Predictive Modeling

Name: Sanjana K Venkatesh
Date: 07/06/2024

Table of Contents

Executive Summary 5

Data Description 5

Exploratory Data Analysis 6

Let us check the types of variables in the data frame. 6

Check for missing values in the dataset: 7

Check for duplicate values in the dataset: 7

Treat the bad data i.e missing values in the dataset: 7

Overview of Linear Regression Model 25

1. Simple Linear Regression 25
2. Multiple Linear regression 25

Steps in Linear Regression 25

Assumptions of Linear Regression 25

Model building: 25

Linearity: 28

Test for Normality: 28

HOMOSCEDASTICITY 30

Predictions 30

The final result summary 31

The linear equation 31

Actionable Business Insights and recommendation: 31

Executive Summary 32

Data Description 32

Exploratory Data Analysis 33

Let us check the types of variables in the data frame. 33

Check for missing values in the dataset: 33

Check for duplicate values in the dataset: 33

Treat the bad data i.e missing values in the dataset: 34

Overview of Logistic Regression 43

Steps for Logistic Regression 43

Steps Model building: 43

Actionable Insights and recommendation: 46

Overview of LINEAR DISCRIMINANT ANALYSIS 47

Steps to Build LDA Model: 47

Model Building steps:.....47
Generate the coefficients and LDA function.....49
Observation and insights.....50
Overview of Decision Tree.....51
Model building.....51
Model Evaluation.....52
Actionable Insights and recommendations:.....54
Compare the Model.....55

List of figures

Univariant analysis.....8
Correlation plot:.....14
Outlier treatment.....14
Univariant analysis for Categorical variable.....22
Bivariant analysis.....22
Categorical vs numerical.....24
Univariant Analysis.....35
Correlation plot: Relation between numeric variables.....35
Outlier treatment:.....36
Univariant analysis for Categorical variable.....38
Bivariant analysis.....41
Categorical vs numerical.....42
Classification report for train data.....45
Classification report for test data.....46
Confusion Matrix for train data.....47
Classification report for train data.....47
AUC for the Training Data.....48
Confusion Matrix for test data.....48
Classification report for test data.....48
AUC-ROC for test data.....49
AUC and ROC Curve on training data.....52
AUC and ROC Curve on test data.....52
Confusion Matrix for the training data.....53
Classification report on training data.....53
Confusion matrix on test data.....54

The Classification report on test data 54

List of Tables

Sample of the dataset: 6

Sample of the dataset: 32

Check for summary statistics 7

Check the summary statistics 33

Executive Summary

The comp-activ database comprises activity measures of computer systems. Data was gathered from a Sun Sparcstation 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 analyze various system attributes to understand their influence on the system's 'usr' mode.

Data Description

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

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

3. scall - Number of system calls of all types per second

4. sread - Number of system read calls per second .

5. swrite - Number of system write calls per second .

6. fork - Number of system fork calls per second.

7. exec - Number of system exec calls per second.

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

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

10. pgout - Number of page out requests per second

11. ppgout - Number of pages, paged out per second

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

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

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

15. pgin - Number of page-in requests per second

16. ppgin - Number of pages paged in per second

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

18. vflt - Number of page faults caused by address translation .

19. 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.)

20. freemem - Number of memory pages available to user processes

Sample of the dataset:

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freesw
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	4670	1730
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	1869
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	1021
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	1863
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	1760

Dataset is in the shape of 8192, 22.

Exploratory Data Analysis

Let us check the types of variables in the data frame.

```
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
```

Total 22 columns with 5 rows. Out of 22 columns, 13 are float, 8 are integer and 1 is object data type.

Check for summary statistics

	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

summary of the statistics ranges from 0 to max 2526649.00 in different columns.

Check for missing values in the dataset:

From the above results we can see that there missing data for rchar and wchar we need to trat the bad data.

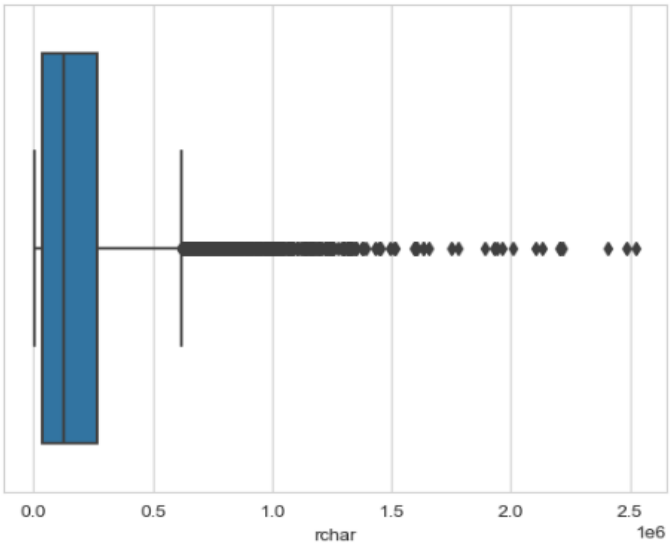
Check for duplicate values in the dataset:

There are no duplicates in the dataset.

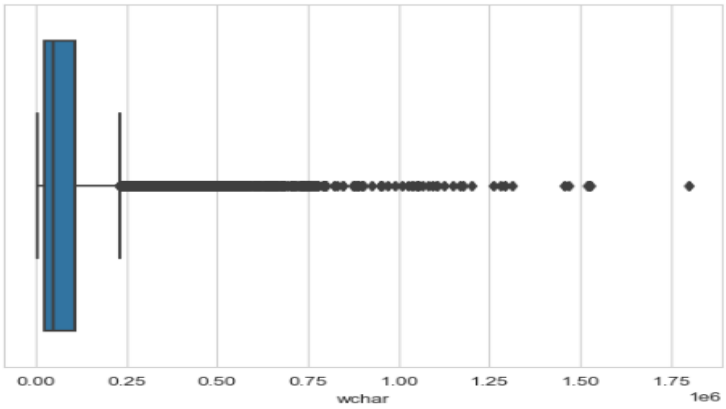
Treat the bad data i.e missing values in the dataset:

Plot the boxplot to check if the dataset is uniformly distributed. If it is normally distributed then we could go for mean imputation for missing values. In our dataset looks like there are outliers hence it can't be normally distributed. Hence will impute median for both rchar and wchar.

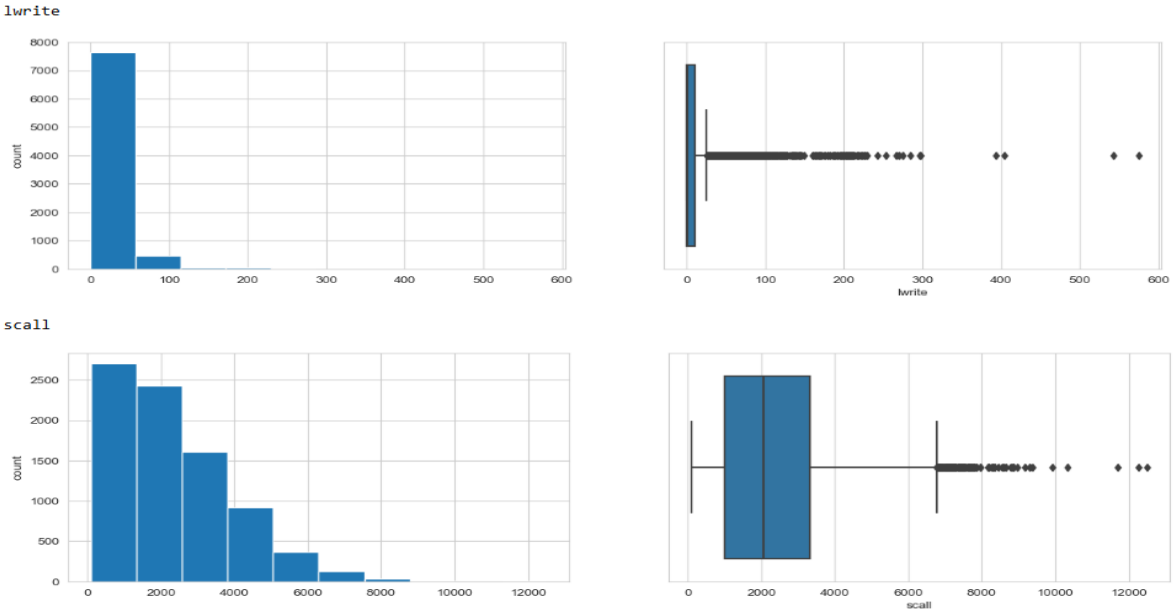
rchar:



wchar:



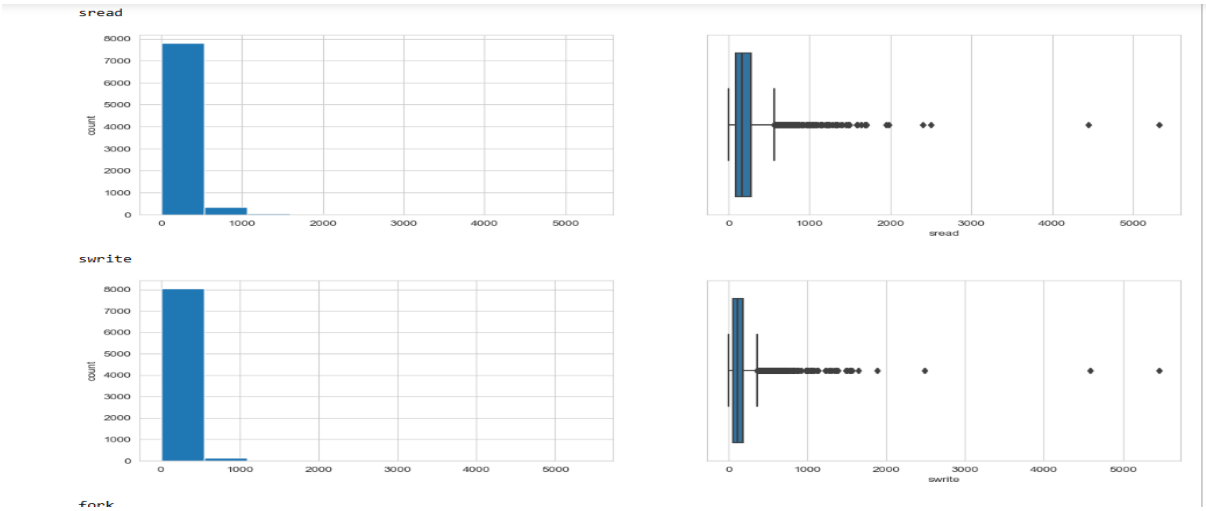
Univariant analysis



Observations:

lwrite has too many outliers. The max value is 6K+ for writes between system memory and user memory.

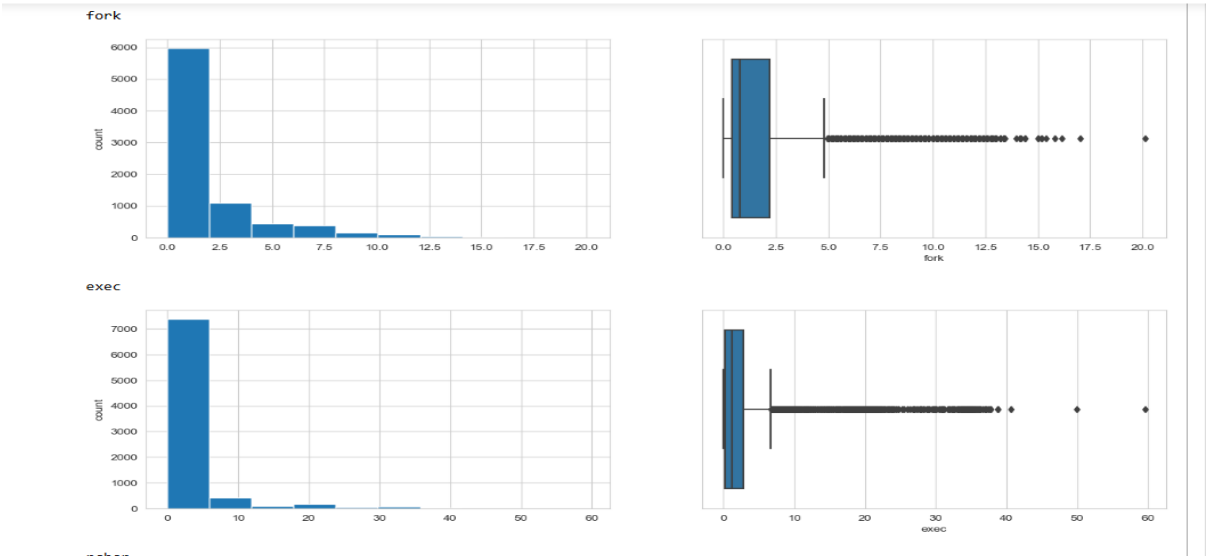
Scall has median around 2000 but it also has outliers. Which is 12k+.



Observations:

Number of system read calls per second has very minimal mean and median is around 1. Very few outboxed outliers. But it has crossed 6k+.

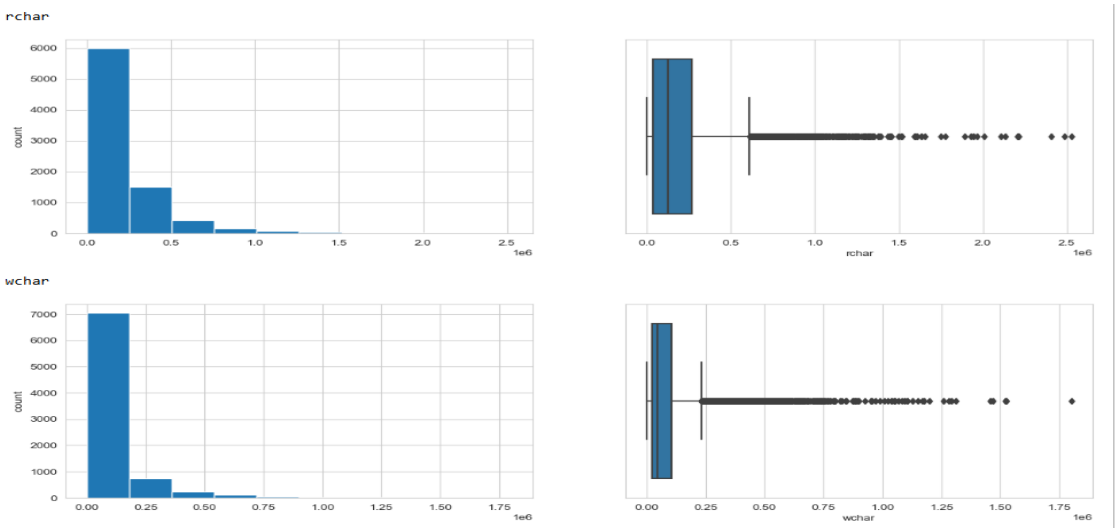
Number of system write calls per second(Swrite) is showing same behaviour has sread.



Observation:

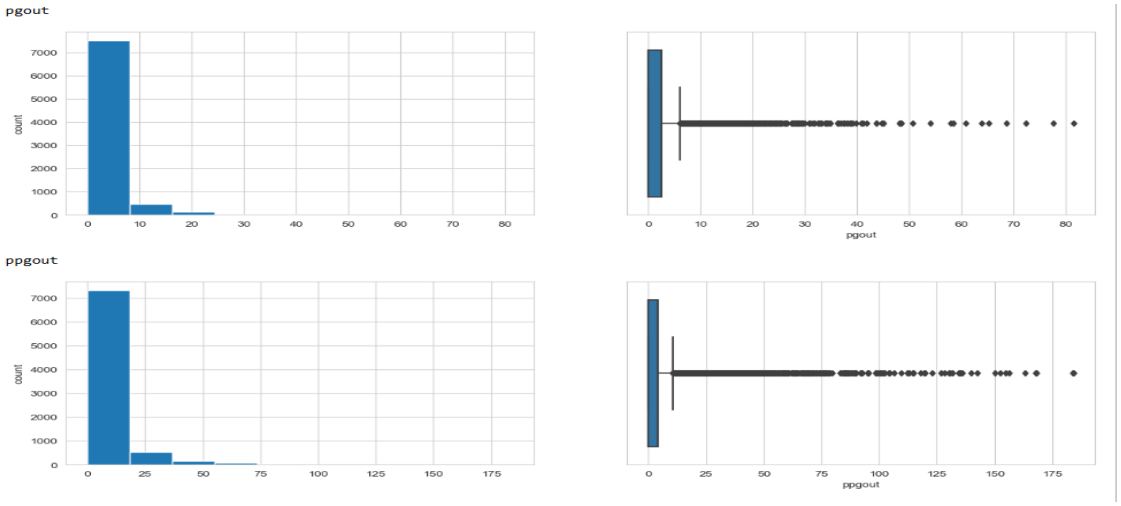
The fork is having wide range of observations from 0 to 12.5.Huge outliers found.

The exec is having less observations only few calls have reached 7k+.



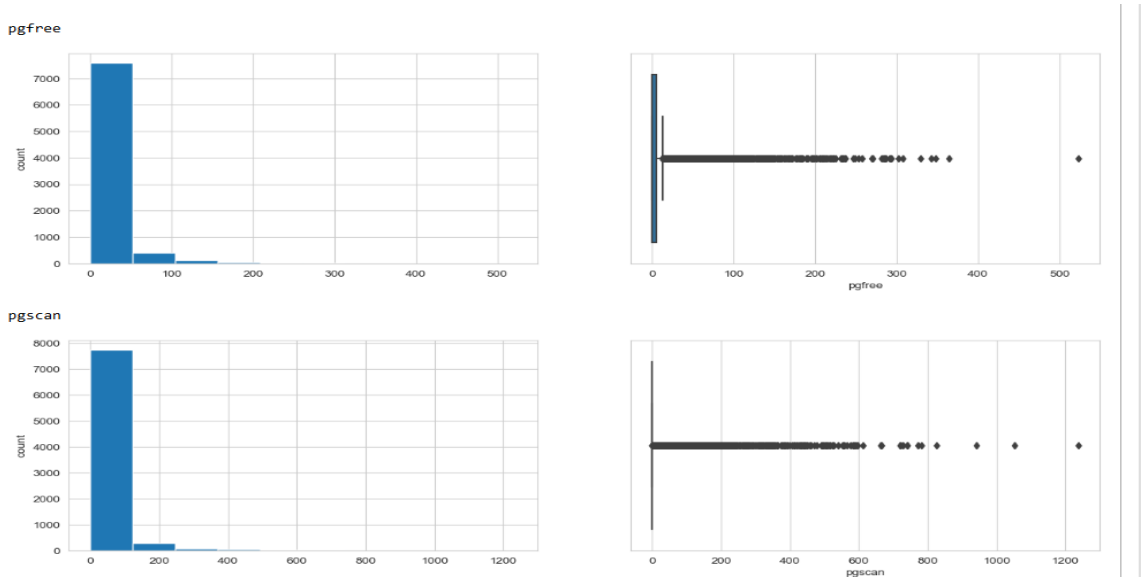
Observations:

Rchar and wchar both are almost having similar behaviour. Except that wchar has less outliers compared to rchar. The characters transferred is around 1.



