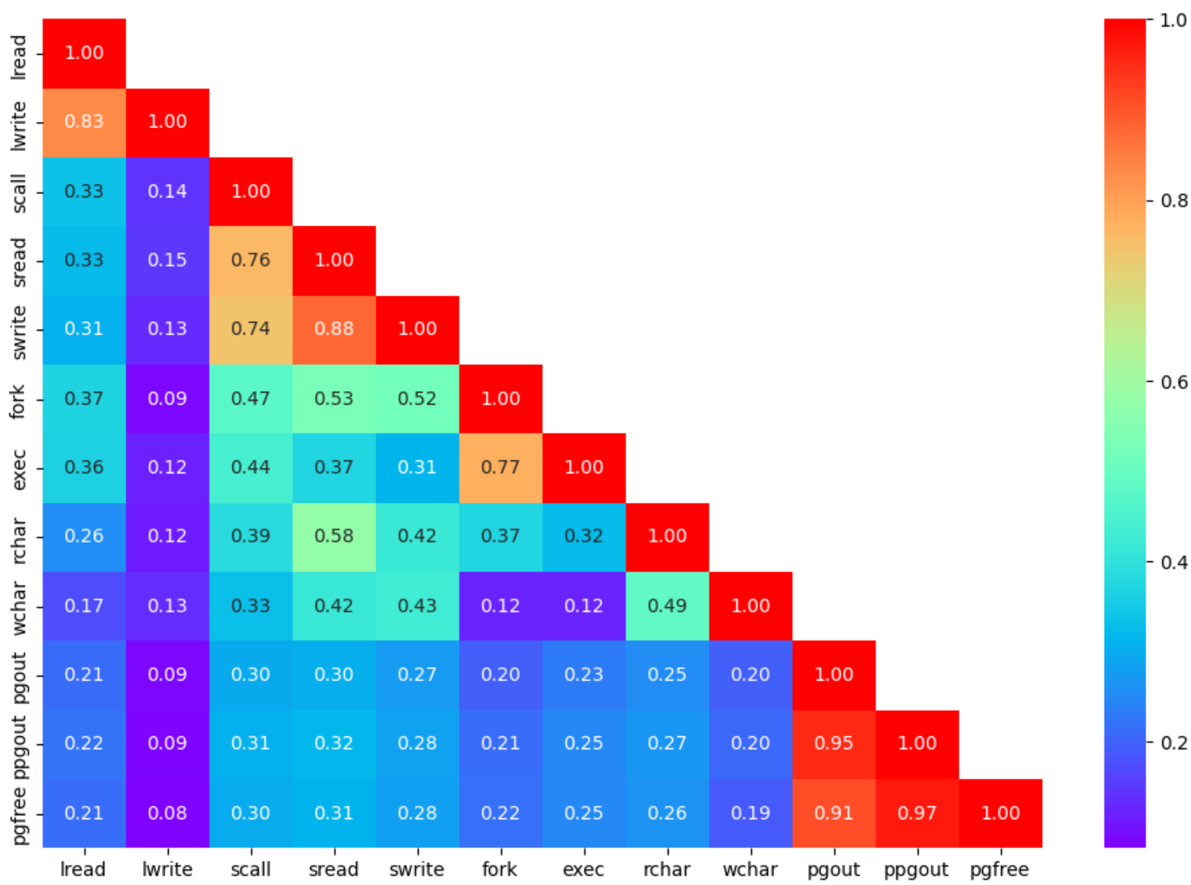
- We have to treat the outliers present so we have to treat the outliers first to move further with dataset . We have used IQR to treat the outliers.

After treating the outliers we have checked the boxplot that if there is any outliers present . So we need to plot the boxplot as shown below



After treating we found that there is no outliers present so we are good to go further with our dataset.

Heatmap (To check the correlation)



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.

- Value count checked for runqsz (unique values for categorical variables)

```
Not_CPU_Bound     4331
CPU_Bound         3861
Name: runqsz, dtype: int64
```

- Categorical variables been changed into dummy variable
- Dataset after encoding

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap	usr
0	1	0	2147	79	68	0.2	0.2	40671.000000	53995.0	0.0	...	0.0	0.0	1.6	2.6	16.00	26.40	0	4670	1730946	95
1	0	0	170	18	21	0.2	0.2	448.000000	8385.0	0.0	...	0.0	0.0	0.0	0.0	15.63	16.83	1	7278	1869002	97
2	15	3	2162	159	119	2.0	2.4	197385.728363	31950.0	0.0	...	0.0	1.2	6.0	9.4	150.20	220.20	1	702	1021237	87
3	0	0	160	12	16	0.2	0.2	197385.728363	8670.0	0.0	...	0.0	0.0	0.2	0.2	15.60	16.80	1	7248	1863704	98
4	5	1	330	39	38	0.4	0.4	197385.728363	12185.0	0.0	...	0.0	0.0	1.0	1.2	37.80	47.60	1	633	1760253	90

5 rows × 22 columns

- Datatypes after encoding we checked
  We found that there is no object data type present . Please refer the below snip

```
<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     8192 non-null   float64
 8   wchar     8192 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   int8   
 19  freemem   8192 non-null   int64  
 20  freeswap  8192 non-null   int64  
 21  usr       8192 non-null   int64  
dtypes: float64(13), int64(8), int8(1)
memory usage: 1.3 MB
```

- Check for multicollinearity using Variance Inflation Factor (VIF)

```
lread ---> 9.326562900259399
lwrite ---> 6.435149952798939
scall ---> 9.006255898359603
sread ---> 18.562678659324423
swrite ---> 16.862194869787487
fork ---> 24.981567255591003
exec ---> 5.916552039350398
rchar ---> 4.287093542695739
wchar ---> 3.393450625593844
pgout ---> 16.192842817459816
ppgout ---> 42.81334193258917
pgfree ---> 24.0386514399627
pgscan ---> nan
atch ---> 2.723143483602382
pgin ---> 23.066405205476794
```

From the above Variance Inflation Factor (VIF) that is one of the method that is used to check the multi-correlation between them. As we know that if they are correlated that is the VIF is more than 10 then it is consider as highly correlated and is not ideal for linear regression . If we have VIF equal to 5 or more than that it is consider as moderately correlated and if it is equal to 1 then there is no multicollinearity between them.

By analysing the above VIF factor as shown in snip above we can say that it is **moderately correlated**.

We have to split the dataset into 70:30 ie. 70% train data and 30% test data set and then we have fit the dataset

- Summary of the OLS Regression Model

OLS Regression Results

Dep. Variable:	usr	**R-squared:**	0.796
Model:	OLS	**Adj. R-squared:**	0.795
Method:	Least Squares	**F-statistic:**	1116.
Date:	Sun, 05 Nov 2023	**Prob (F-statistic):**	0.00
Time:	05:20:10	**Log-Likelihood:**	-16656.
No. Observations:	5734	**AIC:**	3.335e+04
Df Residuals:	5713	**BIC:**	3.349e+04
Df Model:	20		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	84.1314	0.316	266.122	0.000	83.512	84.751
lread	-0.0634	0.009	-7.064	0.000	-0.081	-0.046
lwrite	0.0480	0.013	3.660	0.000	0.022	0.074
scall	-0.0007	6.28e-05	-10.576	0.000	-0.001	-0.001
sread	0.0003	0.001	0.336	0.737	-0.002	0.002
swrite	-0.0055	0.001	-3.805	0.000	-0.008	-0.003
fork	0.0296	0.132	0.225	0.822	-0.229	0.288
exec	-0.3211	0.052	-6.219	0.000	-0.422	-0.220
rchar	-5.212e-06	4.87e-07	-10.696	0.000	-6.17e-06	-4.26e-06
wchar	-5.346e-06	1.03e-06	-5.179	0.000	-7.37e-06	-3.32e-06
pgout	-0.3669	0.090	-4.077	0.000	-0.543	-0.190
ppgout	-0.0786	0.079	-0.999	0.318	-0.233	0.076
pgfree	0.0853	0.048	1.786	0.074	-0.008	0.179
pgscan	6.241e-15	3.76e-17	165.982	0.000	6.17e-15	6.31e-15
atch	0.6304	0.143	4.414	0.000	0.350	0.910
pgin	0.0198	0.028	0.695	0.487	-0.036	0.076
ppgin	-0.0672	0.020	-3.406	0.001	-0.106	-0.029
pflt	-0.0336	0.002	-16.954	0.000	-0.037	-0.030
vflt	-0.0055	0.001	-3.831	0.000	-0.008	-0.003
runqsz	1.6137	0.126	12.807	0.000	1.367	1.861
freemem	-0.0005	5.07e-05	-9.022	0.000	-0.001	-0.000
freeswap	8.829e-06	1.9e-07	46.463	0.000	8.46e-06	9.2e-06

Omnibus:	1102.551	**Durbin-Watson:**	2.016
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2367.549
Skew:	-1.118	**Prob(JB):**	0.00
Kurtosis:	5.216	**Cond. No.**	7.02e+22

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 2.32e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

- Coefficient of determination

```
The variation in the independent variable which is explained by the dependent variable is 79.6157 %
```

