**PREDICTIVE MODELLING**

**PROJECT**

**LINEAR REGRESSION**

**LOGISTIC REGRESSION**

**LDA & CART**

BUSINESS

REPORT

**Prepared by: AMAYIKA**

**Great Learning.**

## Email- Kashyap.amayika009@gmail.com

December 1st, 2022

Table of Contents

[Problem -1 – 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. . 6](#_bookmark0)

[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.. 8](#_bookmark1)

[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. 17](#_bookmark2)

[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.. 21](#_bookmark3)

[1.4 Inference: Basis on these predictions, what are the business insights and recommendations.. 30](#_bookmark4)

[Problem -2 Logistic Regression, LDA, 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..   32](#_bookmark18)

[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.. 32](#_bookmark21)

[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. 47](#_bookmark23)

[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.? 51](#_bookmark28)

[2.4 Inference: Basis on these predictions, what are the insights and recommendations. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.. 57](#_bookmark18)

List of Tables

[Table 1. Data set of comp-active databases (1-5th rows) 7](#_bookmark7)

[Table 2. Information of data set 8](#_bookmark8)

[Table 3. Value count of object data type variable 9](#_bookmark9)

[Table 4. Description of data set of company 9](#_bookmark10)

[Table 5. Missing values of the data set 9](#_bookmark13)

[Table 6. Missing values percentage in the data set 17](#_bookmark15)

[Table 7. Information after missing value imputation 17](#_bookmark16)

[Table 8. Data set rows after missing value imputation 18](#_bookmark17)

[Table 9. Description after missing values imputation 19](#_bookmark19)

[Table 10. Data set after encoding 22](#_bookmark20)

[Table 11. Data set after encoding continued 22](#_bookmark22)

[Table 12. Coefficients of variables 23](#_bookmark25)

[Table 13. OLS for Regression model before multicollinearity Treatment 24](#_bookmark7)

[Table 14. VIF Values for variables 28](#_bookmark8)

[Table 15. OLS for Regression model after multicollinearity Treatment 29](#_bookmark9)

[Table 16. VIF Values for variables after multicollinearity Treatment 30](#_bookmark10)

[Table 17. Parameters of the fitted values 30](#_bookmark13)

[Table 18. Beta coefficients after multicollinearity Treatment 31](#_bookmark15)

[Table 19. Contraceptive data set top 5 rows and last 5 rows 33](#_bookmark16)

[Table 20. Information of data set 34](#_bookmark17)

[Table 21. Null values of the of data set 35](#_bookmark19)

[Table 22. Description of data set 35](#_bookmark20)

[Table 23. Variables after null value imputation 36](#_bookmark22)

[Table 24. Description after null value imputation 37](#_bookmark25)

[Table 25. Contraceptive data set after encoding 47](#_bookmark7)

[Table 26. Predictive probabilities on Training data for LR 48](#_bookmark8)

[Table 27. Predictive probabilities on TEST data for LR 48](#_bookmark9)

[Table 28. Predictive probabilities on Training data for LDA 52](#_bookmark9)

[Table 29. Predictive probabilities on Training data for LDA 52](#_bookmark10)

[Table 30. Classification Report on Training data LR 52](#_bookmark13)

[Table 31. Classification Report on Testing data LR 53](#_bookmark15)

[Table 32. Classification Report on Training data for LDA 54](#_bookmark16)

[Table 33. Classification Report on Testing data for LDA 55](#_bookmark17)

[Table 34. Importance of features in CART model in Training data 57](#_bookmark19)

[Table 35. Classification Report on Training data for CART 57](#_bookmark16)

[Table 36. Classification Report on Testing data for CART 58](#_bookmark17)

List of Equations and Formulas

[Equation 1. MAE Formula 10](#_bookmark8)

[Equation 2. RMSE Formula 10](#_bookmark9)

[Equation 3. IQR Formula 10](#_bookmark10)

[Equation 4. VIF formula 38](#_bookmark13)

[Equation 5. Log loss function of logistic regression 55](#_bookmark15)

List of Figures

[Figure 1. Null values percentage 10](#_bookmark9)

[Figure 2. Histograms of the variables. 11](#_bookmark10)

[Figure 3. Boxplots of the variables with outliers’ 12](#_bookmark13)

[Figure 4. Pair plot of the variables 14](#_bookmark15)

[Figure 5. Heat map of the variables 15](#_bookmark16)

[Figure 6. Bivariate analysis – Bar graph 16](#_bookmark17)

[Figure 7. Bivariate analysis – scatter plot 16](#_bookmark19)

[Figure 8. Boxplots of the variables without outliers 20](#_bookmark20)

[Figure 9. Boxplots of the variables after outlier treatment 21](#_bookmark20)

[Figure 10. Fitted vs Residual plot before multicollinearity 27](#_bookmark16)

[Figure 11. Fitted vs Residual plot after multicollinearity treatment 31](#_bookmark17)

[Figure 12. Null value percentage 35](#_bookmark19)

[Figure 13. Boxplots before outliers’ treatment 36](#_bookmark20)

[Figure 14. Boxplots after outliers’ treatment 37](#_bookmark22)

[Figure 15. Histogram of wife’s age 38](#_bookmark17)

[Figure 16. Boxplot of wife’s age 38](#_bookmark19)

[Figure 17. Histogram of Number of children born 39](#_bookmark20)

[Figure 18. Boxplot of Number of children born 39](#_bookmark22)

[Figure 19. Histogram of Husband’s Occupation 40](#_bookmark22)

[Figure 20. Boxplot of Husband’s Occupation 40](#_bookmark17)

[Figure 21. Histogram of Wife Education 41](#_bookmark19)

[Figure 22. Histogram of Husband’s Education 41](#_bookmark20)

[Figure 23. Histogram of Wife Religion 41](#_bookmark22)

[Figure 24. Histogram of Wife Working frequency 42](#_bookmark17)

[Figure 25. Histogram of Wife Education 42](#_bookmark19)

[Figure 26. Histogram of Media Exposure 43](#_bookmark20)

[Figure 27. Histogram of target variable 44](#_bookmark22)

[Figure 28. Bivariate analysis of target variable and Husband Education 44](#_bookmark22)

[Figure 29. Multivariate analysis of target variable 45](#_bookmark25)

[Figure 30. Pir plot of Numerical variables 46](#_bookmark22)

[Figure 31. Heat map of Numerical variables 46](#_bookmark25)

[Figure 32. Tree flow chart of CART model 51](#_bookmark22)

[Figure 33. AUC-ROC Curve of LR – Training data 53](#_bookmark25)

[Figure 34. AUC-ROC Curve of LR – Testing data 54](#_bookmark22)

[Figure 35. AUC-ROC Curve of LDA – Training data 55](#_bookmark25)

[Figure 36. AUC-ROC Curve of LDA – Testing data 55](#_bookmark25)

List of Dictionary

1. lread - Reads (transfers per second) between system memory and user memory.
2. lwrite - writes (transfers per second) between system memory and user memory.
3. scall - Number of system calls of all types per second.
4. sread - Number of systems read calls per second.
5. swrite - Number of systems write calls per second.
6. fork - Number of system fork calls per second.
7. exec - Number of system exec calls per second.
8. rchar - Number of characters transferred per second by system read calls.
9. wchar - Number of characters transferred per second by system write calls.
10. pgout - Number of pages out requests per second.
11. ppgout - Number of pages, paged out per second.
12. pgfree - Number of pages per second placed on the free list.
13. pgscan - Number of pages checked if they can be freed per second.
14. atch - Number of page attaches (satisfying a page fault by reclaiming a page in memory) per second.
15. pgin - Number of page-in requests per second.
16. ppgin - Number of pages paged in per second.
17. pflt - Number of page faults caused by protection errors (copy-on-writes).
18. vflt - Number of page faults caused by address translation.
19. runqsz - Process run queue size (The number of kernel threads in memory that are waiting for a CPU to run.
20. freemem - Number of memory pages available to user processes.
21. freeswap - Number of disk blocks available for page swapping.
22. usr - Portion of time (%) that CPUs run in user mode
23. Wife's age (numerical)
24. Wife's education (categorical) 1=uneducated, 2, 3, 4=tertiary
25. Husband's education (categorical) 1=uneducated, 2, 3, 4=tertiary
26. Number of children ever born (numerical)
27. Wife's religion (binary) Non-Scientology, Scientology
28. Wife's now working? (binary) Yes, No
29. Husband's occupation (categorical) 1, 2, 3, 4(random)
30. Standard-of-living index (categorical) 1= overflow, 2, 3, 4=high
31. Media exposure (binary) Good, not good
32. Contraceptive method used (class attribute) No, Yes

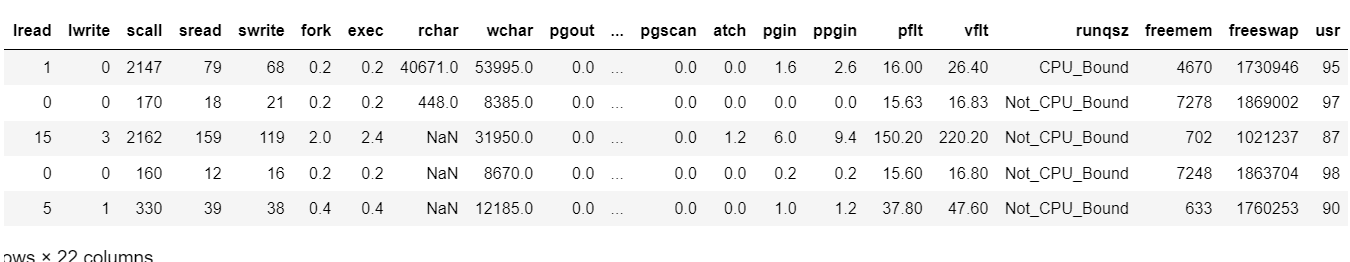
# PROBLEM – 1

# LINEAR REGRESSION

The comp-active databases are 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.

Table 1 – Data set of comp-active databases (1st – 5 Rows)



**System measures used (dictionary):**

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 systems read calls per second.

swrite - Number of systems 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 transferred per second by system write calls.

pgout - Number of pages 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.

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

**Part 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.**

**Exploratory Data Analysis**

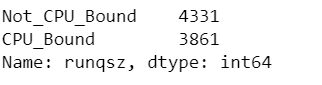
So now we will start exploring the data first and we will get to know the structure and observations in the data set along with the information and description of the data. we will check the missing values and 5-point summary of the dataset and then impute the same if there are any missing values or outliers present.

Table 2 – Information of dataset (Last – 5 Rows)



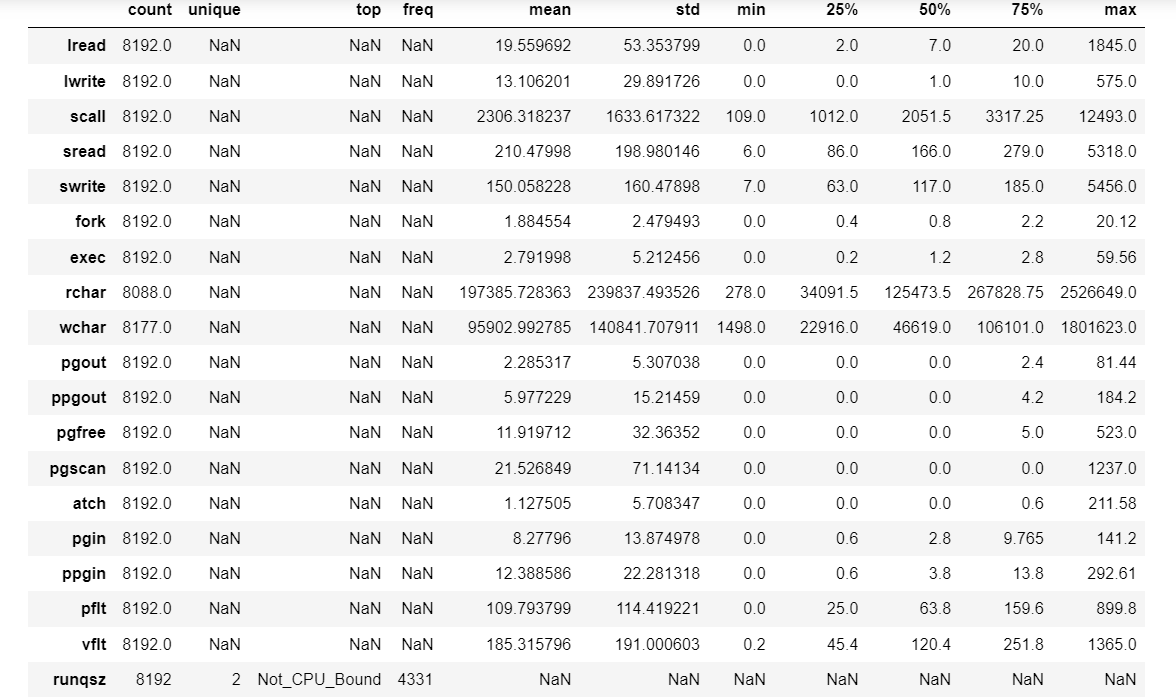
The data set contains 8192 rows and 22 columns, whose descriptions are mentioned above. Here we can see that 21 out of 22 columns are numerical in nature with either Float or Int data type. 'runqsz' which represents Process run queue size is in the 'object' data type format. We will check the features and their value counts and then encode them using one hot encoding.

