**Predictive Learning Project**

**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.

DATA DICTIONARY:  
-----------------------  
System measures used:

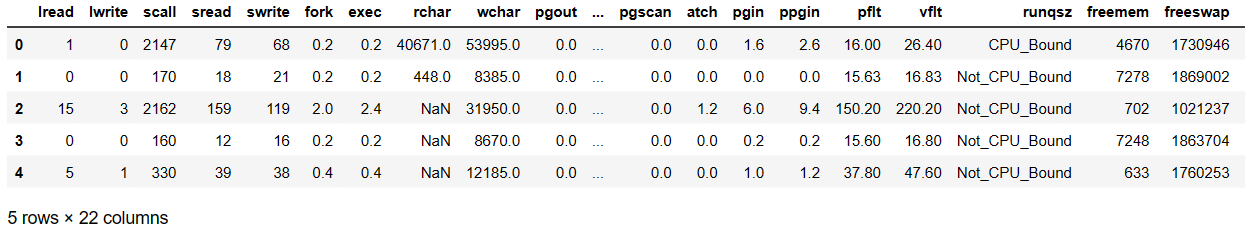
lread - Reads (transfers per second ) between system memory and user memory  
lwrite - writes (transfers per second) between system memory and user memory  
scall - Number of system calls of all types per second  
sread - Number of system read calls per second .  
swrite - Number of system write calls 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

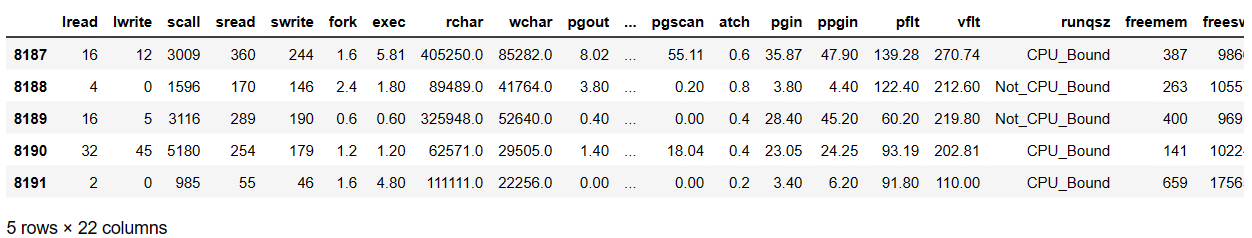
* 1. **Read the data and do exploratory data analysis. Describe the data briefly. (Check the Data types, shape, EDA, 5 point summary). Perform Univariate, Bivariate Analysis, Multivariate Analysis.**

**Ans:-** We have loaded the all the required packages and loaded compactiv Data file using Pandas.

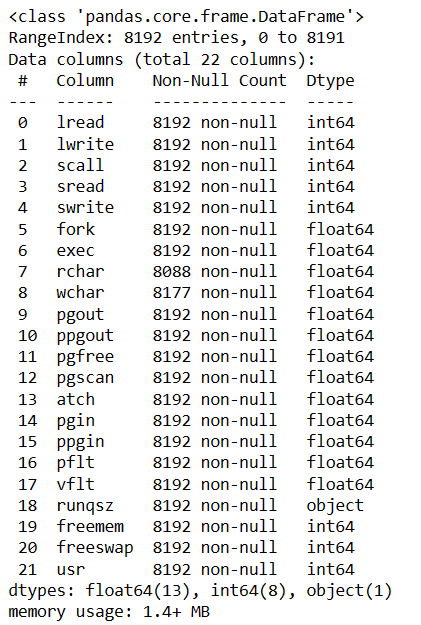
Dataset has 8192 rows and 22 columns.

We have viewed first and last few rows using head() and tail() functions respectively.





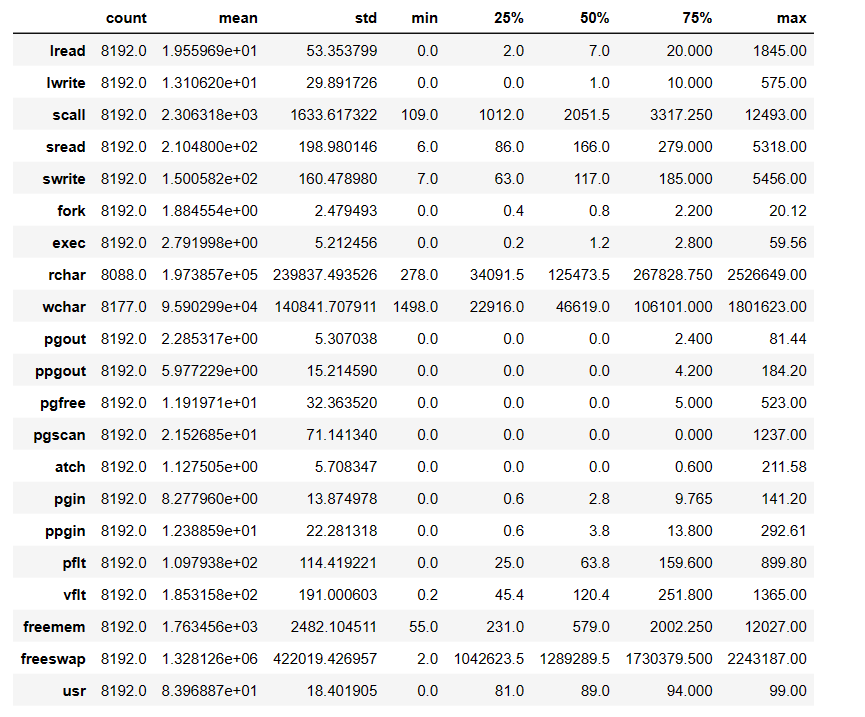
We can view dataset information using info()



There are 13 variables with float datatype,8 with integer data type and 1 with categorical datatype.

There are null values in rchar and wchar columns

Dataset can be described using describe() function.

****

Insights:-

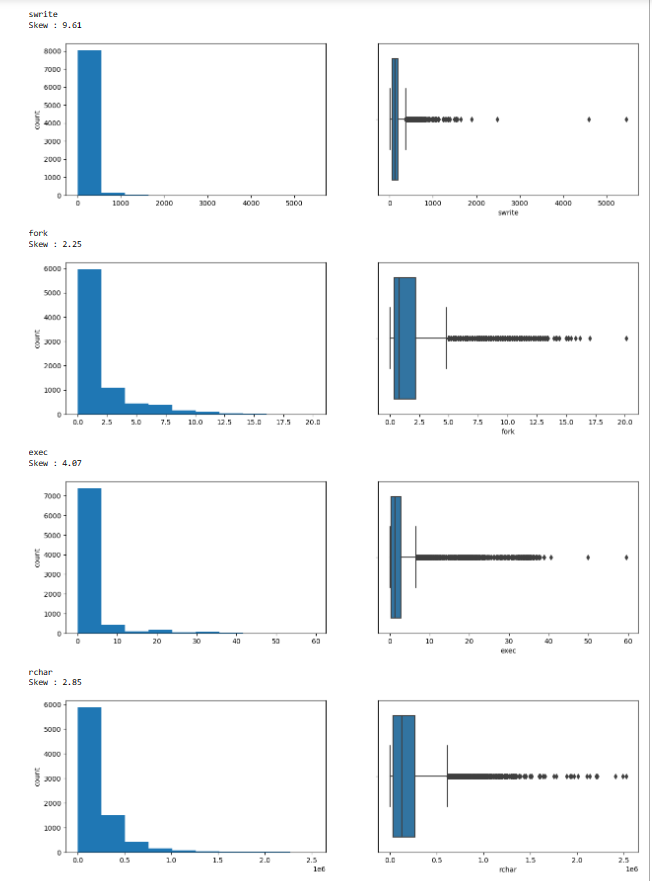
1. Mean,standard deviation, count etc can be seen using describe() function for all the numerical fields.
2. There are minimum 0 reads and write transferred between system memory and user’s memory per second. 1845 read are transferred at max and 575 write transferred at max between system memory and user’s memory per second.
3. Almost every other variable has 0 minimum except for a few.
4. CPU runs on an average of 8 times in user mode.
5. On an average Number of pages checked if they can be freed or not is around 2. At max they can be checked around 1237 times.
6. At max, number of memory pages available to user processes are around 12k and at max number of disk blocks available for page swapping is 99.
7. In this way by looking at description we can understand above information.

Let us plot graphs for All Numeric and Categoric variables individually which is known as Univariate Ananlysis.



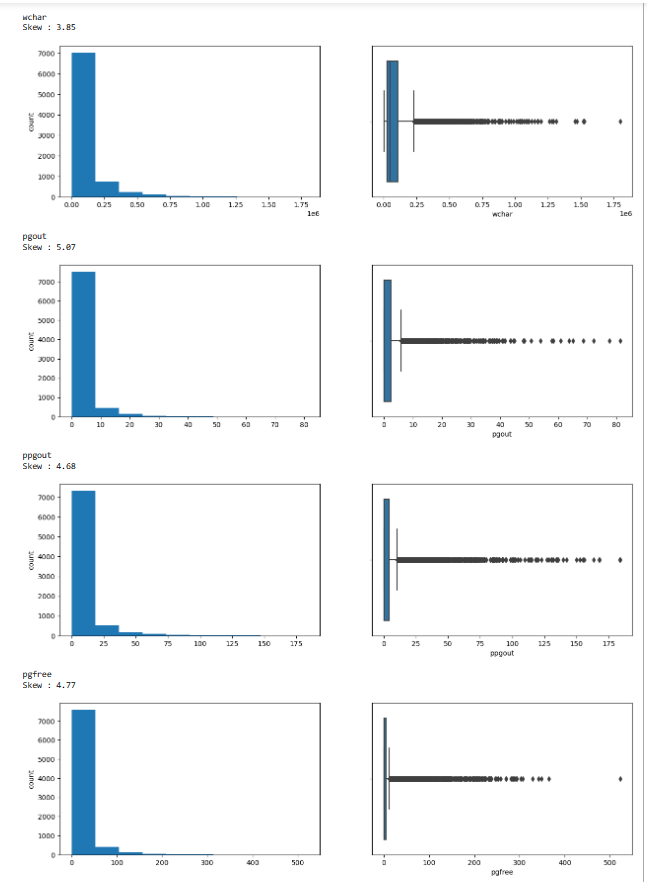
Insights:-

1. All of the variables above have skewed distribution.
2. All the variables are right skewed.
3. Iread is highly right skewed with skewness of 13.9.
4. All the variables have outliers.



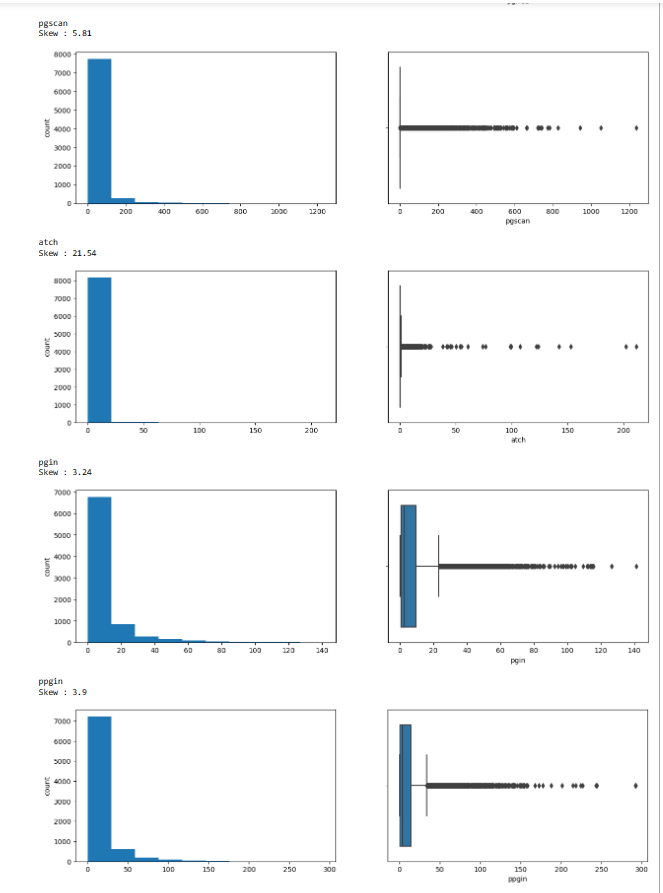
Insights:-

1. All 4 variables above have skewed distribution and they all are right skewed.
2. All the variables show outliers.



Insights:-

1. All the above 4 variables have skewed distribution.
2. All of them have right skewed distribution.
3. All of them show Outliers.



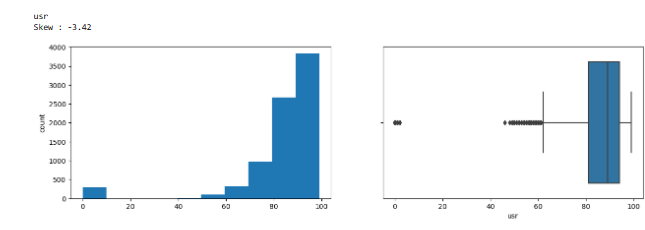
Insights:

1. All the above 4 variables have skewed distribution.
2. All of them have right skewed distribution.
3. All of them show Outliers.
4. Atch shows very high right skewed distribution.