- Get the RMSE on training data

```
The Root Mean Square Error (RMSE) of the model is for the training set is 4.4190166755430935
```

- Get the RMSE on test dataset

```
The Root Mean Square Error (RMSE) of the model is for testing set is 4.652920160995275
```

Using Linear Model from Sci-kit learn library

- We have fitted the dataset using sci-kit learn library
- Get the score on training set i.e R square

```
The coefficient of determination R^2 of the prediction on Train set 0.7961565330395103
```

t the score on test set

- Get the score on test set i.e R square

```
The coefficient of determination R^2 of the prediction on Test set 0.7676695029858367
```

- Get the RMSE on test set

```
The Root Mean Square Error (RMSE) of the model is for testing set is 4.652920160995946
```

- Get the RMSE on train set

```
The Root Mean Square Error (RMSE) of the model is for train set is 4.419016675543094
```

Inference::

1. We can see that from both the models that is OLS and sci-kit we see that the values remain the same as below::

   RMSE on train set is 0.7961565330395103
   RMSE on test set is 0.7676695029858367
   Rsquare on test set is 4.652920160995946
   Rsquare on train set is 4.419016675543094
   Coefficient of determination is 79.6157 %

2. We can say that both the model is best in terms of performance.
3. R-squared value or both test and train is 0.76 and 0.79 respectively, which indicates that more than 75% of observed variance can be explained by model's inputs

- We can write the Linear regression as:

  (84.1314) * const + (-0.0634) * lread + (0.0480) * lwrite + (-0.0007) * scall + (0.0003) * sread + (-0.0055)* swrite + (-0.0296) * fork + (-0.3211) * exec + (--5.212e-06) * rchar + (--5.346e-06) * wchar + (-0.3669) * pgout +(-0.0786) * ppgout + (0.0853) * pgfree + (6.241e-15) * pgscan + (0.6304) * atch + (0.0198) * pgin + (-0.058) *ppgin + (-0.0314) * pflt + (-0.0055) * vflt + (-0.0005) * freemem + (0.0) * freeswap + (1.6137) *runqsz_Not_CPU_Bound

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

usr = 0.0405676330815176 + 0.18940457506323255 * (lwrite) + 0.0016996629317421739 * (scall) + 0.01062705578998727 * (sread) + 0.03120726140328898 * (swrite) + -4.231812819055642 * (fork) + 1.0100421766330803 * (exec) + 3.4822544027262732e-06 * (rchar) + 2.1100167546548864e-05 * (wchar) + 1.187104251670321 * (pgout) + -1.8389318151987566 * (ppgout) + 0.914582080869214 * (pgfree) + -2.9836628921562222e-15 * (pgscan) + 4.06363294599411 * (atch) + 0.5787638692335303 * (pgin) + -0.2512282171035673 * (ppgin) + -0.06887191627623623 * (pflt) + 0.03828409162454253 * (vflt) + 13.436506742209744 * (runqsz) + -0.0015677884113311758 * (freemem) + 4.98035940246531e-05 * (freeswap)

- **1 unit increase in the lwrite lead to 0.2 times increase in the usr**
- **1 unit increase in the swrite lead to 0.03 times increase in the usr**
- **if forl decreases the usr by a factor of 4.23**
- **if pgout it decreases the usr by a factor of 1.83**

Problem 2: Logistic Regression, LDA and CART

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

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

Data Dictionary:

1. Wife's age (numerical)
2. Wife's education (categorical) 1=uneducated, 2, 3, 4=tertiary
3. Husband's education (categorical) 1=uneducated, 2, 3, 4=tertiary
4. Number of children ever born (numerical)
5. Wife's religion (binary) Non-Scientology, Scientology
6. Wife's now working? (binary) Yes, No
7. Husband's occupation (categorical) 1, 2, 3, 4(random)
8. Standard-of-living index (categorical) 1=verlow, 2, 3, 4=high
9. Media exposure (binary) Good, Not good
10. Contraceptive method used (class attribute) No,Yes

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

We have loaded the dataset given to us i.e. Contraceptive method dataset and then we have read the dataset after that

- Top 5 rows displayed by head

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

id_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_method_used
Secondary	3.0	Scientology	No	2	High	Exposed	No
Secondary	10.0	Scientology	No	3	Very High	Exposed	No
Secondary	7.0	Scientology	No	3	Very High	Exposed	No
Primary	9.0	Scientology	No	3	High	Exposed	No
Secondary	8.0	Scientology	No	3	Low	Exposed	No

- Bottom 5 rows displayed by tail

	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
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

- Shape of the dataset

```
(1473, 10)
```

- Information of dataset by info

```
<bound method DataFrame.info of       Wife_age Wife_ education Husband_education  No_of_children_born  \
0         24.0         Primary         Secondary                  3.0
1         45.0      Uneducated         Secondary                 10.0
2         43.0         Primary         Secondary                  7.0
3         42.0       Secondary           Primary                  9.0
4         36.0       Secondary         Secondary                  8.0
...        ...             ...               ...                  ...
1468      33.0        Tertiary          Tertiary                  NaN
1469      33.0        Tertiary          Tertiary                  NaN
1470      39.0       Secondary         Secondary                  NaN
1471      33.0       Secondary         Secondary                  NaN
1472      17.0       Secondary         Secondary                  1.0

     Wife_religion Wife_Working  Husband_Occupation Standard_of_living_index  \
0      Scientology           No                   2                     High
1      Scientology           No                   3                Very High
2      Scientology           No                   3                Very High
3      Scientology           No                   3                     High
4      Scientology           No                   3                      Low
...            ...          ...                 ...                      ...
1468   Scientology          Yes                   2                Very High
1469   Scientology           No                   1                Very High
1470   Scientology          Yes                   1                Very High
1471   Scientology          Yes                   2                      Low
1472   Scientology           No                   2                Very High

     Media_exposure  Contraceptive_method_used
0           Exposed                         No
1           Exposed                         No
2           Exposed                         No
3           Exposed                         No
4           Exposed                         No
...             ...                        ...
1468        Exposed                        Yes
1469        Exposed                        Yes
1470        Exposed                        Yes
1471        Exposed                        Yes
1472        Exposed                        Yes

[1473 rows x 10 columns]>
```

- Describe the dataset shared

	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

- Check the number of null values present in dataset

```
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
```

We need to remove the null values by imputing it with mean values. After imputing the dataset has no null values

```
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
```

- Check the presence of Duplicates rows.

```
80
```

We found that there is 80 duplicates present in dataset.We need to remove the same to move further

- We have dropped the duplicates values . After that the shape of the dataset becomes

```
(1473, 10)
```

- No duplicity is there now

```
0       False
1       False
2       False
3       False
4       False
        ...
1468    False
1469    False
1470    False
1471    False
1472    False
Length: 1473, dtype: bool
```

- We have divided the dataset into numerical and categorical

The first row displayed the numerical

The second row displayed the categorical variables

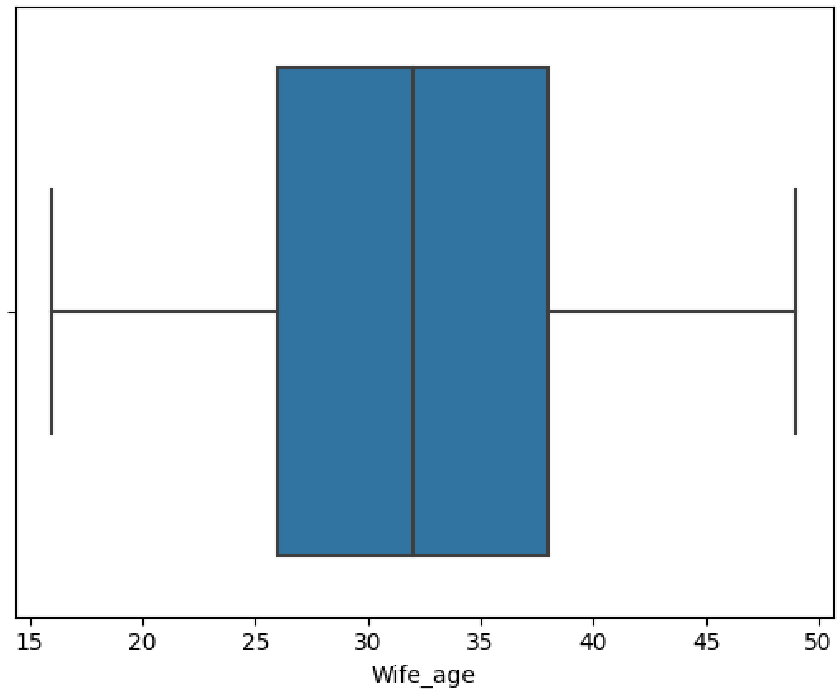
```
['Wife_ education', 'Husband_education', 'Wife_religion', 'Wife_Working', 'Standard_of_living_index', 'Media_exposure ', 'Contraceptive_method_used']
['Wife_age', 'No_of_children_born', 'Husband_Occupation']
```