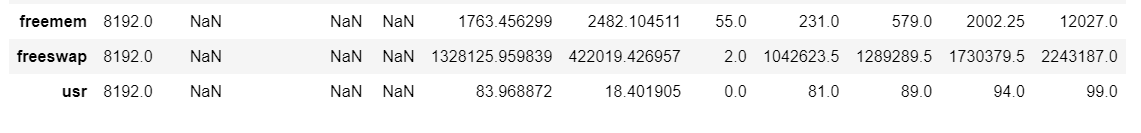
Table 3 – Value count of Object data type Variable



Since this variable is in binary form, we can change it in integer datatype so that it can have values 0 or 1 for 'Not CPU Bound' and 'CPU Bound' respectively. So, we can create Dummy variables using One hot Encoding method.

Table 4 – Description of Company





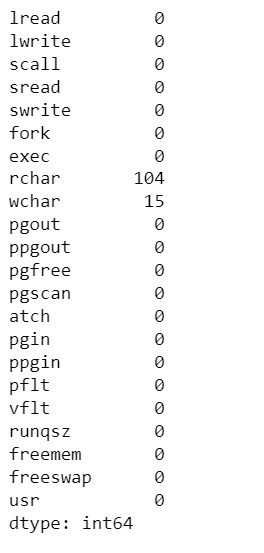
**Summary of dataset**

1-The mean and median are approximately equal for all the variables except for "rchar’, "wchar", "freeswap", "scall", "sread”, "freemem", and target variable "usr".

2- All variables are continuous variables except for "runqsz" which is categorical in nature.

3- Almost all the variables have minimum amount values as zero, except for "rchar', "wchar", "freeswap", "scall", "sread", "swrite" and "freemem".

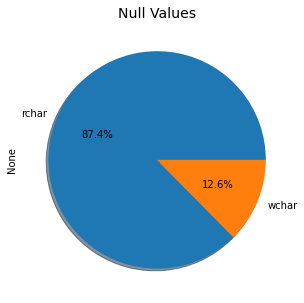
Table 5 – Missing values in the data set



The null values present in the data set is less than 1% of the total observations present. so, we will treat this data rather than dropping these observations.

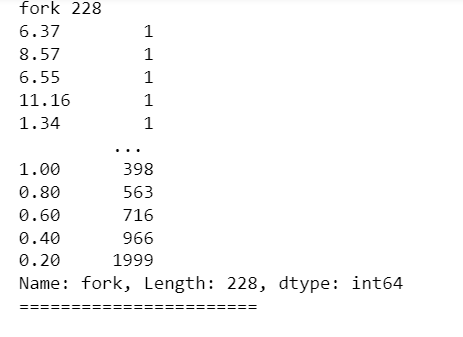
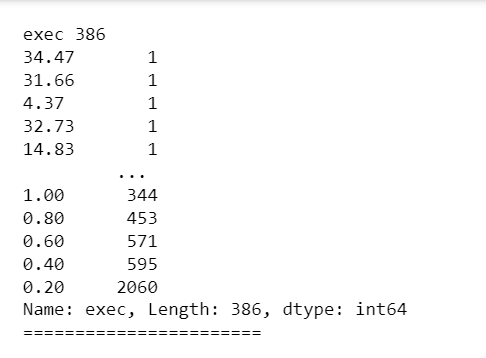
1. Observed null values in 2 fields rchar and wchar.

2. We can impute the null values with median value of the data set as the variables are continuous in nature.

 fig -1 Null values percentage

Now checking the duplicate rows in the dataset is the next step and if we find any duplicate values, we will drop those columns else we will move forward with our further analysis. So, we used the corresponding duplicated function and it gives out that we have zero (0) duplicate rows in our dataset.

**UNIVARIATE ANALYSIS:**

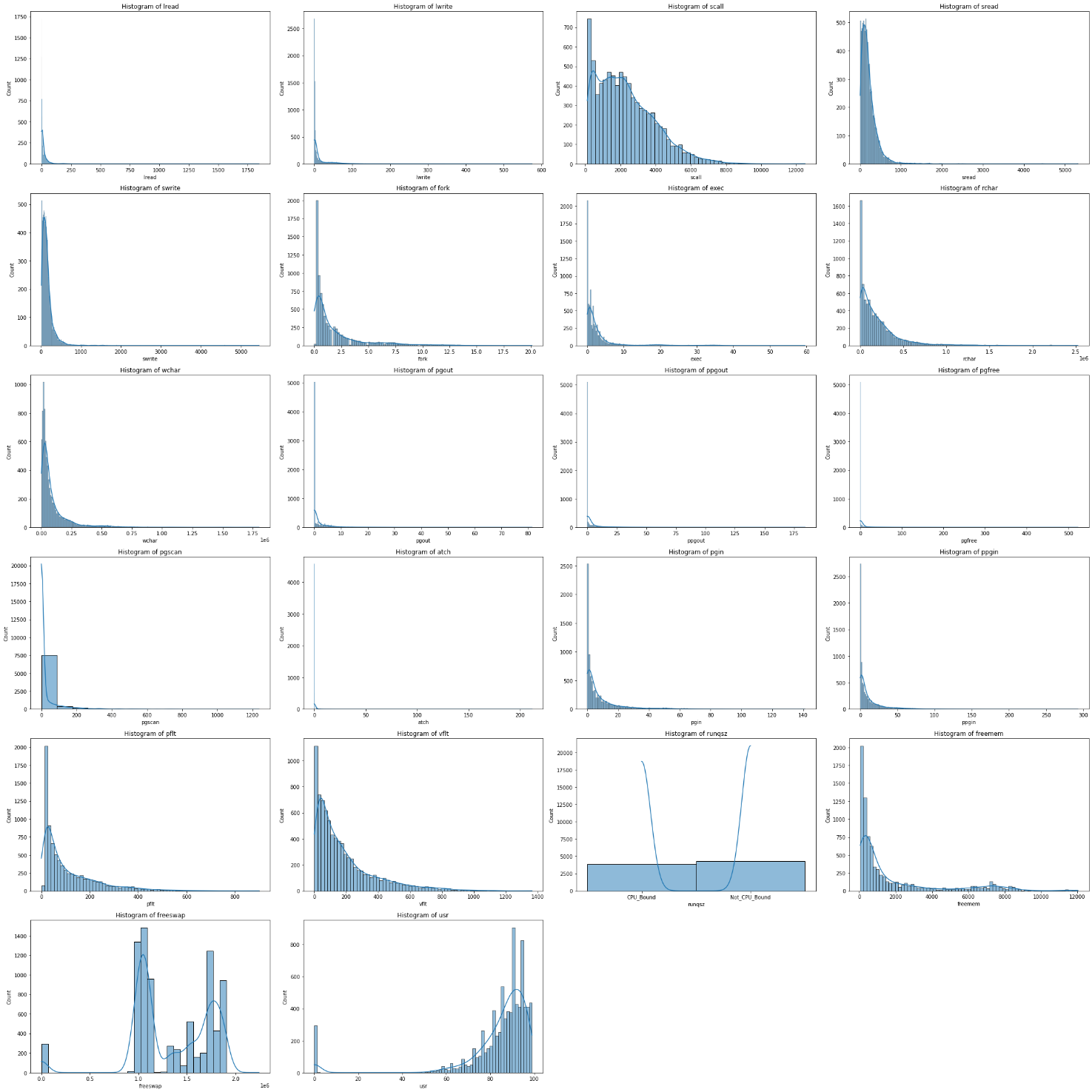


fig -2 Histograms of the variables

**Inference:**

1-The data for almost all the columns shows that the data is positively skewed except for "usr" and there could be possibilities of multimode as there are multiple peaks seen in the case of "scall" and "freeswap".

2-The data for "freeswap" is almost normally distributed with two peaks and distributed between the range of 0.0

and 2.0.

3-The data for target variable "usr" shows that the data is negatively skewed.

4-The data for "runqsz" shows that the data has only two observations "CPU Bound" and "CPU not Bound" where the amount of "CPU Bound" is slightly less than "CPU not Bound".

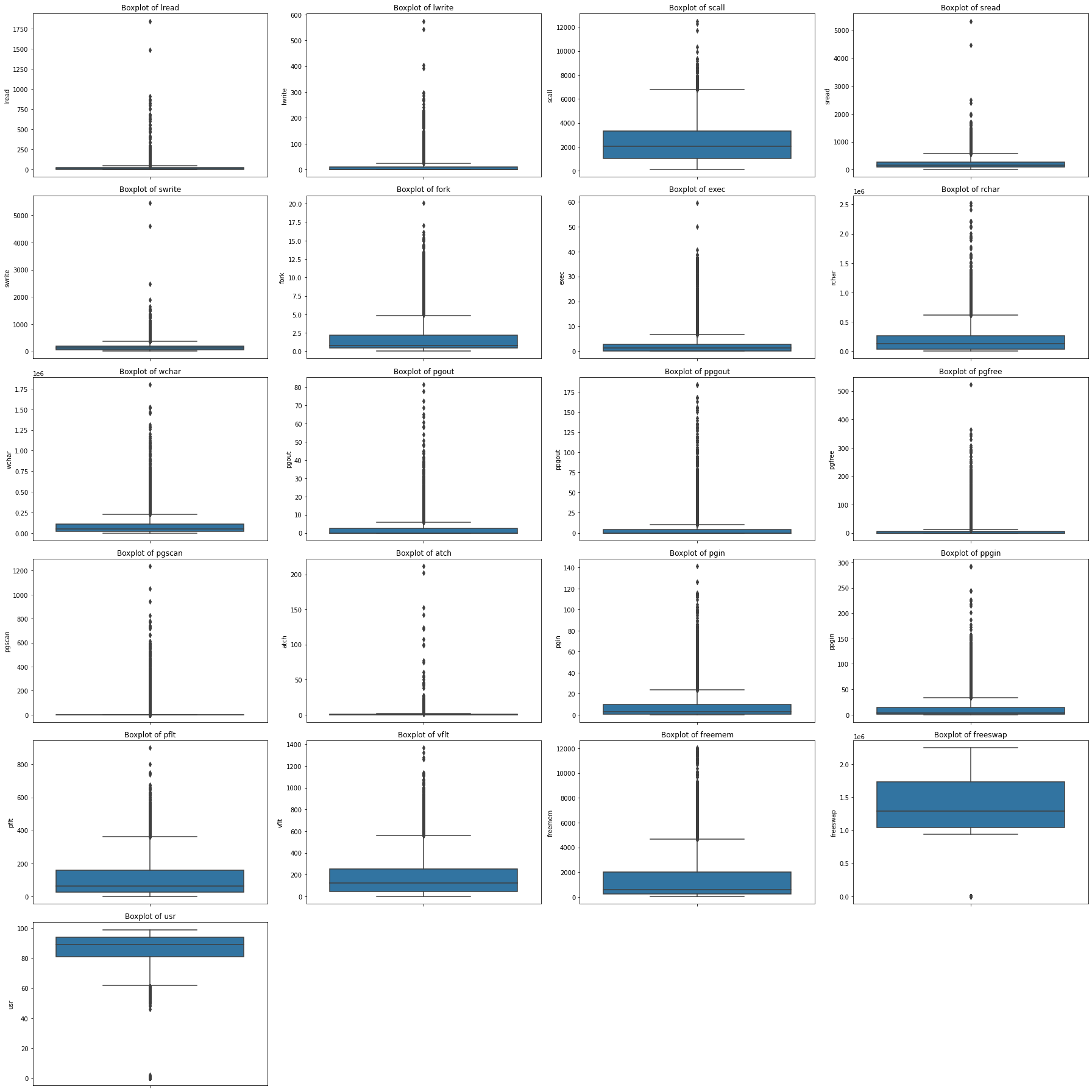
****

fig -3 – Boxplots of variables with outliers

**Observations:**

1. Almost all the continuous variables have outliers except for "freeswap" which has no outlier present.

2. We can treat these outliers using the IQR approach

3. The target variable "usr" has outliers less than the minimum value as we have already seen that it is a negatively skewed curve.

4. In this case, it is not necessary to scale the data as, we'll get an equivalent solution whether we apply some kind of linear scaling or not. For example, to find the best parameter values of a linear regression model, there is a closed-form solution, called the Normal Equation. If our implementation makes use of that equation, there is no stepwise optimization process, so feature scaling is not necessary.

**Bivariate and Multivariate Analysis:**

We can do the bivariate and multivariate analysis using count plot and bar plot of various variables with respect to each other and also with the target variable.