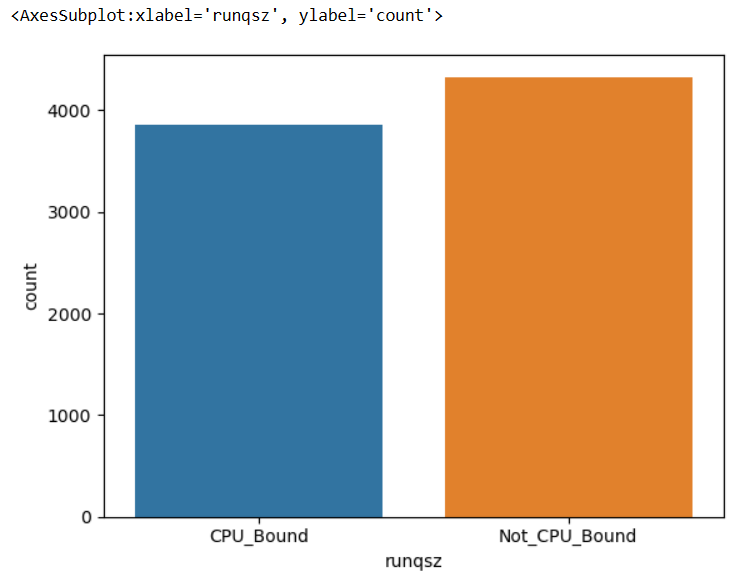
Insights:

1. All of the above 4 distributions have skewed distribution.
2. All the variables have right skewed distribution except for freeswap which has left skewed distribution with skewness of -0.79.
3. All of them show outliers.



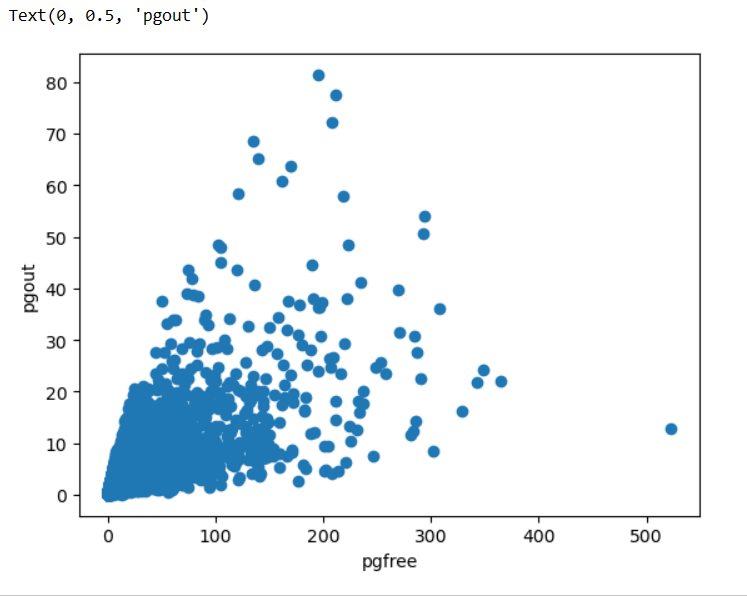
Usr has left skewed distribution with skewness of -3.42. It also shows outliers.

Univariate for Categorical variable

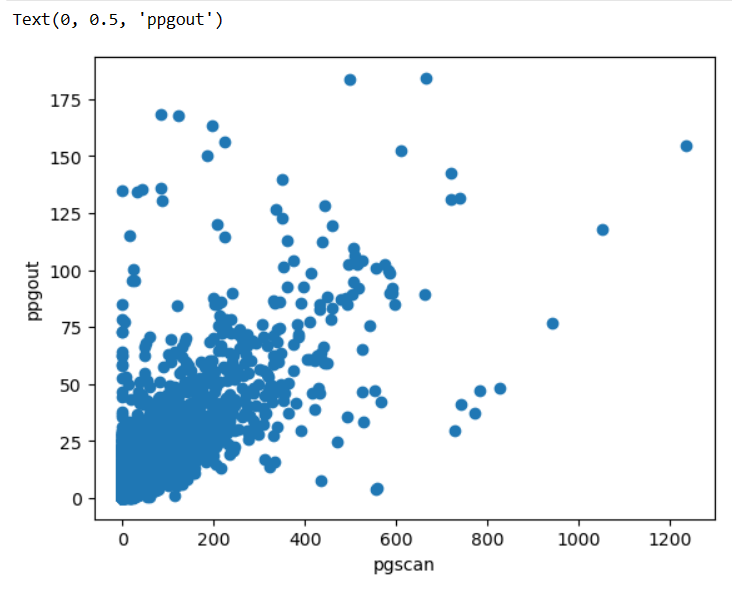


Runqsz is the number of kernel threads in memory that are waiting for a CPU to run.  
Typically, this value should be less than 2. Consistently higher values mean that the system might be CPU-bound. Process run queue size is not bound to CPU in most of the cases than CPU bound cases.

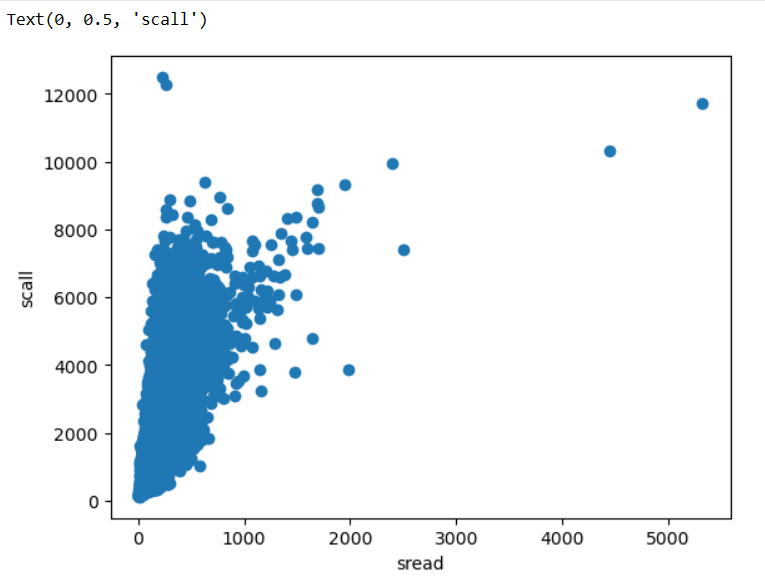
Bivariate Analysis



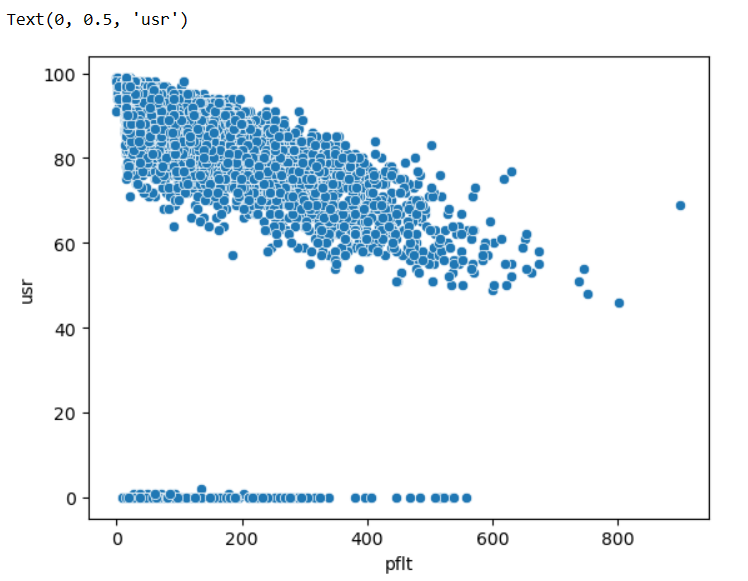
Number of page out requests per second is positively related to Number of pages per second placed on the free list.



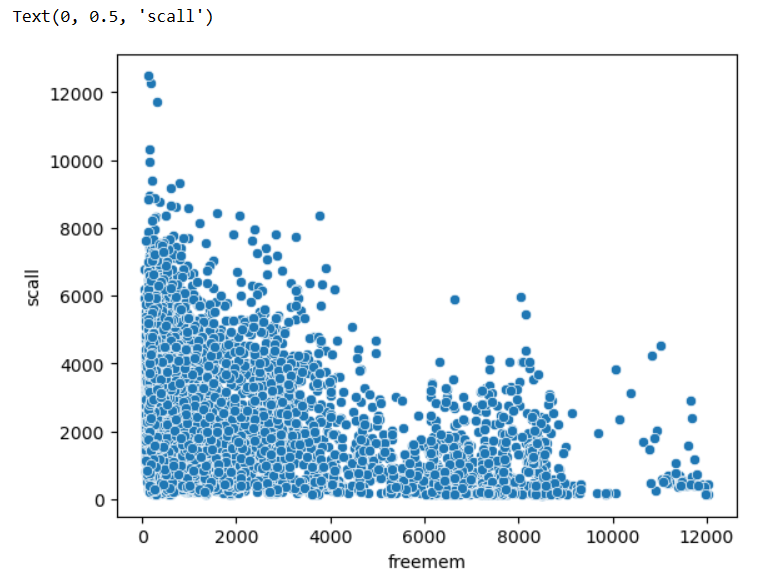
Number of pages paged out per seconds and Number of pages checked if they can be freed per second are positively correlated.



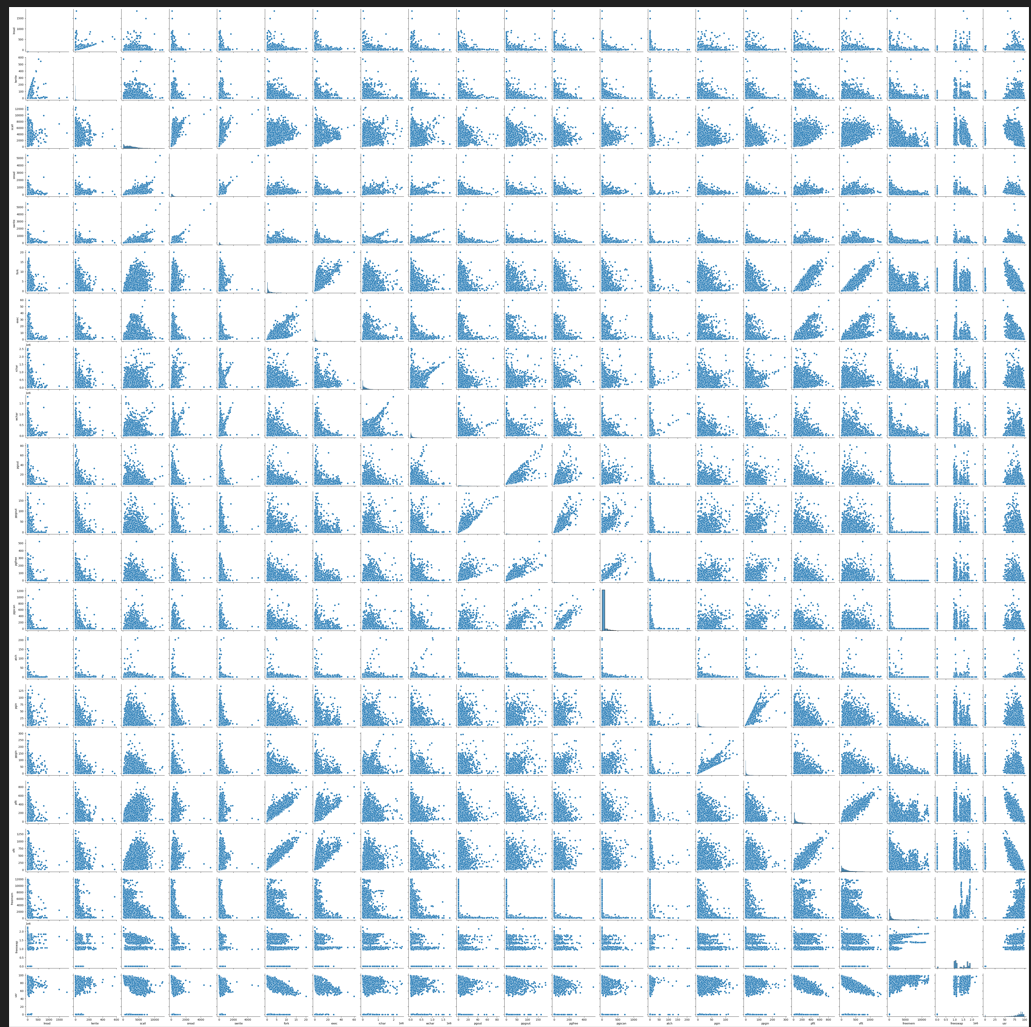
The more the number of system calls of all type per second ,the more will be system read calls per seconds and wise versa.

