Predictive modelling

Aisha Khan

Contents

[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. 4](#_Toc126536247)

[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. 18](#_Toc126536248)

[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. 18](#_Toc126536249)

[4. Inference: Basis on these predictions, what are the business insights and recommendations. 23](#_Toc126536250)

[5. 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. 25](#_Toc126536251)

[6. 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. 31](#_Toc126536252)

[7. 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. 35](#_Toc126536253)

[8. Inference: Basis on these predictions, what are the insights and recommendations. 36](#_Toc126536254)

**Table of Figures**

[Fig. 1. Missing values 6](#_Toc126536263)

[Fig. 2. Univariate analysis graphs 16](#_Toc126536264)

[Fig. 3. Bivariate Analysis Graphs 17](file:///C:\Users\anamk\Downloads\Predictive%20Modelling.docx#_Toc126536265)

[Fig. 4. Normal Distribution of Residual 20](file:///C:\Users\anamk\Downloads\Predictive%20Modelling.docx#_Toc126536266)

[Fig. 5. Normal distribution of residuals with transformed data 22](file:///C:\Users\anamk\Downloads\Predictive%20Modelling.docx#_Toc126536267)

[Fig. 6. Univariate analysis 28](file:///C:\Users\anamk\Downloads\Predictive%20Modelling.docx#_Toc126536268)

[Fig. 7. Multivariate analysis 30](file:///C:\Users\anamk\Downloads\Predictive%20Modelling.docx#_Toc126536269)

[Fig. 8. Pair plot 31](file:///C:\Users\anamk\Downloads\Predictive%20Modelling.docx#_Toc126536270)

[Fig. 9. Confusion matrix for Logistic Regression 32](file:///C:\Users\anamk\Downloads\Predictive%20Modelling.docx#_Toc126536271)

[Fig. 10. Sample decision tree diagram. 34](file:///C:\Users\anamk\Downloads\Predictive%20Modelling.docx#_Toc126536272)

**List of Tables**

[Table 1. Data types 4](#_Toc126536273)

[Table 2. First 5 rows 4](file:///C:\Users\anamk\Downloads\Predictive%20Modelling.docx#_Toc126536274)

[Table 3. Last 5 rows 4](file:///C:\Users\anamk\Downloads\Predictive%20Modelling.docx#_Toc126536275)

[Table 4. Summary of the Data 5](#_Toc126536276)

[Table 5. Linear regression model with original data 19](file:///C:\Users\anamk\Downloads\Predictive%20Modelling.docx#_Toc126536277)

[Table 6. Regressional model with transformed data. 21](#_Toc126536278)

[Table 7. Data types 25](#_Toc126536279)

[Table 8. First 5 rows of the data 26](#_Toc126536280)

[Table 9. Last five rows of the data 27](#_Toc126536281)

[Table 10. Summary of the data 27](#_Toc126536282)

[Table 11. Classification of LDA model results for Train data 32](#_Toc126536283)

[Table 12. Classification of LDA model results for Test data 33](#_Toc126536284)

[Table 13. Confusion matrix for LDA 33](#_Toc126536285)

**LINEAR REGRESSION**

**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

# 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.

The required packages were loaded, work directory was set and data was loaded.

Dataset has 8192 rows and 22 features with below data types bifurcation:

|  |  |
| --- | --- |
| **Data Type** | **Count of Columns** |
| float64 | 2 |
| int64 | 8 |
| object | 13 |
| **Grand Total** | **22** |

Table . Data types

Data Exploration was performed using the following functions:

1. Head
2. Tail
3. Shape
4. Summary
5. Check Duplicates
6. Null Values

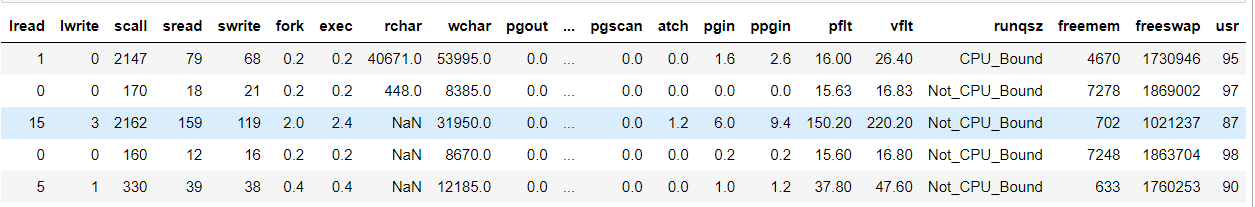
1.Head

Table . First 5 rows

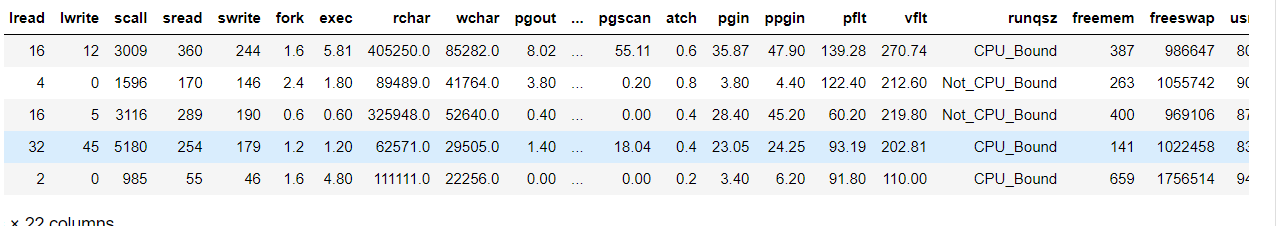
2.Tail

Table . Last 5 rows

1. Shape.

Dataset has 8192 rows and 22 features

1. Summary**:**

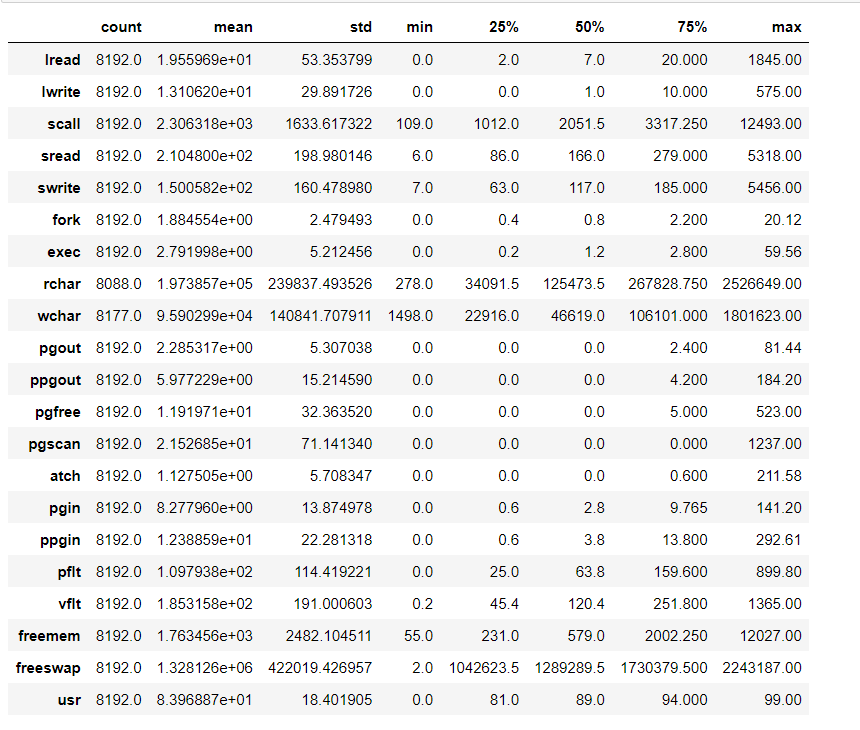
****

Table . Summary of the Data

5.Check Duplicates:

No Duplicates were observed in the dataset.