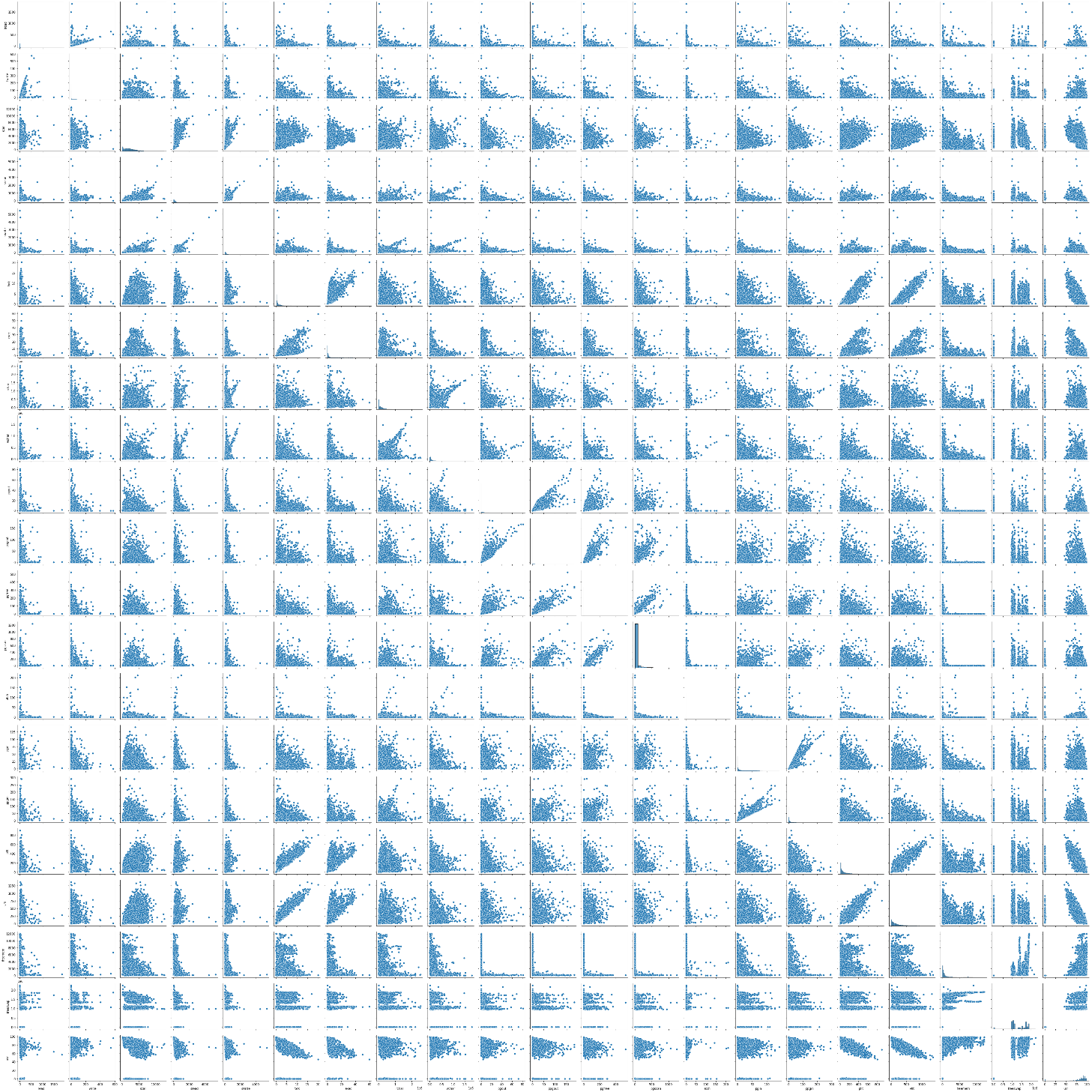


fig -4 Pair plot of all the variables

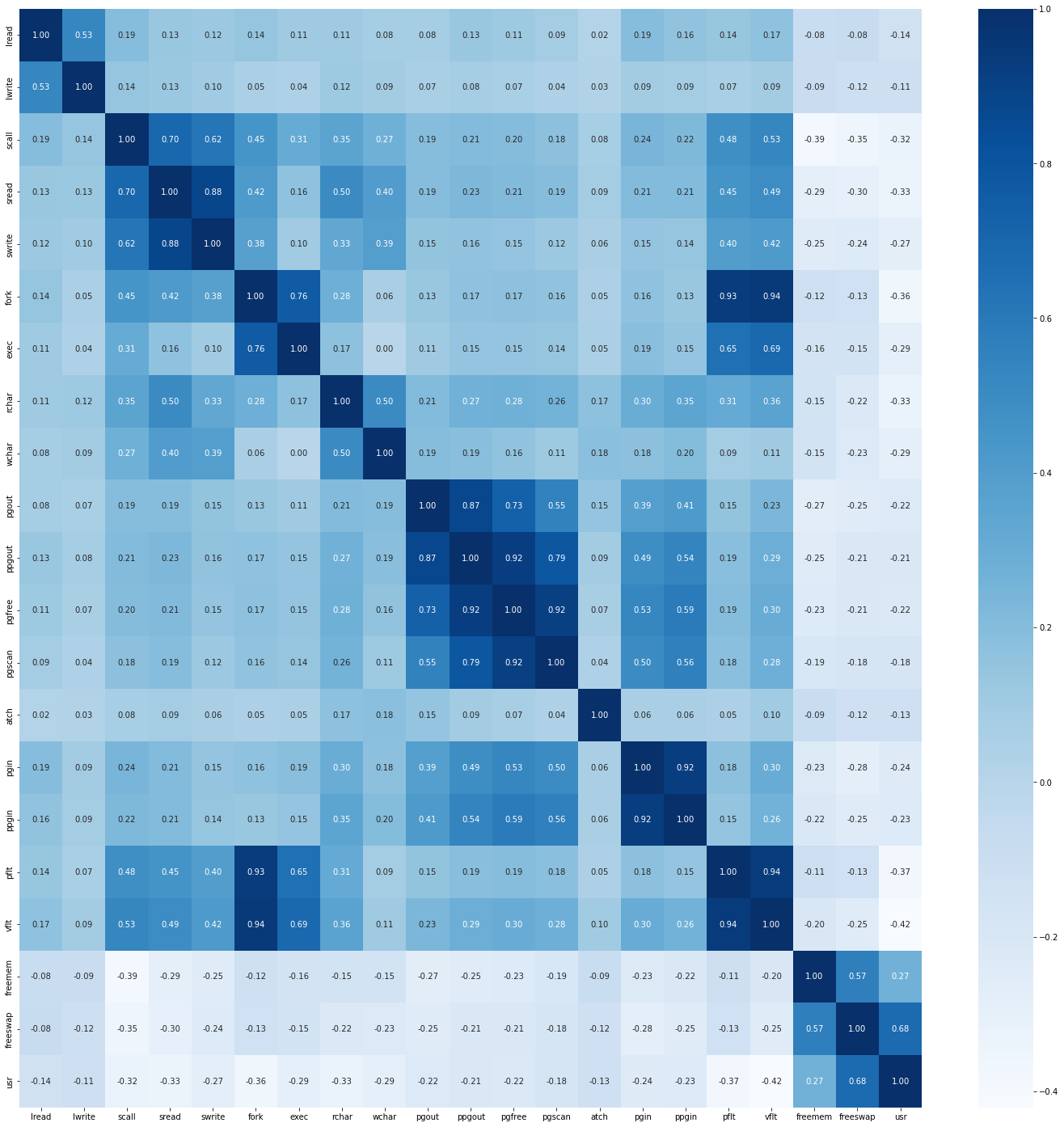


fig -5 Heat map of all the variables



Fig 6 – Bivariate analysis – Bar graphs

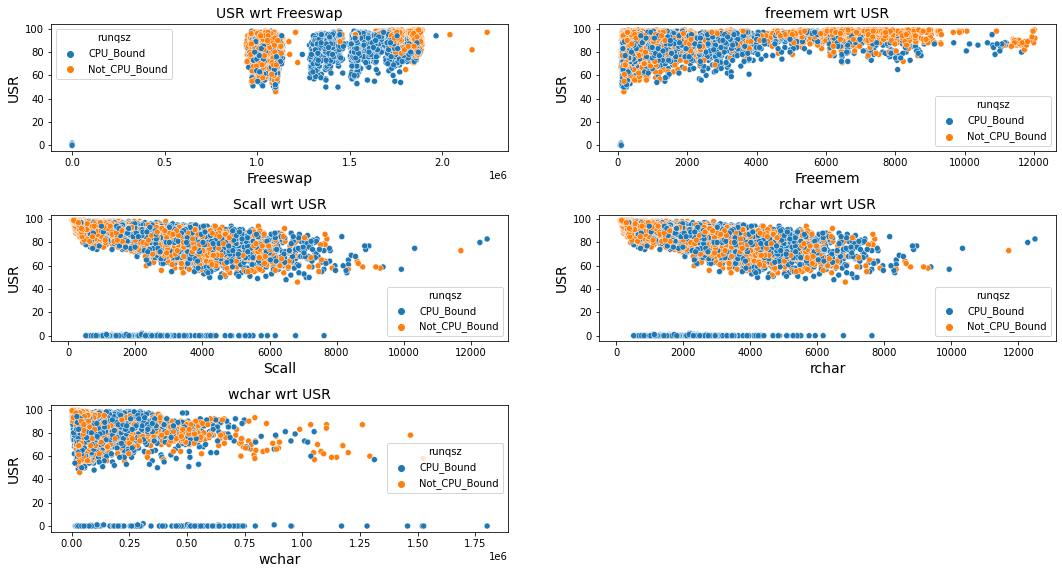


Fig 7 – Bivariate analysis – Scatter plots

**Observations based on pair plot:**

Inference: As per the Heat Map, we can conclude that the following variables are highly correlated:

- freemem and usr

- freeswap and usr

- ppgout and pgfree

- ppgout and pgout

- fork with pflt and pflt

1. 'usr’ is the target variable or dependent variable and rest are predictor variables or also known as independent variables.

2. Looking into the fields in the univariate analysis, we see there are outliers that needs to be treated.

3. Multivariate analysis indicates that there is strong positive correlation between the target variable "usr" and the predictor variables "freemem" and "freeswap" and also some of the independent variables like "ppgout" with "pgout" and "pgfree".

4. "pflt" and "vflt" has very high positive correlation with "fork".

5. For bivariate analysis We can plot Bar plots and Count plots with target variables and these independent variables to see how they are behaving with the target variable.

**Part 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.**

Table 6 – Missing values % in the data set

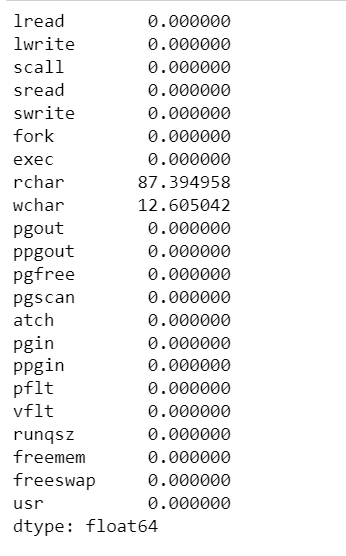
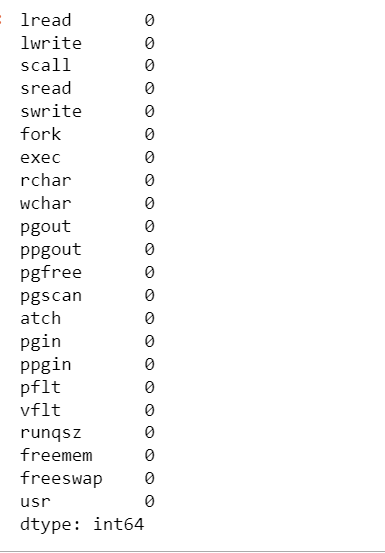


