Predictive Modeling-Coded project -Business Report (By Sowmya Subramaniam)

TABLE OF CONTENTS:

1. LINEAR REGRESSION

1.1 Read the data and do exploratory data analysis.
a) Reading the data and performing basic analysis
b) EDA (Univariate, Bivariate, Multivariate analysis)
c) Data insights

1.2 Impute null values if present, check for 0 values. Also check for outliers and duplicates if there.
a) Check for missing values and treating them
b) Plotting linear relationship between usr and other features c) Check for 0 values in independent variables
d) Check for Outliers
e) Check for duplicates
f) Data insights

1.3 Encode the data (having string values) for Modelling. Split the data into train and test . Apply Linear regression using scikit learn.
a) Dummy encoding
b) Train-Test split

c) Decision tree model
d) Decision tree feature importance
e) Linear regression model
f) R-square, RMSE of model
g) Data insights

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

—----------

2.1 Data Ingestion:

Read the dataset. Do the descriptive statistics and EDA.
a) Reading the data and performing basic analysis
b) EDA (Univariate, Bivariate, Multivariate analysis)
c) Data insights
d) Check for missing values and treating them
e) Check for Duplicate records
f) EDA 2.2 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. a) Data encoding
b) Train-Test split
c) Logistic regression model

d) LDA
e) CART

2.3 Performance Metrics:
a) Predictions on train and test data
b) Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model
c) Comparison of models

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

==========

PROBLEM 1

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, and Multivariate Analysis. Solution:

DATASET:

The dataset contains 8192 records having 22 features in which 21 are descriptive features and one target variable.

GOAL:
to find out how each attribute affects the system to be in 'usr' mode using a list of system attributes.

TARGET FEATURE:
The response variable is “usr” which denotes portion of time(%) cpu runs in user mode. Loading and viewing the dataset. Printing the first five rows using head ():

First 5 records

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap	usr
0	1	0	2147	79	68	0.2	0.2	40671.0	53995.0	0.0	...	0.0	0.0	1.6	2.6	16.00	26.40	CPU_Bound	4670	1730946	95
1	0	0	170	18	21	0.2	0.2	448.0	8385.0	0.0	...	0.0	0.0	0.0	0.0	15.63	16.83	Not_CPU_Bound	7278	1869002	97
2	15	3	2162	159	119	2.0	2.4	NaN	31950.0	0.0	...	0.0	1.2	6.0	9.4	150.20	220.20	Not_CPU_Bound	702	1021237	87
3	0	0	160	12	16	0.2	0.2	NaN	8670.0	0.0	...	0.0	0.0	0.2	0.2	15.60	16.80	Not_CPU_Bound	7248	1863704	98
4	5	1	330	39	38	0.4	0.4	NaN	12185.0	0.0	...	0.0	0.0	1.0	1.2	37.80	47.60	Not_CPU_Bound	633	1760253	90

5 rows × 22 columns

Last 5 records using tail()

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap	usr
8187	16	12	3009	360	244	1.6	5.81	405250.0	85282.0	8.02	...	55.11	0.6	35.87	47.90	139.28	270.74	CPU_Bound	387	986647	80
8188	4	0	1596	170	146	2.4	1.80	89489.0	41764.0	3.80	...	0.20	0.8	3.80	4.40	122.40	212.60	Not_CPU_Bound	263	1055742	90
8189	16	5	3116	289	190	0.6	0.60	325948.0	52640.0	0.40	...	0.00	0.4	28.40	45.20	60.20	219.80	Not_CPU_Bound	400	969106	87
8190	32	45	5180	254	179	1.2	1.20	62571.0	29505.0	1.40	...	18.04	0.4	23.05	24.25	93.19	202.81	CPU_Bound	141	1022458	83
8191	2	0	985	55	46	1.6	4.80	111111.0	22256.0	0.00	...	0.00	0.2	3.40	6.20	91.80	110.00	CPU_Bound	659	1756514	94

5 rows × 22 columns

Checking the shape of dataset:

The dataset has 8192 rows and 22 columns

Checking the datatypes:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8192 entries, 0 to 8191
Data columns (total 22 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   lread     8192 non-null   int64
 1   lwrite    8192 non-null   int64
 2   scall     8192 non-null   int64
 3   sread     8192 non-null   int64
 4   swrite    8192 non-null   int64
 5   fork      8192 non-null   float64
 6   exec      8192 non-null   float64
 7   rchar     8088 non-null   float64
 8   wchar     8177 non-null   float64
 9   pgout     8192 non-null   float64
 10  ppgout    8192 non-null   float64
 11  pgfree    8192 non-null   float64
 12  pgscan    8192 non-null   float64
 13  atch      8192 non-null   float64
 14  pgin      8192 non-null   float64
 15  ppgin     8192 non-null   float64
 16  pflt      8192 non-null   float64
 17  vflt      8192 non-null   float64
 18  runqsz    8192 non-null   object
 19  freemem   8192 non-null   int64
 20  freeswap  8192 non-null   int64
 21  usr       8192 non-null   int64
dtypes: float64(13), int64(8), object(1)
memory usage: 1.4+ MB
```

There are three different types of datatypes. Dtypes: float64(13), int64(8), object(1). Most of the columns in the data are numeric in nature ('int64' or 'float64' type). The “runqsz” is a string column ('object' type)

Describing the summary of the dataset.

The five number summary involves the calculation of 5 summary statistical quantities, namely:

Median: The middle value in the sample, also called 50th percentile or the 2 nd quartile.

1 st Quartile: The 25th percentile.

3 rd Quartile: The 75th percentile.

Minimum value: The lowest value.

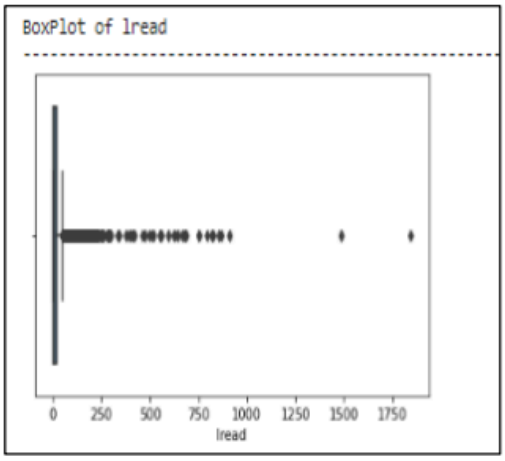
Maximum value: The highest value.

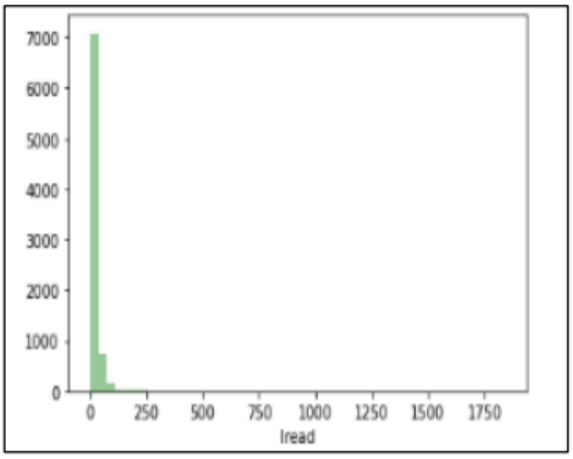
OUTPUT:

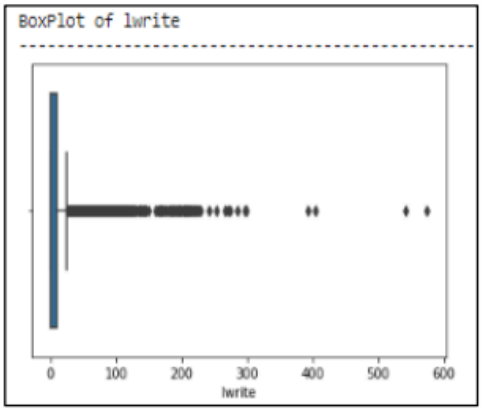
	count	mean	std	min	25%	50%	75%	max
lread	8192.0	1.955969e+01	53.353799	0.0	2.00	7.0	20.000	1845.00
lwrite	8192.0	1.310620e+01	29.891726	0.0	0.00	1.0	10.000	575.00
scall	8192.0	2.306318e+03	1633.617322	109.0	1012.00	2051.5	3317.250	12493.00
sread	8192.0	2.104800e+02	198.980146	6.0	86.00	166.0	279.000	5318.00
swrite	8192.0	1.500582e+02	160.478980	7.0	63.00	117.0	185.000	5456.00
fork	8192.0	1.884554e+00	2.479493	0.0	0.40	0.8	2.200	20.12
exec	8192.0	2.791998e+00	5.212456	0.0	0.20	1.2	2.800	59.56
rchar	8192.0	1.964728e+05	238446.012054	278.0	34860.50	125473.5	265394.750	2526649.00
wchar	8192.0	9.581275e+04	140728.464118	1498.0	22977.75	46619.0	106037.000	1801623.00
pgout	8192.0	2.285317e+00	5.307038	0.0	0.00	0.0	2.400	81.44
ppgout	8192.0	5.977229e+00	15.214590	0.0	0.00	0.0	4.200	184.20
pgfree	8192.0	1.191971e+01	32.363520	0.0	0.00	0.0	5.000	523.00
pgscan	8192.0	2.152685e+01	71.141340	0.0	0.00	0.0	0.000	1237.00
atch	8192.0	1.127505e+00	5.708347	0.0	0.00	0.0	0.600	211.58
pgin	8192.0	8.277960e+00	13.874978	0.0	0.60	2.8	9.765	141.20
ppgin	8192.0	1.238859e+01	22.281318	0.0	0.60	3.8	13.800	292.61
pflt	8192.0	1.097938e+02	114.419221	0.0	25.00	63.8	159.600	899.80
vflt	8192.0	1.853158e+02	191.000603	0.2	45.40	120.4	251.800	1365.00
freemem	8192.0	1.763456e+03	2482.104511	55.0	231.00	579.0	2002.250	12027.00
freeswap	8192.0	1.328126e+06	422019.426957	2.0	1042623.50	1289289.5	1730379.500	2243187.00
usr	8192.0	8.396887e+01	18.401905	0.0	81.00	89.0	94.000	99.00

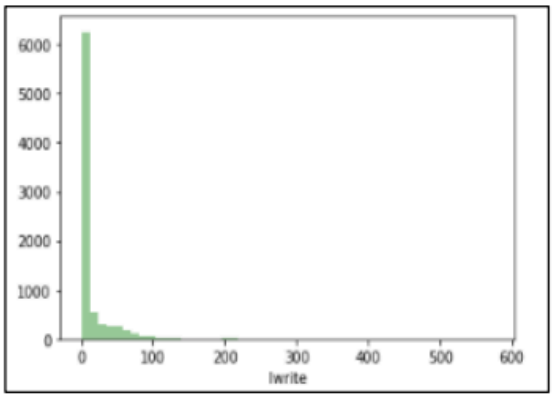
Boxplots are used to graphically represent the distribution of the data using Five Number summary values. It is one of the most efficient ways to detect outliers in our dataset

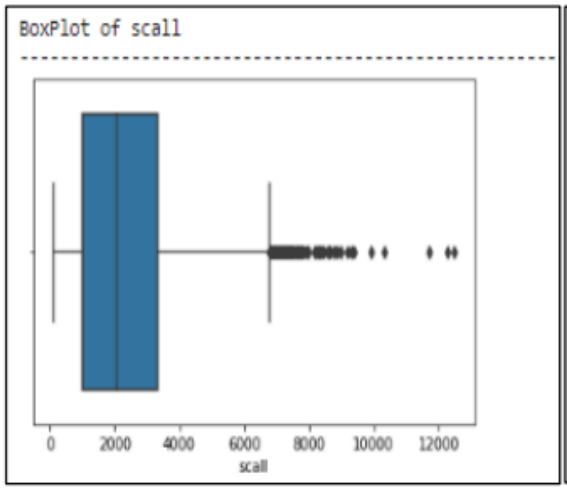
UNIVARIATE ANALYSIS:

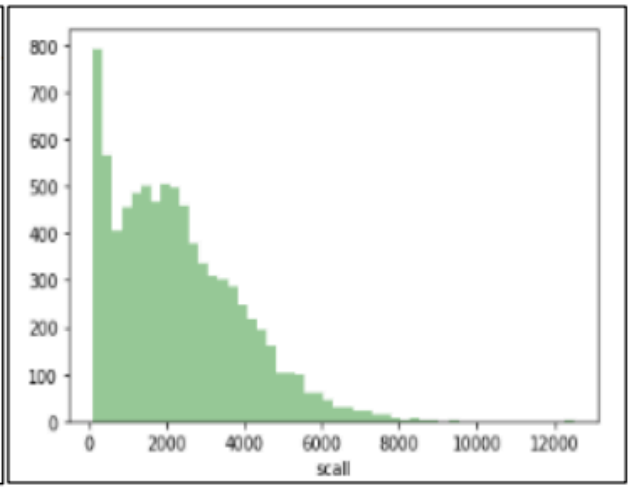


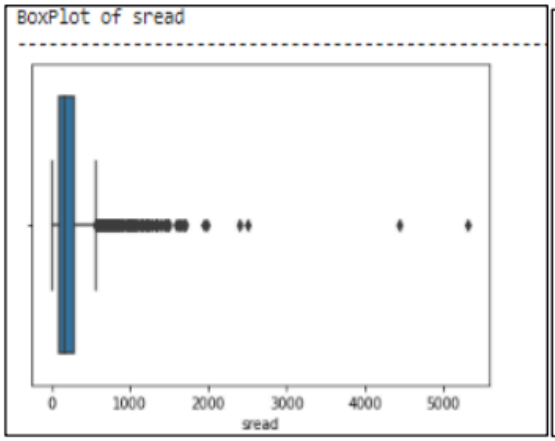


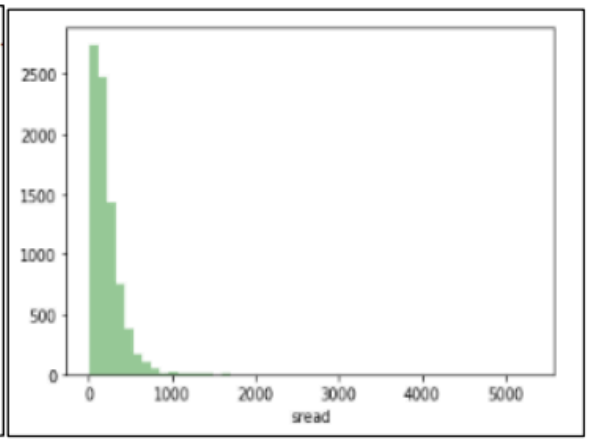


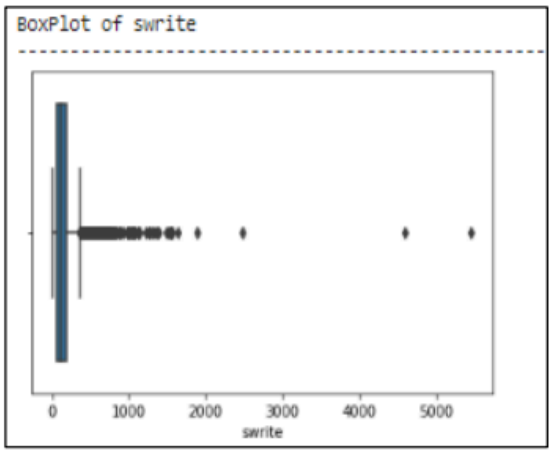


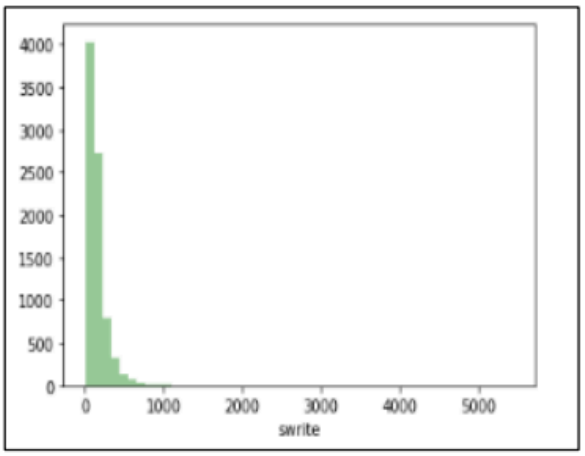

BoxPlot of scall
0
2000
4000
6000
8000
10000
12000
scall

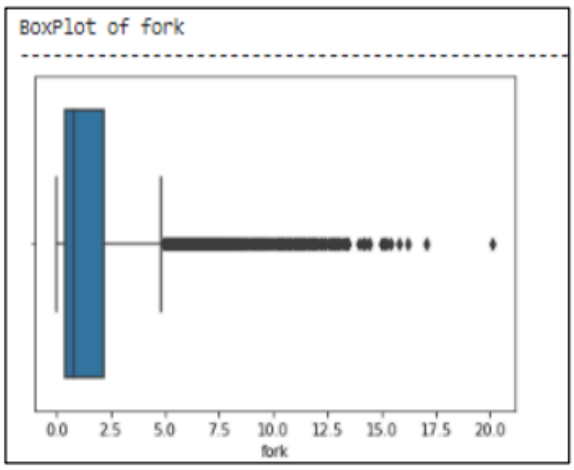

800
700
600
500
400
300
200
100
0
0
2000
4000
6000
8000
10000
12000
scall

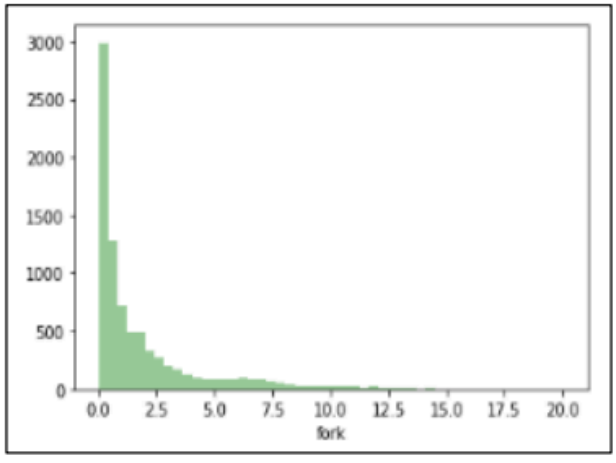

BoxPlot of sread
0
1000
2000
3000
4000
5000
sread

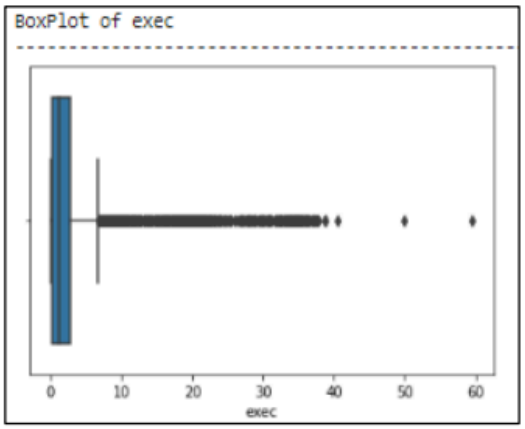

2500
2000
1500
1000
500
0
0
1000
2000
3000
4000
5000
sread

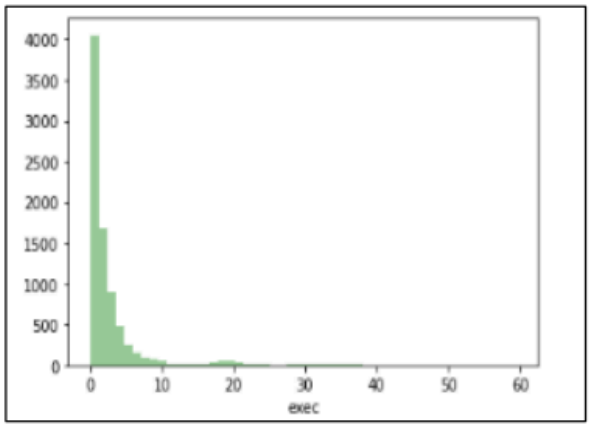

BoxPlot of swrite
0
1000
2000
3000
4000
5000
swrite

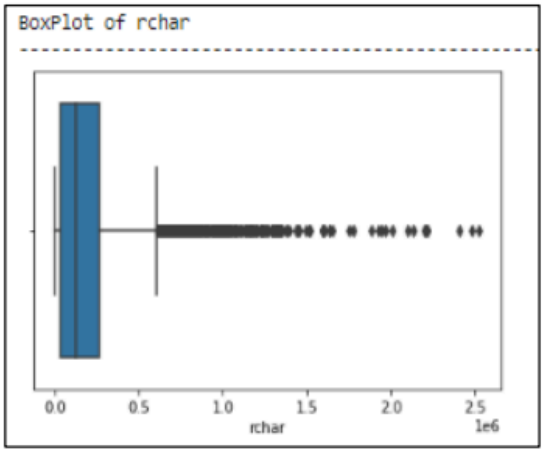

4000
3500
3000
2500
2000
1500
1000
500
0
0
1000
2000
3000
4000
5000
swrite

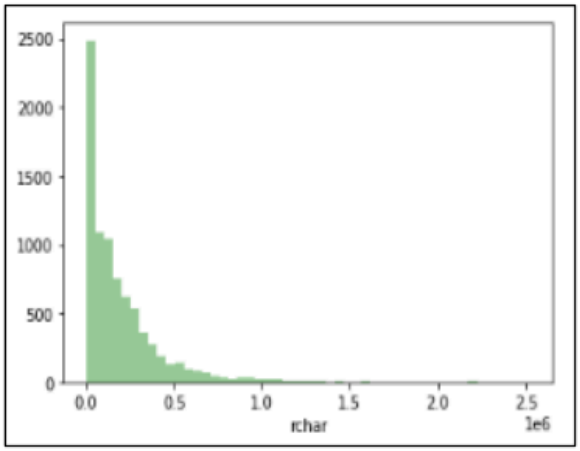

BoxPlot of fork
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
fork

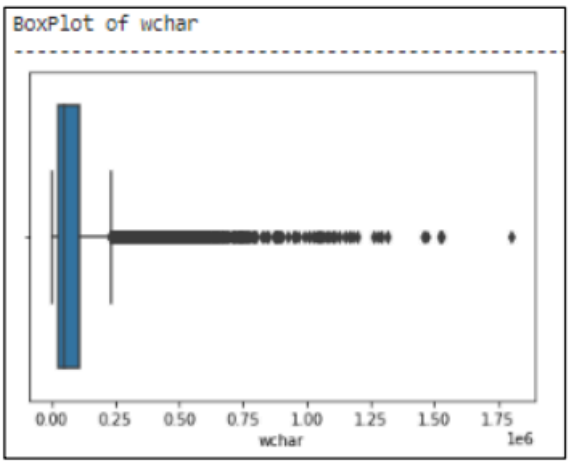

3000
2500
2000
1500
1000
500
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
fork

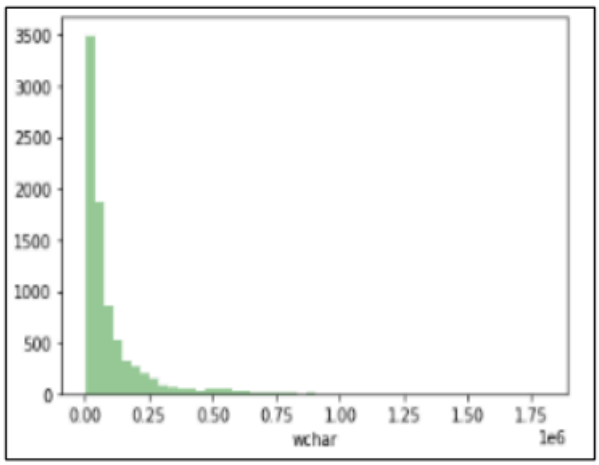

BoxPlot of exec
0
10
20
30
40
50
60
exec

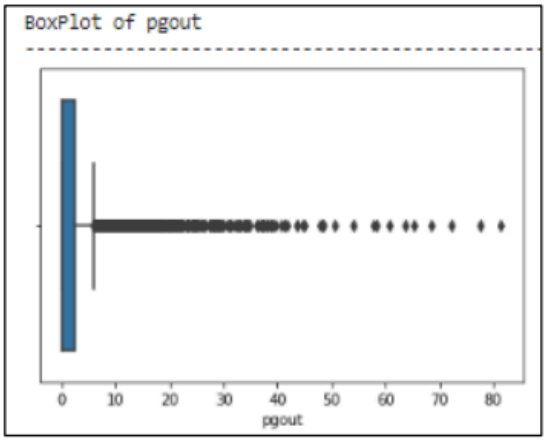

4000
3500
3000
2500
2000
1500
1000
500
0
0
10
20
30
40
50
60
exec

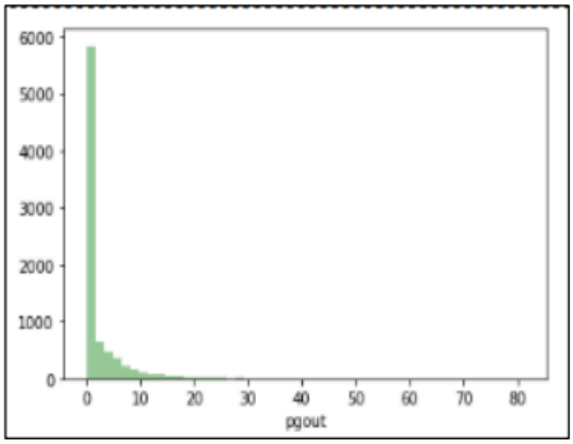

BoxPlot of rchar
0.0
0.5
1.0
1.5
2.0
2.5
rchar
1e6

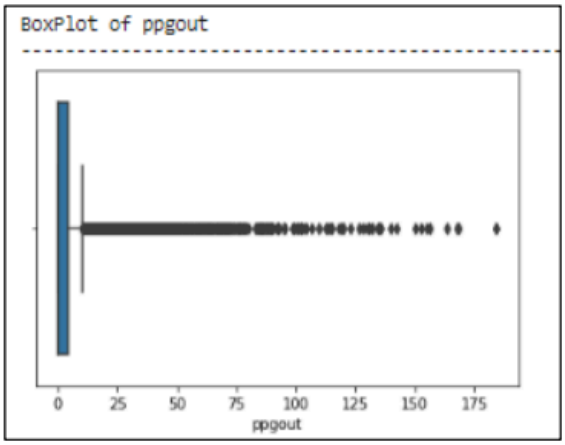

2500
2000
1500
1000
500
0
0.0
0.5
1.0
1.5
2.0
2.5
rchar
1e6

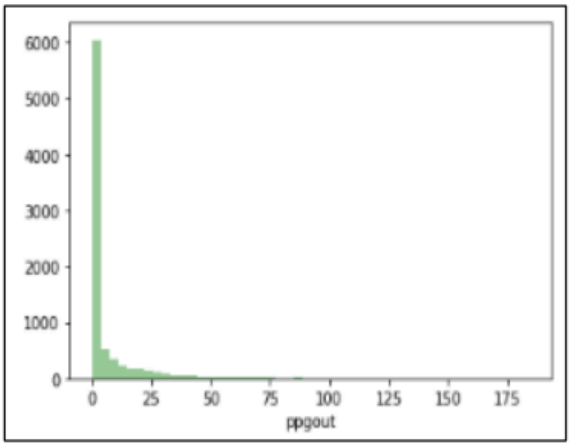

BoxPlot of wchar
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
wchar
1e6

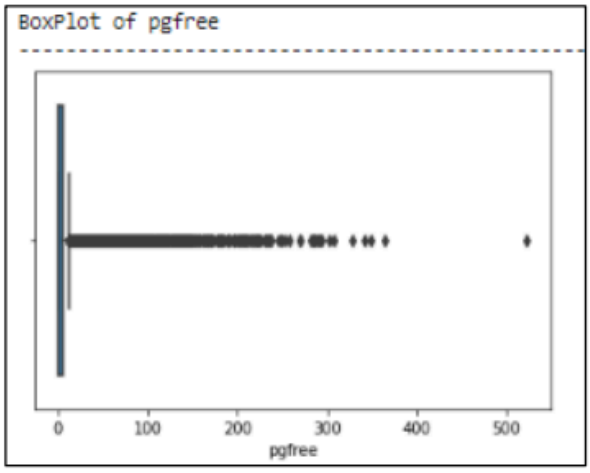

3500
3000
2500
2000
1500
1000
500
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
wchar
1e6

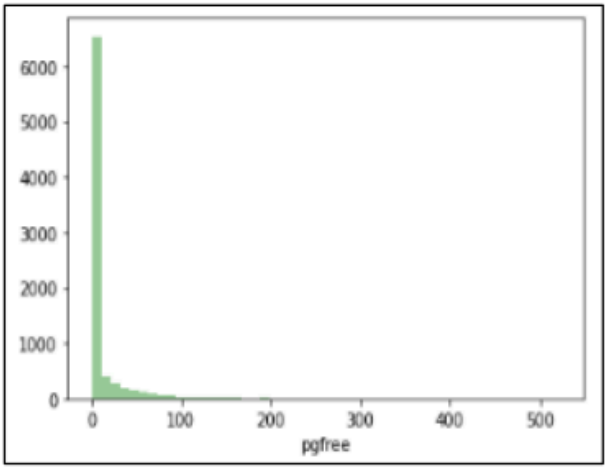

BoxPlot of pgout
0 10 20 30 40 50 60 70 80
pgout

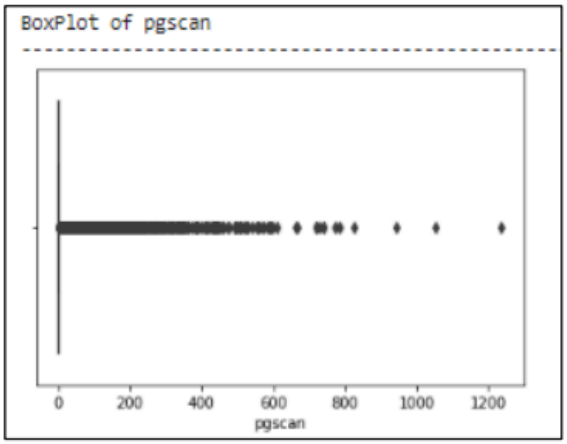

6000
5000
4000
3000
2000
1000
0
0 10 20 30 40 50 60 70 80
pgout

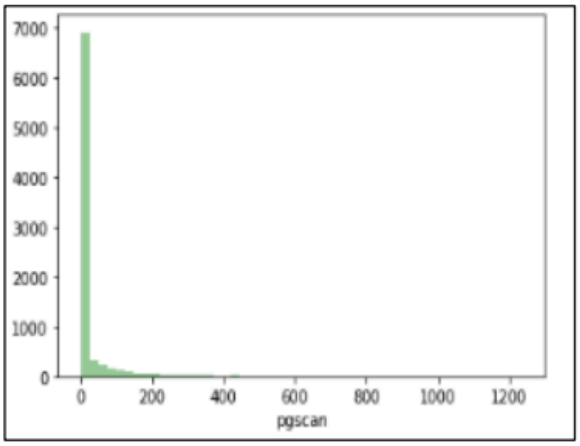

BoxPlot of ppgout
0 25 50 75 100 125 150 175
ppgout

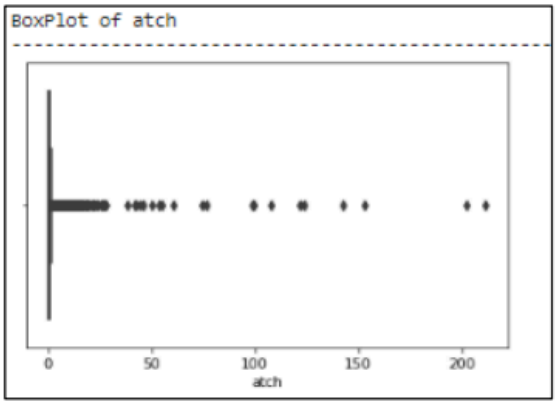

6000
5000
4000
3000
2000
1000
0
0 25 50 75 100 125 150 175
ppgout

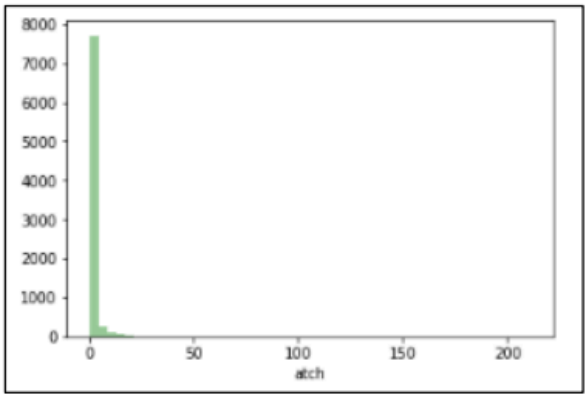

BoxPlot of pgfree
0 100 200 300 400 500
pgfree

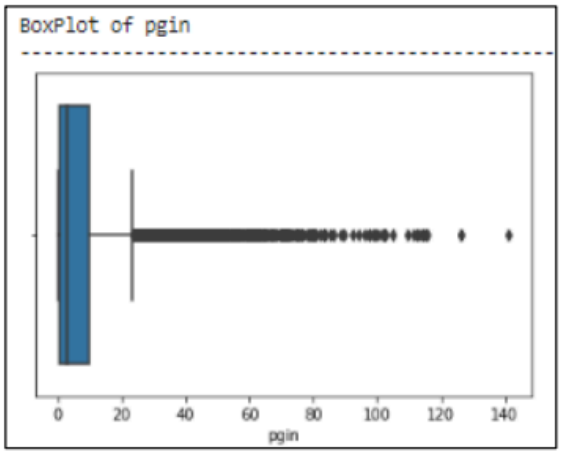

6000
5000
4000
3000
2000
1000
0
0 100 200 300 400 500
pgfree

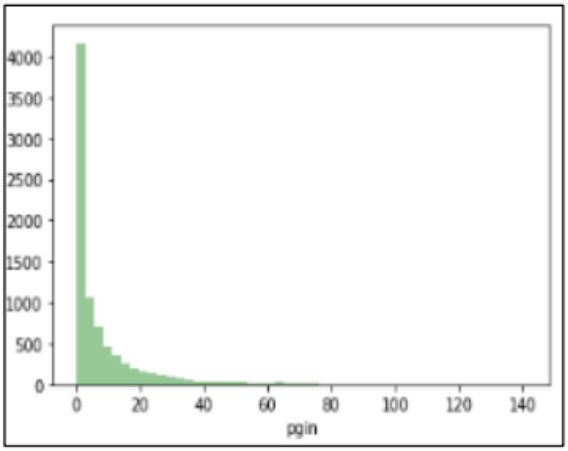

BoxPlot of pgscan
0
200
400
600
800
1000
1200
pgscan

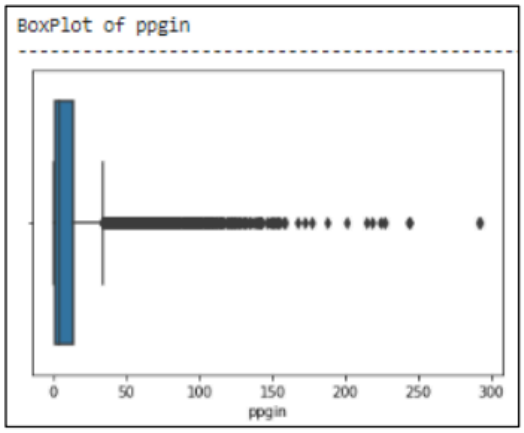

7000
6000
5000
4000
3000
2000
1000
0
0
200
400
600
800
1000
1200
pgscan

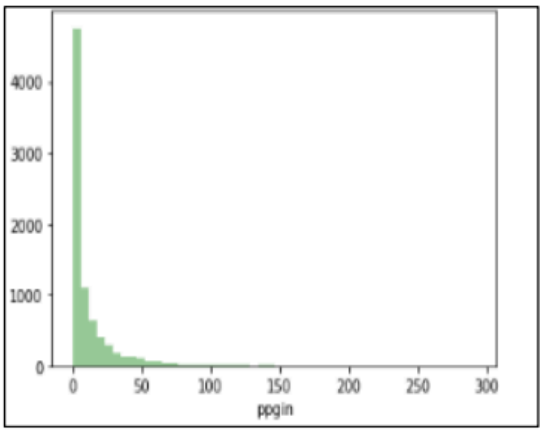

BoxPlot of atch
0
50
100
150
200
atch

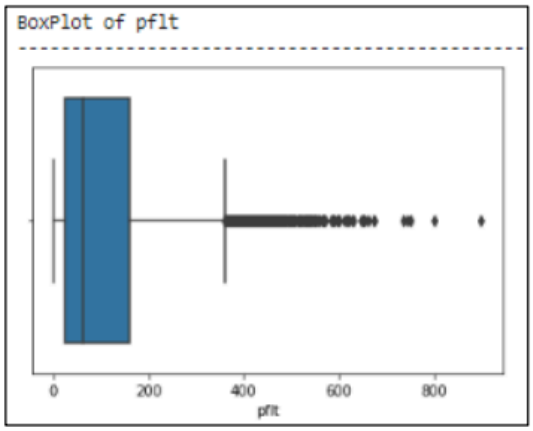

8000
7000
6000
5000
4000
3000
2000
1000
0
0
50
100
150
200
atch

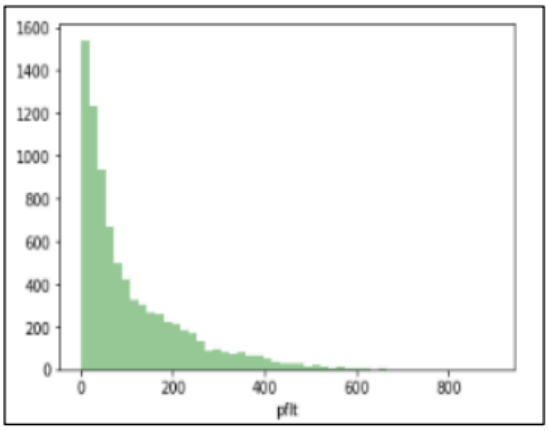

BoxPlot of pgin
0
20
40
60
80
100
120
140
pgin

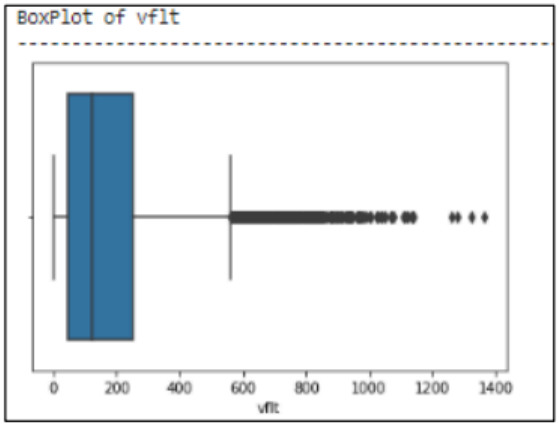

4000
3500
3000
2500
2000
1500
1000
500
0
0
20
40
60
80
100
120
140
pgin

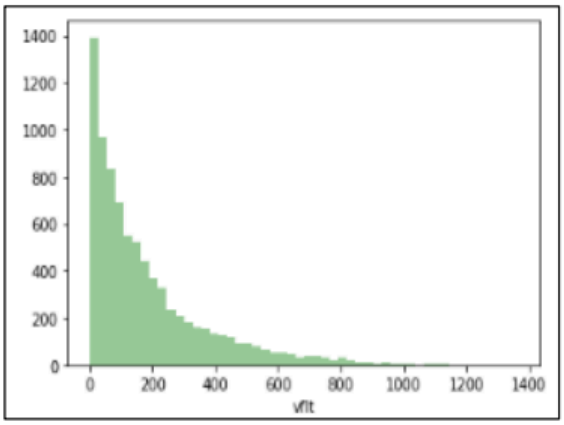

BoxPlot of ppgin
0
50
100
150
200
250
300
ppgin


4000
3000
2000
1000
0
0
50
100
150
200
250
300
ppgin

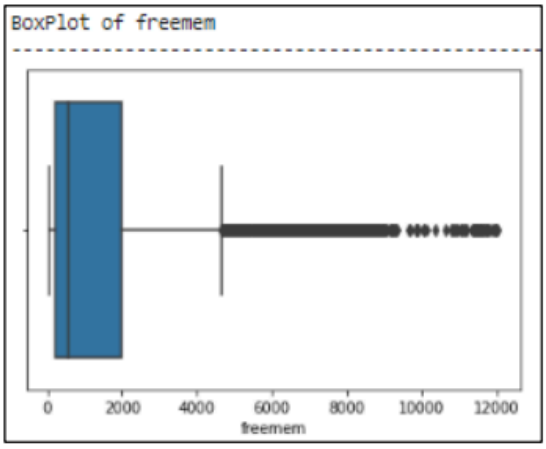

BoxPlot of pflt
0
200
400
600
800
pflt

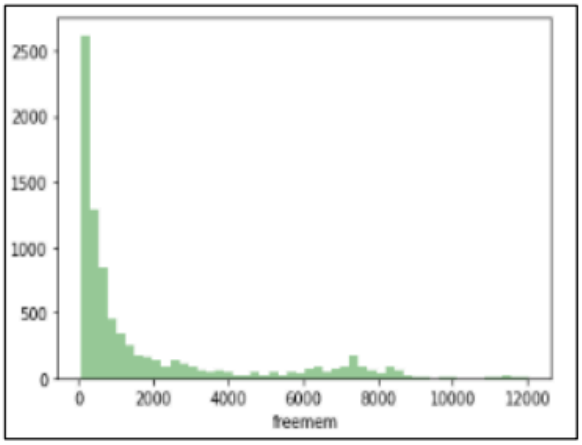

1600
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
pflt


BoxPlot of vflt
0
200
400
600
800
1000
1200
1400
vflt

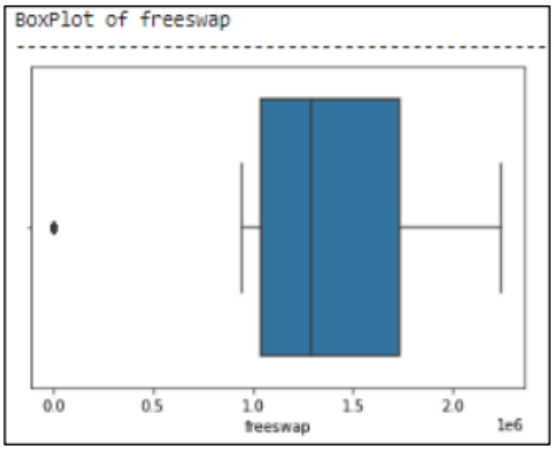

1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
vflt

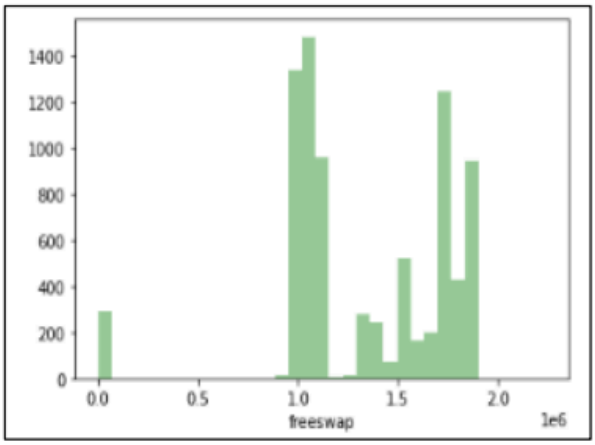
BoxPlot of freemem



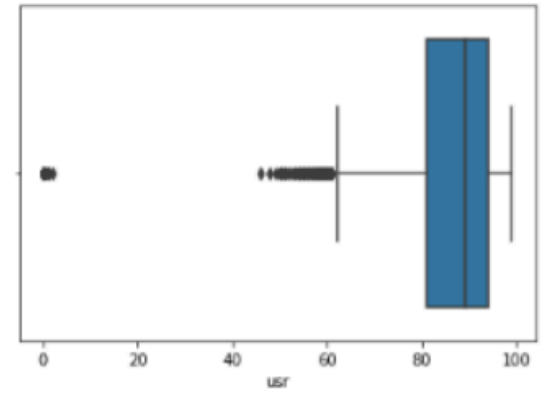


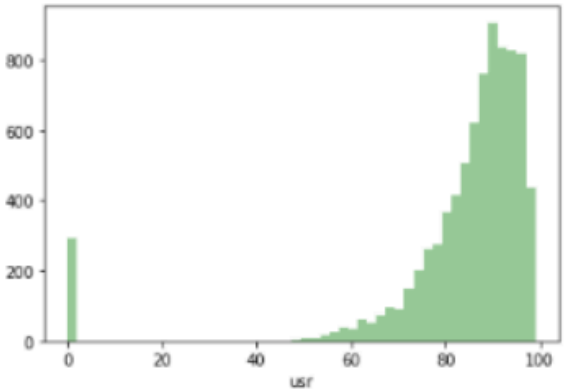
BoxPlot of freeswap





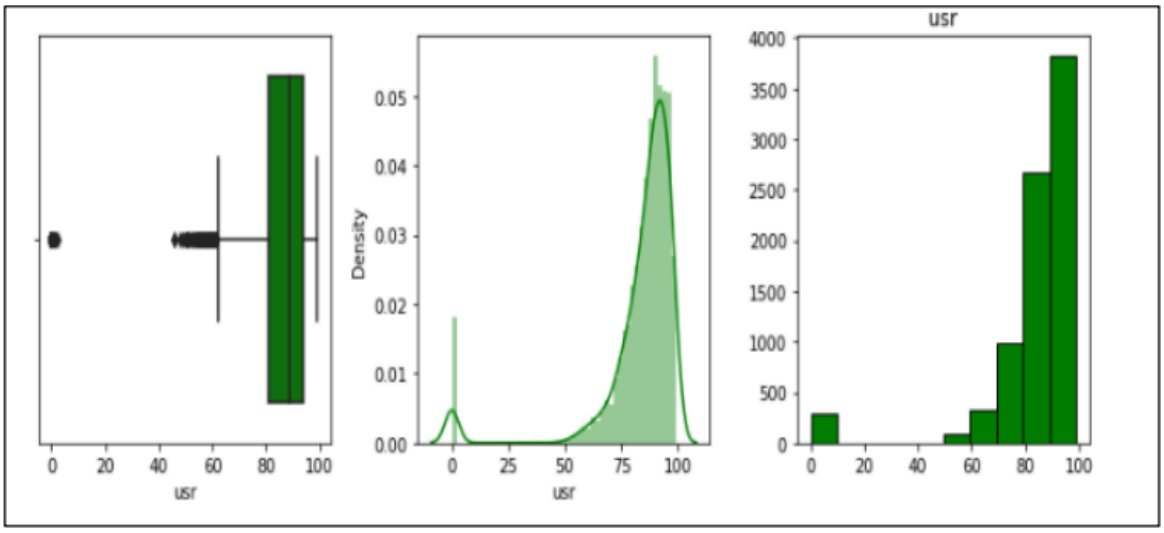
BoxPlot of usr



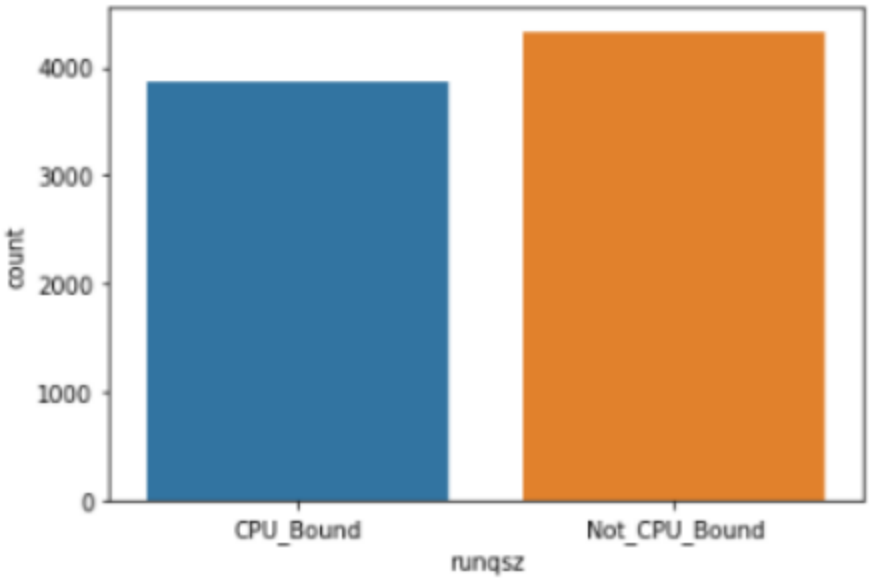


Data is right skewed for all the numeric fields except freeswap and usr

Boxplots, Distplots and Histplot of “usr” field

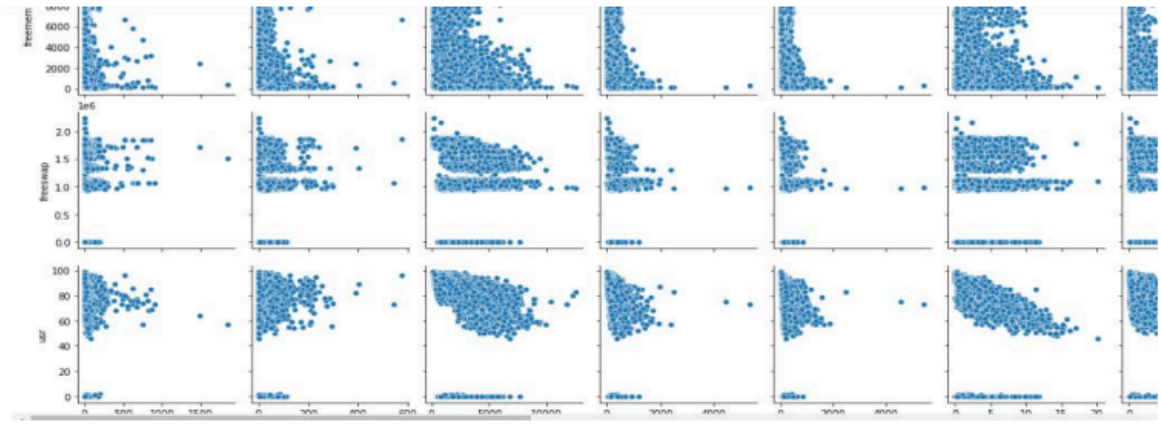


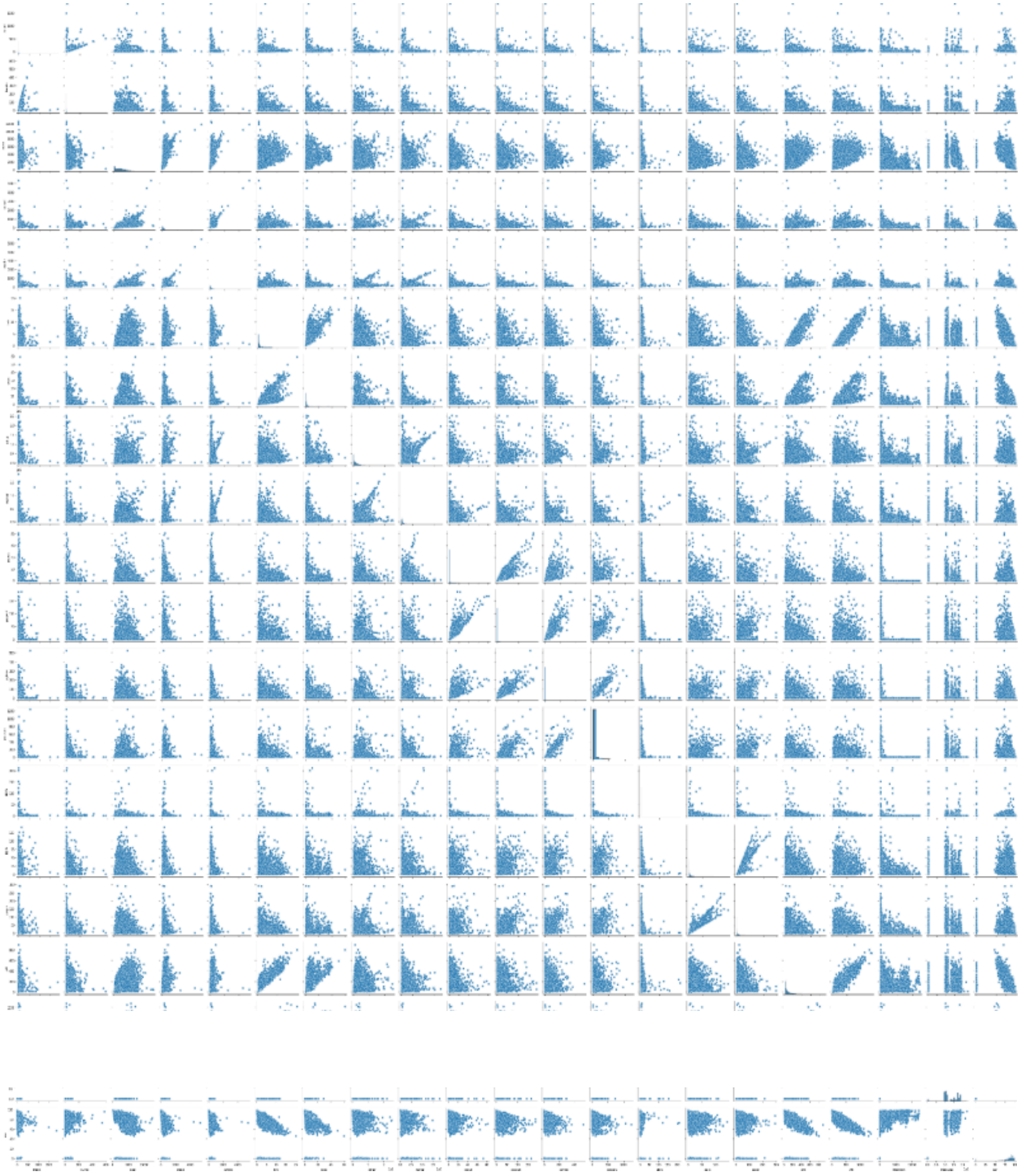
Countplot of “runqsz” field



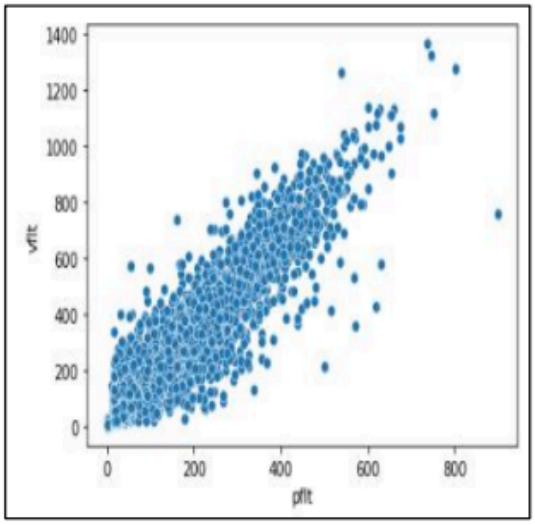
BIVARIATE ANALYSIS

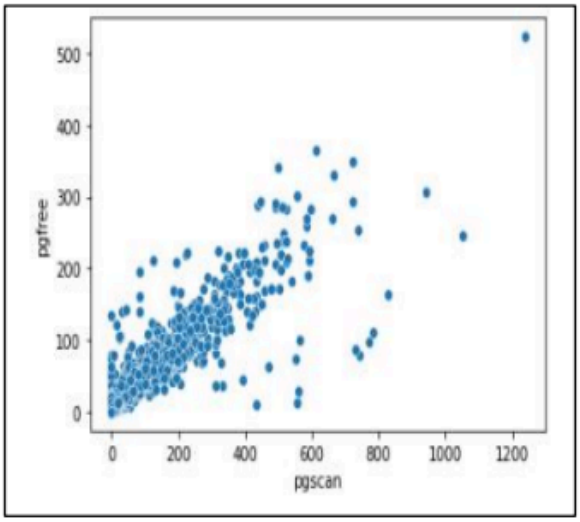
Bivariate analysis using Pairplot

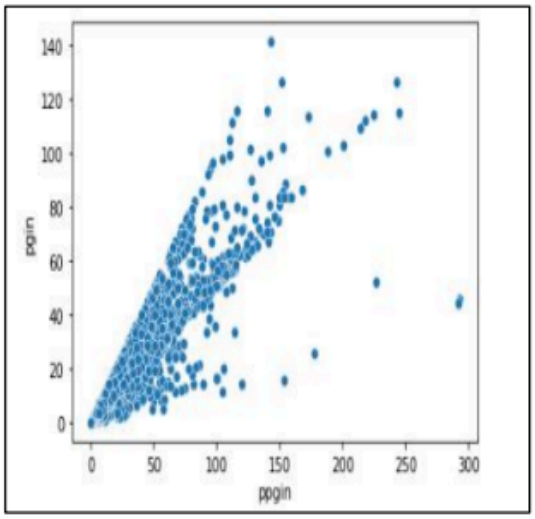


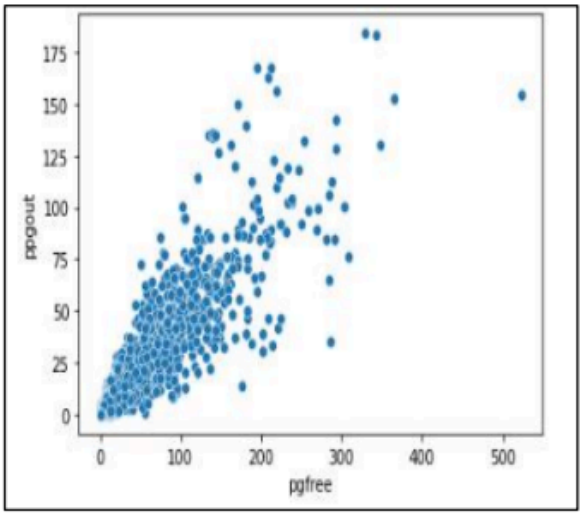


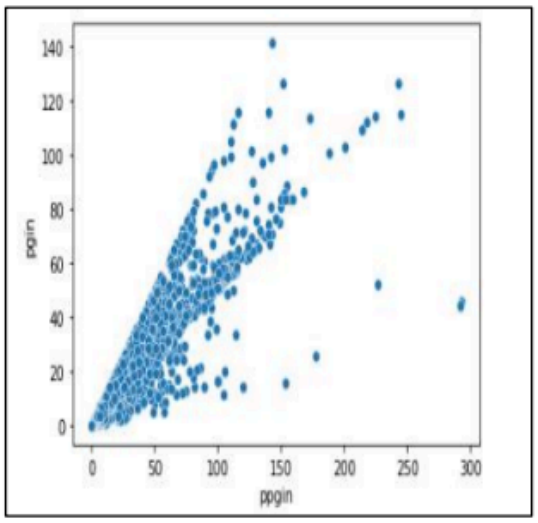
STRONG CORRELATIONS BETWEEN SOME FEATURES:

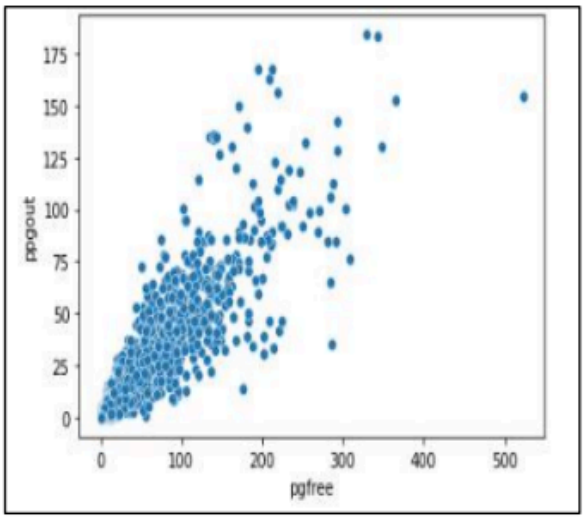

vfilt
pfilt
0
200
400
600
800
1000
1200
1400

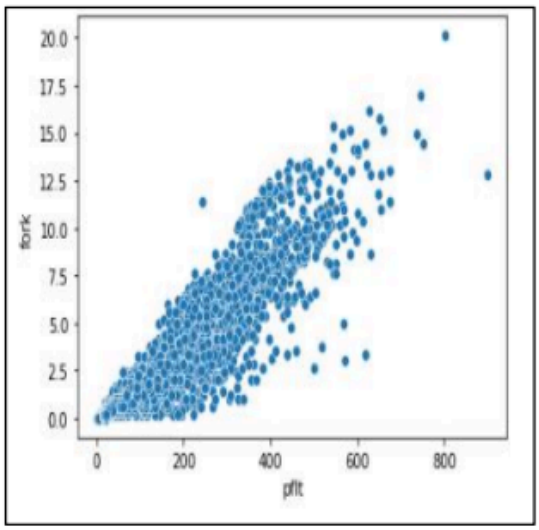

pgfree
pgscan
0
100
200
300
400
500
600
800
1000
1200

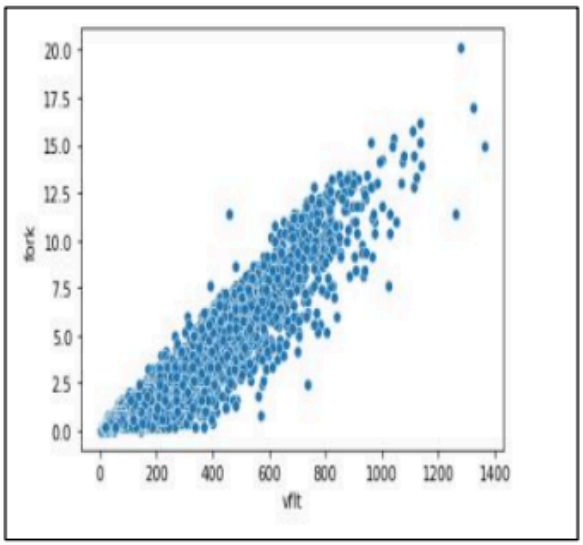

pgin
ppgin
0
20
40
60
80
100
120
140
50
150
200
250
300

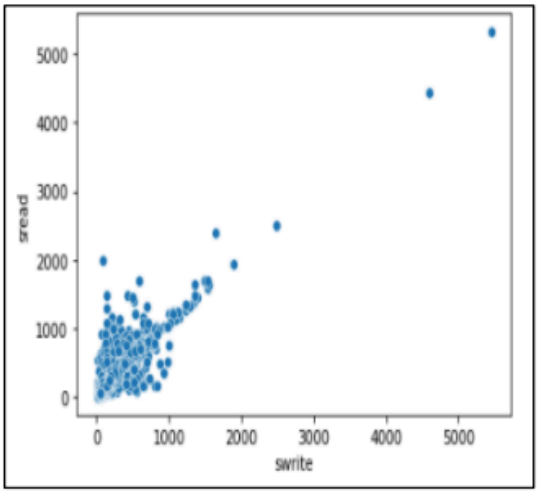

ppgout
pgfree
0
25
50
75
100
125
150
175
200
300
400
500

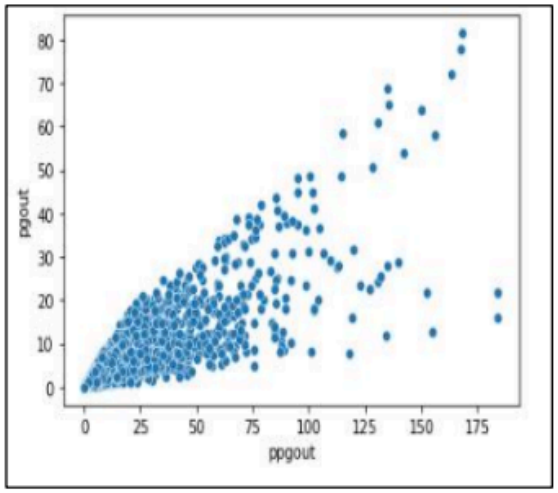

pgin
ppgin


ppgout
pgfree


fork
pflt


fork
vflt


sread
swrite


pgout
ppgout

Strongest positive correlationship is between systems attributes pflt and vflt, which is around 0.94.

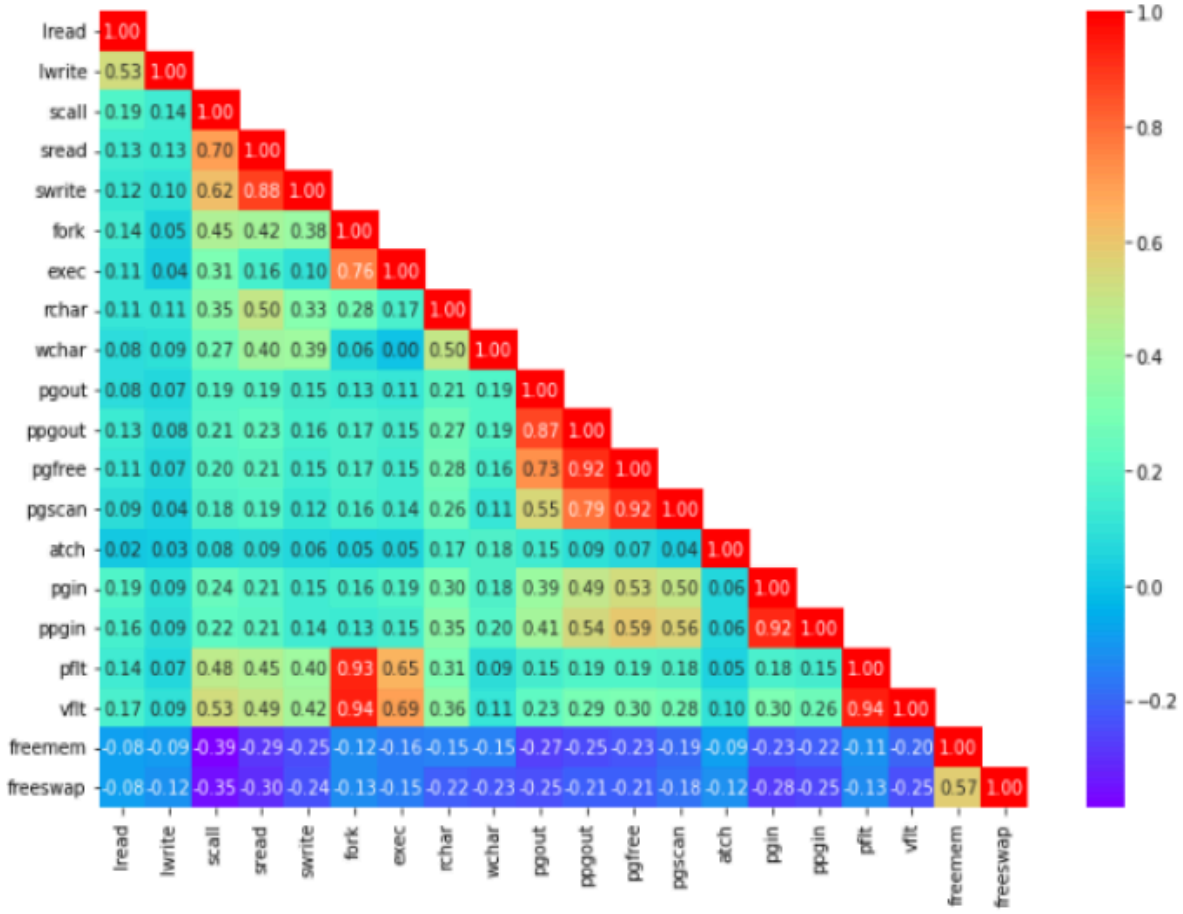
Followed by pflt - fork and vflt – fork having 0.93.

Pgscan – pgfree = 0.92

Swrite – sread =0.88,

ppgout – pgout =0.87

Heatmap



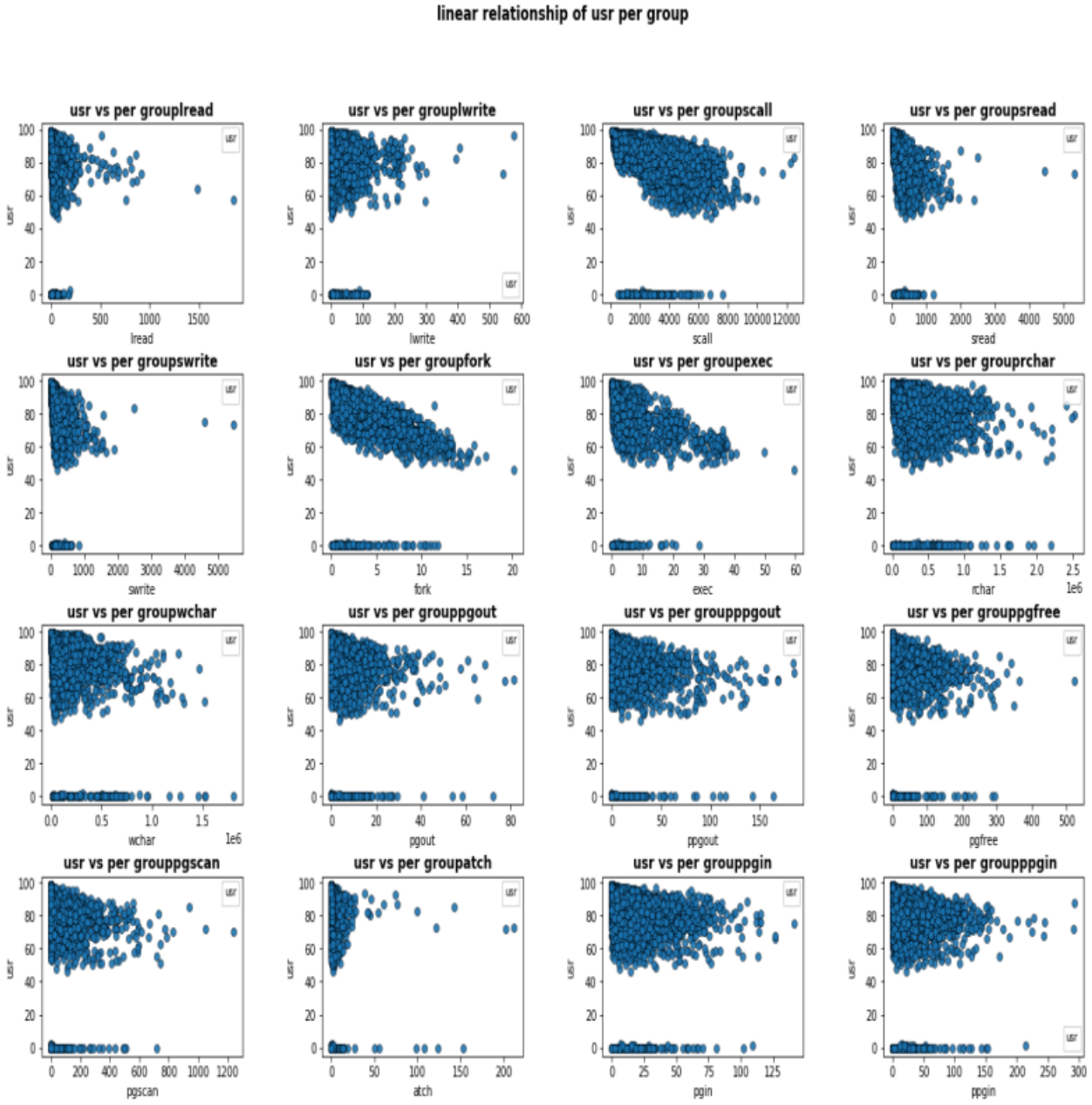
Pearson Correlation and Linear Regression

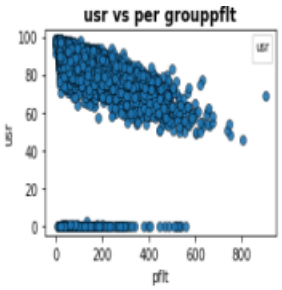
Linear relationship between target variable “usr” with the other system features

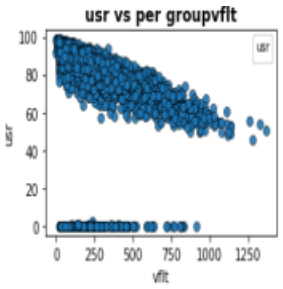
```
linear relationship between usr and lread :  -0.1414
linear relationship between usr and lwrite :  -0.1112
linear relationship between usr and scall :  -0.3232
linear relationship between usr and sread:  -0.3322
linear relationship between usr and swrite :  -0.2723
linear relationship between usr and fork :  -0.3633
linear relationship between usr and exec :  -0.2885
linear relationship between usr and rchar :  -0.3263
linear relationship between usr and wchar :  -0.289
linear relationship between usr and pgout :  -0.2219
linear relationship between usr and ppgout :  -0.2123
linear relationship between usr and pgfree :  -0.2163
linear relationship between usr and pgscan :  -0.1815
linear relationship between usr and atch :  -0.1251
linear relationship between usr and pgin :  -0.2417
linear relationship between usr and ppgin :  -0.2337
linear relationship between usr and pflt :  -0.3725
linear relationship between usr and vflt :  -0.4207
linear relationship between usr and freemem :  0.2703
linear relationship between usr and freeswap :  0.6785
```

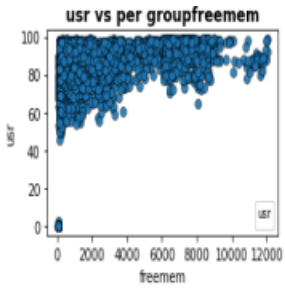
From the figure given above its clear that there is a positive relationship between “usr” and freeswap followed by freemen.

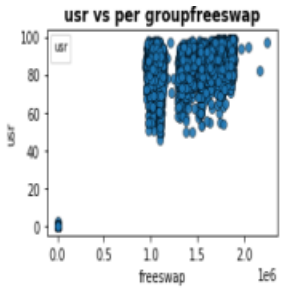
Scatterplot is plotted showing relationship between usr and other variables.











INSIGHTS:

1. Data consists of both categorical and numerical variables. 2. There are total 8192 rows and 22 columns in the dataset. Out of 22 columns only 1 column is of object data type, 8 columns are of integer type and remaining 13 are of float data type.

3. “usr” is the target variable and all other are predictor variables.

4. Upon performing univariate analysis, we find lot of outliers that needs to be treated.

5. Bivariate and multivariate analysis suggests that there is a strong positive correlation between the target variable ‘usr’ and the independent variables freemem and freeswap

==========

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.

Missing values in the data with percentage

lread	0
lwrite	0
scall	0
sread	0
swrite	0
fork	0
exec	0
rchar	104
wchar	15
pgout	0
ppgout	0
pgfree	0
pgscan	0
atch	0
pgin	0
ppgin	0
pflt	0
vflt	0
runqsz	0
freemem	0
freeswap	0
usr	0

lread	0.000000
lwrite	0.000000
scall	0.000000
sread	0.000000
swrite	0.000000
fork	0.000000
exec	0.000000
rchar	1.269531
wchar	0.183105
pgout	0.000000
ppgout	0.000000
pgfree	0.000000
pgscan	0.000000
atch	0.000000
pgin	0.000000
ppgin	0.000000
pflt	0.000000
vflt	0.000000
runqsz	0.000000
freemem	0.000000
freeswap	0.000000
usr	0.000000

Missing values in the data(after null value treatment)

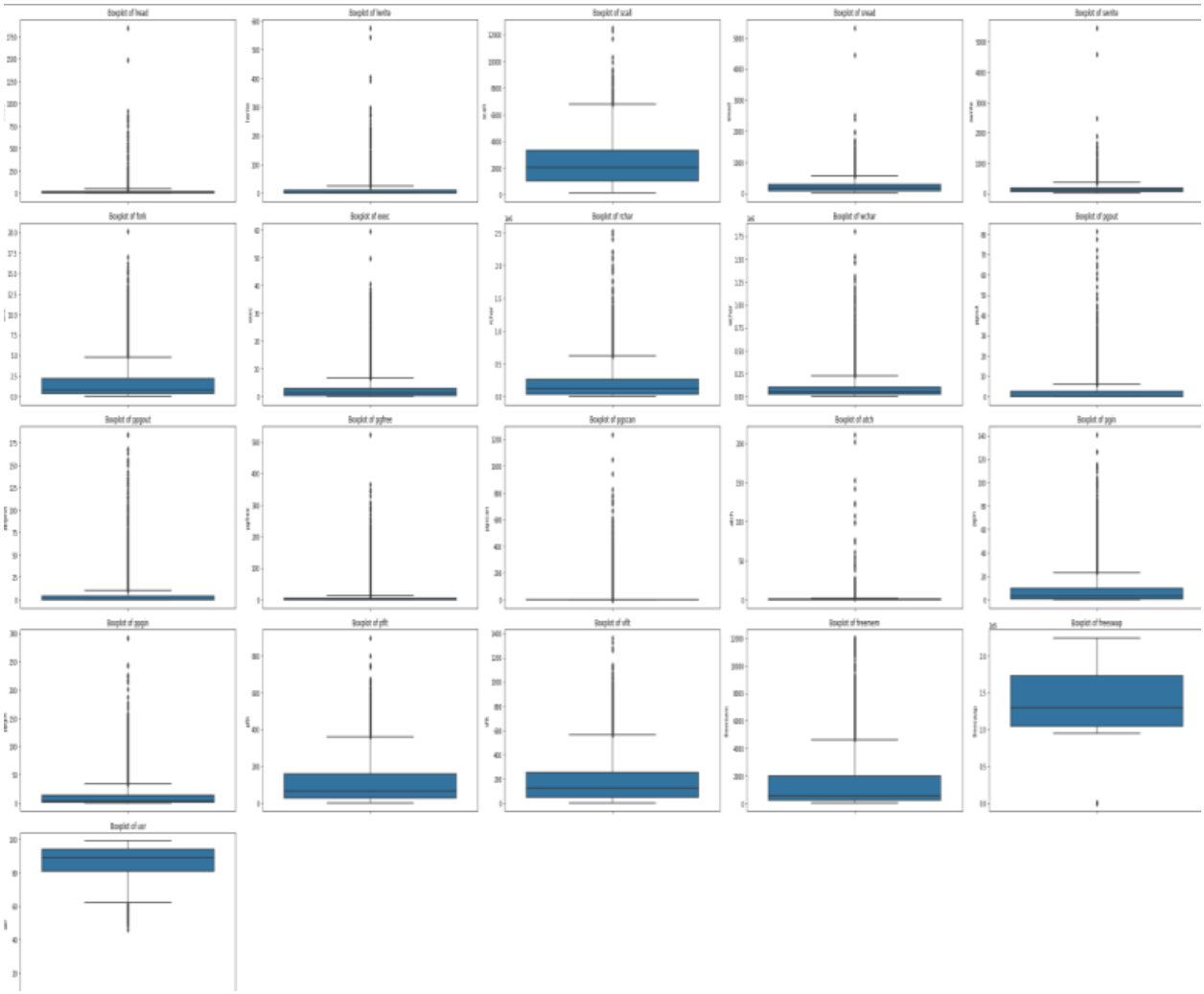
```
lread           0
lwrite          0
scall           0
sread           0
swrite          0
fork            0
exec            0
rchar           0
wchar           0
pgout           0
ppgout          0
pgfree          0
pgscan          0
atch            0
pgin            0
ppgin           0
pflt            0
vflt            0
runqsz          0
freemem         0
freeswap        0
usr             0
```

Zero values in the data

