**PREDICTIVE MODELLING**

**CONTENTS:**

|  |  |  |
| --- | --- | --- |
| S. no | content | Pg. no |
|  | **Problem 1**: Linear Regression | 5 |
| 1.1 | **Read the data and do exploratory data analysis. Describe the data briefly. (Check the Data types, shape, EDA, 5-point summary). Perform Univariate, Bivariate Analysis, Multivariate Analysis.** | 6 |
| 1.2 | **Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of creating new features if required. Also check for outliers and duplicates if there.** | 8 |
| 1.3 | **Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from stats model. Create multiple models and check the performance of Predictions on Train and Test sets using R squared, RMSE & Adj R square. Compare these models and select the best one with appropriate reasoning.** | 15 |
| 1.4 | **Inference: Basis on these predictions, what are the business insights and recommendations.** | 24 |
|  |  |  |
|  | **Problem 2:** Logistic Regression, LDA and CART | 25 |
|  |  |  |
| 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.** | 26 |
| 2.22 | **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.** | 32 |
| 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.** | 32 |
| 2.4 | **Inference: Basis on these predictions, what are the insights and recommendations.** | 44 |

**LIST OF THE TABLES**

|  |  |  |
| --- | --- | --- |
| S. No | List of tables | Pg. no |
|  | Head of the comp-activ data | 6 |
|  | Basic information of comp-activ data | 6 |
|  | Null values in comp-activ data | 7 |
|  | Statistical information | 7 |
|  | Null values after treatment | 8 |
|  | Variance Influence Factor | 14 |
|  | Coefficients of train data | 16 |
|  | OLS results model1 | 18 |
|  | Coefficients of model2 | 20 |
|  | OLS results model2 | 21 |
|  | Coefficients of model3 | 22 |
|  | OLS results model 3 | 23 |
|  | Head of the contraceptive prevalence survey | 25 |
|  | Basic information of contraceptive prevalence survey | 26 |
|  | Null values before treatment | 26 |
|  | Null values after treatment | 26 |
|  | Statistical description | 27 |
|  | Classification report for training data | 33 |
|  | Classification report for testing data | 33 |
|  | Coefficients of | 35 |
|  | Classification report for training data | 37 |
|  | Classification report for testing data | 37 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| S.no | List of Figures | Pg. no |
|  | Outliers check before treatment | 9 |
|  | Outliers check after treatment | 10 |
|  | Univariate Analysis | 11 |
|  | Bivariate Analysis | 12 |
|  | Correlation plot | 14 |
|  | Correlation after dropping | 15 |
|  | Feature importance for model 1 | 16 |
|  | Regression plot | 19 |
|  | Feature importance of model 2 | 20 |
|  | Feature importance of model 3 | 22 |
|  | Correlation among variables | 27 |
|  | Outliers check before treatment | 27 |
|  | Outliers check after treatment | 28 |
|  | Bar plot for all categorical data | 28 |
|  | Bivariate analysis | 30 |
|  | Confusion matrix | 32 |
|  | Confusion matrix for training data | 34 |
|  | Confusion matrix for testing data | 34 |
|  | Confusion matrix of LDA | 36 |
|  | ROC of LDA | 38 |
|  | Feature importance of without regularizing data | 39 |
|  | Feature importance of with regularizing data | 40 |
|  | Measuring AUC – ROC for training data | 40 |
|  | Measuring AUC – ROC for testing data | 41 |
|  | Confusion matrix for training data | 41 |
|  | Confusion matrix for testing data | 42 |

**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:

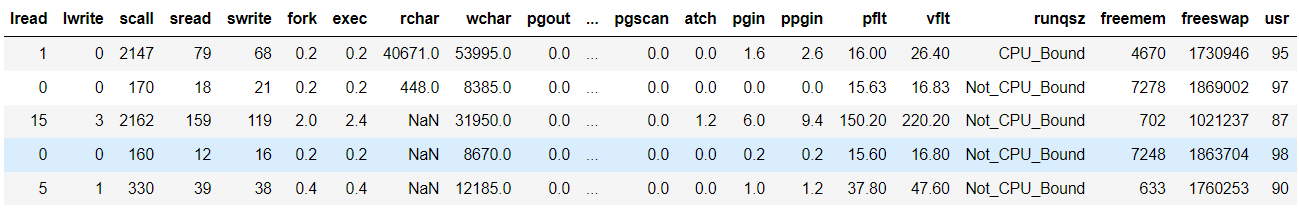
lread - Reads (transfers per second ) between system memory and user memory  
lwrite - writes (transfers per second) between system memory and user memory  
scall - Number of system calls of all types per second  
sread - Number of system read calls per second .  
swrite - Number of system write calls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

To check the data, we must import all the libraries and load the dataset.

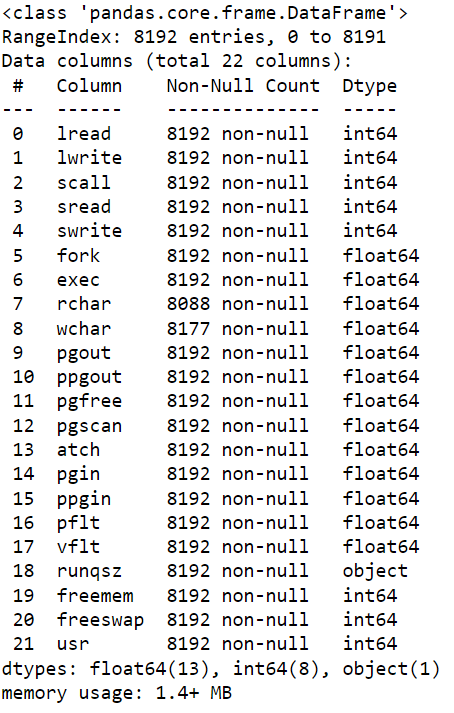
First, see the first five rows of the data. This helps to know about the data and columns.

Table 1: Head of the comp-active data



Next, check the basic information of the dataset, this can be done by using info() function.

Table 2: Basic information of comp-active data



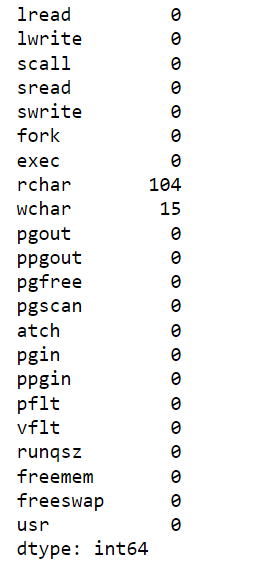
From the above table, we can understand that there are 8192 rows and 22 columns.

We could see that there are object, int, and float types.

These object-type variable to be treated.

Let us check the null values in the dataset.

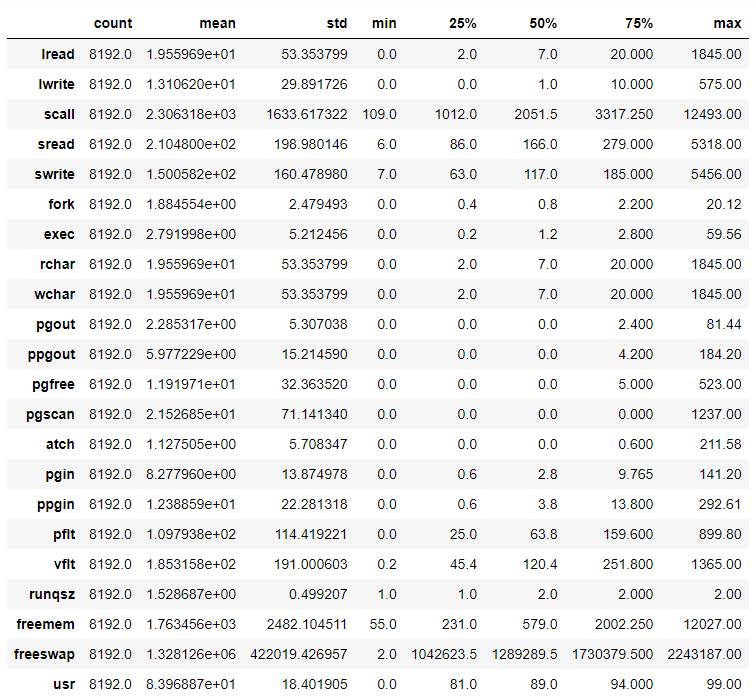
Table 3: Null values in comp-active data



Here there are null values in rchar, wchar, these are to be treated.

Check the statistical description of the data.

Table 3: statistical information of comp-active data



From the statistical information, we can understand few variables have many zeroes as their values.

This can be understood by looking into the statistical information on the variables -pgout, ppgout, pgfree, pgscan, atch.

These have almost 50% of zeroes in their columns.

Therefore we can drop these columns

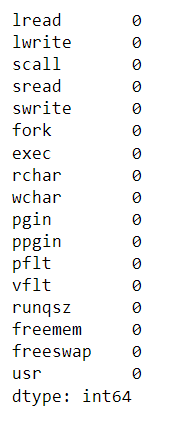
**1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of creating new features if required. Also check for outliers and duplicates if there.**

**Treating Null values**:

We have seen null values in the dataset, they can be treated in many ways.

Here, the null values are filled by the median of the data.

Table 4: Null values after the treatment



**Columns from object type to numerical type:**

For runqsz, as there are only two unique values ['CPU\_Bound','Not\_CPU\_Bound'], these values are replaced by 1,2 respectively.

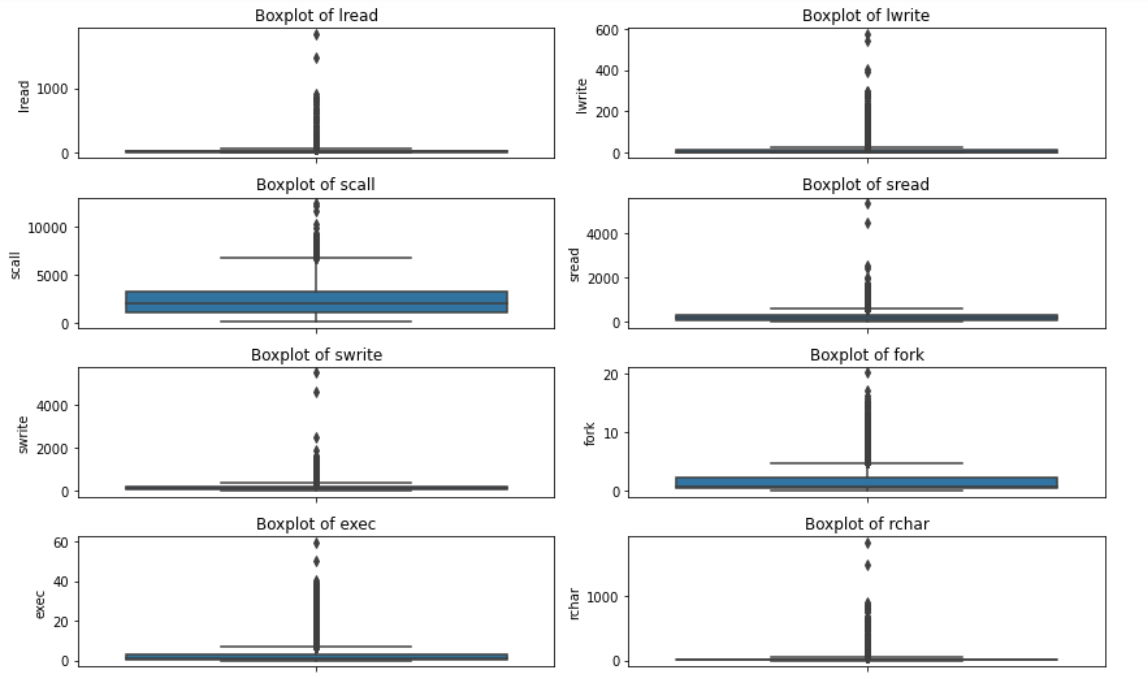
After treating null values rchar and wchar are converted into object type.

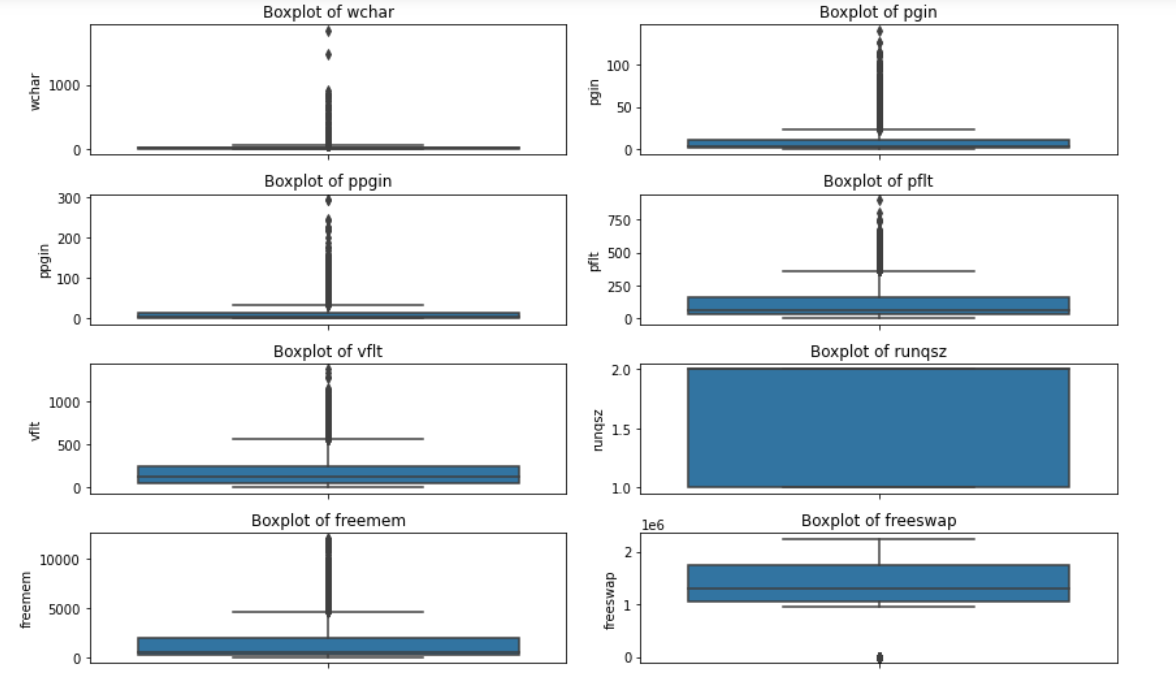
Even after checking the unique values for rchar and wchar as there are no object type data found to treat, these columns are directly changed to int datatype.

