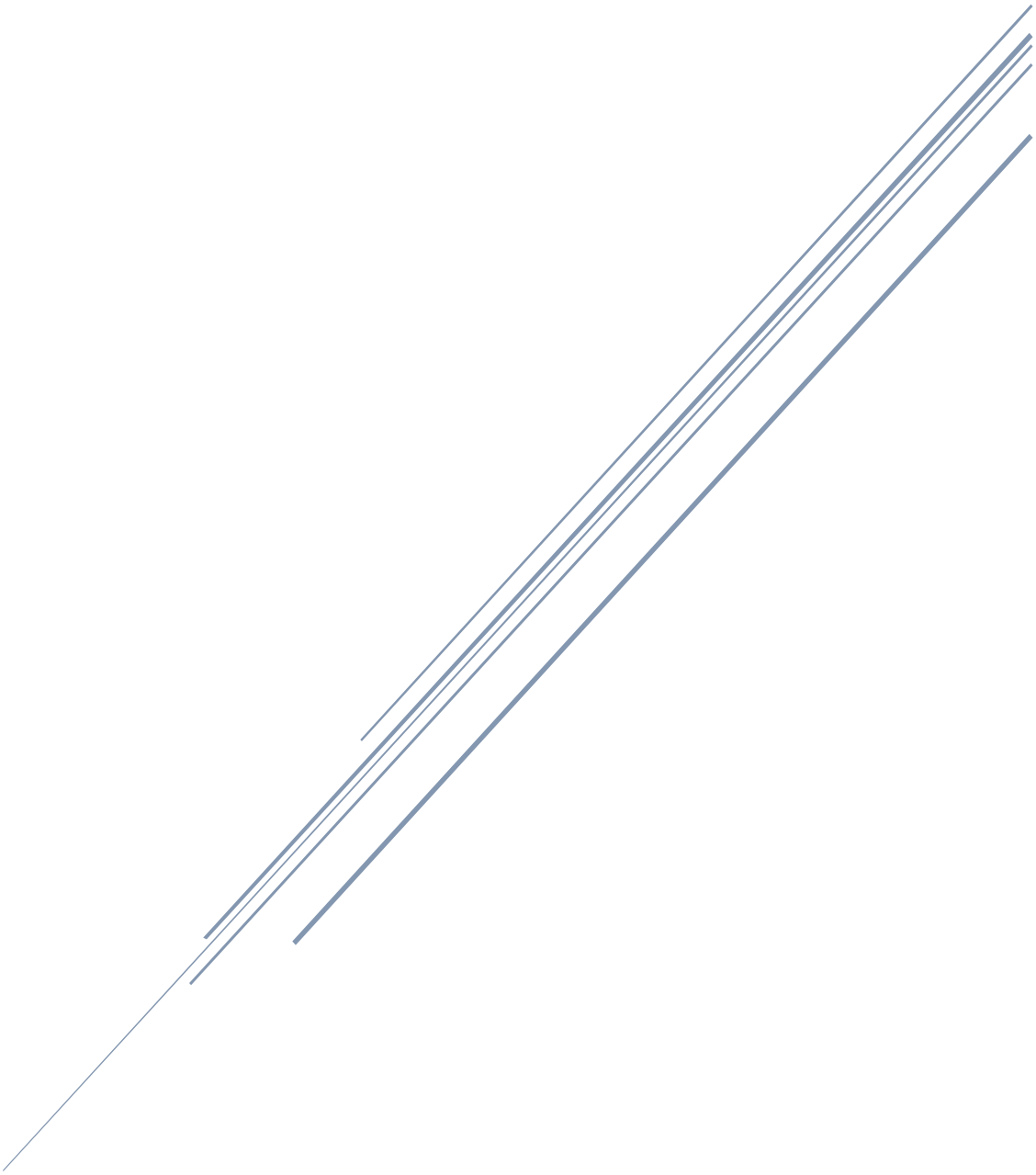
PREDICTIVE MODELLING

Gudla Sai Sirinvas



Predictive Modeling

Contents

Table of Figures

List of Tables

LINEAR REGRESSION

Problem 1: Linear Regression

The comp-activ databases is a collection of a computer systems activity measures . The data was collected from a Sun Sparcstation 20/712 with 128 Mbytes of memory running in a multi-user university department. Users would typically be doing a large variety of tasks ranging from accessing the internet, editing files or running very cpu-bound programs.

As you are a budding data scientist you thought to find out a linear equation to build a model to predict 'usr'(Portion of time (%) that cpus run in user mode) and to find out how each attribute affects the system to be in 'usr' mode using a list of system attributes.

DATA DICTIONARY:

System measures used:

lread - Reads (transfers per second) between system memory and user memory
lwrite - writes (transfers per second) between system memory and user memory
scall - Number of system calls of all types per second
sread - Number of system read calls per second .
swrite - Number of system write calls per second .
fork - Number of system fork calls per second.
exec - Number of system exec calls per second.
rchar - Number of characters transferred per second by system read calls
wchar - Number of characters transfreed per second by system write calls
pgout - Number of page out requests per second
ppgout - Number of pages, paged out per second
pgfree - Number of pages per second placed on the free list.
pgscan - Number of pages checked if they can be freed per second
atch - Number of page attaches (satisfying a page fault by reclaiming a page in memory) per second
pgin - Number of page-in requests per second
ppgin - Number of pages paged in per second
pflt - Number of page faults caused by protection errors (copy-on-writes).
vflt - Number of page faults caused by address translation .
runqsz - Process run queue size (The number of kernel threads in memory that are waiting for a CPU to run.
Typically, this value should be less than 2. Consistently higher values mean that the system might be CPU-bound.)
freemem - Number of memory pages available to user processes
freeswap - Number of disk blocks available for page swapping.
---
usr - Portion of time (%) that cpus run in user mode

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.

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

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

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

Table 1. Data types

Data Exploration was performed using the following functions:

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

	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

Table 2. First 5 rows

1.Head

2.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

Table 3. Last 5 rows

1. Shape.

Dataset has 8192 rows and 22 features

2. Summary**:**

	count	mean	std	min	25%	50%	75%	max
lread	8192.0	1.955969e+01	53.353799	0.0	2.0	7.0	20.000	1845.00
lwrite	8192.0	1.310620e+01	29.891726	0.0	0.0	1.0	10.000	575.00
scall	8192.0	2.306318e+03	1633.617322	109.0	1012.0	2051.5	3317.250	12493.00
sread	8192.0	2.104800e+02	198.980146	6.0	86.0	166.0	279.000	5318.00
swrite	8192.0	1.500582e+02	160.478980	7.0	63.0	117.0	185.000	5456.00
fork	8192.0	1.884554e+00	2.479493	0.0	0.4	0.8	2.200	20.12
exec	8192.0	2.791998e+00	5.212456	0.0	0.2	1.2	2.800	59.56
rchar	8088.0	1.973857e+05	239837.493526	278.0	34091.5	125473.5	267828.750	2526649.00
wchar	8177.0	9.590299e+04	140841.707911	1498.0	22916.0	46619.0	106101.000	1801623.00
pgout	8192.0	2.285317e+00	5.307038	0.0	0.0	0.0	2.400	81.44
ppgout	8192.0	5.977229e+00	15.214590	0.0	0.0	0.0	4.200	184.20
pgfree	8192.0	1.191971e+01	32.363520	0.0	0.0	0.0	5.000	523.00
pgscan	8192.0	2.152685e+01	71.141340	0.0	0.0	0.0	0.000	1237.00
atch	8192.0	1.127505e+00	5.708347	0.0	0.0	0.0	0.600	211.58
pgin	8192.0	8.277960e+00	13.874978	0.0	0.6	2.8	9.765	141.20
ppgin	8192.0	1.238859e+01	22.281318	0.0	0.6	3.8	13.800	292.61
pflt	8192.0	1.097938e+02	114.419221	0.0	25.0	63.8	159.600	899.80
vflt	8192.0	1.853158e+02	191.000603	0.2	45.4	120.4	251.800	1365.00
freemem	8192.0	1.763456e+03	2482.104511	55.0	231.0	579.0	2002.250	12027.00
freeswap	8192.0	1.328126e+06	422019.426957	2.0	1042623.5	1289289.5	1730379.500	2243187.00
usr	8192.0	8.396887e+01	18.401905	0.0	81.0	89.0	94.000	99.00

Table 4. Summary of the Data

5.Check Duplicates:

No Duplicates were observed in the dataset.

6.Null Values:

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

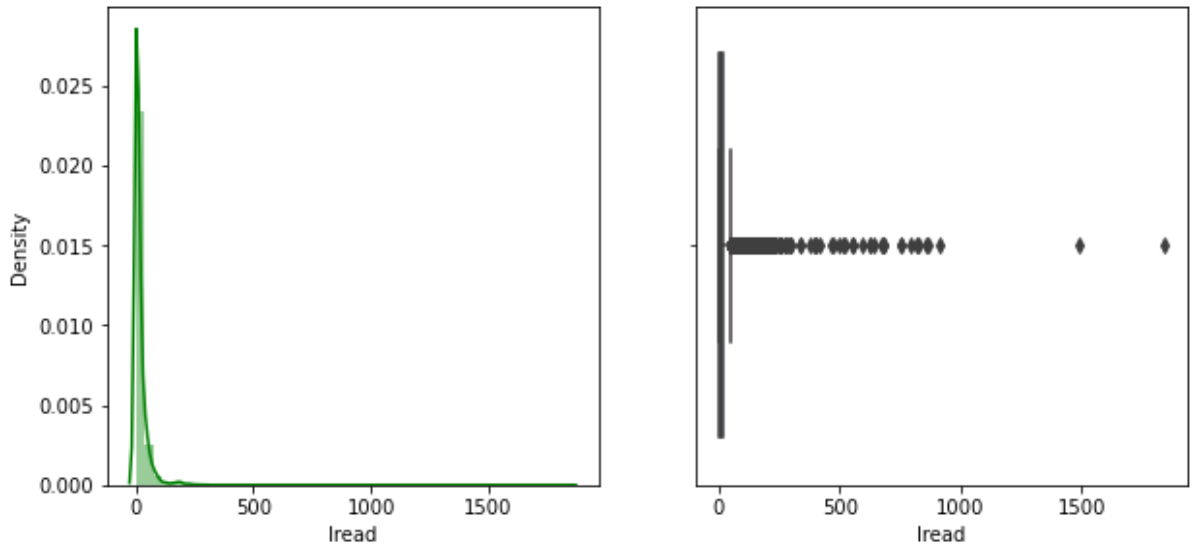
- **Rchar**
- **Wchar**

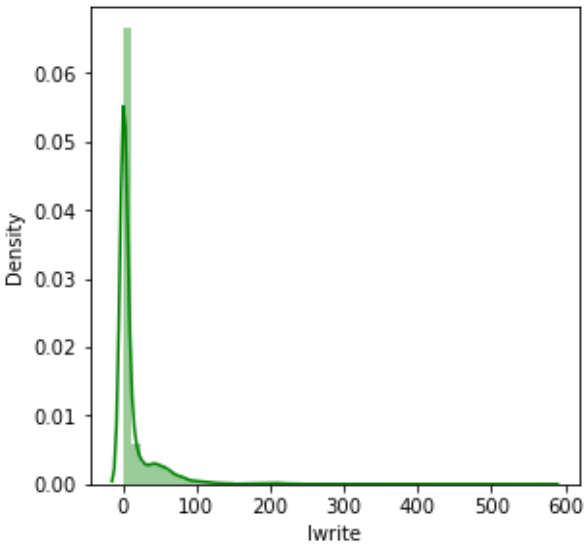
```
lread          0
lwrite         0
scall          0
sread          0
swrite         0
fork           0
exec           0
rchar        104
wchar         15
pgout          0
ppgout         0
pgfree         0
pgscan         0
atch           0
pgin           0
ppgin          0
pflt           0
vflt           0
runqsz         0
freemem        0
freeswap       0
usr            0
dtype: int64
```