6.Null Values:

On checking for missing entries/null values it was observed that 2 numeric variables includes null values which are:

* **Rchar**
* **Wchar**

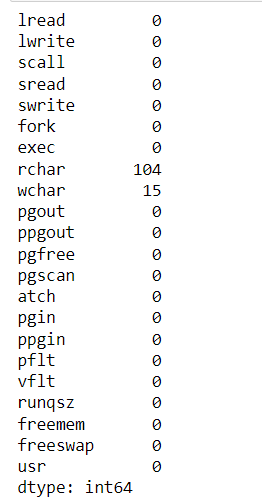
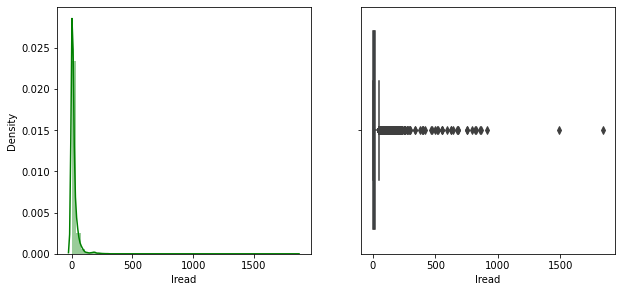
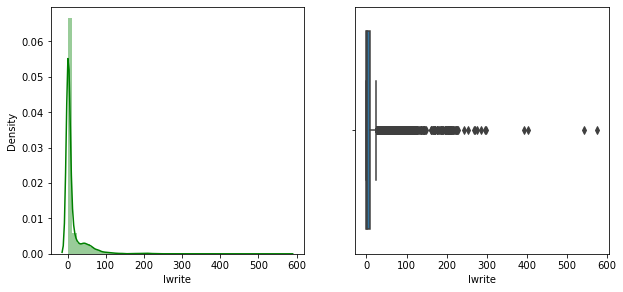


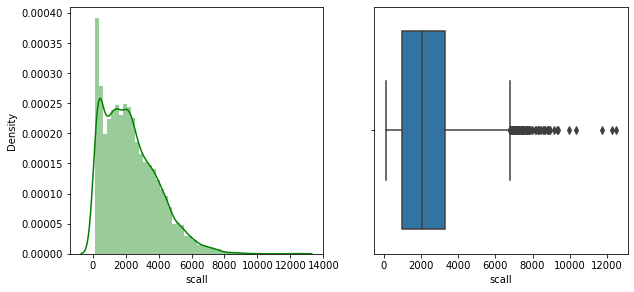
Fig. . Missing values

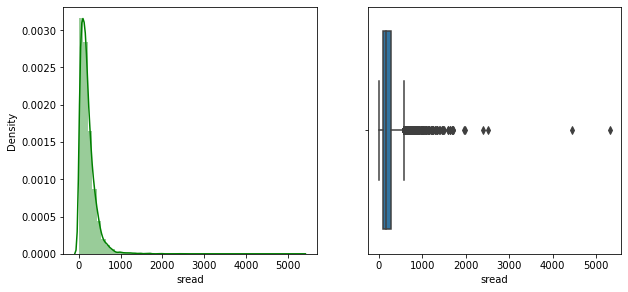
The missing values were treated by calculating its mean values.