```
Count of zeros in Column  lread :  675
Count of zeros in Column  lwrite :  2684
Count of zeros in Column  scall :  0
Count of zeros in Column  sread :  0
Count of zeros in Column  swrite :  0
Count of zeros in Column  fork :  21
Count of zeros in Column  exec :  21
Count of zeros in Column  rchar :  0
Count of zeros in Column  wchar :  0
Count of zeros in Column  pgout :  4878
Count of zeros in Column  ppgout :  4878
Count of zeros in Column  pgfree :  4869
Count of zeros in Column  pgscan :  6448
Count of zeros in Column  atch :  4575
Count of zeros in Column  pgin :  1220
Count of zeros in Column  ppgin :  1220
Count of zeros in Column  pflt :  3
Count of zeros in Column  vflt :  0
Count of zeros in Column  runqsz :  0
Count of zeros in Column  freemem :  0
Count of zeros in Column  freeswap :  0
Count of zeros in Column  usr :  0
```

There are no duplicate rows in the data

Outliers in the data



DATA INSIGHTS:

1. Data has null (missing) values in two fields, namely ‘rchar’, ‘wchar’.
2. Missing values got treated by imputing median values.
3. Outliers are present in almost all numeric features.
4. Records with zero values were not removed, as it might not have an impact on model
5. There are no duplicates records in the given data set

Train Test Split

Train test dataset split in the ratio 70:30

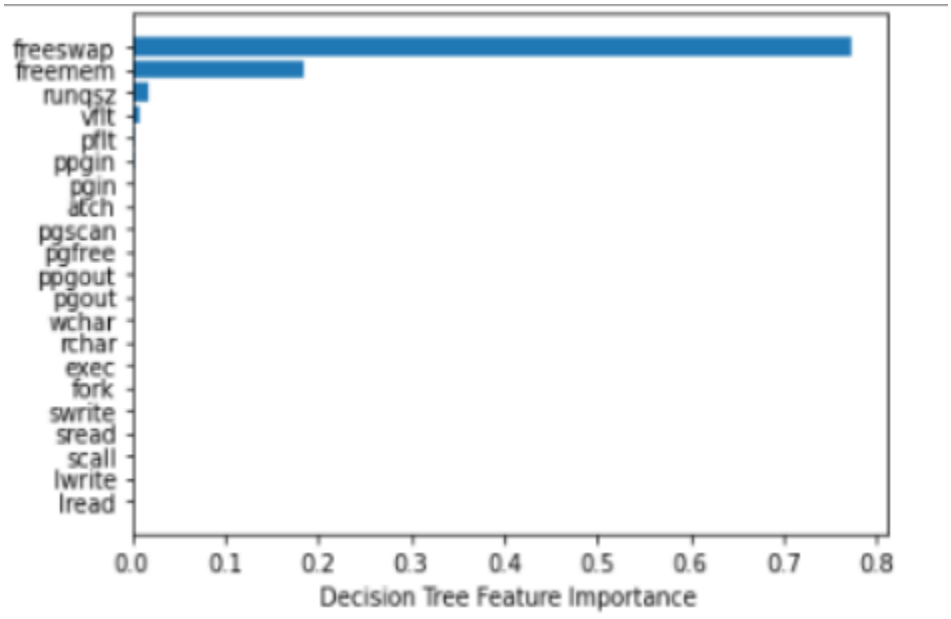
```
(5734, 22)
(5734, 1)
(2458, 22)
(2458, 1)
```

Train and Test RMSE and Train and Test Scores.

	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	10.813214	11.594014	0.64284	0.631217
Decision Tree Regressor	0.000000	3.519098	1.00000	0.966024

The Decision Tree, is over-fitting because of the difference in values between train and test RMSE.,

Feature importance shown by decision tree



Coefficients of independent variables after train test split