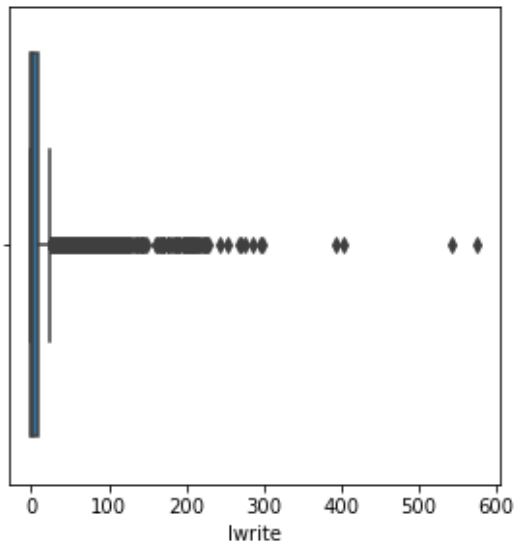
Fig. 1. Missing values

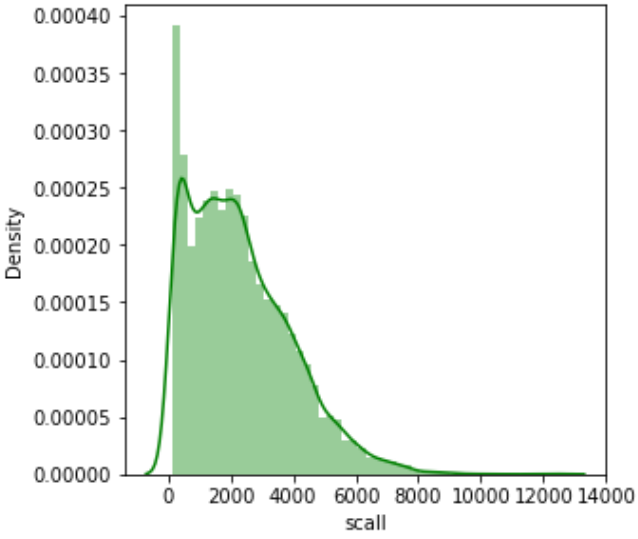
The missing values were treated by calculating its mean values.

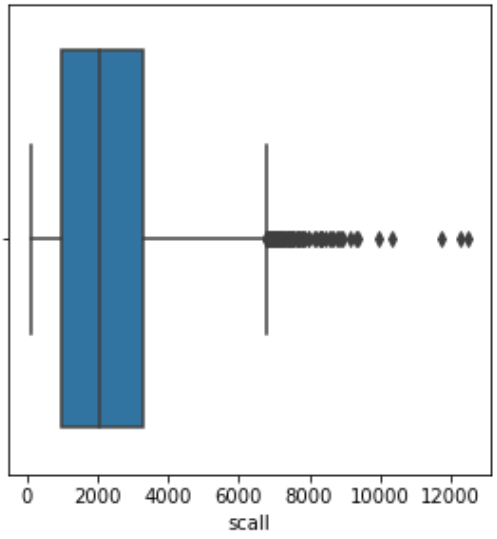
UNIVARIATE ANALYSIS:

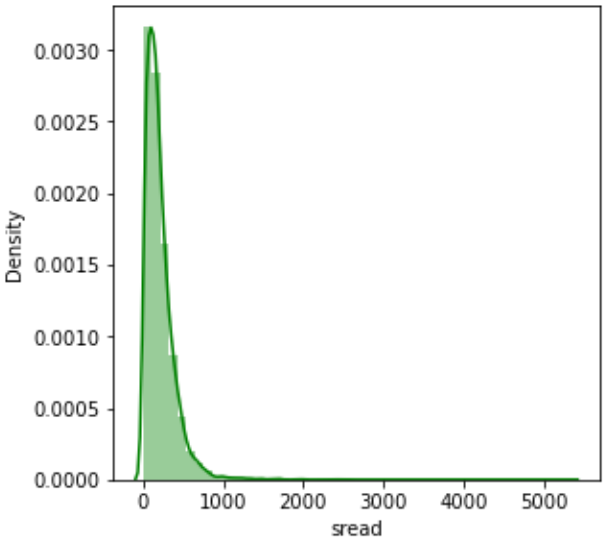


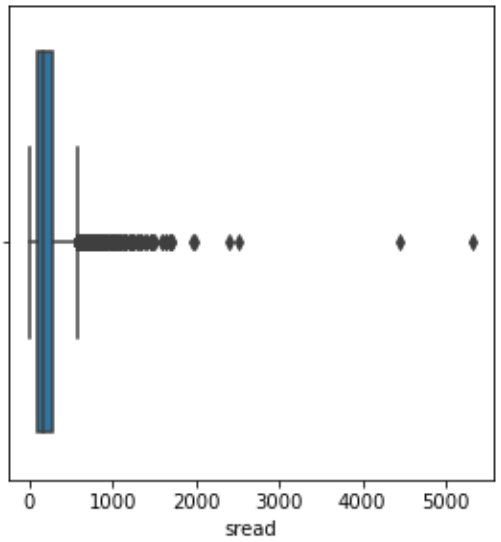

Density
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
100
200
300
400
500
600
lwrite

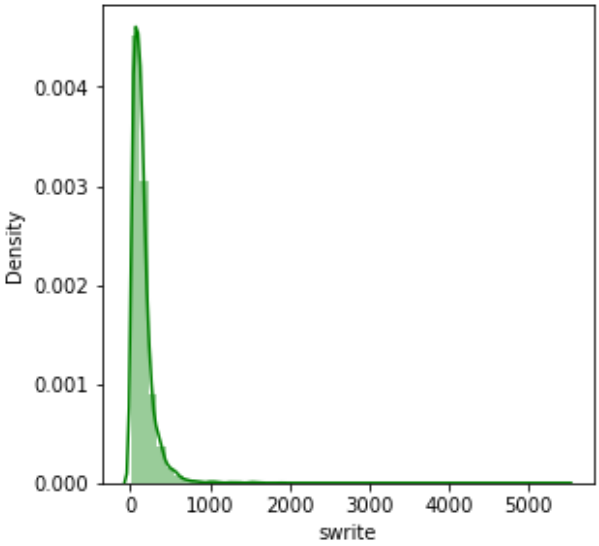

0
100
200
300
400
500
600
lwrite

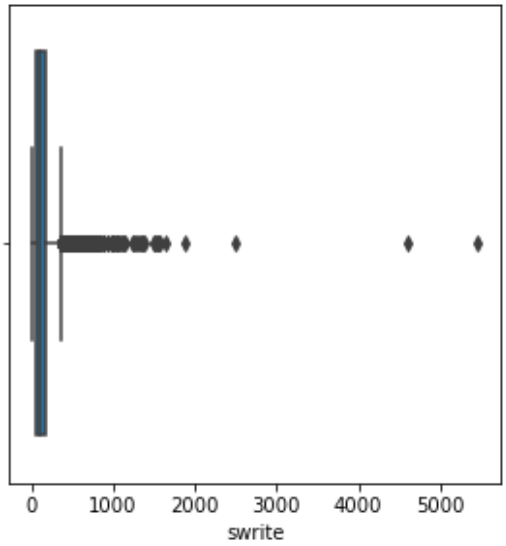

Density
0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
2000
4000
6000
8000
10000
12000
14000
scall

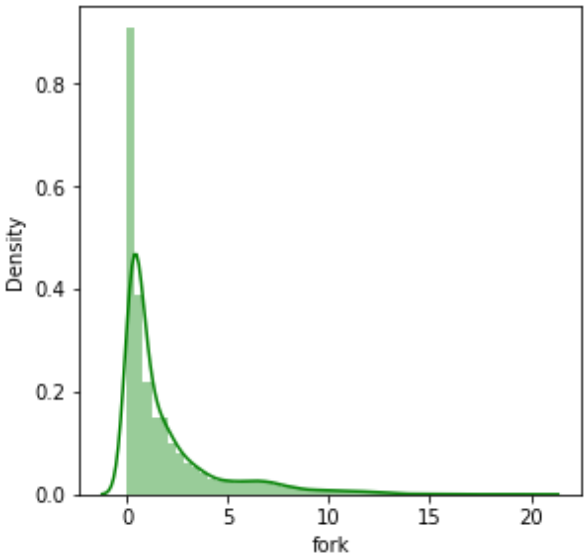

0
2000
4000
6000
8000
10000
12000
scall

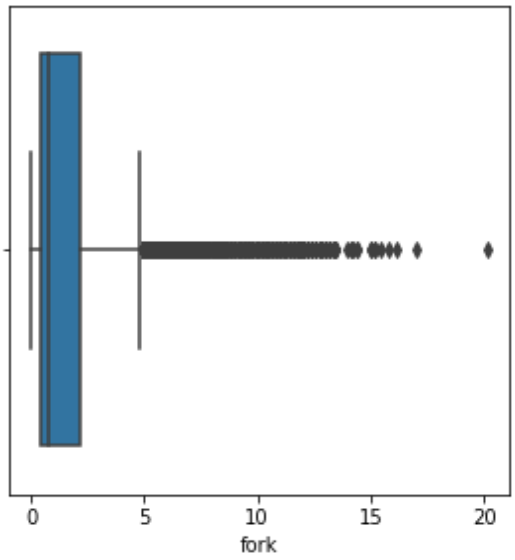

Density
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
1000
2000
3000
4000
5000
sread