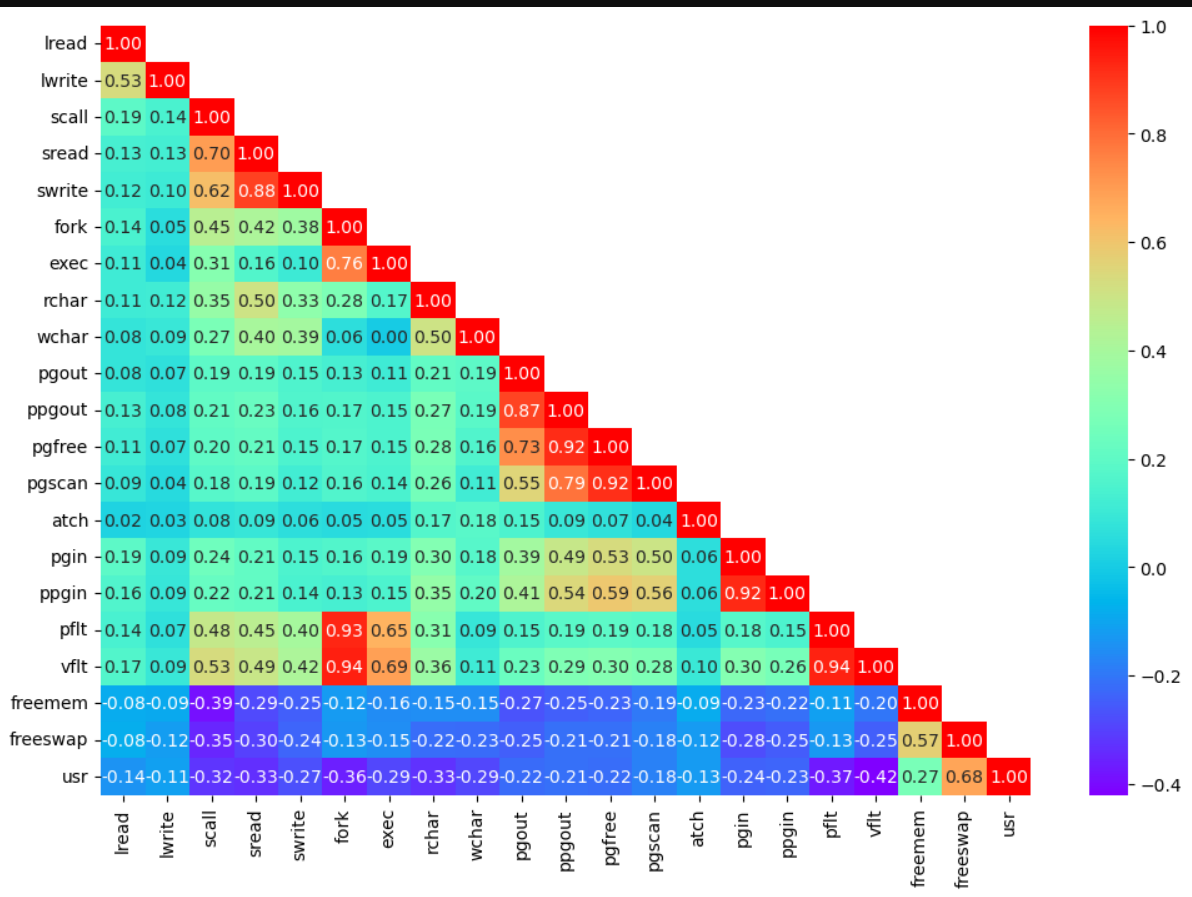


Number of page faults caused by protection errors (copy-on-writes) and Portion of time (%) that cpus run in user mode are negatively related.



Scall and Freemem are mildly positively related to each other.





Insights:-

1. There is high positive correlation between pgfree and ppgout and pgscan and pgfree resp.
2. Pflt and vflt are highly positively correlated with fork.
3. Pgin and ppgin are moderately positively correlated.
4. Usr and vflt are higly negatively correlated with each other means if Portion of time (%) that cpus run in user mode goes down that Number of page faults caused by address translation goes up.
5. Iwrite and Iread has moderate positive correltion.
   1. **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.**

**Ans:-**

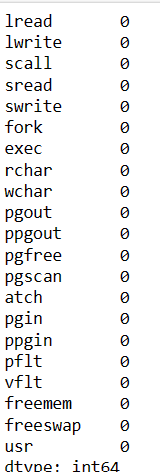
****

There are 104 null values in rchar and 15 null values in wchar.

Wchar and rchar have integer datatypes so we have replace null values in them with median value of each.

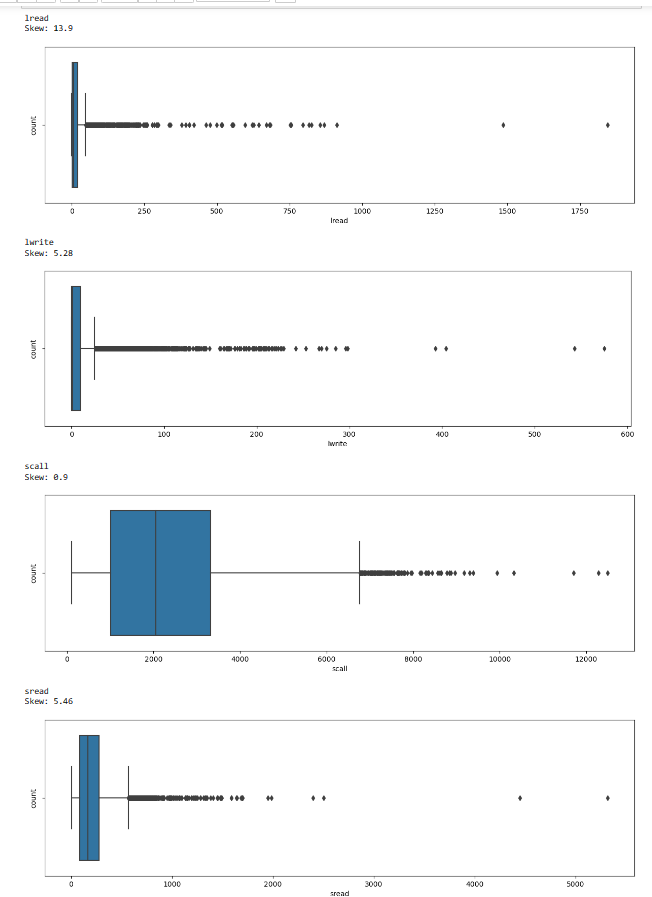
There are no duplicate values.

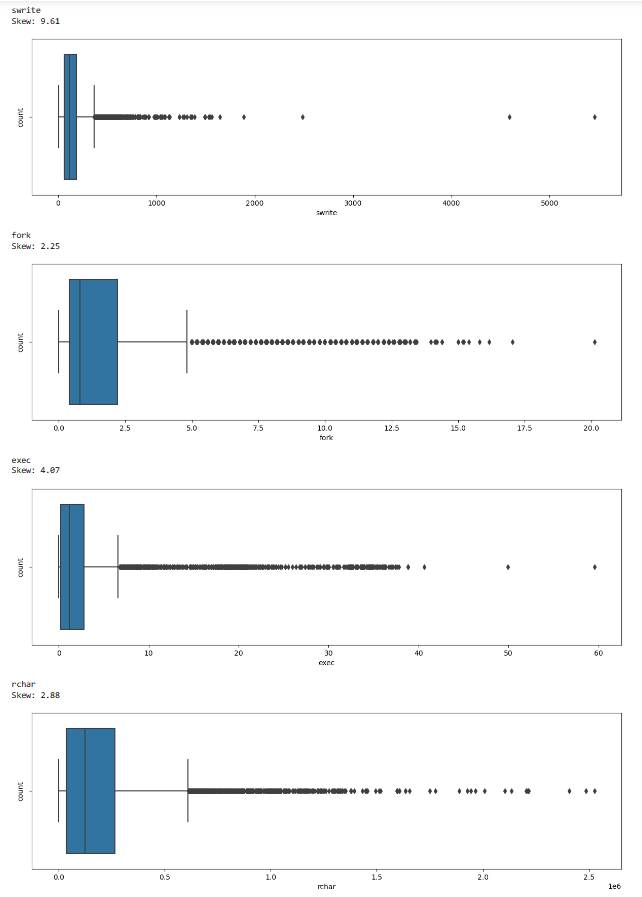
We don’t need to create any new features.

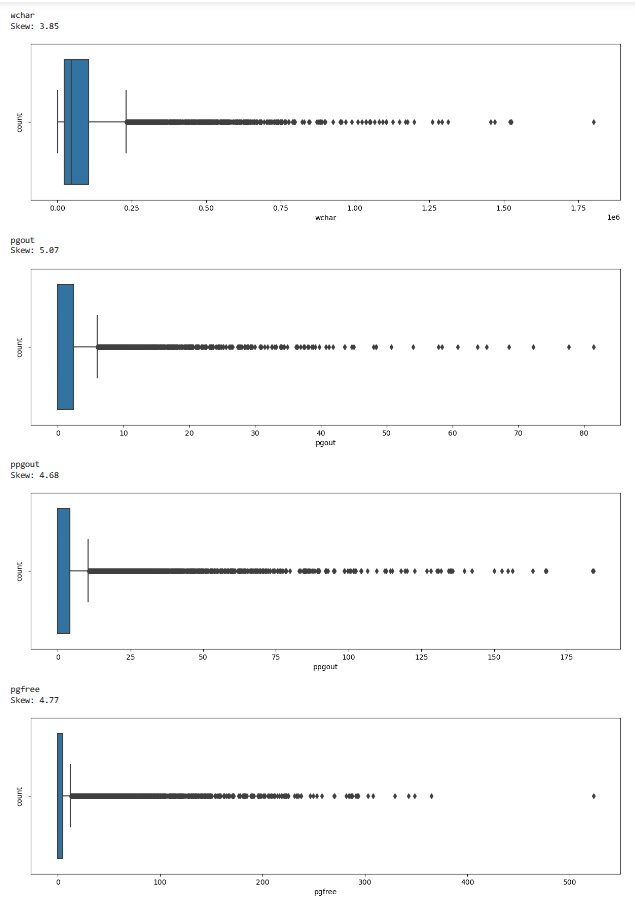


There are many entries in most of the variables with 0 as a value. We should not treat or drop them as they could be valid entries from computer.

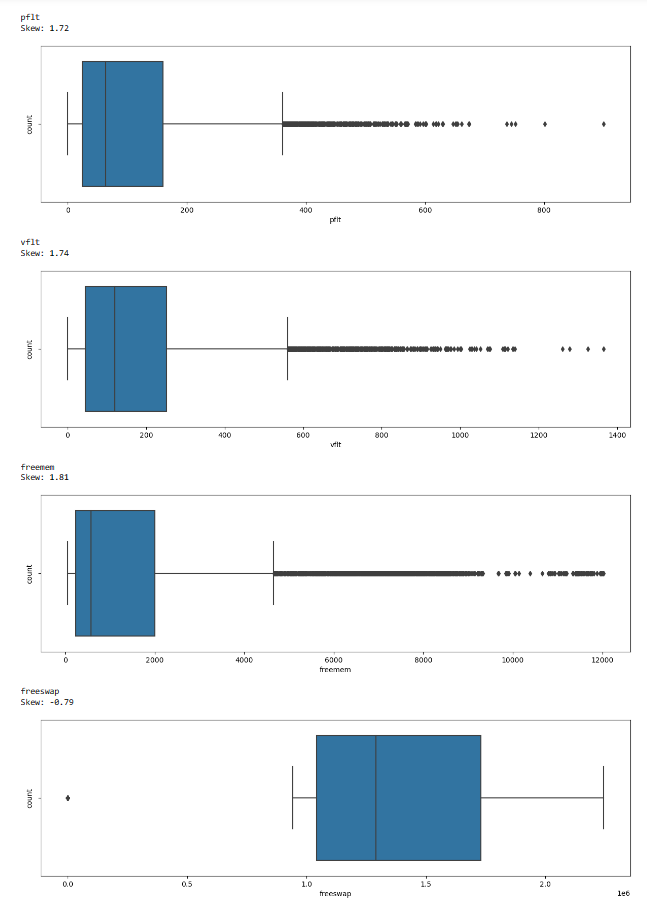
Checked for Outliers if any .

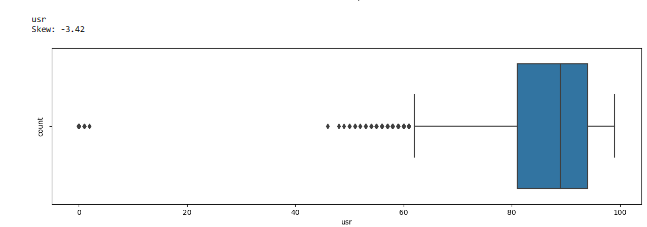








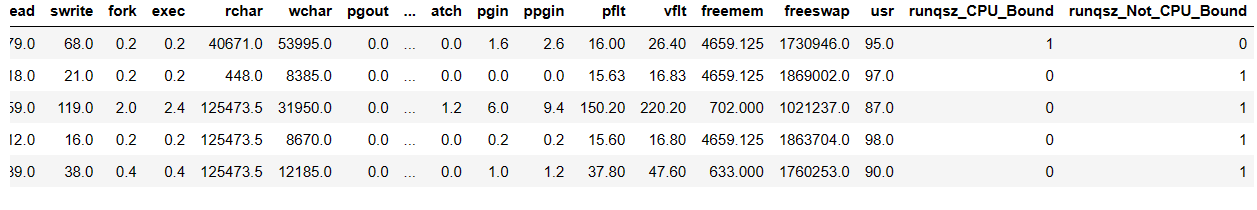




Insights:-

1. We can see that outliers are present on all the variables above.
2. Most of the variables have right skewed distribution except for usr and freeswap which show left skewed distribution.
3. We have treated Outliers using IQR method by replacing the ourliers with lower values and upper values depending upon if the distribution is left skewed or right skewed.
4. Now there are no outliers present in the data.
   1. **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.**

**Ans:-** We have performed one hot encoding method on categorical variable runqsz.

Runqsz is divided into two columns:- 1st is runqsz\_CPU\_Bound and runqsz\_Not\_CPU\_Bound.