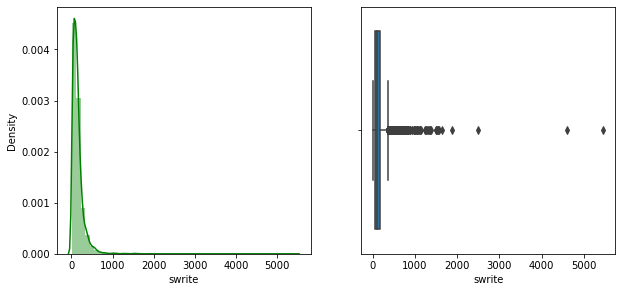
**UNIVARIATE ANALYSIS:**

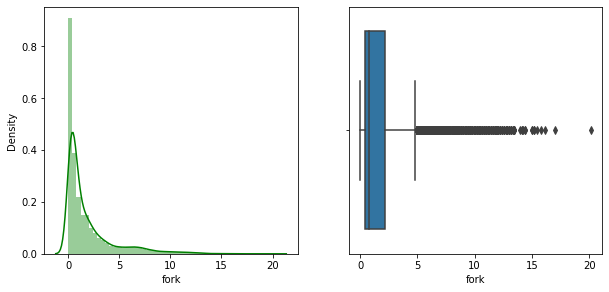


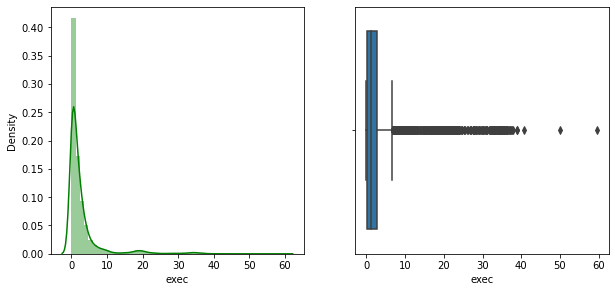


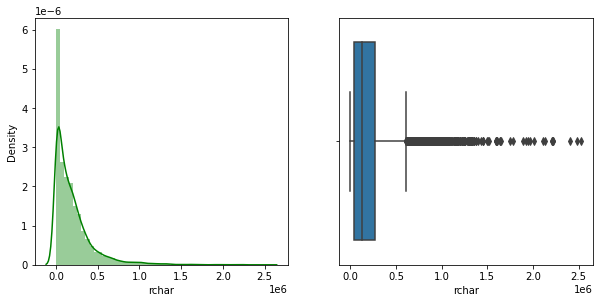


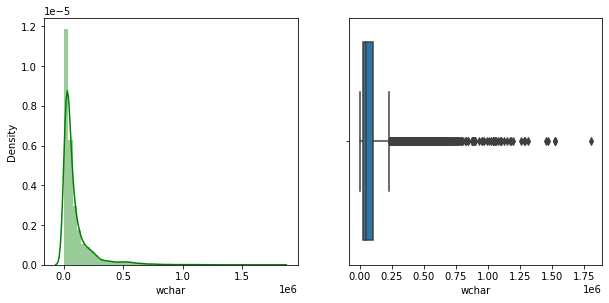


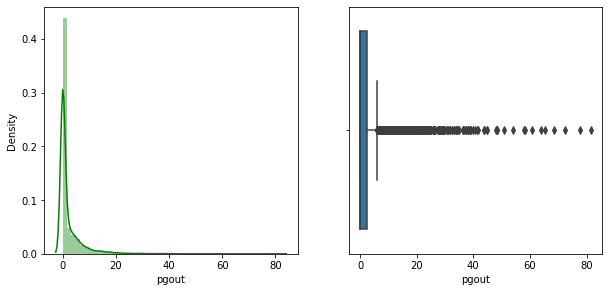


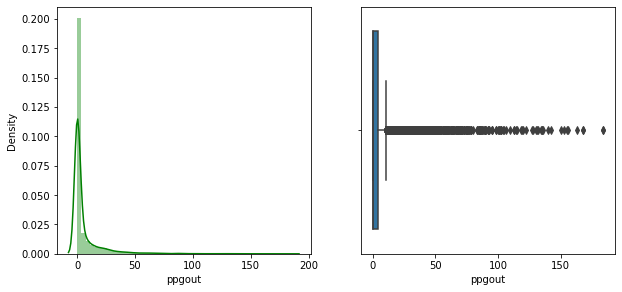


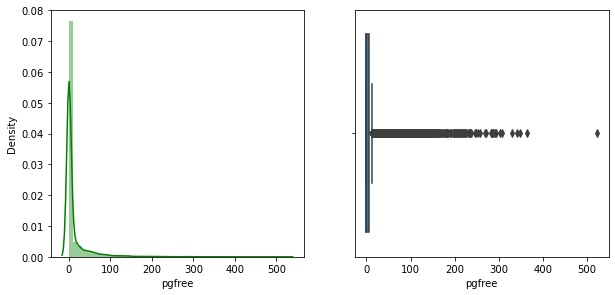


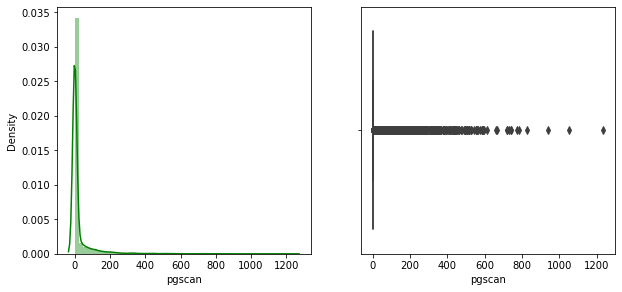


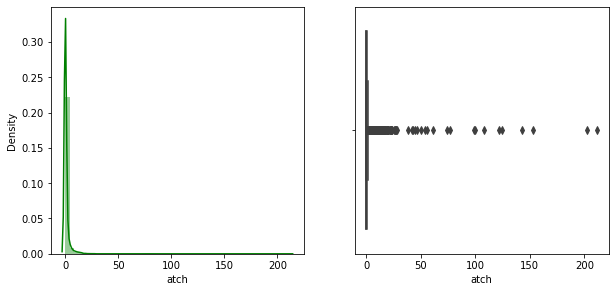


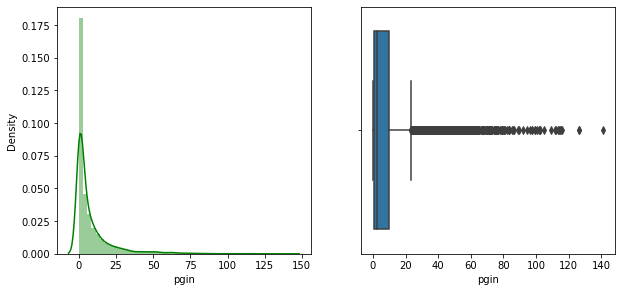


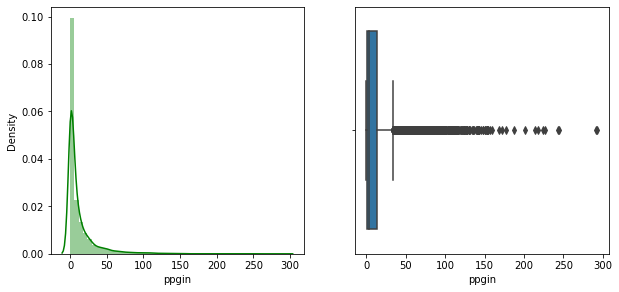


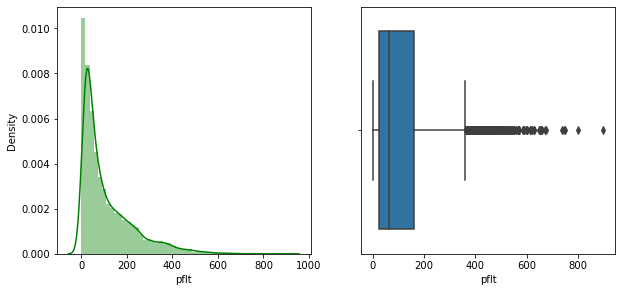


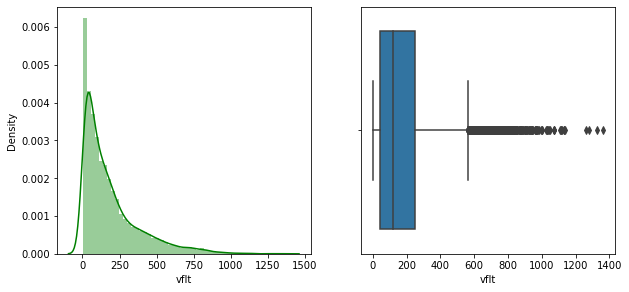


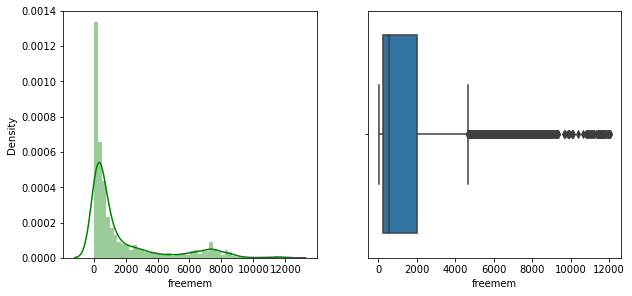


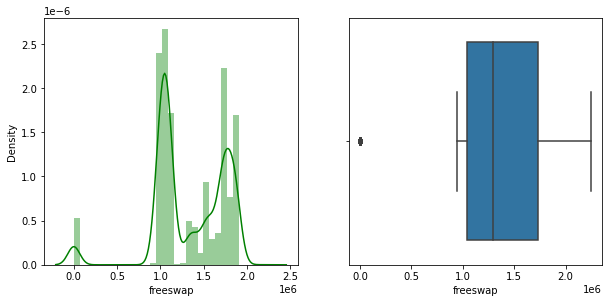


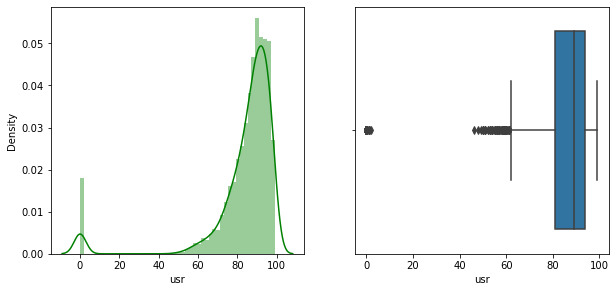












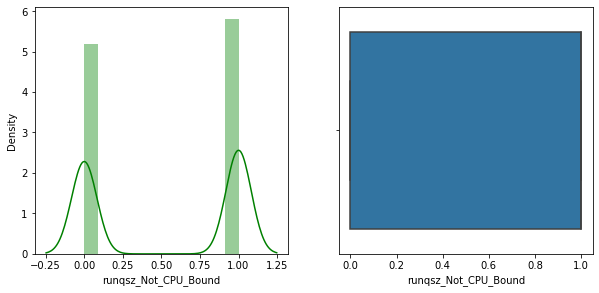


Fig. . Univariate analysis graphs

From the above graphs we can say that no variable is normally distributed.

Also it is observed that variables lread,lwrite,scall,sread,swrite,fork,exec,rchar,wchar,pgout,ppgout,pgscan,atch,pgin,ppgin,pfit,vflt,freemem & usr has outliers which were further treated by capping due to their effect on the further model building.