- Information of dataset by info

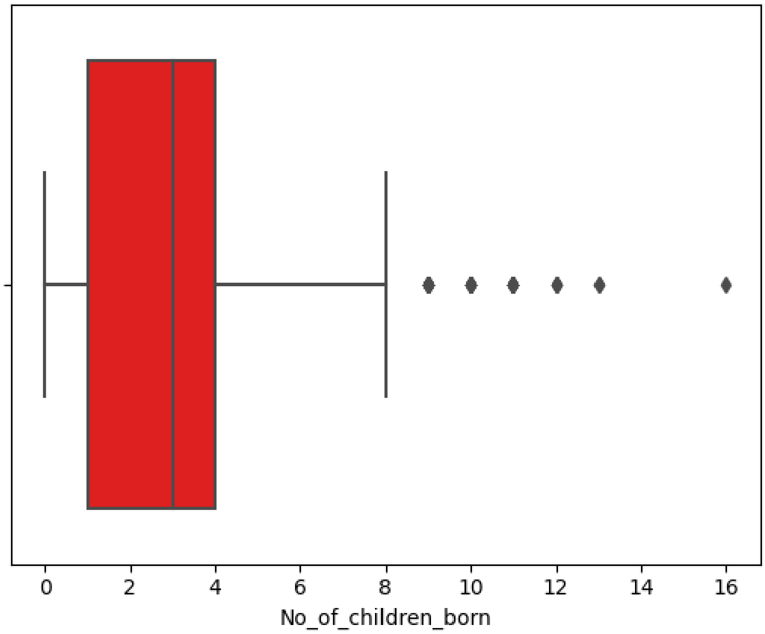
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1473 entries, 0 to 1472
Data columns (total 10 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   Wife_age                   1473 non-null   float64
 1   Wife_ education            1473 non-null   object 
 2   Husband_education          1473 non-null   object 
 3   No_of_children_born        1473 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
```

- Checking the outliers in the numerical variables if present and if present we need to remove the same.

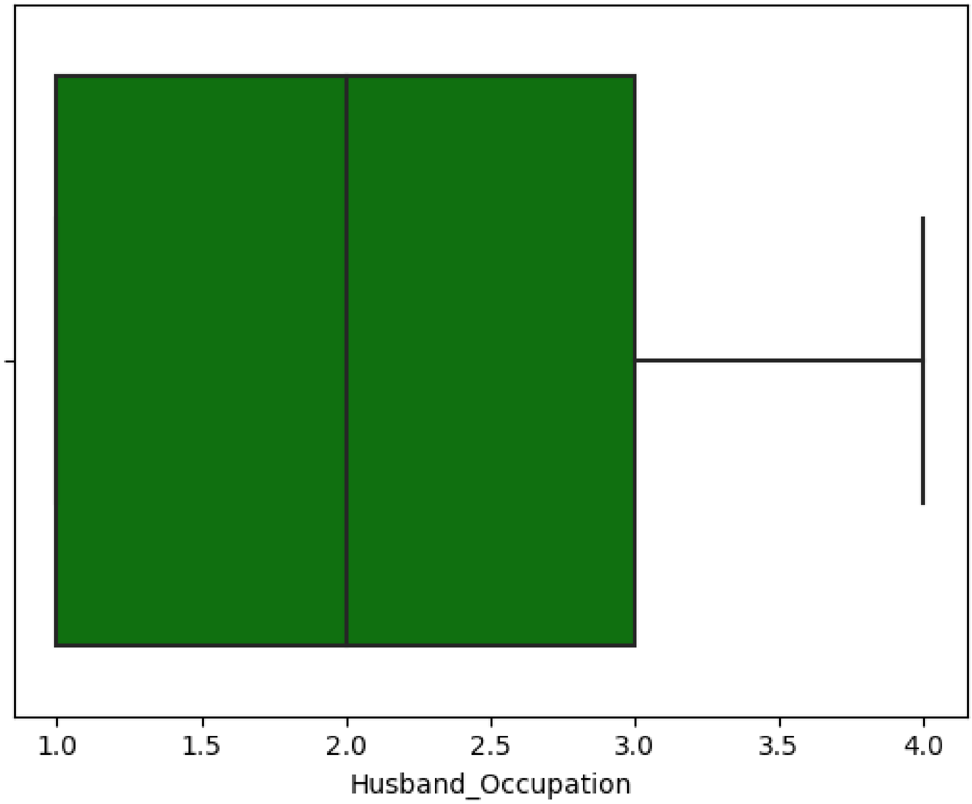
- WIFE AGE



- NUMBER OF CHILDREN

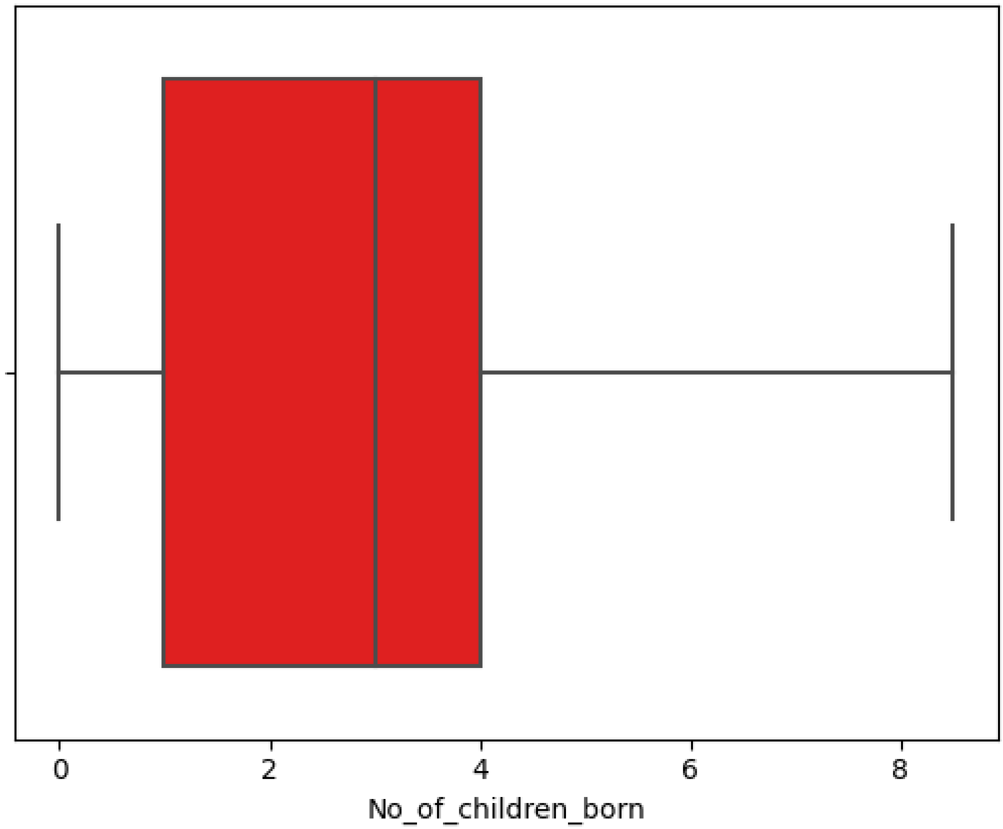


- HUSBAND OCCUPATION



We found that there is outliers present in number of children column and we need to treat the same to move further with the dataset.

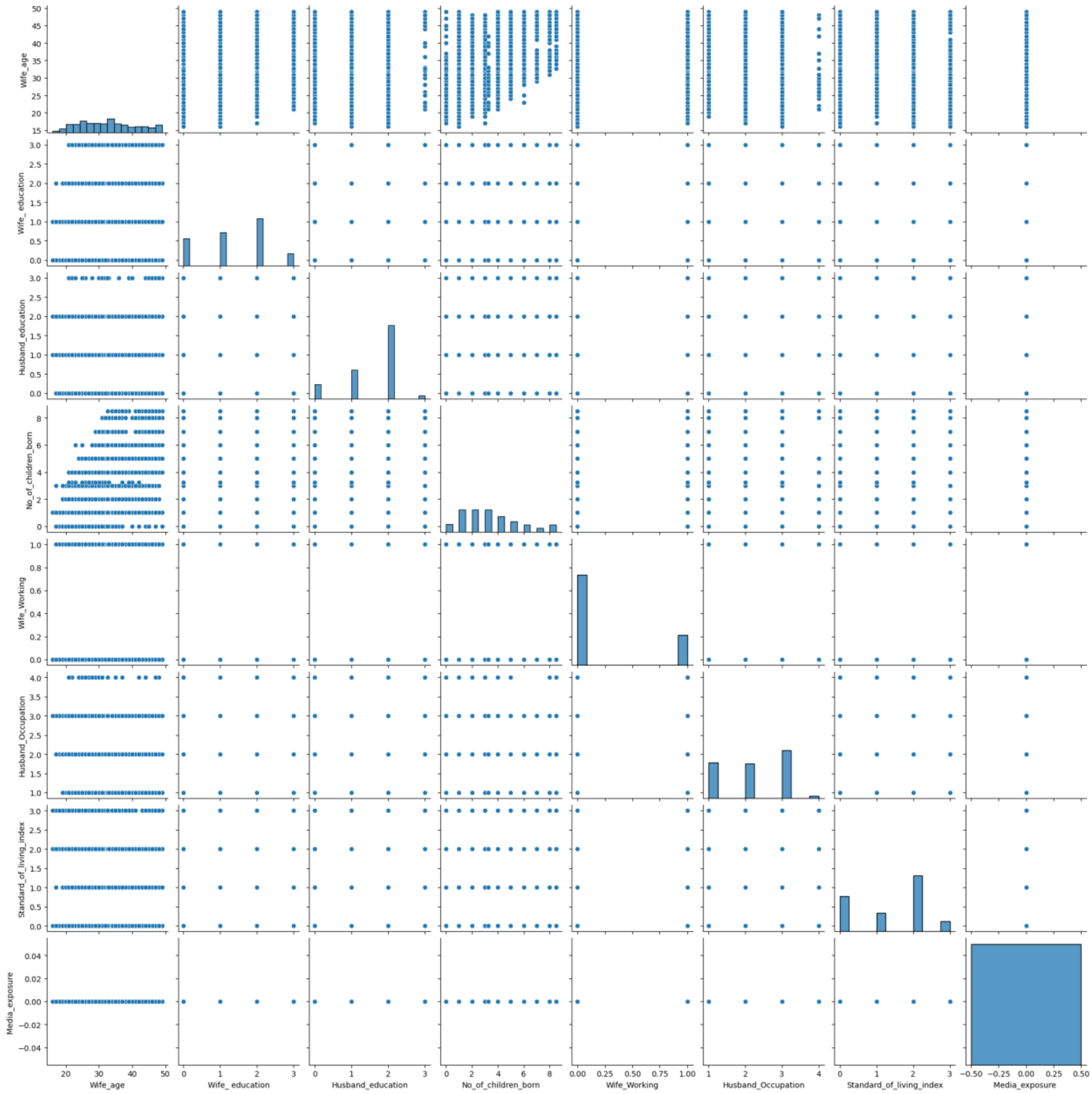