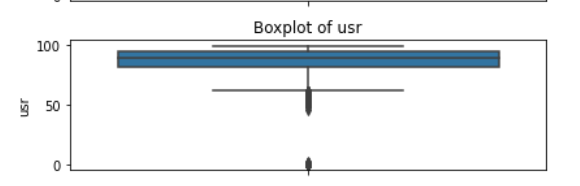
**Outlier Treatment:**

Check for the outliers which can be done by boxplot.

Figure 1: Outlier Check before treatment





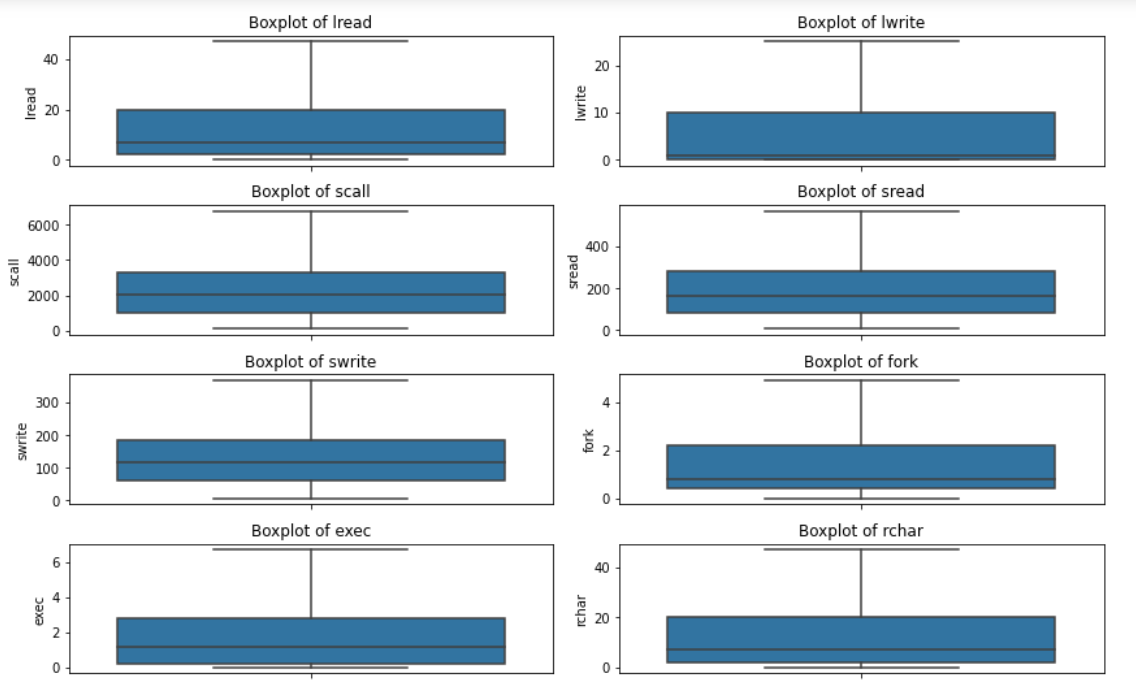


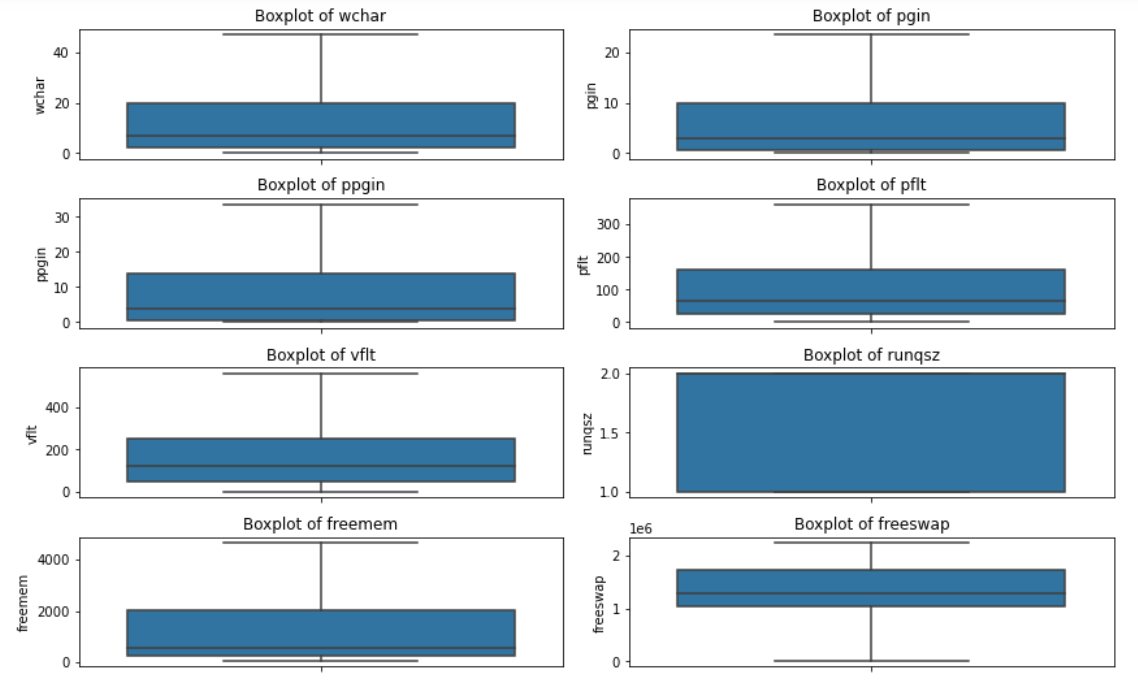
Here we can observe outliers in all the variables.These are to be treated.

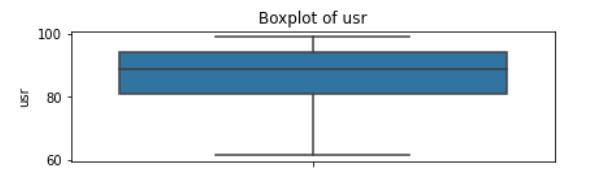
The data points which are greater than the maximum value are compressed to the maximum value.

The data points which are lesser than the minimum value are compressed to the minimum value.

Figure 2: Outlier check after treatment





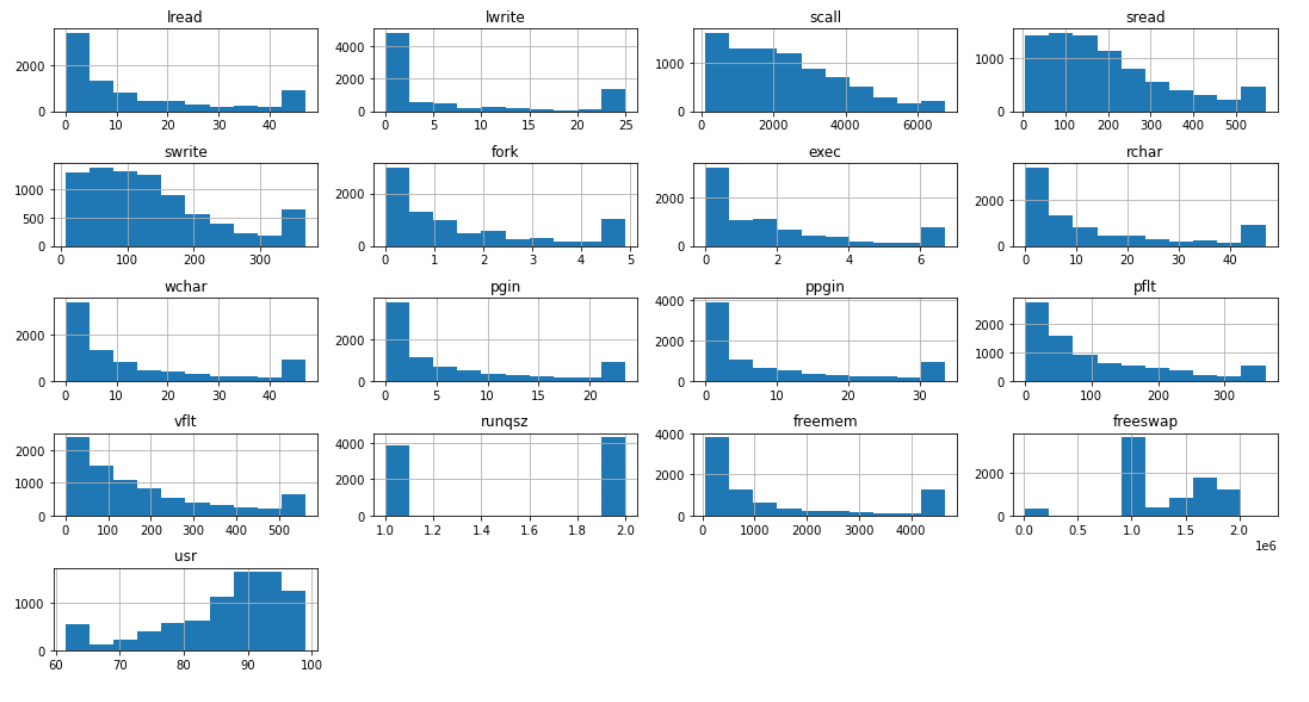


**Univariate Analysis:**

We have converted all the variables to numeric type, hence histogram is plotted to all the variables.

We could see some skewness in the data.

Figure 3: Univariate Analysis



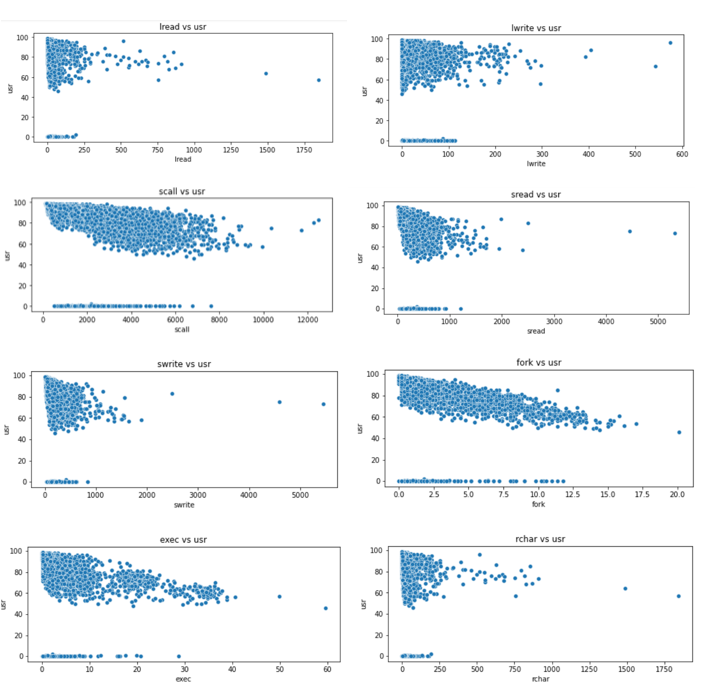
Analysis:

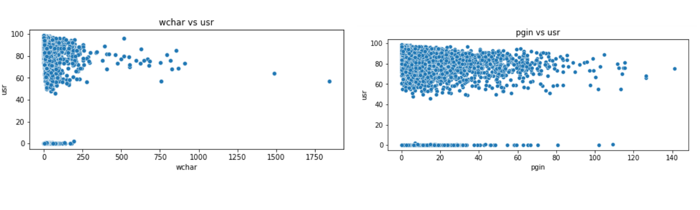
We could see some skewness in the data

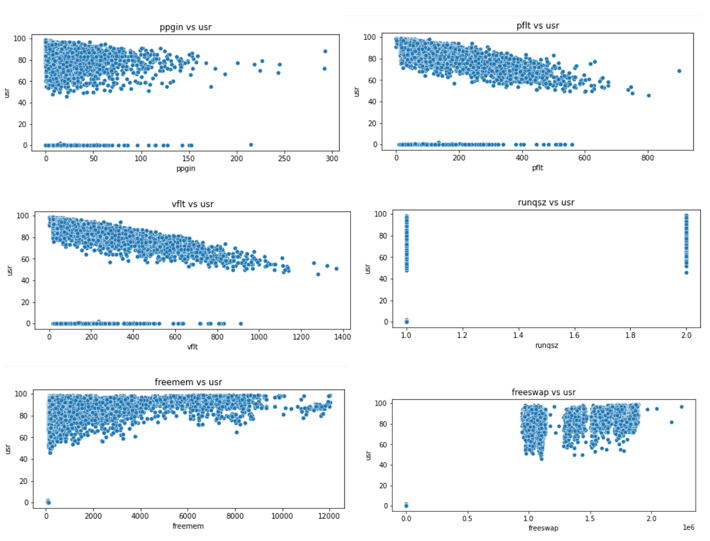
As runqsz is 1 and 2 the data is two pillars.

**Bivariate Analysis:**

Figure 4: Bivariate Analysis







**Analysis:**

1. We could see all the variables has some linear relation between them except for freeswap and runqsz
2. Even the plots we could see the zero values for all the variables.
3. All the variables have more density at the high rates of usr.