Table 7 – After missing values imputation

****

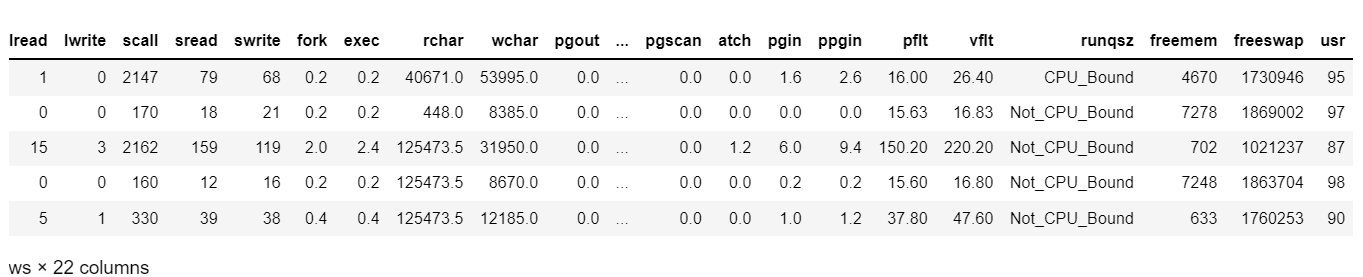
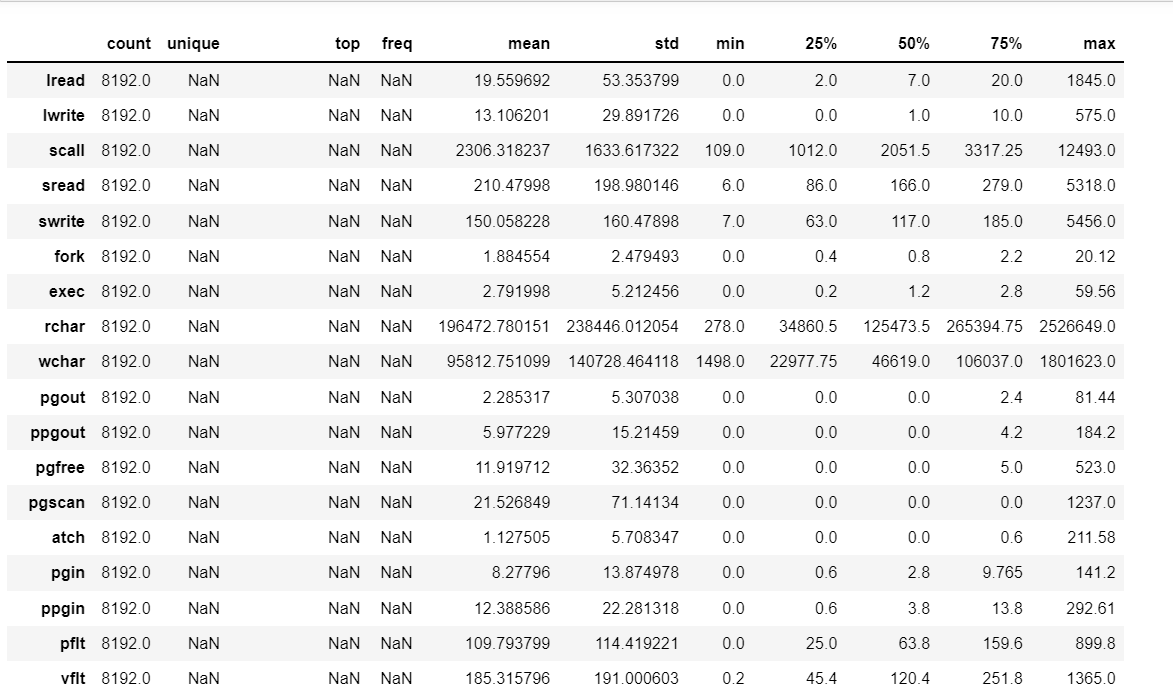
****

Table 8 – Data set After missing values imputation

**Summary after missing values imputation:**

****

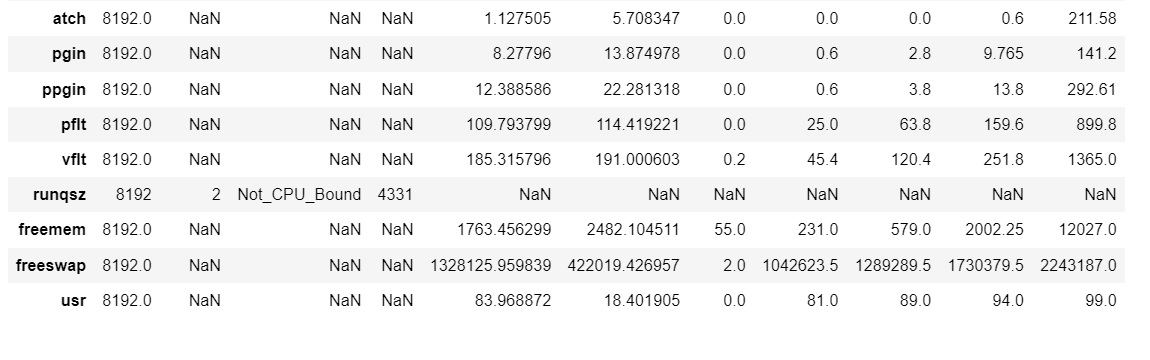
****

Table 9 – Description After missing values imputation

Number of duplicate rows = 0

**Treating the outliers:**

As we have already seen that outliers are present in the dataset, we need to treat those outliers as they can affect the modelling algorithm. So, we will treat the outliers using IQR method and then again check the 5-point summary for the continuous variables.

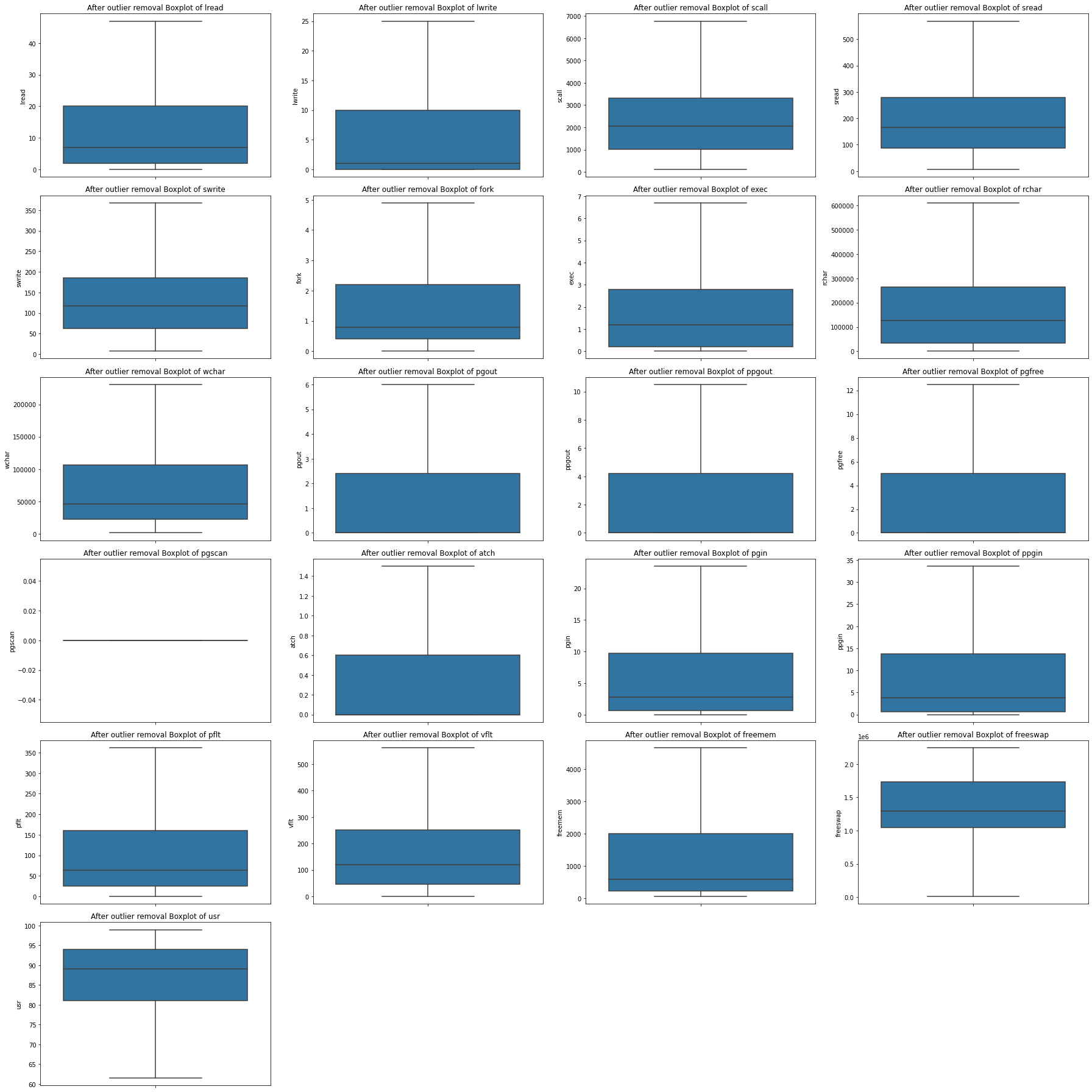


Fig 8 – Boxplots after outliers’ treatment

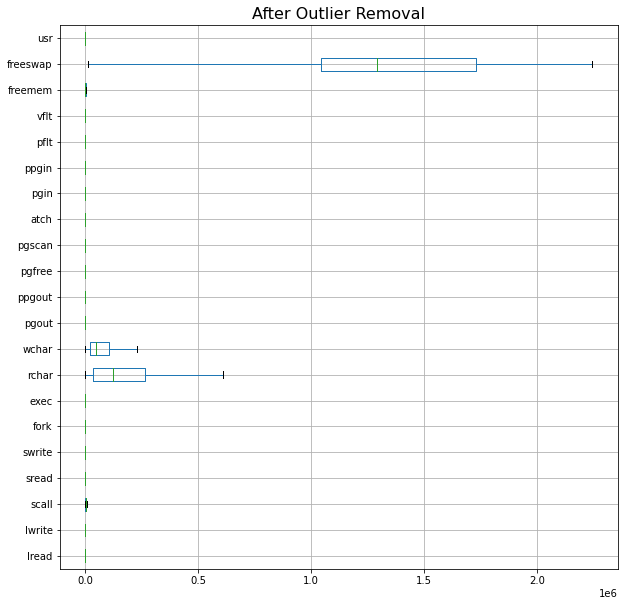
****

Fig 9 – Boxplots after outliers Treatment

For values which are equal to Zero (0)

* After treating the outliers, we can see that almost all the variables have their minimum values as zero (0) except for variables like "scall", "rchar", "wchar", "freemem", "freeswap".
* Removing the records with 0 values will not be a good idea in this case as it might not have an impact on the model building

**Part 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 Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.**

Encoding the data:

There are two types of categorical data:

- Ordinal: Order based like 'good',' bad',' worst' like Clothing sizes

- Nominal: Without any order or ranks like city names, Genders, etc.

Here we will use "One Hot Encoding" because we don’t have the problem of high dimensionality in this case as it’s a Boolean type of categorical variable.

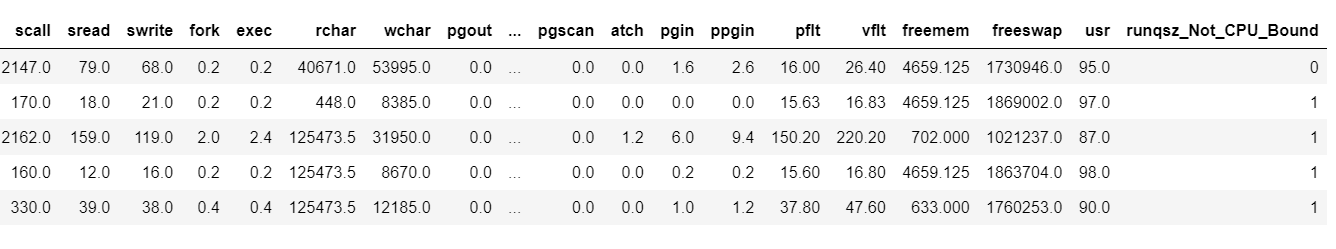


Table 10 – Data set After Encoding the data set

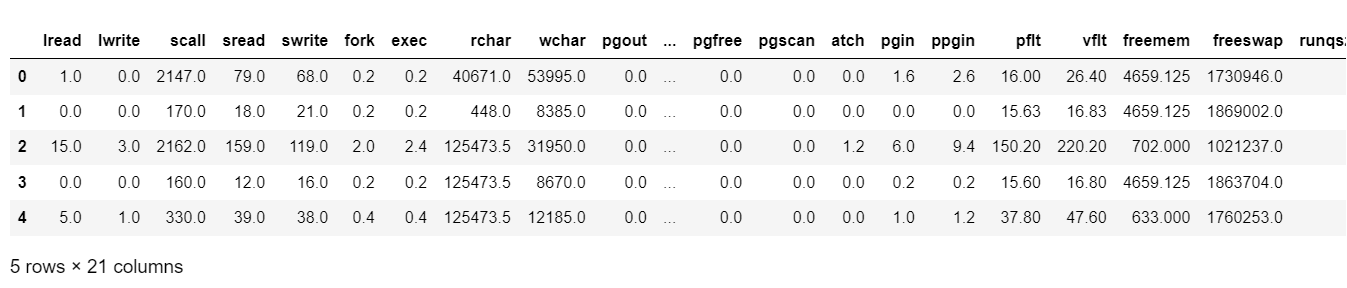


Table 11 – Data set After Encoding the data set continued

### Linear Regression Model –

Evaluation of Linear regression model-Evaluation helps to judge the performance of any machine learning model that would provide best results to our test data. Fundamentally three types of evaluation metrics are used to evaluate linear regression model. -R2 measure (discussed with least square method)-Mean Absolute Error (MAE)-Root Mean Square Error (RMSE) Mean Absolute Error (MAE)-Mean Absolute Error is the average of the difference between actual and predicted value of target variable.