Treating the outliers with IQR method and after treating we need to plot the boxplot again for the number of children column to check whether there is any outliers present after treating the same.

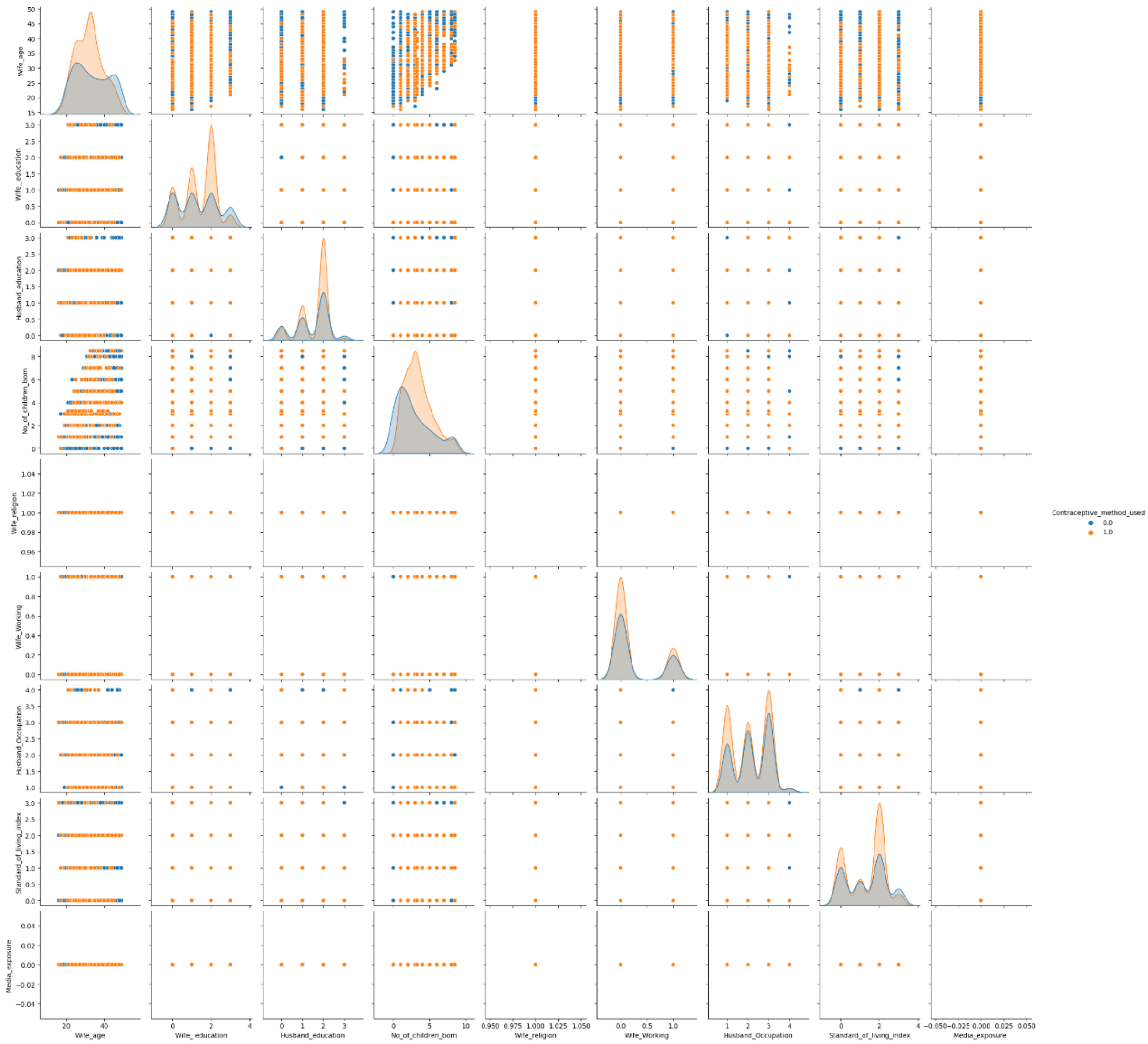


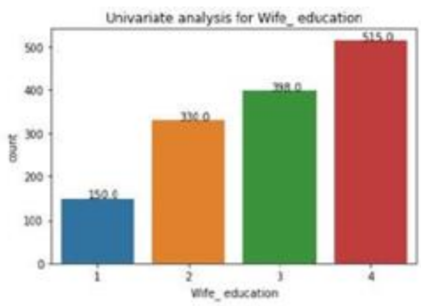
- We need to do the univariate, bi-variate and multi variate analysis for the given dataset.

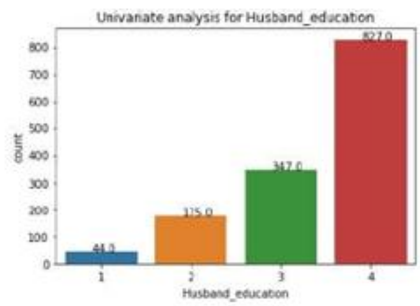
- BIVARIATE ANALYSIS

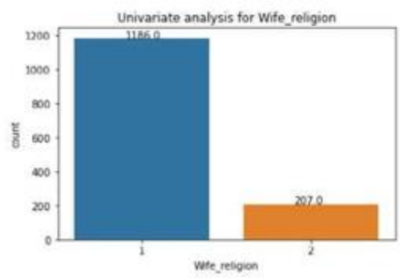
- **PAIRPLOT**

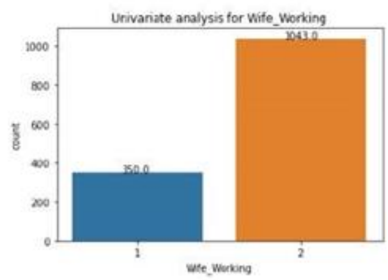


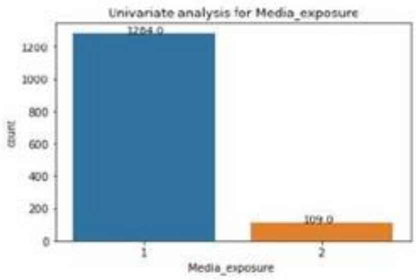

Wife_age
Wife_education
Husband_education
No_of_children_born
Wife_religion
Wife_Working
Husband_Occupation
Standard_of_living_index
Media_exposure
Contraceptive_method_used
0.0
1.0

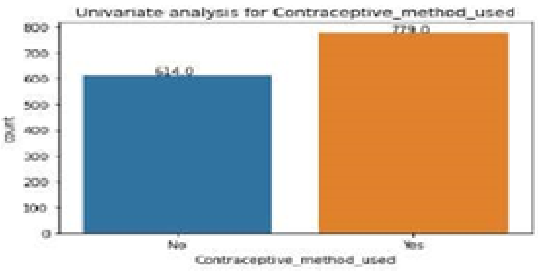

Univariate analysis for Wife_ education
150.0
330.0
398.0
515.0
count
Wife_ education

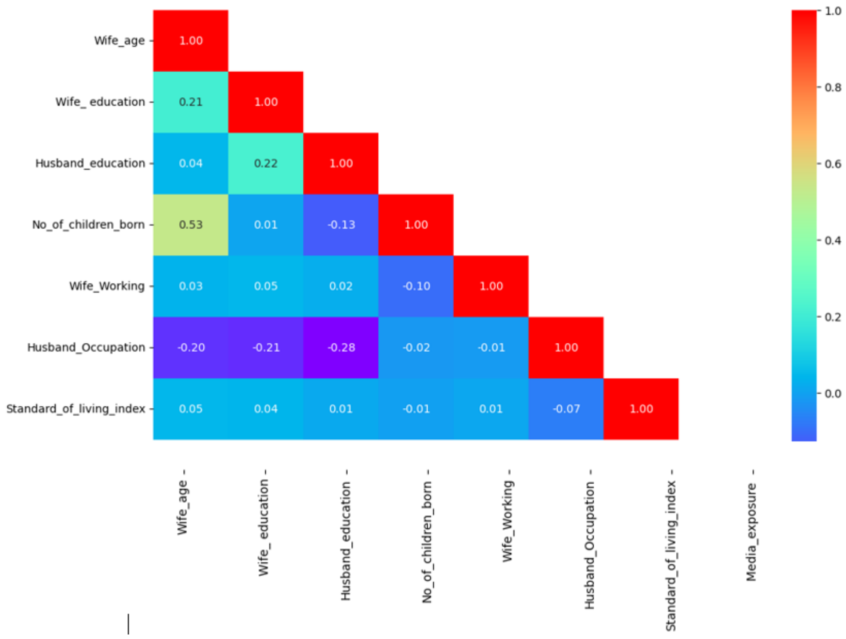

Univariate analysis for Husband_education
44.0
175.0
347.0
827.0
count
Husband_education