**CORRELATION :**

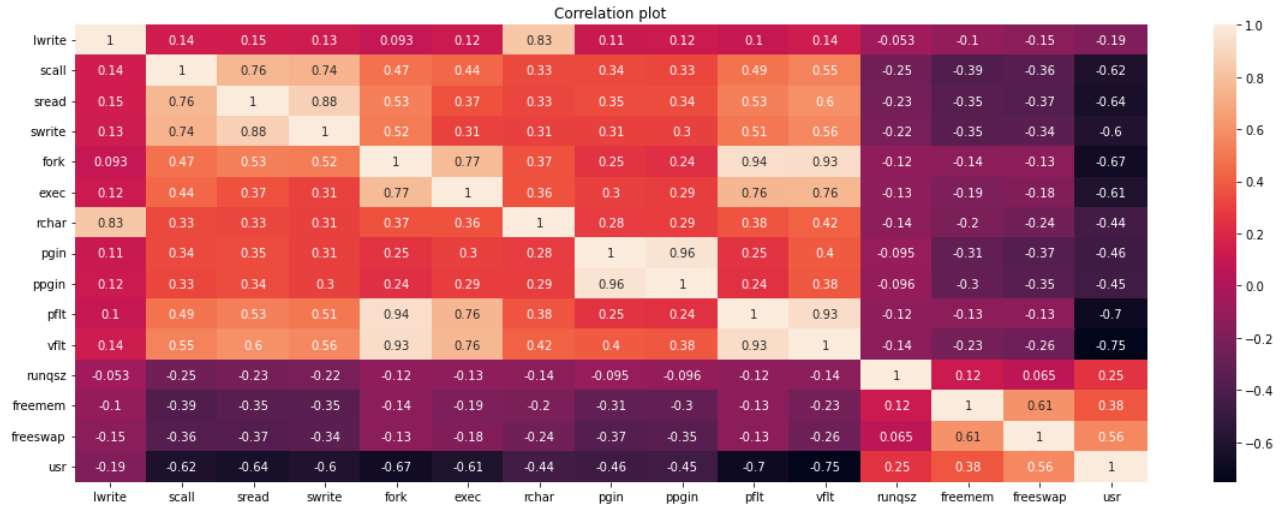
Table 5: Variance Influence Factor

****

* A value of 1 indicates there is no correlation between a given explanatory variable and any other explanatory variables in the model.
* A value between 1 and 5 indicates moderate correlation between a given explanatory variable and other explanatory variables in the model, but this is often not severe enough to require attention.
* A value greater than 5 indicates potentially severe correlation between a given explanatory variable and other explanatory variables in the model. In this case, the coefficient estimates and p-values in the regression output are likely unreliable.

Correlation plot:

Figure 5: Correlation plot

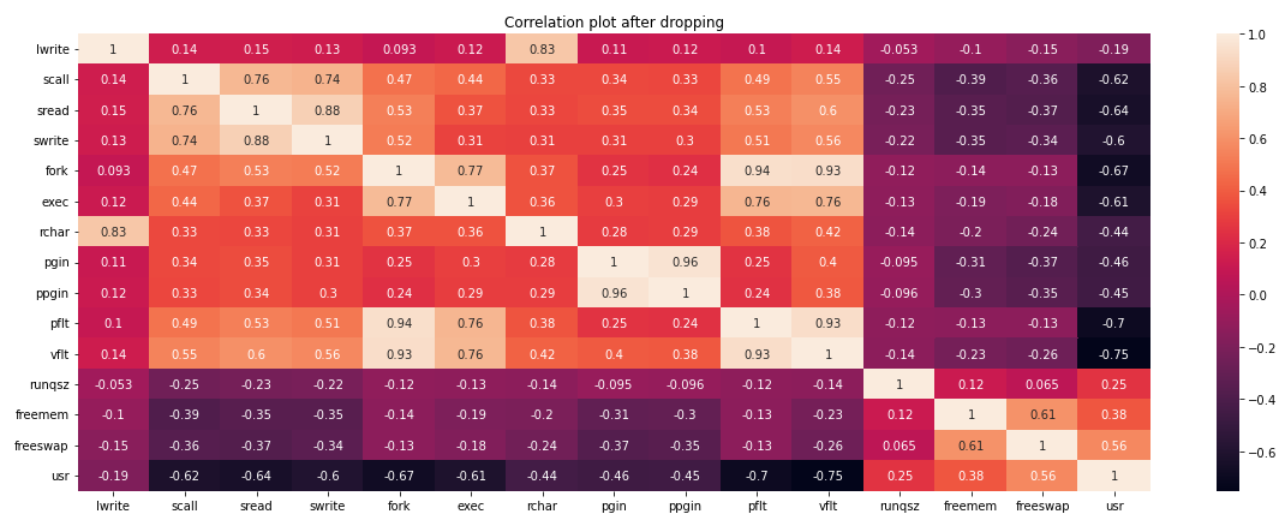


As there is much correlation between a few variables, we can drop those vaariables.

PCA or variance inflation factor can be used but as they’re more than 90% correlation. We can drop these columns.

* lread, rchar, wchar have a 100% correlation. Hence two of these variables can be dropped.
* vflt, pflt, fork has almost 94% correlation. Therefore two of these variables can also be dropped.
* pgscan, pgfree,ppgout has a 92% correlation. So, two of these variables can be dropped.
* For now lets only drop the columns with 100% correlation and check the models.

Figure 6: Correlation after dropping



**1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.**

Manipulation of data is done Let’s create a Linear Regression model for this data.

**LINEAR REGRESSION:**

Linear regression analysis is **used to predict the value of a variable based on the value of another variable**.

The variable you want to predict is called the dependent variable.

The variable you are using to predict the other variable's value is called the independent variable.

In this data set, according to the business need the dependent variable is **‘usr’** and all other remaining variables are dependent variables.

Let’s split the data into dependent and independent data. Here I used **‘x’** for the independent variables and **‘y’** for the dependent variable.

To train and test the data, we divide the data into two parts in the 70:30 ratio.

70% of the data is to train the model and 30% of the data is to test the model.To build a Linear Regression model, we should import the required libraries.

To this model we fit the train data. Here this model fits a straight line to cover the maximum of the data.

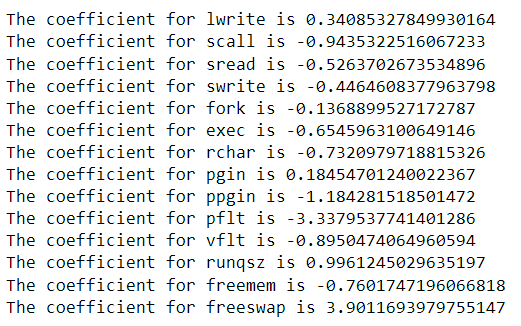
The equation will be:

Y = b0 + b1\*a1 + b2\*a2 + b3\*a3…

b0 is the intercept and b1,b2,b3.. are the coefficients of the variables a1,a2,a3 respectively.

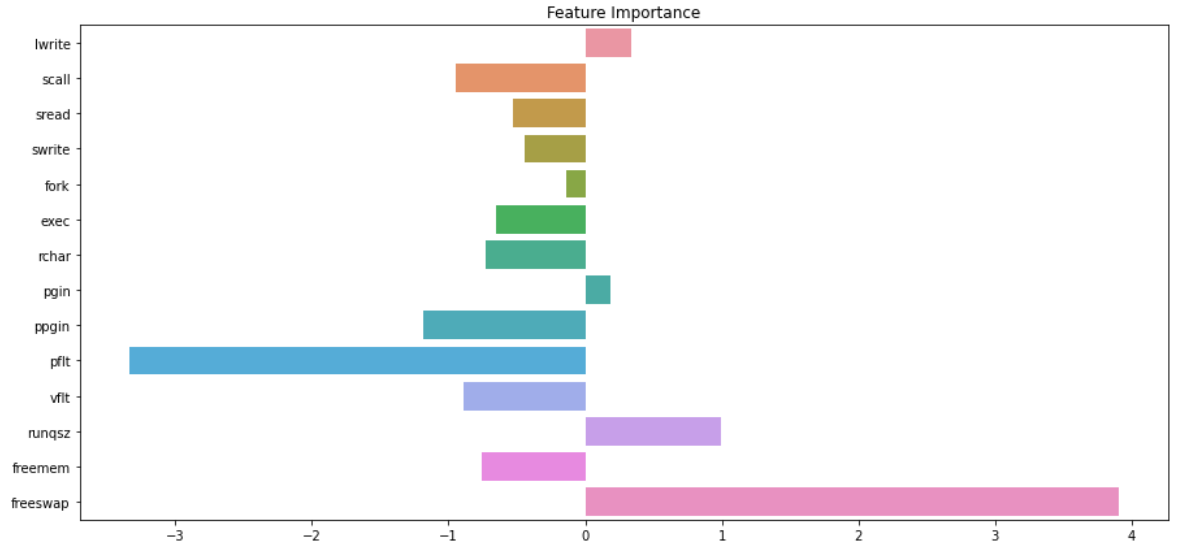
After fitting the model to the train data, the coefficients are as following:

Table 6: Coefficients of train data



**The intercept for the model is 86.28313568189745**

Figure 7: Feature Importance



R-Squared is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable.

In other words, r-squared shows how well the data fit the regression model

**R square value is about 0.78, Which is 78% of the data explained in this model**.

The RMSE (Root Mean Square Error) estimates the deviation of the actual y-values from the regression line.

Another way to say this is that it estimates the standard deviation of the y-values in a thin vertical rectangle. where ei = yi - yi^. The RMSE can be computed more simply as

RMSE = SDy √(1 - r2).

**The RMSE for the training data is 4.526**

**The RMSE for the testing data is 4.727**

**Ordinary Least Squares regression (OLS):**

Ordinary Least Squares regression (OLS) is a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable.

From this OLS, we can understand R square value, coefficients of the data and Probability.

P values and coefficients in regression analysis work together to tell you which relationships in your model are statistically significant and the nature of those relationships.

The linear regression coefficients describe the mathematical relationship between each independent variable and the dependent variable.

The p values in regression help determine whether the relationships that you observe in your sample also exist in the larger population.

The linear regression p value for each independent variable tests the null hypothesis that the variable has no correlation with the dependent variable.

If there is no correlation, there is no association between the changes in the independent variable and the shifts in the dependent variable. In other words, there is insufficient evidence to conclude that there is an effect at the population level.

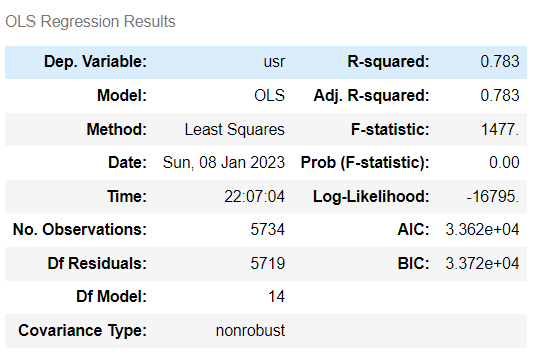
If the p-value for a variable is less than your significance level, your sample data provide enough evidence to reject the null hypothesis for the entire population.

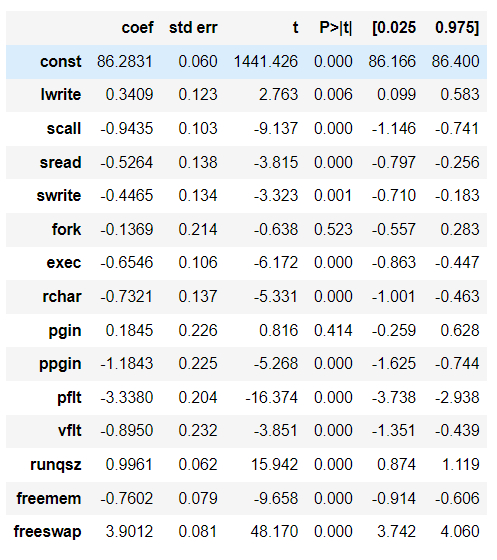
That is if p value is greater than 0.05, they are considered as insignificant. So, we drop those variables.

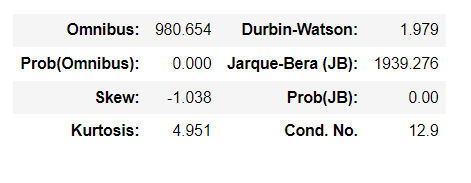
* A low P-value (< 0.05) means that the coefficient is likely not to equal zero.
* A high P-value (> 0.05) means that we cannot conclude that the explanatory variable affects the dependent variable (here: if Average\_Pulse affects Calorie\_Burnage).
* A high P-value is also called an insignificant P-value.

The below is the OLS Result for first model:

Table 7: OLS results

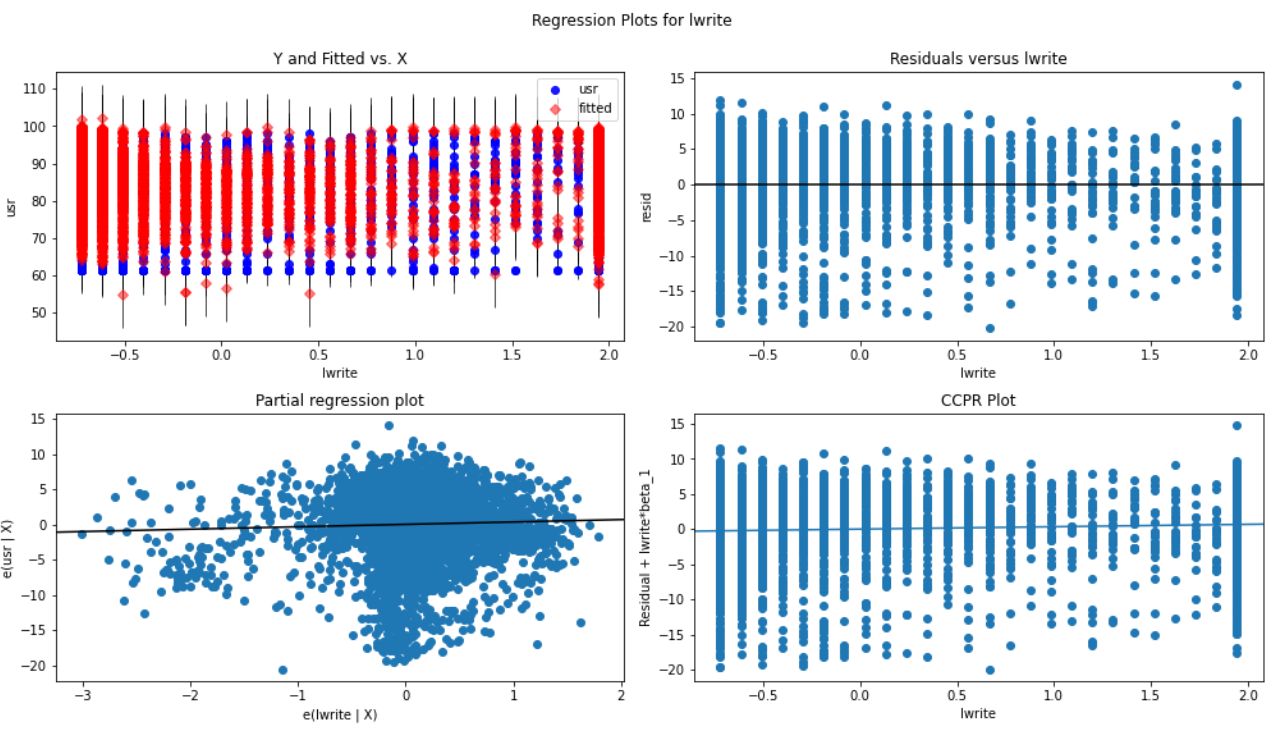






.