True value in each newly created column are given 1 and False values are given 0.

1. We have copied predictor variables into X dataframe and target variable into y dataframe.
2. Data was then split into 70:30 ration into training and testing set.
3. Training set of predictor variables have 5734 rows and 21 columns and testing set of predictor variables have 2458 rows and 21 columns.
4. Training and testing set of tagert variable has 1column and same number of rows as training and testing set of predictor variables.
5. **USING LINEAR MODEL FROM SCI-KIT LEARN LIBRARY**
6. We have built Linear Regression model and fitted training set data into it.
7. Coefficients for each of the independent attributes in training and testing data was calculated.
8. 
9. We can see that freeswap and runqzs\_Not\_CPU\_Bound can directly impact the values of target variables as they have high positive values of coefficients.
10. The intercept for our model is 84.92938971974552.
11. We can check performance of the Model by calculating R square, RMSE values for both training and test data.
12. For training data





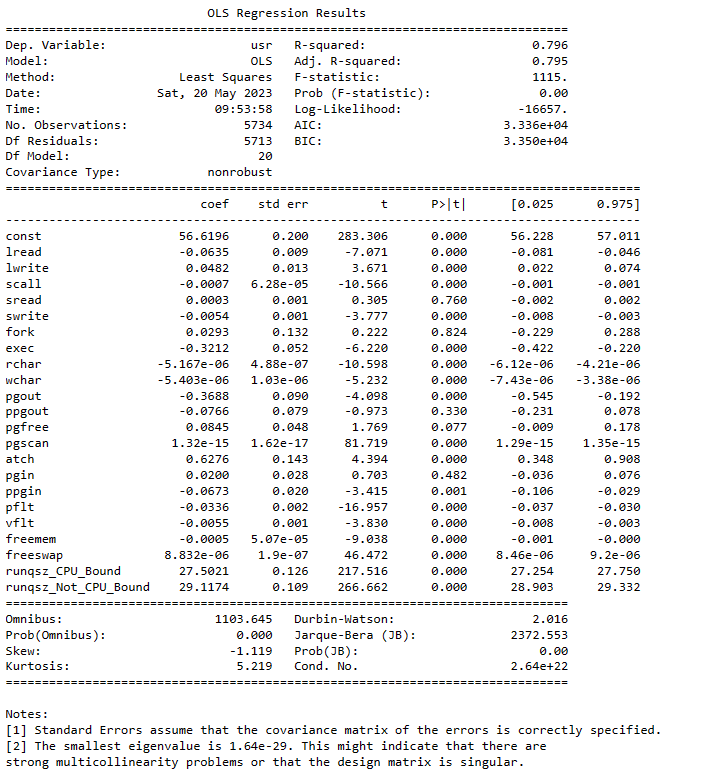
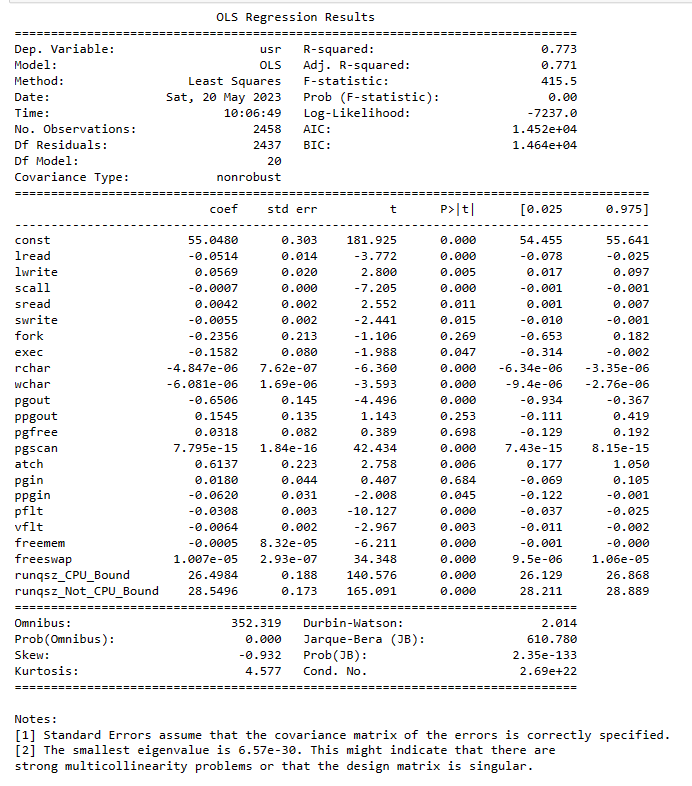
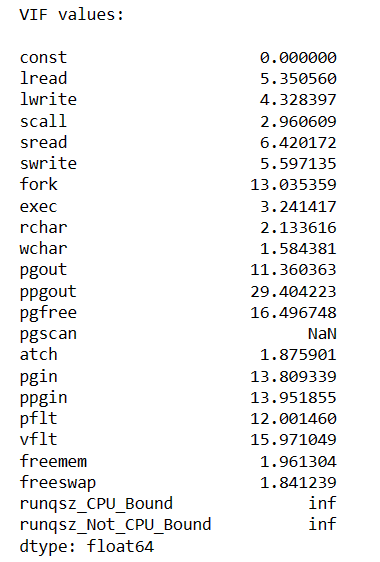
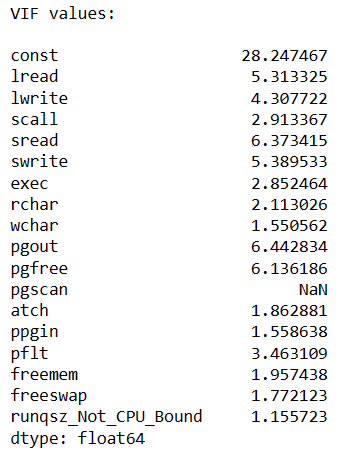
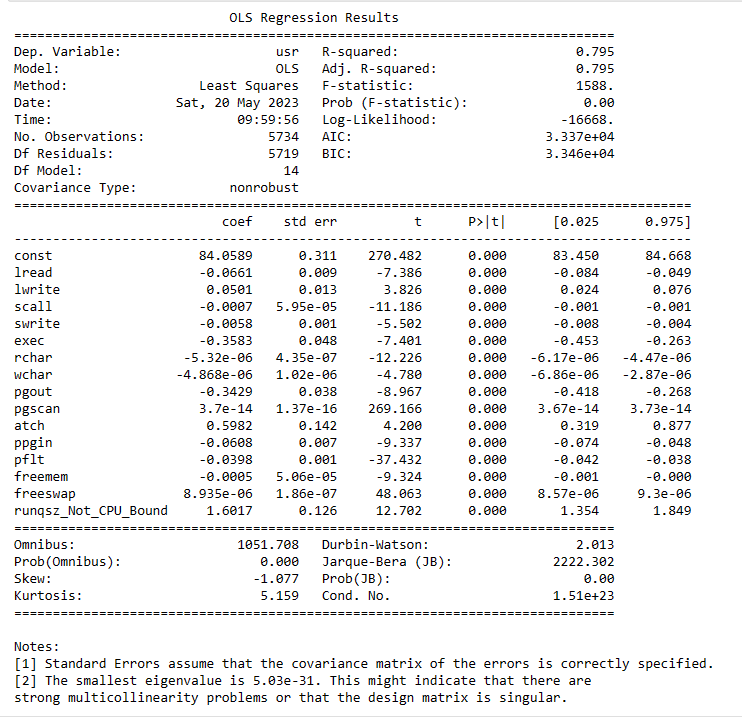


1. This means that 79% of the variation in the usr is explained by the predictors in the model for train set.
2. For testing data

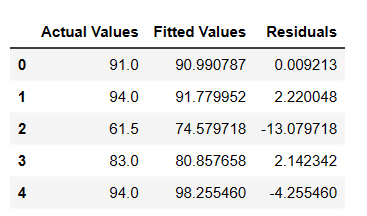






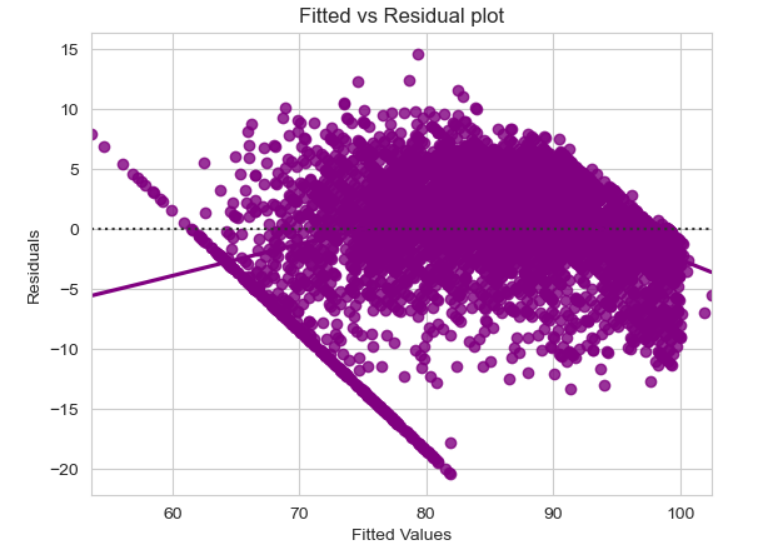
1. This means that 76% of the variation in the usr is explained by the predictors in the model for the test set.
2. We can say that this model is neither overfitted nor underfitted as there is not much difference in R square value.
3. **USING STATSMODEL OLS**
4. Lets build OS statsmodel and then fit training dataset into it.
5. For training data
6. 
7. For testing data
8. 
9. We can see that for training set
10. R square is 0.796 and Adj R square is 0.795
11. For testing set
12. R square is 0.773 and Adj R square is 0.771.
13. Lets check for multicollinearity in the training dataset using VIF method.
14. 
15. We can see that VIF value for ppgout is the highest and there are several other variables which have VIF values greater than 10. Multicollinearity can affect performance of the model.
16. lets remove multicollinear columns one by one and observe effect on our predictive model
17. In the end we have removed runqsz\_CPU\_Bound, ppgout, vflt, pgin, fork as dropping these variables were not affecting R square and Adj R square value.
18. We have used VIF method to drop these values and checked for OLS summary after dropping each variables one by one.
19. 
20. Now there are no VIF values which are affecting performance of the model.
21. sread and pgfree has p-value greater than 0.05,so we can drop it because it is not significant in predicting usr.
22. After dropping all the multicollinear variables OLS regression result looks like below.
23. 
24. There is not much change in R square and Adj R square.