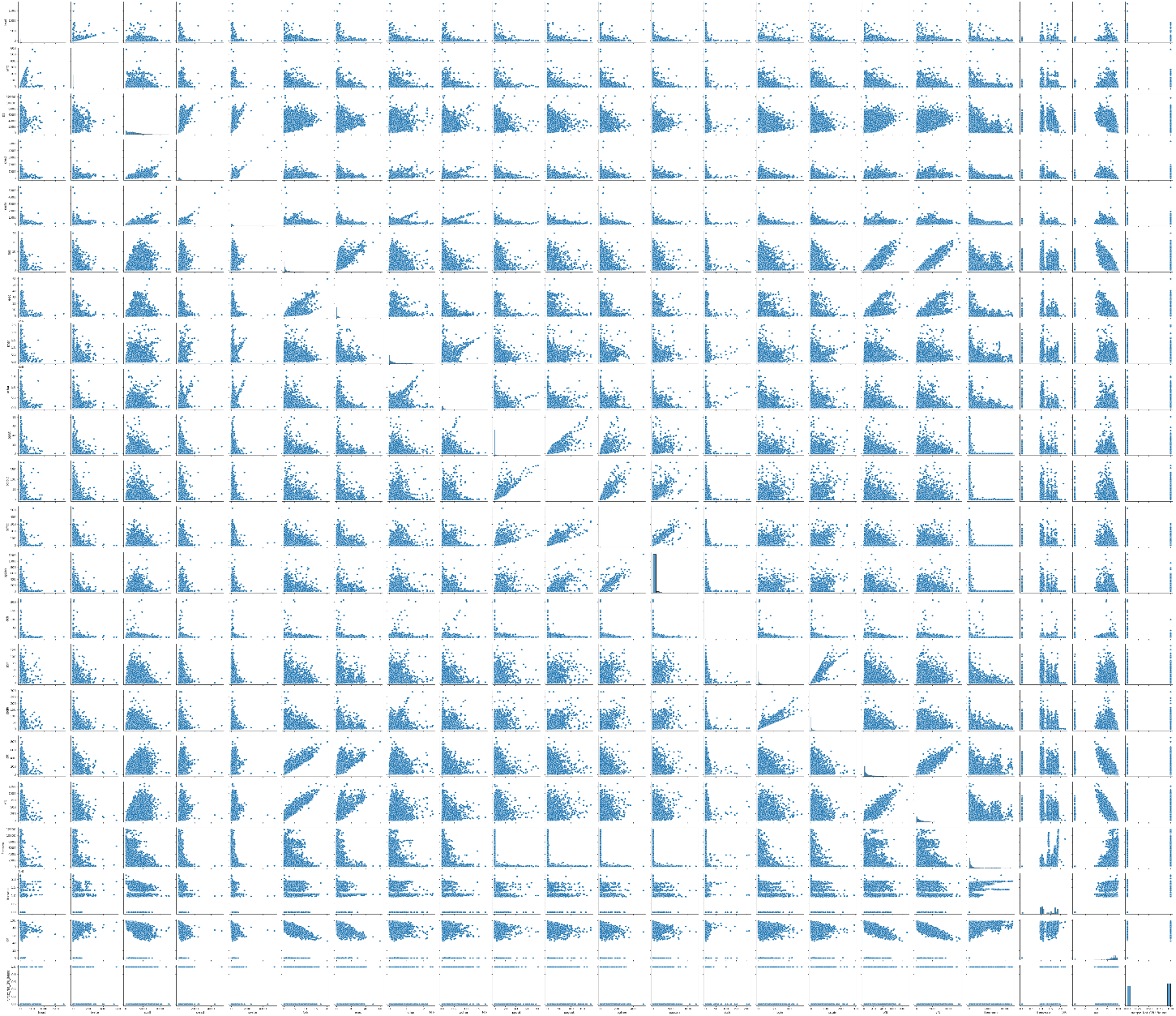
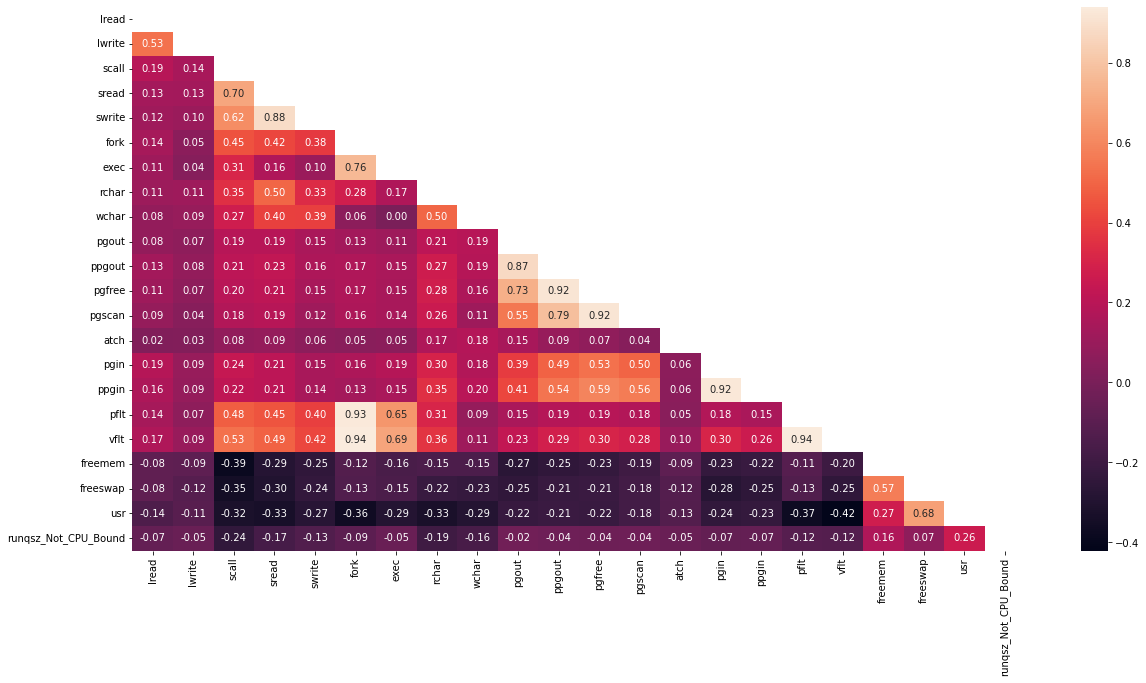
**BIVARIATE ANALYSIS:**

Fig. . Bivariate Analysis Graphs

**Correlation between the dependent variable with independents variables are being observed with the heat map.**

# 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.

Null values were observed in the data which were further treated using mean as the variable with missing values was a continuous variable.

Values which are equal to 0 do have significance in the data, hence no 0’s were removed.

Outliers were present in the below variables which were treated by capping.

Variables with outliers:

lread,lwrite,scall,sread,swrite,fork,exec,rchar,wchar,pgout,ppgout,pgscan,atch,pgin,ppgin,pfit,vflt,freemem & usr

Duplicate values were not observed in the dataset.

# 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.

Data encoding was done by using the getdummies function with which all object datatypes were converted to categorical data.

Data was split into train & test with 70% & 30% spread respectively.

On performing transformation on the original dataset it was observed that the transformed data was better in performance compared to the original data.

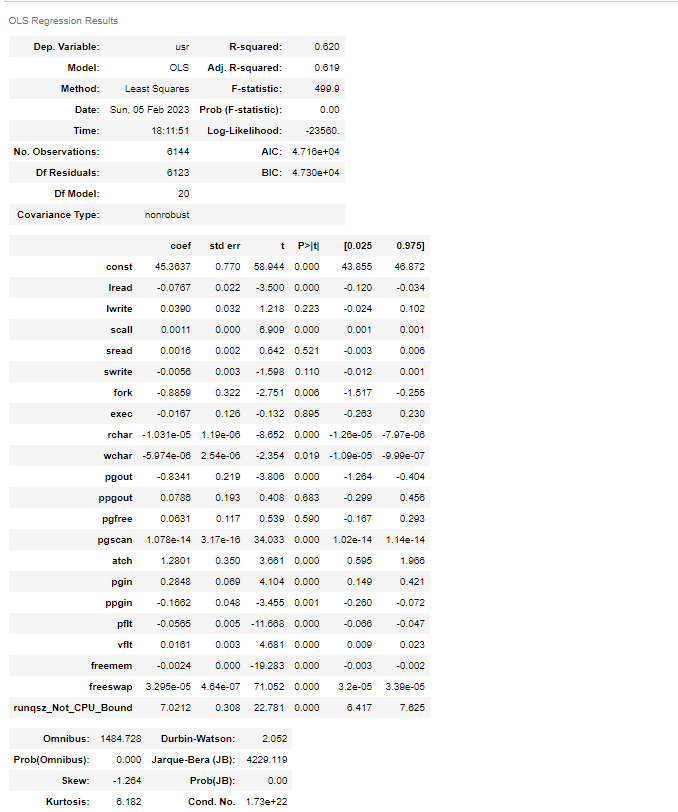
****

Table . Linear regression model with original data

**R2 value for = 0.62**

**From t**he above tables it is observed that:

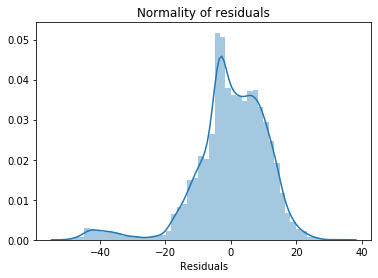
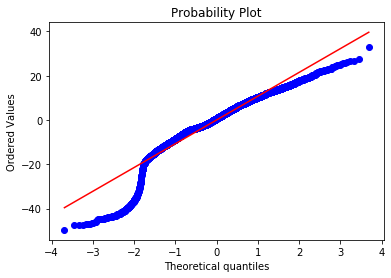
R2 value: 0.62(62%) which is low for a model to be processed further.