Figure 8: Regression Plot

****

This is regression plot for lwrite.

First figure tells about the independent vs dependent variables, that is lwrite vs usr

Second figure tells about the residuals versus lwrite

Third figure tells about fitted versus lwrite.

Forth figure tells about fitted plus residual versus lwrite (CCPR plot)

**The Linear regression equation for first is :**

usr = b0 \* const + b1 \* lwrite + b2 \* scall + b3 \* sread + b4 \* swrite + b5 \* fork + b6 \* exec + b7 \* rchar + b8 \* pgin + b9 \* ppgin + b10 \* pflt + b11 \* vflt + b12 \* runqsz + b13 \* freemem + b14 \* freeswap

**usr = (86.283) \* const + (0.341) \* lwrite + (-0.944) \* scall + (-0.526) \* sread + (-0.446) \* swrite + (-0.137) \* fork + (-0.655) \* exec + (-0.732) \* rchar + (0.185) \* pgin + (-1.184) \* ppgin + (-3.338) \* pflt + (-0.895) \* vflt + (0.996) \* runqsz + (-0.76) \* freemem + (3.901) \* freeswap**

This model explains 78% of the data, but there are variables which are insignificant

As we could see p value greater for pgin and fork, they should be dropped one after one and build models.

**MODEL 2:**

For model 2, will drop ‘pgin’ variable and build model.

Here as same, we split and divide the data to train and test. After fitting the coefficients for model2 is as follows.

Table 8: Coefficents of model2

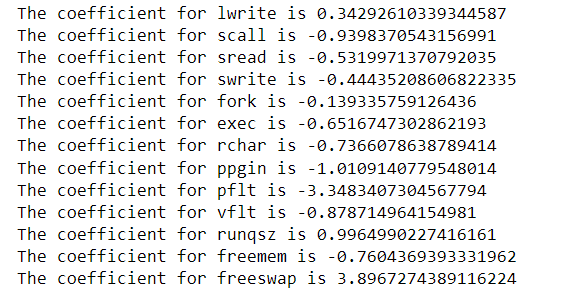
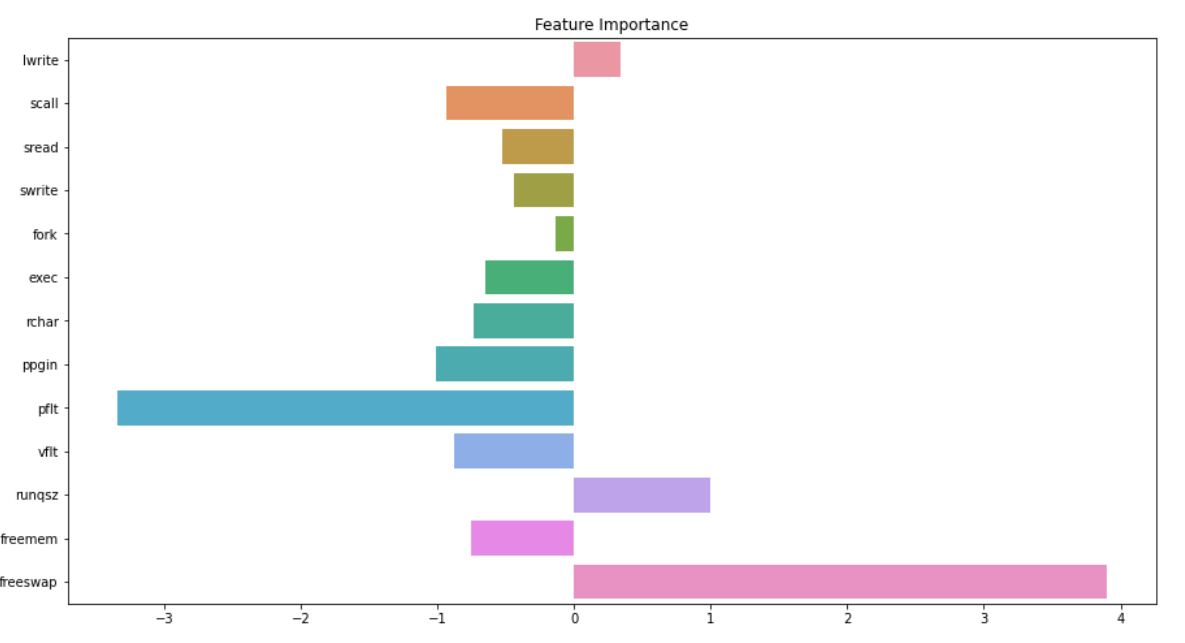


Figure 9: Feature importance for model2



**The intercept for our model is 86.28313568189745**

**The R Square value for training data is 0.7833**

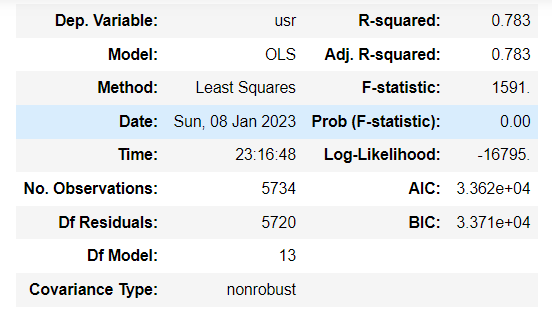
**The R square value for testing data is 0.7635**

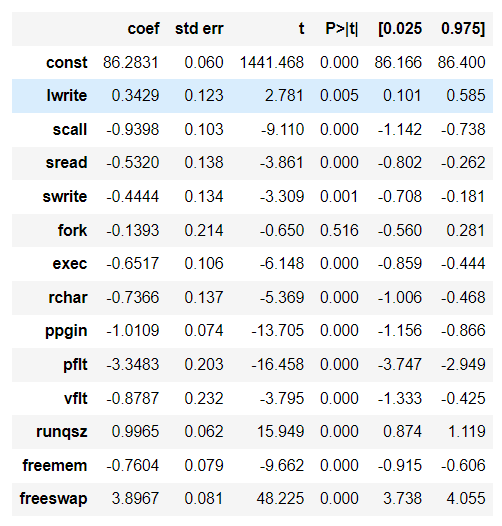
**The RMSE for the training data is 4.527**

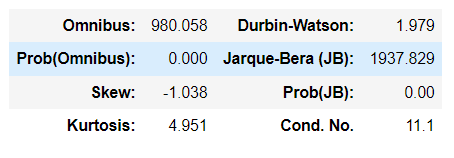
**The RMSE for the testing data is 4.750**

OLS Results:

Table 9: OLS Results

****





There is one variable with high p value, which is to be dropped.

Let’s drop fork variable and build other model.

**The Linear regression equation for second model is**

**Usr =(86.283) \* const + (0.341) \* lwrite + (-0.944) \* scall + (-0.526) \* sread + (-0.446) \* swrite + (-0.137) \* fork + (-0.655) \* exec + (-0.732) \* rchar + (0.185) \* pgin + (-1.184) \* ppgin + (-3.338) \* pflt + (-0.895) \* vflt + (0.996) \* runqsz + (-0.76) \* freemem + (3.901) \* freeswap.**

**MODEL3:**

For model 3, will drop **‘pgin’, ‘fork’** variables and build the model.

Here as same, we split and divide the data to train and test. After fitting the coefficients for model2 is as follows.

Table 10: Coefficients for model 3

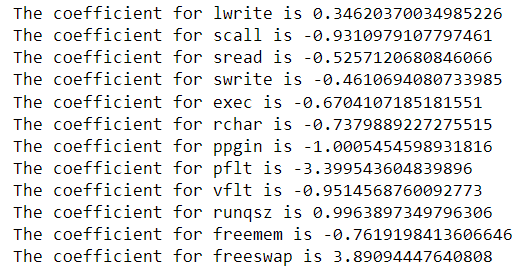
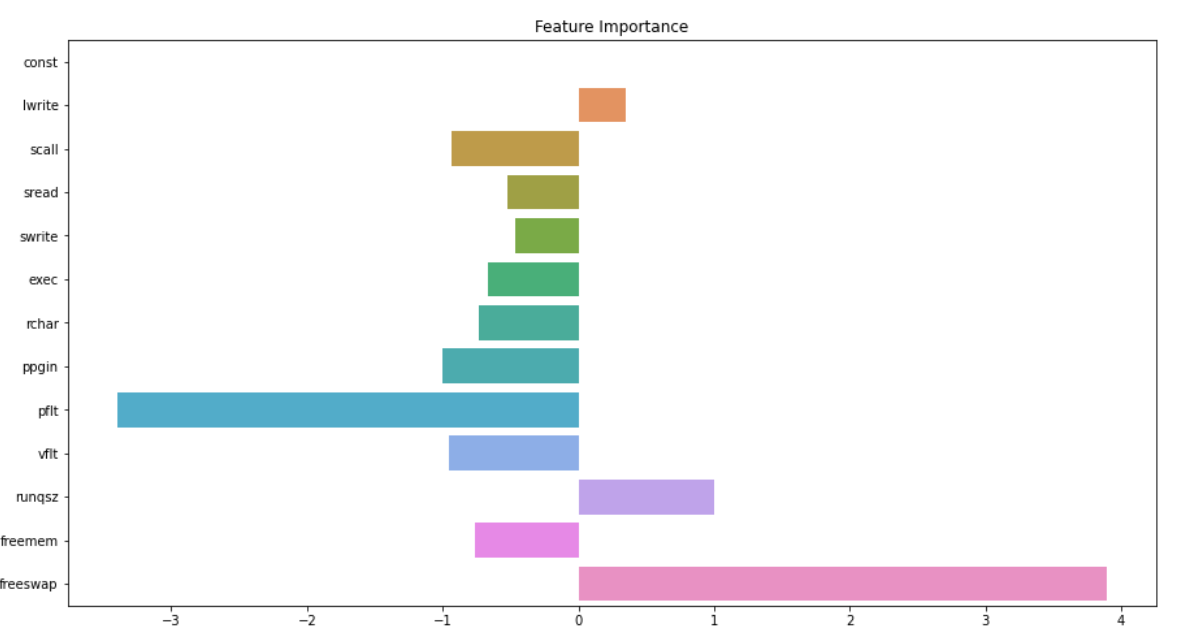


Figure 10: Feature Importance for model3

****

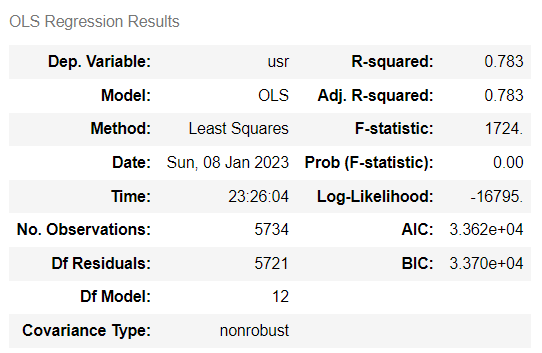
**The intercept for our model is 86.28313568189745**

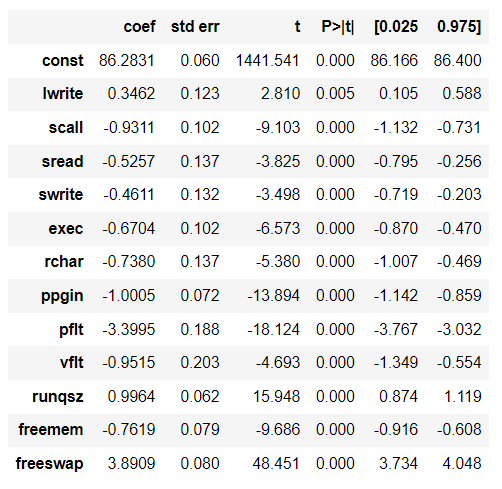
**The R Square value for training data is 0.7833**

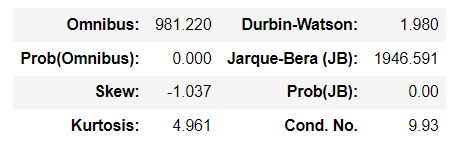
**The R square value for testing data is 0.7636**

**The RMSE for the training data is 4.527**

**The RMSE for the testing data is 4.752**

Table 11: OLS Results for model 3 ****





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

The model 3 is our final and best model for the data.

78% of the variance is explained.

Coefficients are not too small.

P values for all the variables are significant.

The final linear regression equation is:

Usr = (86.283) \* const + (0.346) \* lwrite + (-0.931) \* scall + (-0.526) \* sread + (-0.461) \* swrite + (-0.67) \* exec + (-0.738) \* rchar + (-1.001) \* ppgin + (-3.4) \* pflt + (-0.951) \* vflt + (0.996) \* runqsz + (-0.762) \* freemem + (3.891) \* freeswap ​

There are three positive values which increases the portion of time in user mode. As 1 unit of lwrite increases then usr increases by 0.346.

There are few negative values which decreases the portion of time in user mode. As 1 unit of scall increases then usr decreases by 0.931.