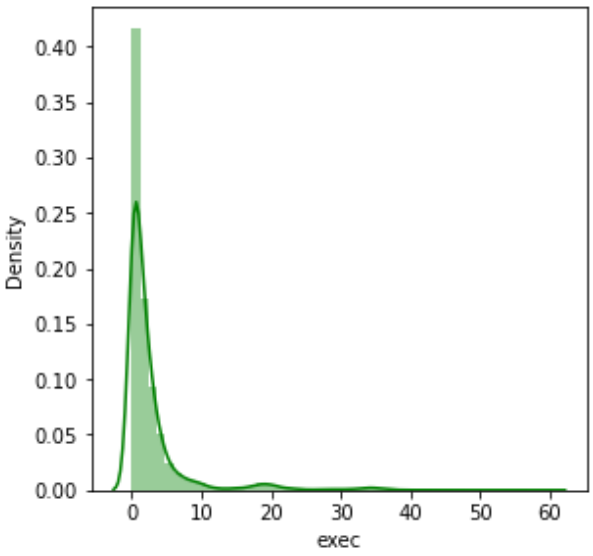

0
1000
2000
3000
4000
5000
sread

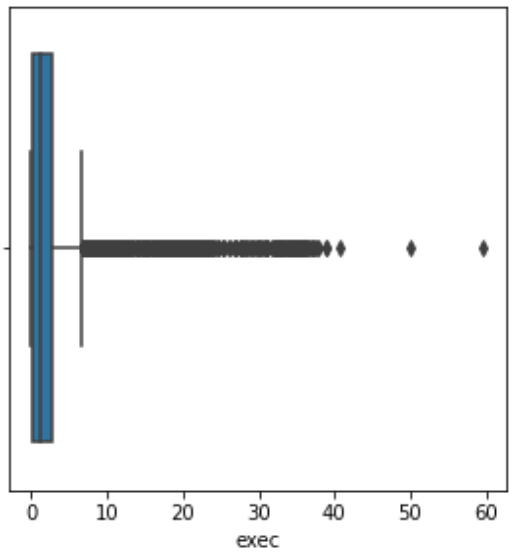

0.004
0.003
0.002
0.001
0.000
Density
0
1000
2000
3000
4000
5000
swrite

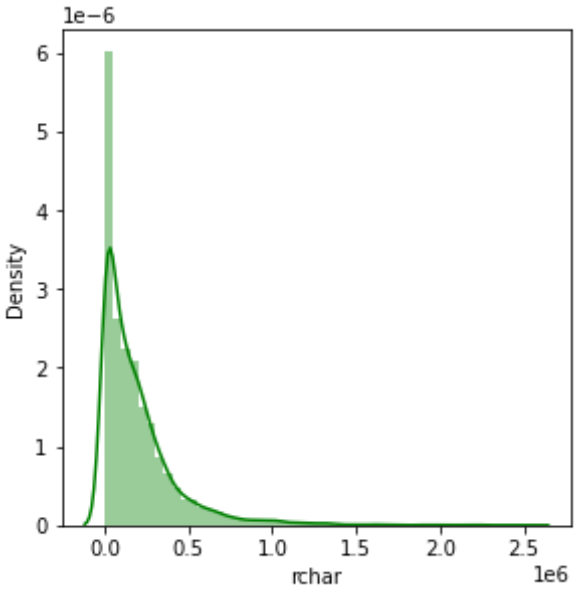

0
1000
2000
3000
4000
5000
swrite

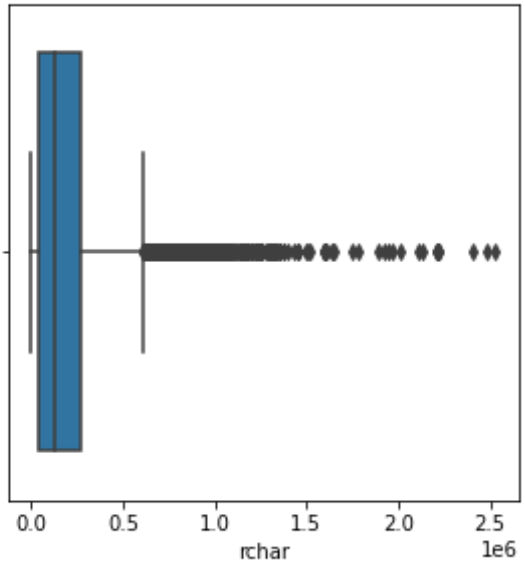

0.8
0.6
0.4
0.2
0.0
Density
0
5
10
15
20
fork

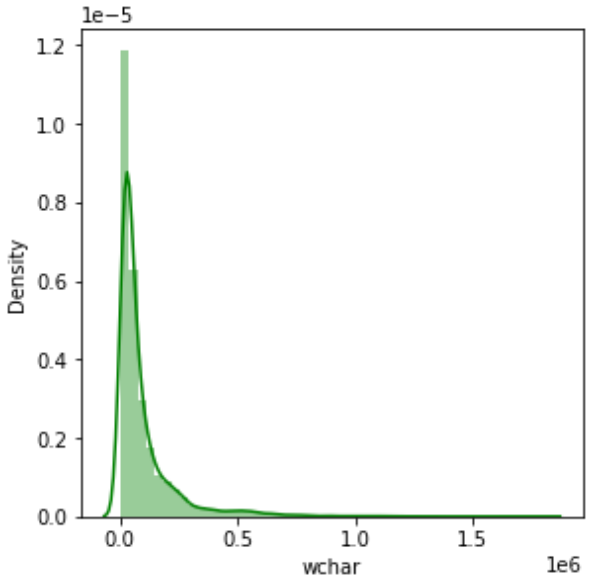

0
5
10
15
20
fork

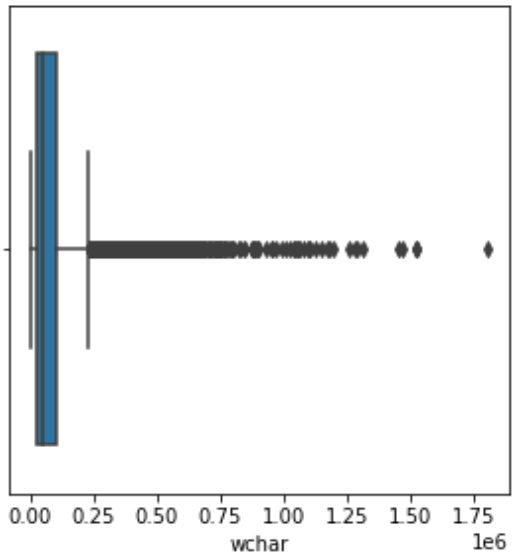

Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
10
20
30
40
50
60
exec

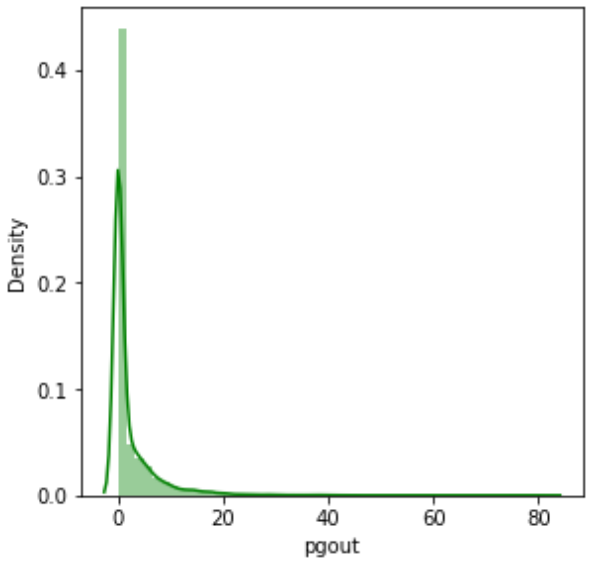

0
10
20
30
40
50
60
exec

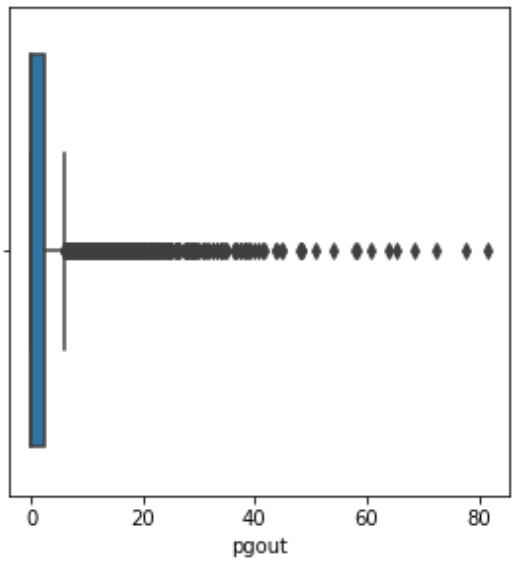

1e−6
Density
6
5
4
3
2
1
0
0.0
0.5
1.0
1.5
2.0
2.5
rchar
1e6

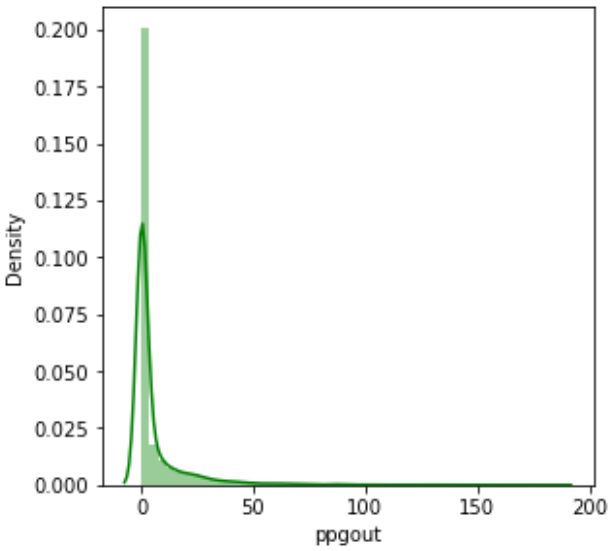

0.0
0.5
1.0
1.5
2.0
2.5
rchar
1e6

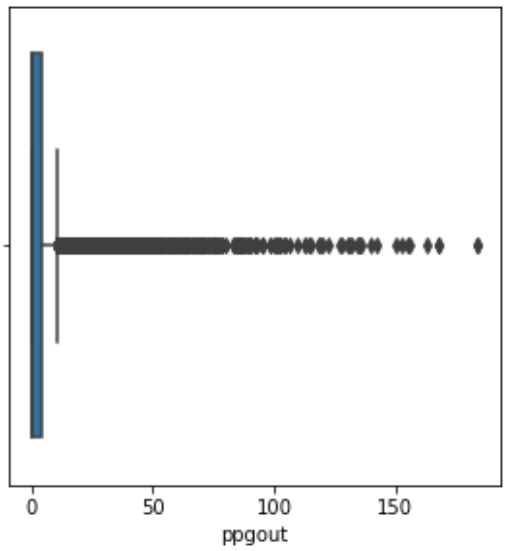

1e−5
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Density
0.0
0.5
1.0
1.5
1e6
wchar

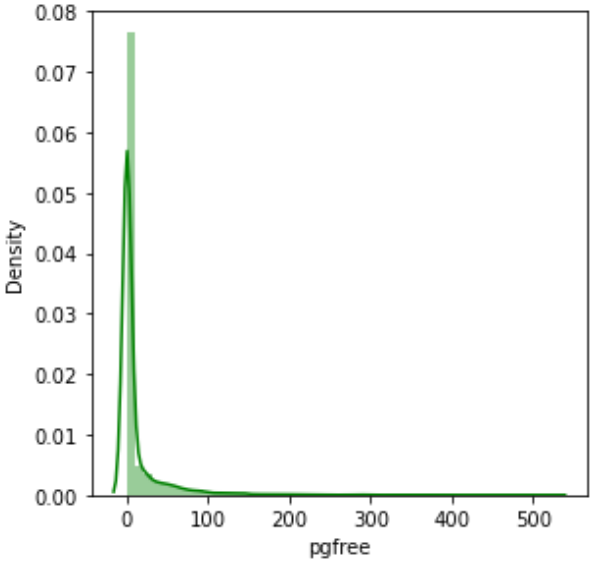

0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
1e6
wchar

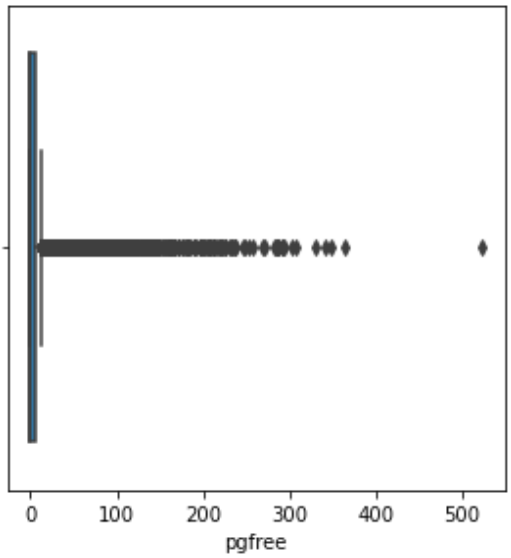

0.4
0.3
0.2
0.1
0.0
Density
0
20
40
60
80
pgout

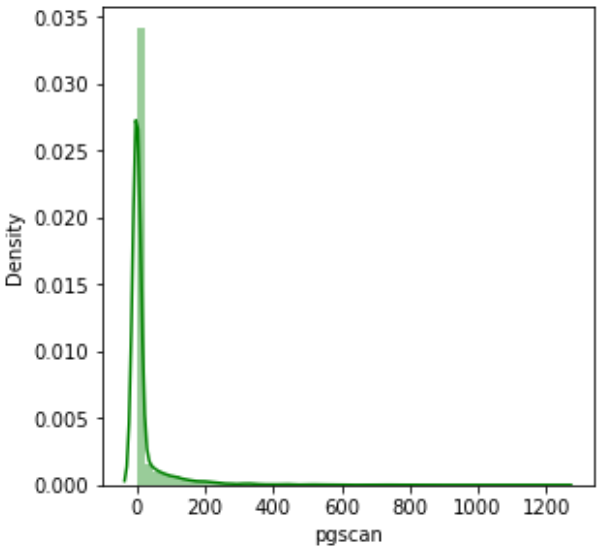

0
20
40
60
80
pgout

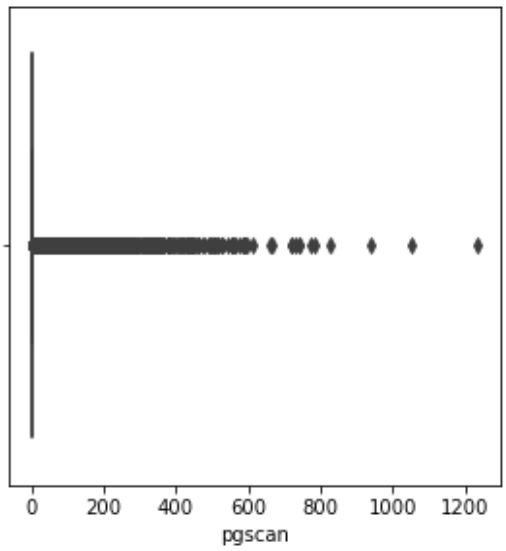

0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Density
0
50
100
150
200
ppgout