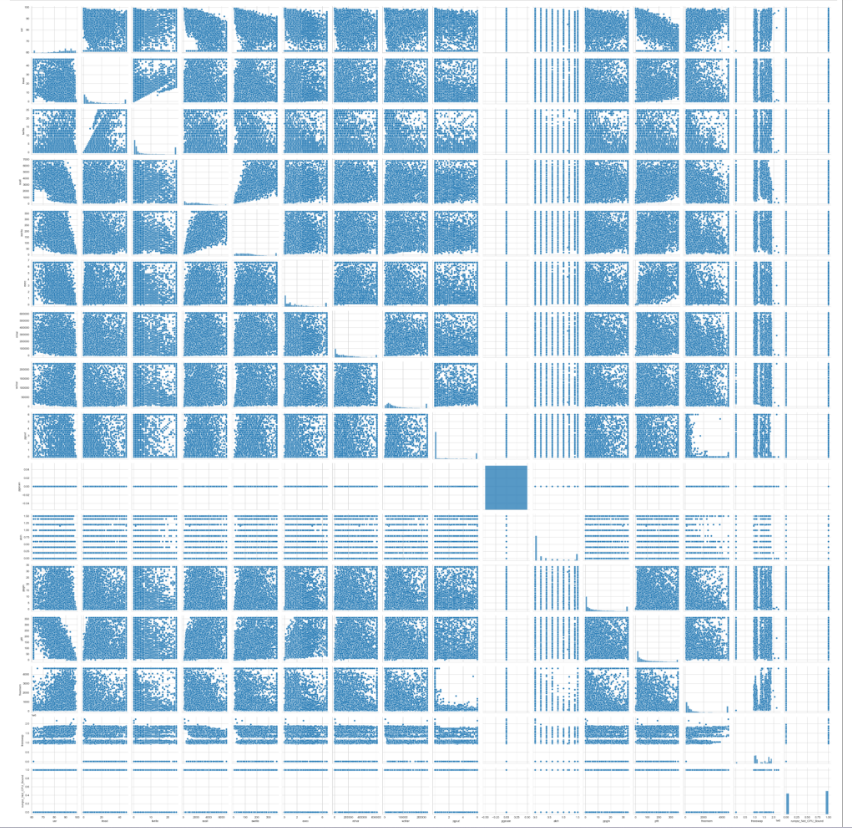
Assumptions of Linear Regression



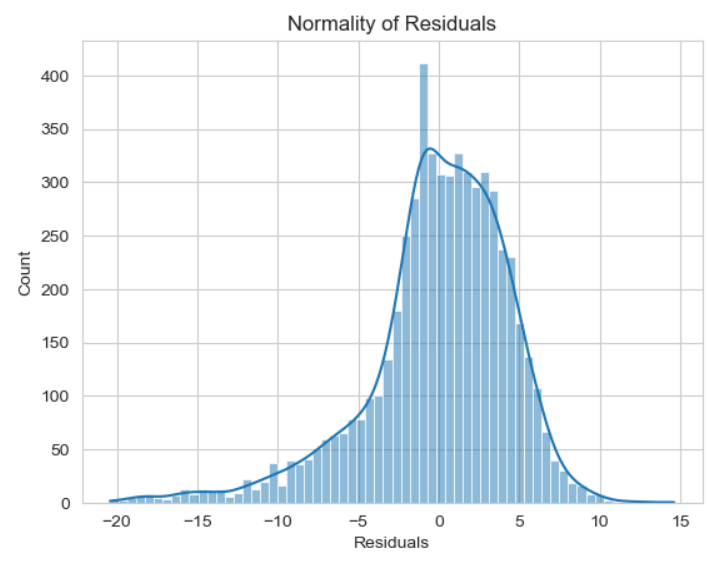
Actual values,Fitted values and Residuals look like this.

Lets plot fitted values vs Residual values

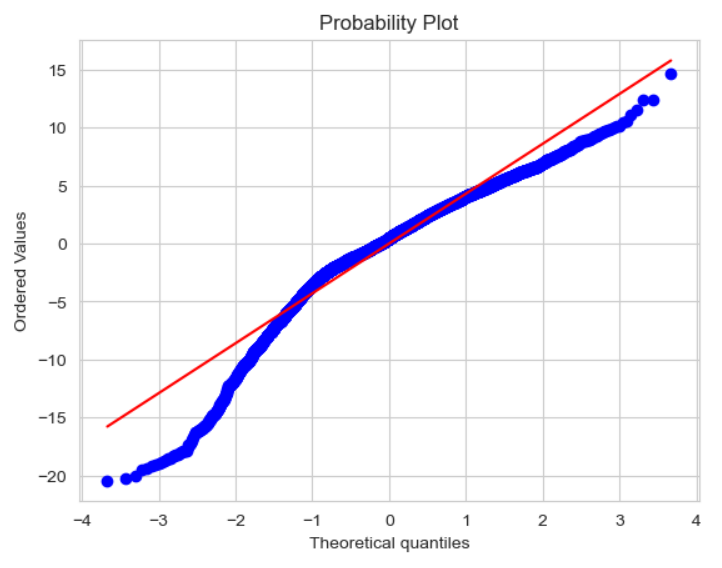
1)

2) 

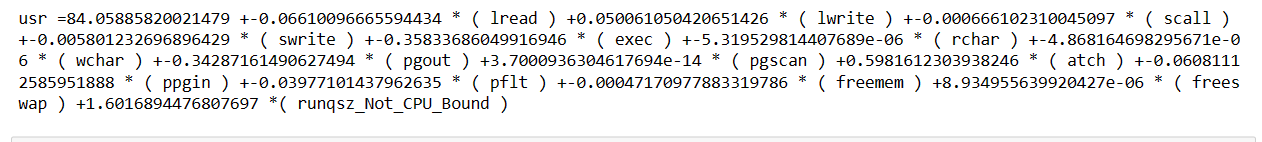
Above pairplot is plotted using preditore variables with no multicollinearity.

3)

Above graph shows that residuals shows almost normal distribution but perfect.

1. 
2. 
3. since p\_value >0.05, we can say that residuals are homoscedastic.
4. So now we have removed all the multicollinear columns from test data.
5. 
6. 
7. 
8. 
9. There is not much change in R square value after removing multicollinearity.
10. We can say Sci-Kit learn library performs better than OLS model.
    1. **Inference: Basis on these predictions, what are the business insights and recommendations.Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.**

**Ans:-**

1. We have loaded the dataset after importing certain required packages.
2. There are 8192 rows and 22 columns.
3. There are 13 variables with float data type,8 with integer datatype and 1 categorical datatype.
4. We have plotted the data using various plotting techniques.
5. There were no duplicate values in the data
6. There are few null values in rchar and wchar which were replaced by median of each columns.
7. There were Outliers in the data which were treated using IQR method.
8. We have used two methods to check performance of linear regression model.
9. Dataset with predictor variables was copied in X and target variable was copied in Y dataframe.
10. Data was splits into training and testing data in 70:30 ratio resp.
11. 1st method was Linear model using sci kit learn library.
12. We have fitted data into this model and calculated R square and RMSE for both testing and training data.
13. 79% variance was calculated by training data and 77% variance was calculated using testing data.
14. This model was neither overfitting nor underfiiting.
15. 2nd model was OLS model.
16. Data was fiited into OLS model.
17. We have seen that there is multicollinearity in the data.
18. We have used VIF method to remove multicollinearity in the data.
19. Data with p-score greater than 0.05 was removed.
20. There was not much change in the R sqr and Adj R sqr values before and after removing multicollinearity.
21. Linear equation for above model is
22. 
23. We have also checked for assumptions of linear regression.
24. Usr is highly affected by freeswap and pgscan.
25. All the other variables does not affect the usr value.

**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:**

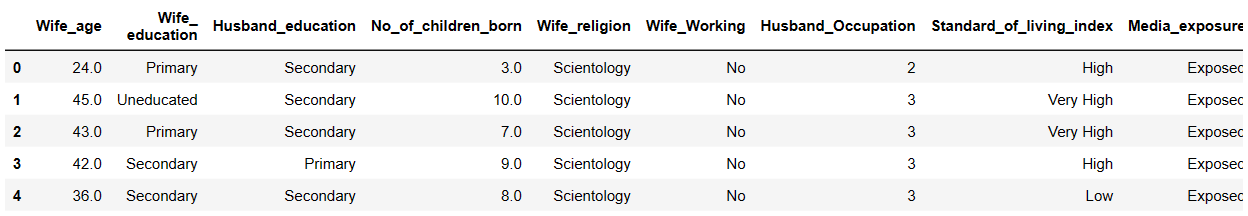
1. Wife's age (numerical)  
2. Wife's education (categorical) 1=uneducated, 2, 3, 4=tertiary  
3. Husband's education (categorical) 1=uneducated, 2, 3, 4=tertiary  
4. Number of children ever born (numerical)  
5. Wife's religion (binary) Non-Scientology, Scientology  
6. Wife's now working? (binary) 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

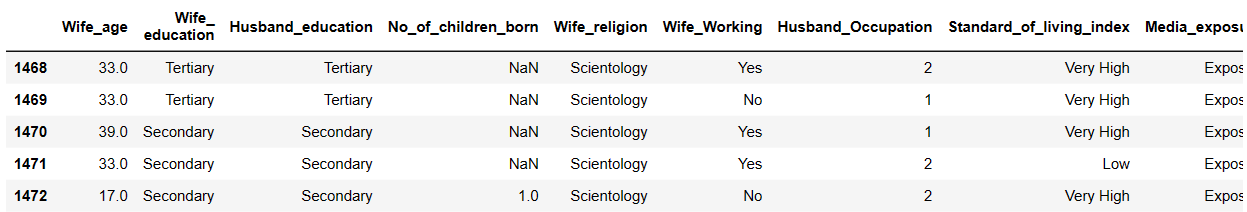
**2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, check for duplicates and outliers and write an inference on it. Perform Univariate and Bivariate Analysis and Multivariate Analysis.**

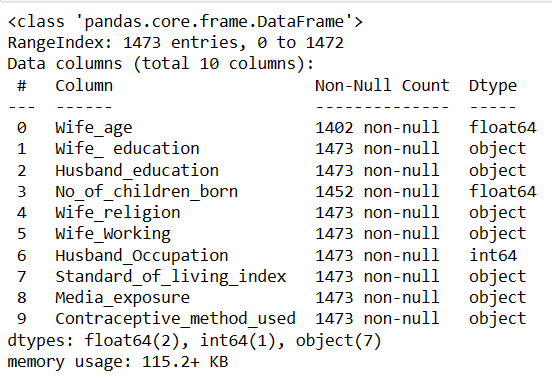
**Ans:-** We have loaded the all the required packages and loaded Contraceptive\_method\_dataset Data file using Pandas.

Dataset has 1473 rows and 10 columns.

We have viewed first and last few rows using head() and tail() functions respectively.



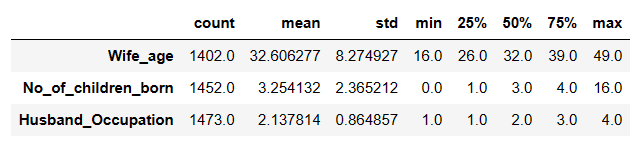




There are few null values in Wife\_age and No\_of\_children\_born.

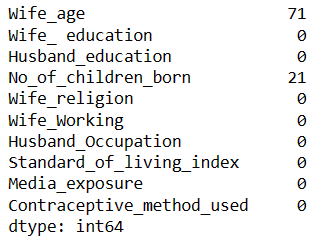
There are 2 variables with float datatype, 1with integer and rest other are with object datatype.

We can describe numerical data using describe() function.



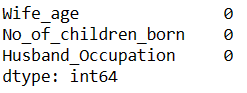
Insights:-

1. Maximum value for wife\_age is 49 and minimum is 16.
2. On an average 3 children are born. Minimum 0 and maximum 16 children are born.
3. Husband has maximum 4 occupations and minimum 1.



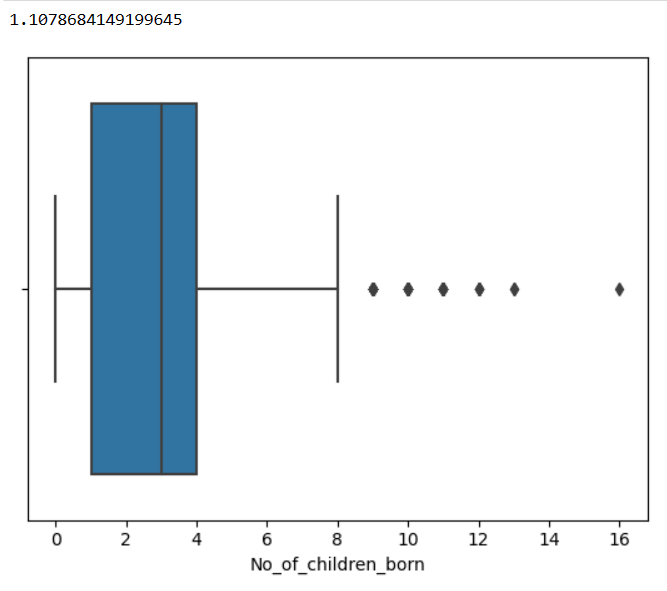
There are 71 null values in Wife\_age column and 21 in No\_of\_children\_born.

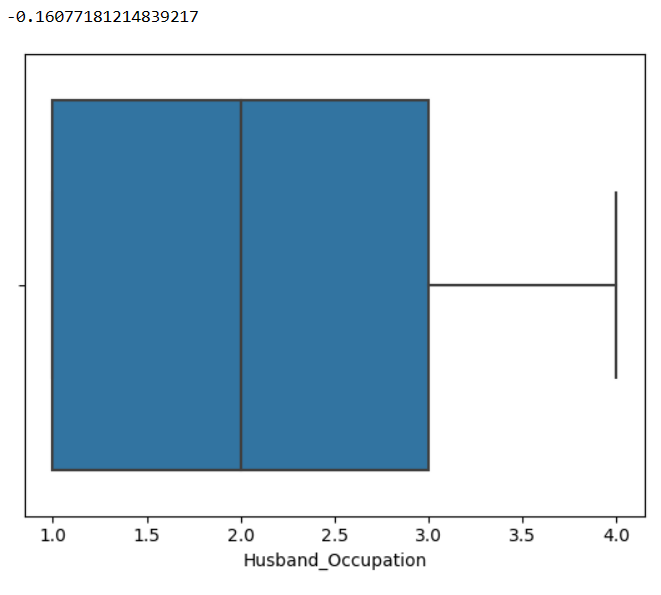
We have treated this null values using median() of each column.