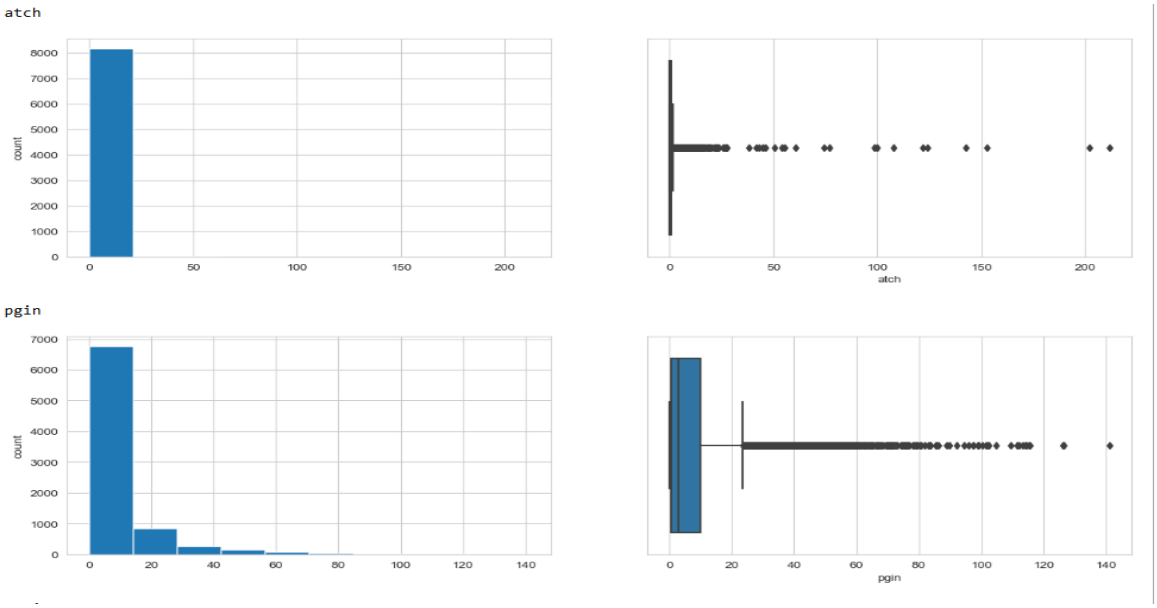
Observations:

Pgout and ppgout both are right skewed. Data is not symmetric and there is 0.5 median.



Observations:

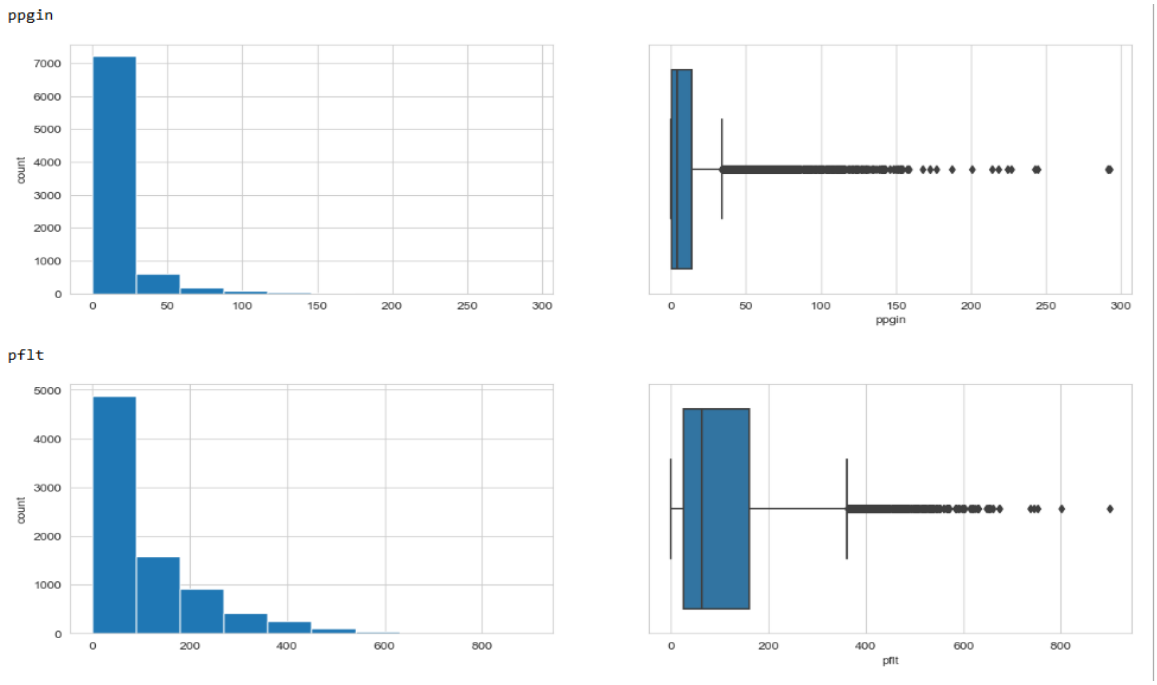
Pgfree and pgscan right skewed. Both pgscan and pgfree has 0 median with least mean like 1.



Observations:

atch	8192.0	1.127505e+00	5.708347	0.0	0.0	0.0	0.600
pgin	8192.0	8.277960e+00	13.874978	0.0	0.6	2.8	9.765

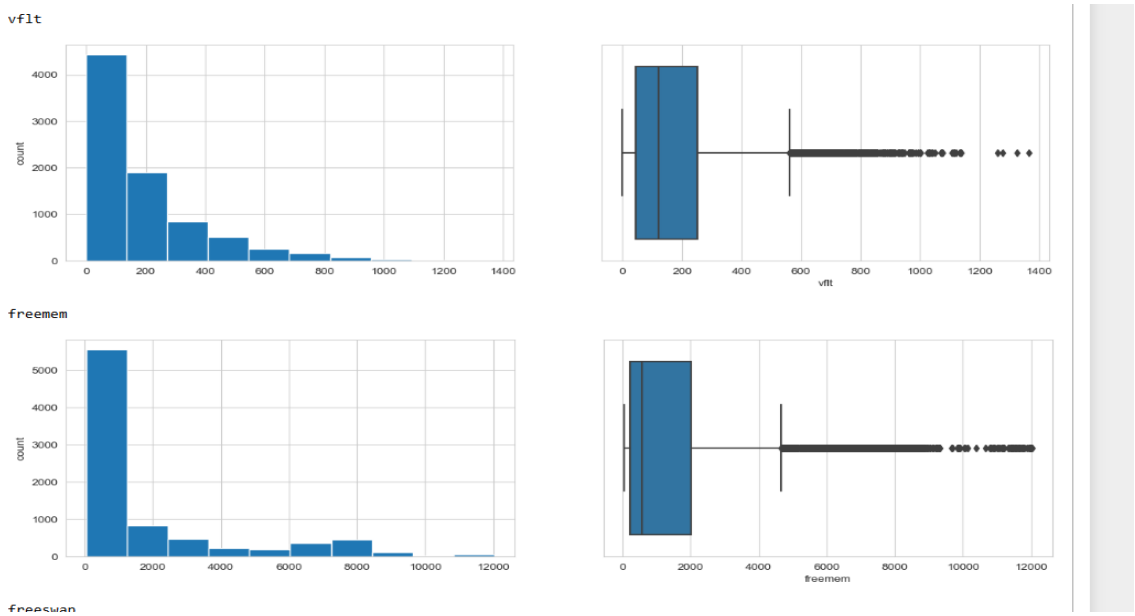
The mean value is 8 for pgin. The observations are not normally distributed for both.The atch has 0 median.



Observations:

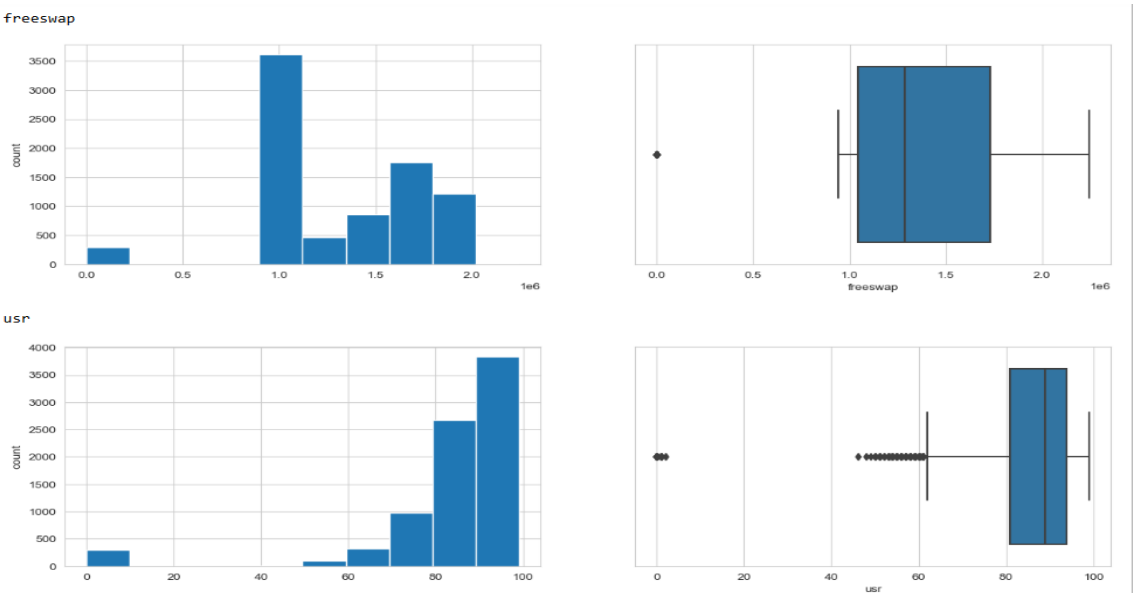
Ppgin has min value of 0 with max value has 292.

Pflt has median of 62.8 and the data is right skewed.



Observations:

Vflt and freemem both are not normally distributed. The value for vflt is spread from min,max(0.2,1365). The freemem i.e the number of memory pages available to user process is very less.

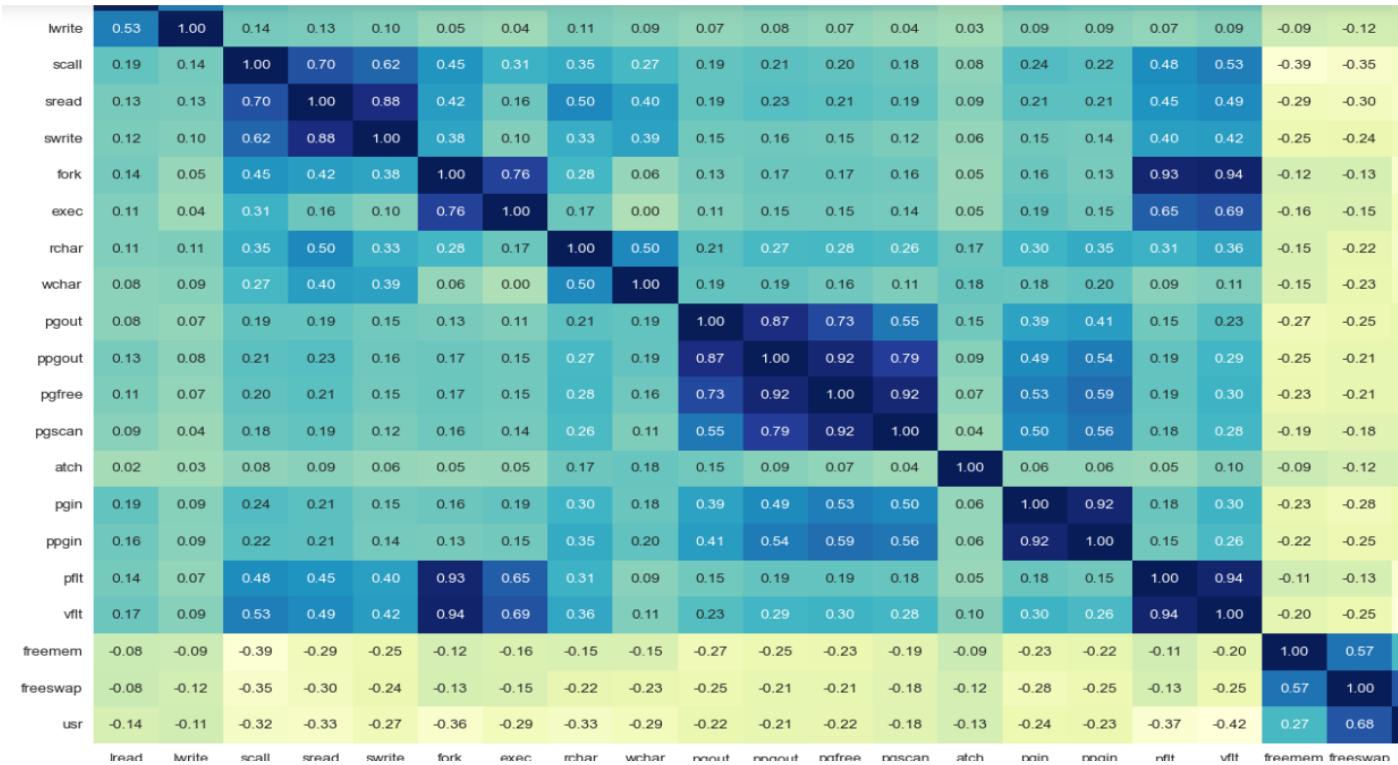


Observations:

The usr is the target variable. freewasp and usr both are left skewed. We can see very few data points in freewap and usr in the range 0,1 and 0-50.

Correlation plot:

Relation between all numeric variables.

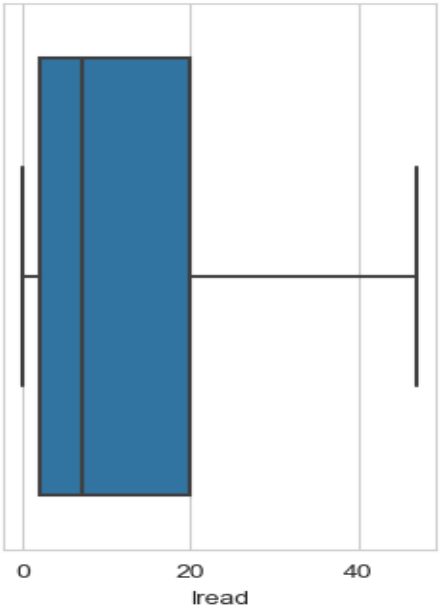


From the correlation plot, we can see that various attributes of the car are highly correlated to each other. Correlation values near to 1 or -1 are highly positively correlated and highly negatively correlated respectively. Correlation values near to 0 are not correlated to each other.

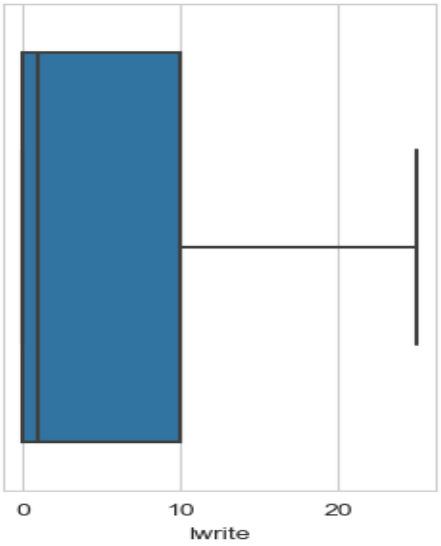
Outlier treatment

After the box plot technique of outlier treatment, the outliers are removed. It uses a technique of lower index and upper index for eliminating the outliers.

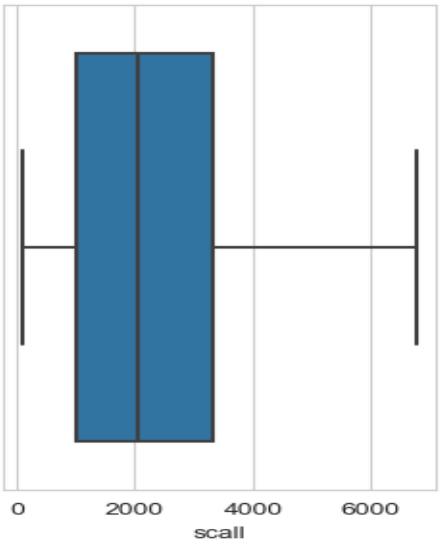
lower range -25.0 and upper range 47.0



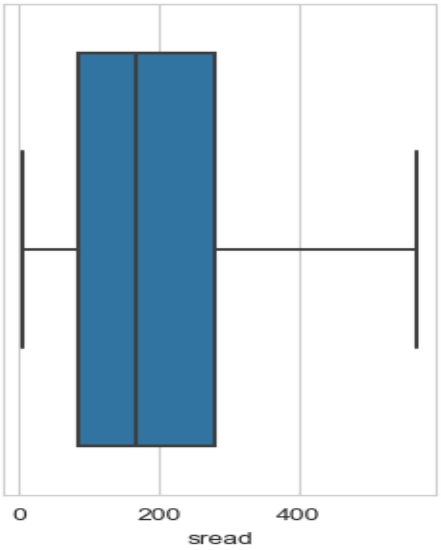
lower range -15.0 and upper range 25.0



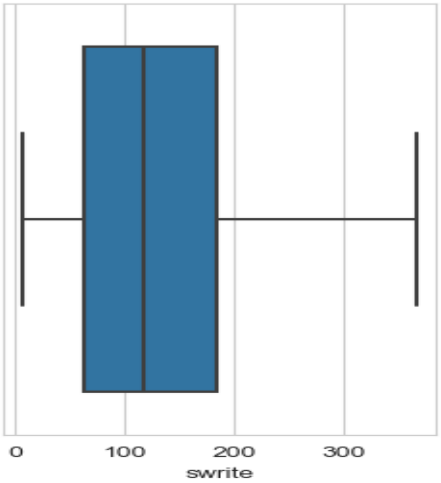
lower range -2445.875 and upper range 6775.125



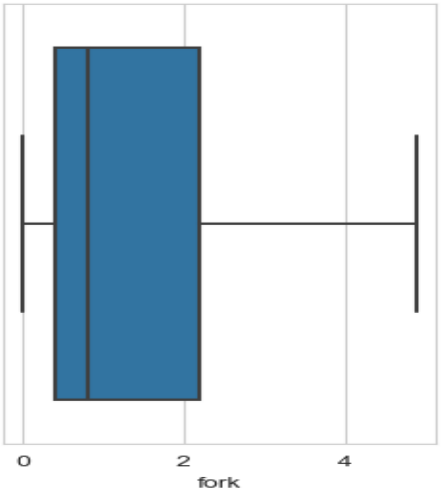
lower range -203.5 and upper range 568.5



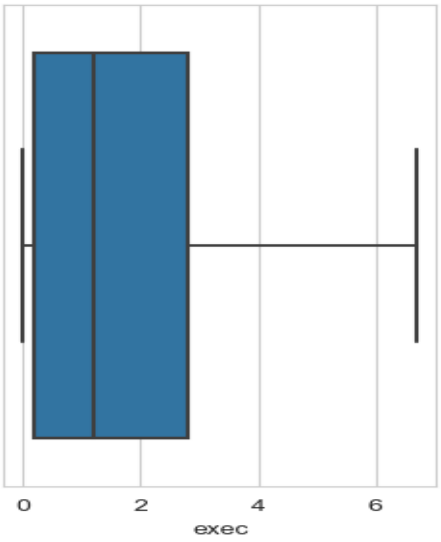
lower range -120.0 and upper range 368.0



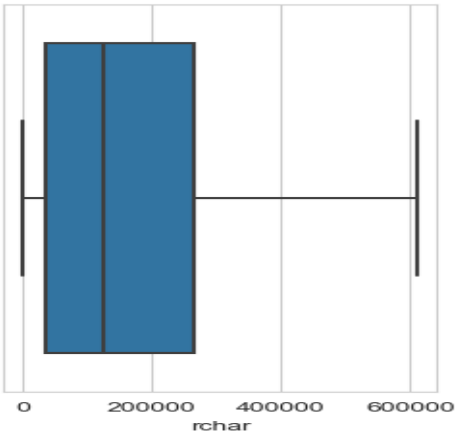
lower range -2.3000000000000003 and upper range 4.9