𝑀𝐴𝐸=1𝑛∑|𝑦𝑖−𝑦ℎ𝑎𝑡𝑖|𝑛𝑖=1 – Equation 1

Root Mean Square Error (RMSE)-defined as:

𝑅𝑀𝑆𝐸=√1𝑛∑(𝑦𝑖−𝑦ℎ𝑎𝑡𝑖)2𝑛𝑖=1 – Equation 2

Pros and cons of Linear Regression: Pros-Linear regression models are very simple and easy to implement. These models are said to be most interpretable. Cons-Linear regression models are largely affected by the presence of outlier in training data.

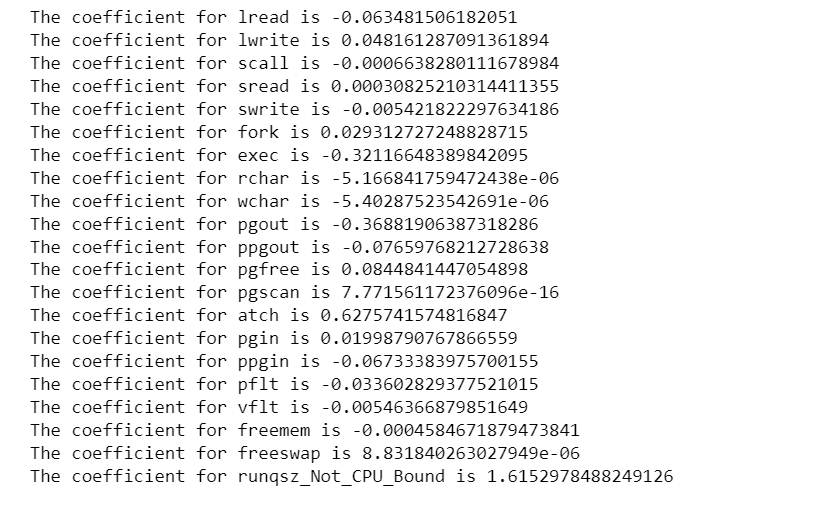
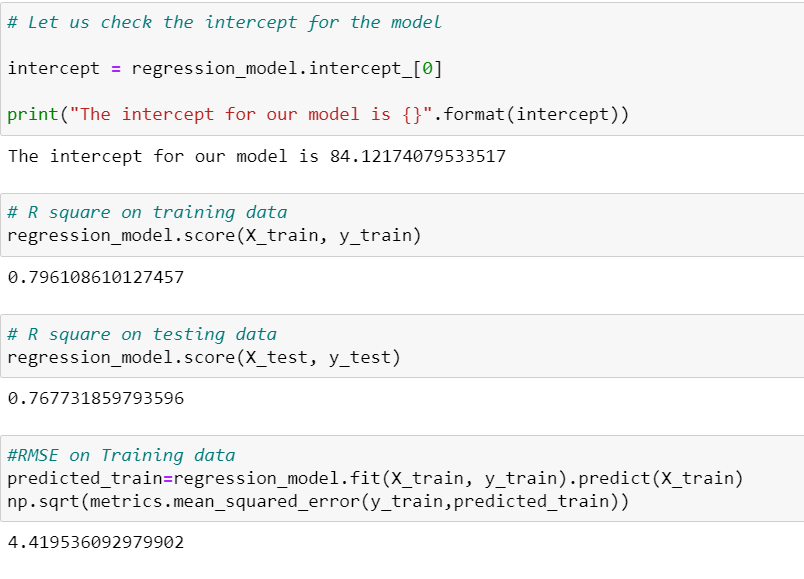


Table 12 – Coefficients of variables

**Performance Metrics**

Performance measures are a way to evaluate and compare our business models and to decide which model works well for the business scenario. To evaluate our Linear Regression Model, we will take two measures namely, R square and RMSE which we will compute for both train and test datasets. R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables in a regression model.

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.



### fig -2

### Insights:

### - R-squared is always between 0 and 100%:

### - 0% indicates that the model explains none of the variability of the response data around its mean.

### - 100% indicates that the model explains all the variability of the response data around its mean. In general, the higher the R-squared, the better the model fits data.

### - In this case, R-squared value for both test and train is 0.76 and 0.79 respectively, which indicates that more than 75% of observed variance can be explained by model’s inputs.

### Let’s make it more clear for each continuous variables, so that we can figure out the no of outliers, and treat them.

### Linear Regression using stats models (OLS)

Table 13 – OLS for regression model before multicollinearity treatment

### 

### 

### 

### Notes: 1. Standard Errors assume that the covariance matrix of the errors is correctly specified. 2. The smallest eigenvalue is 2.58e-28. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

### 3. The R-squared value tells us that our model can explain 79.6% of the variance in the training set.

### Interpretation of Coefficients:

### - The coefficients tell us how one unit change in X can affect y.

### - The sign of the coefficient indicates if the relationship is positive or negative.

### - Multicollinearity occurs when predictor variables in a regression model are correlated. This correlation is a problem because predictor variables should be independent. If the collinearity between variables is high, we might not be able to trust the p-values to identify independent variables that are statistically significant.

### - When we have multicollinearity in the linear model, the coefficients that the model suggests are unreliable.

### The Sklearn model:

### - The R-squared / Co-efficient of determinant (Training data) = 0.796

### - The R-squared / Co-efficient of determinant (Test data) = 0.756

### - RMSE/ Root mean squared error (Training data) = 4.419

### - RMSE/ Root mean squared error (Test data) = 4.612

### - As the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.

### The Stats models:

### - The R-squared / Co-efficient of determinant and Adjusted R –squared are same = 0.796

### - For now, rchar, wchar, pgout, pgscan, depth, pflt, vflt, freemem, freeswap are good predictor of usr.

### - As we can see variable scall, exec, pgout, ppgout and swrite has negative co-efficient, which means higher the value of scall, exec, pgout, ppgout and swrite, the usr will be lower.

### Z score (Scaled dataset):

### - VIF (Variation Inflation Factor): - There is high multi-collinearity between the independent variables.

### Interpretation of p-values (P > |t|)

### For each predictor variable there is a null hypothesis and alternate hypothesis.

### - Null hypothesis: Predictor variable is not significant

### - Alternate hypothesis: Predictor variable is significant

### - (P > |t|) gives the p-value for each predictor variable to check the null hypothesis

### - If the level of significance is set to 5% (0.05), the p-values greater than 0.05 would indicate that the corresponding predictor variables are not significant.

### - However, due to the presence of multicollinearity in our data, the p-values will also change.

### - We need to ensure that there is no multicollinearity in order to interpret the p-values.

### Insights:

### - Null hypothesis (Ho) is true i.e., there is no relation between dependent and independent variable in the universe, where usr is dependent variable and others like lread, lwrite, ppgout, pgout, rchar, wchar are independent variables.

### - For example, all the variables are showing p value below the 0.05 (alpha), but some variables like fork, ppgout, pgin is showing some p value greater than 0.05, if more sample is collected, we can predict better if these variables are good predictor of usr or not.

### - For now, rchar, wchar, pgout, pgscan, depth, pflt, vflt, freemem, freeswap are good predictor of usr. As we can see variable scall, exec, pgout, ppgout and swrite has negative co-efficient, which means higher the value of scall, exec, pgout, ppgout and swrite, the usr will be lower.

**RMSE on train data = 4.41953609297990**

**RMSE on test data = 4.6522957041930155**

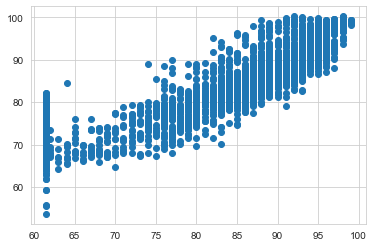
****

Fig 10 – Fitted vs Residual plot before multicollinearity Treatment

Here in our model, we can observe that few of the coefficients are positive and few are negative in nature. The positive coefficients shows that Target variable (usr) increases with 1 unit when these coefficients increase with 1 unit. So, following are observations based on the model

- The coefficient of constant that is b0 is equals to 84.12

- When lread increases by 1 unit, usr decreases by 0.06 units keeping all other predictors constant.

- When lwrite increases by 1 unit, usr increases by 0.05 units keeping all other predictors constant.

- When atch increases by 1 unit, usr increases by 0.63 units keeping all other predictors constant.

- When pgout decreases by 1unit, usr decreases by 0.37 units keeping all other predictors constant.

- Similarly, 1 unit increase of ppgin, decreases the value of usr by 0.07 units

### Insights:

### 1.) We know that if P>|t| result, we can infer that the variables like sread, fork, ppgout, pgfree and pgin are the statistically insignificant variables, as their p-value is greater than 0.05.

### 2.) Omnibus test checks the normality of the residuals once the model is deployed. If the value is zero, then it means the residuals are perfectly normal. Here, in the example prob (Omnibus) is 0 indicating that there is 0% chance that the residuals are normally distributed. For a model to be robust, besides checking R-squared and other rubrics, the residual distribution is also required to be normally distributed.

### 3.) Hence, we can say that our model is not robust and may not be fit.

### 4.) Also, we can observe there are very strong multi collinearity present in the data set as the values are exceeding 10 or close to 0. Ideally, the values should be within 1 to 5

### VIF

### Equation3

### Reviewing Linear Regression:

### 

Table 14 – VIF values for variables

### Observations:

### 1. The VIF values indicate that the features pgin, ppgin, pflt, vflt, const - 29.229332, lread - 5.350560, lwrite-4.328397, scall - 2.960609, sread- 6.420172, swrite-5.597135, fork- 13.035359, exec - 3.241417, pgout-11.360363, ppgout - 29.404223, pgfree are correlated with one or more independent features.

### 2. Multicollinearity affects only the specific independent variables that are correlated. Therefore, in this case, we can trust the p-values of rchar, wchar, freemem-1.961304, freeswap-1.841239, runqsz Not CPU Bound - 1.156815variables.

### 3. To treat multicollinearity, we will have to drop one or more of the correlated features.

### 4. We will drop the variable that has the least impact on the adjusted R-squared of the model.

### 5. “pgscan" whose value shows "NaN" which means VIF very high almost near to infinity.

### Let's drop multicollinear columns one by one and observe the effect on our predictive model.

### In case of all the above variables “fork”, "pgin", "ppgin", "vflt" and "pflt" the value of R- squared remained the same. Since there is a very small effect (0.002) or no effect on adj. R-squared after dropping the all the above columns, we can remove it from the training set and then again evaluate the model.

Table 15 – OLS model after multicollinearity treatment

### 

### 

### New VIF values after dropping almost 12 columns which showed no effect on R- squared value and adjusted R- squared value. We are able to obtain the following VIFs values. VIF for all the features is <2, except for "pgscan" whose value shows "NaN" which means VIF very high almost near to infinity. So, we can drop that value as well to check the model performance.

### 

Table 16 – VIF values after treatment of multicollinearity

**Part 1(4) – 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.**

**Before dropping the columns, this states our linear model:**

**usr (target variable)** = b0 + (b1) \* lread + (b2) \* lwrite + (b3) \* scall + (b4) \* sread + (b5) \* swrite + (b6) \* fork + (b7) \* exec + (b8) \* rchar + (b9) \* wchar + (b10) \* pgout + (b11) \* ppgout + (b12) \* pgfree + (b13) \* pgscan + (b14) \* atch + (b15) \* pgin + (b16) \* ppgin + (b17) \* pflt + (b18) \* vflt + (b19) \* freemem + (b20) \* freeswap + (b21) \* runqsz\_Not\_CPU\_Bound

**which can be written as:**

(84.12) \* const + (-0.06) \* lread + (0.05) \* lwrite + (-0.0) \* scall + (0.0) \* sread + (-0.01) \* swrite + (0.03) \* fork + (-0.32) \* exec + (-0.0) \* rchar + (-0.0) \* wchar + (-0.37) \* pgout + (-0.08) \* ppgout + (0.08) \* pgfree + (0.0) \* pgscan + (0.63) \* atch + (0.02) \* pgin + (-0.07) \* ppgin + (-0.03) \* pflt + (-0.01) \* vflt + (-0.0) \* freemem + (0.0) \* freeswap + (1.62) \* runqsz\_Not\_CPU\_Bound

**Insights:**

• When lwrite increases by 1 unit, usr increases by 0.05 units keeping all other predictors constant.

• When atch increases by 1 unit, usr increases by 0.63 units keeping all other predictors constant.