Fig. . Normal Distribution of Residual

The residuals for the original data are not normally distributed which can be observed in the above graphs. Hence transformation is needed for the better model delivery.

After transformation it is observed that:

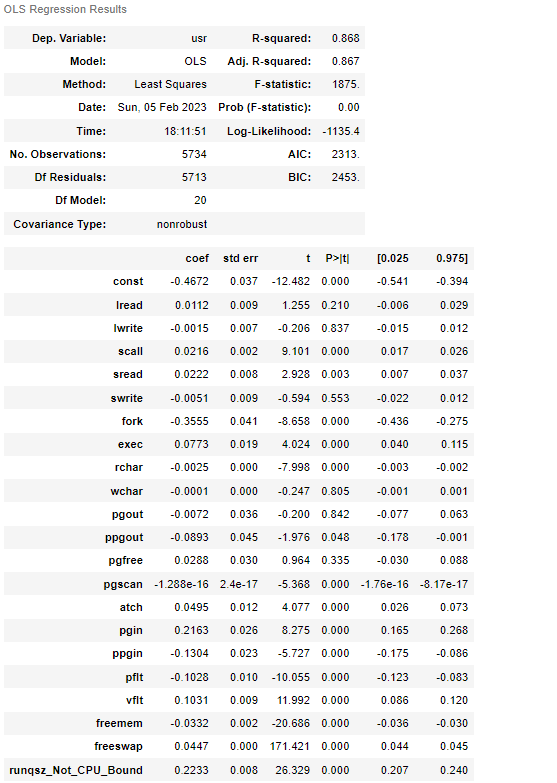


Table . Regressional model with transformed data.

R2 = 0.868 (for the transformed data)

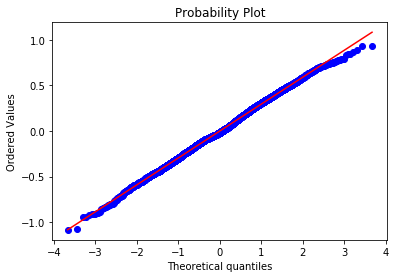
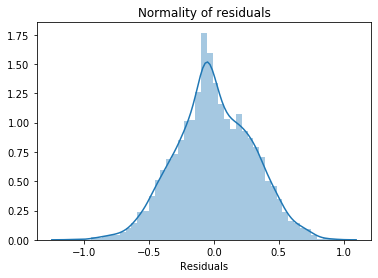
****

Fig. . Normal distribution of residuals with transformed data

The residuals for the transformed data are now normally distributed which can be observed in the above graphs. The R2 values has increased to 0.867(87%) which states that the transformed model is delivering better outcomes.

As transformation required for the given model, will run the model on test data too for more inferences.

Let us observe the different values of train and test from the model on transformed data,

R2 for Train data = 0.867

R2 for Test data = 0.88

Adj. R-squared for Train data: 0.866

Adj. R-squared for Test data: 0.879

RMSE value for Train data : 0.29

RMSE value for Test data : 0.31

R2, Adj. R2, RMSE value for train and test are nearly equal to each other.

The regression model with original data have R2 value is 0.62 (approx.). But for the transformed data model R2  value is 0.867.

From this it is observed that percentage(%) variability of usr (Portion of time (%) that CPUs run in user mode) with all the independence variable is more in Transformed data model than the original data model.

So, we can consider the regression model made with Transformed data as the apt model for this data set.

Therefore, the linear regression equation for the given data set is,

usr = **-0.44630643867097636 + 0.005574039028878019 \* ( lwrite ) + 0.024552830539764835 \* ( scall ) + 0.00995061092301841 \* ( swrite ) + ( -0.34625960283487744) \* ( fork ) + 0.07282623310597805 \* ( exec ) + (-0.0023597025694529336) \* ( rchar ) + (-0.08211331397359337 )\* ( pgout ) + 0.0506282519769673 \* ( atch ) + 0.06934455145127787 \* ( pgin ) + -0.10390486913144725 \* ( pflt ) + 0.10668668134488296 \* ( vflt ) + -0.03406748500220984 \* ( freemem ) + 0.04457112055308339 \* ( freeswap ) + ( -0.03322907797377543) \* ( runqsz\_Not\_CPU\_Bound )**

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

Based on the model results, the attributes that are mostly affects the computer systems are

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

vflt - Number of page faults caused by address translation

**scall –** Number of system calls of all types 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

**pgout –** Number of page out requests 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

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

**freemem –** Number of memory pages available to user processes

**freeswap –** Number of disk blocks available for page swapping.

**runqsz-** Process run queue size (The number of kernel threads in memory that are waiting for a CPU to run.

Company should concentrate on these variables to get better portion of time that CPU runs in user mode.

* EDA process to deep dive in the data and understand each variable significance and the correlation with each other
* Data cleaning by imputing null values, treating outliers to avoid any biasness in the data
* Encoding the string values to satisfy the data requirement while running the model. (numerical values/ categorical acceptable for running the models)
* Application of linear regression model by using scikit learn for both original data and transformed data.
* After model fit it is concluded that transformed model is the best model using R square, RMSE, Adj R square values.
* Dimensionality reduction has been taken by using VIF values.
* The best fit line was drawn with the linear equation.

**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

# 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.

The required packages were loaded, work directory was set and data was loaded.

Dataset has 1473 rows and 10 features with below data types bifurcation:

|  |  |
| --- | --- |
| **Data Type** | **Count of Columns** |
| float64 | 2 |
| int64 | 1 |
| object | 7 |
| **Grand Total** | **10** |
|  |  |

Table . Data types

Data Exploration was performed using the following functions:

1. Head
2. Tail
3. Shape
4. Summary
5. Check Duplicates
6. Null Values

**1.Head:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Wife\_age** | **Wife\_education** | **Husband\_education** | **No\_of\_children\_born** | **Wife\_religion** | **Wife\_Working** | **Husband\_Occupation** | **Standard\_of\_living\_index** | **Media\_exposure** | **Contraceptive\_method\_used** |
| 24 | Primary | Secondary | 3 | Scientology | No | 2 | High | Exposed | No |
| 45 | Uneducated | Secondary | 10 | Scientology | No | 3 | Very High | Exposed | No |
| 43 | Primary | Secondary | 7 | Scientology | No | 3 | Very High | Exposed | No |
| 42 | Secondary | Primary | 9 | Scientology | No | 3 | High | Exposed | No |
| 36 | Secondary | Secondary | 8 | Scientology | No | 3 | Low | Exposed | No |

Table . First 5 rows of the data

**2.Tail:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Wife\_age** | **Wife\_education** | **Husband\_education** | **No\_of\_children\_born** | **Wife\_religion** | **Wife\_Working** | **Husband\_Occupation** | **Standard\_of\_living\_index** | **Media\_exposure** | **Contraceptive\_method\_used** |
| 33 | Tertiary | Tertiary | NaN | Scientology | Yes | 2 | Very High | Exposed | Yes |
| 33 | Tertiary | Tertiary | NaN | Scientology | No | 1 | Very High | Exposed | Yes |
| 39 | Secondary | Secondary | NaN | Scientology | Yes | 1 | Very High | Exposed | Yes |
| 33 | Secondary | Secondary | NaN | Scientology | Yes | 2 | Low | Exposed | Yes |
| 17 | Secondary | Secondary | 1 | Scientology | No | 2 | Very High | Exposed | Yes |

Table . Last five rows of the data

**3. Shape:**

The dataset has 1473 rows and 10 variables.