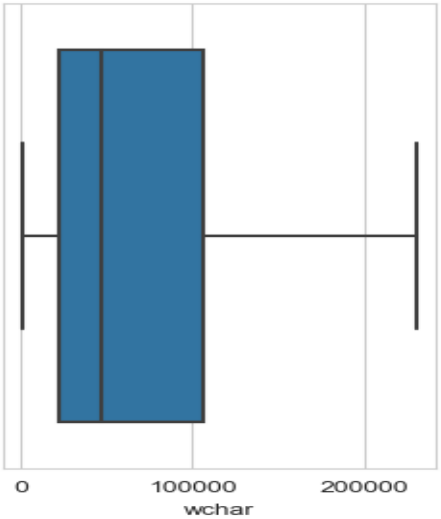
lower range -3.6999999999999993 and upper range 6.699999999999999



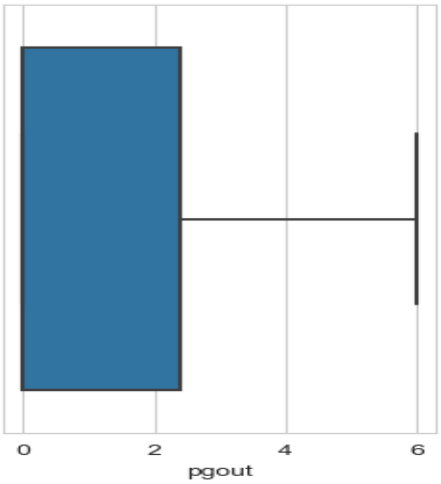
lower range -310940.875 and upper range 611196.125



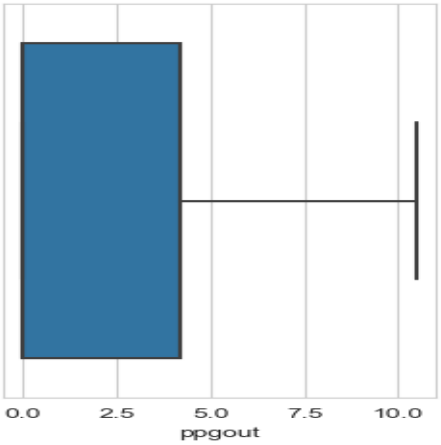
lower range -101611.125 and upper range 230625.875



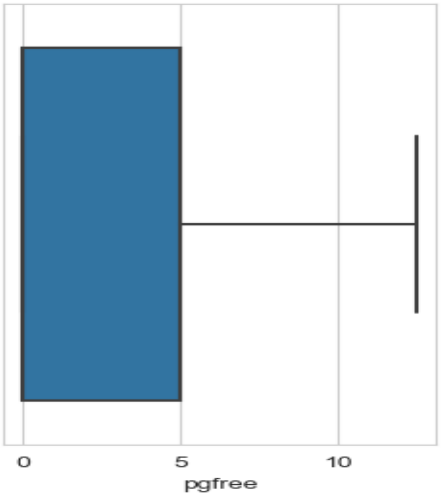
lower range -3.5999999999999996 and upper range 6.0



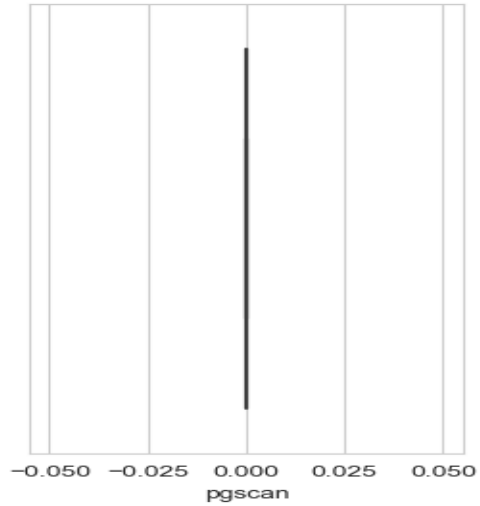
lower range -6.300000000000001 and upper range 10.5



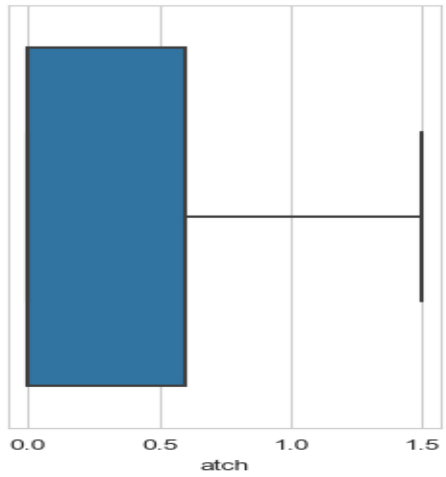
lower range -7.5 and upper range 12.5



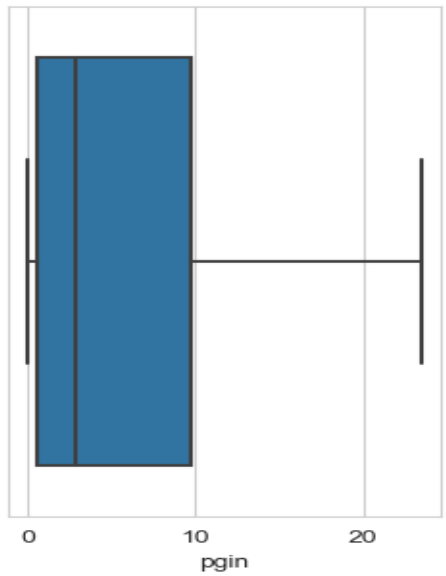
lower range 0.0 and upper range 0.0



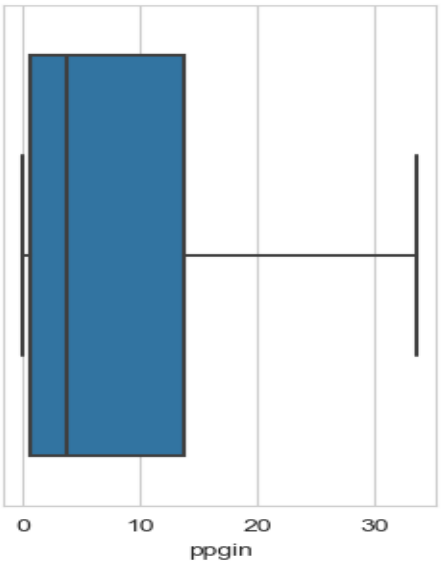
lower range -0.8999999999999999 and upper range 1.5



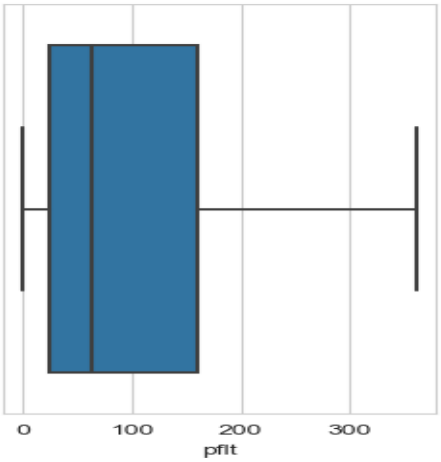
lower range -13.147500000000003 and upper range 23.512500000000003



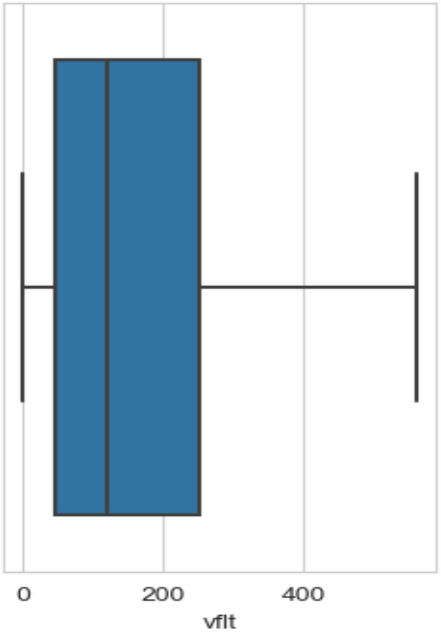
lower range -19.2 and upper range 33.6



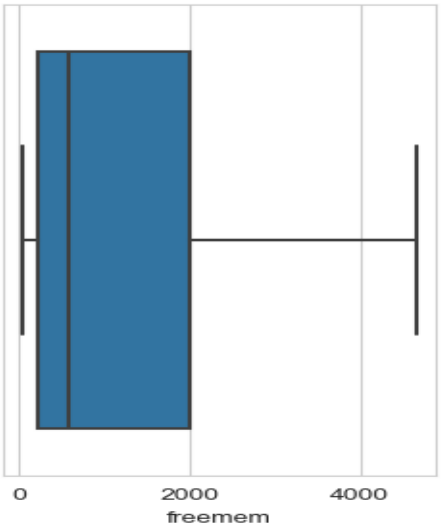
lower range -176.89999999999998 and upper range 361.5



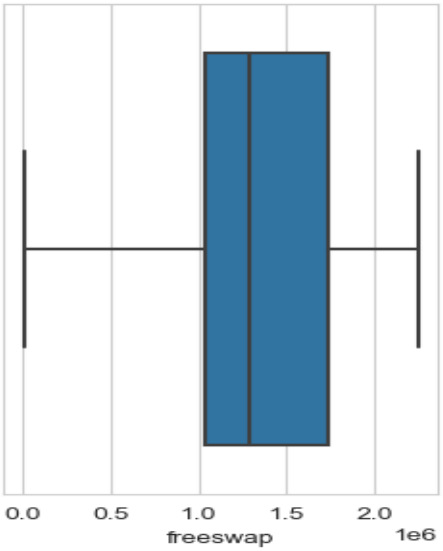
lower range -264.20000000000005 and upper range 561.4000000000001



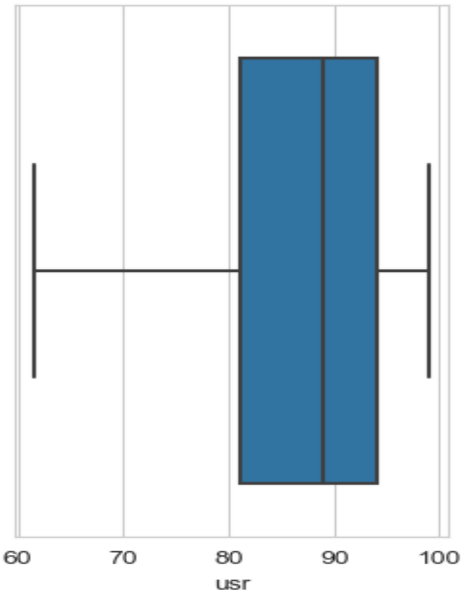
lower range -2425.875 and upper range 4659.125



lower range 10989.5 and upper range 2762013.5

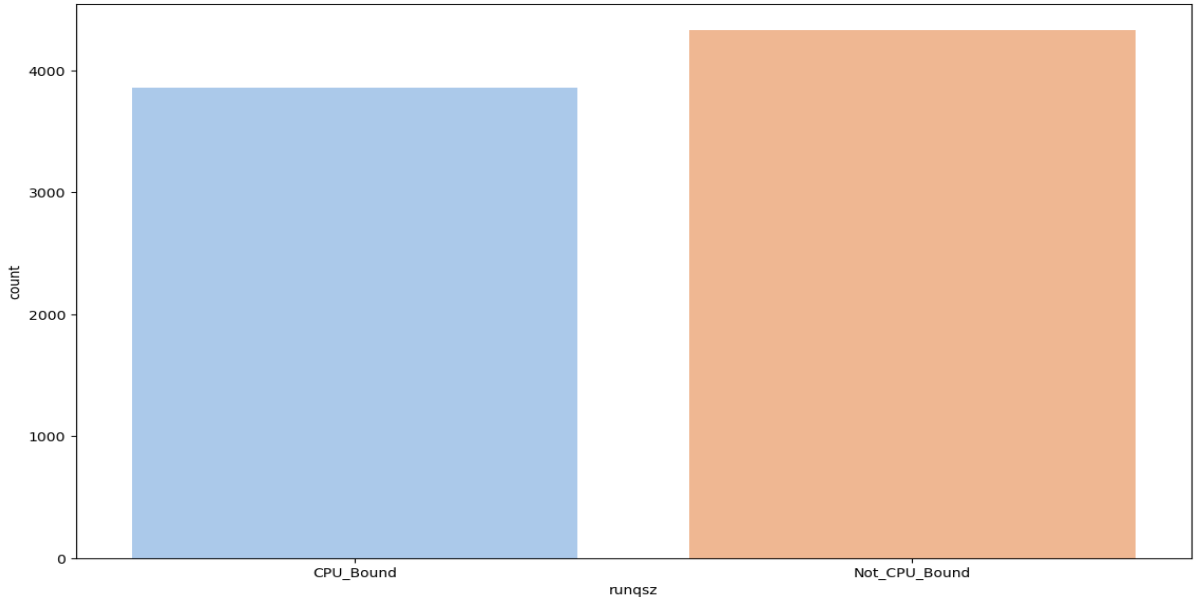


lower range 61.5 and upper range 113.5



Univariant analysis for Categorical variable

runqsz

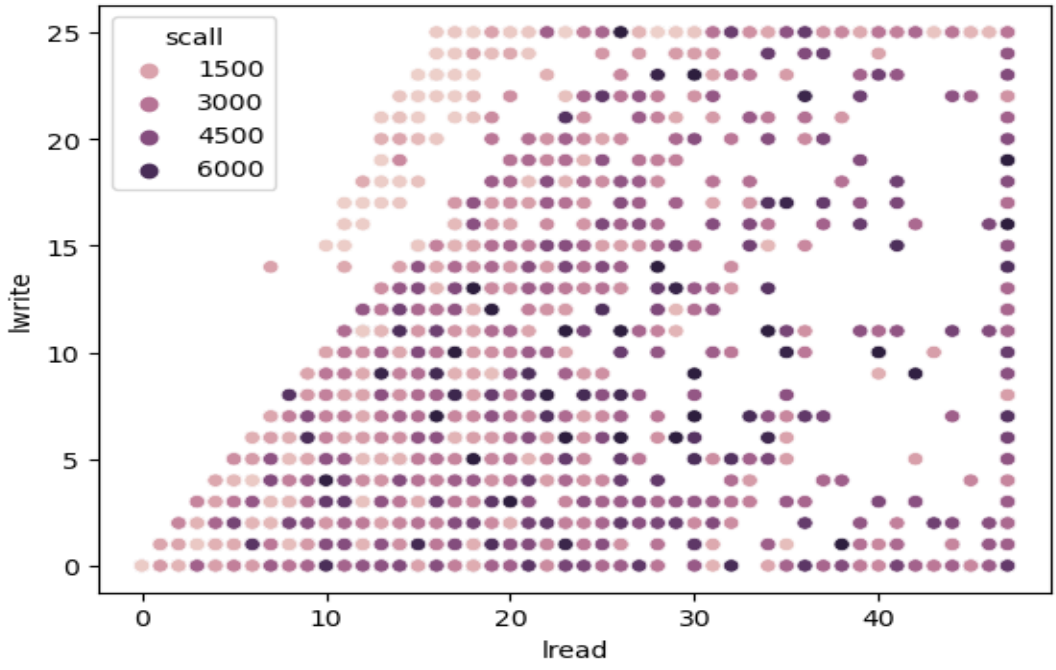


Observations:

We have one categorical variable runqsz. The number of kernel threads in memory that are waiting for a CPU to run is around 3800 and other way its around 4500.

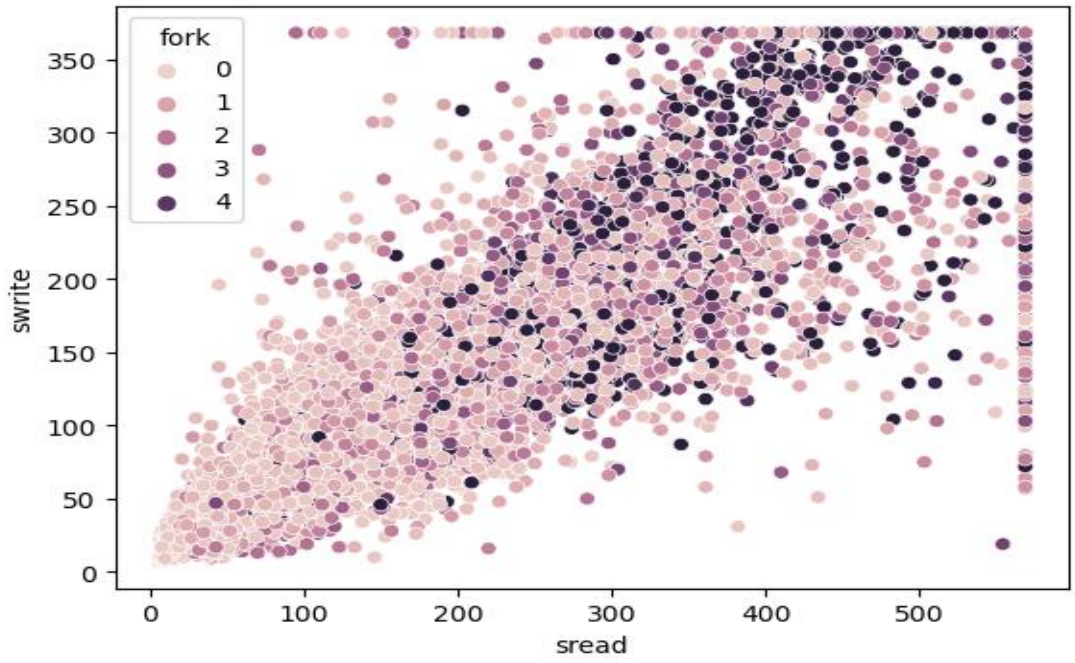
Bivariant analysis

<Axes: xlabel='lread', ylabel='lwrite'>



Observations:

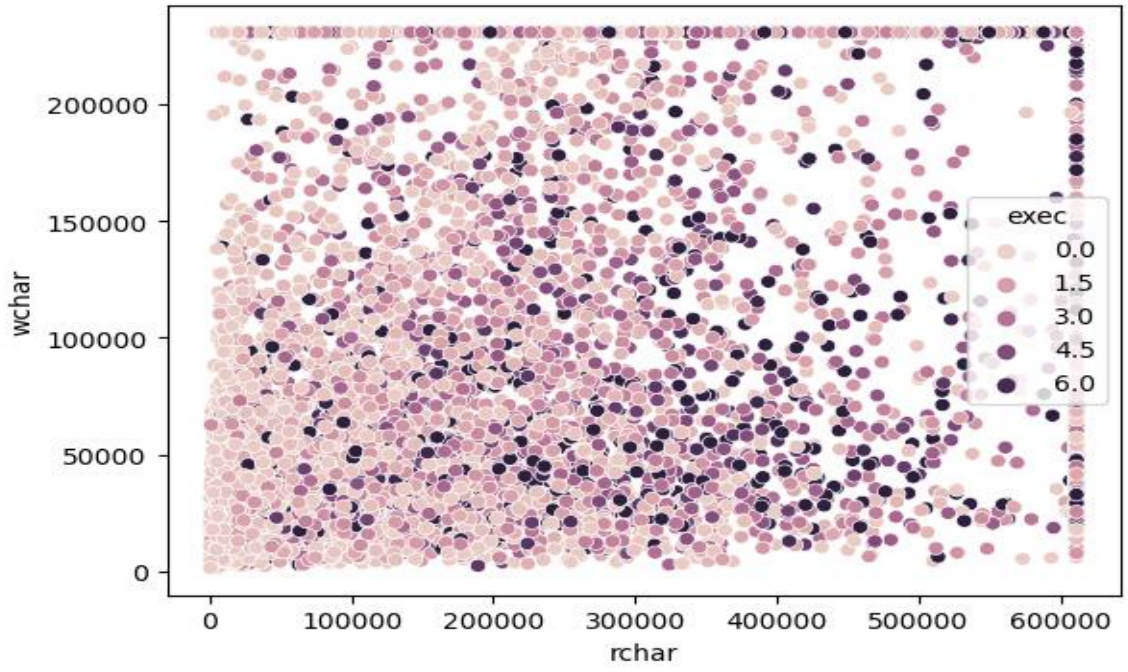
The correlation is increasing as and when the no of writes is increasing. Its slightly positively correlated.