• When pgout decreases by 1unit, usr decreases by 0.37 units keeping all other predictors constant

**After solving the multicollinearity problem, we are getting the following parameters and model:**

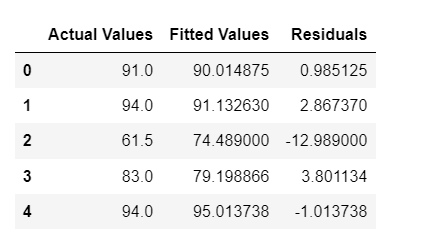
****

Table 17– Parameters of fitted values

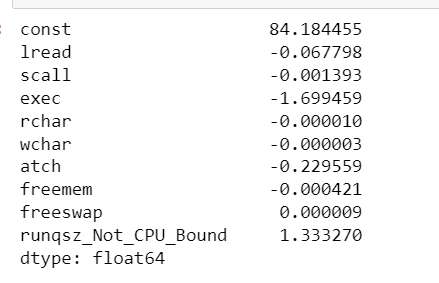
****

Table 18 – Beta coefficients after multicollinearity Treatment

**After removing the multi collinearity and dropping the columns these are what we are getting as coefficients:**

USR = 84.18445516289617 + -0.06779791940844981 \* (lread) + -0.0013932765906504916 \* (scall) + -1.6994587566206565 \* (exec) + -1.00812025182409e-05 \* (rchar) + -2.70530481458774e-06 \* (wchar) + -0.22955897717411577 \* (atch) + -0.00042099213129056525 \* (freemem) + 8.589364183271158e-06 \* (freeswap) + 1.3332700789577303 \* (runqsz\_Not\_CPU\_Bound)

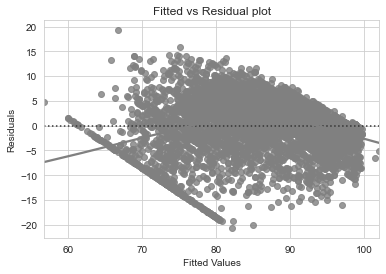
****

Fig 11 – Fitted vs Residual plot before multicollinearity Treatment

We observe that the pattern has slightly decreased and the data points seems to be randomly distributed and the prediction error should not be linked to the magnitude of the value predicted.

**Inferences of the model:**

1. We can see that RMSE on the train and test sets are comparable. So, our model is not suffering from overfitting.

2. MAE indicates that our current model is able to predict usr within a mean error of 3.3 units on the test data and 4.03 on the train data. Hence, we can conclude the model "ols\_res11" is good for prediction as well as inference purposes.

3. When "scall" increases by 1 unit, usr decreases by 0.005 units keeping all other predictors constant.

4. When "atch" increases by 1 unit, usr decreases by 0.35 units keeping all other predictors constant.

5. When "exec" increases by 1unit, usr decreases by 1.8 units keeping all other predictors constant

6. 8.941837446447745e-06 times "freeswap" and 1.451351637618937 times "runqsz\_Not\_CPU\_Bound", shows that both the variables contribute positively to usr.

7. We have already removed around 8 variables from the data set because of the multicollinearity issue using VIF factor method and hence our model is not overfitting now.

**So, WE can conclude that overall, 5 variables which actually plays an important role in case of defining the usr is given as follows**

- usr = a + b1\*x1 + b2\*x2 + b3\*x3 + b4\*x4 + b5\*x5

- where x1 = amount of free memory, x2 = amount of free swap space, x3 = number of processes, x4 = number of users, and x5 = load average.

- b1, b2, b3, b4, and b5 are the coefficients that determine how each attribute affects the system to be in 'usr' mode.

# PROBLEM STATEMENT– 2

## **About Data**

# 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.

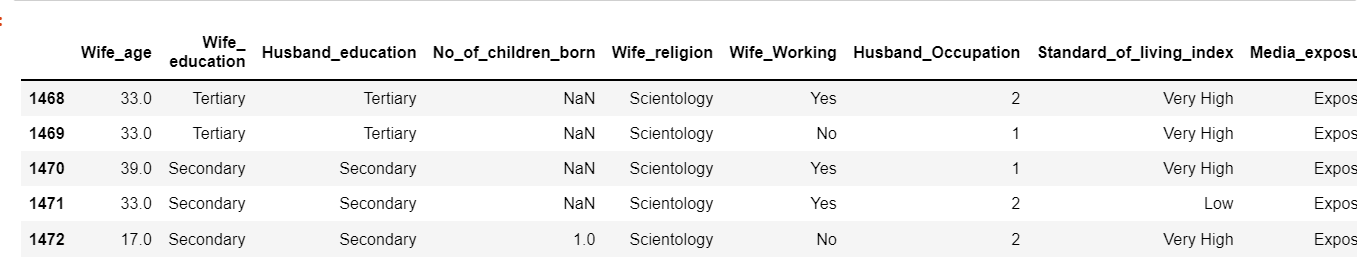
The problem is to predict do/don't they use a contraceptive method of choice based on their demographic and socio-economic characteristics.

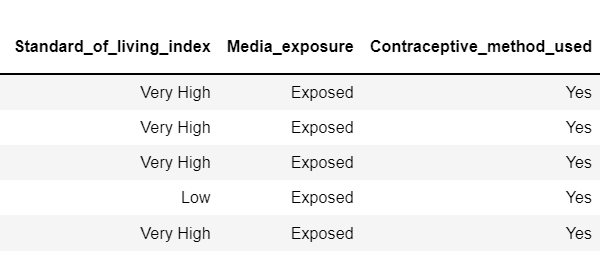
### Part 2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, check for duplicates and outliers and write an inference on it. Perform Univariate and Bivariate Analysis and Multivariate Analysis.?

**EDA Exploratory Data Analysis**

# 

# Table-18-Contraceptive data set – Top 5 rows



Table-19-Contraceptive data set – last 5 rows

**Information and Null Values counts of the complete dataset**:

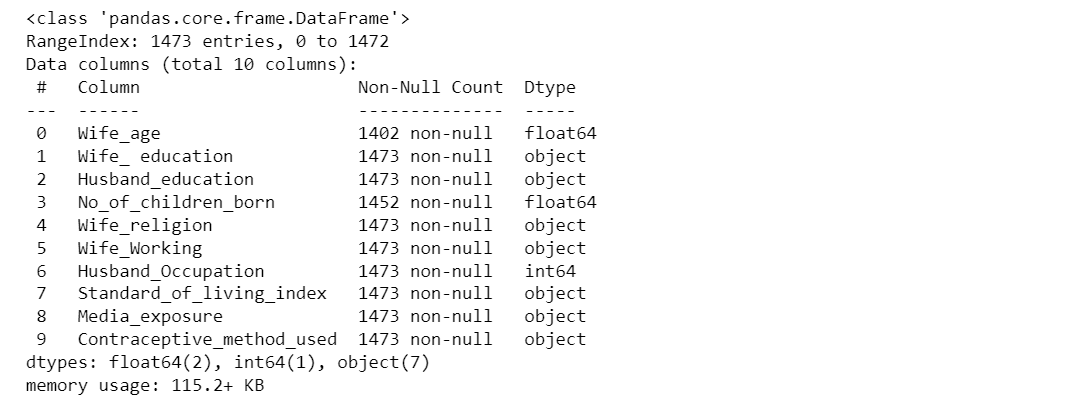
****

Table-20-Information of data set

**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= overflow, 2, 3, 4=high

9. Media exposure (binary) Good, not good

10. Contraceptive method used (class attribute) No, Yes

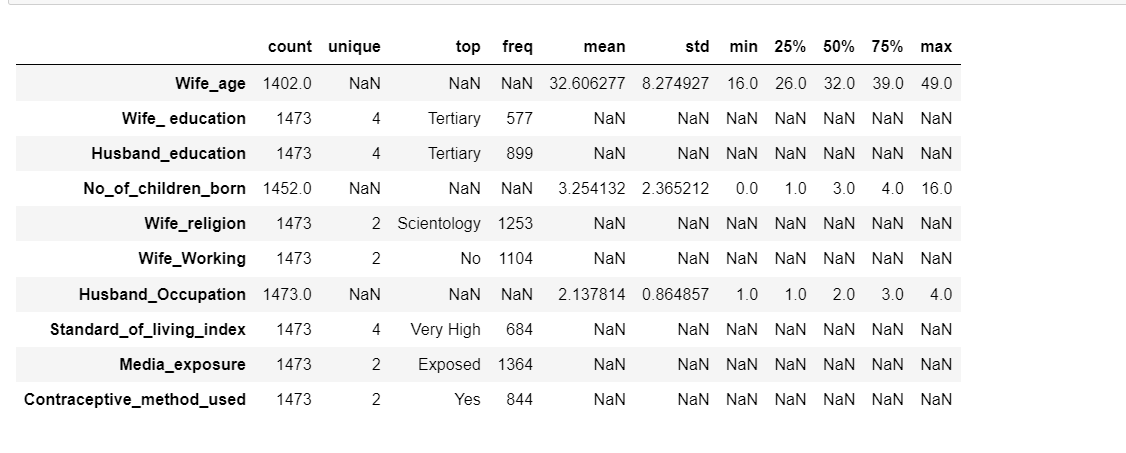


Table-21-Missing values in the data set

**Description of the complete dataset**:

Now let’s check the null and duplicate values of the given sample. If null values or duplicate data are present in the sample data set, we will treat them and if not, we will further proceed with the outlier’s detection and their treatment.

Table -22, Description of data set.



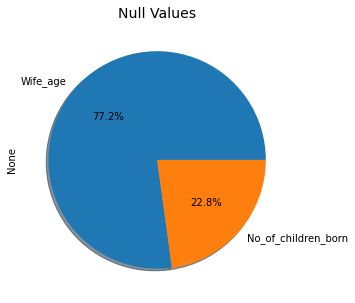


Fig 12 – Null values Percentage

**Observations:**

**Insights based on EDA**

* There is total 1473 rows and 10 columns in the dataset.
* Out of 10, 7 columns are of object type,1 columns of integer type and remaining 2 are of float type data.
* The variable "wife age" has 71 missing values and the variable "Number of children born" has 21 missing values. We need to impute these missing values using statistical central tendency method like mean or median. Here we are using Median method to impute the missing values.
* Contraceptive method used is the dependent variable or the target variable.
* There are 85 duplicate rows in the data set which has to be dropped first for further analysis.



Table-23- variables after null values implementation

* Checking duplicate rows - Number of duplicate rows = 85
* Let’s drop the duplicate values - Number of duplicate rows = 0

### Outliers and their Treatment:

### 

Fig 13– Boxplots before outliers Treatment

### Looking into the fields in the univariate analysis, we see outliers is present only in the field "number of children". The other two continuous variables "wife age" and "Husband Occupation" are not having any outliers. So, we will have to treat the outliers for further analysis.

### Removing the outliers using IQR method:

### IQR = Q3- Q1

### upper= Q3+(1.5 \* IQR)

### lower= Q1-(1.5 \* IQR)- Equation 4

### 

Fig 14– Boxplots After outliers Treatment

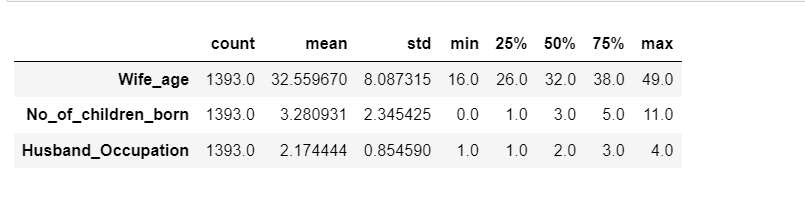


Table-24 Description after null values Imputation

### Univariate analysis:

### Let us define a function 'univariateAnalysis\_numeric' to display information as part of univariate analysis of numeric variables. The function will accept column name and number of bins as arguments.

### The function will display the statistical description of the numeric variable, histogram or density plot to view the distribution and the box plot to view 5-point summary and outliers if any.

**Plotting the histogram and boxplots of “All the numerical variables together”**

### 

### 

Fig 15– Histograms of Wife Age

### 

Fig 16– Boxplots of wife’s Age

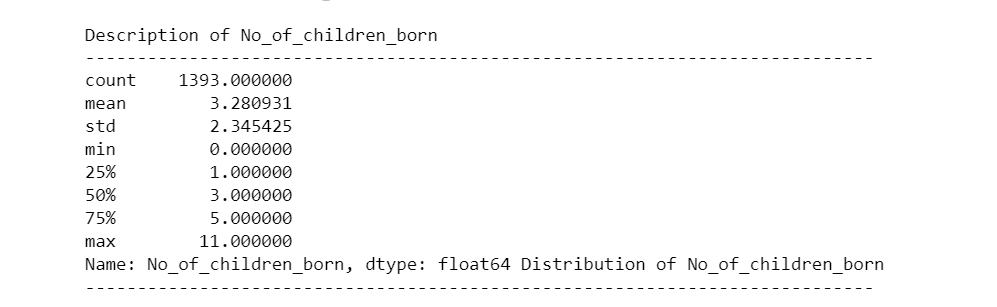
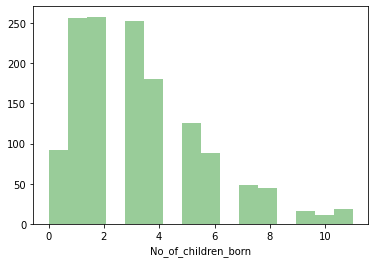
 