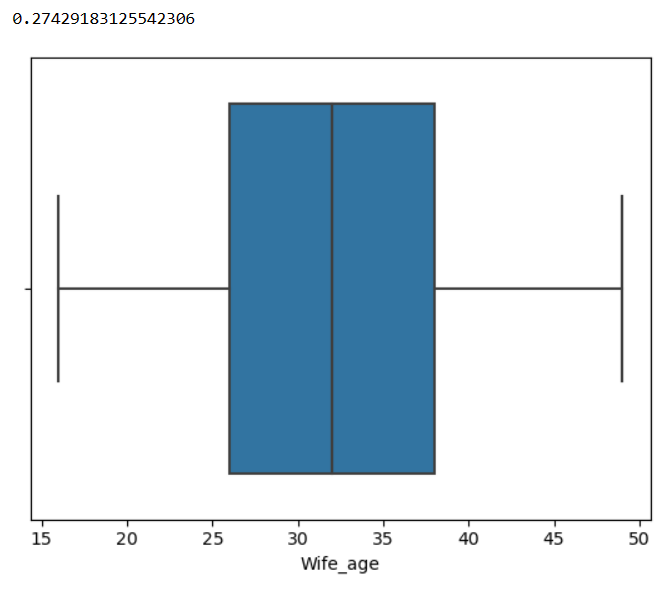


Lets visualize data using various visualization techniques.

1. Numeric variables

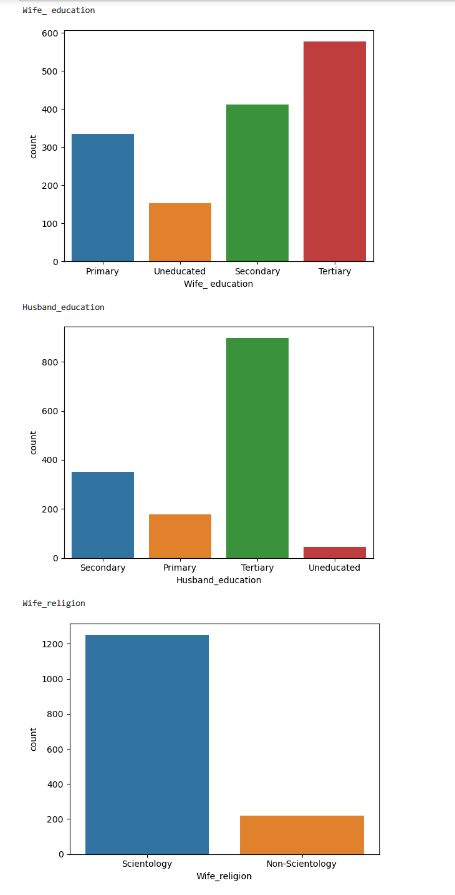






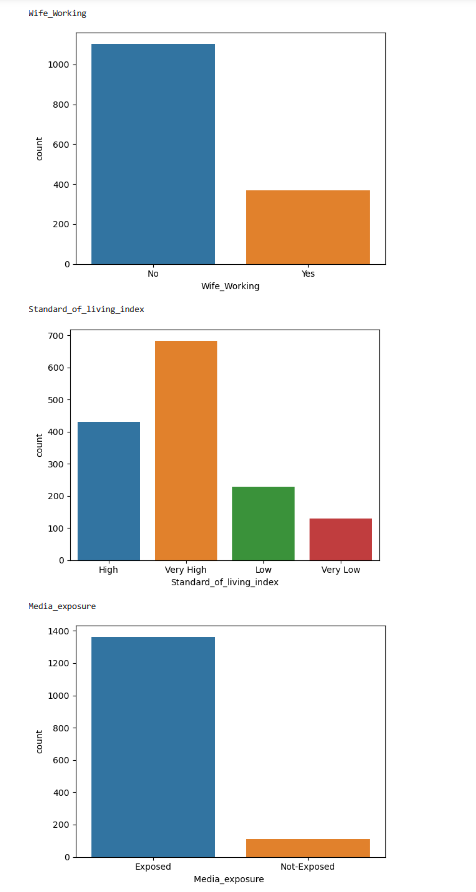
Insights:-

1. All the observation above have skewed distribution.
2. All the observations are right skewed except for Husband\_occupation which is left skewed.
3. Only No\_Of\_Child\_born have outliers.
4. Catergorical variables.



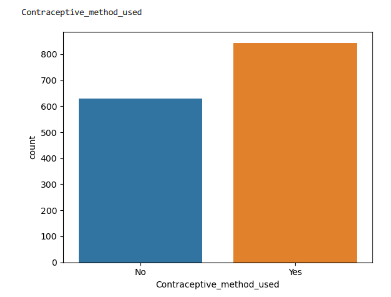
Insights:-

1. Most of the Wives are educated. Mostly wives have tertiary education followed by secondary, primary and uneducated.
2. Most of the Husbands also are educated. Mostly husbands have tertiary education followed by secondary, primary and uneducated.
3. Most wives believe in Scientology .



Insights:-

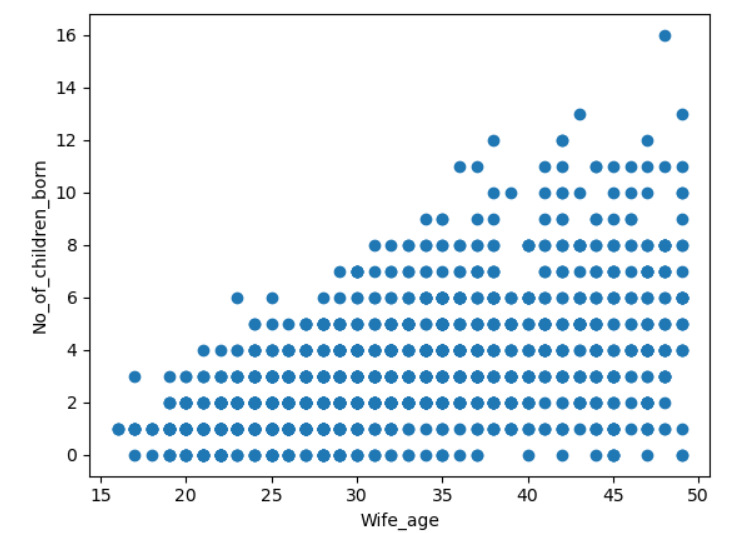
1. Most of the wives aree non working
2. Standard of living is very high followed by High,Low and very low.
3. Media Exposure is too much than non exposed.



1. Most of the wives use Contraceptive method.

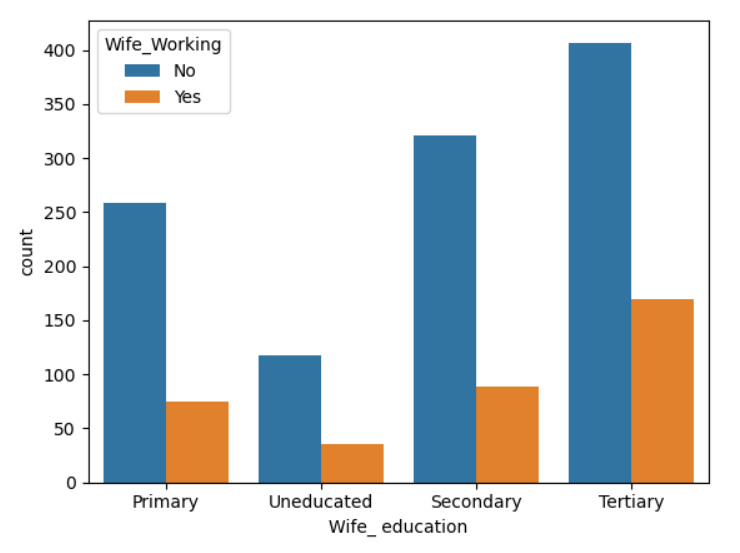
Bivariate

1. Numeric vs Numeric

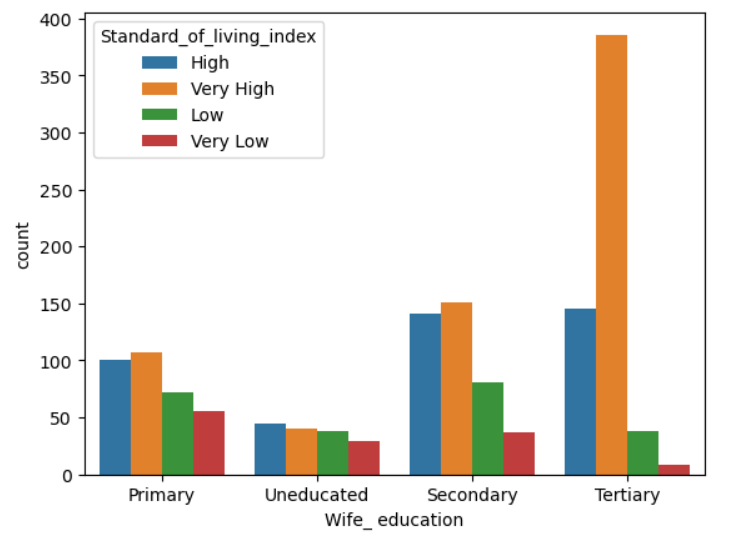


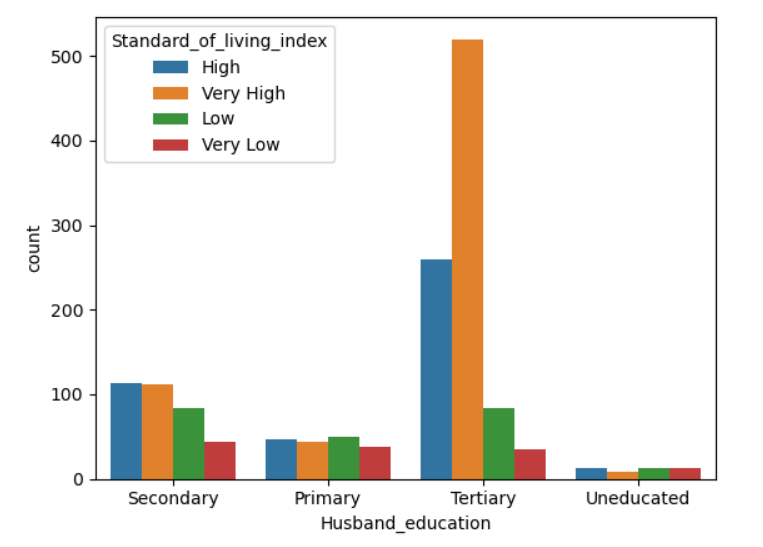
Wife\_age and No\_of\_child\_born are moderately related to each other.

1. Catergoric vs Categoric

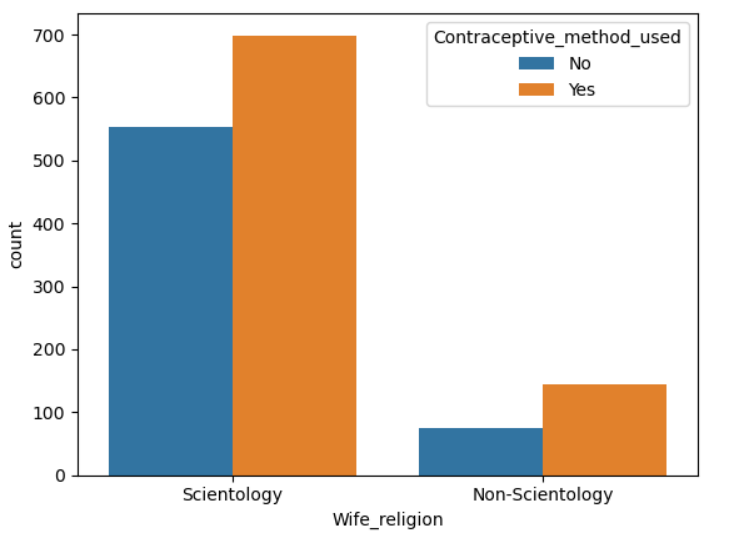


Above graph shows that wife with tertiary education are mosty working followed by secondary,Primary and Uneducated.