Insights from our final model:

1. There is a major increment in the portion of time in user mode by Number of disk blocks available for page swapping.
2. There is some increment in the portion of time in user mode by The number of kernel threads in memory that are waiting for a CPU to run.
3. There is a major decrement in the portion of time in user mode by Number of page faults caused by protection errors.
4. There is a decrement in the portion of time in user mode by Number of pages paged in per second.

### **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:**

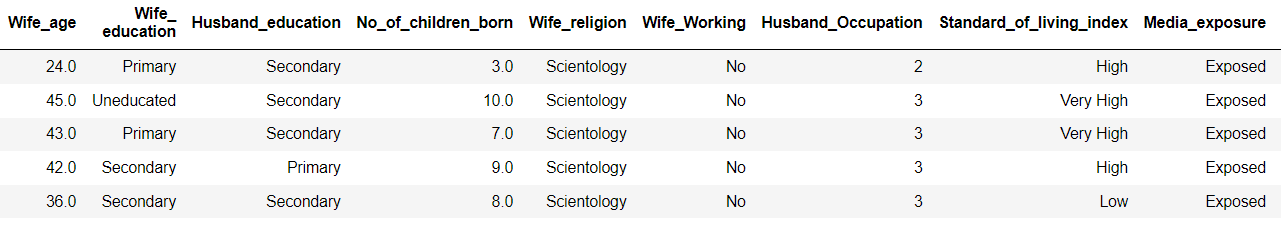
Wife's age (numerical)  
2. Wife's education (categorical) 1=uneducated, 2, 3, 4=tertiary  
3. Husband's education (categorical) 1=uneducated, 2, 3, 4=tertiary  
4. Number of children ever born (numerical)  
5. Wife's religion (binary) Non-Scientology, Scientology  
6. Wife's now working? (binary) Yes,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

To check the data, we must import all the libraries and load the dataset.

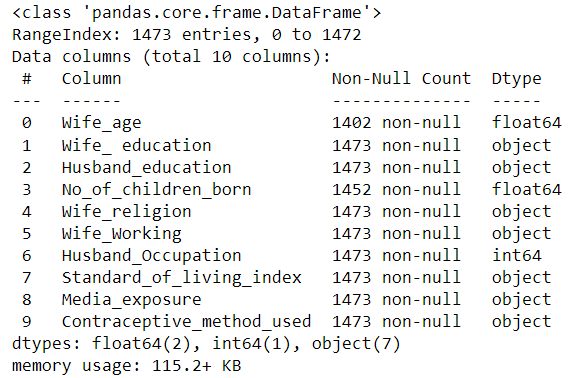
First, see the first five rows of the data. This helps to know about the data and columns.

Table 12: Head of Contraceptive Prevalence Survey



Next, check the basic information of the dataset, this can be done by using info() function.

Table 13: Basic information of Contraceptive Prevalence Survey

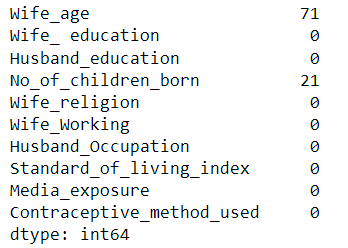


There are object datatypes, int and float datatypes.

There are null values in two of the column.

**Null values treatment:**

Table 14: Null values before treatment



The null values are clearly in columns wife age, No of children born.

These null values treated using median. These null values will be filled with median of that specific column.

Table 15: Null values after treatment

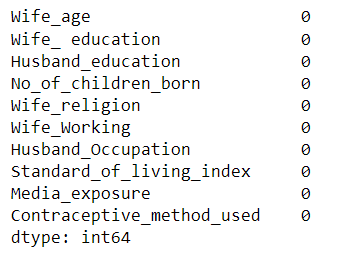
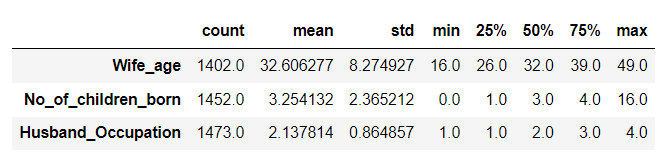
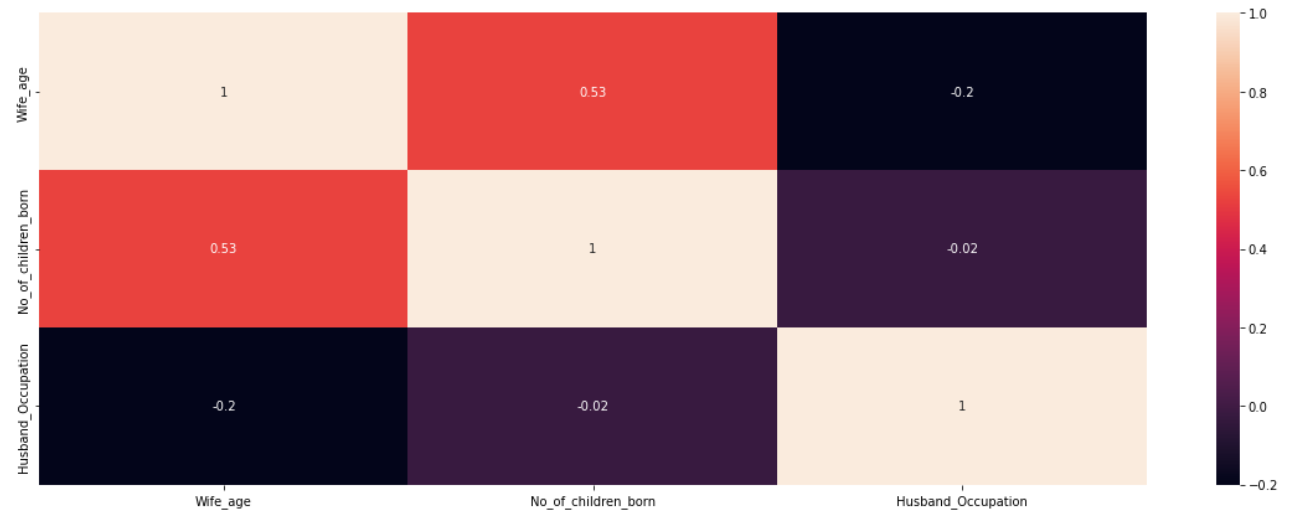


Table 16: Statistical description



The description of only numeric variables are shown.

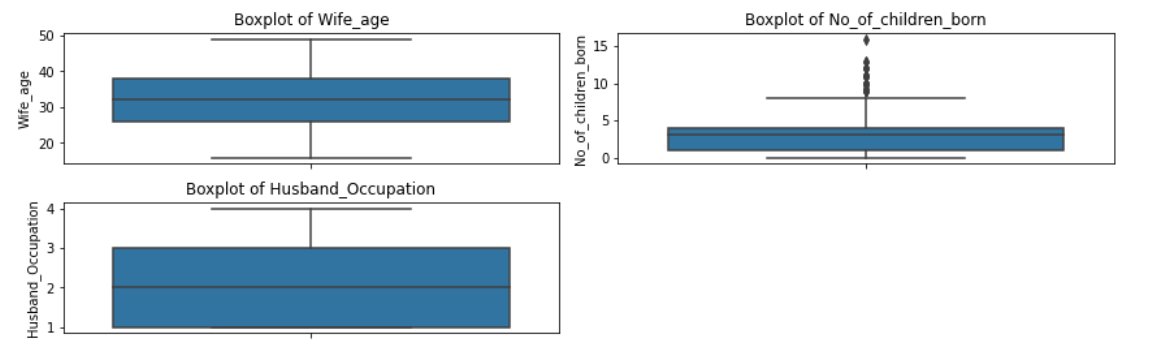
Figure 11: Correlation among variables



The correlation among variables is good.

**Outlier treatment:**

Figure 12: Outliers before treatment



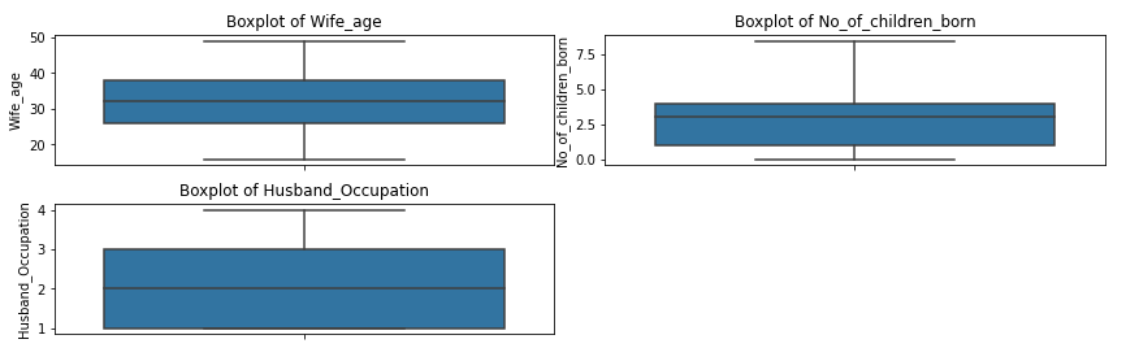
Outliers can be seen through boxplot. There are outliers in No of children born.

This have to be treated.

The data points which are greater than the maximum value are compressed to the maximum value.

The data points which are lesser than the minimum value are compressed to the minimum value.

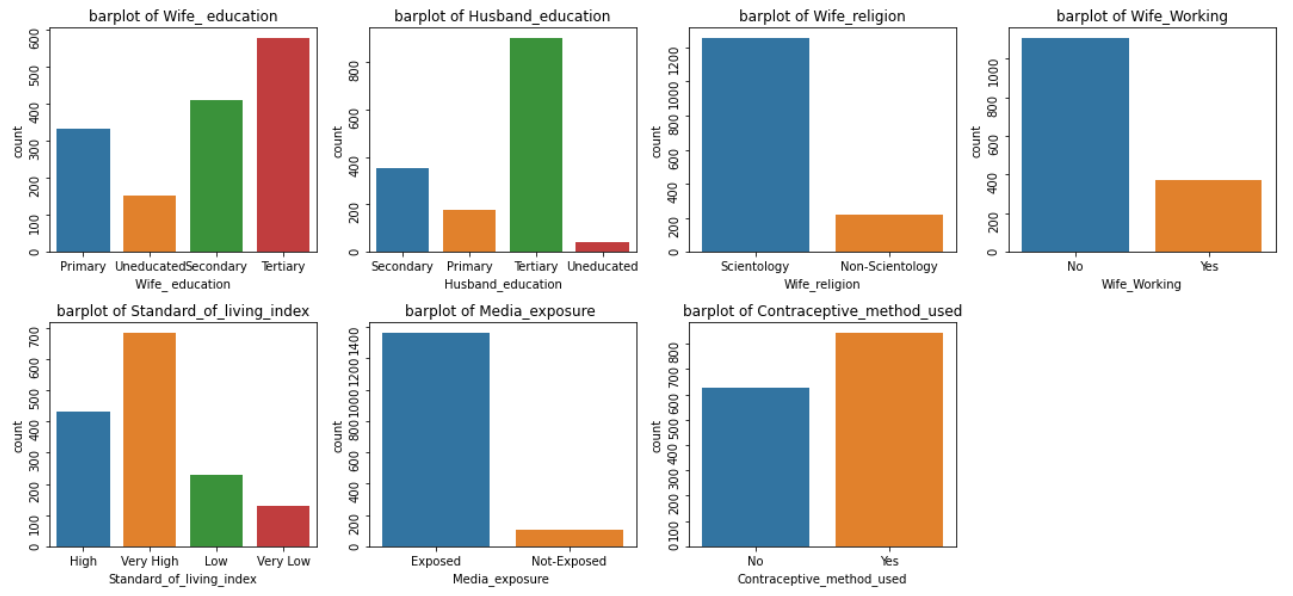
Figure 13: Outliers after treatment



Now, there are no outliers.

**Univariate Analysis:**

Figure 14: Barplot for all the categorical data



**Analysis:**

There are more educated women and men.

When comparing there are more uneducated women than men.

There are more non working women.

Media is more exposed to people.

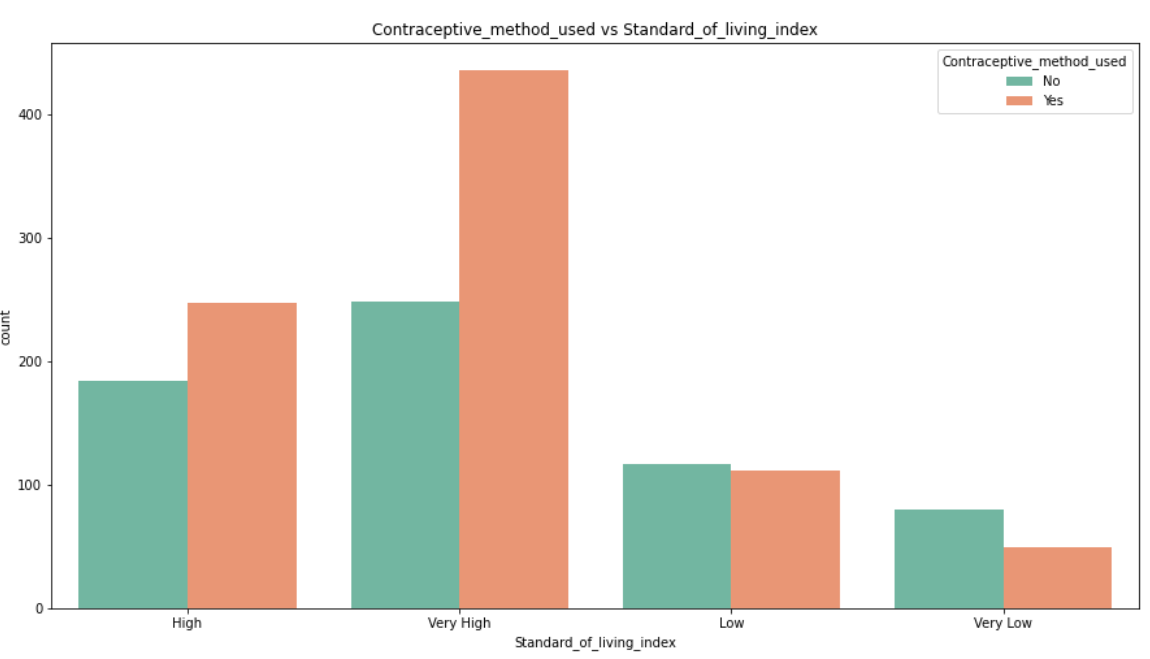
Most of the people are using contraceptive methods.