Observations:

The correlation is increasing as and when the no of writes is increasing. Its higly positively correlated with sread.

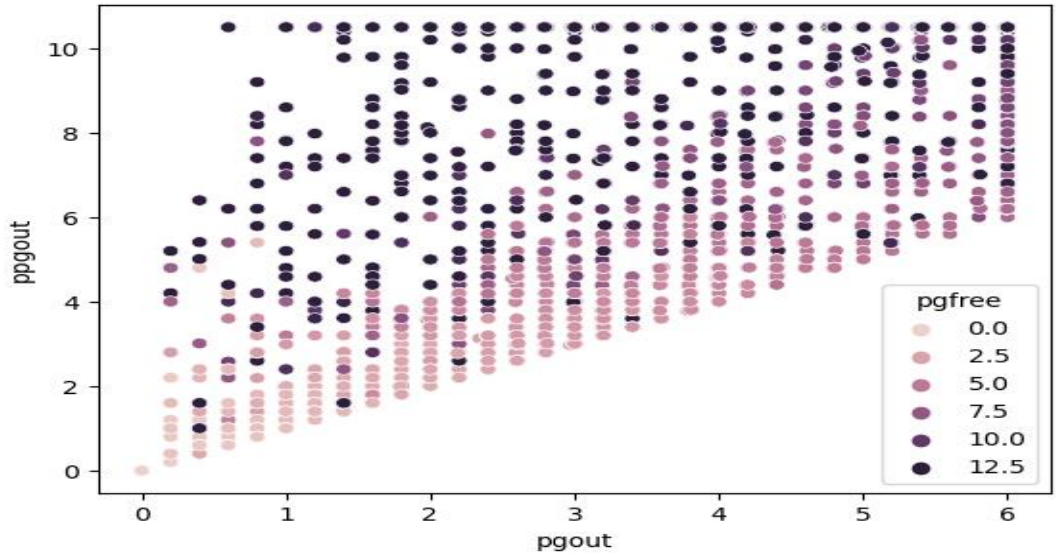
<Axes: xlabel='rchar', ylabel='wchar'>



Observations:

We can see different patterns. When exec is 0.o there is strong correlation. But when exec is 6 its not correlated much.

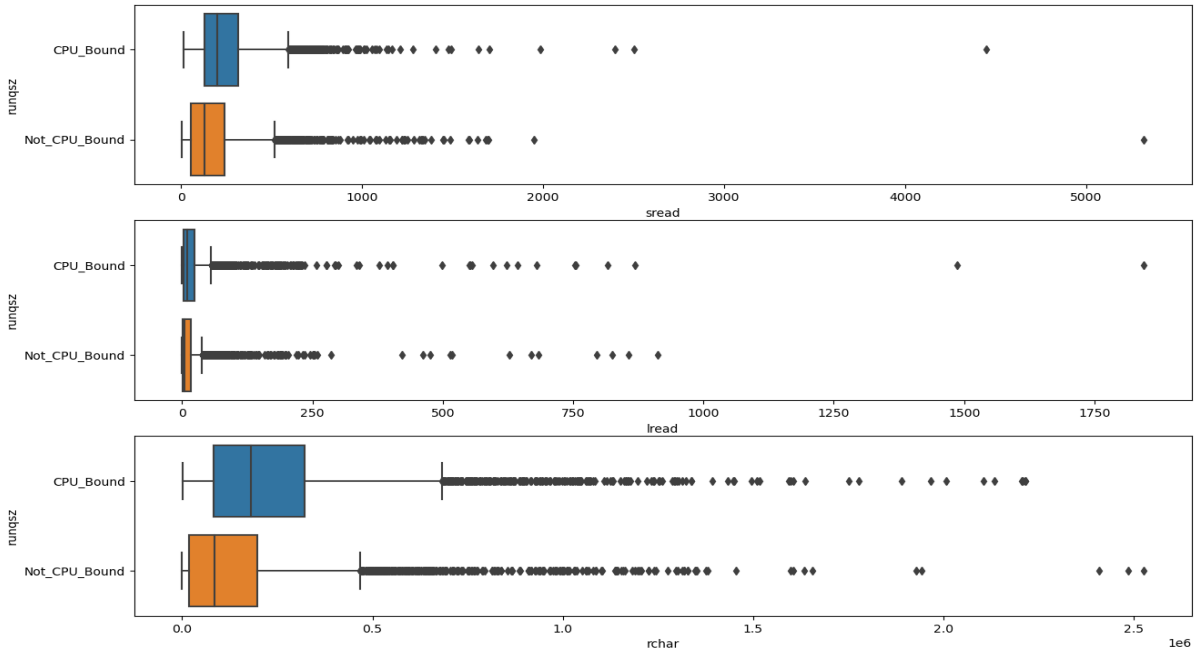
<Axes: xlabel='pgout', ylabel='ppgout'>



Observations:
The number of pages out is correlated.

Categorical vs numerical

<Axes: xlabel='rchar', ylabel='runqsz'>



Overview of Linear Regression Model

Linear regression is a method - Relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the best-fitting linear relationship between the dependent and independent variables.

1. Simple Linear Regression: Where there will be only one independent variable.

   Equation: y=β0+β1x+ϵ

- y is the dependent variable.
- x is the independent variable.
- β0\beta_0β0 is the intercept.
- β1\beta_1β1 is the slope of the line (regression coefficient).
- ϵ\epsilonϵ is the error term (residual).

2. Multiple Linear regression: Where there will be more than one independent variable.

   Equation: y=β0+β1x1+β2x2+…+βnxn+……+ ϵ

Steps in Linear Regression:

1. Model building : Fit a linear regression model.
2. Model Evaluation: Using R^2, adjusted R^2, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).
3. Prediction: Use the model built for prediction with new data.

Assumptions of Linear Regression:

1. Linearity
2. Independence
3. Homoscedasticity
4. Normality
5. No Multicollinearity

Model building:

1. Create dummy variables for the categorical variables.

..	pgscan	atch	pgin	ppgin	pflt	vflt	freemem	freeswap	usr	runqsz_Not_CPU_Bound
..	0.0	0.0	1.6	2.6	16.00	26.40	4670	1730946	95	0
..	0.0	0.0	0.0	0.0	15.63	16.83	7278	1869002	97	1
..	0.0	1.2	6.0	9.4	150.20	220.20	702	1021237	87	1
..	0.0	0.0	0.2	0.2	15.60	16.80	7248	1863704	98	1
..	0.0	0.0	1.0	1.2	37.80	47.60	633	1760253	90	1

2. On the dummy data, drop the target variable and save to X and pop the target variable to y.
3. Split the data in train and test. Test data as 30% and train data as 70%. Using train_test_split function.
4. Add a intercept to data.
5. Fit the linear model using fit function. Ordinal least square method pass the Y and X train data.
6. OLS regression summary will have all the variables. R2 and adj-R2 and covariance

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.643
Model:	OLS	Adj. R-squared:	0.642
Method:	Least Squares	F-statistic:	489.6
Date:	Wed, 03 Jul 2024	Prob (F-statistic):	0.00
Time:	10:43:20	Log-Likelihood:	-21788.
No. Observations:	5734	AIC:	4.362e+04
Df Residuals:	5712	BIC:	4.377e+04
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	44.6380	0.746	59.831	0.000	43.175	46.101
lread	-0.0199	0.003	-6.214	0.000	-0.026	-0.014
lwrite	0.0048	0.006	0.795	0.427	-0.007	0.017
scall	0.0010	0.000	7.451	0.000	0.001	0.001
sread	-0.0005	0.002	-0.257	0.797	-0.004	0.003
swrite	-0.0020	0.002	-1.018	0.309	-0.006	0.002
fork	-1.7222	0.244	-7.052	0.000	-2.201	-1.244
exec	-0.0896	0.048	-1.879	0.060	-0.183	0.004
rchar	-4.062e-06	8.29e-07	-4.898	0.000	-5.69e-06	-2.44e-06
wchar	-1.164e-05	1.28e-06	-9.118	0.000	-1.41e-05	-9.14e-06
pgout	-0.1739	0.064	-2.717	0.007	-0.299	-0.048
ppgout	0.0989	0.037	2.701	0.007	0.027	0.171
pgfree	-0.0703	0.020	-3.508	0.000	-0.110	-0.031
pgscan	0.0086	0.006	1.362	0.173	-0.004	0.021
atch	-0.0786	0.027	-2.949	0.003	-0.131	-0.026
pgin	0.0913	0.029	3.103	0.002	0.034	0.149
ppgin	-0.0594	0.019	-3.128	0.002	-0.097	-0.022
pflt	-0.0415	0.004	-9.697	0.000	-0.050	-0.033
vflt	0.0223	0.003	6.665	0.000	0.016	0.029
freemem	-0.0016	7.53e-05	-21.489	0.000	-0.002	-0.001
freeswap	3.219e-05	4.54e-07	70.985	0.000	3.13e-05	3.31e-05
runqsz_Not_CPU_Bound	7.7908	0.303	25.693	0.000	7.196	8.385

Omnibus:	1507.319	Durbin-Watson:	2.057
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4768.238
Skew:	-1.333	Prob(JB):	0.00
Kurtosis:	6.585	Cond. No.	7.48e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.48e+06. This might indicate that there are

7. check the VIF of the predictors

```
VIF values:

const                     27.191591
lread                      1.472618
lwrite                     1.405898
scall                      2.414301
sread                      6.836403
swrite                     5.320692
fork                      18.210503
exec                       3.059950
rchar                      1.974726
wchar                      1.553348
pgout                      5.776005
ppgout                    15.906900
pgfree                    20.437584
pgscan                     9.237017
atch                       1.087328
pgin                       8.075699
ppgin                      8.672927
pflt                      11.834374
vflt                      20.233207
freemem                    1.677241
freeswap                   1.761193
runqsz_Not_CPU_Bound       1.118922
dtype: float64
```

8. The highest value needs to be removed to reduce the multicollinearity. Let's remove/drop multicollinear columns one by one and observe the effect on our predictive model.¶

9. The model after reducing the multicollinearity the VIF values will be less than 3 it looks like

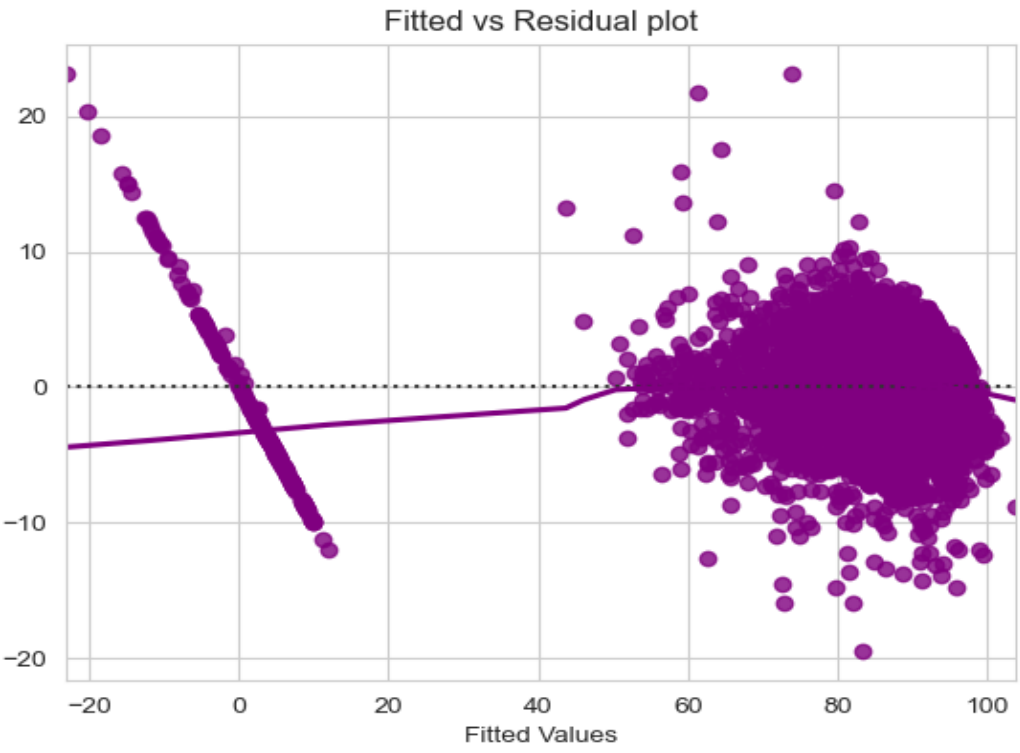
```
VIF values:

const                     25.301809
lread                      1.428892
lwrite                     1.391261
scall                      2.142416
swrite                     1.866997
exec                       1.885393
rchar                      1.602251
wchar                      1.516248
pgout                      2.313369
pgfree                     2.730511
atch                       1.062284
pgin                       1.580404
pflt                       2.327540
freemem                    1.676020
freeswap                   1.654054
runqsz_Not_CPU_Bound       1.117147
dtype: float64
```

10. Check on Assumptions of Linear Regression

	Actual Values	Fitted Values	Residuals
0	91	95.050641	-4.050641
1	94	94.596232	-0.596232
2	0	3.871835	-3.871835
3	83	79.334276	3.665724
4	94	95.181988	-1.181988

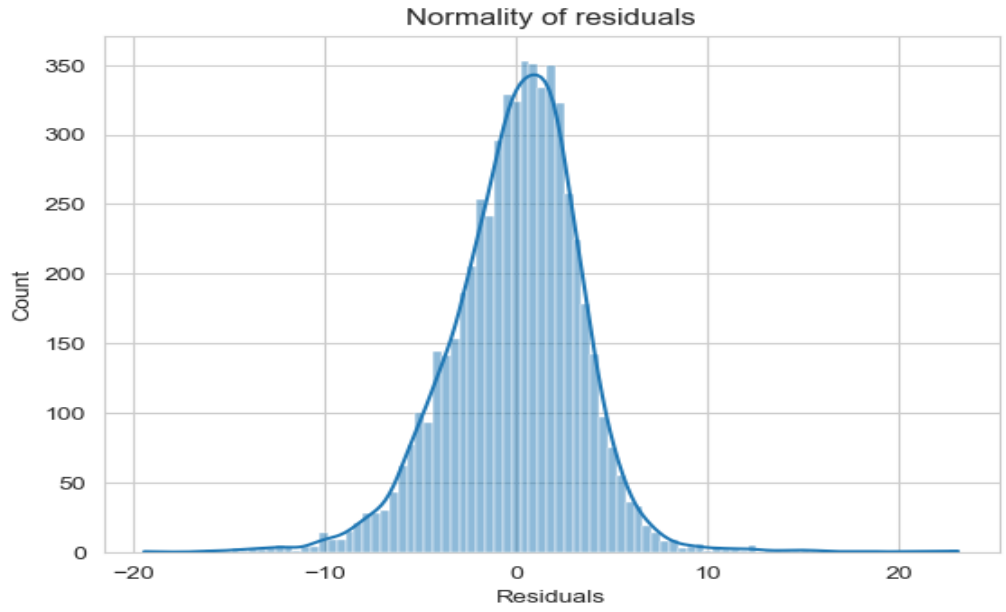
Linearity:

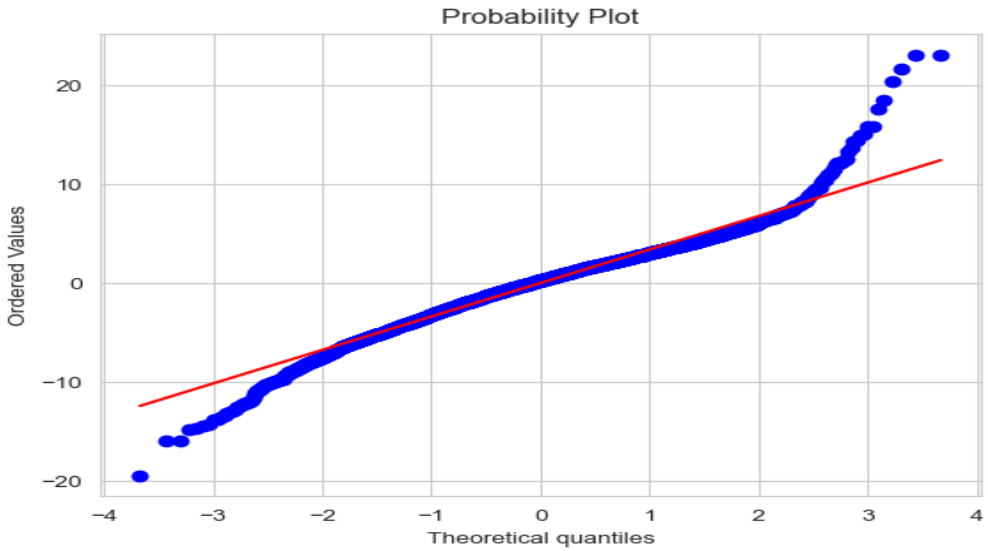
Plot the fitted value vs residuals



Test for Normality:

The plot is slightly left skewed. Plot the QQ plot for residuals to check the normal assumptions.





Few of the 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.

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

HOMOSCEDASTICITY

Homoscedacity - If the variance of the residuals are 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.

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.

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

We transformed freeswap into sqare that resulted in below.

Use transformation like square to transform the residuals. After Transformation the datapoints i.e p value is > 0.05 hence its homoscedastic.