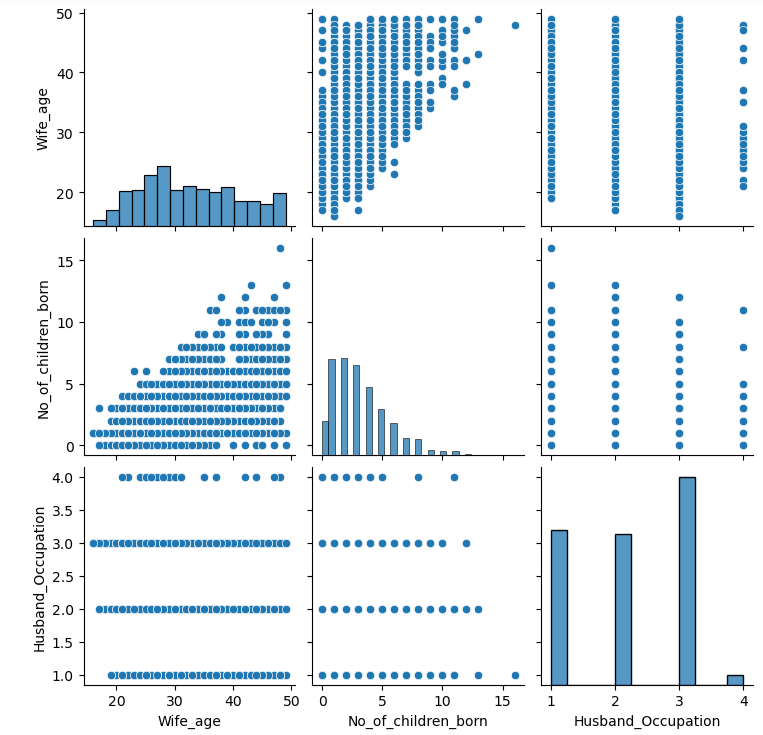


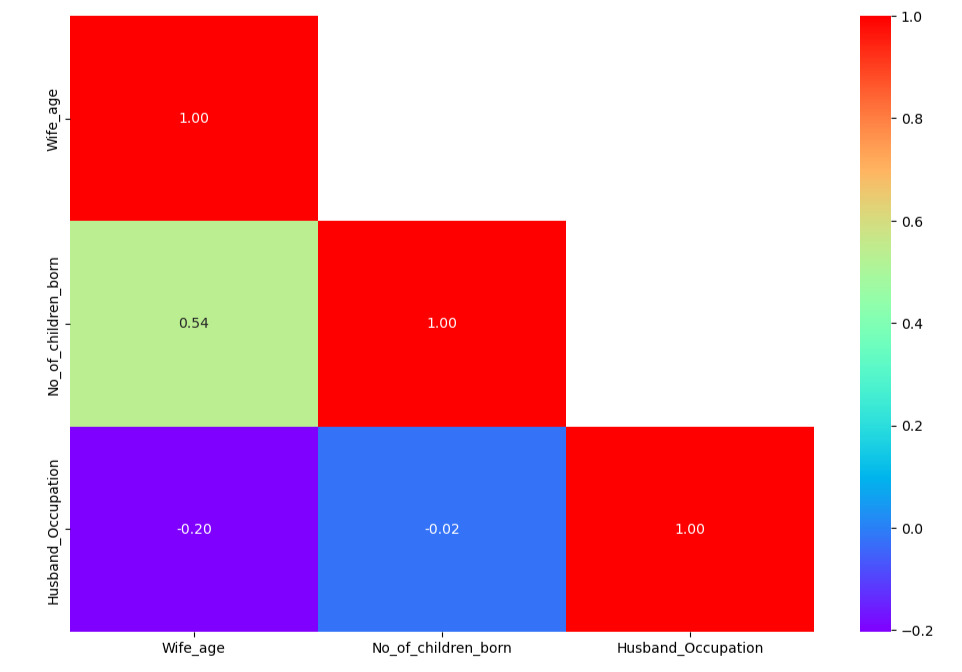


Wives and husband with tertiary education have very high standard of living.It means that education of either men or women affects standard of living.



Wives with scientology as religion generally use Contraceptive method.





No\_of\_Children\_born and Wife\_age is mildly related to each other. All other variables are not related to a certain extend.

Checking for Outliers



There are Outliers in No\_of\_children\_born varaiable.We have treated this outlier using IQR method.

There are 85 duplicate values. We have dropped these values.

**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.**

**Ans:-**

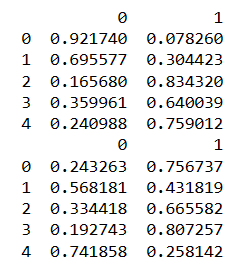
We have Encoded all the categorical variables by replacing values of the variables with certain numbers.



We have copied predictor variable into X dataframe and target variable into y dataframe.

Then we have split the data into train and test into 70:30 ratio.

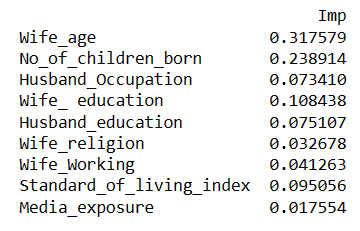
Logistic Regression

1. We have applied Logistic Regression to the training data and testing data
2. Then we have predicted test and training data
3. We have calculated predicted class and their probability
4. 

LDA

1. We have fitted training data into LDA model.
2. We have calculated intercept value and coefficient for linear discriminant function.
3. 
4. 
5. Training and test data class and probability prediction values are calculated.

CART

1. We have fitted Training data into decision tree classifier.
2. Decision tree was drawn which was not regularized.
3. 
4. Variable importance is calculated. We can see that Wife\_age is of most importance. And Wife\_religion of least importance.
5. Then regularized decision tree model was made with depth of 7.
6. Class predicted values were calculated using regularized decision tree model.

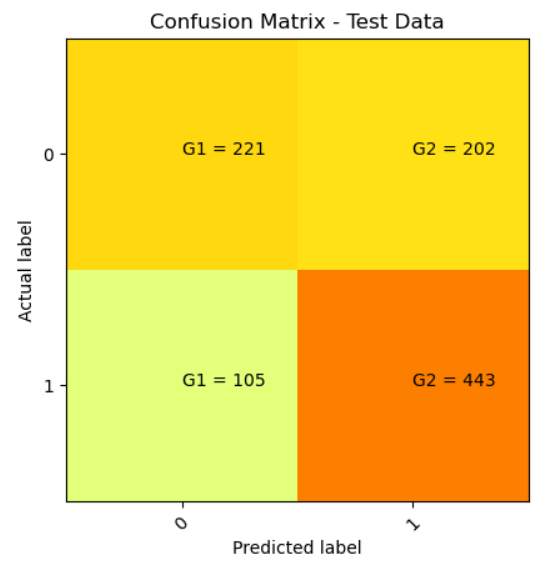
**2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.**

**Ans:-**

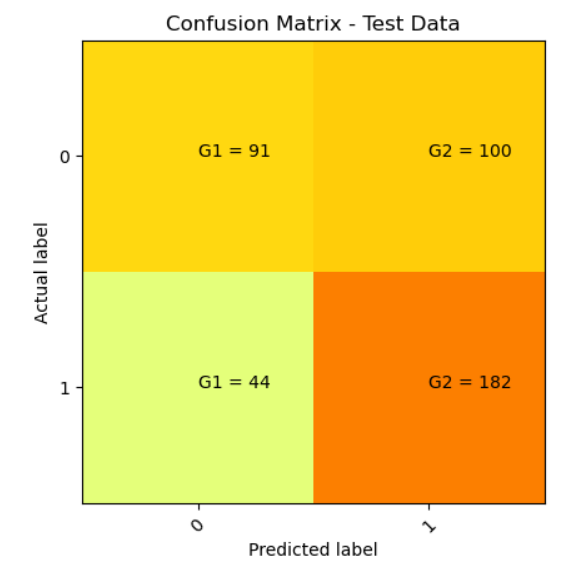
**LOGISTIC REGRESSION**

1. **CONFUSION MATRIX**

**TRAIN DATA**

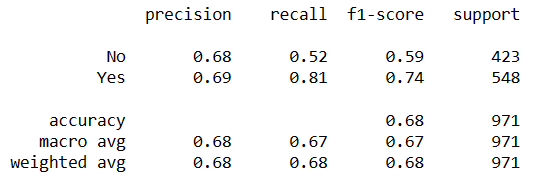


TEST DATA

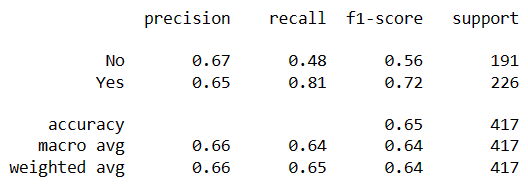


1. CLASSIFICATION REPORT

TRAIN DATA

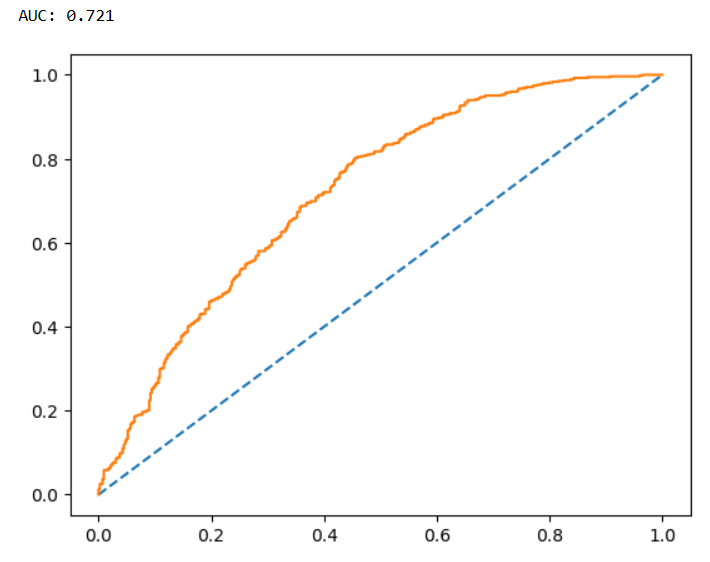


TEST DATA

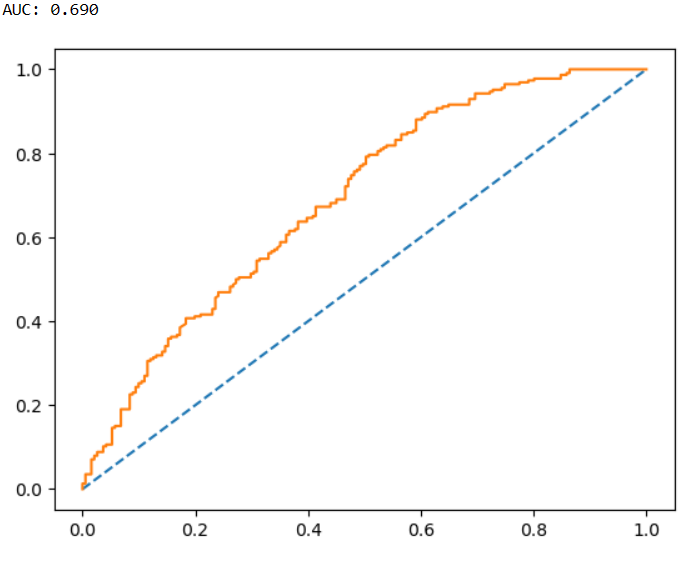


1. AUC AND ROC CURVE

TRAIN DATA



TEST DATA



INSIGHTS:-

1. For predicting contraceptive methods not used (Label No ):
2. Precision (68%) – 68% of wives predicted to not use contraceptive out of all wives predicted to not use contraceptive methods.
3. Recall (52%) – Out of all the wives actually not using contraceptive methods, 52% of wives have been predicted correctly.
4. For predicting contraceptive methods used (Label Yes):
5. Precision (68%) – 68% of wives predicted to use contraceptive out of all wives predicted to use contraceptive methods.
6. Recall (81%) – Out of all the wives actually not using contraceptive methods, 81% of wives have been predicted correctly.

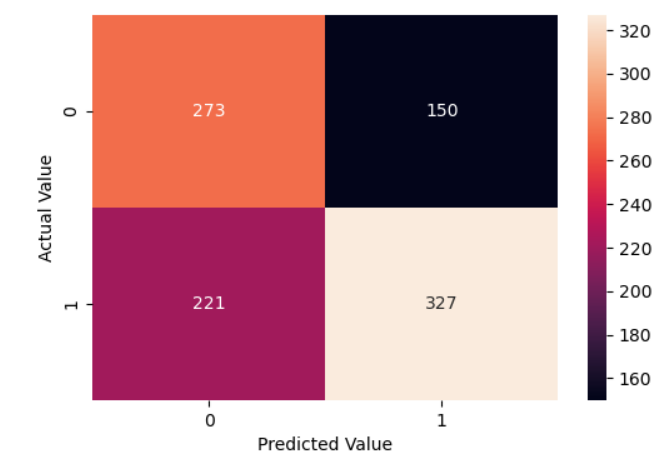
### Overall accuracy of the model – 68 % of total predictions are correct

1. Accuracy, AUC, Precision and Recall for test data is almost inline with training data. This proves no overfitting or underfitting has happened, and overall the model is a good model for classification.