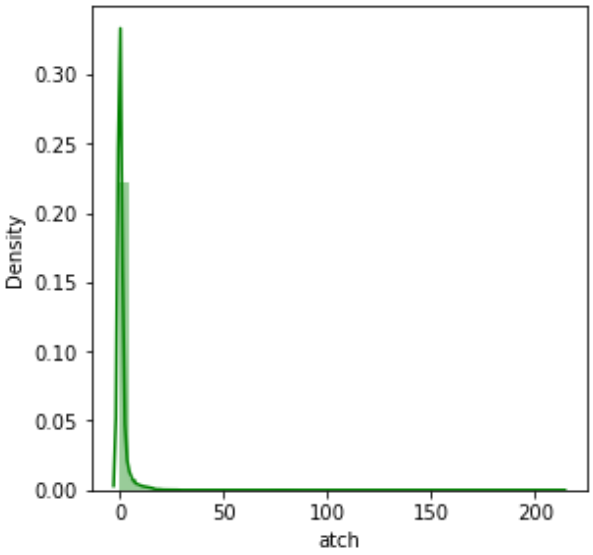

0
50
100
150
ppgout

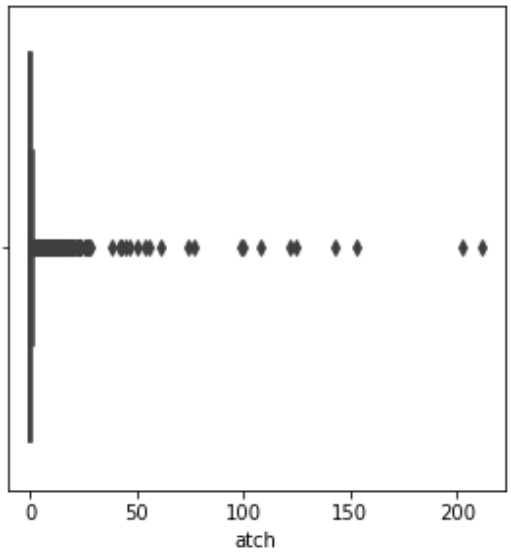

0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Density
0
100
200
300
400
500
pgfree

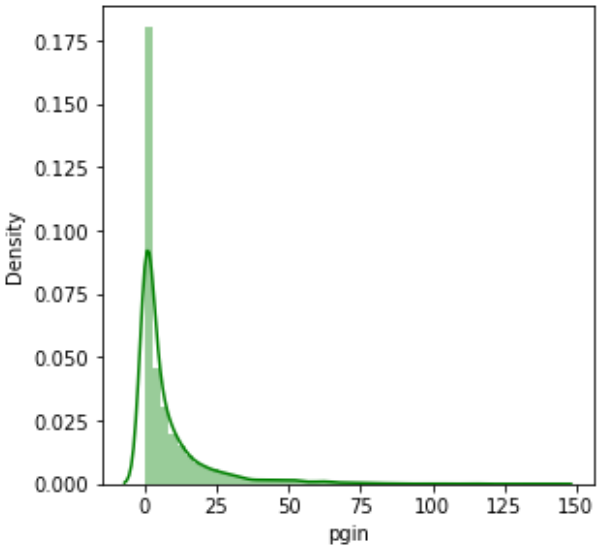

0
100
200
300
400
500
pgfree

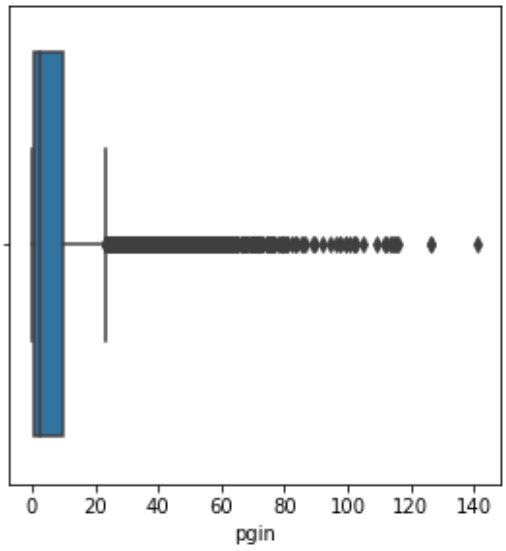

0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
0
200
400
600
800
1000
1200
pgscan

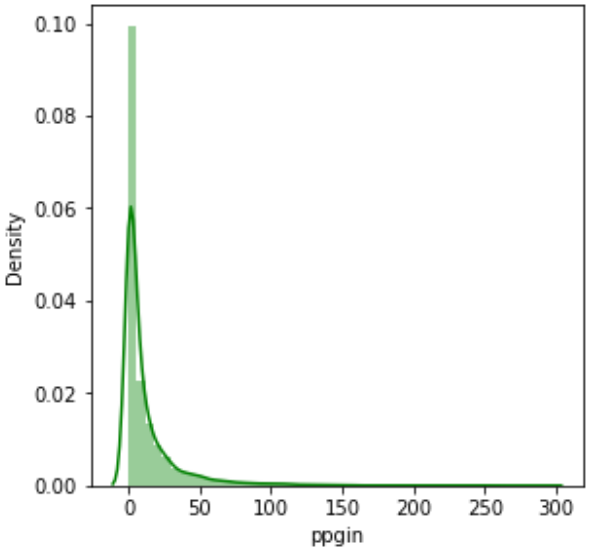

0
200
400
600
800
1000
1200
pgscan

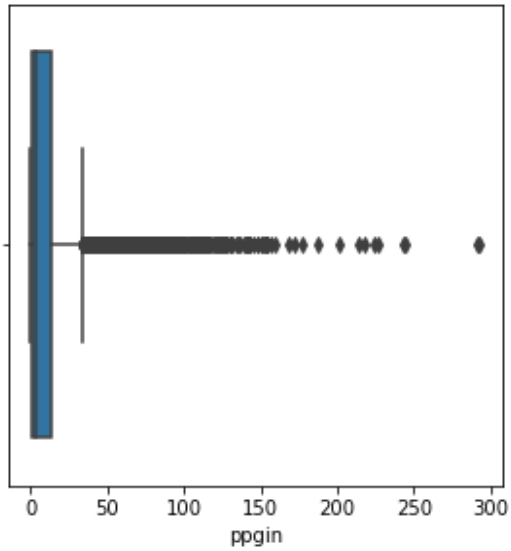

0.30
0.25
0.20
0.15
0.10
0.05
0.00
Density
0
50
100
150
200
atch

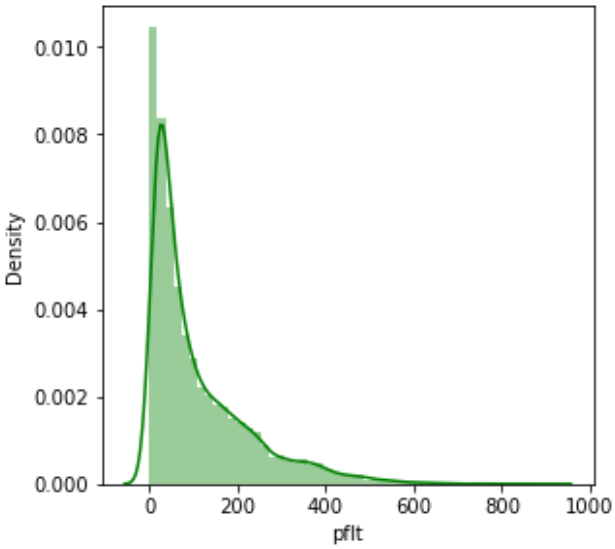

0
50
100
150
200
atch

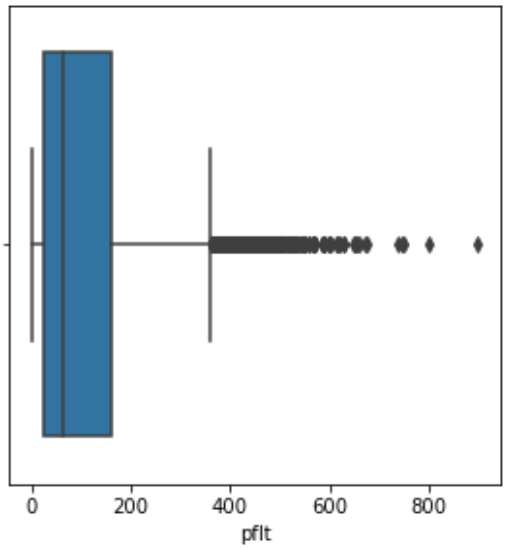

0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Density
0
25
50
75
100
125
150
pgin