```
The coefficient for lread is -0.019887328309043474
The coefficient for lwrite is 0.00479371418053189
The coefficient for scall is 0.0010081513380151988
The coefficient for sread is -0.00047060977558386377
The coefficient for swrite is -0.002040967864439776
The coefficient for fork is -1.7222497016995837
The coefficient for exec is -0.08962698046525462
The coefficient for rchar is -4.061601767872802e-06
The coefficient for wchar is -1.1639177541802703e-05
The coefficient for pgout is -0.17390380486767593
The coefficient for ppgout is 0.09893911749504092
The coefficient for pgfree is -0.07033919333981715
The coefficient for pgscan is 0.008623331383079188
The coefficient for atch is -0.07856282405694104
The coefficient for pgin is 0.09126693493613791
The coefficient for ppgin is -0.05938114917450077
The coefficient for pflt is -0.04150868976786212
The coefficient for vflt is 0.022283906660960827
The coefficient for runqsz is -7.7907927112044595
The coefficient for freemem is -0.0016171671325389517
The coefficient for freeswap is 3.2192359542447795e-05
```

RMSE is an absolute measure of fit

R2 Score = Explained Variation/Total Variation

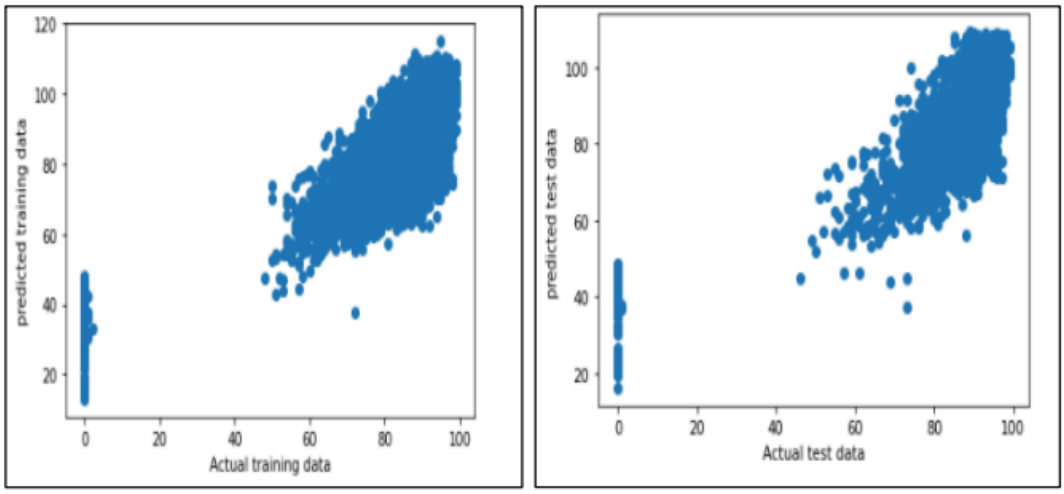
The higher the R2 Score value, the better the model fits the data.

Usually, its value ranges from 0 to 1

```
The coefficient of determination R^2 of the prediction on Test set 0.6312171006119699

The Root Mean Square Error (RMSE) of the model is for testing set is 11.59401399232658

The coefficient of determination R^2 of the prediction on Train set 0.6428396267060905

The Root Mean Square Error (RMSE) of the model is for training set is 10.813213974052196
```

```
The intercept for our model is 52.428777248033086
```

Scatterplots for the predictions



DATA INSIGHTS:

1. Using the p>|t| result, we can say that the variables like lwrite, sread, swrite, pgscan are statistically insignificant variables as ther p-value is greater than 0.05.
2. Omnibus test checks the normality of the residuals once the model is deployed. Here prob(omnibus) is 0 indicating that there is 0% chance that the residuals are normally distributed. For a model to be robust the residual distribution is also required to be normal ideally apart from checking rsquared and other parameters.
3. This indicates our model is not robust and not fit.
4. Also there are very strong multicollinearity present in the dataset

INFERENCES

The final Linear Regression equation is: (52.43) * const + (-0.02) * lread + (0.0) * lwrite + (0.0) * scall + (-0.0) * sread + (- 0.0) * swrite + (-1.72) * fork + (-0.09) * exec + (-0.0) * rchar + (-0.0) * wchar + (- 0.17) * pgout + (0.1) * ppgout + (-0.07) * pgfree + (0.01) * pgscan + (-0.08) * atch + (0.09) * pgin + (-0.06) * ppgin + (-0.04) * pflt + (0.02) * vflt + (-7.79) * runqsz + (-0.0) * freemem + (0.0) * freeswap

- When ppgout increases by 1 unit, usr increases by 0.1 units keeping all other predictors constant.
- When pgscan increases by 1 unit, usr increases by 0.01 units keeping all other predictors constant
- When pgin increases by 1 unit, usr increases by 0.09 units keeping all other predictors constant
- When vflt increases by 1 unit, usr increases by 0.02 units keeping all other predictors constant

++

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

GOAL:

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

DATASET:

The dataset contains 10 features including demographic and socio-economic characteristics of 1473 married women in Indonesia, which is obtained from National Indonesia Contraceptive Prevalence Survey. The dataset has 9 descriptive features and one target variable.

TARGET FEATURE:

The response variable is “Contraceptive method used” having two classes.