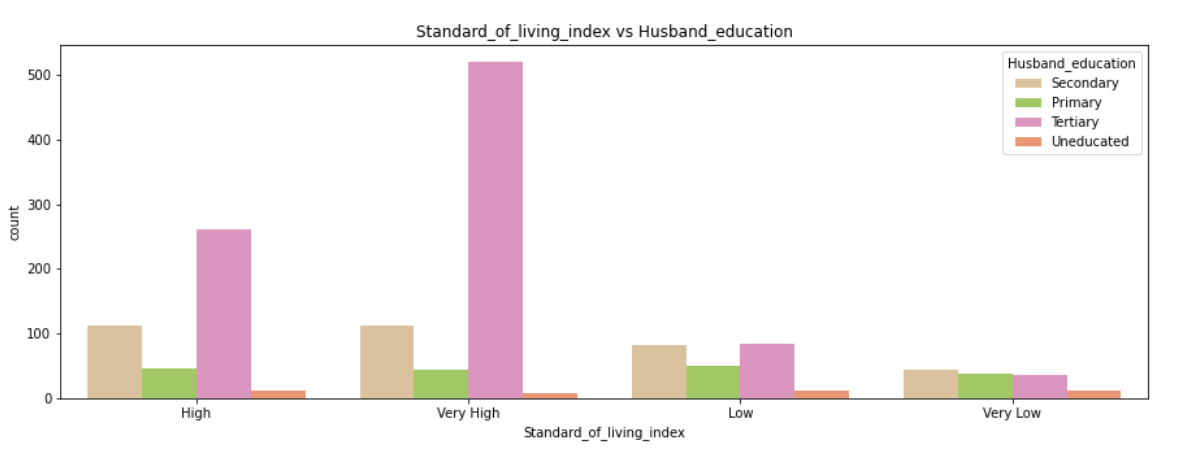
People are preferring very high living standards.

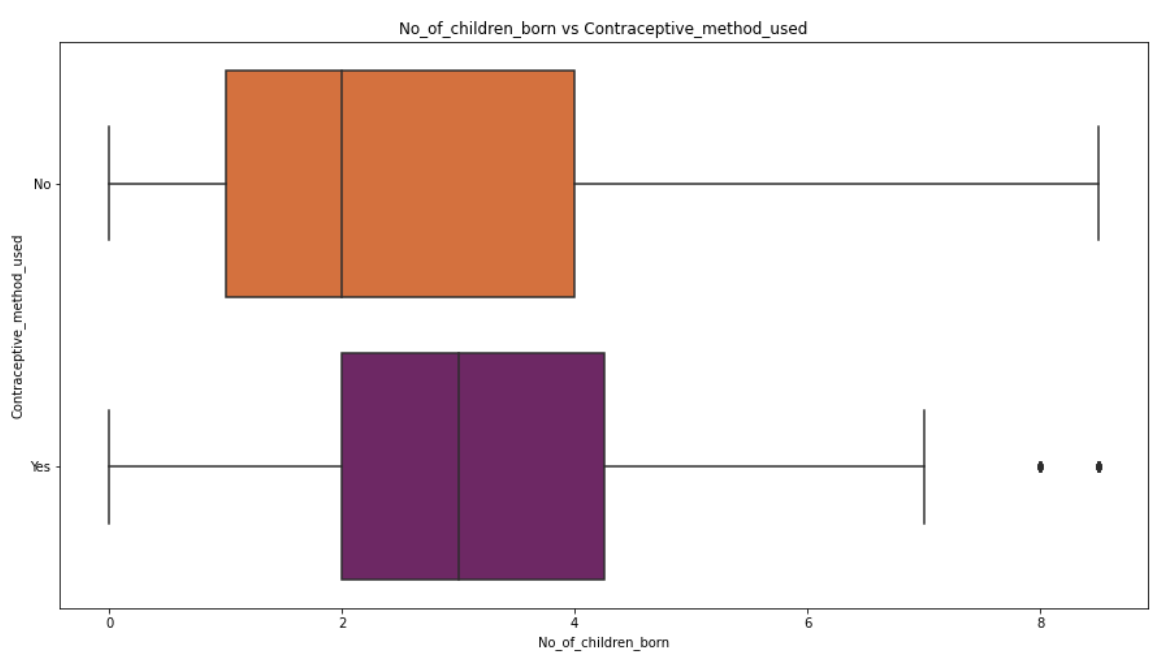
**Bivariate Analysis:**

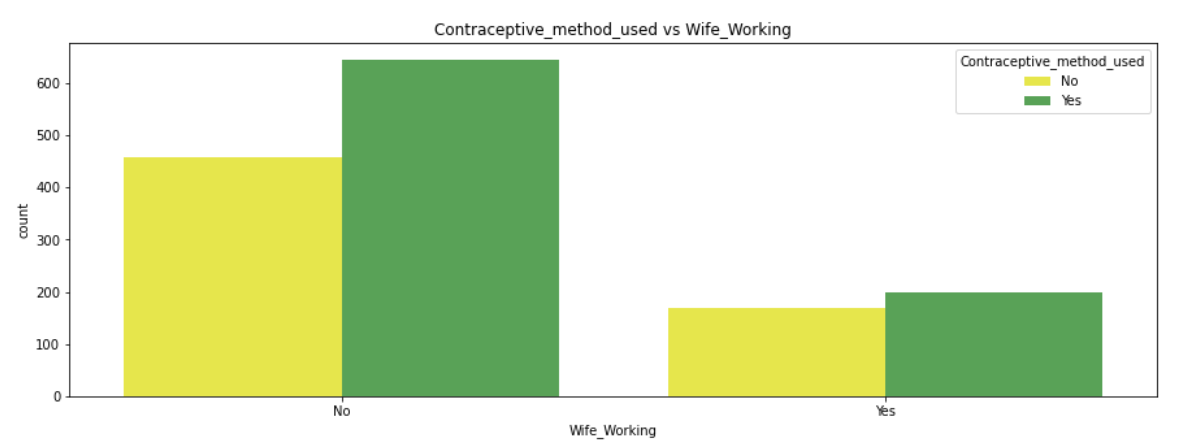
Figure 15: Bivariate analysis between variables

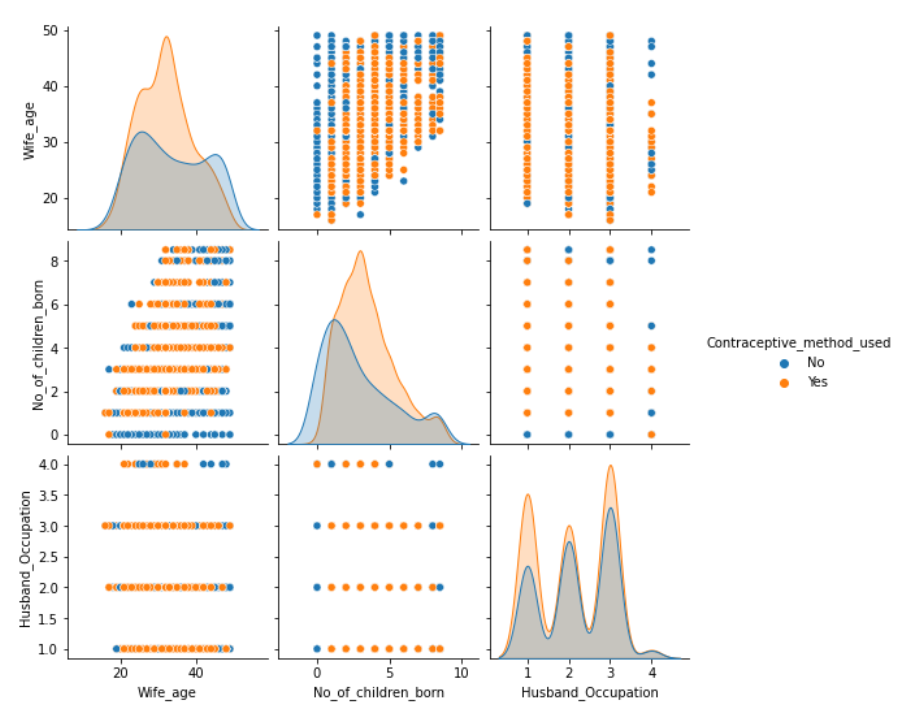












**Analysis:**

1. For husband’s occupation and standard of living we used violinplot the width determines the strength.
2. People with high and very high standard of living use contraceptive method.
3. High and very high standard of living people have high standards in studying.
4. People with more number of children are more likely to use contraceptive methods.
5. Working wives are more likely to use contraceptive methods.
6. As wife age increases, they are more likely to use contraceptive methods.

**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.**

## Logistic Regression

## Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.

The variable you want to predict is called the dependent variable.

The variable you are using to predict the other variable's value is called the independent variable.

In this data set, according to the business need the dependent variable is ‘Contraceptive method used’ and all other remaining variables are dependent variables.

Let’s split the data into dependent and independent data. Here I used ‘x’ for the independent variables and ‘y’ for the dependent variable.

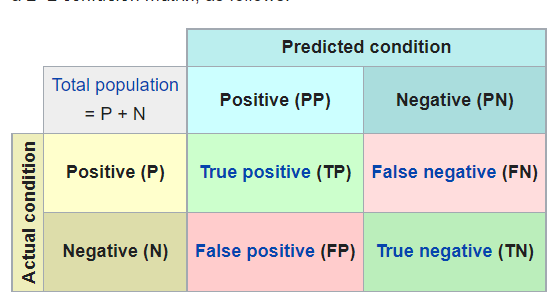
To train and test the data, we divide the data into two parts in the 70:30 ratio.

## 70% of the data is to train the model and 30% of the data is to test the model.

## Here the data must be divided as array, to perform logistic regression.

## After fitting the model, we calculate the precision, accuracy, recall for training and test data.

Figure 16: Confusion matrix



The above table has the following cases:

* True Negative: Model has given prediction No, and the real or actual value was also No.
* True Positive: The model has predicted yes, and the actual value was also true.
* False Negative: The model has predicted no, but the actual value was Yes, it is also called as Type-II error.
* False Positive: The model has predicted Yes, but the actual value was No. It is also called a Type-I error.

**Classification Accuracy**:

It defines how often the model predicts the correct output. It can be calculated as the ratio of the number of correct predictions made by the classifier to all number of predictions made by the classifiers. The formula is given below:  
Confusion Matrix in Machine Learning**Precision:**

It can be defined as the number of correct outputs provided by the model or out of all positive classes that have predicted correctly by the model.

How many of them were actually true. It can be calculated using the below formula:  
Confusion Matrix in Machine Learning**Recall:** It is defined as the out of total positive classes, how our model predicted correctly. The recall must be as high as possible.  
Confusion Matrix in Machine Learning

Table 17: Classification report for training data

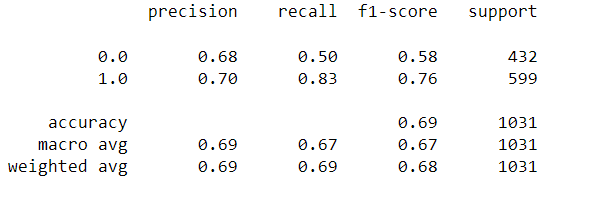
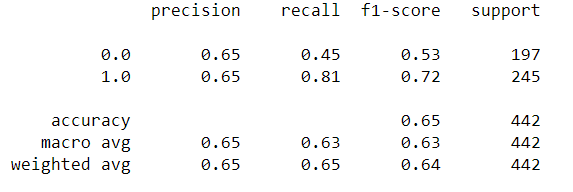


Table 18: Classification report for testing data



We could see all the parameters are good.

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Accuracy = (198 + 218) / (198+218+214+103) = 0.65

Precision = TP/(TP + FP)

Precision = 198 / (198 + 214) = 0.65

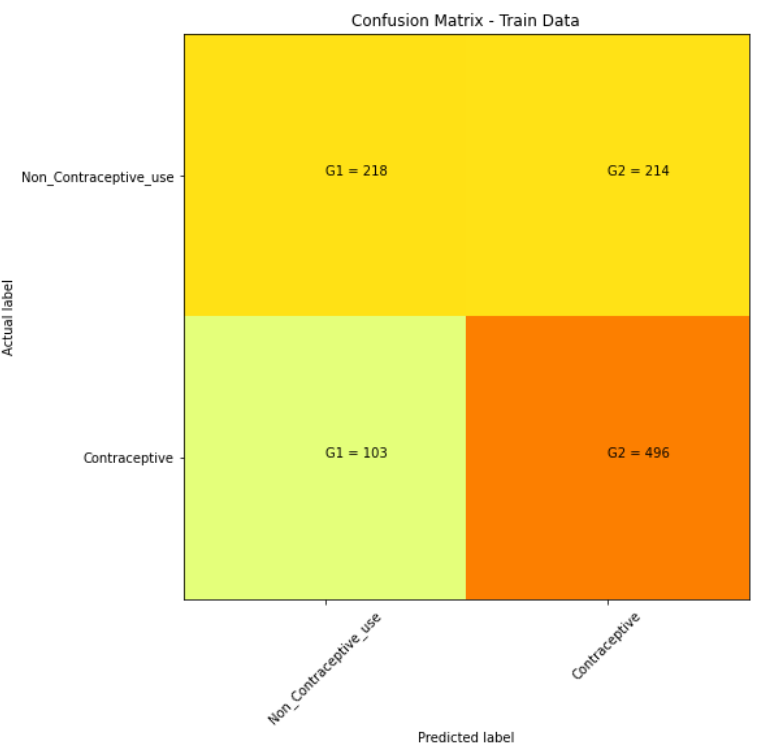
Recall = TP / (TP + FN)

Recall = 198 / (198 + 103) = 0.45

A confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm.

a table of confusion (sometimes also called a confusion matrix) is a table with two rows and two columns that reports the number of true positives, false negatives, false positives, and true negatives. This allows more detailed analysis than simply observing the proportion of correct classifications (accuracy). Accuracy will yield misleading results if the data set is unbalanced; that is, when the numbers of observations in different classes vary greatly.