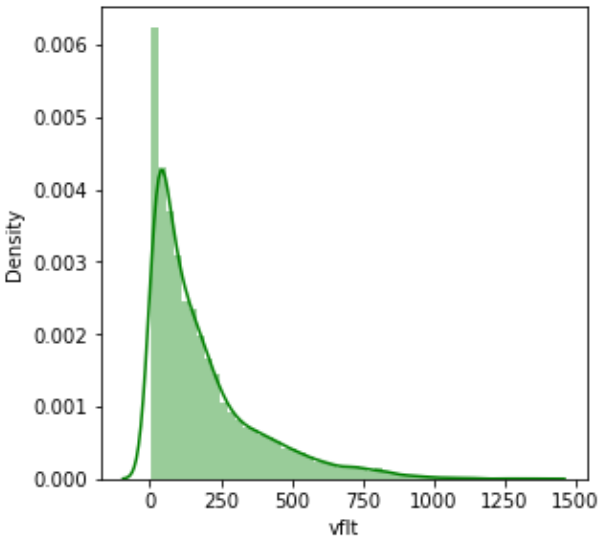

0
20
40
60
80
100
120
140
pgin

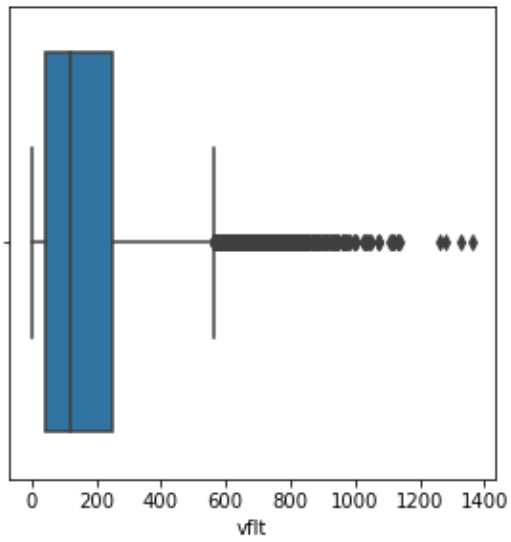

0.10
0.08
0.06
0.04
0.02
0.00
Density
0
50
100
150
200
250
300
ppgin

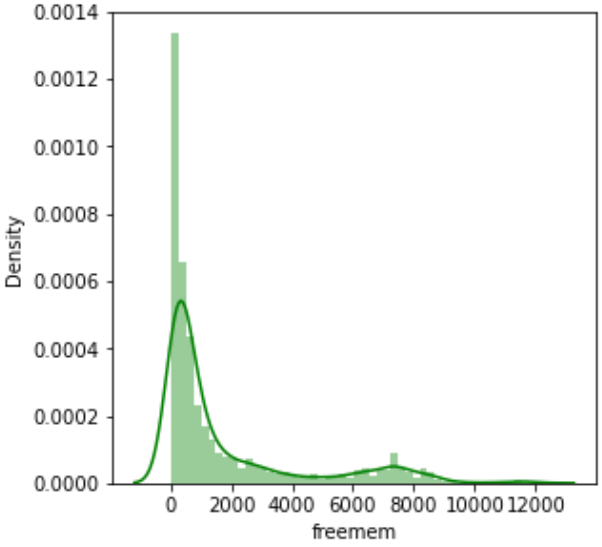

0
50
100
150
200
250
300
ppgin

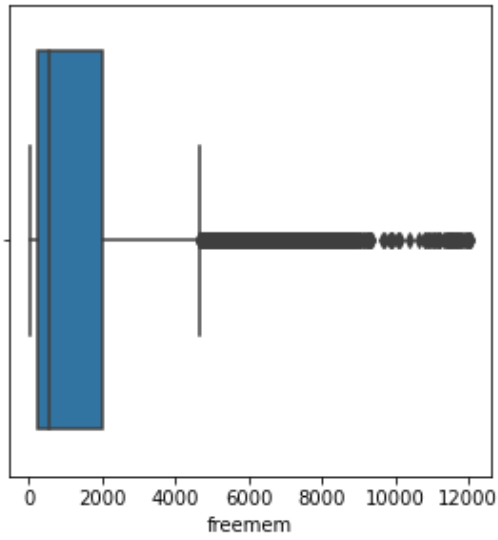

0.010
0.008
0.006
0.004
0.002
0.000
Density
0
200
400
600
800
1000
pflt

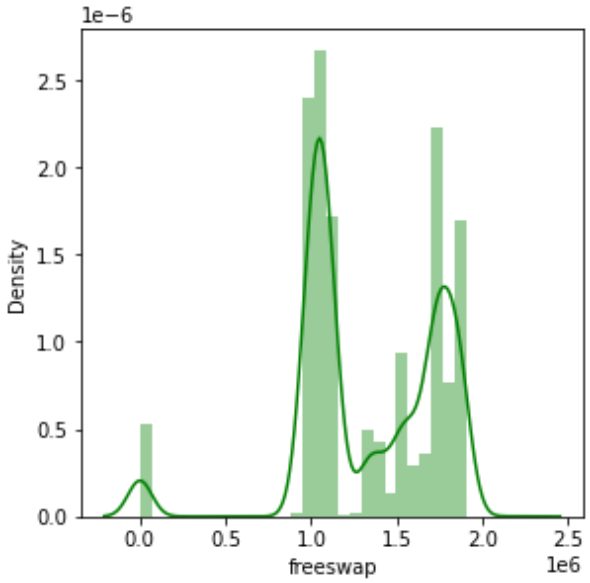

0
200
400
600
800
pflt

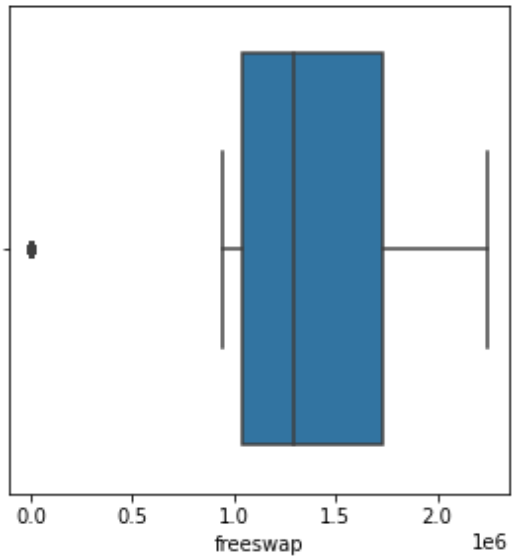

0.006
0.005
0.004
0.003
0.002
0.001
0.000
Density
0
250
500
750
1000
1250
1500
vflt

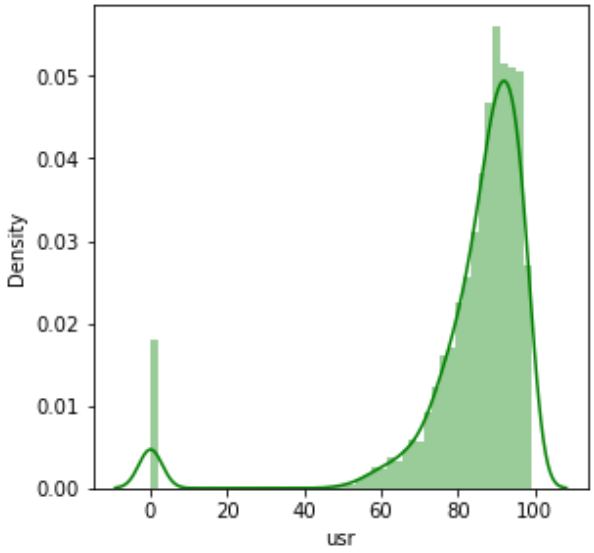

0
200
400
600
800
1000
1200
1400
vflt

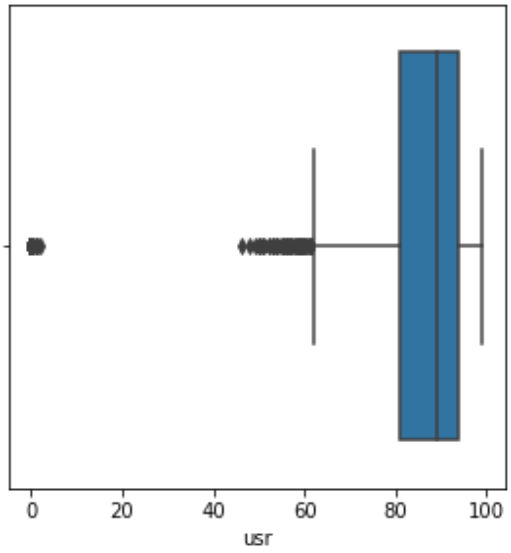

0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Density
0
2000
4000
6000
8000
10000
12000
freemem


0
2000
4000
6000
8000
10000
12000
freemem


1e−6
2.5
2.0
1.5
1.0
0.5
0.0
Density
0.0
0.5
1.0
1.5
2.0
2.5
freeswap
1e6


0.0
0.5
1.0
1.5
2.0
freeswap
1e6


0.05
0.04
0.03
0.02
0.01
0.00
Density
0
20
40
60
80
100
usr


0
20
40
60
80
100
usr

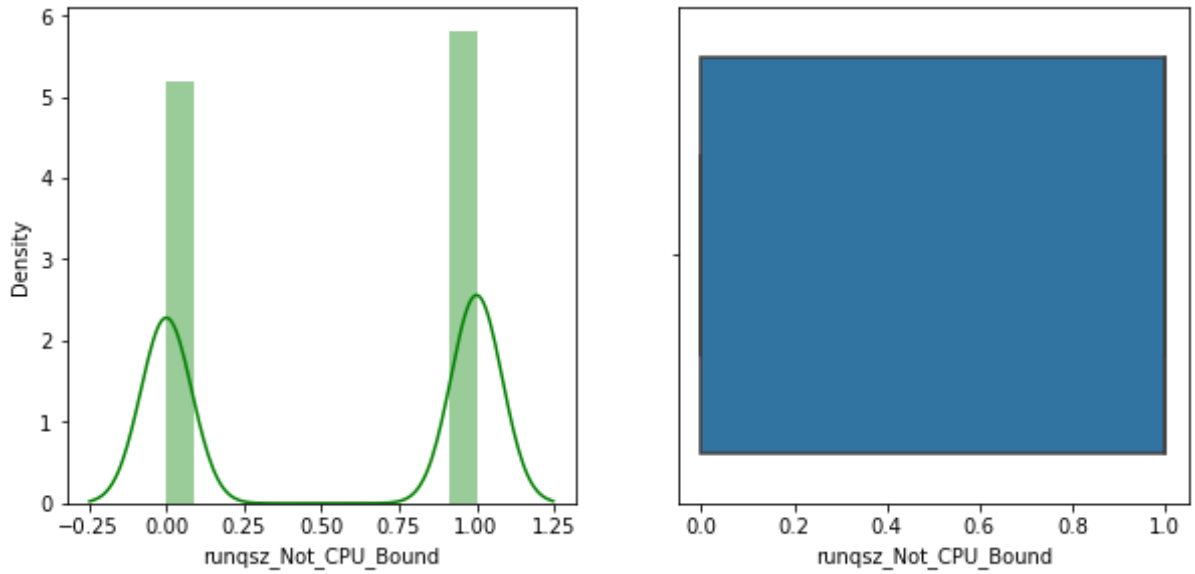


Fig. 2. Univariate analysis graphs

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

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

BIVARIATE ANALYSIS:

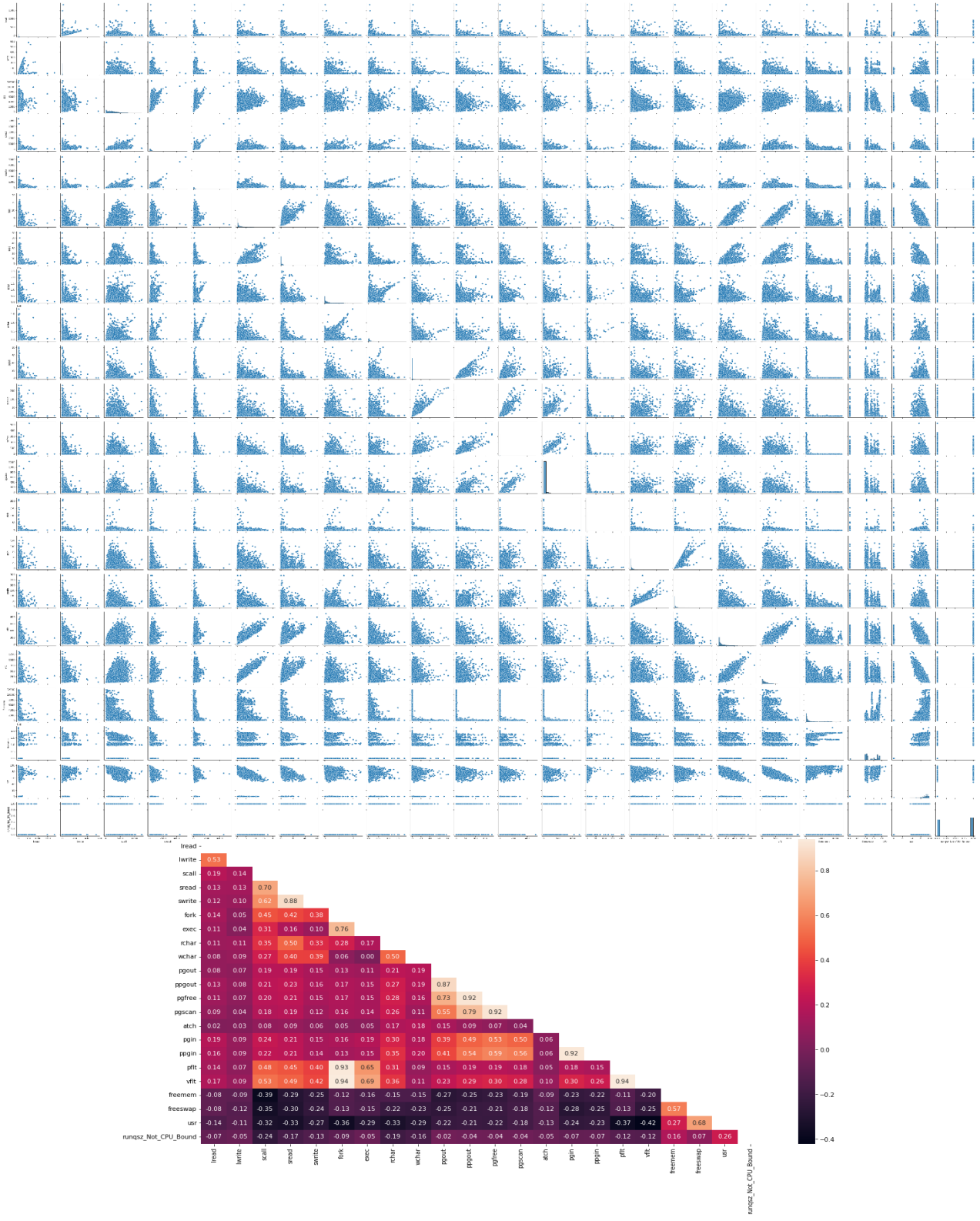


Fig. 3. Bivariate Analysis Graphs

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

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.**

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

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

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

Variables with outliers:

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

Duplicate values were not observed in the dataset.

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.**

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

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

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

```
Output exceeds the size limit. Open the full output data in a text editor
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.620
Model:                            OLS   Adj. R-squared:                  0.619
Method:                 Least Squares   F-statistic:                     499.9
Date:                Sun, 05 Feb 2023   Prob (F-statistic):               0.00
Time:                        18:11:51   Log-Likelihood:                -23560.
No. Observations:                6144   AIC:                         4.716e+04
Df Residuals:                    6123   BIC:                         4.730e+04
Df Model:                          20
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         45.3637      0.770     58.944      0.000      43.855      46.872
lread         -0.0767      0.022     -3.500      0.000      -0.120      -0.034
lwrite         0.0390      0.032      1.218      0.223      -0.024       0.102
scall          0.0011      0.000      6.909      0.000       0.001       0.001
sread          0.0016      0.002      0.642      0.521      -0.003       0.006
swrite        -0.0056      0.003     -1.598      0.110      -0.012       0.001
fork          -0.8859      0.322     -2.751      0.006      -1.517      -0.255
exec          -0.0167      0.126     -0.132      0.895      -0.263       0.230
rchar      -1.031e-05   1.19e-06     -8.652      0.000   -1.26e-05   -7.97e-06
...
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 4.07e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
"""
```

Table 5. Linear regression model with original data

R^2 value for = 0.62

From the above tables it is observed that:

R^2 value: 0.62(62%) which is low for a model to be processed further.

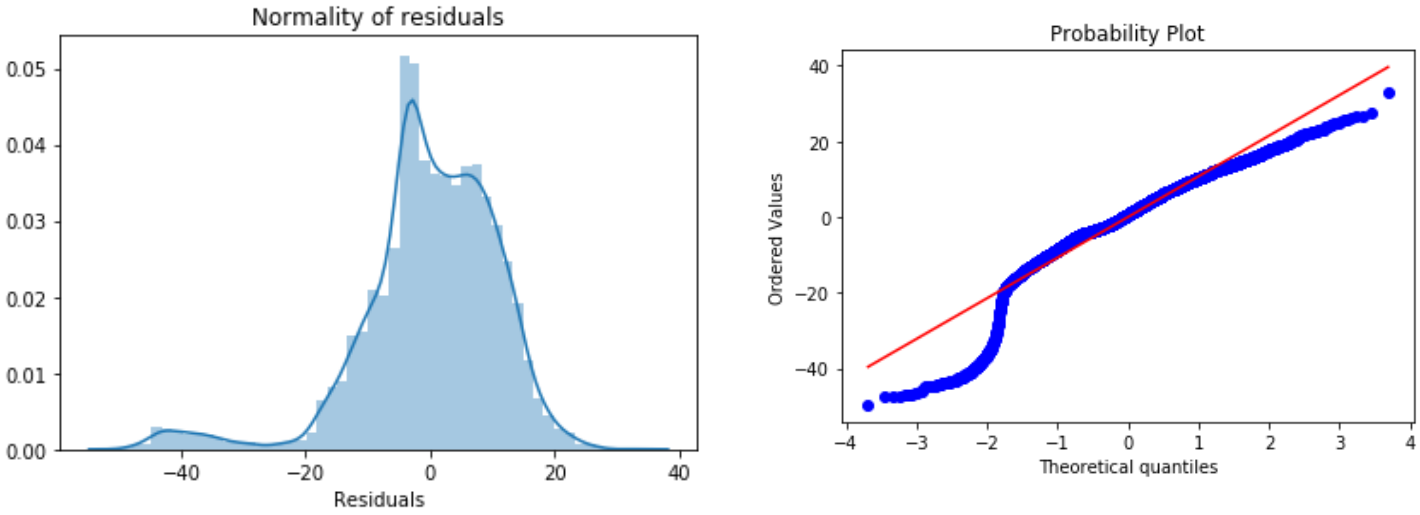


Fig. 4. Normal Distribution of Residual

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

After transformation it is observed that:

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.867
Model:	OLS	Adj. R-squared:	0.866
Method:	Least Squares	F-statistic:	2655.
Date:	Sun, 05 Feb 2023	Prob (F-statistic):	0.00
Time:	19:20:46	Log-Likelihood:	-1160.3
No. Observations:	5734	AIC:	2351.
Df Residuals:	5719	BIC:	2450.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.4463	0.037	-12.056	0.000	-0.519	-0.374
lwrite	0.0056	0.004	1.456	0.145	-0.002	0.013
scall	0.0246	0.002	10.935	0.000	0.020	0.029
swrite	0.0100	0.006	1.606	0.108	-0.002	0.022
fork	-0.3463	0.041	-8.444	0.000	-0.427	-0.266
exec	0.0728	0.019	3.796	0.000	0.035	0.110
rchar	-0.0024	0.000	-9.427	0.000	-0.003	-0.002
pgout	-0.0821	0.009	-9.480	0.000	-0.099	-0.065
atch	0.0506	0.012	4.157	0.000	0.027	0.075
pgin	0.0693	0.006	11.155	0.000	0.057	0.082
pflt	-0.1039	0.010	-10.149	0.000	-0.124	-0.084
vflt	0.1067	0.008	12.566	0.000	0.090	0.123
freemem	-0.0341	0.002	-21.404	0.000	-0.037	-0.031
freeswap	0.0446	0.000	170.948	0.000	0.044	0.045
runqsz_Not_CPU_Bound	0.2219	0.008	26.301	0.000	0.205	0.238

Omnibus:	5.303	Durbin-Watson:	2.040
Prob(Omnibus):	0.071	Jarque-Bera (JB):	4.781
Skew:	-0.013	Prob(JB):	0.0916
Kurtosis:	2.861	Cond. No.	1.35e+03

Table 6. Regressional model with transformed data after dimensional reduction using VIF values .

R^2 = 0.868 (for the transformed data)

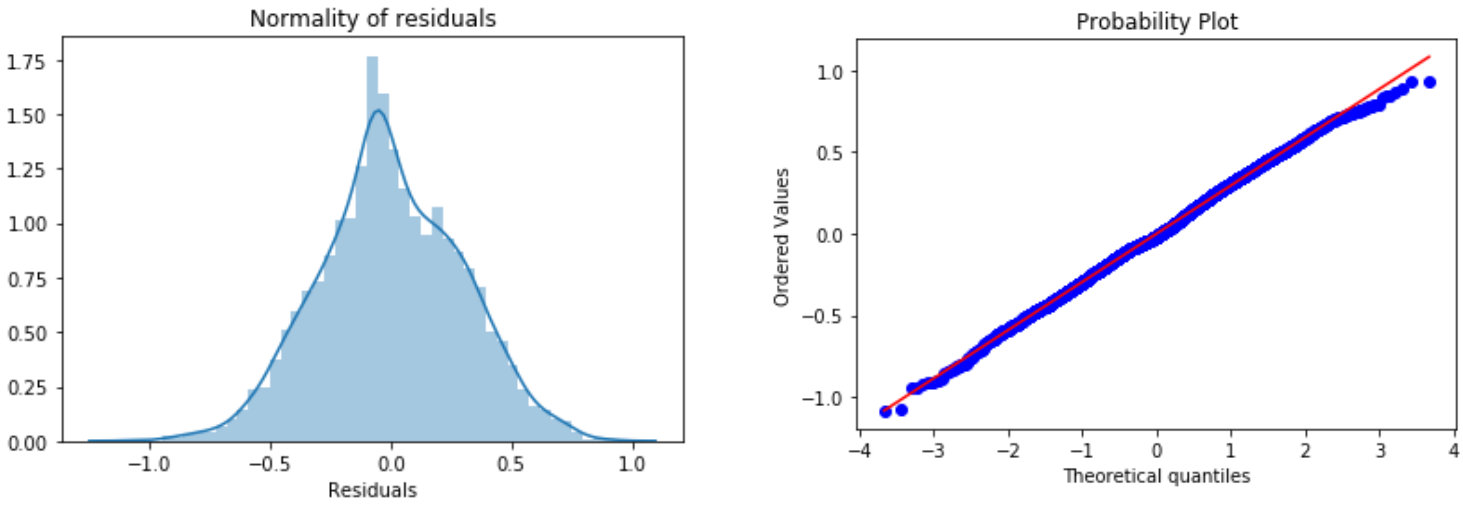


Fig. 5. Normal distribution of residuals with transformed data

The residuals for the transformed data are now normally distributed which can be observed in the above graphs. The R^2 values has increased to 0.867(87%) which states that the transformed model is delivering better outcomes.

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

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

R^2 for Train data = 0.867

R^2 for Test data = 0.88

Adj. R-squared for Train data: 0.866

Adj. R-squared for Test data: 0.879

RMSE value for Train data : 0.29

RMSE value for Test data : 0.31

R^2, Adj. R^2, RMSE value for train and test are nearly equal to each other.

The regression model with original data have R^2 value is 0.62 (approx.). But for the transformed data model R^2 value is 0.867.

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

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

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

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

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

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

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

vflt - Number of page faults caused by address translation

scall – Number of system calls of all types per second

swrite - Number of system write calls per second .

fork – Number of system fork calls per second.

exec – Number of system exec calls per second.
rchar - Number of characters transferred per second by system read calls

pgout – Number of page out requests per second

atch – Number of page attaches (satisfying a page fault by reclaiming a page in memory) per second

pgin - Number of page-in requests per second

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

freemem – Number of memory pages available to user processes

freeswap – Number of disk blocks available for page swapping.

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

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

- EDA process to deep dive in the data and understand each variable significance and the correlation with each other
- Data cleaning by imputing null values, treating outliers to avoid any biasness in the data
- Encoding the string values to satisfy the data requirement while running the model. (numerical values/ categorical acceptable for running the models)

- Applied linear regression model by using scikit learn for both original data and transformed data.
- Justified the transformed model is the best model using R square, RMSE, Adj R square values.
- Dimensionality reduction has been taken by using VIF values
- Find the best fit line for the linear regression and given with equation

Problem 2: Logistic Regression, LDA and CART

You are a statistician at the Republic of Indonesia Ministry of Health and you are provided with a data of 1473 females collected from a Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of the survey.