```
Contraceptive_method_used
Yes    774
No     614
Name: Contraceptive_method_used,
```

DESCRIPTIVE FEATURES

Following are the variables in the dataset.
Wife’s age : numerical
Wife’s education : categorical(1=uneducated, 2, 3, 4=tertiary)
Husband’s education : categorical(1=uneducated, 2, 3, 4=tertiary)
Number of children ever born : numerical
Wife’s religion : binary(Non-Scientology, Scientology)
Wife’s now working : binary(Yes, No)
Husband’s occupation : categorical(1, 2, 3, 4(random))
Standard of living index : categorical(1=very low, 2, 3, 4=high)
Media exposure : binary(Good, Not good)

Printing the first five rows using head () to check that whether the features and descriptions outlined in the document are aligning with the dataset

	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
0	24.0	Primary	Secondary	3.0	Scientology	No	2	High	Exposed	No
1	45.0	Uneducated	Secondary	10.0	Scientology	No	3	Very High	Exposed	No
2	43.0	Primary	Secondary	7.0	Scientology	No	3	Very High	Exposed	No
3	42.0	Secondary	Primary	9.0	Scientology	No	3	High	Exposed	No
4	36.0	Secondary	Secondary	8.0	Scientology	No	3	Low	Exposed	No

Printing the last 5 rows

	Wife_age	Wife_education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_metho
1466	42.0	Primary	Tertiary	3.0	Scientology	No	2	Very High	Exposed	
1468	33.0	Tertiary	Tertiary	3.0	Scientology	Yes	2	Very High	Exposed	
1470	39.0	Secondary	Secondary	3.0	Scientology	Yes	1	Very High	Exposed	
1471	33.0	Secondary	Secondary	3.0	Scientology	Yes	2	Low	Exposed	
1472	17.0	Secondary	Secondary	1.0	Scientology	No	2	Very High	Exposed	

Checking the shape of dataset:

```
(1473, 10)
```

```
The number of rows are  1473
The number of columns are 10
```

The dataset has 1473 rows and 10 columns.

Checking the datatypes:

There are three different datatypes. dtypes: float64(2), int64(1), object(7)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1473 entries, 0 to 1472
Data columns (total 10 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Wife_age                  1402 non-null   float64
 1   Wife_ education           1473 non-null   object
 2   Husband_education         1473 non-null   object
 3   No_of_children_born       1452 non-null   float64
 4   Wife_religion             1473 non-null   object
 5   Wife_Working              1473 non-null   object
 6   Husband_Occupation        1473 non-null   int64
 7   Standard_of_living_index  1473 non-null   object
 8   Media_exposure            1473 non-null   object
 9   Contraceptive_method_used 1473 non-null   object
dtypes: float64(2), int64(1), object(7)
memory usage: 115.2+ KB
```

Describing the summary of the dataset:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Wife_age	1402.0	NaN	NaN	NaN	32.606277	8.274927	16.0	26.0	32.0	39.0	49.0
Wife_ education	1473	4	Tertiary	577	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_education	1473	4	Tertiary	899	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_of_children_born	1452.0	NaN	NaN	NaN	3.254132	2.365212	0.0	1.0	3.0	4.0	16.0
Wife_religion	1473	2	Scientology	1253	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Wife_Working	1473	2	No	1104	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_Occupation	1473.0	NaN	NaN	NaN	2.137814	0.864857	1.0	1.0	2.0	3.0	4.0
Standard_of_living_index	1473	4	Very High	684	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Media_exposure	1473	2	Exposed	1364	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Contraceptive_method_used	1473	2	Yes	844	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Checking the missing values:

```
Wife_age                    71
Wife_ education              0
Husband_education            0
No_of_children_born         21
Wife_religion                0
Wife_Working                 0
Husband_Occupation           0
Standard_of_living_index     0
Media_exposure               0
Contraceptive_method_used    0
dtype: int64
```

```
Wife_age                    4.820095
Wife_ education             0.000000
Husband_education           0.000000
No_of_children_born         1.425662
Wife_religion               0.000000
Wife_Working                0.000000
Husband_Occupation          0.000000
Standard_of_living_index    0.000000
Media_exposure              0.000000
Contraceptive_method_used   0.000000
dtype: float64
```

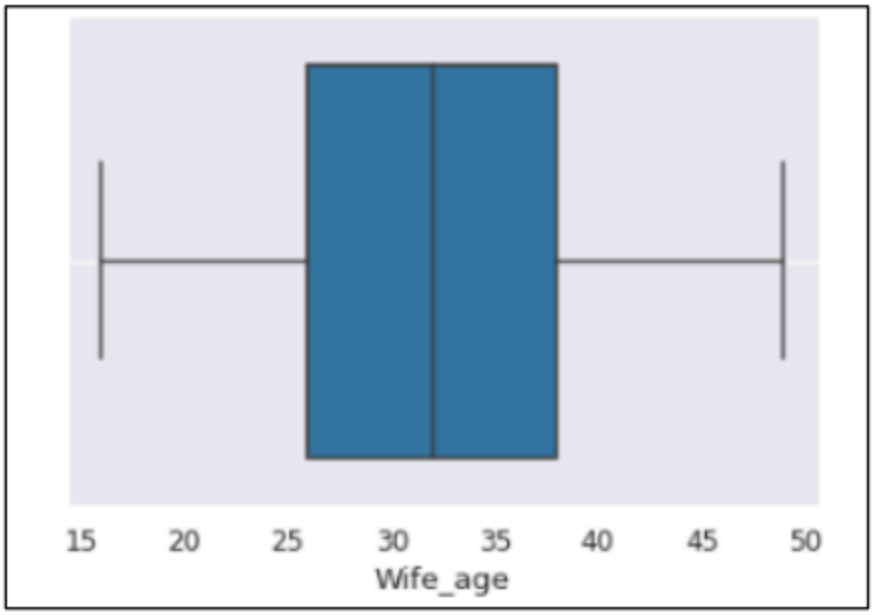
There are missing values present in the “wife’s age” and “No. of children born” variables of the dataset.Approx. 5% of missing values are there in wife’s age field and 1% in the latter

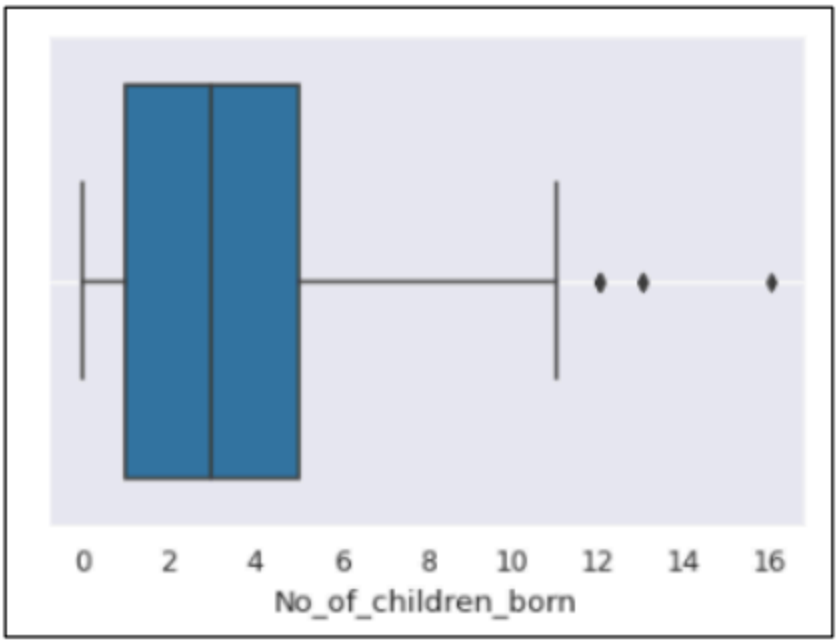
Checking for the duplicate values:

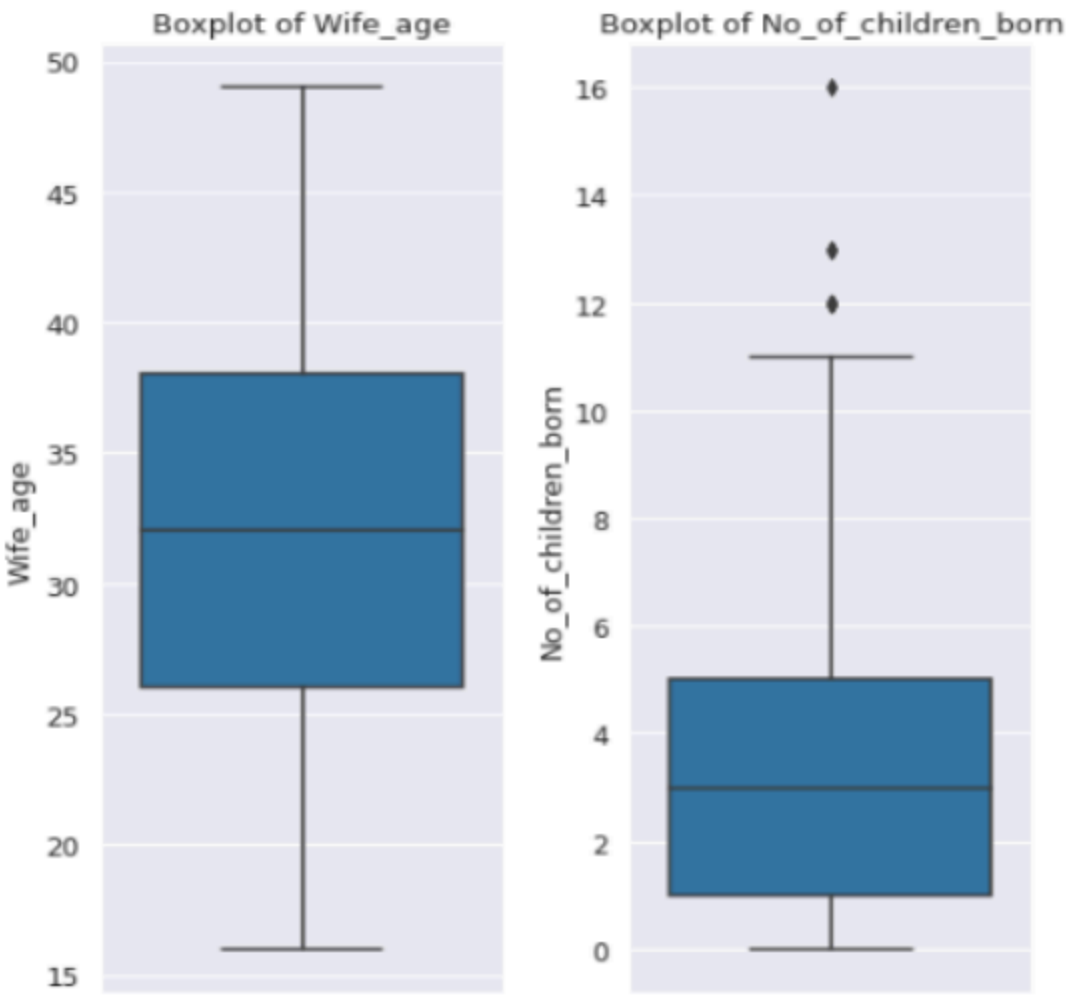
There are 85 rows in the data containing duplicate values

Checking for outliers:

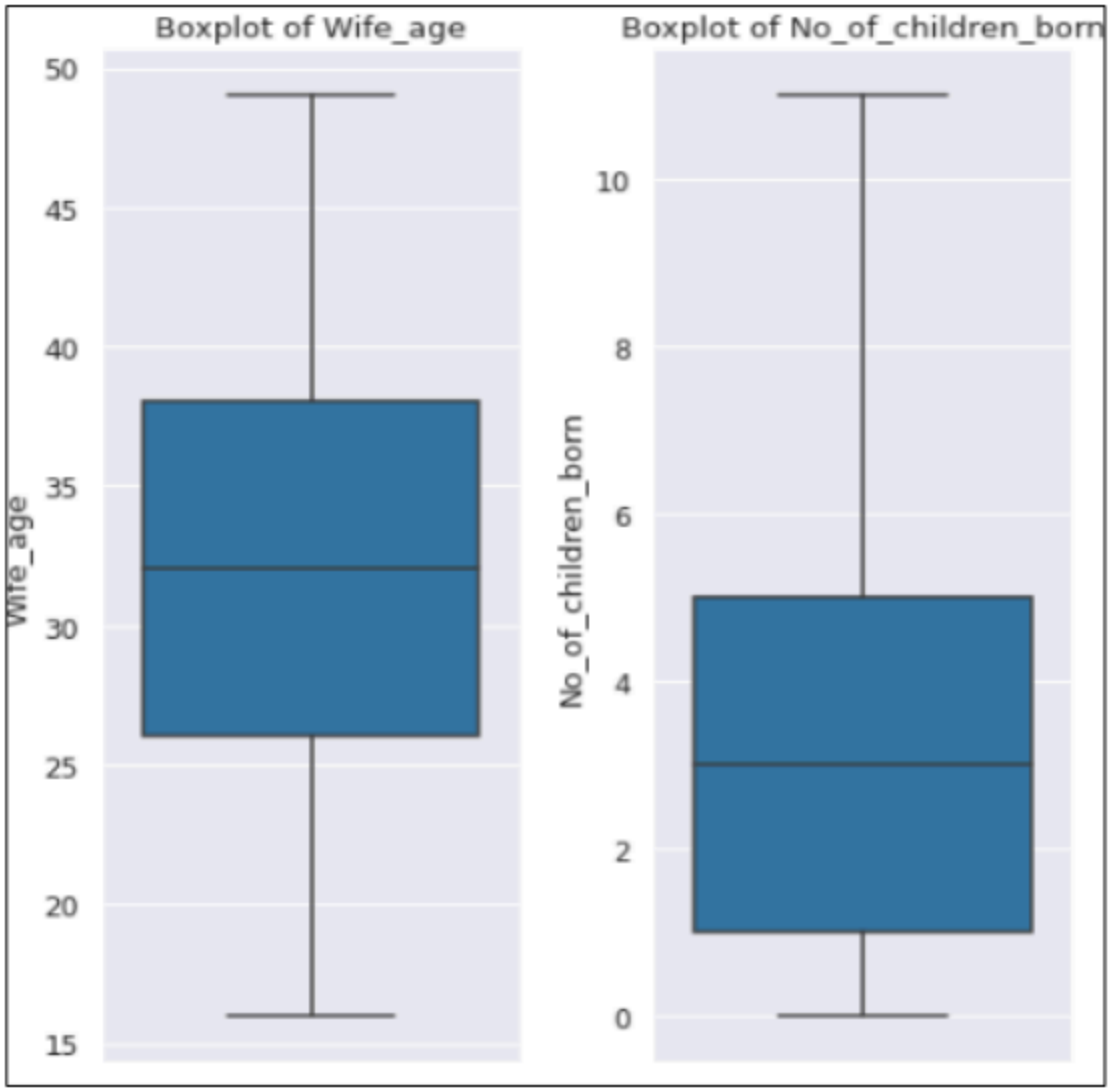
Box Plots with outliers


15
20
25
30
35
40
45
50
Wife_age

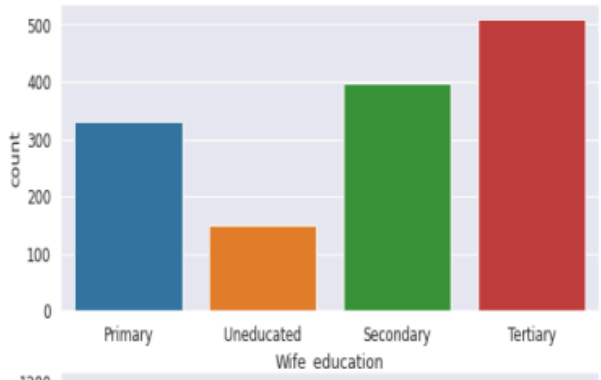

0
2
4
6
8
10
12
14
16
No_of_children_born

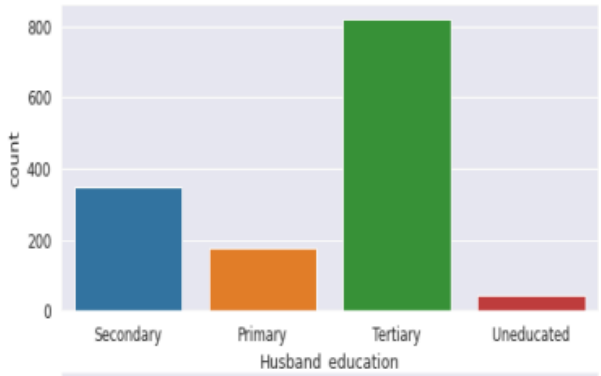


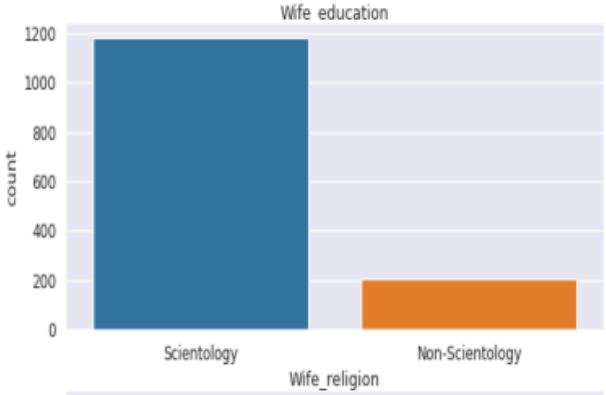
Boxplots (after outliers treatment)

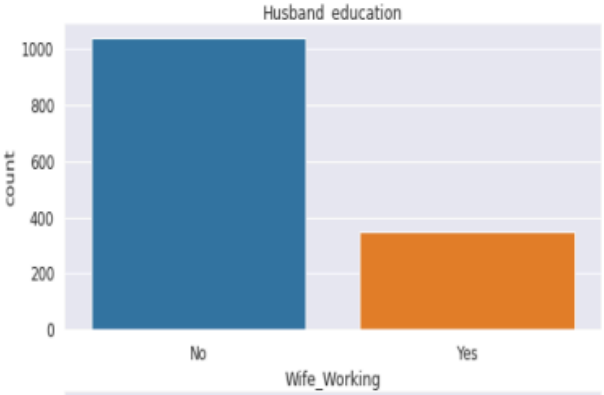


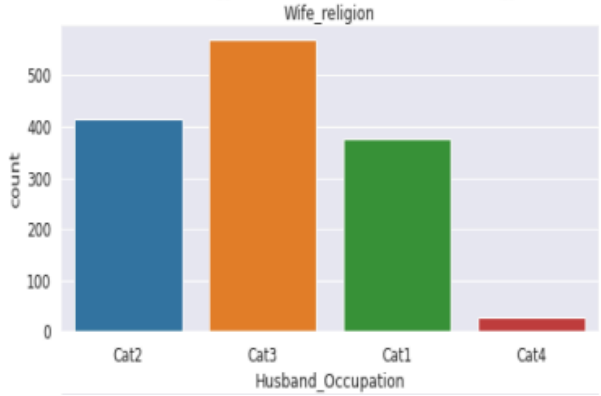
UNIVARIATE ANALYSIS:

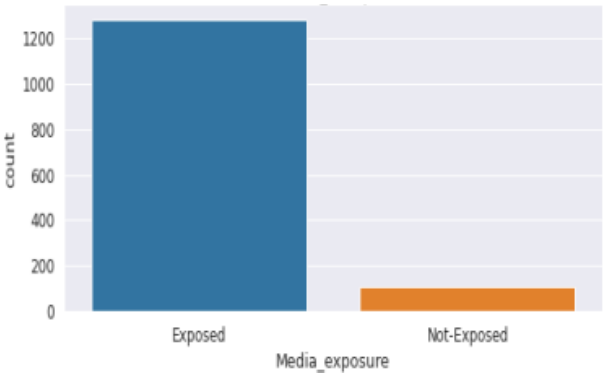


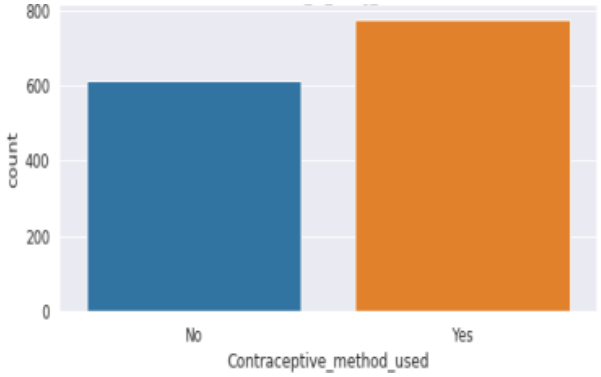








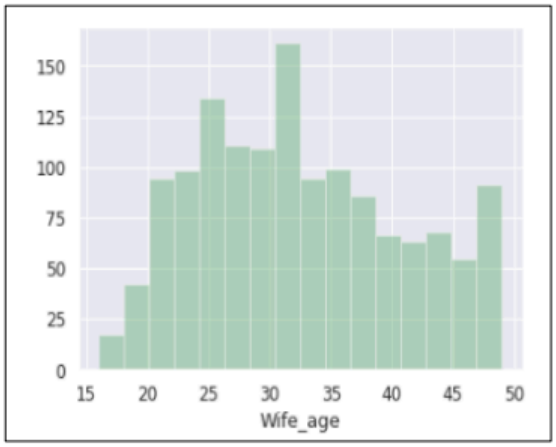


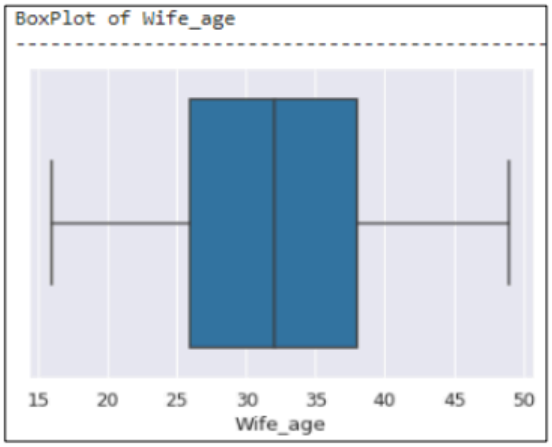


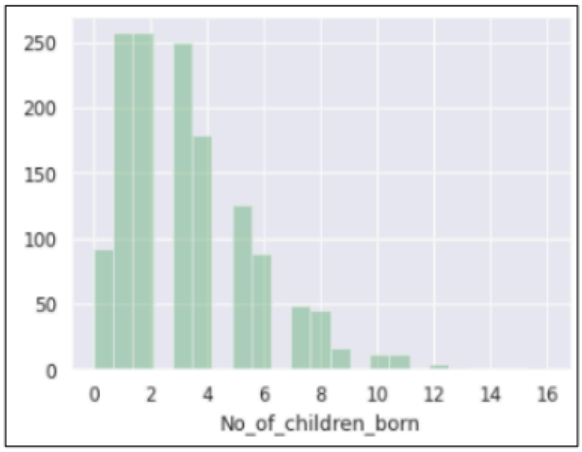
- Both wife and husband tertiary education level is more.

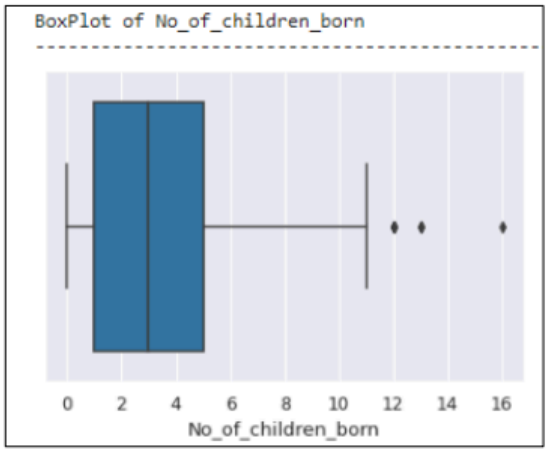
- Scientologist wives are more.
- Majority wives are not working.
- Most of the males or husband has category 3 occupation.
- Media exposure is more

Distplots and Boxplots of numerical fields

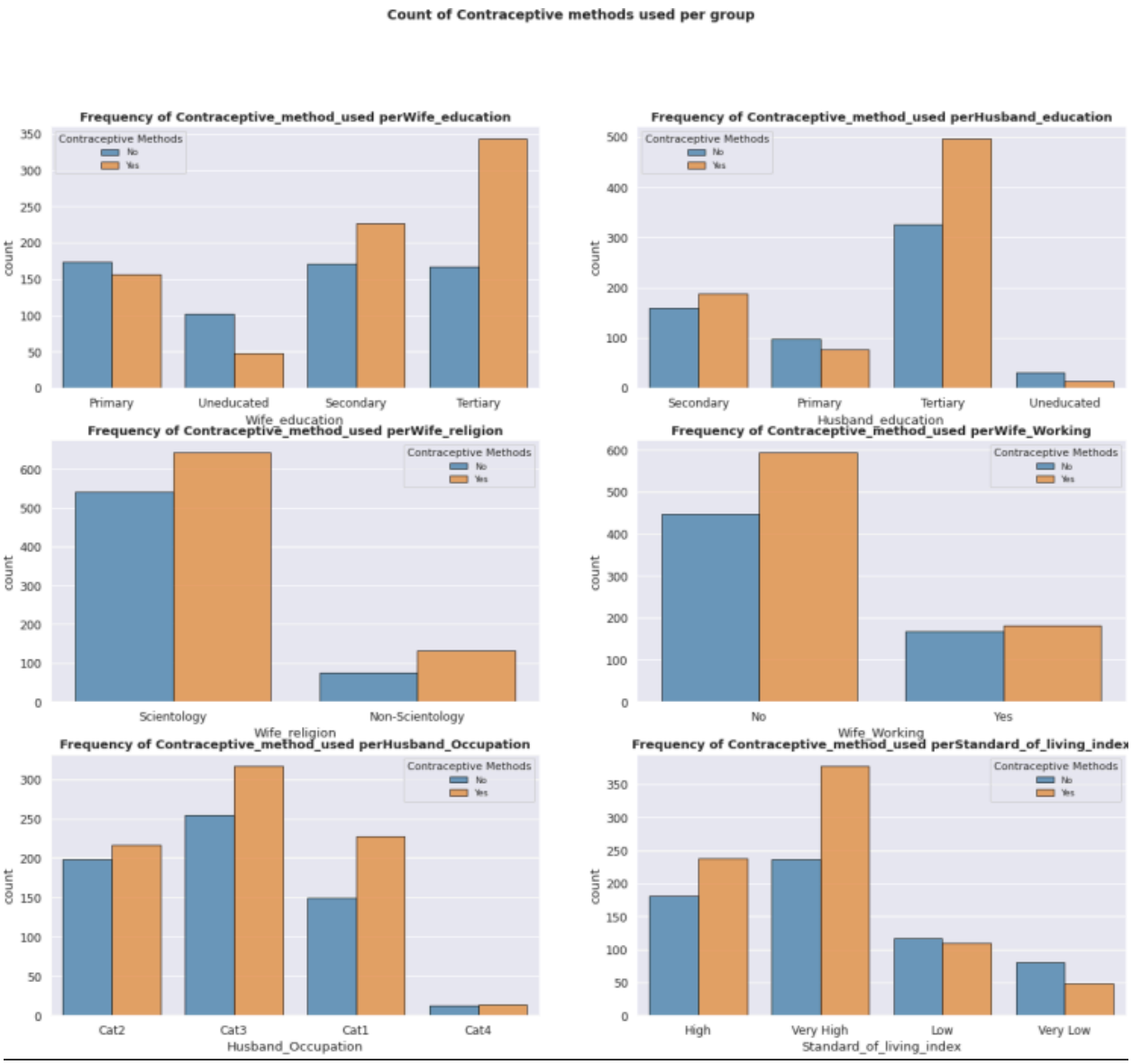








BIVARIATE ANALYSIS

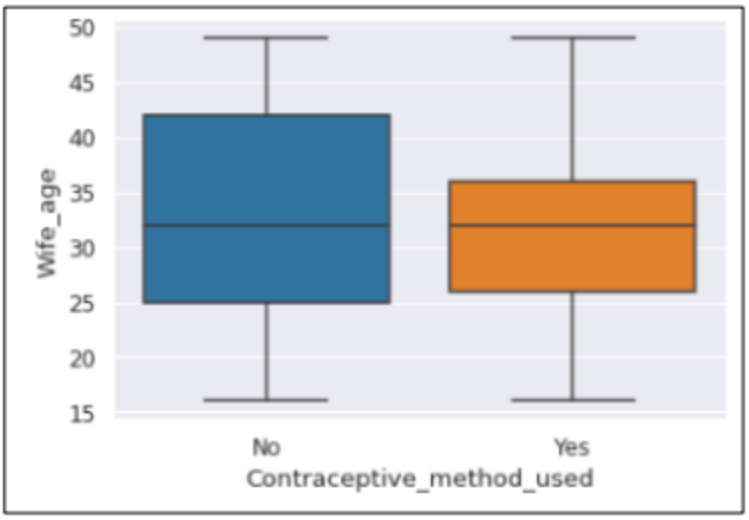


Contraceptive is used by majority of the highly educated wives.

People having very high standard of living index uses more.

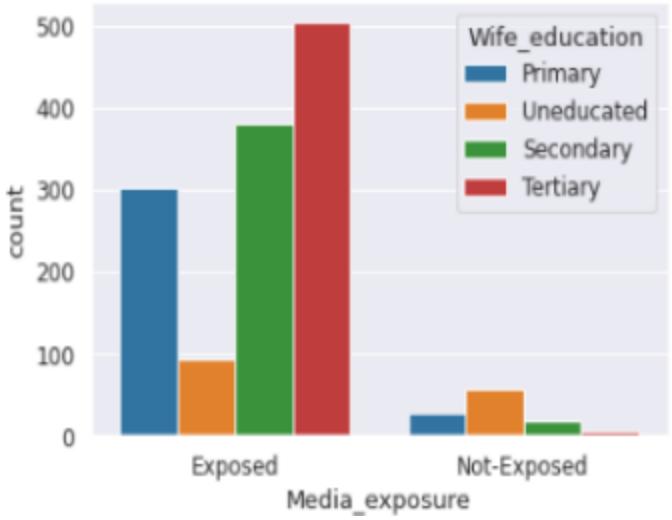
No. of children born vs contraceptive used

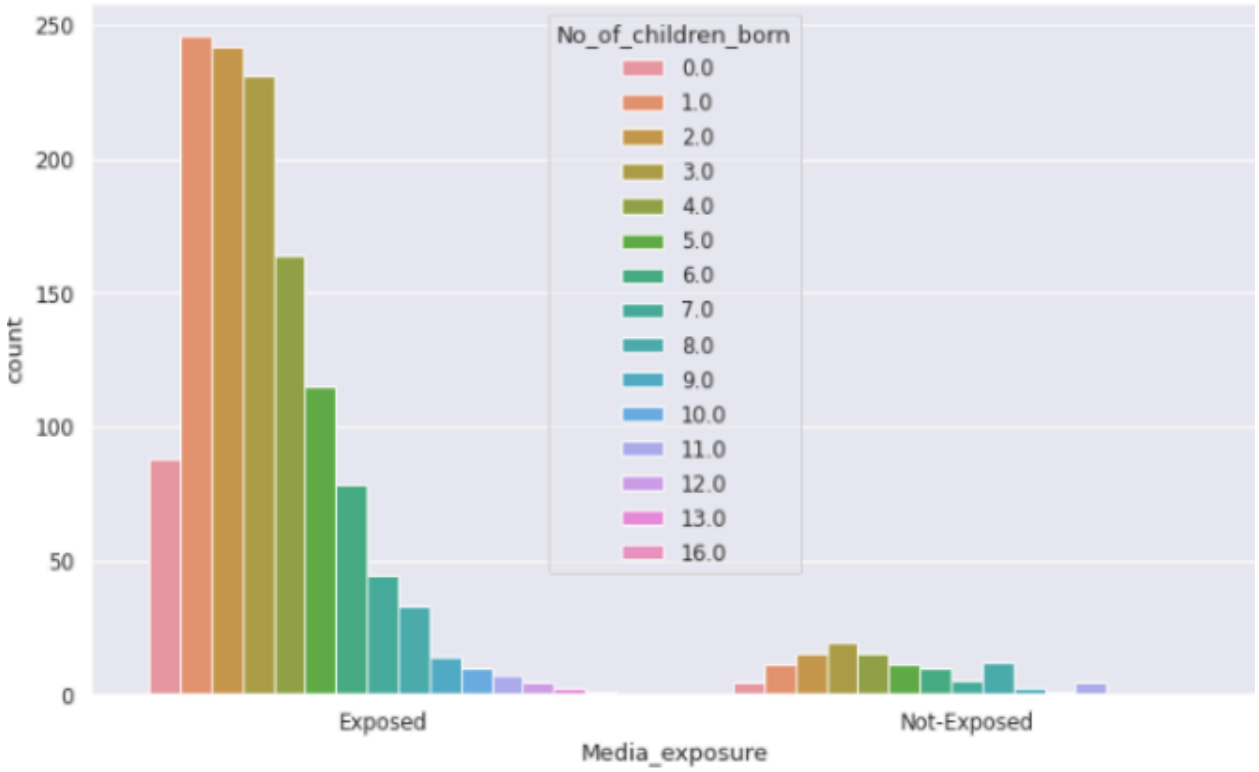
Wife's age vs contraceptive used

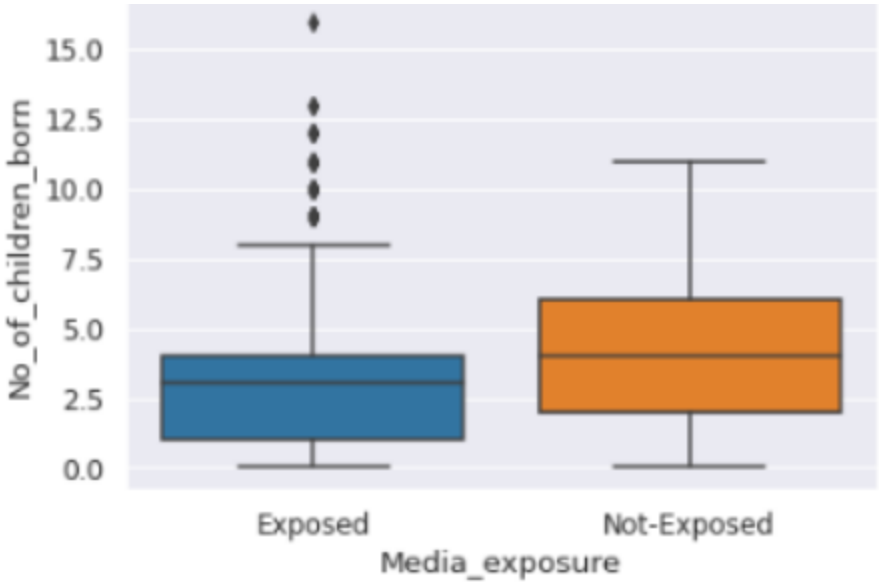


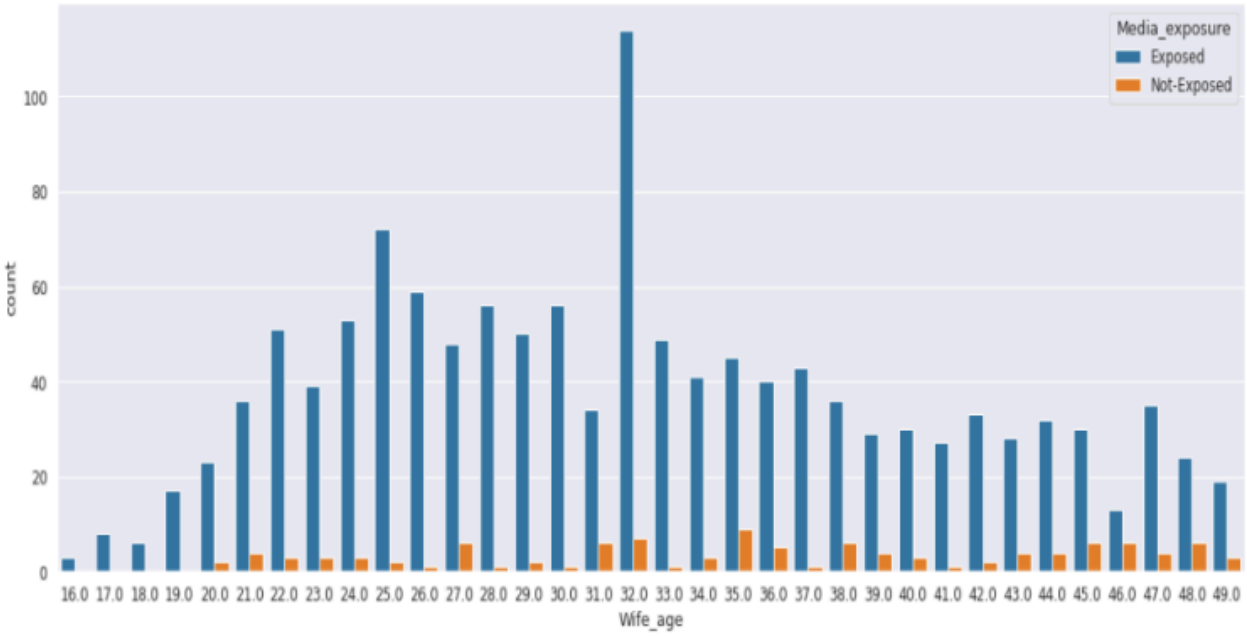
Wife’s Working and Wife’s education vs standard of living index

Wife’s education, No. of children born, and Wife’s age, vs Media Exposure

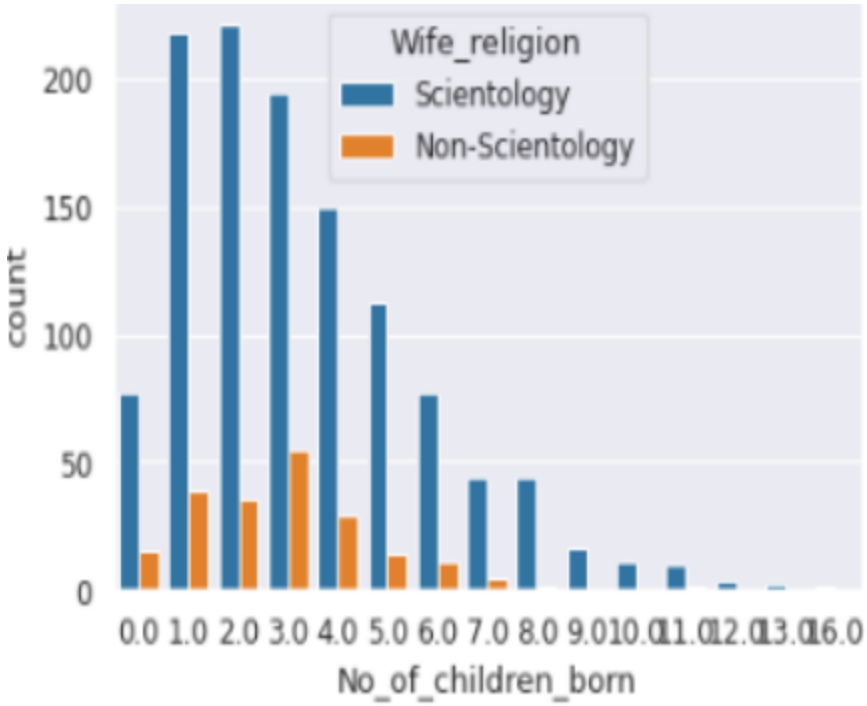

Wife_education
Primary
Uneducated
Secondary
Tertiary
count
500
400
300
200
100
0
Exposed
Not-Exposed
Media_exposure


No_of_children_born
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
16.0
count
250
200
150
100
50
0
Exposed
Not-Exposed
Media_exposure