Fig 17– Histogram No of children Born

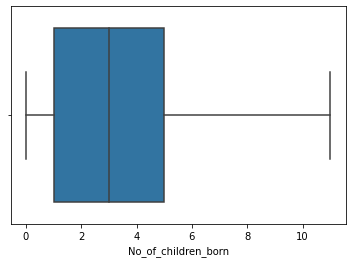
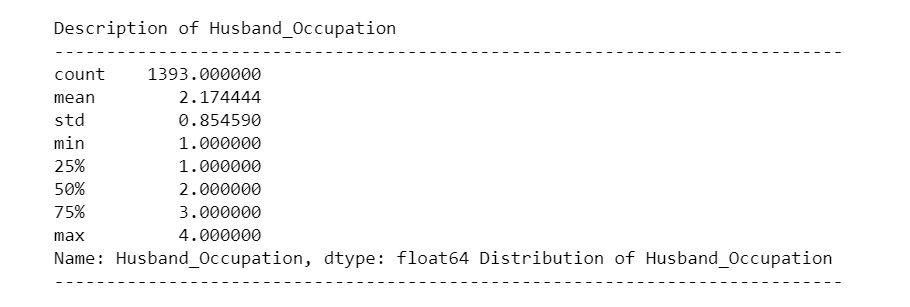


Fig 18– Boxplots No of children Born



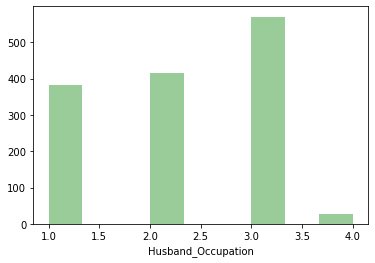


Fig 19– Histogram of Husband’s Occupation

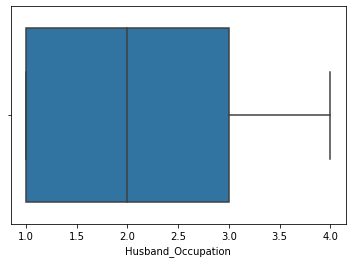


Fig 20 – Boxplot of Husband’s Occupation

Details of Wife\_ education

----------------------------------------------------------------

Tertiary 515

Secondary 398

Primary 330

Uneducated 150

Name: Wife\_ education, dtype: int64

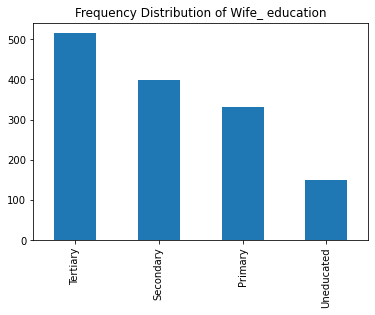


Fig 21 – Histogram of Wife Education

Details of Husband education

----------------------------------------------------------------

Tertiary 827

Secondary 347

Primary 175

Uneducated 44

Name: Husband education, dtype: int64

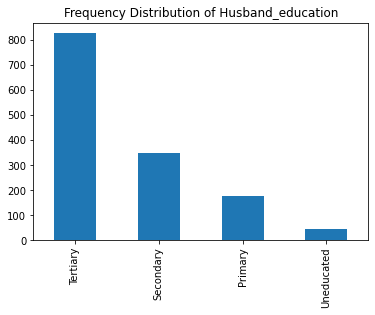


Fig 22 – Histogram of husband Education

Details of Wife religion

----------------------------------------------------------------

Scientology 1186

Non-Scientology 207

Name: Wife religion, dtype: int64

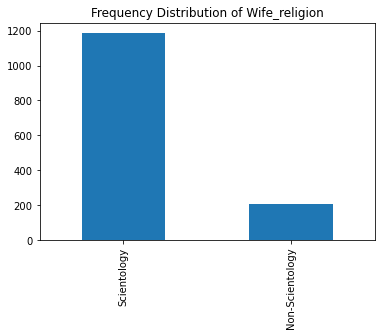


Fig 23 – Histogram of Wife Religion

Details of Wife Working

----------------------------------------------------------------

No 1043

Yes 350

Name: Wife Working, dtype: int64

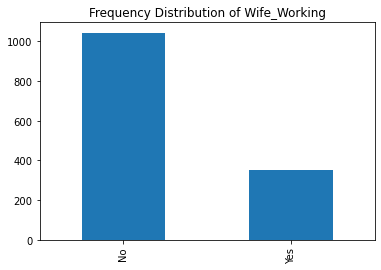


Fig 24 – Histogram of Wife Working frequency

Details of Standard of living index

----------------------------------------------------------------

Very High 618

High 419

Low 227

Very Low 129

Name: Standard of living index, dtype: int64

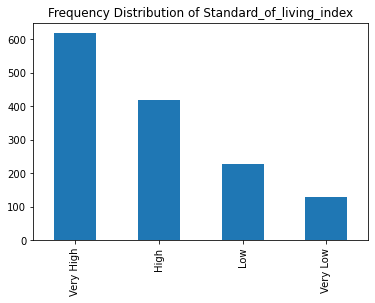


Fig 25 – Histogram of Wife Education

Details of Media exposure

----------------------------------------------------------------

Exposed 1284

Not-Exposed 109

Name: Media exposure, dtype: int64

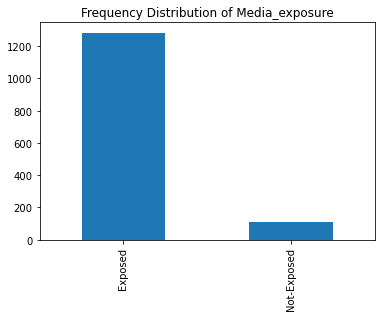


Fig 26 – Histogram of Media Exposure

Details of Contraceptive method used

----------------------------------------------------------------

Yes 779

No 614

Name: Contraceptive method used, dtype: int64

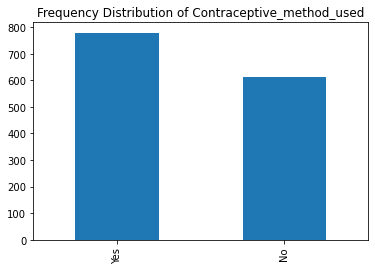


Fig 27 – Histogram of Target variable

### 

Fig 28 – Bivariate analysis of target variable and husband’s occupation

### 

Fig 29 – Multivariate analysis of target variable

### Insights:

### - We can see that in to the boxplot between target variable "Contraceptive method used" and the "No of children born", we see that, "No of children born" is high in the case of use of contraception used

### - Boxplot of Contraceptive Method Used and Husband Occupation shows that only Uneducated husbands have some effect on the method of contraceptive otherwise all the other categories like "secondary", "primary" and "tertiary" educated husbands show same variation.

### - Those who have very low standard of Living are having medians low as compared to other standards when compared to both the contraceptive methods used

### Bivariate and Multivariate analysis:

### 

Fig 30 – Pair plot of numerical variables

### 

### Fig 31 – Heat map of numerical variables

### Insights:

### - Bivariate and multivariate analysis indicates that there is slight positive correlation between the field’s “wife age” and “no of children born”. So, the pair plot shows almost no correlation between variables and hence there is no problem of multicollinearity.

### - But we don’t want high positive or high negative correlation because that will make the beta coefficients unreliable.

### - We also notice that there is almost no correlation between Husband's Occupation and Number of children Born.

### Part 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.?

### Data is split into 70:30 (Logistic Regression): -

### The split data of both independent and dependent trained data (70%) is fit into the logistic regression model to predict target variable

### Logistic regression uses newton-cg (solver), iteration inputs to execute the model

### Data consists of both categorical and numerical values.

### There are total of 1393 rows and 16 columns in the dataset after encoding the data set.

### Now all the variables are converted into Numerical data type for further analysis.

### Table-25-Contraceptive data set – After encoding

### 

### 

### Logistic Regression Model:

### We will create a model based on logistic regression and calculate the coefficients of the variables along with the predicted probabilities. Then based on that model we will interpret about our model weather the model is a good fit or overfit and underfit.

### Predicted probabilities on Training data set is as follows for the first five rows:

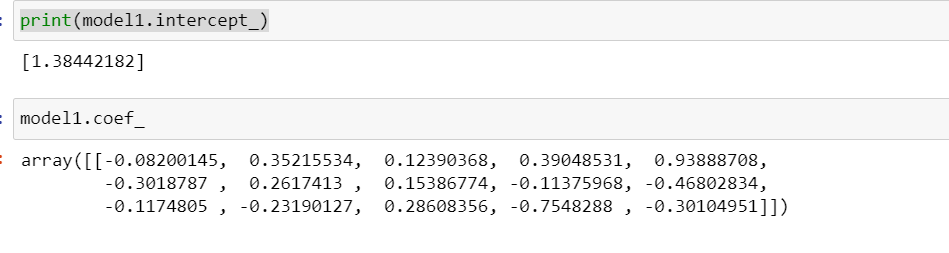
### 

### Table-26-Predicted probabilities on Train data for LR

### Predicted probabilities on Testing data set is as follows for the first five rows:

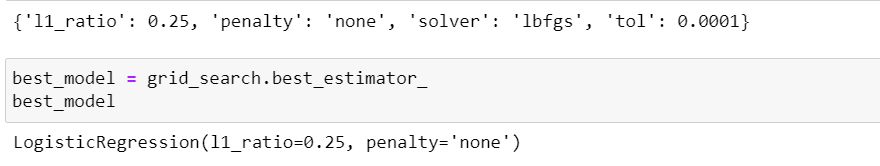
### 

### Table-27-Predicted probabilities on Test data for LR

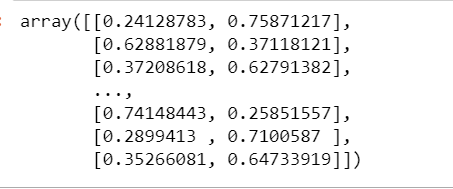


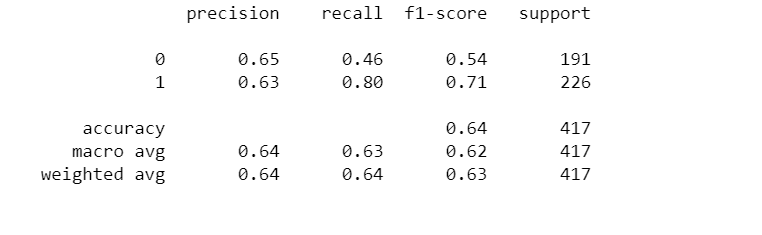
### Grid Search CV-

### 



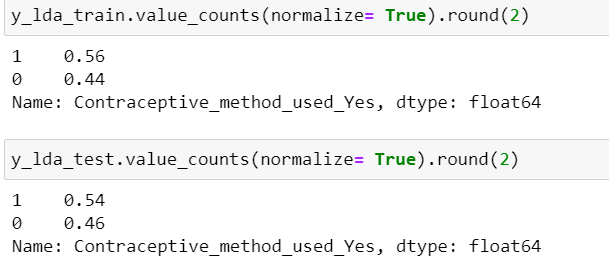
Best model predicted probability on training data:

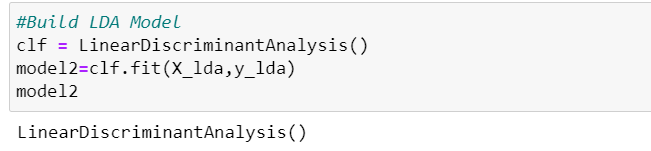




**Linear Discrimination Analysis:**

**Checking the class proportions of the train and test data set for LDA model**

****

****

### Class Label Predictions for LDA

### Class Label Predictions for LDA

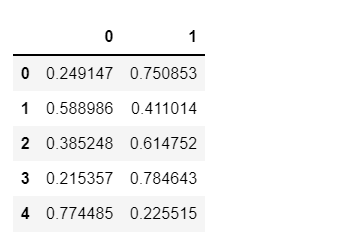


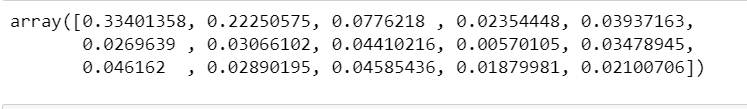
Table-28, Predicted Probabilities of LDA on TRAIN and Test data



**CART Model:**

****

Fig 32 – Tree flow chart of the cart model



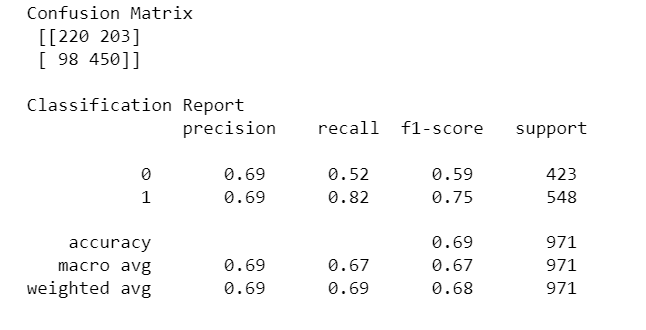
**Part 2.3 -Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.?**

****

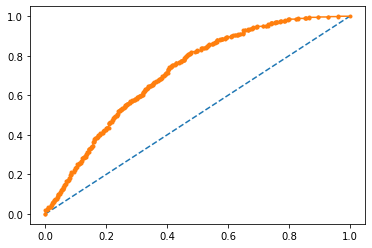
### Table-29-Predicted probabilities on Train data for LDA