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

Data Dictionary:

1. Wife's age (numerical)
2. Wife's education (categorical) 1=uneducated, 2, 3, 4=tertiary
3. Husband's education (categorical) 1=uneducated, 2, 3, 4=tertiary
4. Number of children ever born (numerical)
5. Wife's religion (binary) Non-Scientology, Scientology
6. Wife's now working? (binary) Yes, No
7. Husband's occupation (categorical) 1, 2, 3, 4(random)
8. Standard-of-living index (categorical) 1=verlow, 2, 3, 4=high
9. Media exposure (binary) Good, Not good
10. Contraceptive method used (class attribute) No,Yes

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

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

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

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

Table 7. Data types

Data Exploration was performed using the following functions:

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

1.Head:

Wife_age	Wife_education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_method_used
24	Primary	Secondary	3	Scientology	No	2	High	Exposed	No
45	Uneducated	Secondary	10	Scientology	No	3	Very High	Exposed	No
43	Primary	Secondary	7	Scientology	No	3	Very High	Exposed	No
42	Secondary	Primary	9	Scientology	No	3	High	Exposed	No
36	Secondary	Secondary	8	Scientology	No	3	Low	Exposed	No

Table 8. First 5 rows of the data

2.Tail:

Wife_age	Wife_education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_method_used
33	Tertiary	Tertiary	NaN	Scientology	Yes	2	Very High	Exposed	Yes
33	Tertiary	Tertiary	NaN	Scientology	No	1	Very High	Exposed	Yes
39	Secondary	Secondary	NaN	Scientology	Yes	1	Very High	Exposed	Yes
33	Secondary	Secondary	NaN	Scientology	Yes	2	Low	Exposed	Yes

17	Secon dary	Second ary	1	Scien tolog y	No	2	Very High	Expos ed	Yes

Table 9. Last five rows of the data

3. Shape:

The dataset has 1473 rows and 10 variables.

4. Summary:

	count	mean	std	min	25%	50%	75%	max
Wife_age	1259.0	32.516283	8.303964	16.0	26.0	32.0	39.0	49.0
No_of_children_born	1301.0	3.327440	2.431604	0.0	1.0	3.0	5.0	16.0
Husband_Occupation	1322.0	2.208018	0.843461	1.0	1.0	2.0	3.0	4.0

Table 10. Summary of the data

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

5. Checking Null Values:

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

```
Wife_age                    63
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
```

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

UNIVARIATE ANALYSIS:

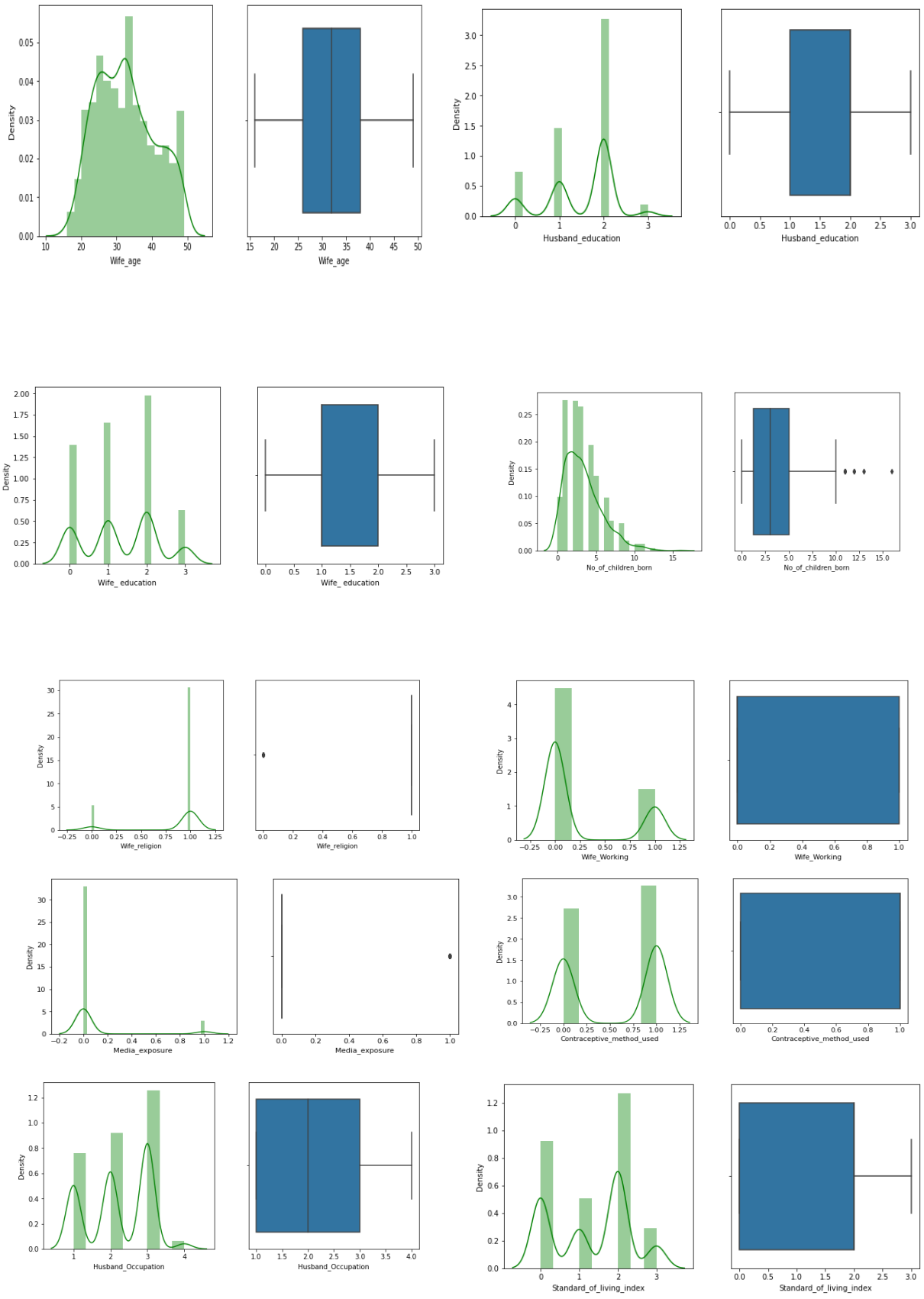


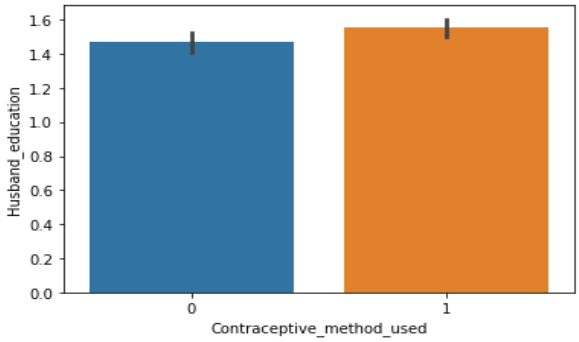
Fig. 6. Univariate analysis

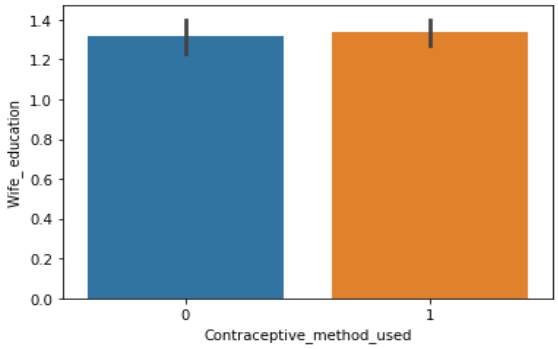
From the above graphs we can say that:

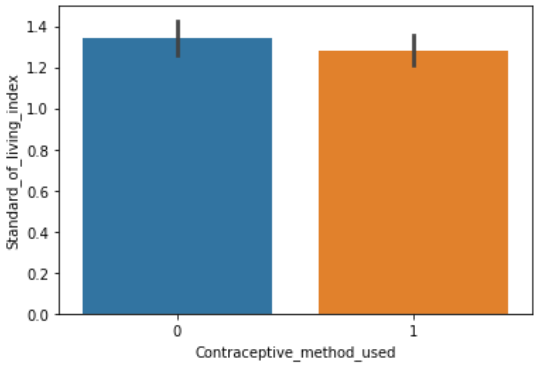
The dataset does not have normal distribution in any variable.

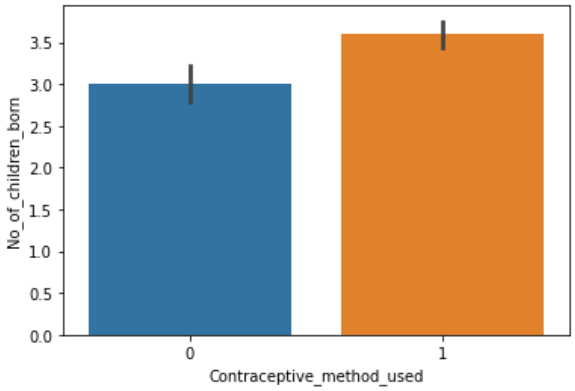
Outliers are observed in No of_children_born, wife_religion, Media_exposure.

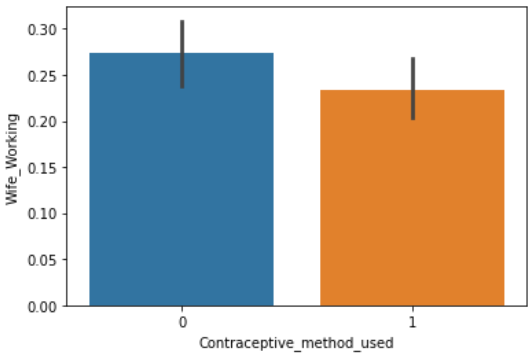
MULTIVARIATE ANALYSIS:

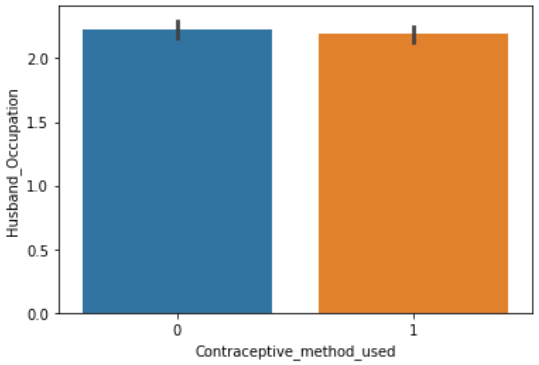












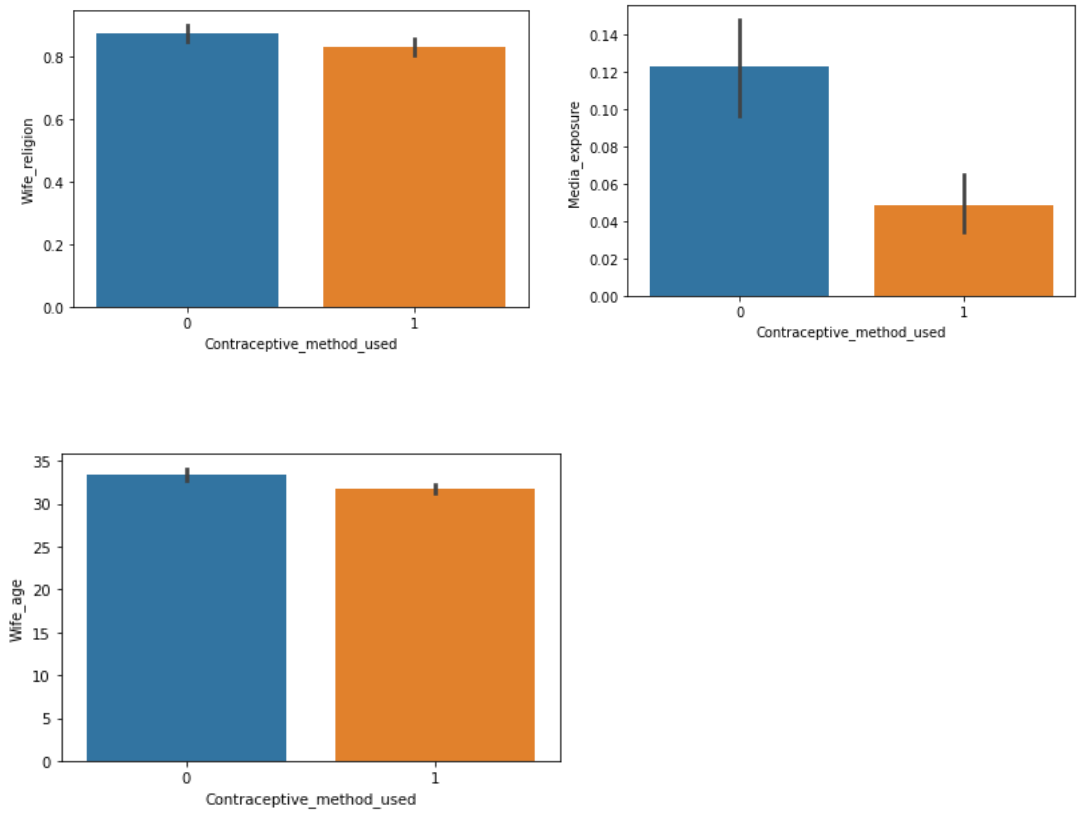


Fig. 7. Multivariate analysis

From the above graphs we can say that:

3. Age group of wife with contraceptive method used(Y/N) is between 30-35. Not much difference was observed in both the cases.
4.

BIVARIATE ANALYSIS:

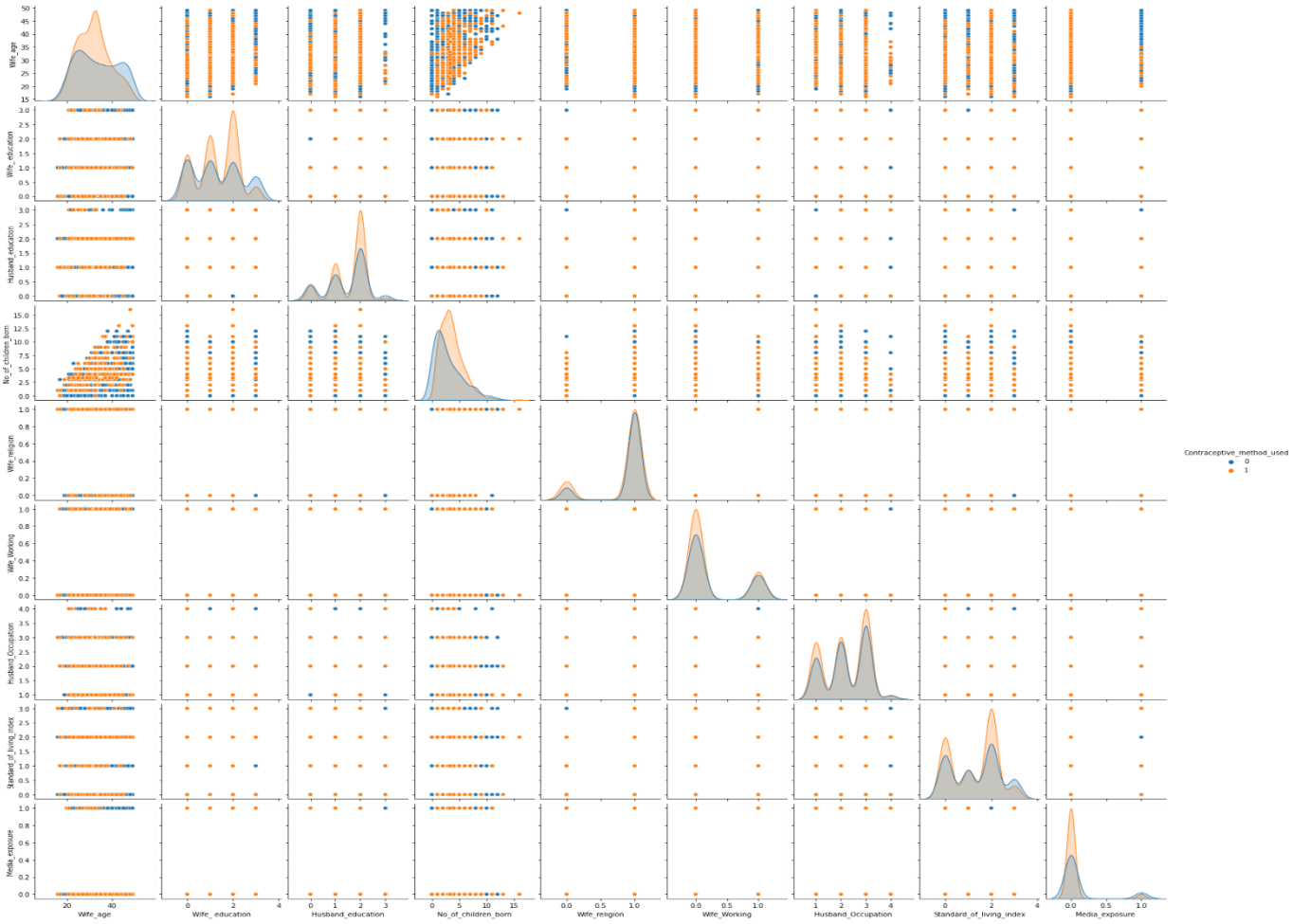


Fig. 8. Pair plot

6. Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis) and CART.

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

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

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

On applying multiple models we found the following results:

LOGISTIC REGRESSION:

Accuracy observed: ~61%.

	precision	recall	f1-score	support
0	0.64	0.41	0.50	187
1	0.60	0.80	0.69	210
accuracy			0.61	397
macro avg	0.62	0.60	0.59	397
weighted avg	0.62	0.61	0.60	397

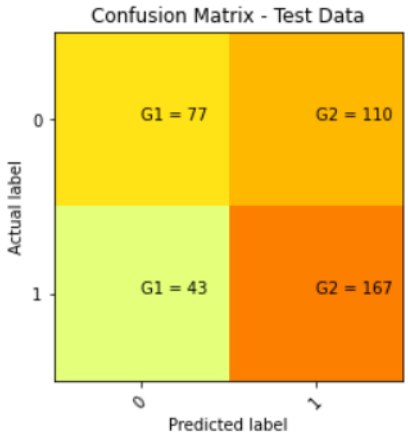


Fig. 9. Confusion matrix for Logistic Regression

LDA:

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

Table 11. Classification of LDA model results for Train data

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

Table 12. Classification of LDA model results for Test data

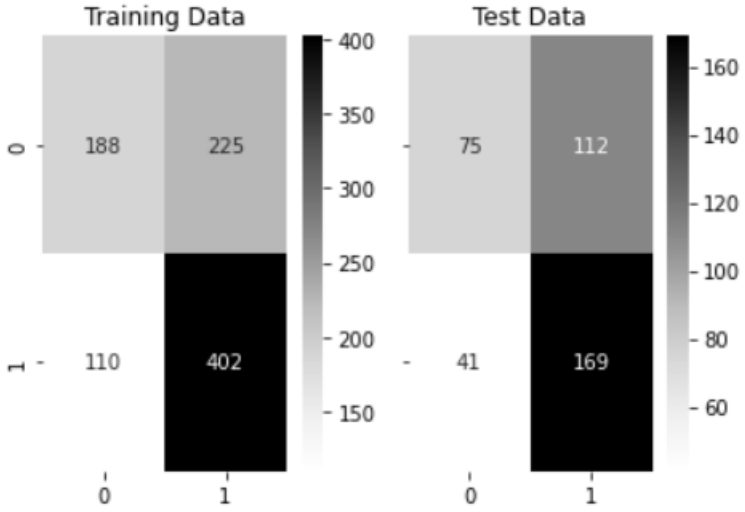


Table 13. Confusion matrix for LDA

CART:

	Imp
Wife_age	0.307761
Wife_ education	0.092646
Husband_education	0.073109
No_of_children_born	0.261894
Wife_religion	0.035091
Wife_Working	0.023390
Husband_Occupation	0.087982
Standard_of_living_index	0.099668
Media_exposure	0.018459

After Pruning:

	Imp
Wife_age	0.307761
Wife_ education	0.092646
Husband_education	0.073109
No_of_children_born	0.261894
Wife_religion	0.035091
Wife_Working	0.023390
Husband_Occupation	0.087982
Standard_of_living_index	0.099668
Media_exposure	0.018459

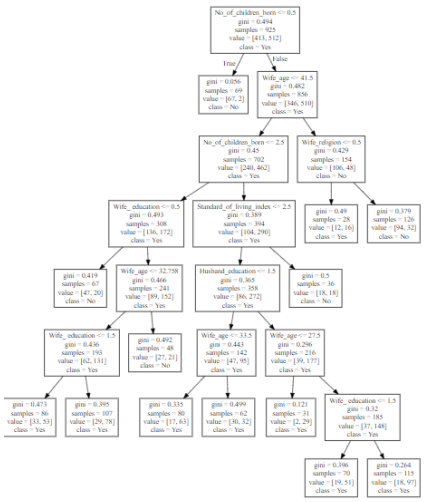
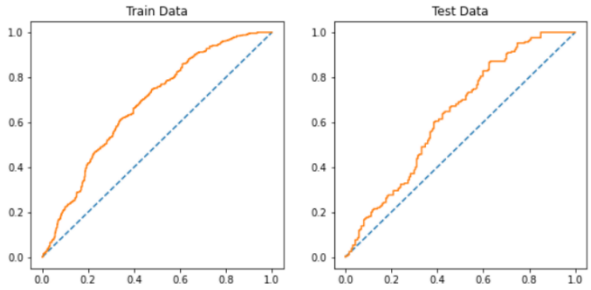


Fig. 10. Sample decision tree diagram.

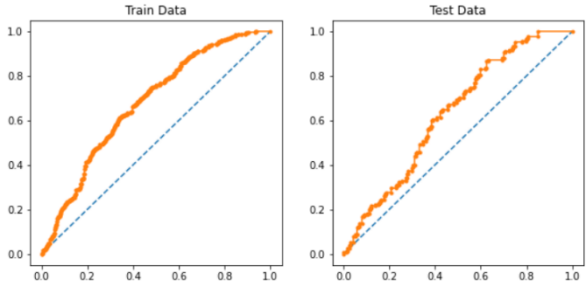
7. **Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.**

AUC & ROC- LOGISTIC REGRESSION:



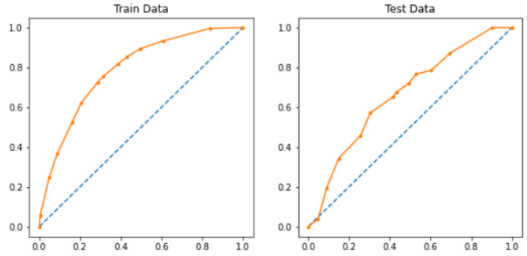

```
AUC LOGISTIC TRAIN DATA: 0.678
AUC LOGISTIC TEST DATA: 0.632
```

AUC & ROC: LDA:




```
AUC CART TRAIN DATA: 0.678
AUC CART TEST DATA: 0.632
```

AUC & ROC: CART:




```
AUC CART TRAIN DATA: 0.788
AUC CART TEST DATA: 0.662
```

On observing the values and graphs given above,

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

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

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

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

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

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