4. Summary:

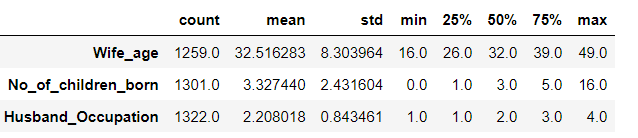
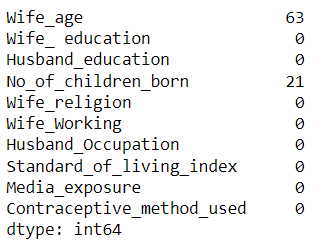


Table . Summary of the data

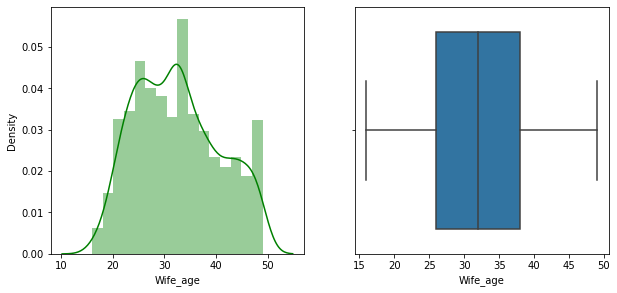
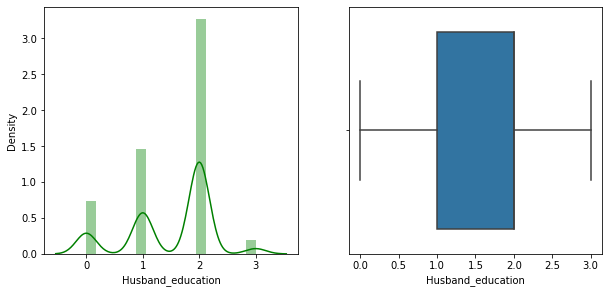
We can observe that the average age group of females is 32 with minimum age being 16 and maximum age with 49 years.

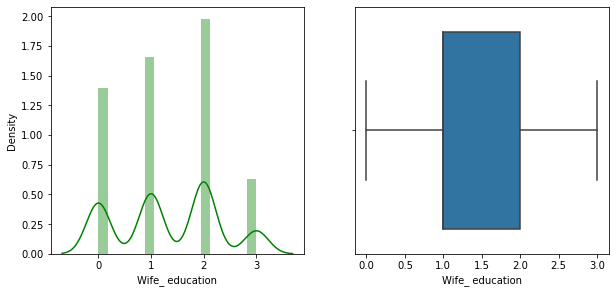
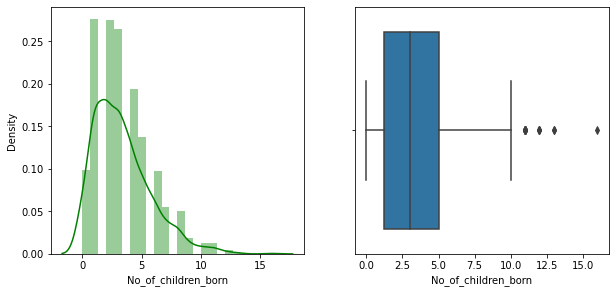
5. Checking Null Values:

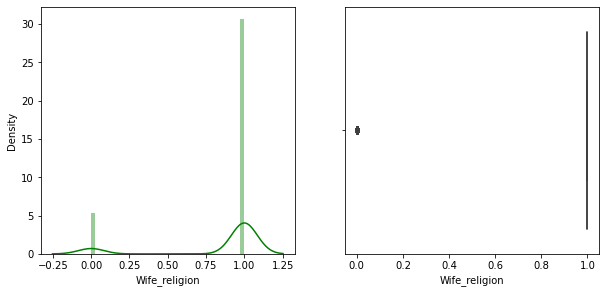
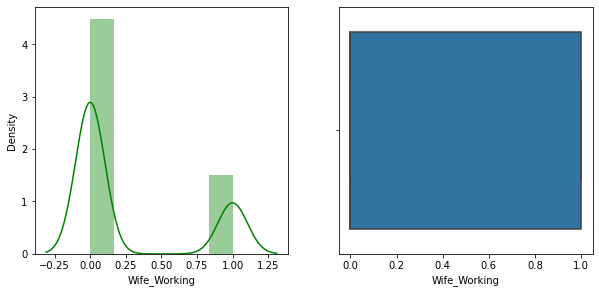
On checking for null values , it was observed that there are 2 variables where null value is found which are as follows:

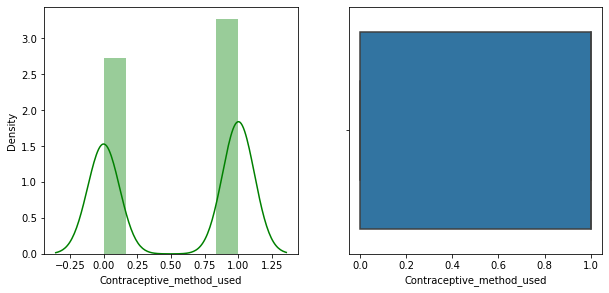


As both the parameters were numeric , null value treatment was done by calculating the mean of the variables.

**UNIVARIATE ANALYSIS:**







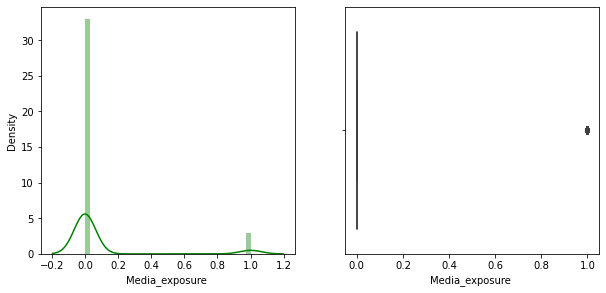
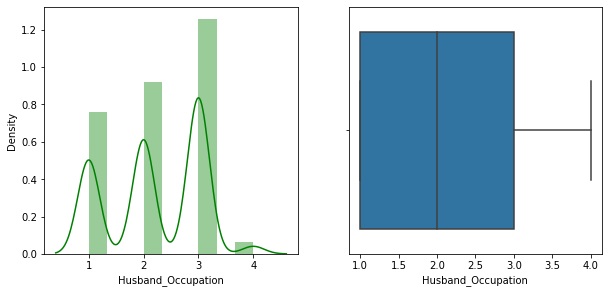
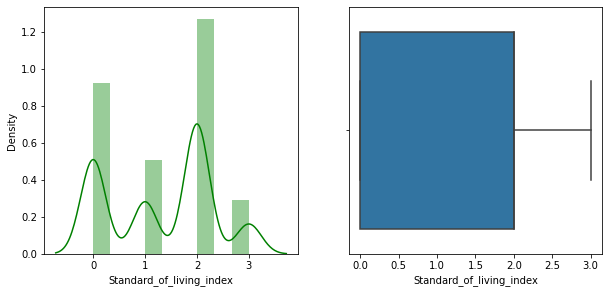


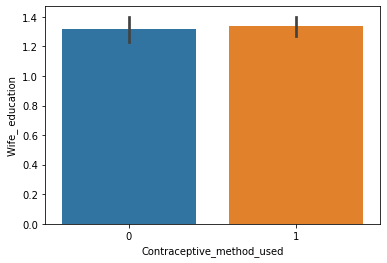
Fig. . Univariate analysis

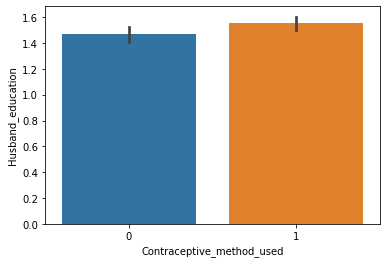
From the above graphs we can say that:

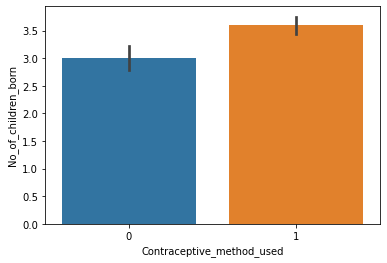
The dataset does not have normal distribution in any variable.