Univariate analysis for Wife_religion
1186.0
207.0
count
Wife_religion

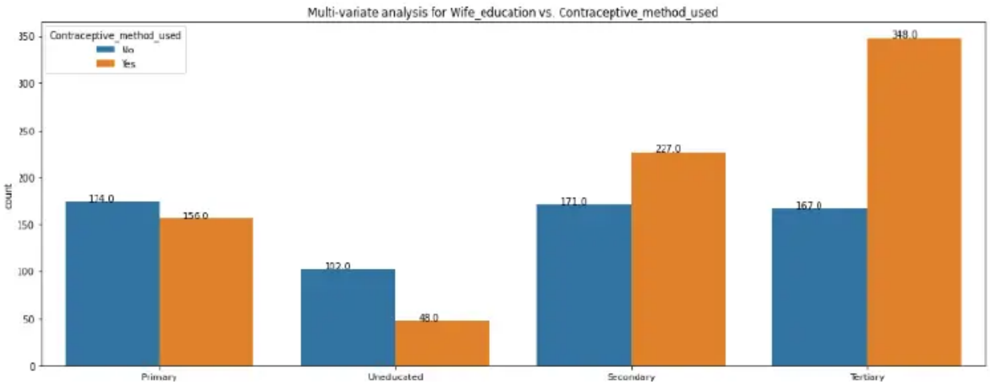

Univariate analysis for Wife_Working
350.0
1043.0
count
Wife_Working

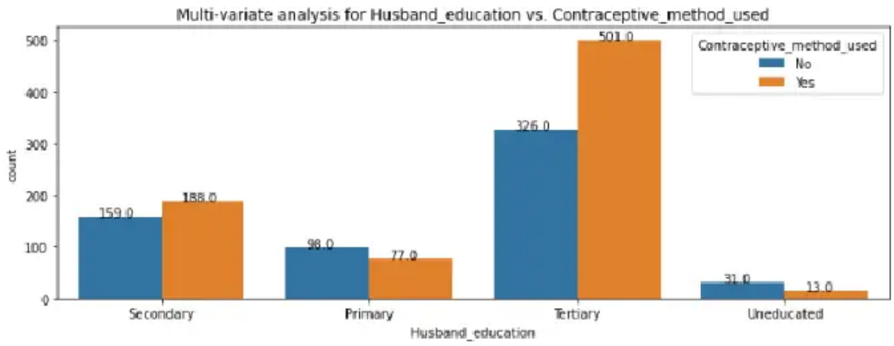


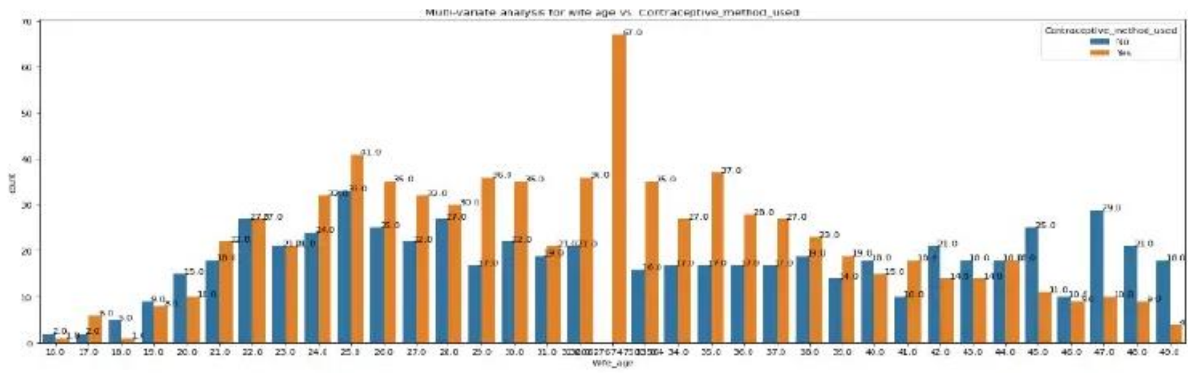


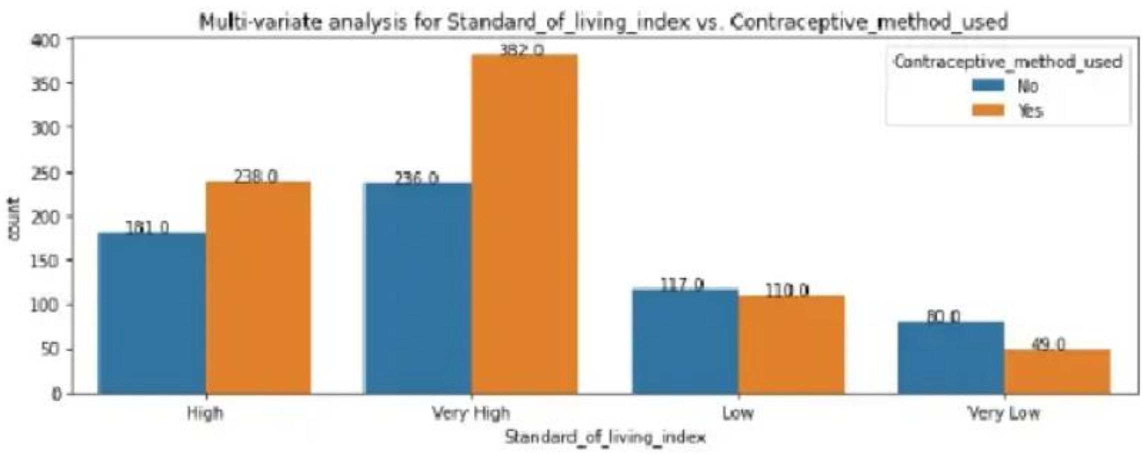


- **MULTIVARIATE ANALYSIS**


Multi-variate analysis for Wife_education vs. Contraceptive_method_used
Contraceptive_method_used
No
Yes
count
174.0
156.0
102.0
48.0
171.0
227.0
167.0
348.0
Primary
Uneducated
Secondary
Tertiary


Multi-variate analysis for Husband_education vs. Contraceptive_method_used
Contraceptive_method_used
No
Yes
count
159.0
188.0
98.0
77.0
326.0
501.0
31.0
13.0
Secondary
Primary
Tertiary
Uneducated
Husband_education



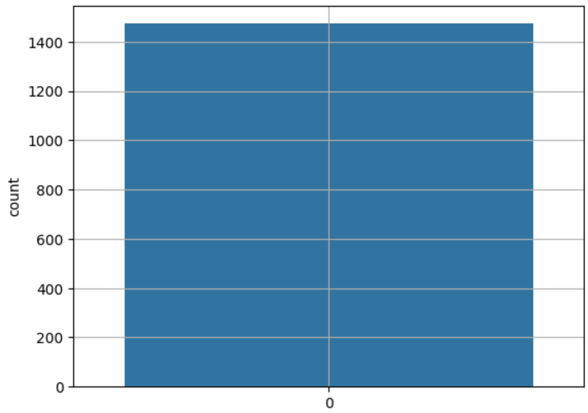

Multi-variate analysis for Standard_of_living_index vs. Contraceptive_method_used
Contraceptive_method_used
No
Yes
count
181.0
238.0
236.0
382.0
117.0
110.0
80.0
49.0
High
Very High
Low