Predictions

Params

```
const                  1.387265e+01
lread                 -9.433174e-03
lwrite                -4.684806e-03
scall                 -1.984157e-03
swrite                -1.827653e-03
exec                  -3.366683e-01
rchar                 -1.248302e-06
wchar                 -4.922668e-06
pgout                 -2.373899e-02
pgfree                -6.277765e-03
atch                   1.555257e-02
pgin                  -5.771292e-02
vflt                  -2.160753e-02
freemem                6.946097e-04
freeswap               1.272636e-04
runqsz_Not_CPU_Bound   1.864810e+00
lread_sq              -2.030707e-06
freeswap_sq           -4.623125e-11
```

The final result summary

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.964
Model:	OLS	Adj. R-squared:	0.964
Method:	Least Squares	F-statistic:	9549.
Date:	Sun, 07 Jul 2024	Prob (F-statistic):	0.00
Time:	15:04:08	Log-Likelihood:	-15214.
No. Observations:	5734	AIC:	3.046e+04
Df Residuals:	5717	BIC:	3.058e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	13.8770	0.270	51.416	0.000	13.348	14.406
lread	-0.0115	0.001	-11.515	0.000	-0.014	-0.010
lwrite	-0.0031	0.002	-1.640	0.101	-0.007	0.001
scall	-0.0020	4.28e-05	-46.290	0.000	-0.002	-0.002
swrite	-0.0018	0.000	-4.795	0.000	-0.003	-0.001
exec	-0.3358	0.013	-26.446	0.000	-0.361	-0.311
rchar	-1.248e-06	2.39e-07	-5.224	0.000	-1.72e-06	-7.79e-07
wchar	-4.922e-06	4.05e-07	-12.168	0.000	-5.72e-06	-4.13e-06
pgout	-0.0237	0.013	-1.840	0.066	-0.049	0.002
pgfree	-0.0063	0.002	-2.690	0.007	-0.011	-0.002
atch	0.0155	0.008	1.851	0.064	-0.001	0.032
pgin	-0.0571	0.004	-13.722	0.000	-0.065	-0.049
vflt	-0.0216	0.000	-52.555	0.000	-0.022	-0.021
freemem	0.0007	2.6e-05	26.771	0.000	0.001	0.001
freeswap	0.0001	4.41e-07	288.633	0.000	0.000	0.000
runqsz_Not_CPU_Bound	1.8630	0.100	18.701	0.000	1.668	2.058
freeswap_sq	-4.623e-11	2e-13	-230.814	0.000	-4.66e-11	-4.58e-11

Omnibus:	393.915	Durbin-Watson:	2.006
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1834.502
Skew:	-0.135	Prob(JB):	0.00
Kurtosis:	5.758	Cond. No.	1.31e+13

The linear equation:

usr = 13.87704181325071 + -0.011539814453154796 * (lread) + -0.00312871982
06714426 * (lwrite) + -0.0019833547009546204 * (scall) + -0.001817790078149
5213 * (swrite) + -0.33582150503431696 * (exec) + -1.2475911941737436e-06 *
(rchar) + -4.922359043197979e-06 * (wchar) + -0.023684172733254558 * (pgo
ut) + -0.006315744294497376 * (pgfree) + 0.01550951818904095 * (atch) + -0.
0571324104859776 * (pgin) + -0.02157824897176316 * (vflt) + 0.000694868783
96285 * (freemem) + 0.0001272527125047064 * (freeswap) + 1.8630417754721
984 * (runqsz_Not_CPU_Bound) + -4.622657406683585e-11 * (freeswap_sq)

Actionable Business Insights and recommendation:

- R-squared of the model is 0.964 and adjusted R-squared is 0.964, which shows that the model is able to explain ~96% variance in the data. This is quite good.
- A unit increase in the atch will result in a 0.0155 unit increase in the usr, all other variables remaining constant.
- The usr of runqsz_Not_CPU_Bound will be 1.8630 units higher than runqsz_Not_CPU not bound.
- When rchar and wchar decrease then usr decreases by a factor of -4.922e-06 and 1.248e-06
- The MAE on train data is 2 and on test its 3.That means we have reduced the no of errors. The RMSE on train data is 3 and on test its 3.

Executive Summary

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.

Introduction

Data Description

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

Sample of the dataset:

	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_
0	24.0	Primary	Secondary	3.0	Scientology	No	2	High	
1	45.0	Uneducated	Secondary	10.0	Scientology	No	3	Very High	
2	43.0	Primary	Secondary	7.0	Scientology	No	3	Very High	
3	42.0	Secondary	Primary	9.0	Scientology	No	3	High	
4	36.0	Secondary	Secondary	8.0	Scientology	No	3	Low	
...	...	...	...	...	...	...	...	...	
1468	33.0	Tertiary	Tertiary	NaN	Scientology	Yes	2	Very High	
1469	33.0	Tertiary	Tertiary	NaN	Scientology	No	1	Very High	
1470	39.0	Secondary	Secondary	NaN	Scientology	Yes	1	Very High	
1471	33.0	Secondary	Secondary	NaN	Scientology	Yes	2	Low	
1472	17.0	Secondary	Secondary	1.0	Scientology	No	2	Very High	

1473 rows × 10 columns

Dataset is in the shape of 1473, 10.

Exploratory Data Analysis

Let us check the types of variables in the data frame.

```
<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
```

Total 10 columns with 1473 rows. Out of 10 columns, 2 are float, 1 is integer and 7 is object data type.

Check the summary statistics

	count	mean	std	min	25%	50%	75%	max
Wife_age	1402.0	32.606277	8.274927	16.0	26.0	32.0	39.0	49.0
No_of_children_born	1452.0	3.254132	2.365212	0.0	1.0	3.0	4.0	16.0
Husband_Occupation	1473.0	2.137814	0.864857	1.0	1.0	2.0	3.0	4.0

Check for missing values in the dataset:

From the above results we can see that there missing data for Wife_age and No_of_childern_born we need to treat the bad data.

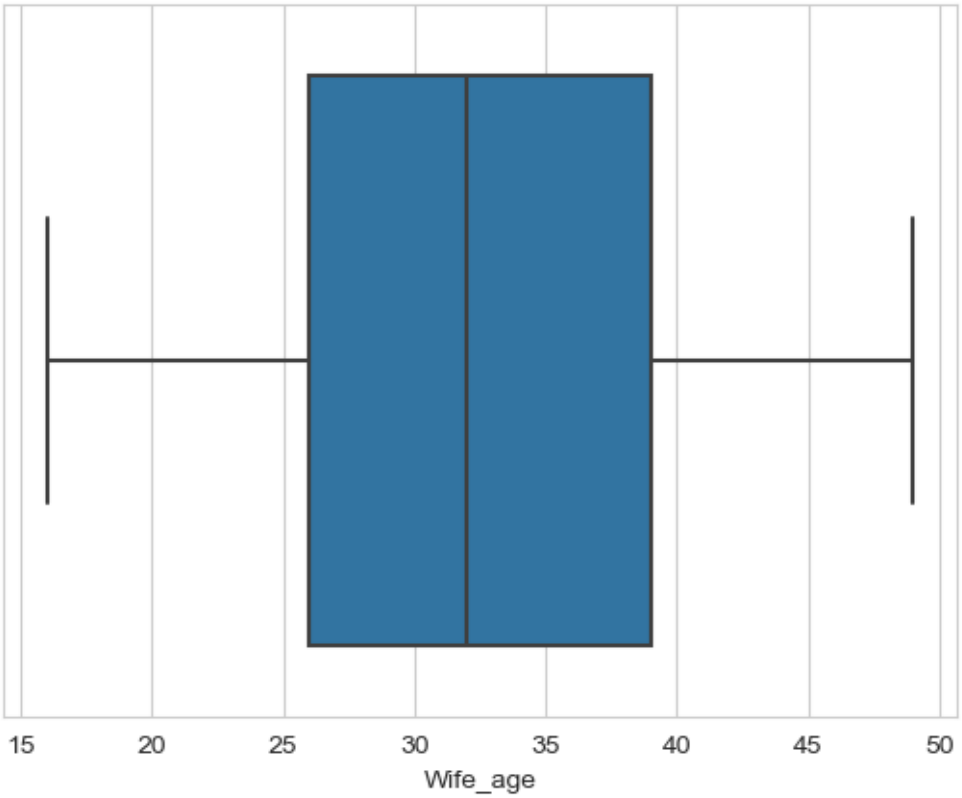
Check for duplicate values in the dataset:

There are 80 duplicates in the dataset. Drop the duplicates.

Treat the bad data i.e missing values in the dataset:

Wife_age:

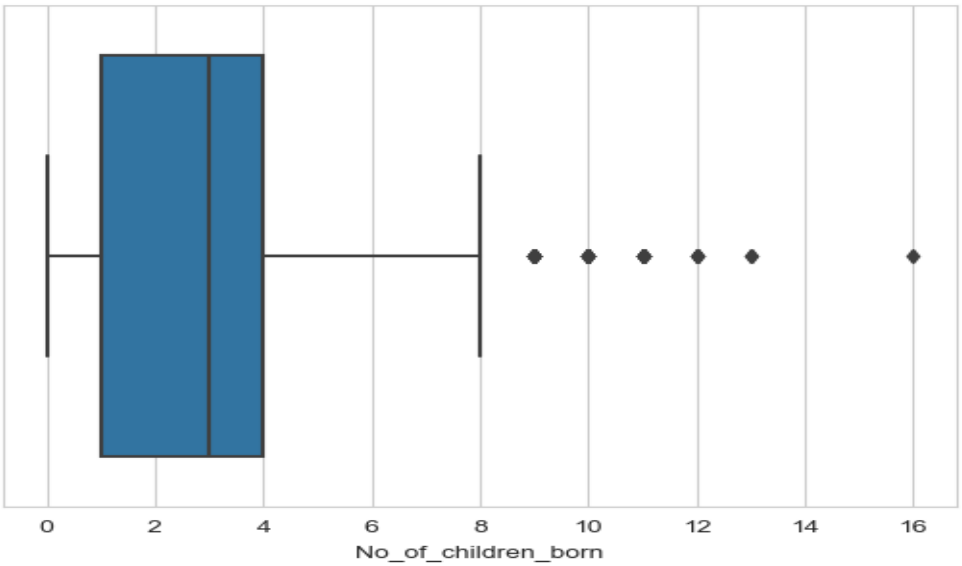
<Axes: xlabel='Wife_age'>



Its normally distributed hence we can impute Wife_age with mean.

No_of_children_born:

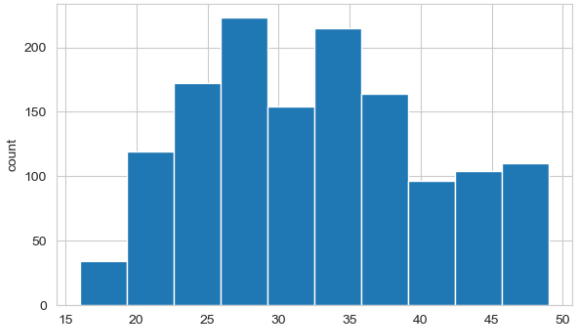
<Axes: xlabel='No_of_children_born'>

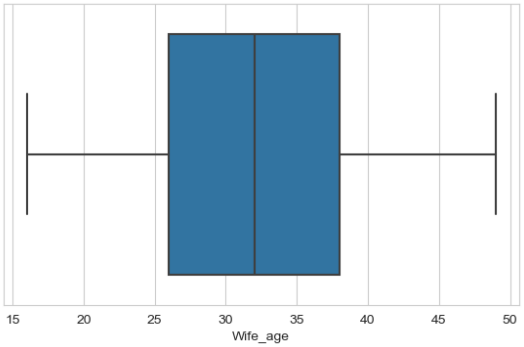


Its not normally distributed hence we can impute No_of_children_born with median not with mean.

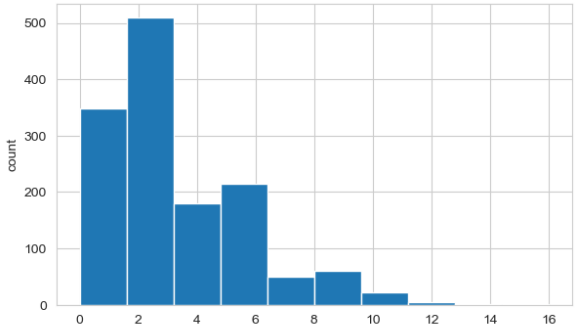
Univariant Analysis

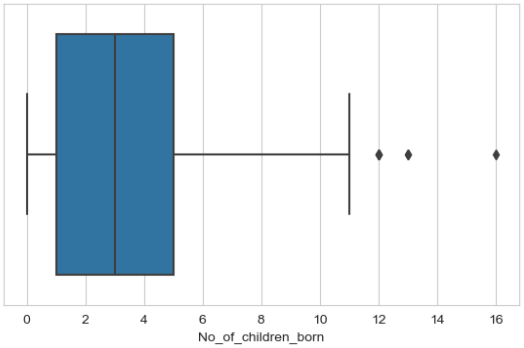
Wife_age



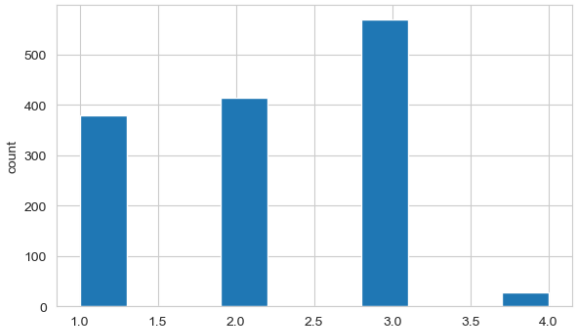


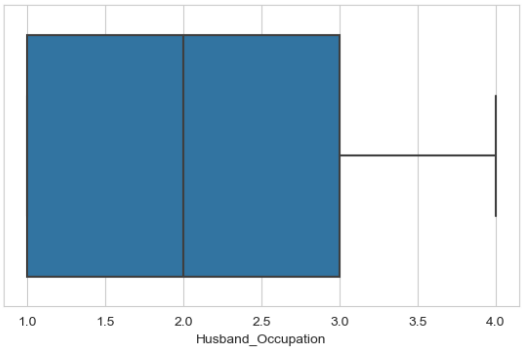
No_of_children_born





Husband_Occupation





Observation:

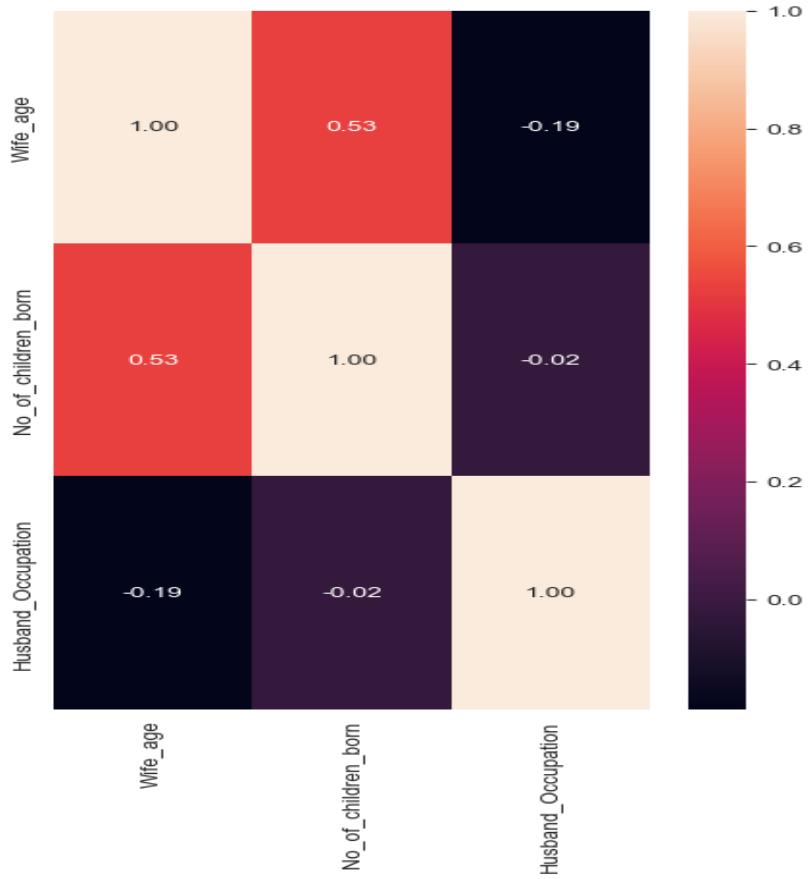
The Wife_age is normally distributed. The median is 32.

The No_of_children_born have outliers. Mean is around 3 and max is going to 16 which is out of normal range.

Husband_Occupation doesn't have Q1.

Correlation plot: Relation between numeric variables.

<Axes: >

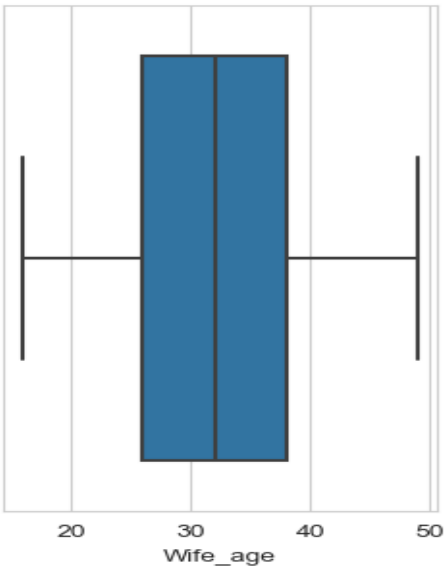


From the correlation plot, we can see that various attributes of the car are highly correlated to each other. Correlation values near to 1 or -1 are highly positively correlated and highly negatively correlated respectively. Correlation values near to 0 are not correlated to each other.

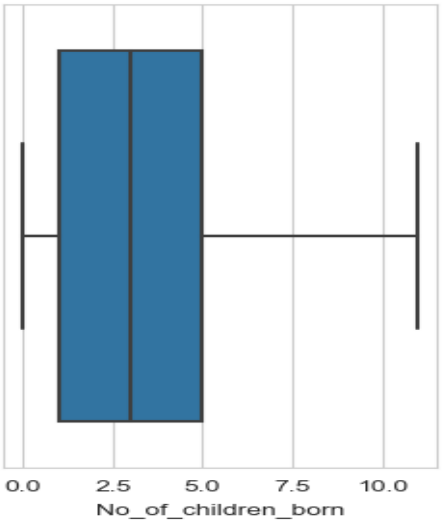
Outlier treatment:

After the box plot technique of outlier treatment, the outliers are removed. It uses a technique of lower index and upper index for eliminating the outliers.