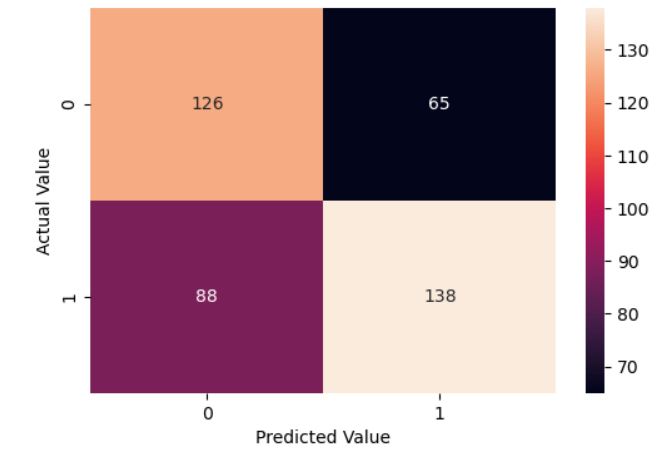
LDA

CONFUSION MATRIX

TRAIN DATA

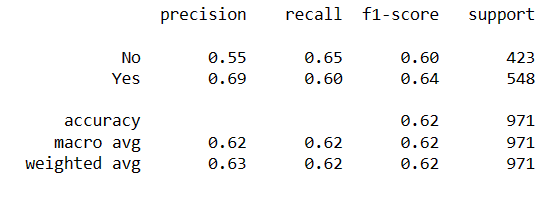


TEST DATA



CLASSIFICATION REPORT

TRAIN DATA

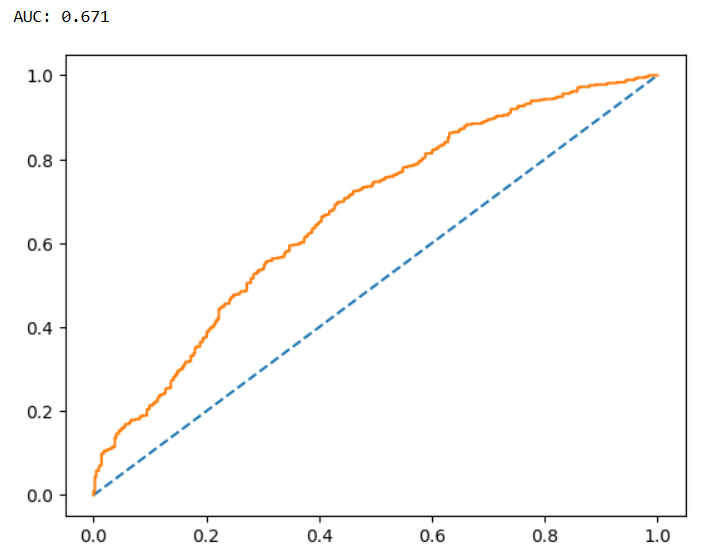


TEST DATA

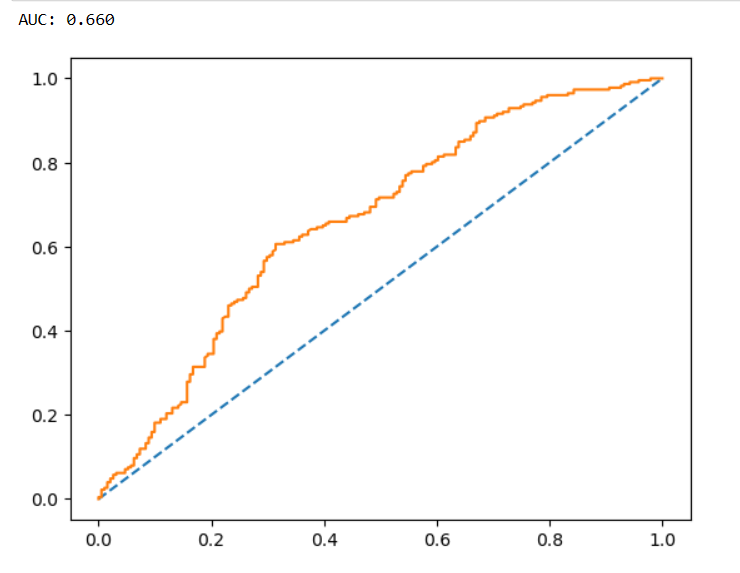


AUC AND ROC CURVE

TRAIN DATA



TEST DATA



INSIGHTS:-

1. For predicting contraceptive methods not used (Label No ):
2. Precision (55%) – 55% of wives predicted to not use contraceptive out of all wives predicted to not use contraceptive methods.
3. Recall (65%) – Out of all the wives actually not using contraceptive methods, 65% of wives have been predicted correctly.
4. For predicting contraceptive methods used (Label Yes):
5. Precision (69%) – 69% of wives predicted to use contraceptive out of all wives predicted to use contraceptive methods.
6. Recall (60%) – Out of all the wives actually not using contraceptive methods, 81% of wives have been predicted correctly.

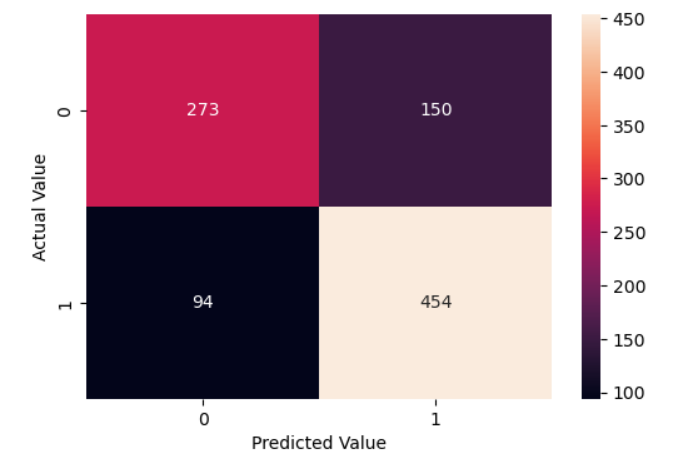
### Overall accuracy of the model – 62 % of total predictions are correct

1. Accuracy, AUC, Precision and Recall for test data is almost inline with training data. This proves no overfitting or underfitting has happened, and overall the model is a good model for classification.

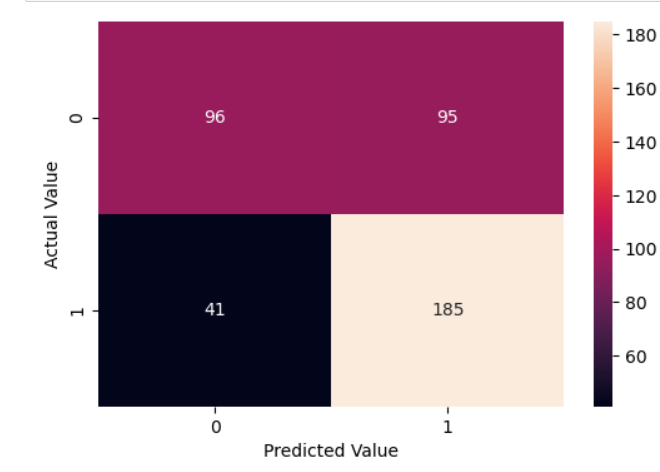
CART

CONFUSION MATRIX

TRAIN DATA

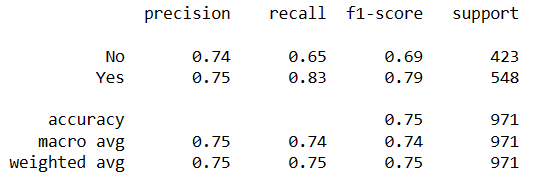


TEST DATA

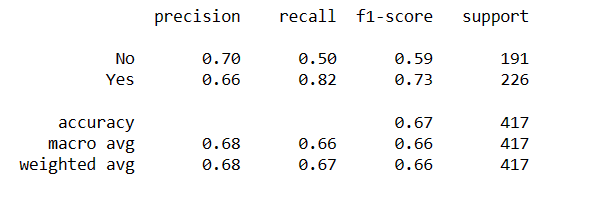


CLASSIFICATION REPORT

TRAIN DATA

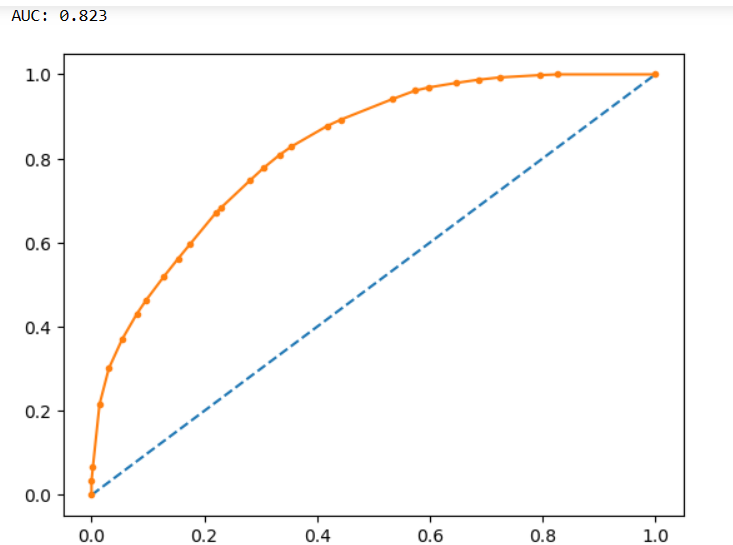


TEST DATA

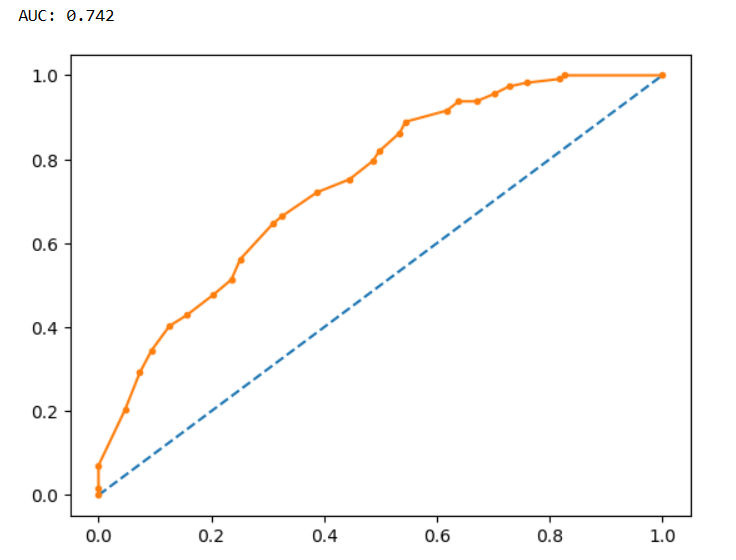


AUC and ROC curve

TRAIN DATA



TEST DATA



Accuracy on the Training Data: 75%  
Accuracy on the Test Data: 67%

AUC on the Training Data: 82.3%  
AUC on the Test: 74.2%

MODEL COMPARISION

1. Accuracy for logistic Regression for TRAINING DATA:- 0.68

Accuracy for LDA for TRAINING DATA:- 0.62

Accuracy for CART for TRAINING DATA:-0.75

1. AUC AND ROC for train for LOGISTIC:-0.72

AUC AND ROC FOR TRAIN FOR LDA:-0.67

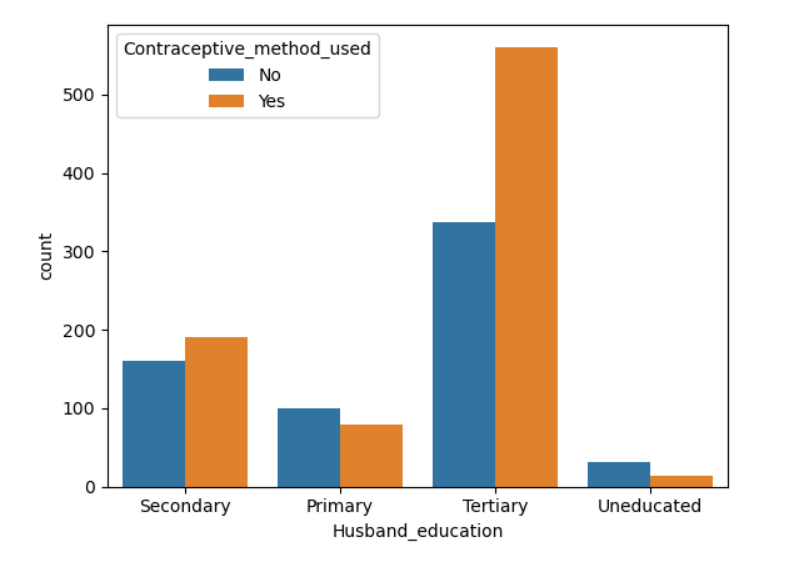
AUC AND ROC for TRAIN for CART:-0.823

Accuracy ,Area under curve and ROC curve shows better result for CART model.

In terms of precision and recall for both the classes CART model covers most of the data. So we can say that CART model should be used in this case to give better result.

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

**ANS:-**

1. We have imported required libraries and loaded dataset
2. There are 1473 rows and 10 columns
3. There are 2 float data type variable,1 integer and 7 object datatype.
4. There were few null values in the data in interger columns which were treated using median of each variable.
5. There were 85 duplicate values which were dropped.
6. Outliers were present in only one variable which was treated.
7. Data for each variable was visualized using various visualization techniques.
8. All the data in catergorical variable was encoded into numeric values.
9. 1st we have used Logistic regression model.
10. Confusion matrix,classification report and AUC and ROC curve was plotted for both test and training data.
11. 2nd we have used LDA model.
12. Confusion matrix,classification report and AUC and ROC curve was plotted for both test and training data.
13. Then at last we have used CART mode.
14. Confusion matrix,classification report and AUC and ROC curve was plotted for both test and training data.
15. Decision tree was plotted using CART model.
16. Accuracy for CART model was highest among 3 models. This is considered best model to see performance matrix in this case.
17. Wives education is most important factor which decides standard of living, their working status.
18. Women who are educated and believes in scientology are more likely to use contraceptive method.
19. Women who use contraceptive methods might affect no of children born.
20. 
21. Husband education also affects women taking contraceptive pills. Educated husband encourage wives to use contraceptive methods more.