Figure 16: Confusion matrix for training data



## Figure 17: Confusion matrix for testing data

## 

The confusion matrix analysis on training data:

* True Negative: Negative value which is correctly mapped. 198
* True Positive: Positive value which is correctly mapped. 218
* False Negative: Positive value, predicted as negative. 214
* False Positive: Negative value, predicted as positive.103

The confusion matrix analysis on testing data:

* True Negative: Negative value which is correctly mapped. 496
* True Positive: Positive value which is correctly mapped. 89
* False Negative: Positive value, predicted as negative. 108
* False Positive: Negative value, predicted as positive. 47

## Linear Discriminate Analysis:

LDA is a supervised classification technique that is considered a part of crafting competitive machine learning models. This category of dimensionality reduction is used in areas like image recognition and predictive analysis in marketing

The variable you want to predict is called the dependent variable.

The variable you are using to predict the other variable's value is called the independent variable.

In this data set, according to the business need the dependent variable is ‘Contraceptive method used’ and all other remaining variables are dependent variables.

Let’s split the data into dependent and independent data. Here I used ‘x’ for the independent variables and ‘y’ for the dependent variable.

To train and test the data, we divide the data into two parts in the 70:30 ratio.

## 70% of the data is to train the model and 30% of the data is to test the model.

## To perform LDA we need import few required libraries.

## Now fit the data to the model.

## The coefficients are as follows:

## Table 19: Coefficients of LDA model

## 

The intercept of the data is 1.57889

Linear Discriminate Analysis equation:

Contraceptive\_method\_used = -0.09 x Wife\_age + 0.38 x No\_of\_children\_born + 0.09 x Husband\_Occupation + 0.63 x Wife\_ education\_Secondary + 1.54 x Wife\_ education\_Tertiary + -0.38 x Wife\_ education\_Uneducated + 0.17 x Husband\_education\_Secondary + -0.14 x Husband\_education\_Tertiary + -0.39 x Husband\_education\_Uneducated + -0.47 x Wife\_religion\_Scientology + -0.13 x Wife\_Working\_Yes + -0.04 x Standard\_of\_living\_index\_Low + 0.3 x Standard\_of\_living\_index\_Very High + -0.4 x Standard\_of\_living\_index\_Very Low + -0.25 x Media\_exposure \_Not-Exposed

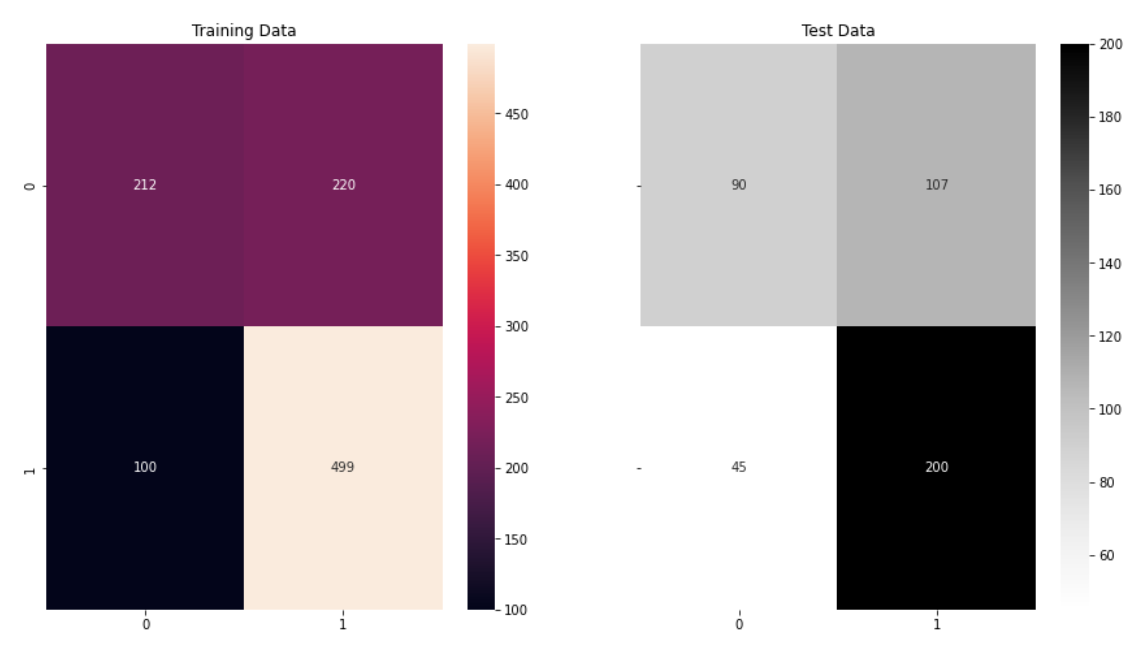
**By the above equation and the coefficients, it is clear that**

predictor Wife\_ education\_Tertiary has the largest magnitude; thus this helps in classifying the best.

predictor Wife\_religion\_Scientology has the smallest magnitude; thus this helps in classifying the least.

Confusion matrix:

Figure 18: Confusion matrices of LDA



The confusion matrix analysis on training data:

* True Negative: Negative value which is correctly mapped. 499
* True Positive: Positive value which is correctly mapped. 212
* False Negative: Positive value, predicted as negative. 220
* False Positive: Negative value, predicted as positive.100

The confusion matrix analysis on testing data:

* True Negative: Negative value which is correctly mapped. 200
* True Positive: Positive value which is correctly mapped. 90
* False Negative: Positive value, predicted as negative. 107
* False Positive: Negative value, predicted as positive.45

**Classification of Training data:**

Table 20: Classification Report of training data

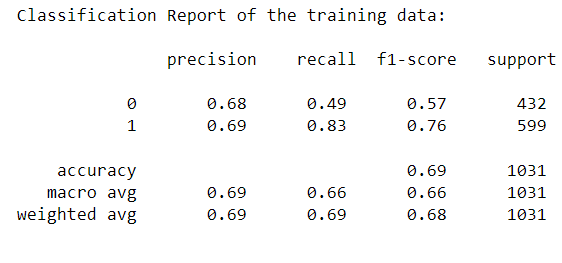


Table 21: Classification Report of testing data

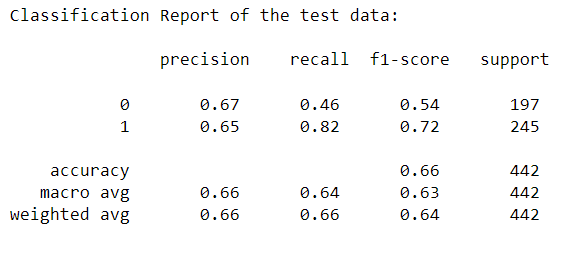
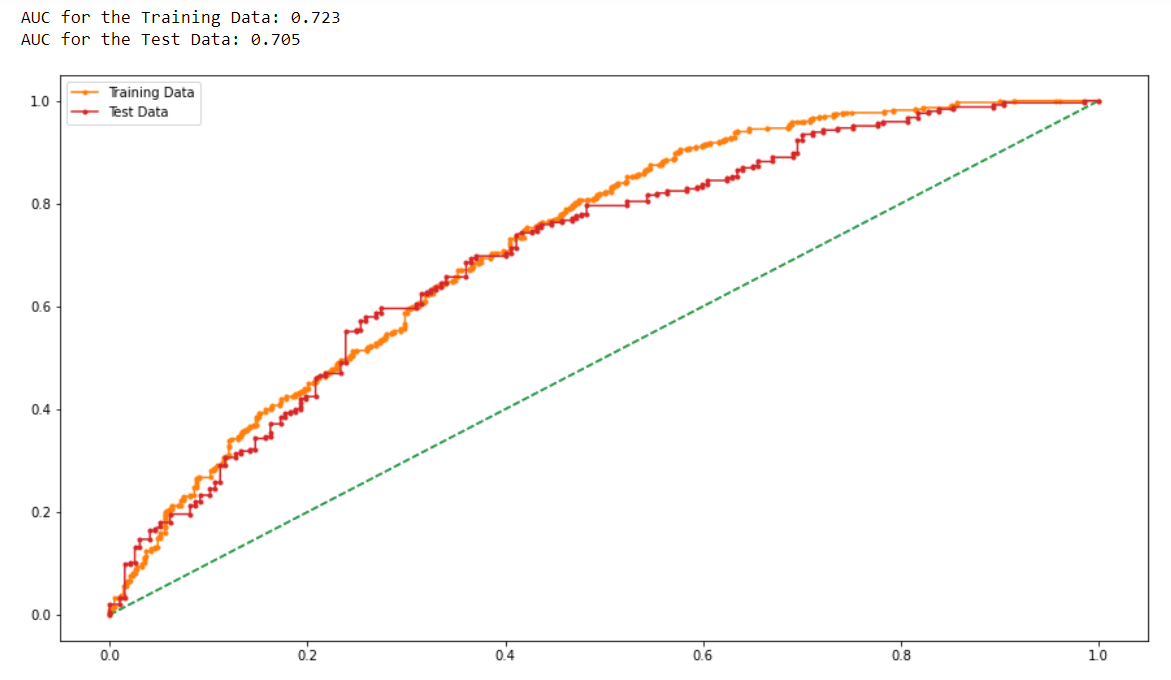


Figure 19: ROC of LDA



ROC stands for Receiver Operating Characteristic curve. This is a graph that shows the performance plotting the true positive rate and the false positive rate.

AUC stands for Area Under the Curve. It is used to measure the entire area under the ROC curve.

Here AUC for train and test data is 72% and 70% respectively which is good in this case.

Almost 70% of the data is covered by this model.

**CART Algorithm:**

CART (Classification and Regression Trees) can be used for both classification and regression problems. The difference lies in the target variable:

With classification, we attempt to predict a class label. In other words, classification is used for problems where the output (target variable) takes a finite set of values

Meanwhile, regression is used to predict a numerical label. This means your output can take an infinite set of values

The variable you want to predict is called the dependent variable.

The variable you are using to predict the other variable's value is called the independent variable.

In this data set, according to the business need the dependent variable is ‘Contraceptive method used’ and all other remaining variables are dependent variables.

Let’s split the data into dependent and independent data. Here I used ‘x’ for the independent variables and ‘y’ for the dependent variable.

To train and test the data, we divide the data into two parts in the 70:30 ratio.

70% of the data is to train the model and 30% of the data is to test the model.

To perform the CART Algorithm we need few packages which are to be imported.

After dividing the data, we fit this data to the algorithm with the classifier as gini.

Gini Index is use as the classifier to classify the data at every stage.

There are three commonly used impurity measures used in binary decision trees: Entropy, Gini index, and Classification Error.

Entropy (a way to measure impurity):

Entropy=−∑jpjlog2pjEntropy=−∑jpjlog2⁡pj

Gini index (a criterion to minimize the probability of misclassification):

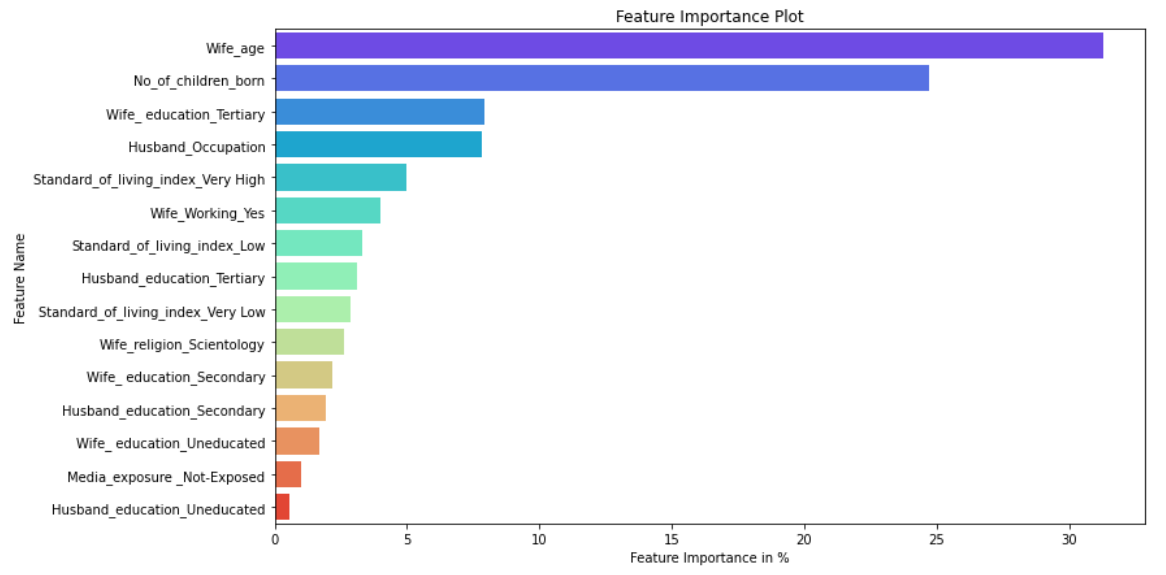
Gini=1−∑jp2j

We fit the data and upload the data into a file and check the tree.

The above code will save a .dot file in your working directory.  
WebGraphviz is Graphviz in the Browser.  
Copy paste the contents of the file into the link below to get the visualization  
<http://webgraphviz.com/>

The dot file is ld\_Tress\_File.

Figure 20: Feature Importance without regularizing



Regularize the data:

We regularize the data as following:

Max depth = 30,

Min samples leaf=50,

Min samples split=100.

After regularizing the data, the feature importance is as follows.