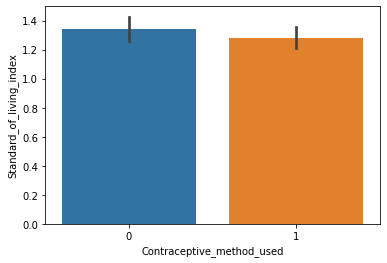
Outliers are observed in No of\_children\_born, wife\_religion, Media\_exposure.

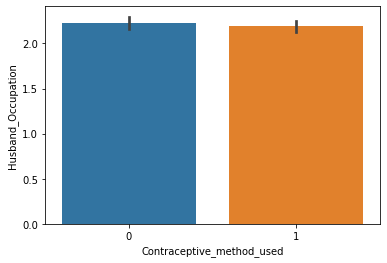
**MULTIVARIATE ANALYSIS:**

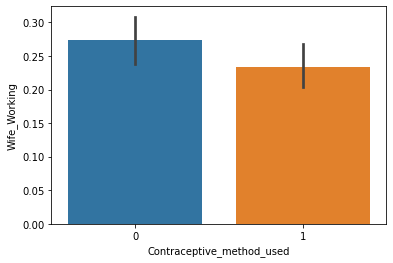


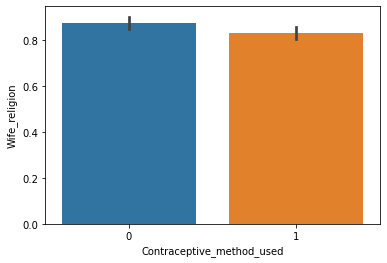
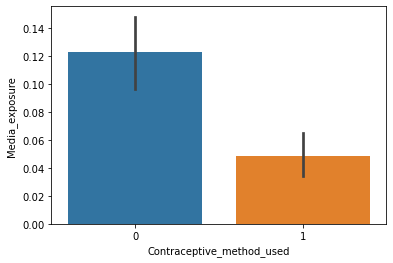












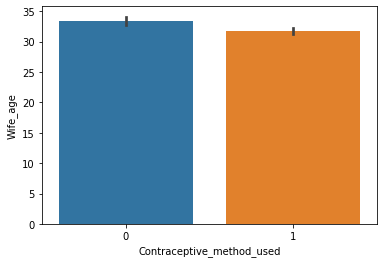


Fig. . Multivariate analysis

From the above graphs we can say that:

1. Age group of wife with contraceptive method used(Y/N) is between 30-35. Not much difference was observed in both the cases.

**BIVARIATE ANALYSIS:**

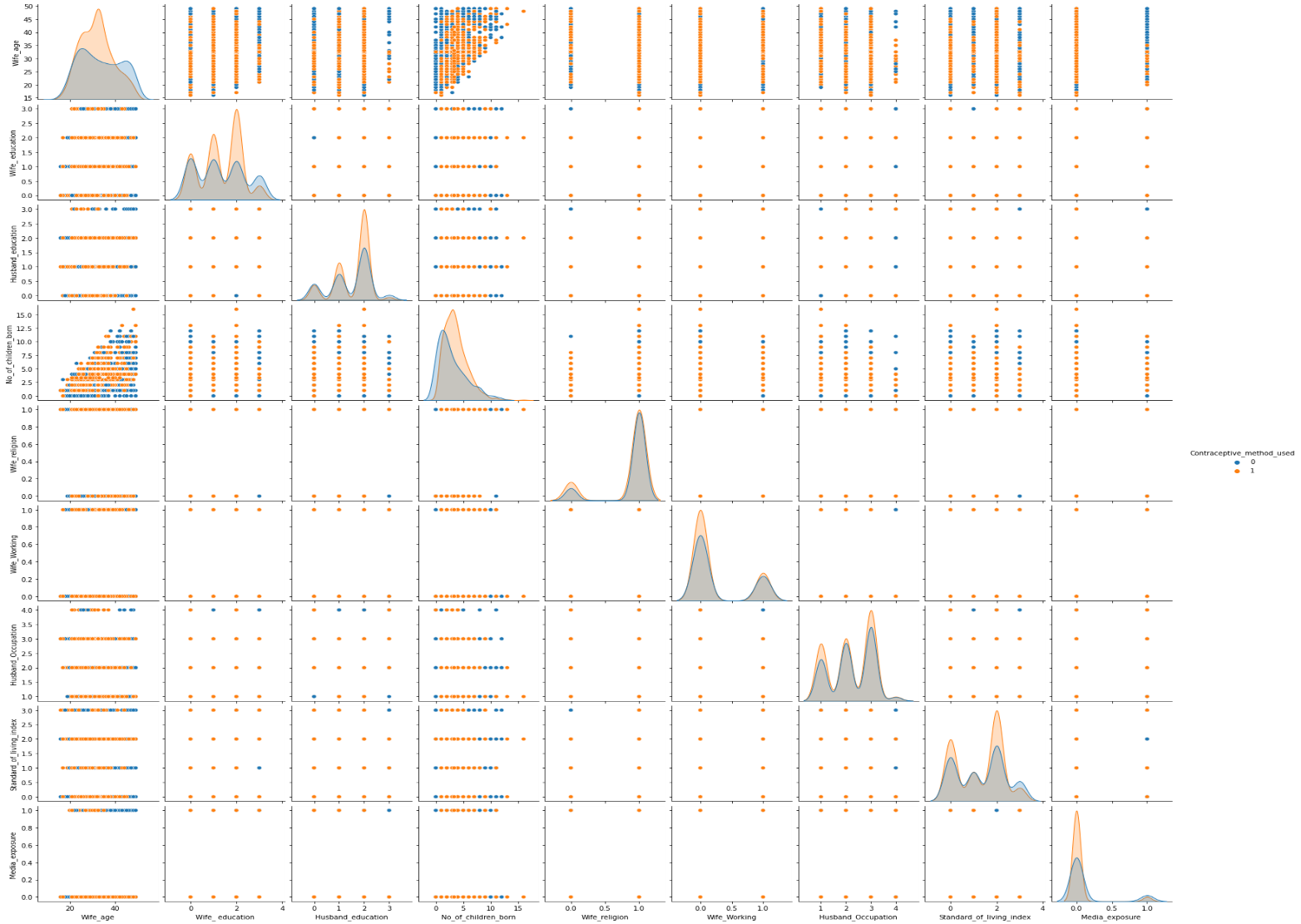


Fig. . Pair plot

# 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.

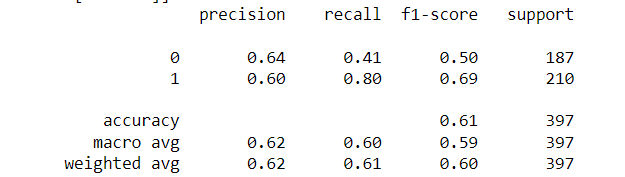
For the above query, as mentioned scaling was not performed.

For the variables that had object datatype encoding was performed to have categoric/continuous values for model building.