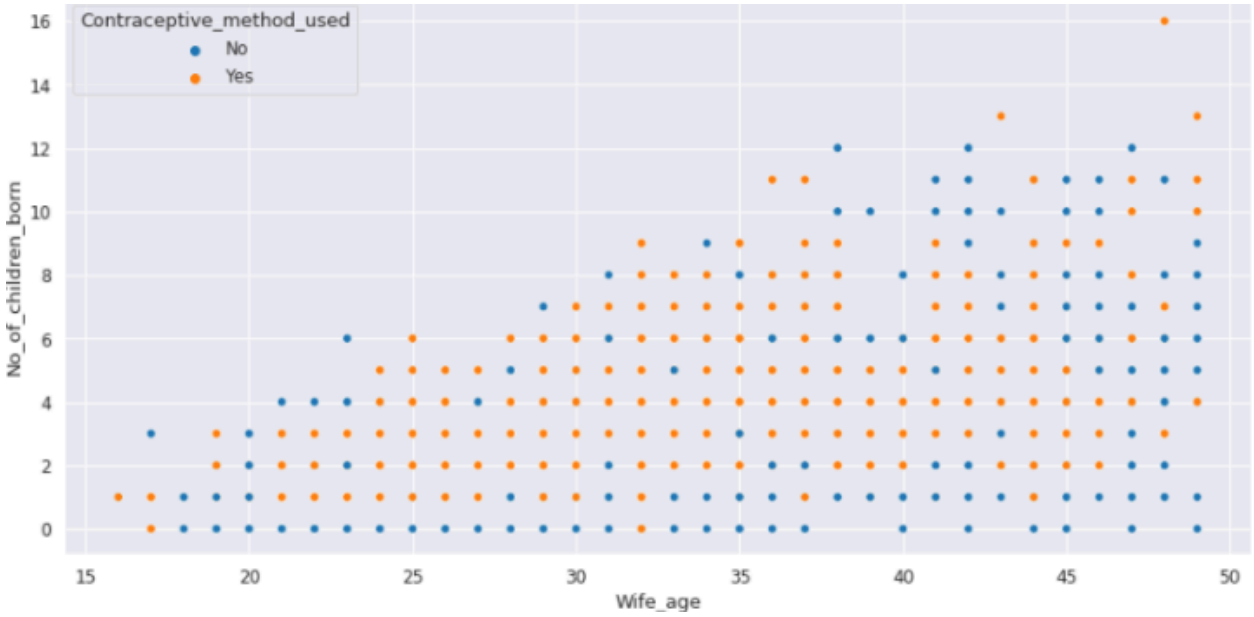


Wife’s Religion vs No. of children born

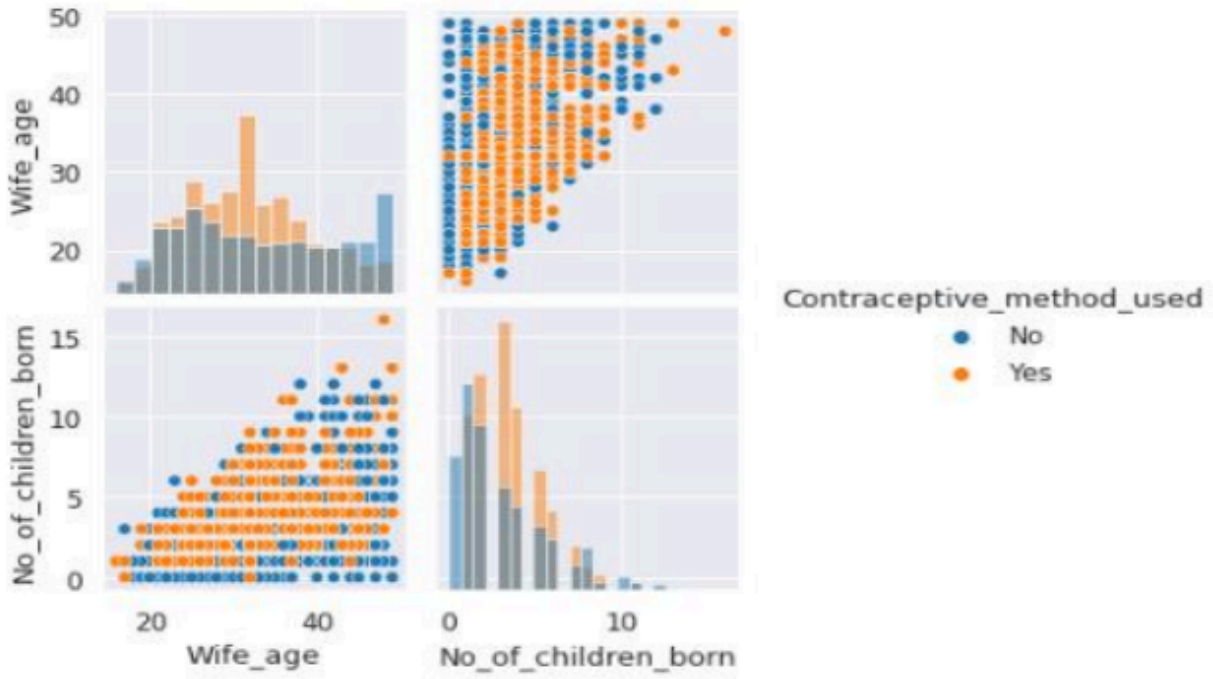


MULTIVARIATE ANALYSIS

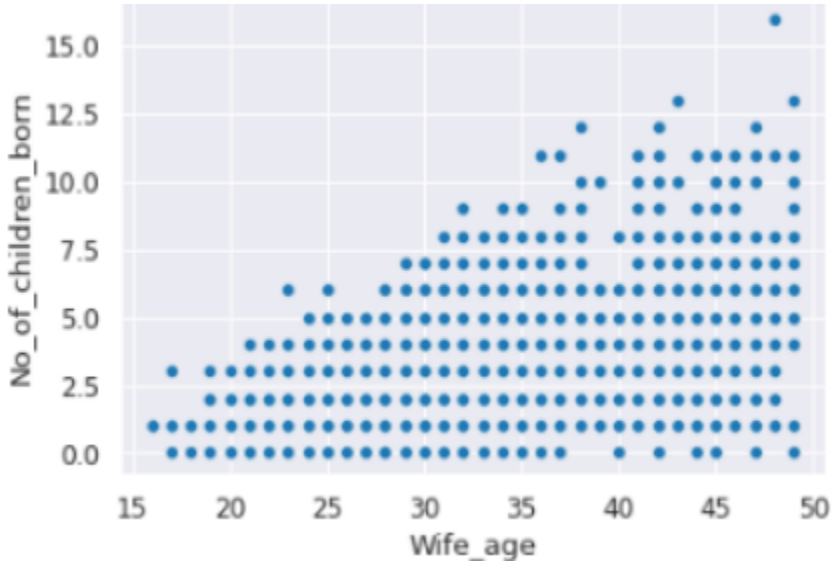
Scatterplot between Wife's age and No. of children by Contraceptive Methods



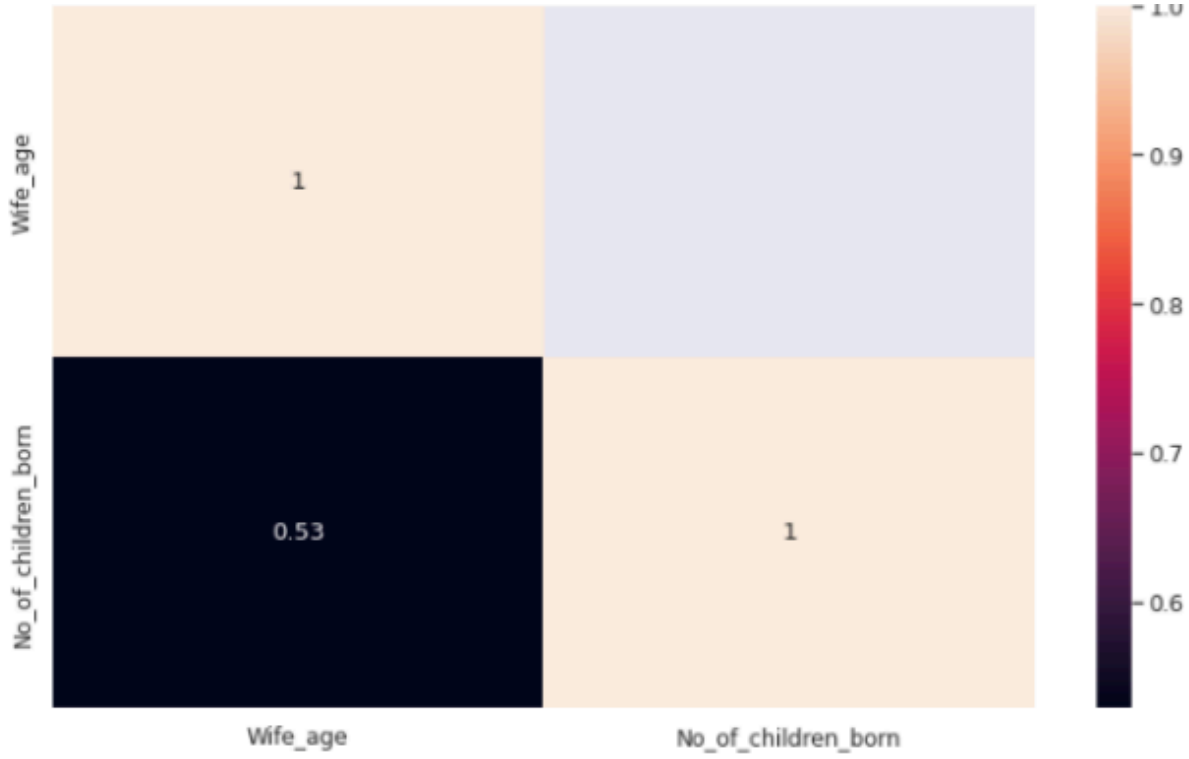
Pairplot between Wife's age and No. of children by Contraceptive Methods



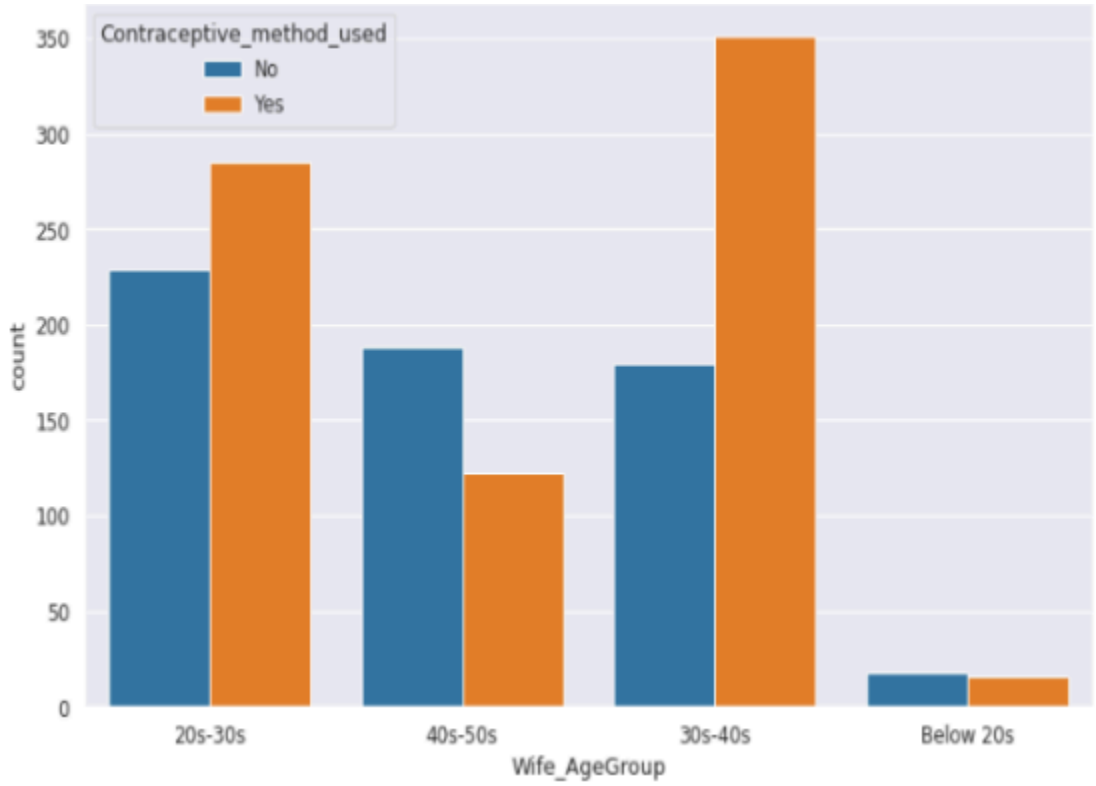
CORRELATION PLOT:

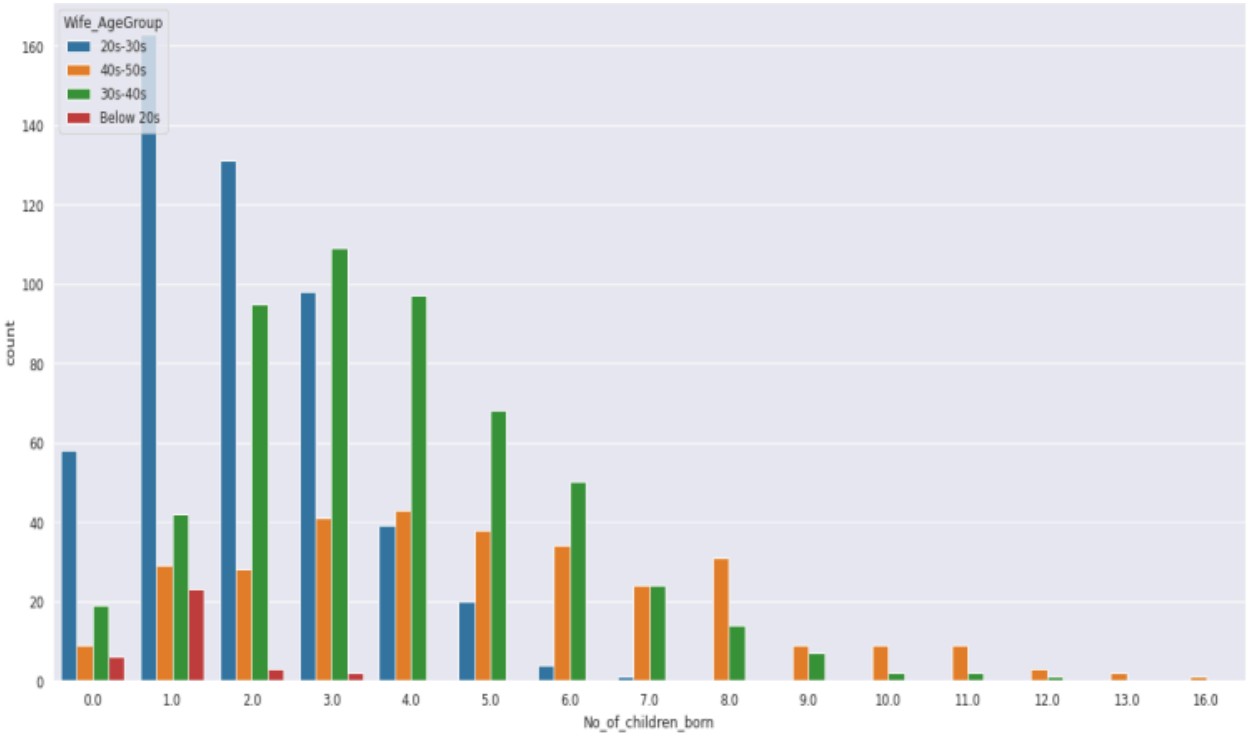


Heatmap

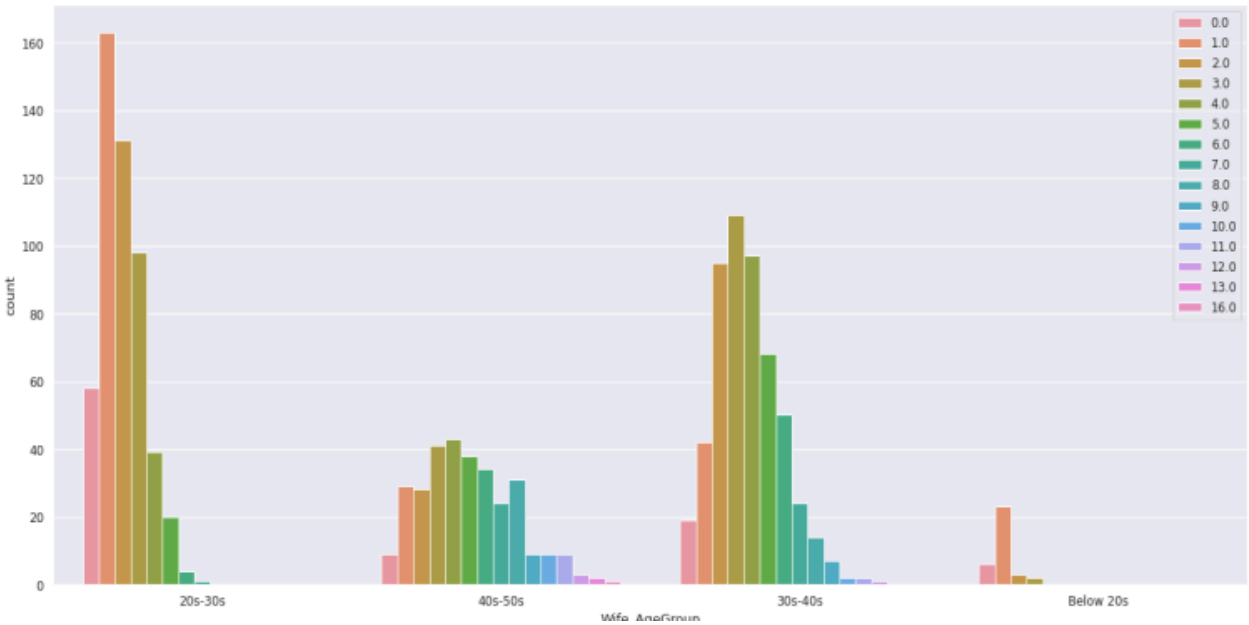


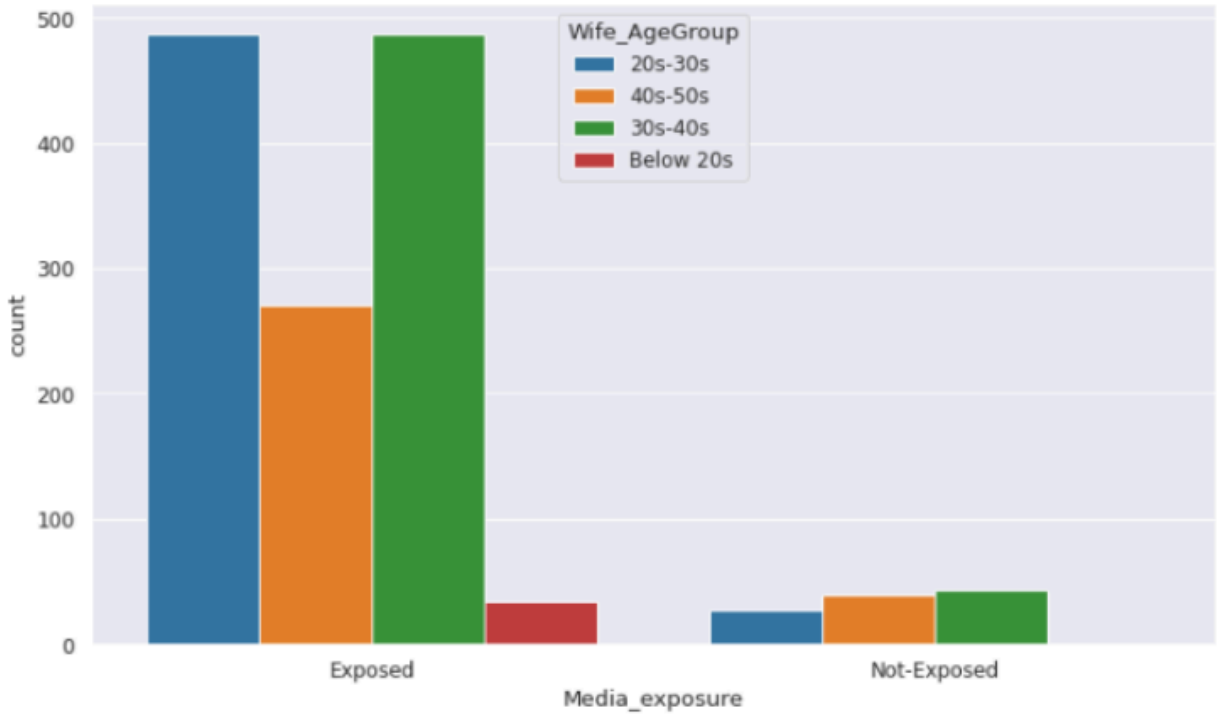
Wife Age group vs contraceptive method used And no. of children born





Wife Age group vs no. of children born and media exposure





INFERENCES:

1. Data consists of both categorical and numerical values.
2. There are total of 1473 rows and 10 columns in the dataset. Out of 22, 7 columns are of object type, 1 columns of integer type and remaining 2 are of float type data.
3. 'contraceptive used' is the target variable and all other are predictor variables.
4. Looking into the fields in the univariate analysis, we see outliers is present only in the field number of children.
5. Looking 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.
6. Bivariate and multivariate analysis indicates that there is strong positive correlation between the field's wife age and no_of_children_born
7. We also notice that there are 80 duplicates records in the given data set and has been removed.
8. Null values identified has been imputed with median

2.2Apply Logistic Regression and LDA (linear discriminant analysis) and CART.