Figure 21: Feature Importance with regularizing

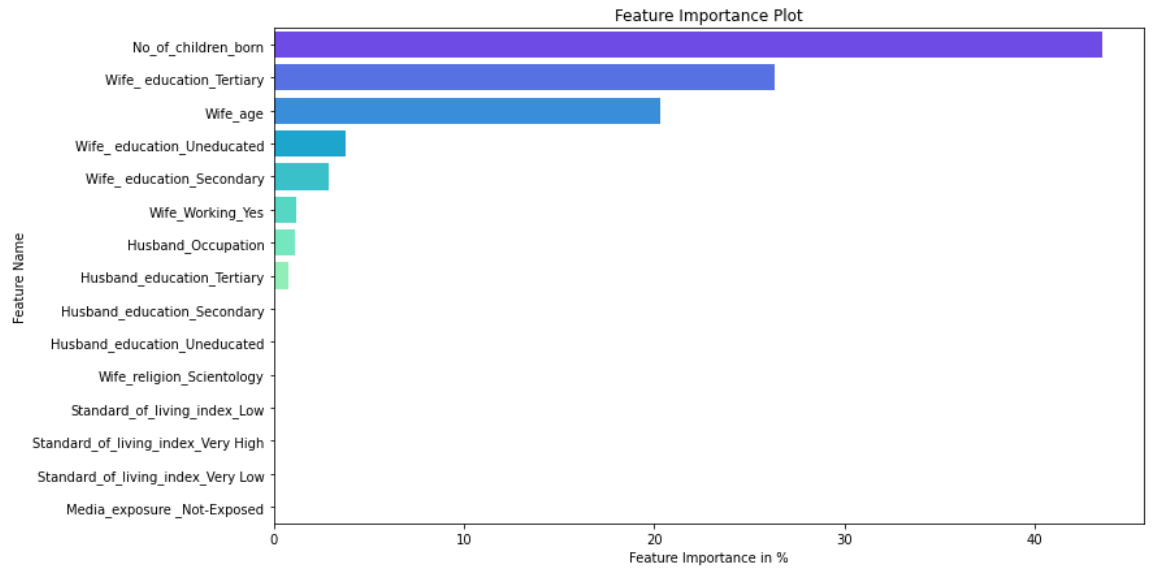
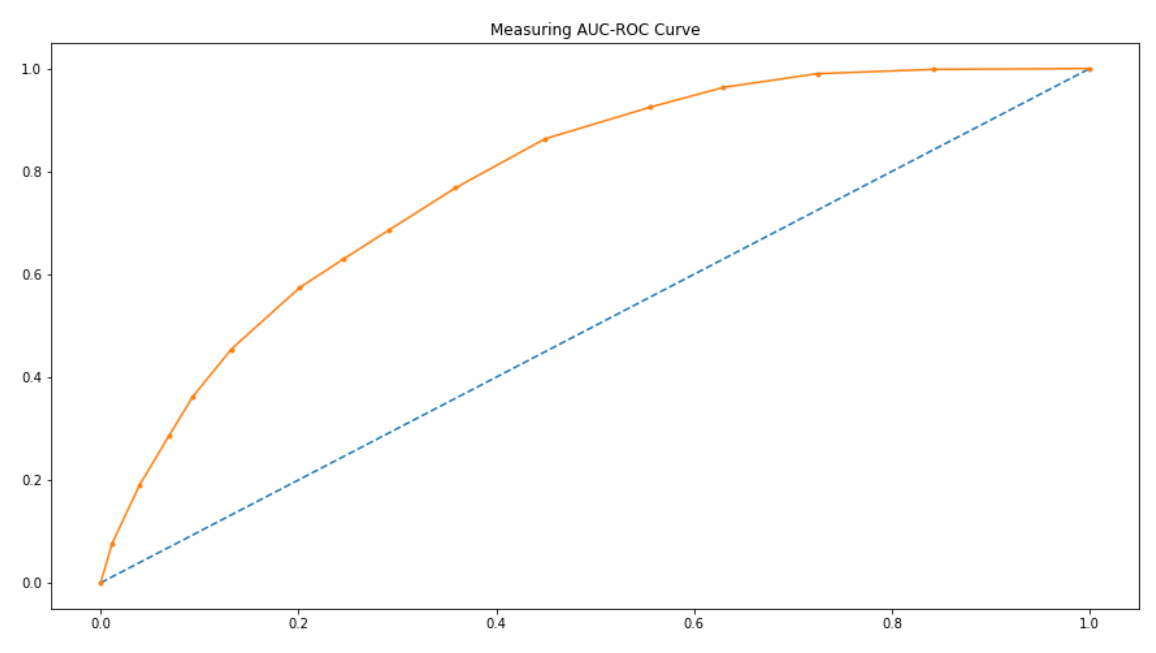
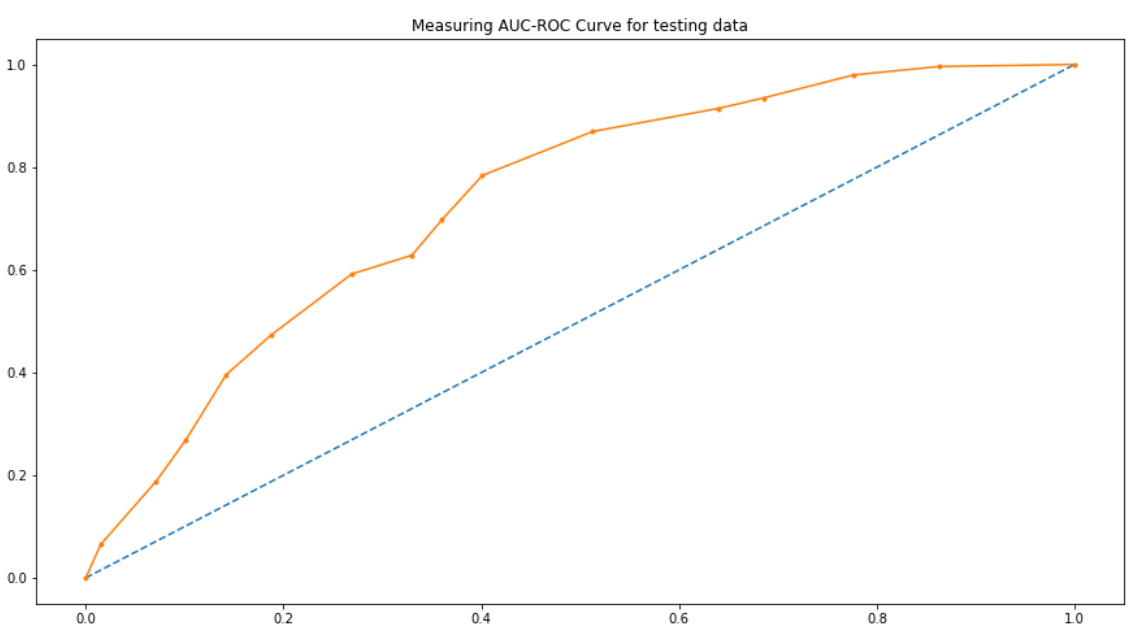


Figure 22: Measuring AUC-ROC Curve for training data



AUC for training data is 0.78

Figure 23: Measuring AUC-ROC Curve for testing data



AUC for training data is 0.73

Figure 24: Confusion Matrix for training data

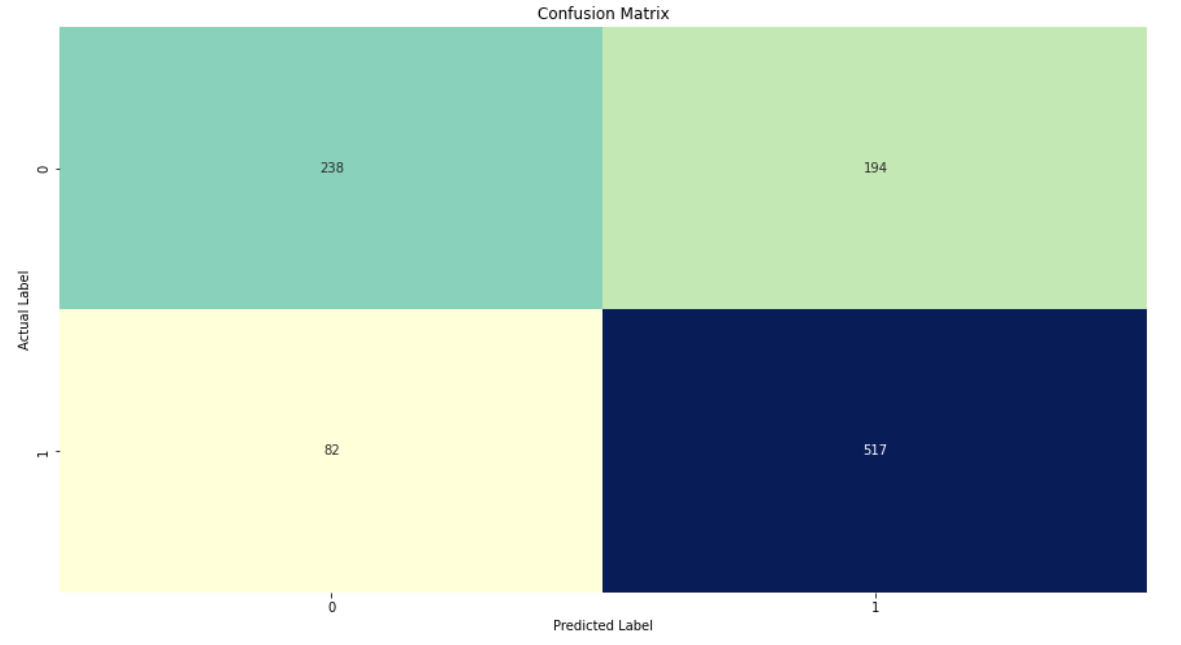
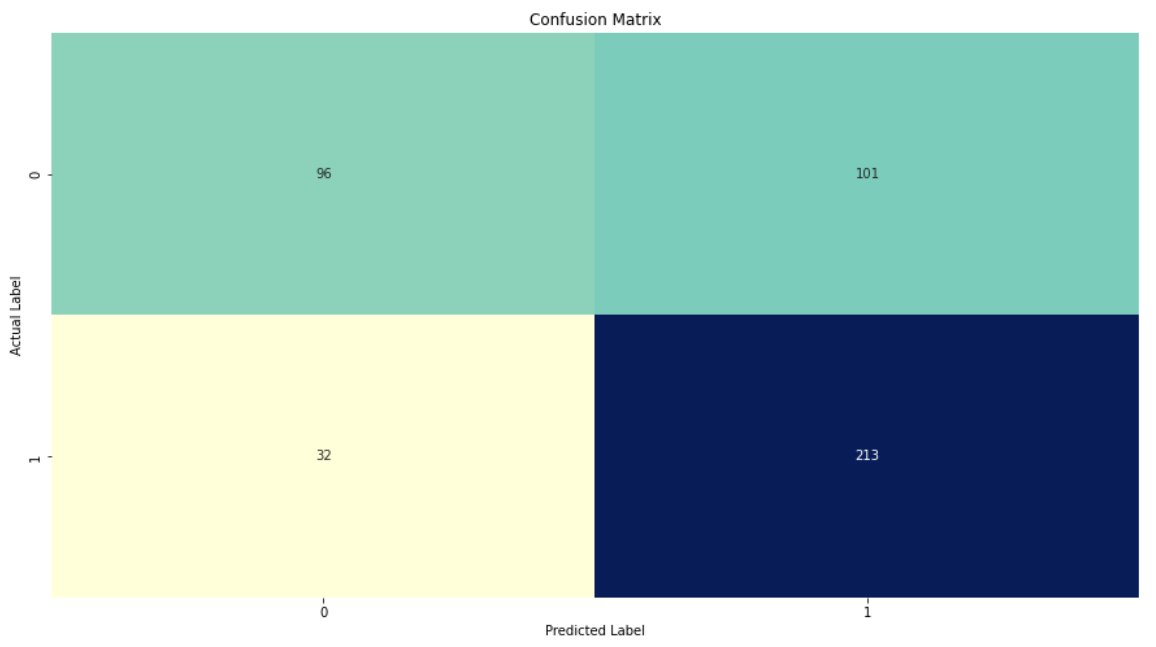


Figure 25: Confusion Matrix for training data



The confusion matrix analysis on training data:

* True Negative: Negative value which is correctly mapped. 517
* True Positive: Positive value which is correctly mapped. 238
* False Negative: Positive value, predicted as negative. 194
* False Positive: Negative value, predicted as positive.84

The confusion matrix analysis on testing data:

* True Negative: Negative value which is correctly mapped. 213
* True Positive: Positive value which is correctly mapped. 96
* False Negative: Positive value, predicted as negative. 101
* False Positive: Negative value, predicted as positive. 32

Table 23: Classification Report of training data

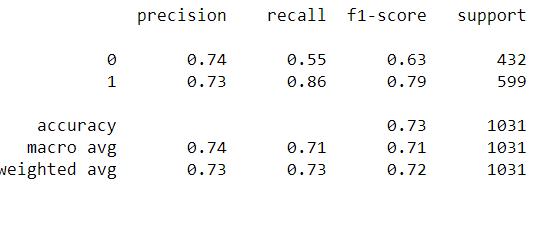
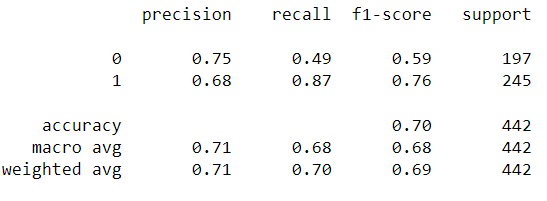


Table 24: Classification Report of testing data



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

From all the models used:

Accuracy gained by Logical regression: 65%

Accuracy gained by LDA: 69%

Accuracy gained by CART algorithm: 70%

AUC for train data by LDA - 72%

AUC for test data by LDA -70%

AUC for train data by CART- 78%

AUC for test data by CART - 73%

When compared to logistic regression, LDA, CART algorithms CART is best model among them.

The accuracy is high for CART algorithm.

By comparing precision, recall also declares CART algorithm as the best model among others to this data.

Number of children plays a major role in this data.

Next comes wife age and wife education.

Contraceptive use depends majorly on the Number of children, wife age and wife education tertiary.