As defined in the above query, train & test data has been split up to 70% & 30% respectively.

On applying multiple models we found the following results:

**LOGISTIC REGRESSION:**

**Accuracy observed:** ~61%.

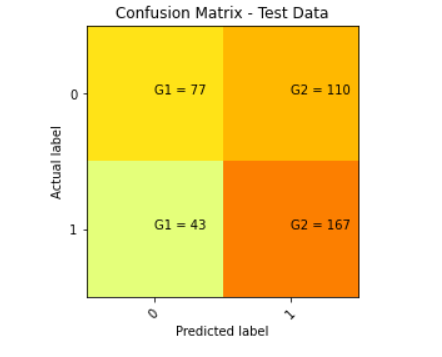


Fig. . Confusion matrix for Logistic Regression

**LDA:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
|  |  |  |  |  |
| 0 | 0.63 | 0.46 | 0.53 | 413 |
| 1 | 0.64 | 0.79 | 0.71 | 512 |
|  |  |  |  |  |
| accuracy |  |  | 0.64 | 925 |
| macroavg | 0.64 | 0.62 | 0.62 | 925 |
| weightedavg | 0.64 | 0.64 | 0.63 | 925 |

Table . Classification of LDA model results for Train data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
|  |  |  |  |  |
| 0 | 0.65 | 0.4 | 0.5 | 187 |
| 1 | 0.6 | 0.8 | 0.69 | 210 |
|  |  |  |  |  |
| accuracy |  |  | 0.61 | 397 |
| macroavg | 0.62 | 0.6 | 0.59 | 397 |
| weightedavg | 0.62 | 0.61 | 0.6 | 397 |

Table . Classification of LDA model results for Test data

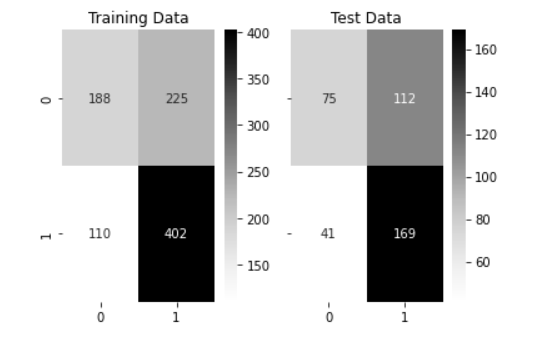
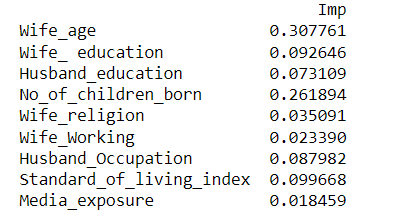


Table . Confusion matrix for LDA

**CART:**

****

**After Pruning:**

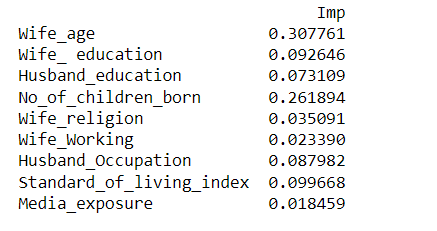
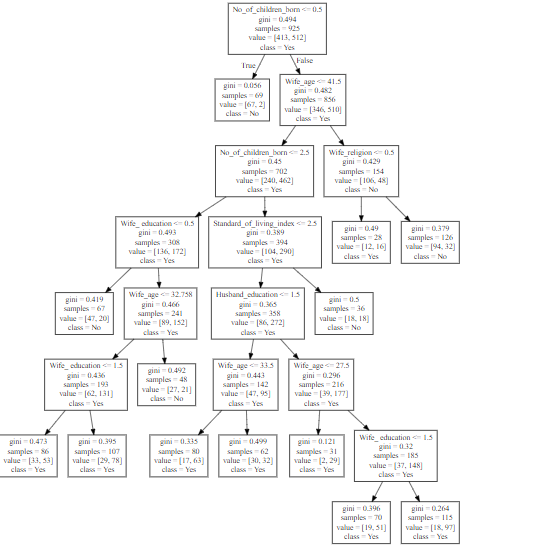
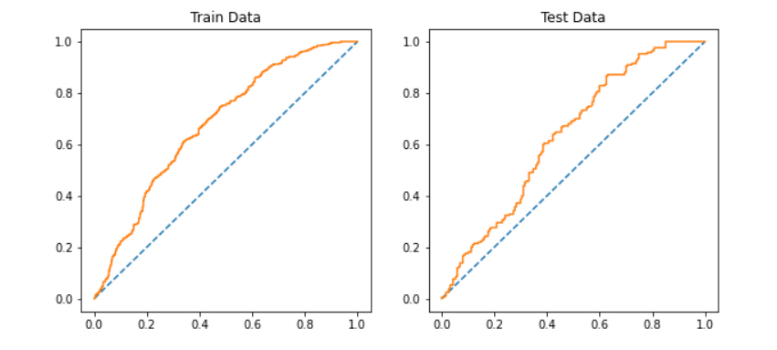
****

Fig. . Sample decision tree diagram.

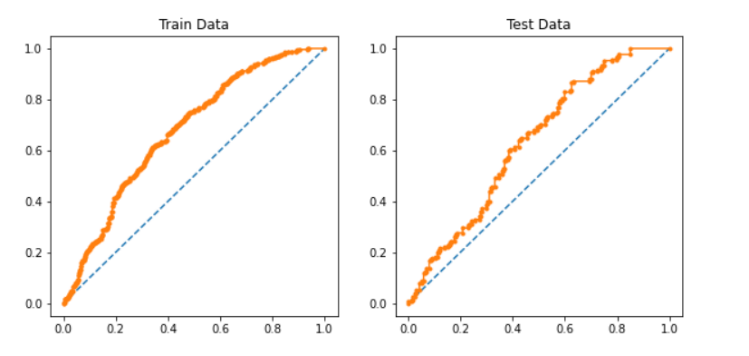
# Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

**AUC & ROC- LOGISTIC REGRESSION:**

****

**AUC LOGISTIC TRAIN DATA: 0.678**

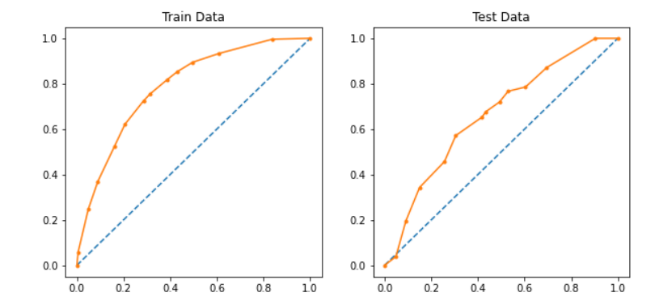
**AUC LOGISTIC TEST DATA: 0.632**

**AUC & ROC: LDA:**

**AUC CART TRAIN DATA: 0.678**

**AUC CART TEST DATA: 0.632**

**AUC & ROC: CART:**

****

**AUC CART TRAIN DATA: 0.788**

**AUC CART TEST DATA: 0.662**

On observing the values and graphs given above,

It is observed that the best fit for our data set is Decision tree-based model i.e.., CART

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

By implying multiple models on the data set we could analyse that the decision tree model was the best fit for the data set as it covered maximum area under the curve while performing ROC.

Also, the AUC of CART for decision tree model was the highest followed by LDA and Logistic regression.

To summarize the project we have done the following to get the best fit model for the dataset:

* EDA process to deep dive in the data and understand each variable significance and the correlation with each other
* Data cleaning by imputing null values, treating outliers to avoid any biasness in the data
* Encoding the string values to satisfy the data requirement while running the model. (numerical values/ categorical acceptable for running the models)
* Running various supervised learning models on the data set like Decision tree (CART), LDA, Logistic regression to attain the desired results.
* Based on the accuracy confirming the CART to be best model for the data set