Converted the target variable “Contraceptive_method_used” into numeric by using Label Encoder function from Sklearn by defining the function label encoder.

Dataset after label encoding

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
24.0	Primary	Secondary	3.0	Scientology	No	Cat2	High	Exposed	0
45.0	Uneducated	Secondary	10.0	Scientology	No	Cat3	Very High	Exposed	0
43.0	Primary	Secondary	7.0	Scientology	No	Cat3	Very High	Exposed	0
42.0	Secondary	Primary	9.0	Scientology	No	Cat3	High	Exposed	0
36.0	Secondary	Secondary	8.0	Scientology	No	Cat3	Low	Exposed	0

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
42.0	Primary	Tertiary	3.0	Scientology	No	Cat2	Very High	Exposed	1
33.0	Tertiary	Tertiary	3.0	Scientology	Yes	Cat2	Very High	Exposed	1
39.0	Secondary	Secondary	3.0	Scientology	Yes	Cat1	Very High	Exposed	1
33.0	Secondary	Secondary	3.0	Scientology	Yes	Cat2	Low	Exposed	1
17.0	Secondary	Secondary	1.0	Scientology	No	Cat2	Very High	Exposed	1

Dataset after dummy encoding

	Wife_age	No_of_children_born	Contraceptive_method_used	Wife_education_Secondary	Wife_education_Tertiary	Wife_education_Uneducated	Husband_education_Secondary
0	24.0	3.0	0	0	0	0	1
1	45.0	10.0	0	0	0	1	1
2	43.0	7.0	0	0	0	0	1
3	42.0	9.0	0	1	0	0	0
4	36.0	8.0	0	1	0	0	1

Husband_education_Secondary	Husband_education_Tertiary	Husband_education_Uneducated	Wife_religion_Scientology	Wife_Working_Yes	Husband_Occupation_Cat2	Husband_Occupation_Cat3
1	0	0	1	0	1	0
1	0	0	1	0	0	1
1	0	0	1	0	0	1
0	0	0	1	0	0	1
1	0	0	1	0	0	1

Husband_Occupation_Cat4	Standard_of_living_index_Low	Standard_of_living_index_Very High	Standard_of_living_index_Very Low	Media_exposure_Not-Exposed
0	0	0	0	0
0	0	1	0	0
0	0	1	0	0
0	0	0	0	0
0	1	0	0	0

TRAIN-TEST SPLIT

The data has been split in the 70:30 (train:test) ratio.

```
(971, 17)
```

```
(417, 17)
```

X train head

	Wife_age	No_of_children_born	Wife_education_Secondary	Wife_education_Tertiary	Wife_education_Uneducated	Husband_education_Secondary	Husband_education_Tertiary
570	36.0	3.0	0	1	0	0	1
1426	32.0	3.0	1	0	0	0	1
406	24.0	3.0	0	0	0	0	0
1083	35.0	5.0	0	1	0	0	1
1094	38.0	6.0	0	0	1	0	0

Husband_education_Uneducated	Wife_religion_Scientology	Wife_Working_Yes	Husband_Occupation_Cat2	Husband_Occupation_Cat3	Husband_Occupation_Cat4	Standard_of_living_index_Low
0	0	0	1	0	0	0
0	1	0	0	1	0	0
0	1	0	1	0	0	0
0	1	0	0	0	0	0
0	1	0	1	0	0	0

Standard_of_living_index_Very High	Standard_of_living_index_Very Low	Media_exposure_Not-Exposed
1	0	0
0	1	0
1	0	0
1	0	0
0	1	1

Formulate a logistic regression model on the train data. Fit the logistic regression model

```
LogisticRegression(max_iter=1000)
```

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.

Decision Tree classifier

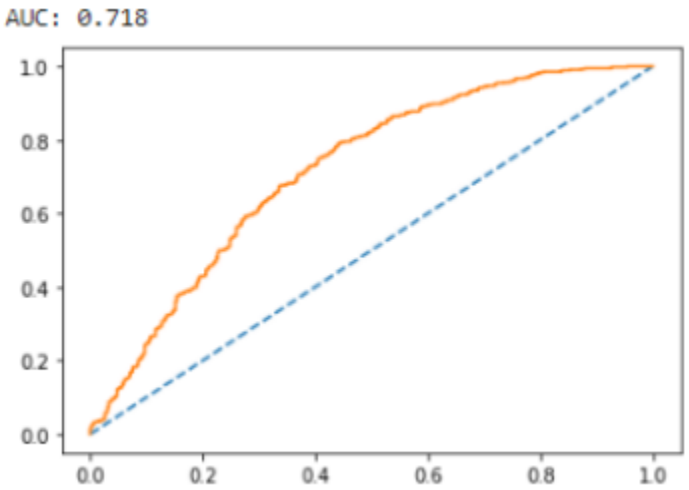
	Train Accuracy	Test Accuracy
Decision Tree Classifier	0.983522	0.585132
LDA	0.681771	0.635492
Logistic Regression	0.681771	0.633094

The Decision Tree Classifier, is under-fitting because train accuracy > test accuracy .,

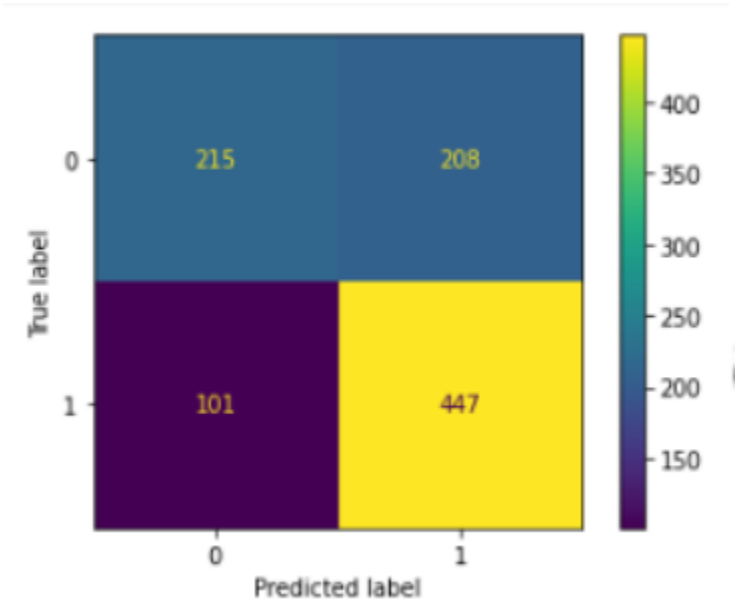
Logistic model

```
# Accuracy - Training Data
0.6817713697219362

Accuracy - Test Data
0.6306954436450839
```

AUC and ROC for the training data



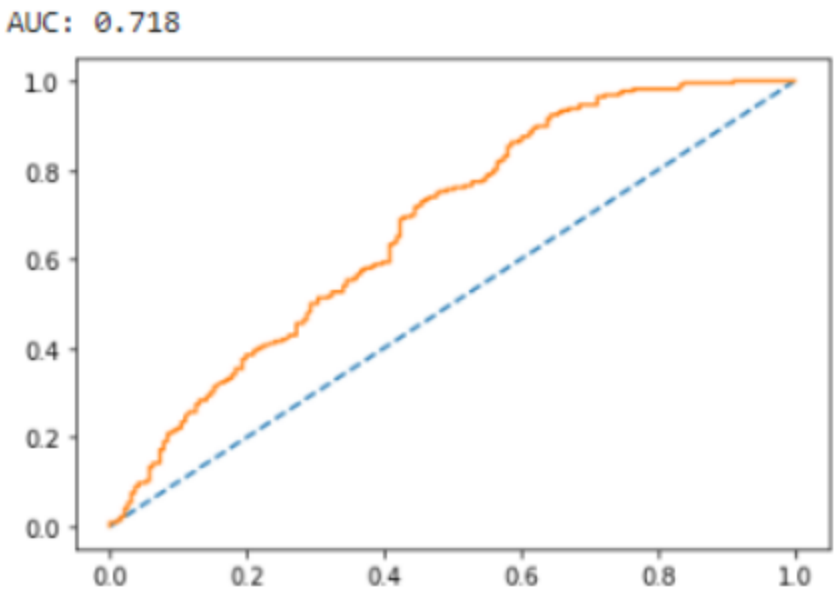
Confusion Matrix for the training data



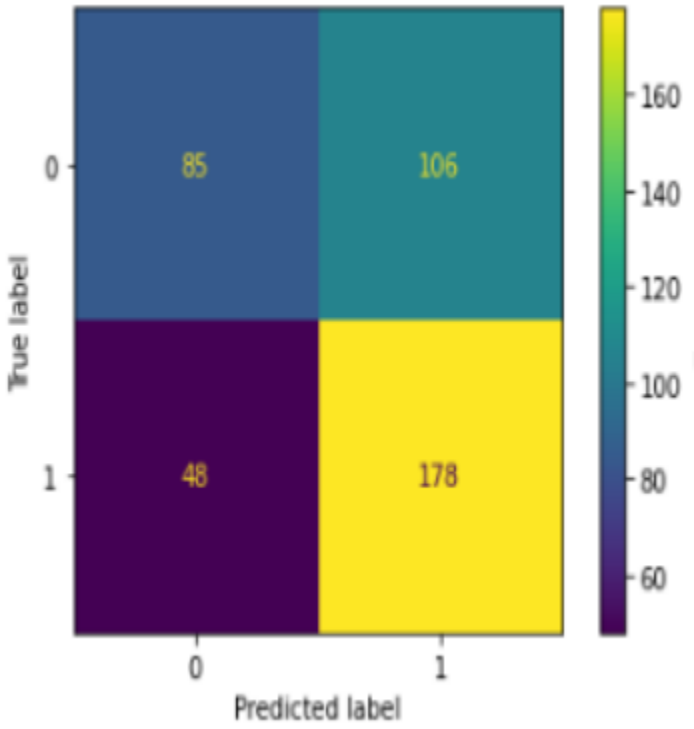
Classification report of train data

	precision	recall	f1-score	support
0	0.68	0.51	0.58	423
1	0.68	0.82	0.74	548
accuracy			0.68	971
macro avg	0.68	0.66	0.66	971
weighted avg	0.68	0.68	0.67	971

AUC and ROC for the test data



Confusion Matrix for the test data



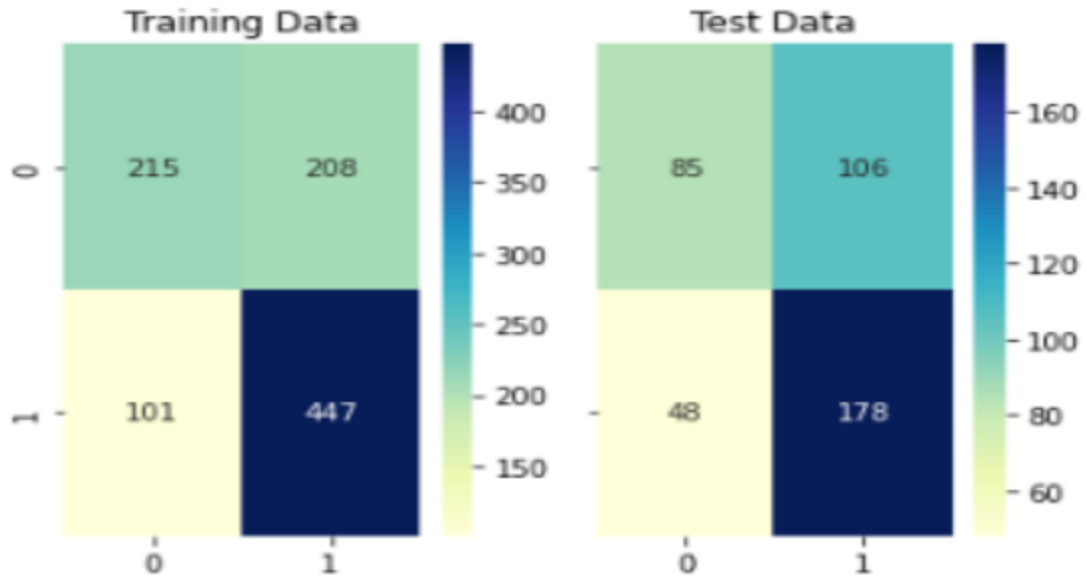
	precision	recall	f1-score	support
0	0.64	0.45	0.52	191
1	0.63	0.79	0.70	226
accuracy			0.63	417
macro avg	0.63	0.62	0.61	417
weighted avg	0.63	0.63	0.62	417

LINEAR DISCRIMINANT ANALYSIS

Generate Coefficients and intercept for the Linear Discriminant Function intercept value:

```
array([1.76720844])
```

```
array([[-0.08166222,  0.33495826,  0.43702965,  0.97255106, -0.26608413,
         0.25179775,  0.0823513 , -0.12386827, -0.52560658, -0.1365502 ,
        -0.11748001,  0.16687103,  0.83330564, -0.26115594,  0.25413658,
        -0.87524849, -0.29513378]])
```



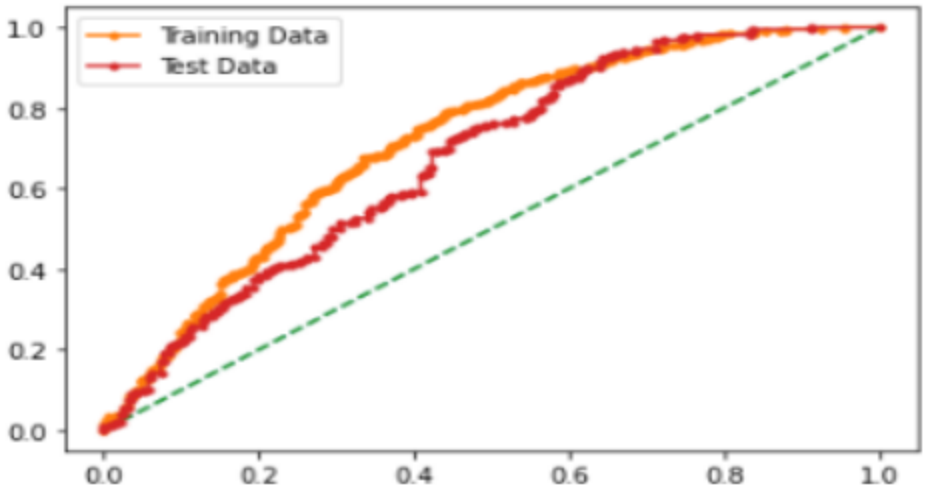
Classification Report of the training data:

	precision	recall	f1-score	support
0	0.68	0.51	0.58	423
1	0.68	0.82	0.74	548
accuracy			0.68	971
macro avg	0.68	0.66	0.66	971
weighted avg	0.68	0.68	0.67	971

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.65	0.45	0.53	191
1	0.63	0.80	0.70	226
accuracy			0.64	417
macro avg	0.64	0.62	0.62	417
weighted avg	0.64	0.64	0.62	417

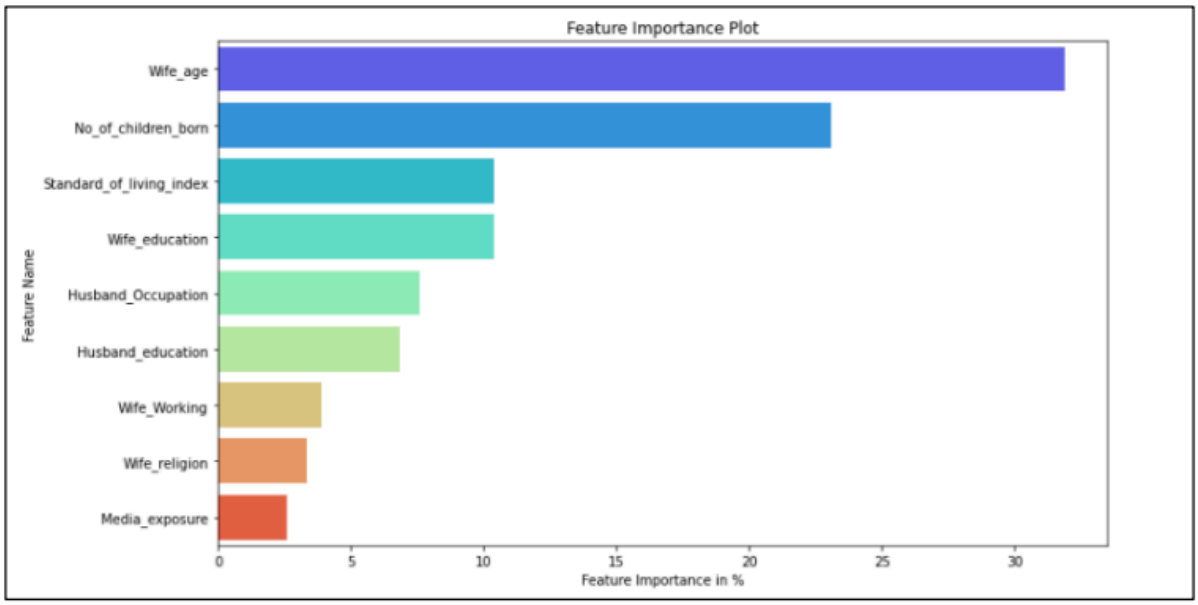
AUC for the Training Data: 0.718
AUC for the Test Data: 0.676



CART:

Feature importance plot

	Imp
Wife_age	0.305473
Wife_education	0.106121
Husband_education	0.068274
No_of_children_born	0.241842
Wife_religion	0.030304
Wife_Working	0.042451
Husband_Occupation	0.076253
Standard_of_living_index	0.108425
Media_exposure	0.020859

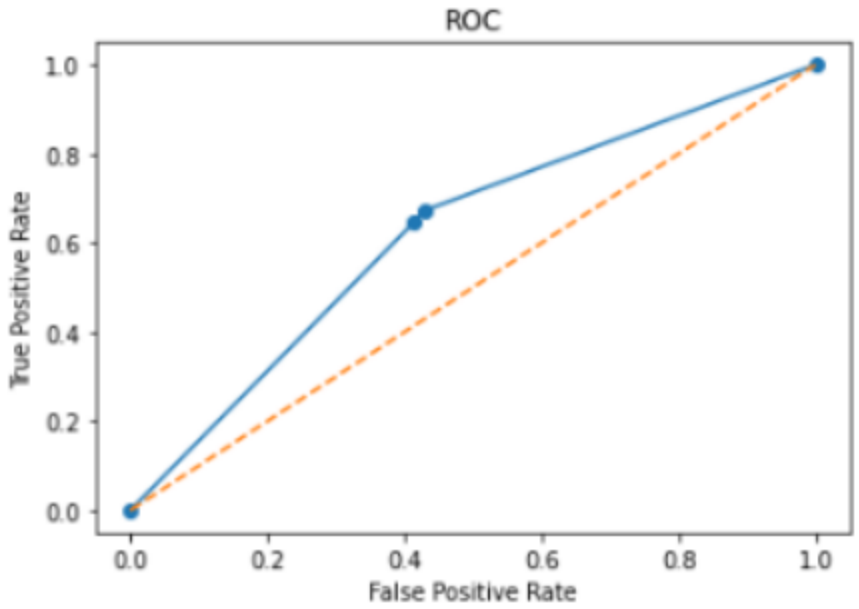


From the above plot we can infer that Wife's age and no. of children born are the most important feature which will help in model building and classification

Classification report of train and test data

	precision	recall	f1-score	support
0	0.74	0.64	0.69	423
1	0.75	0.82	0.78	548
accuracy			0.74	971
macro avg	0.74	0.73	0.74	971
weighted avg	0.74	0.74	0.74	971

	precision	recall	f1-score	support
0	0.70	0.54	0.61	191
1	0.67	0.80	0.73	226
accuracy			0.68	417
macro avg	0.68	0.67	0.67	417
weighted avg	0.68	0.68	0.67	417



INSIGHTS FROM LOGISTIC REGRESSION:

- For predicting the target variable "Contraceptive_method_used " is "No"
- Precison : tells us how many predictions are actually positive out of all the total positive predicted. Precision (64%) – 65% of the people predicted are actually not using contraceptions out of all families predicted to have been not using contraceptions.
-
- For predicting the target variable " Contraceptive_method_used " is "Yes" .
- Precision (63%) – 63% of the people predicted are actually not using contracept ions out of all families predicted to have been not using contraceptions.
-

- Overall accuracy of the model – 63 % of total predictions are correct.

INSIGHTS FROM LDA:

- For predicting the target variable " Contraceptive_method_used " is "No"

- Precision (65%) – 65% of the people predicted are actually not using contracept ions out of all families predicted to have been not using contraceptions.

- For predicting the target variable " Contraceptive_method_used " is "Yes" .

- Precision (68%) – 68% of the people predicted are actually not using contracept ions out of all families predicted to have been not using contraceptions.

- Overall accuracy of the model – 64 % of total predictions are correct.

INSIGHTS FROM CART:

- For predicting the target variable “ Contraceptive_method_used “ is “No”

- Precision (70%) – 70% of the people predicted are actually not using contracept ions out of all families predicted to have been not using contraceptions.

- For predicting the target variable “ Contraceptive_method_used “ is “Yes”

- Precision (67%) – 67% of the people predicted are actually not using contracept ions out of all families predicted to have been not using contraceptions.

- Accuracy of test data is comparatively more in CART(0.68) ,followed by LDA(0.6 4) and LOGISTIC model(0.63).

- AUC is also almost same for all the three models i.e. 0.718 Accuracy, AUC, Precison and Recall for test data is almost inline with training data. This proves no overfitting or underfitting has happened, and overall all the model can be considered suitable for classification.

CONCLUSION: All models gives similar results and hence for given data all models are ideal