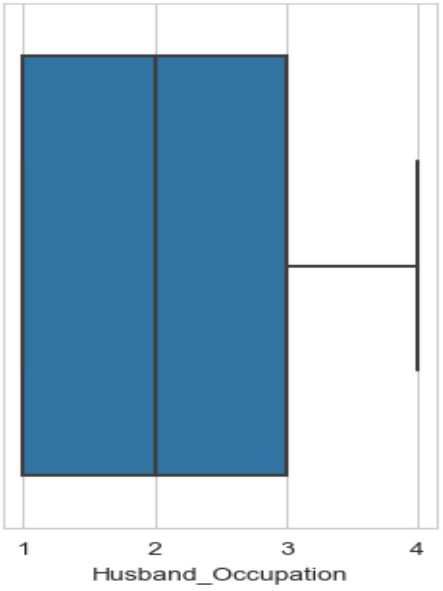
lower range 8.0 and upper range 56.0



lower range -5.0 and upper range 11.0

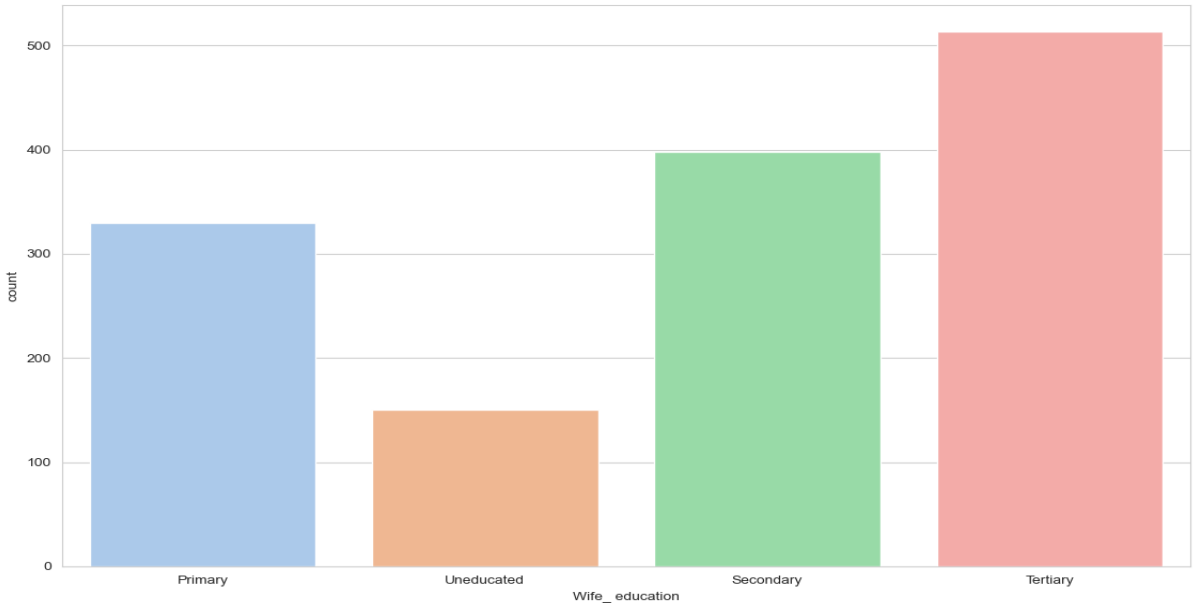


lower range -2.0 and upper range 6.0



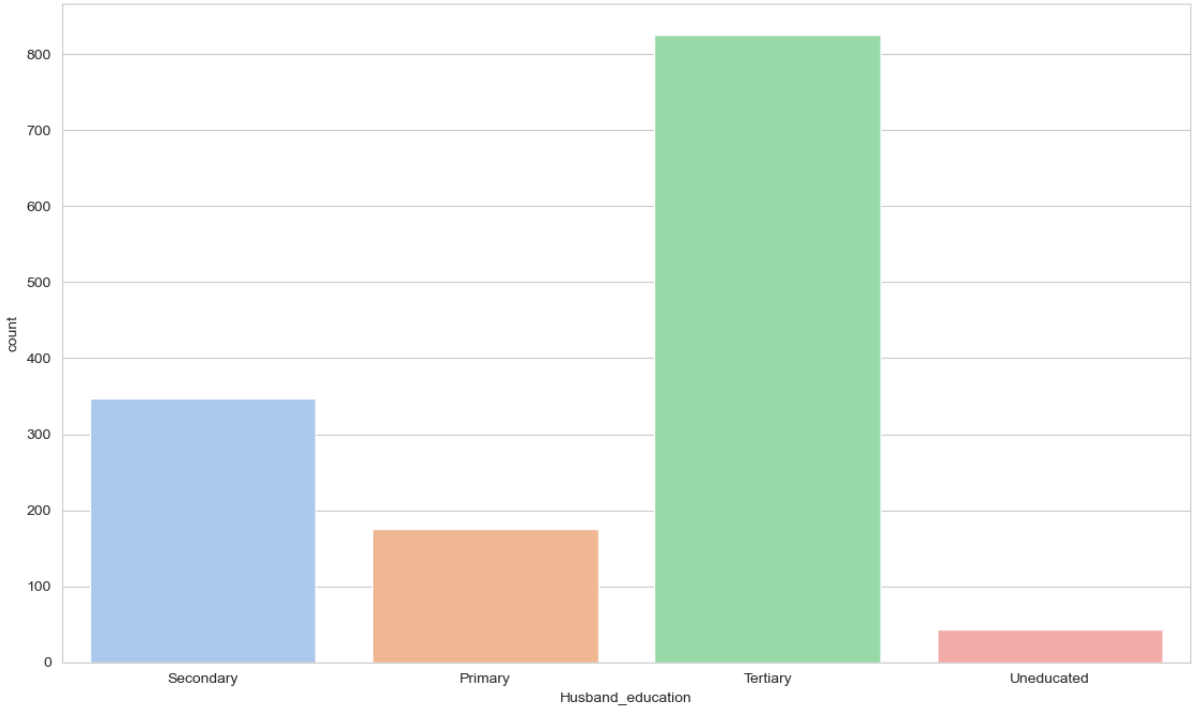
Univariant analysis for Categorical variable

Wife_ education



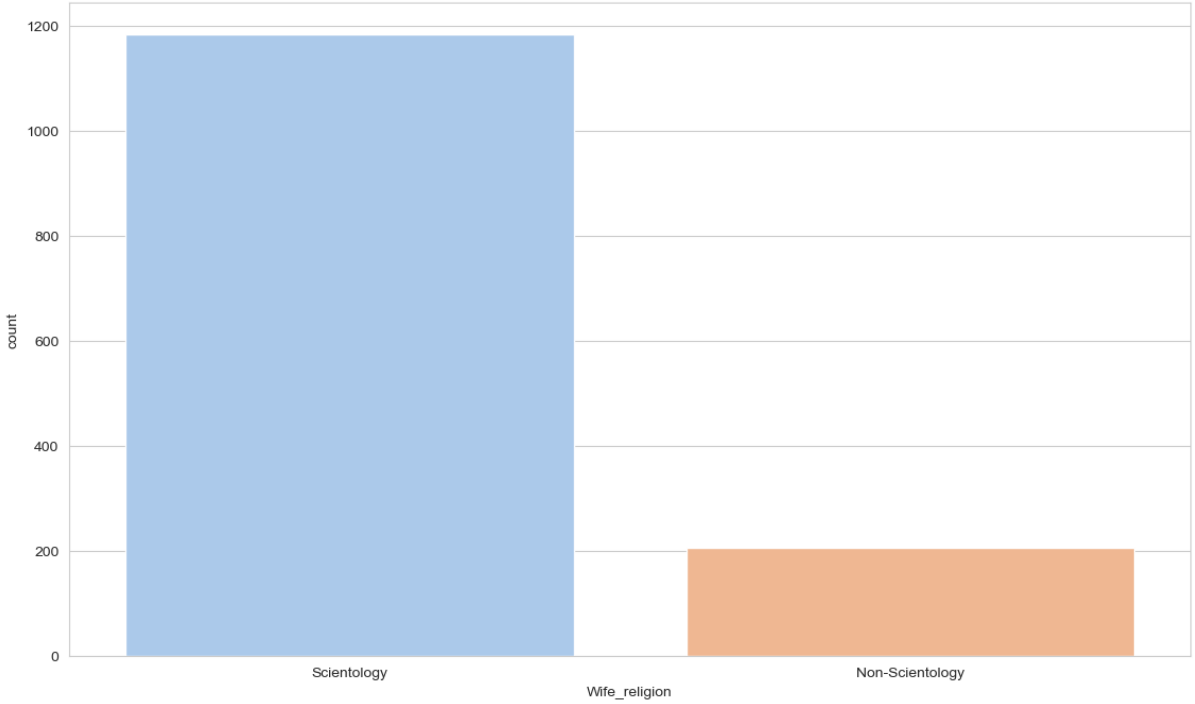
The number of uneducated in women is 150 around. The highest educated women is 500+ which is pretty hight that means womens are more educated. Basic education of primary is done by 350 people.

Husband_education



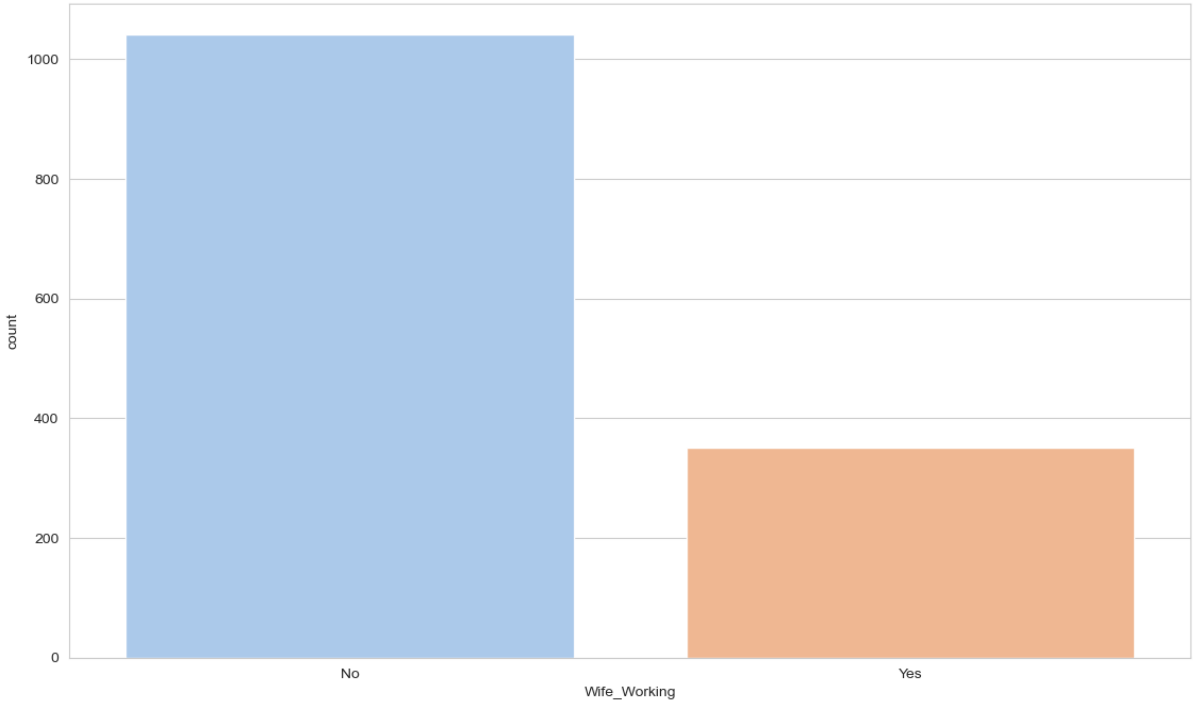
The highest educated women is 800+ which is pretty hight that means Men are more educated in highest degree. There are very few less uneducated men.

Wife_religion



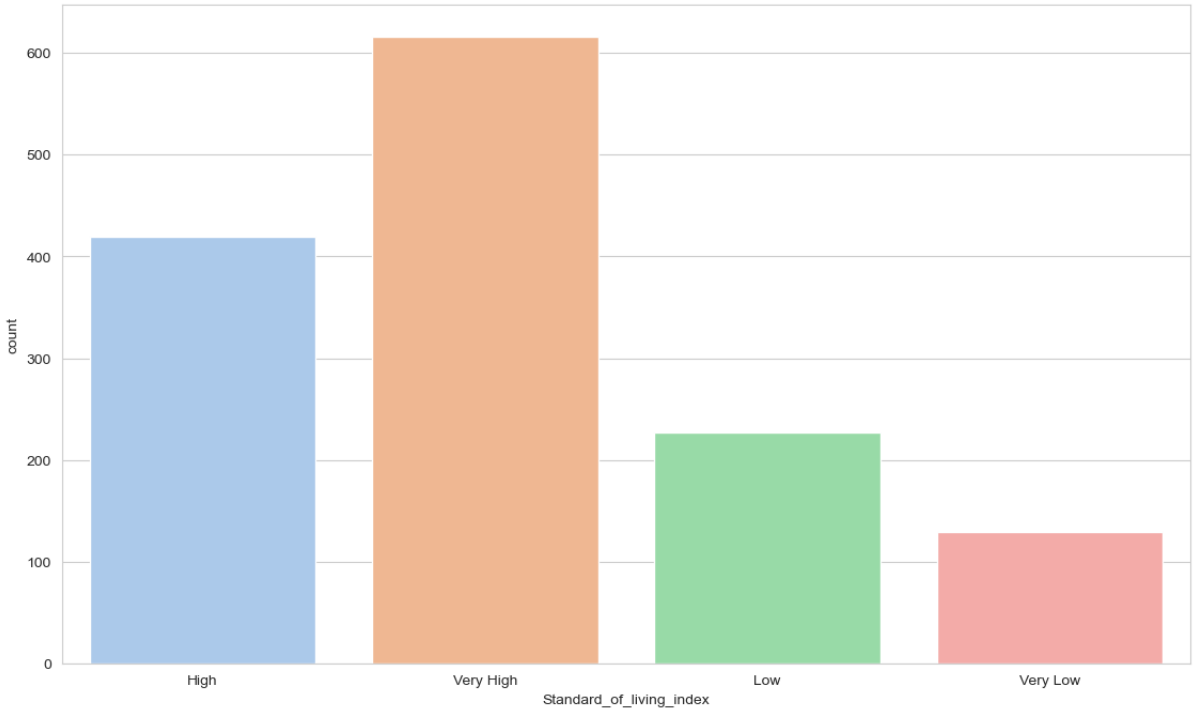
We see scientology more than other.

Wife_Working



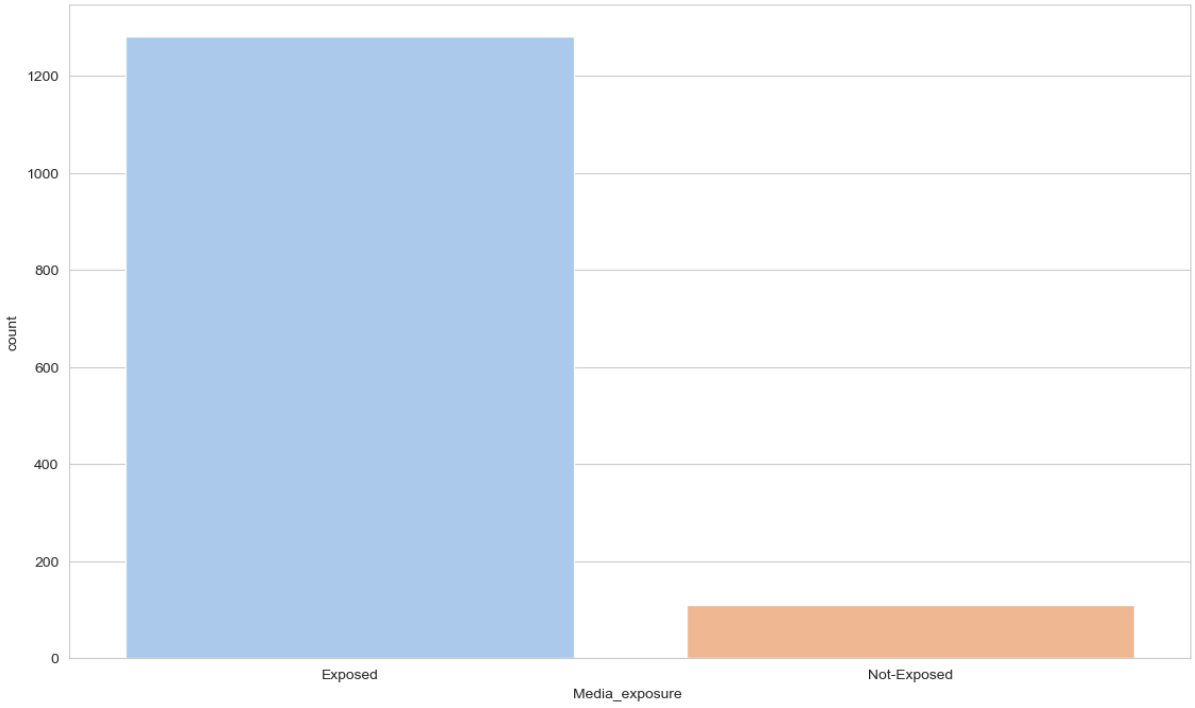
The Wife working count is very less compared to non working women.

Standard_of_living_index



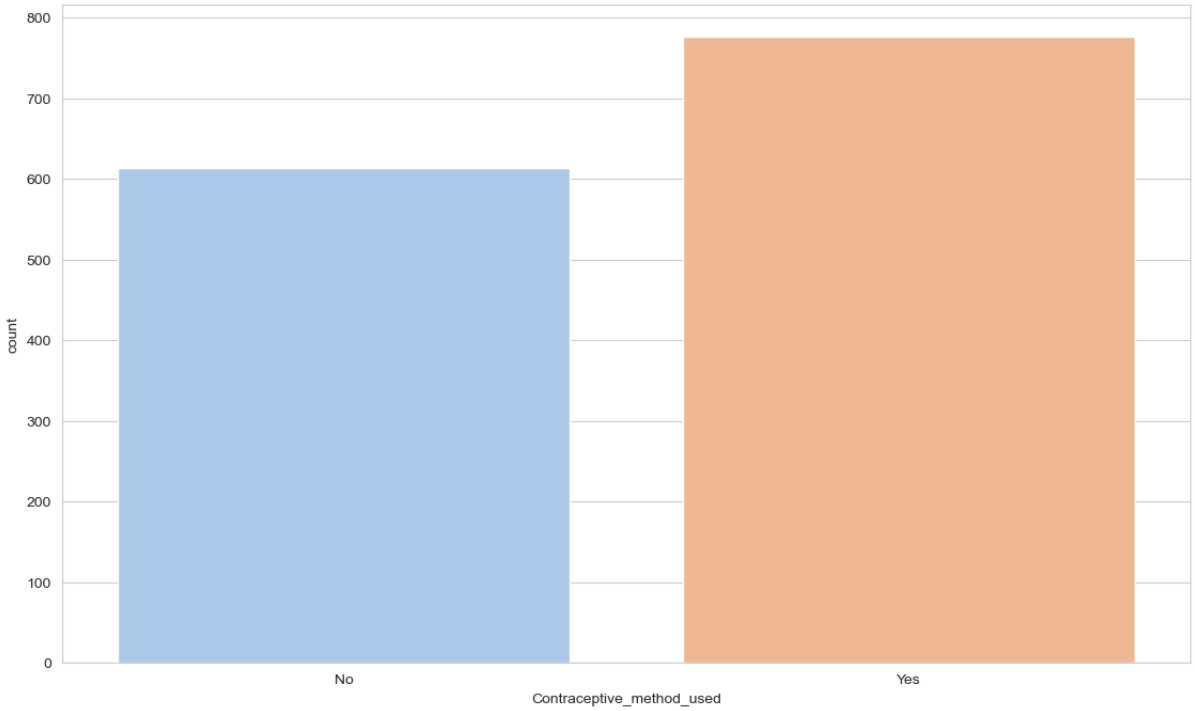
Stndard living index is in all ranges from very hight to very low. Majority lifestyle looks very high.

Media_exposure



More than half of the population is exposed to media to know about contracetive.

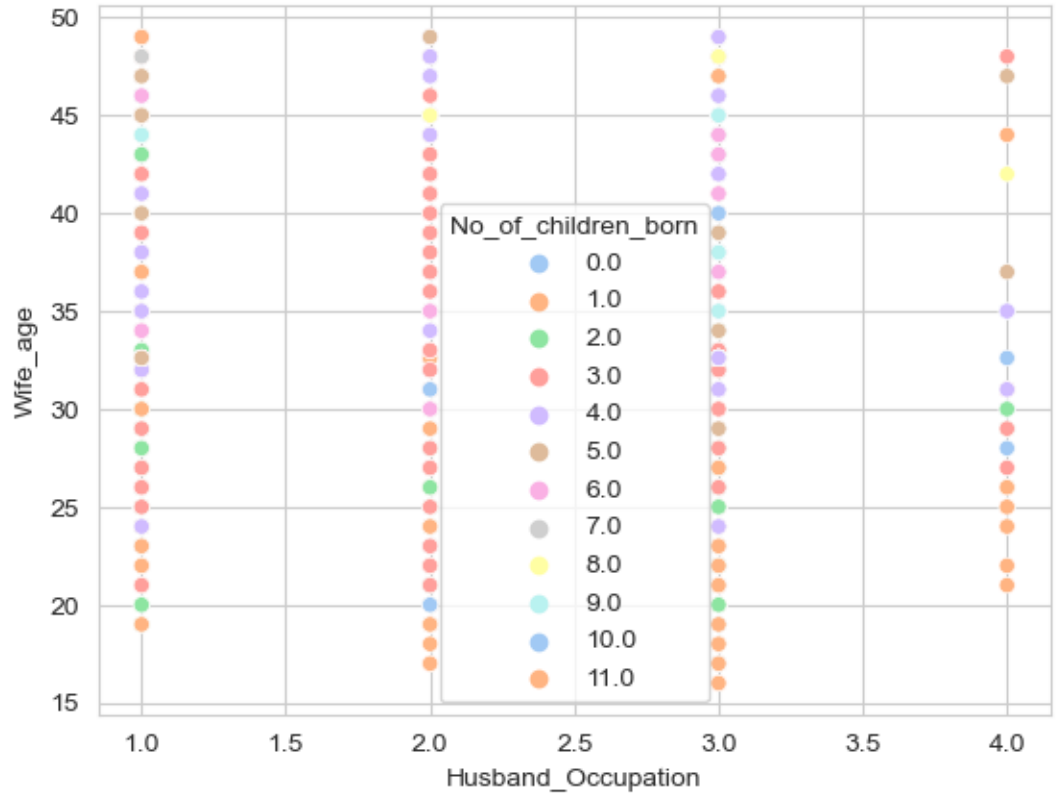
Contraceptive_method_used



Most of the population have opted for Contraceptive.