Very Low
Standard_of_living_index

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 and LDA (linear discriminant analysis) and CART.

We have encoded the dataset given to us as below

	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	0	1	3.0	1	0	2	0	0	0
1	45.0	3	1	10.0	1	0	3	2	0	0
2	43.0	0	1	7.0	1	0	3	2	0	0
3	42.0	1	0	9.0	1	0	3	0	0	0
4	36.0	1	1	8.0	1	0	3	1	0	0

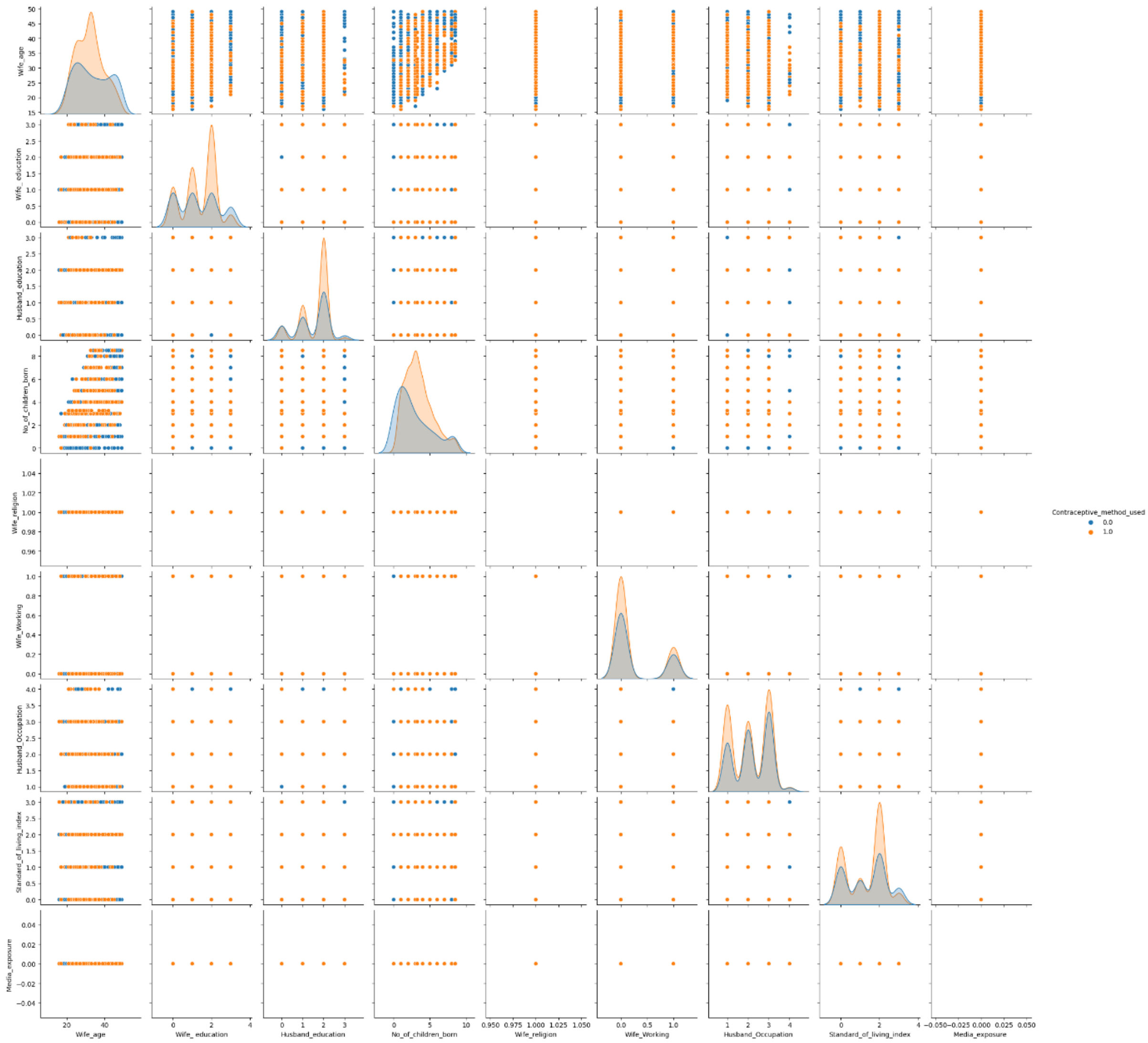
- Count plot for Contraceptive_method_used

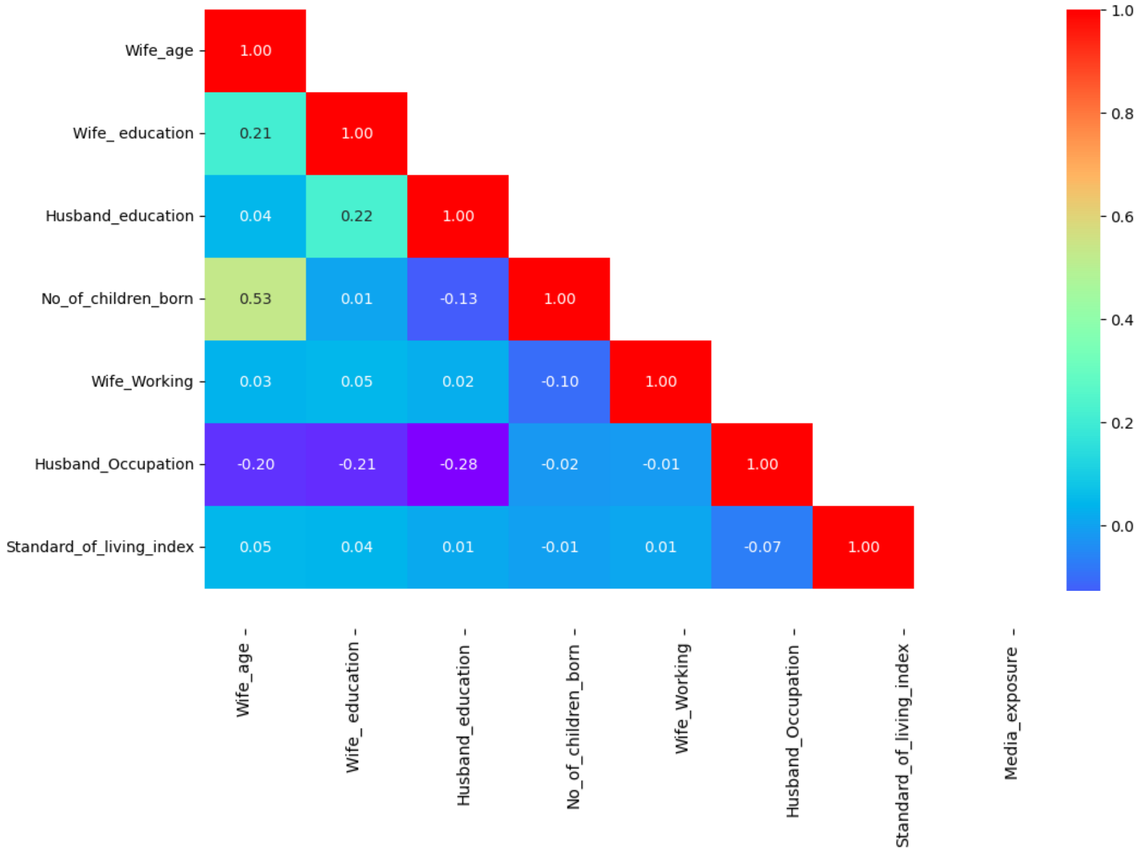


- We have group by Contraceptive_method_used and found the below count

Contraceptive_method_used	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure
0.0	629	629	629	629	629	629	629	629	629
1.0	844	844	844	844	844	844	844	844	844

Pairplot for Contraceptive_method_used





- After encoding we need to split the dataset and train and test i.e. 70% and 30% respectively.

APPLYING LOGISTIC REGRESSION

- We have calculated the average score using logistic regression.