**Evaluation of Logistic regression model:**

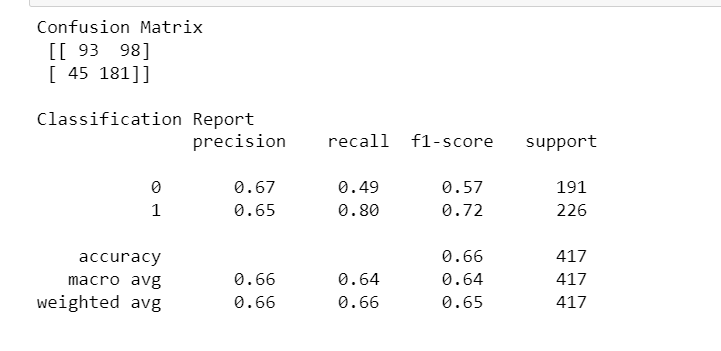
**Confusion Matrix – Training set:**

****

### Table-30-Classification report on training data -LR

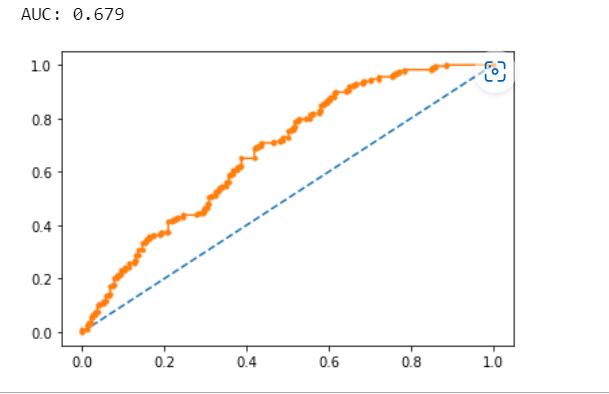
* True negative of logistic Regression model is – 220
* False Positive of Logistic Regression model is – 203
* False Negative of Logistic Regression model is - 98
* True positive of Logistic Regression model is – 450
* **Model score = 0.69001**
* **AUC – ROC curve for the training data set is given as follows where AUC = 0.715**
* ****
* Fig 33. AUC-ROC Curve- LR – Training

**Confusion Matrix – Testing set:**

****

### Table-31-Classification report on Testing data -LR

* True negative of logistic Regression model is – 93
* False Positive of Logistic Regression model is – 98
* False Negative of Logistic Regression model is - 45
* True positive of Logistic Regression model is – 181
* **Model score = 0.66**
* **AUC – ROC curve for the training data set is given as follows where AUC = 0.68**

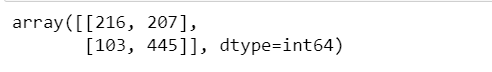
****

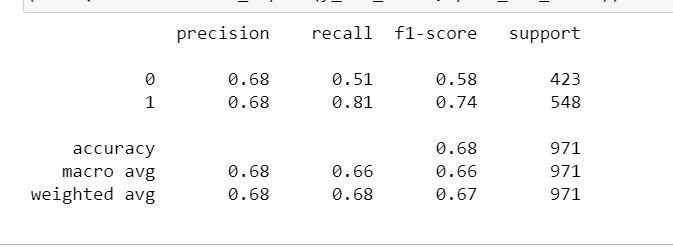
* Fig 34. AUC-ROC Curve- LR – TESTING

**Evaluation of LDA Model:**

Data is split into 70:30 (LDA-Linear discriminant analysis): - The split data of both independent and dependent trained data (70%) is fit into the LDA model to predict target variable.

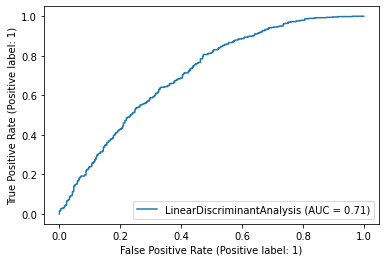
**Confusion Matrix – Training set:**

****

****

### Table-32-Classification report on training data -LDA

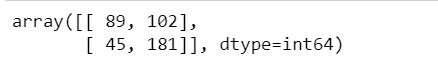
* True negative of logistic Regression model is – 216
* False Positive of Logistic Regression model is – 207
* False Negative of Logistic Regression model is - 103
* True positive of Logistic Regression model is – 445
* **Model score = 0.6801**
* **AUC – ROC curve for the training data set is given as follows where AUC = 0.715**

****

* Fig 35. AUC-ROC Curve- LDA – Training

**Evaluation of LDA Model:**

**Confusion Matrix – Testing set:**

****

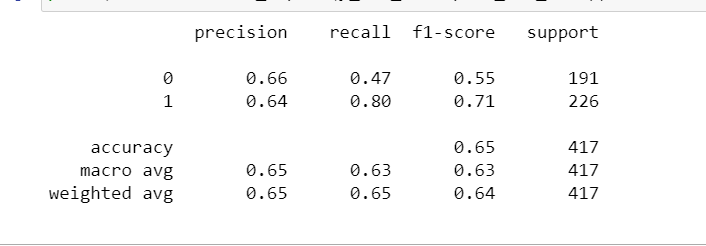
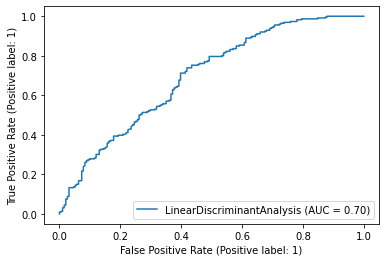
****

Table-33-Classification report on training data -LDA

* True negative of logistic Regression model is – 89
* False Positive of Logistic Regression model is – 102
* False Negative of Logistic Regression model is - 45
* True positive of Logistic Regression model is – 181
* Model score = 0.6501
* AUC – ROC curve for the training data set is given as follows where AUC = 0.715

****

* Fig 36. AUC-ROC Curve- LDA – TESTING

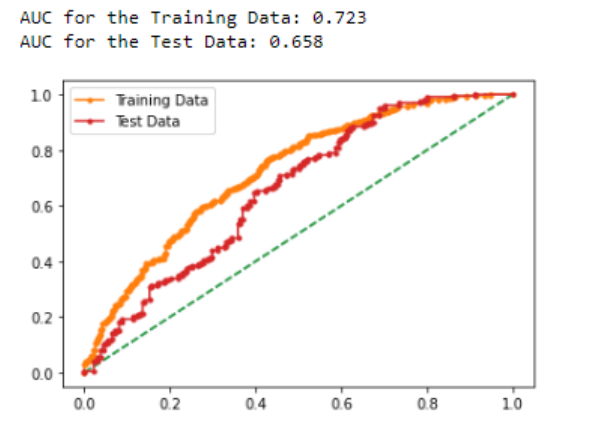
**Insights:**

1. Logistic Regression Model and LDA are able to predict the employee’s behaviour with only 66% and 65% accuracy respectively. Although both of these are not good models for our predictions but in this scenario, we can take Logistic Regression model into consideration.

2. As per the Performance Metrics computed above for all the models it can be evidently seen that the best model for our prediction is Logistic Regression Model over LDA as all the performance metrics are comparatively slightly higher than LDA.

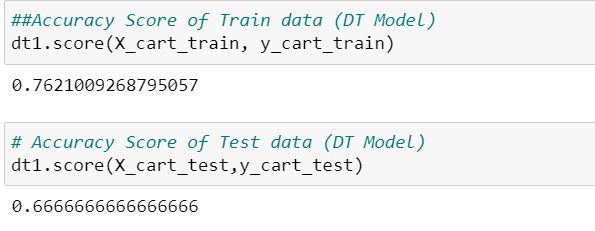
3. The employees opting for the Contraceptive method used as "yes" are correctly identified by approximately 66% of the times by Logistic Regression Model whereas LDA is predicting these types of employees by 65% of the times as observed by area under the curve. So as to increase the use of contraceptive method used as yes, it is concluded that Logistic Regression Model is favoured as it maximises the chances for discriminating between the two classes i.e., between the ones opting for the contraceptive from the ones not opting for contraceptive method. But it also depends on various kinds of social factors

4. Wife age and number of children born are two most important factors that can be deciding factor for the use of contraceptive methods.



**Evaluation of Cart Model:**

Decision Tree Classifier



**Importance of features of various variables:**

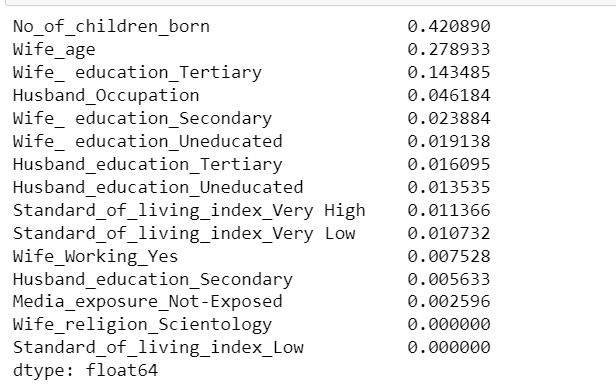
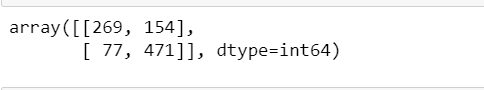
****

Table-34-Impotance of features in Cart model on training data

**Evaluation of Cart Model:**

**Confusion Matrix – Training data:**

****

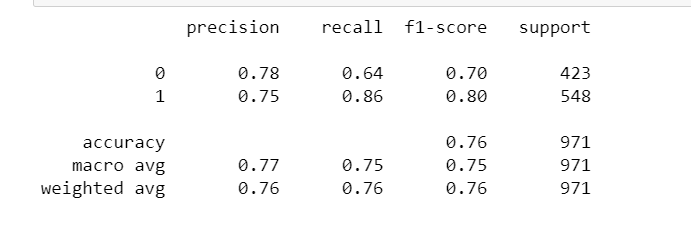
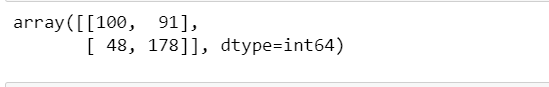
****

Table-35-Classification Report on Training data for CART

Obtaining AUC of DT Model on Test data: 0.8415169712343186

**Confusion Matrix – Testing data:**

****

****

Table-36-Classification Report on Testing data for CART

Obtaining AUC of DT Model on Test data: 0.7085669276745588

Cart is performing slightly better than LDA and Logistic Regression as we can see the accuracy on the test data for Cart model is around 67%

**Model Comparison: On Testing Data and Training Data**

**Logistic regression Accuracy –**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Logistic Train** | **Logistic Test** | **LDA Train** | **LDA Test** | **CART Train** | **CART Test** |
| **Accuracy** | 0.69 | 0.66 | 0.68 | 0.65 | 0.76 | 0.67 |
| **Recall** | 0.82 | 0.8 | 0.81 | 0.8 | 0.86 | 0.79 |
| **F1 Score** | 0.75 | 0.72 | 0.74 | 0.71 | 0.8 | 0.72 |
| **Precision** | 0.69 | 0.65 | 0.68 | 0.64 | 0.75 | 0.66 |
| **AUC** | 0.715 | 0.68 | 0.71 | 0.7 | 0.84 | 0.7 |

**Part 2.4 - Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.?**

**Insights from Logistic regression:**

**For predicting ‘Contraceptive method used’ is “No” (Label 0):**

* Precision (67%) – 69% of the people predicted are actually not using contraceptives out of all families predicted to have been not using contraceptives.
* Recall (49%) – Out of all the people not using contraceptives ,49% of families have been predicted correctly.

**For predicting ‘Contraceptive method used is “Yes” (Label 1):**

* Precision (65%) – 69% of people predicted are actually using contraceptives out of all people predicted to have been not using contraceptives
* Recall (80%) – Out of all the people using contraceptives ,80% of families have been predicted correctly

**Overall accuracy of the model – 66 % for test data which means 66% of total predictions are correct**

* Accuracy, AUC, Precision and Recall for test data is almost in line with training data.
* AUC of Train data = almost 71%
* AUC of Test data = almost 69%
* This proves no overfitting or underfitting has happened, and overall, the model is a good model for classification.

**Insights from LDA:**

**For predicting ‘Contraceptive method used’ is “No” (Label 0):**

* Precision (66%) – 65% of the people predicted are actually not using contraceptives out of all families predicted to have been not using contraceptives.
* Recall (47%) – Out of all the people not using contraceptives ,47% of families have been predicted correctly.

**For predicting ‘Contraceptive method used’ is “Yes” (Label 1):**

* Precision (64%) – 65% of people predicted are actually using contraception’s out of all people predicted to have been not using contraception’s
* Recall (80%) – Out of all the people using contraception’s ,81% of families have been predicted correctly
* Overall accuracy of the model – 65 % of total predictions are correct

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

**CONCLUSION:**

We can see that AUC curve of Both LDA, and CART shows similar results which is around 70% for test data and 68% in case of Logistic regression, hence we can say both LDA and CART can be a good model in this case but when it comes to Accuracy for Logistic Regression and CART are performing better than LDA which has the lowest accuracy on Test data which is around 65% but we can say that for the given data set all the models are almost similar and reasonable.

*Reference – Great Learning lecture videos and Mentors*