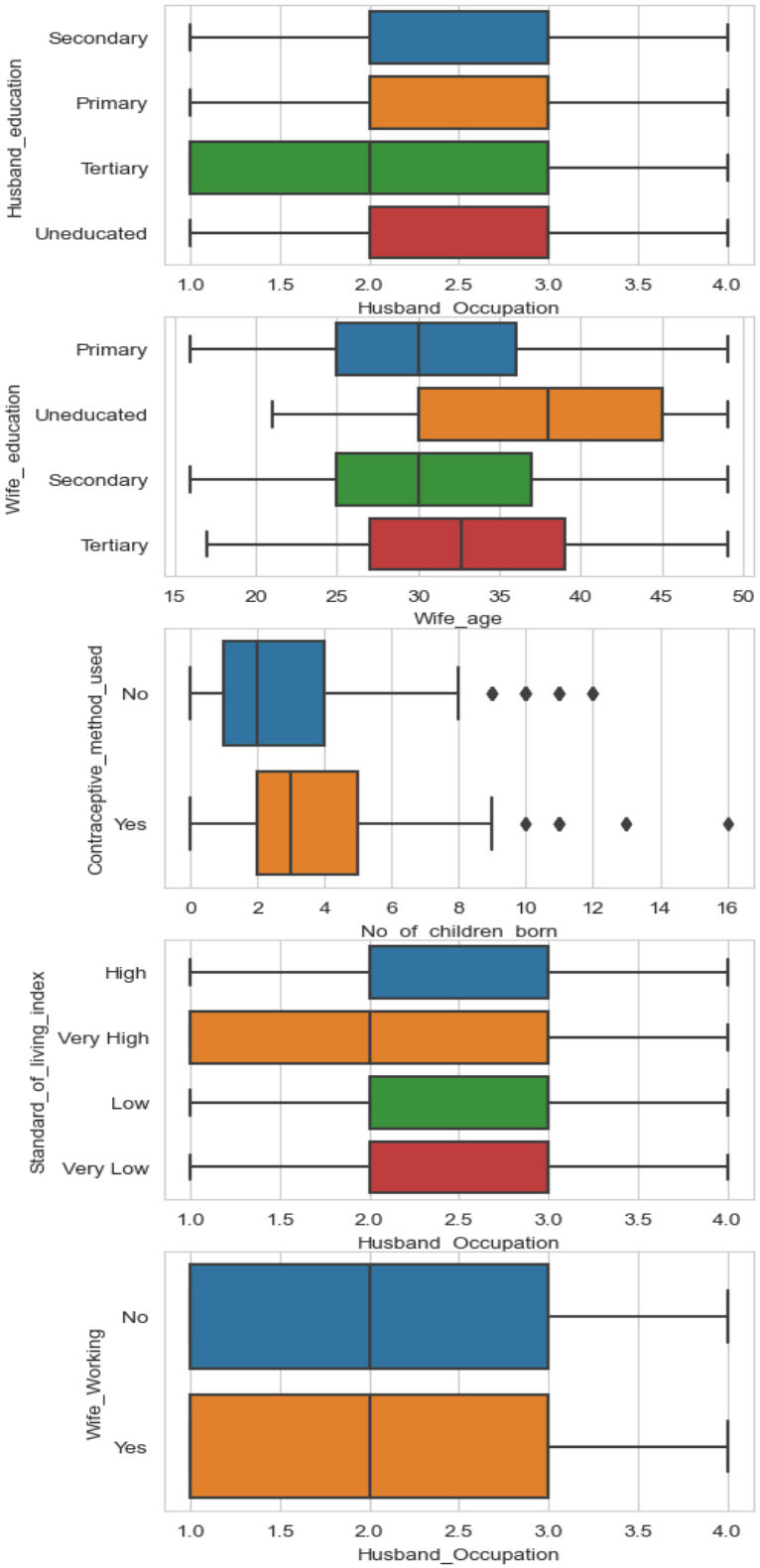
Bivariant analysis

<Axes: xlabel='Husband_Occupation', ylabel='Wife_age'>



Categorical vs numerical

<Axes: xlabel='Husband_Occupation', ylabel='Wife_Working'>



Overview of Logistic Regression

Logistic regression is a method for binary classification problem. The model helps to classifies the dataset into different groups.

Steps for Logistic Regression

1)Model building – Fit the model
2)Model evaluation - Confusion matrix, Classification report, F1 score, recall and precision score.
3)Model prediction – AUC-ROC curve.

Steps Model building:

1) Pre process the data
2) Check for object data type which are categorical transform them using one hot encoding or some grouping basis.
3) Converting the other 'object' type variables as dummy variables

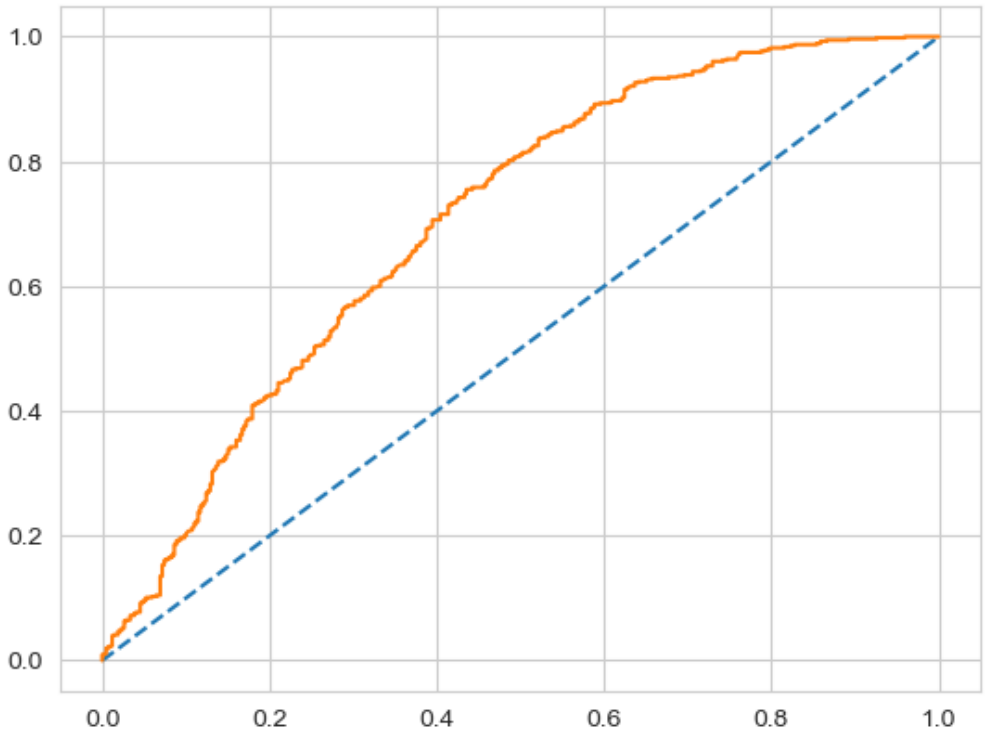
Media_exposure	Contraceptive_method_used	Wife_ education_UnEducated	Husband_education_UnEduc
1	0	0	
1	0	1	
1	0	0	
1	0	0	
1	0	0	

4) Split the Train and test split. With X as independent variable and y as dependant variable.
5) Fit the logistic regression model.

```
LogisticRegression
LogisticRegression()
```

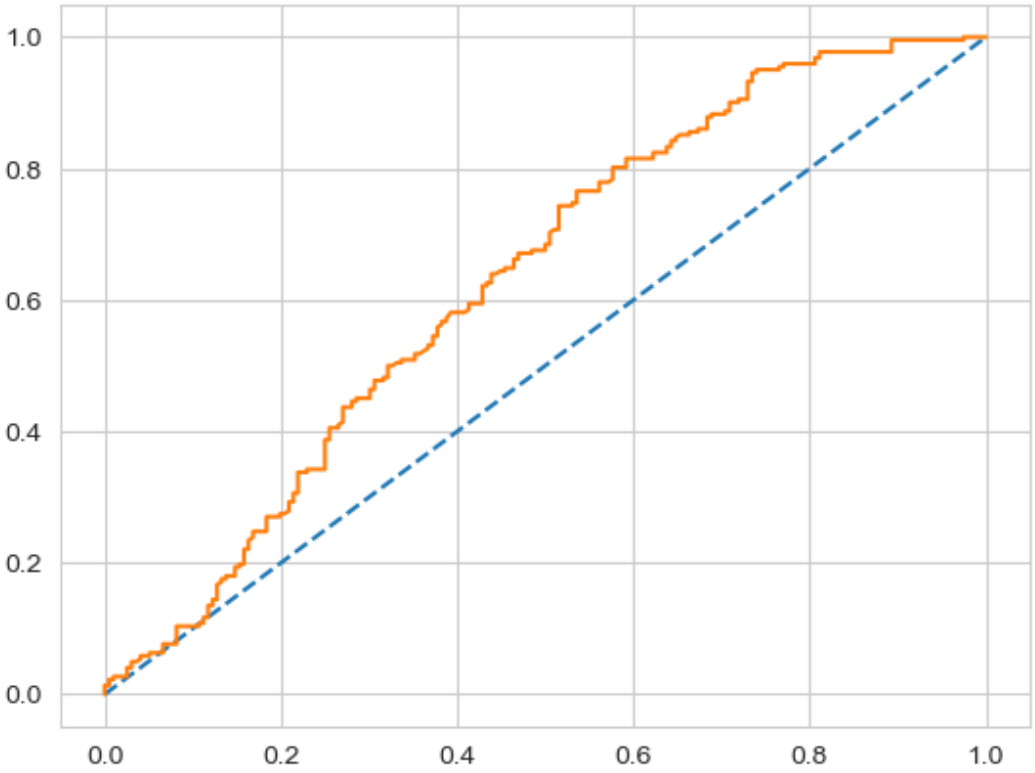
6) Predicting on Training and Test dataset.
7) Evaluate the model. Calculate the mode score to get accuracy from both training and test data. For training the accuracy is 67%. and test data accuracy is 61%
8) AUC-ROC for the train data

```
AUC: 0.703
```

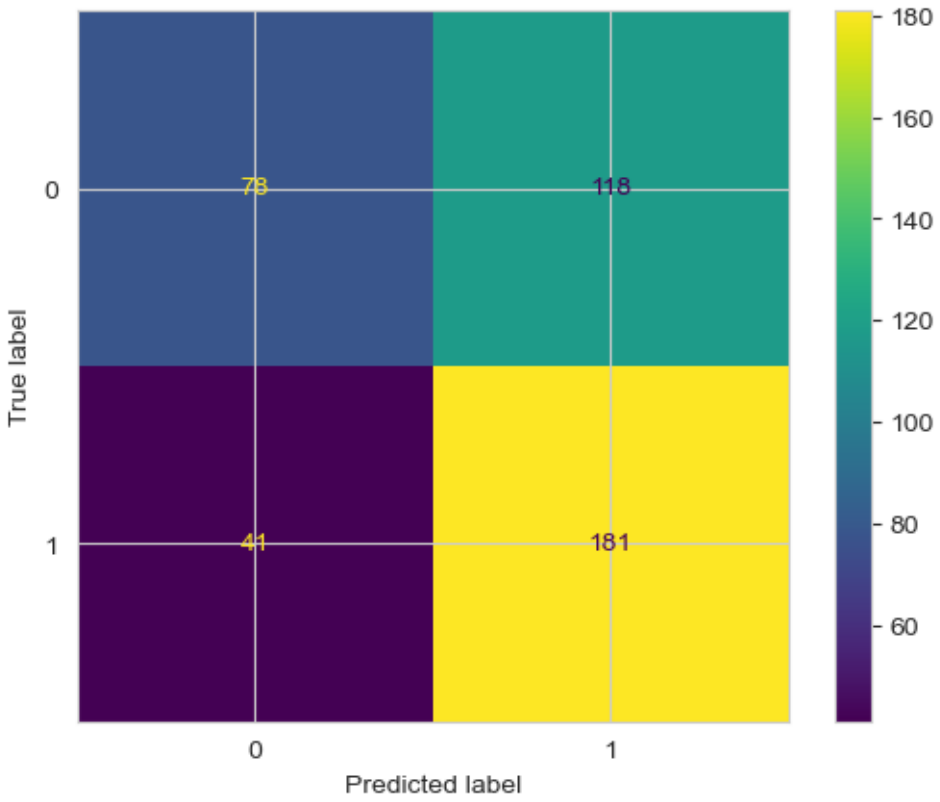


9)AUC-ROC for test data

```
AUC: 0.703
```



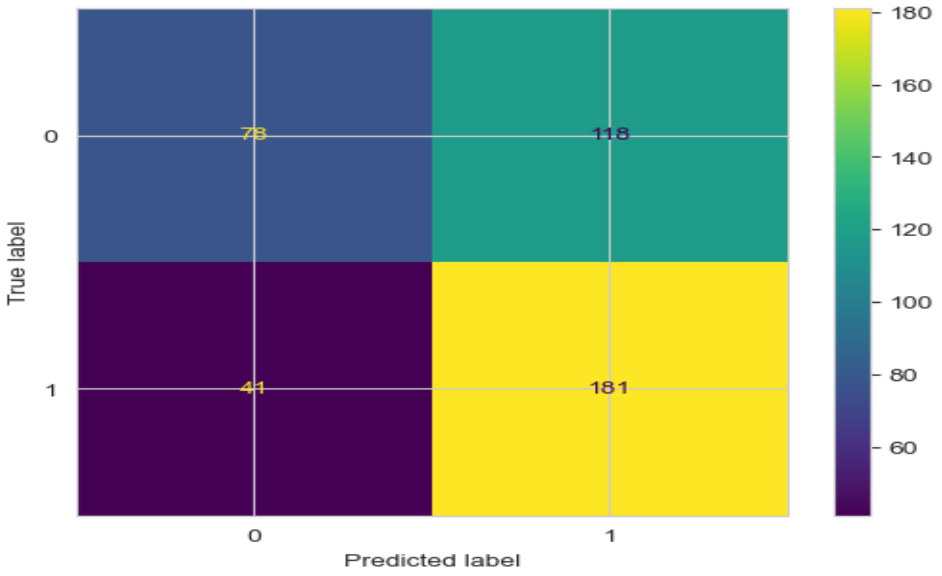
10)Compute the Confusion matrix for training data



Classification report for train data

	precision	recall	f1-score	support
0	0.68	0.48	0.56	418
1	0.68	0.83	0.75	555
accuracy			0.68	973
macro avg	0.68	0.65	0.65	973
weighted avg	0.68	0.68	0.67	973

11)Compute the Matrix for test data



Classification report for test data

	precision	recall	f1-score	support
0	0.66	0.40	0.50	196
1	0.61	0.82	0.69	222
accuracy			0.62	418
macro avg	0.63	0.61	0.60	418
weighted avg	0.63	0.62	0.60	418

Actionable Insights and recommendation:

- There is no overfitting in the dataset. Since train and test dataset val ues are same.
- The Accuracy score is 70%. Can be considered as moderately best model.
- The Recall of class 1 is high 0.83 means the type 2 error is low.
- But for class 0 its low hence we can say the type 2 error exists.
- The Precision for class 1 and 0 is 0.68 which means the ratio of contr aceptive used by total of used and false predicted.
- There is a chance of type 1 error.
- In this case the Type 1 error could be more looked into since if the n umber of false positives increases then chances are No of Childers b orn might increase due to non-usage of contraceptive.

Overview of LINEAR DISCRIMINANT ANALYSIS

LDA is classification technique used to classify the features into two or more classes. It is also used for dimensionality reduction technique.

Assumption of LDA:

- The data is normally distributed.
- Features are linearly separable.
- Homogeneity of variance-covariance.

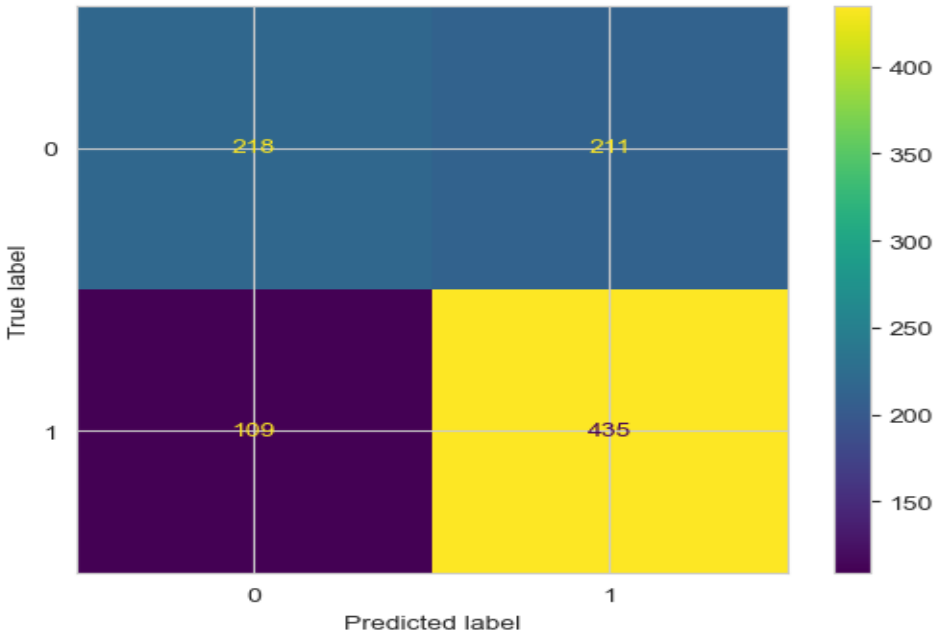
Steps to Build LDA Model:

1)Model building – Fit the model
2)Model evaluation - Confusion matrix, Classification report, F1 score, recall and precision score.
3)Model prediction – AUC-ROC curve.

Model Building steps:

1) Transform the model with still object categorical type into numerical.
2) Split the train and test data in the ratio of 70 and 30.
3) Scale the data using StandardScaler
4) Build a LDA Model and fit the data.
5) Predication on train and test data.

Confusion Matrix for train data.