```
Accuracy Score is  0.6832579185520362
```

- Classification Report and Confusion matrix of test set.

```
Classification Report
              precision    recall  f1-score   support

         0.0       0.72      0.43      0.54       189
         1.0       0.67      0.87      0.76       253

    accuracy                           0.68       442
   macro avg       0.69      0.65      0.65       442
weighted avg       0.69      0.68      0.66       442
```

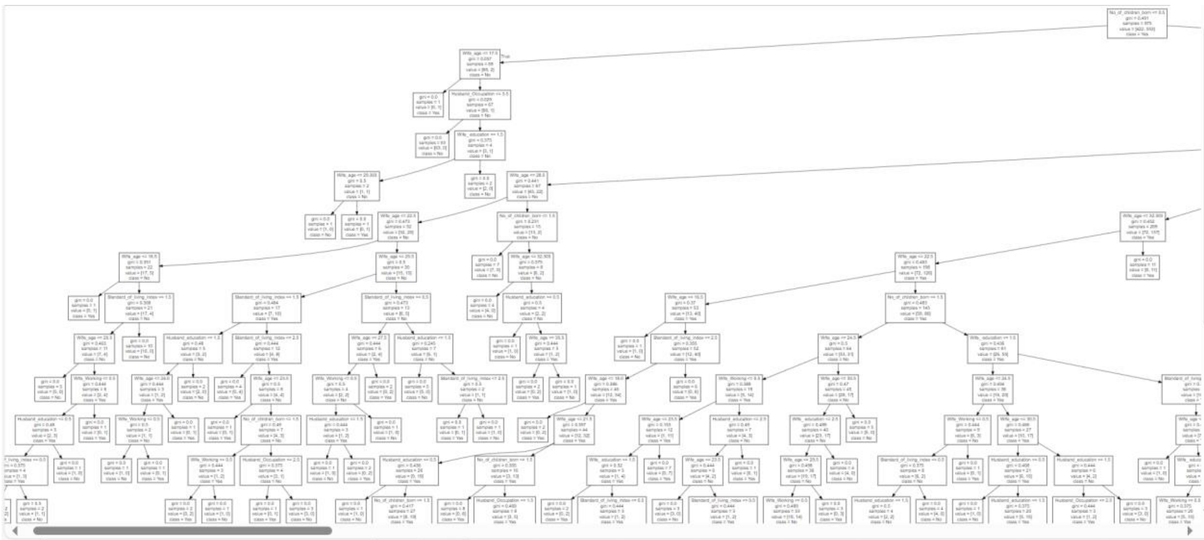
```
Confusion Matrix
 [[ 81 108]
 [ 32 221]]
```

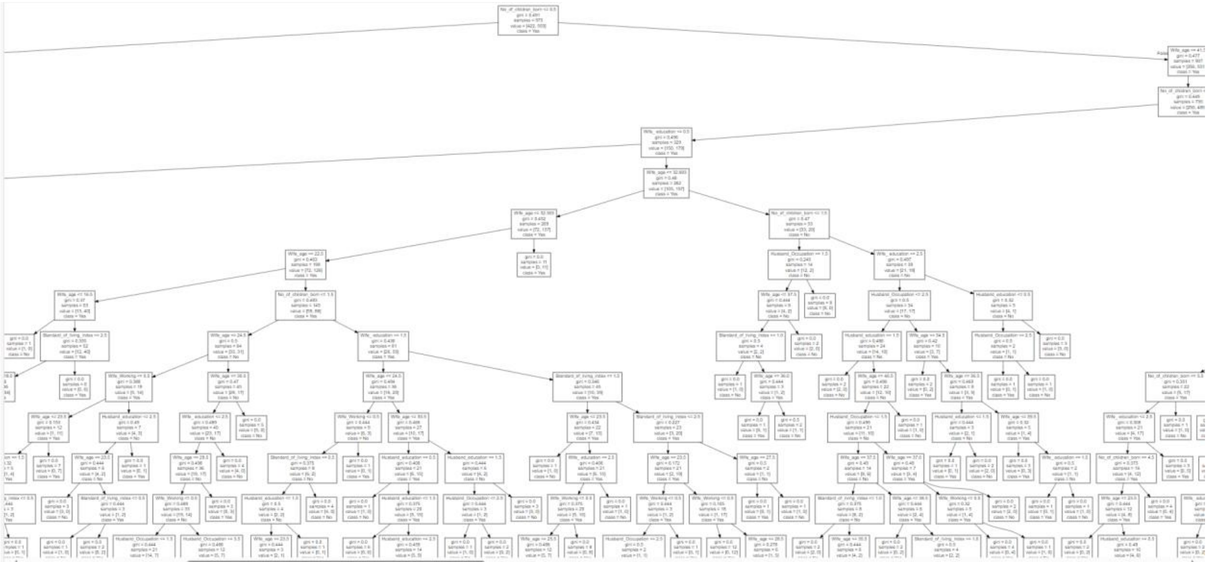
- Classification Report and Confusion matrix of train set.

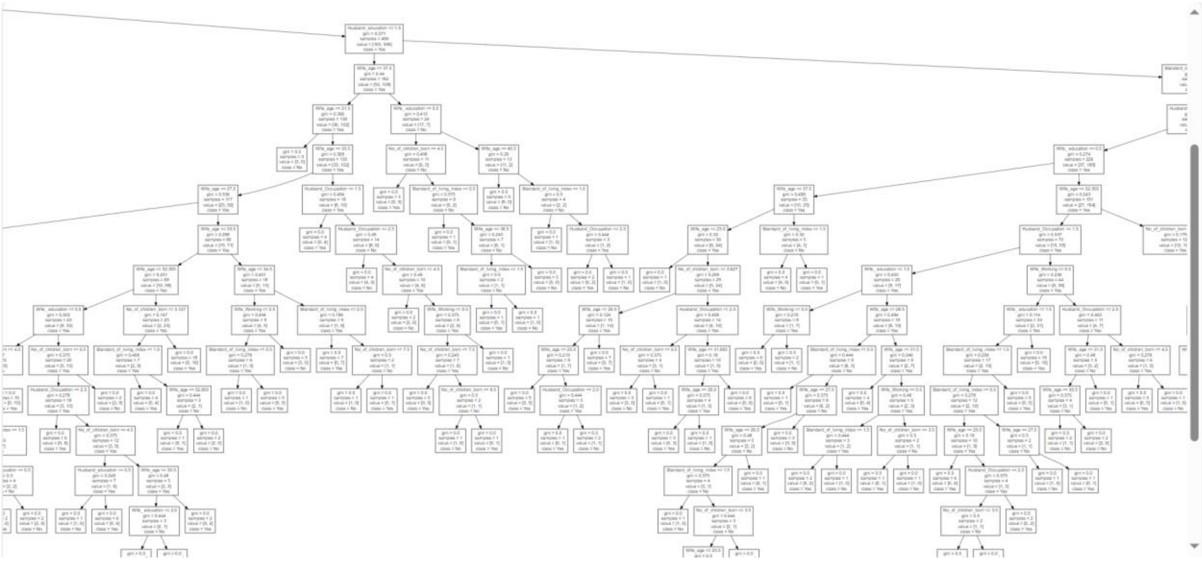
```
Accuracy Score is  0.6508244422890398


Classification Report
              precision    recall  f1-score   support

         0.0       0.72      0.43      0.54       189
         1.0       0.67      0.87      0.76       253

    accuracy                           0.68       442
   macro avg       0.69      0.65      0.65       442
weighted avg       0.69      0.68      0.66       442
```

CART

- We have made decision tree using GRAPHVIZ:







- **Importance of features**

	Imp
Wife_age	0.347470
Wife_ education	0.109216
Husband_education	0.076432
No_of_children_born	0.246126
Wife_religion	0.000000
Wife_Working	0.069886
Husband_Occupation	0.078468
Standard_of_living_index	0.072401
Media_exposure	0.000000

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

APPLYING LOGISTIC REGRESSION

- We have calculated the average score using logistic regression.

```
Accuracy Score is  0.6832579185520362
```

- Classification Report and Confusion matrix of test set.

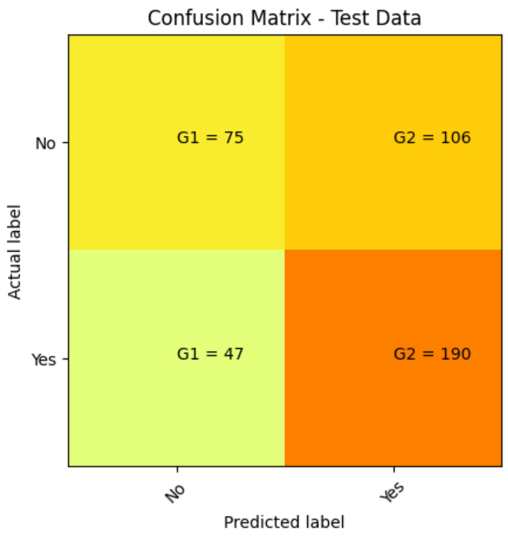
```
Classification Report
              precision    recall  f1-score   support

         0.0       0.72      0.43      0.54       189
         1.0       0.67      0.87      0.76       253

    accuracy                           0.68       442
   macro avg       0.69      0.65      0.65       442
weighted avg       0.69      0.68      0.66       442
```

```
Confusion Matrix
 [[ 81 108]
 [ 32 221]]
```

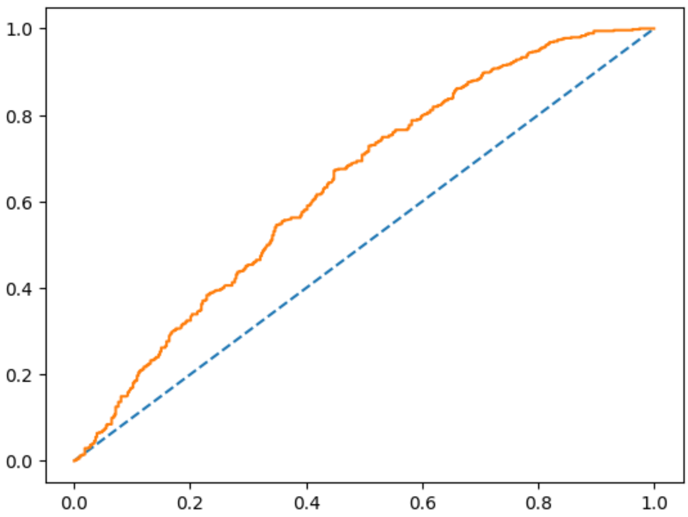
- Classification Report and Confusion matrix of train set.

```
Accuracy Score is  0.6508244422890398
```

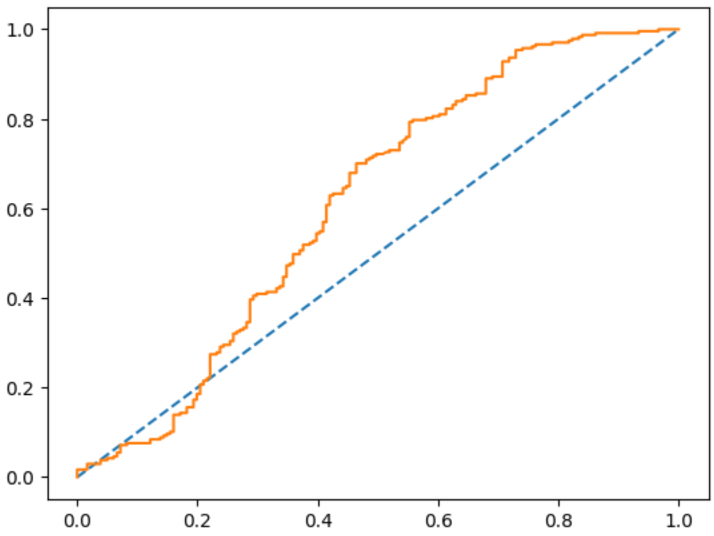
```
Classification Report
              precision    recall  f1-score   support

         0.0       0.72      0.43      0.54       189
         1.0       0.67      0.87      0.76       253

    accuracy                           0.68       442
   macro avg       0.69      0.65      0.65       442
weighted avg       0.69      0.68      0.66       442
```



- ROC CURVE AND AUC of train set



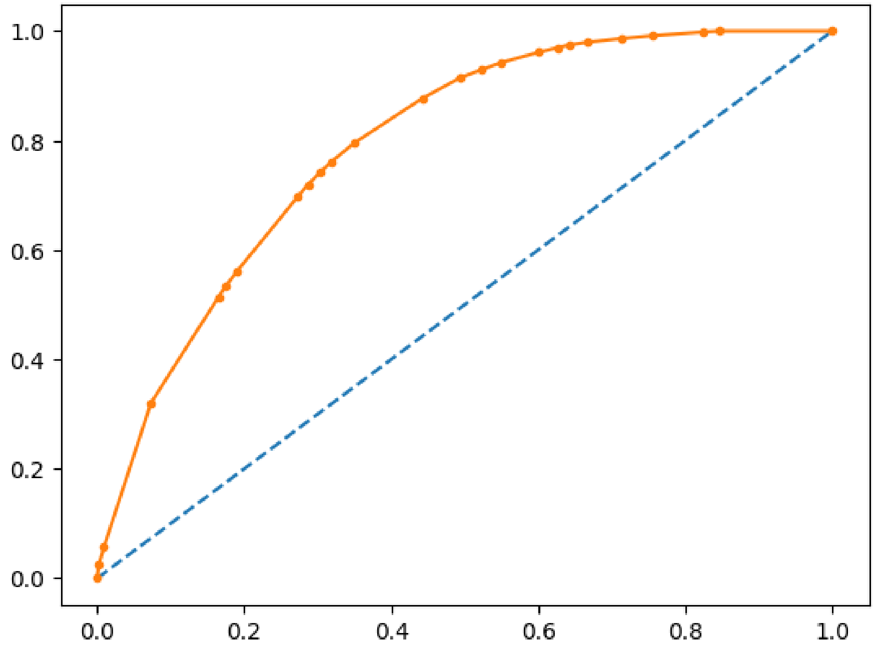
- ROC CURVE AND AUC CURVE OF TEST SET.



CART

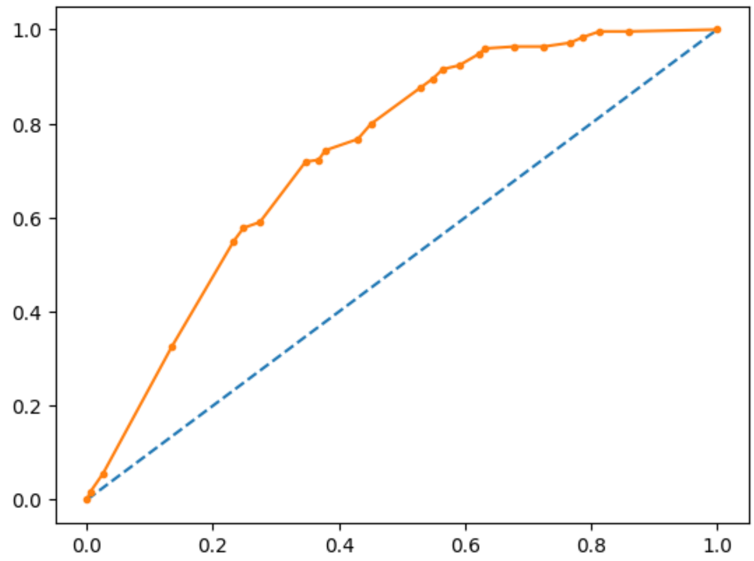
- AUC and ROC for the training data

AUC: 0.794



- AUC and ROC for the test data

AUC: 0.736



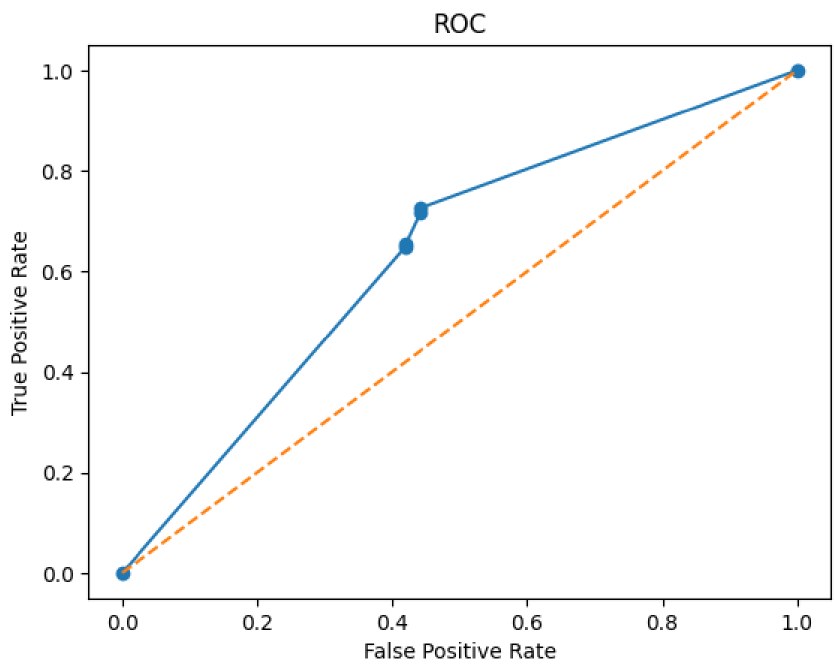
- Classification report of train set

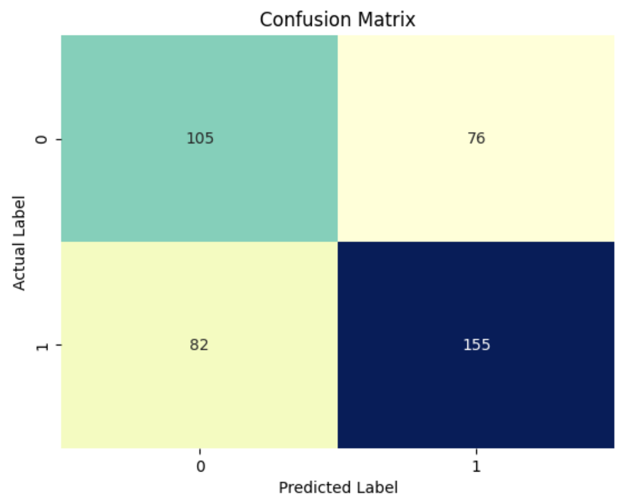
	precision	recall	f1-score	support
0.0	0.77	0.56	0.65	436
1.0	0.73	0.88	0.80	595
accuracy			0.74	1031
macro avg	0.75	0.72	0.72	1031
weighted avg	0.75	0.74	0.73	1031

- Classification report of test set

	precision	recall	f1-score	support
0.0	0.75	0.47	0.58	193
1.0	0.68	0.88	0.77	249
accuracy			0.70	442
macro avg	0.71	0.67	0.67	442
weighted avg	0.71	0.70	0.68	442

- ROC CURVE





- Confusion matrix of train set

```
array([[243, 193],
       [ 73, 522]])
```

- Confusion matrix of train set

```
array([[ 91, 102],
       [ 31, 218]])
```

- Regulalized model score of train data

```
0.7419980601357905
```

- Regularized model score of test data

```
0.6990950226244343
```

- Comparing both the model we found that the results are same. CART and LOGISTIC REGRESSION is training set accuracy is 0.74 which is good and for test set accuracy is 0.70
- AUC for the Training Data: 0.815
- AUC for the Test Data: 0.725

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.

- As per analysis shows that women with a tertiary education and very high standard of living used contraceptive methods Women ranging from 21 to 38 generally use contraceptive methods more
- We also found that usage of contraceptive methods need not depend on their demographic or socioeconomic backgrounds since the use of contraceptive methods were almost the same for both working and non-working women
- We also found that contraceptive method was high for both Scientology and Non-scientology women