Classification report for train data

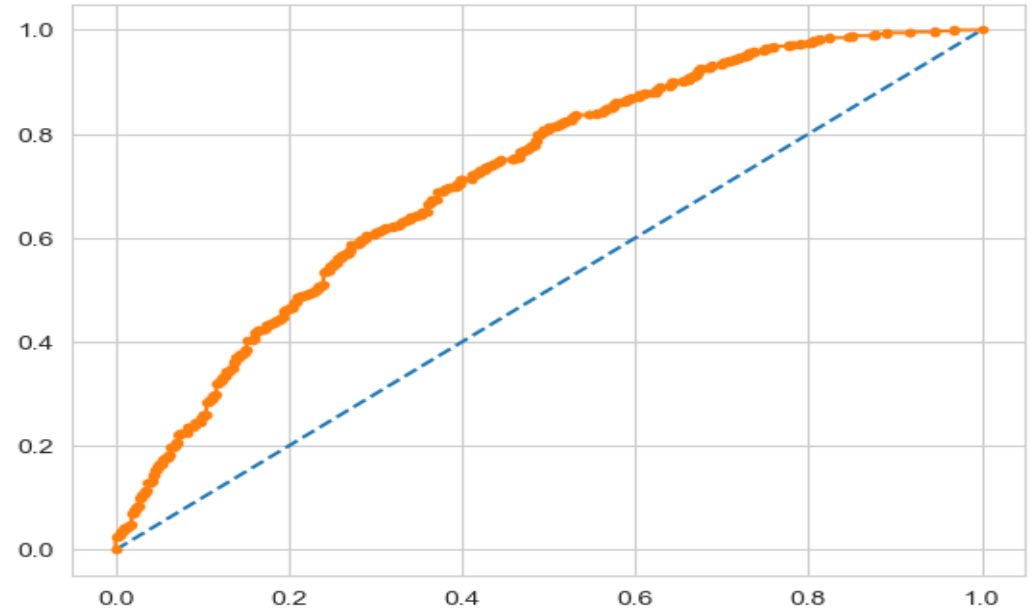
	precision	recall	f1-score	support

0	0.67	0.51	0.58	429
1	0.67	0.80	0.73	544
accuracy			0.67	973
macro avg	0.67	0.65	0.65	973
weighted avg	0.67	0.67	0.66	973

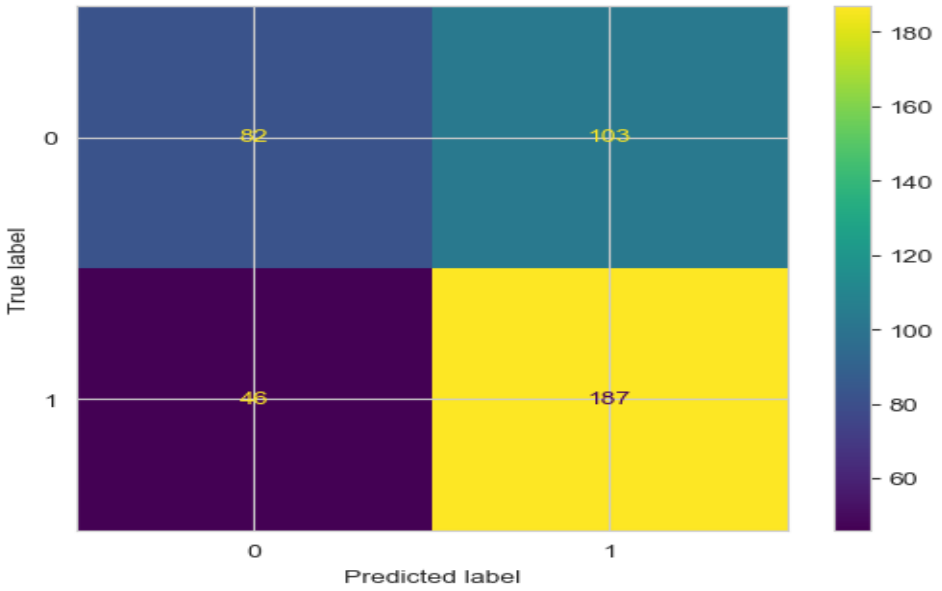
AUC for the Training Data:

0.716

[<matplotlib.lines.Line2D at 0x20671510150>]



Confusion Matrix for test data

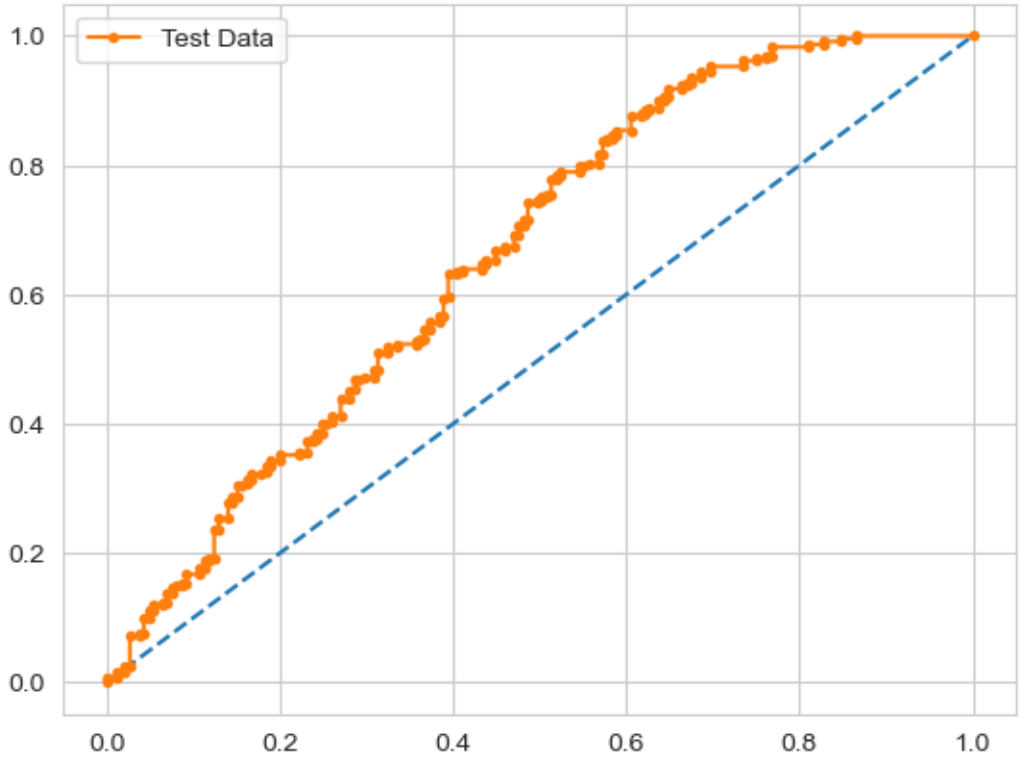


Classification report for test data

	precision	recall	f1-score	support
0	0.64	0.44	0.52	185
1	0.64	0.80	0.72	233
accuracy			0.64	418
macro avg	0.64	0.62	0.62	418
weighted avg	0.64	0.64	0.63	418

AUC-ROC for test data

9) AUC for the Test Data: 0.664



Generate the coefficients and LDA function

array([[-0.59, 0.52, 0.02, 0.74, -0.18, -0.08, 0.15, 0.31, 0.09]])

Index(['Wife_age', 'Wife_ education', 'Husband_education',
'No_of_children_born', 'Wife_religion', 'Wife_Working',
'Husband_Occupation', 'Standard_of_living_index', 'Media_exposure '],
dtype='object')

Observation and insights

By the above equation and the coefficients, it is clear that¶

1) No_of_children_born has the largest magnitude thus this helps in classifying the best.

2) Wife_age has the small magnitude thus this helps in classifying the least.

3) The Recall of class 1 is high 0.80 means the type 2 error is low.But for class 0 its low hence we can say the type 2 error exists.

4) The Precision for class 1 and 0 is 0.67 which means the ratio of contraceptive used by total of used and false predicted.

5) In this case the Type 1 error could be more looked into since if the number of false positives increases then chances are No of Childers born might increase due to non-usage of contraceptive.

Overview of Decision Tree

Decision tree is used for both classification and regression tasks. It splits the data into subsets based on the feature values, creating a tree-like.

1) Decision tree in Python can take only numerical / categorical columns. It cannot take string / object types. So, convert all objects into int8 and float8.
2) Split the data into train and test.

Model building

1) Create a object of DecisionTreeClassifier and fit the model using gini criteria.
2) After model is trained the decision tree looks with many branches since there were no limit set for leaf node.
3) Check on the important variable

	Imp
Wife_age	0.305948
No_of_children_born	0.261204
Wife_ education	0.114722
Standard_of_living_index	0.100121
Husband_education	0.061655
Husband_Occupation	0.054908
Wife_Working	0.049657
Wife_religion	0.041117
Media_exposure	0.010668

4) Regularising the Decision Tree with grid searchCV. The hyper tuning parameters are set to prune the tree.
    a. Parameters like:
        i. Max_depth
        ii. Min_samples
        iii. criteria
5) Generate the new decision tree and check on IMP variables.

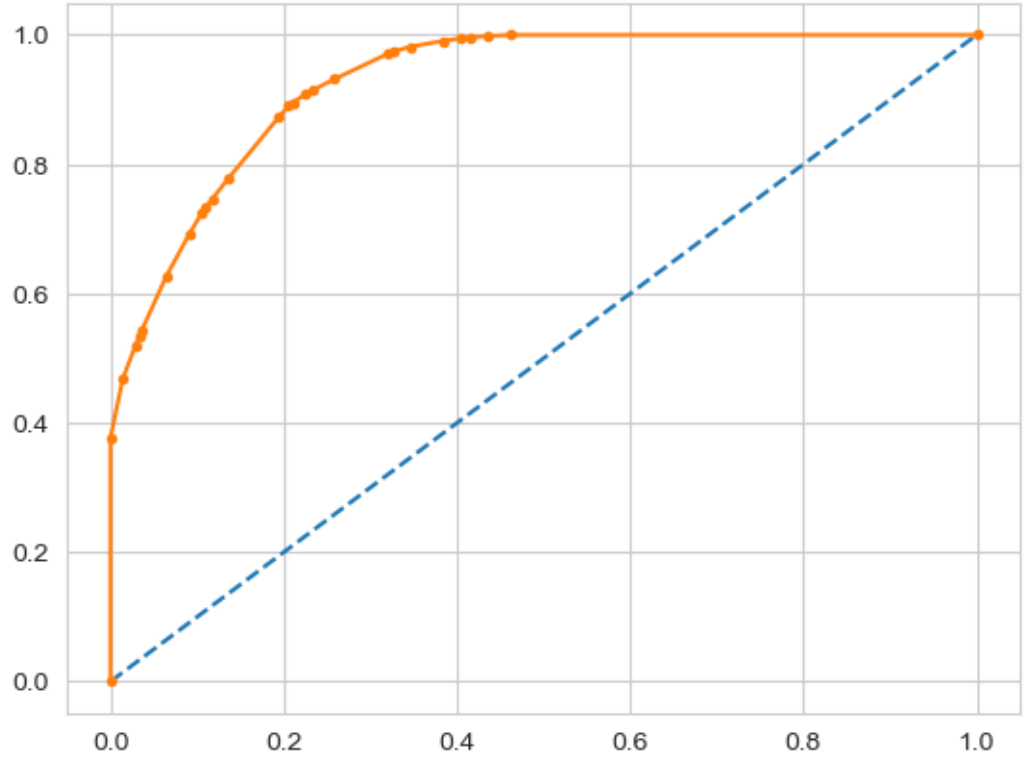
	Imp
No_of_children_born	0.354518
Wife_age	0.278700
Wife_ education	0.145261
Standard_of_living_index	0.074061
Wife_Working	0.040893
Husband_education	0.038280
Husband_Occupation	0.036408
Wife_religion	0.018281
Media_exposure	0.013598

6) Predicting on Training and test data.
7) Predicting the probabilities with proba function

Model Evaluation

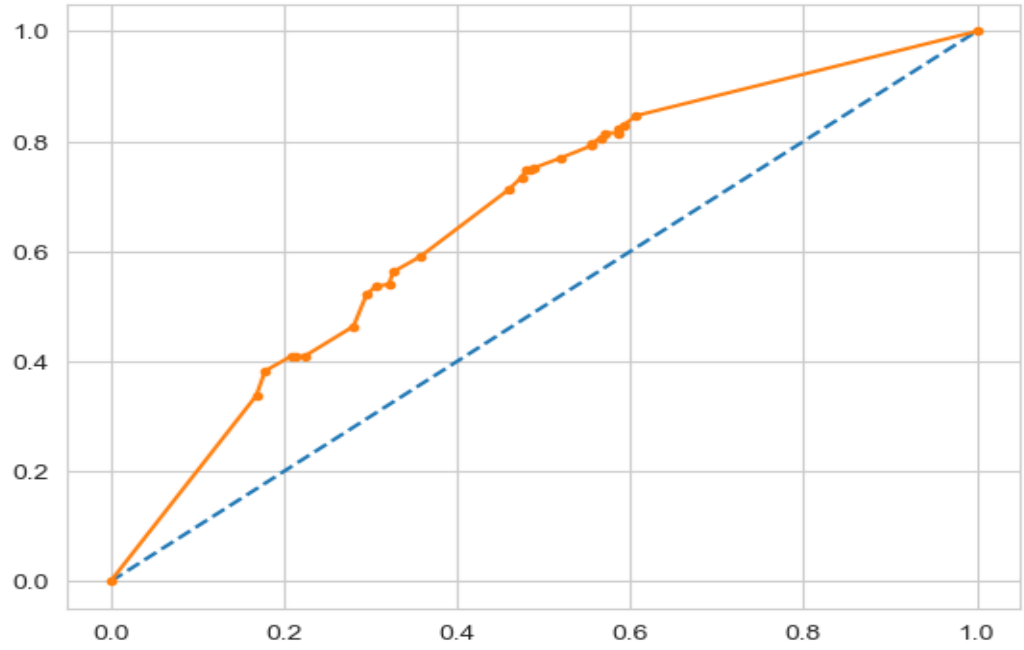
AUC and ROC Curve on training data.

AUC: 0.928

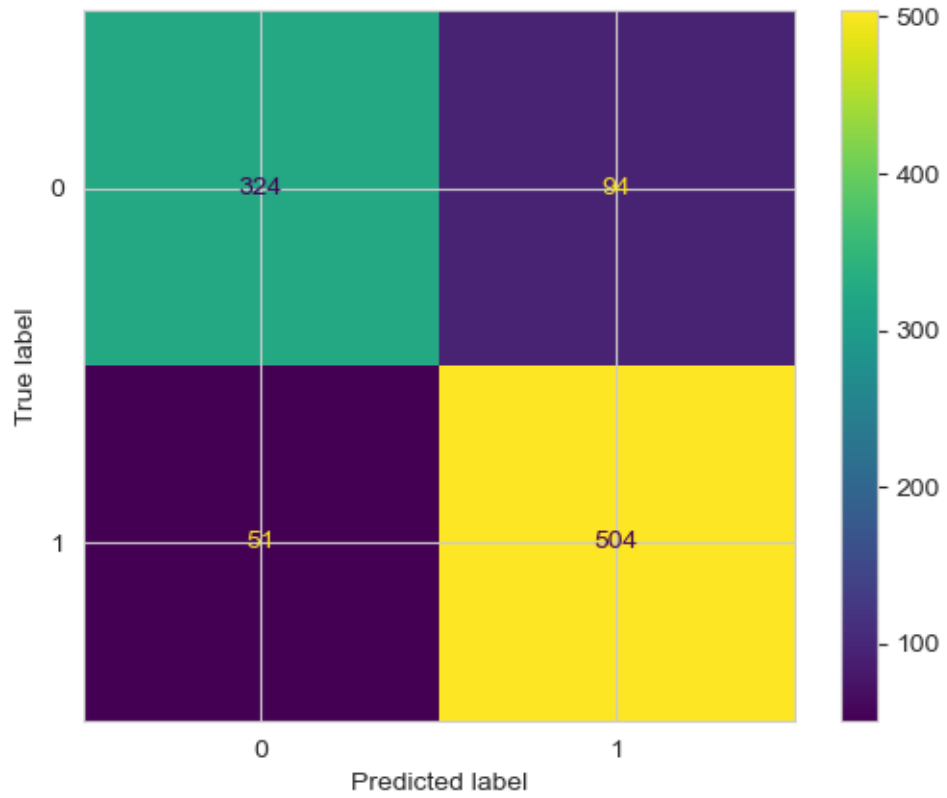


AUC and ROC Curve on test data.

AUC: 0.661



Confusion Matrix for the training data

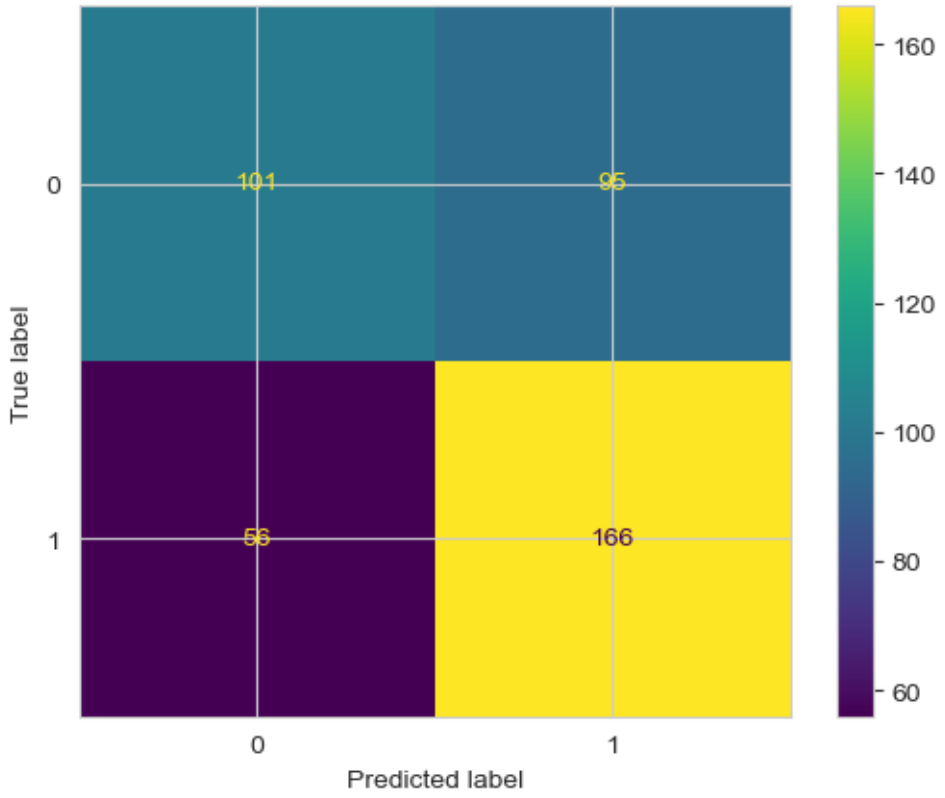


1) Data Accuracy for training is around 85%.

Classification report on training data

	precision	recall	f1-score	support
0	0.86	0.78	0.82	418
1	0.84	0.91	0.87	555
accuracy			0.85	973
macro avg	0.85	0.84	0.85	973
weighted avg	0.85	0.85	0.85	973

Confusion matrix on test data



2) The test data accuracy is around 64%.

The Classification report on test data

	precision	recall	f1-score	support
0	0.64	0.52	0.57	196
1	0.64	0.75	0.69	222
accuracy			0.64	418
macro avg	0.64	0.63	0.63	418
weighted avg	0.64	0.64	0.63	418

Actionable Insights and recommendations:

1) The train data has a accuracy of 85% whereas the test data as around 64% . The model is overfit. Since its not same.
2) The Precision is 86% and 84% for class 0 and 1. Which means the Type 1 error.
3) The recall is pretty high which indicates the type 2 error.
4) It falls into type1 error type.
5) 375 are class 0(no contraceptive used) and 598 are people where contrace ptive used.

6) I recommend to lower the overfit of the model using Lasso and ridge metho ds and then train the datasets.

Compare the Model

The Logistic regression and LDA model performed pretty same. The accuracy of Logistic regression is 70% and can be considered as moderately good model.

Based on AUC-ROC curve the graph is showing 70% for train and 64 for test. Minor difference in train and test data. Still can be considered best model.

The Type 1 error is significantly can affect in this problem. False positive can lead to a situation of not exposed to media and not knowing about the contraceptive and hence the population could increase.

The decision tree had overfitting issue. The tree was pruned for max-leaf and width to reduce the branches. The accuracy was around 92 in train data.

Based on analysis I would recommend Logistic regression as best fit.