PREDICTIVE MODELLING

Linear Regression/Logistic Regression/LDA/CART

NOVEMBER 05, 2023

SHUBHAM KUMAR
Batch: G2- SAT 3PM

Contents

Problem 1

Problem 2

1.1 Read the data and do exploratory data analysis. Describe the data briefly. (Check the Data types, shape, EDA, 5-point summary). Perform Univariate, Bivariate Analysis, Multivariate Analysis.

Solution: -

First Five rows of the data

	0	1	2	3	4
lread	1	0	15	0	5
lwrite	0	0	3	0	1
scall	2147	170	2162	160	330
sread	79	18	159	12	39
swrite	68	21	119	16	38
fork	0.2	0.2	2.0	0.2	0.4
exec	0.2	0.2	2.4	0.2	0.4
rchar	40671.0	448.0	NaN	NaN	NaN
wchar	53995.0	8385.0	31950.0	8670.0	12185.0
pgout	0.0	0.0	0.0	0.0	0.0
ppgout	0.0	0.0	0.0	0.0	0.0
pgfree	0.0	0.0	0.0	0.0	0.0
pgscan	0.0	0.0	0.0	0.0	0.0
atch	0.0	0.0	1.2	0.0	0.0
pgin	1.6	0.0	6.0	0.2	1.0
ppgin	2.6	0.0	9.4	0.2	1.2
pflt	16.0	15.63	150.2	15.6	37.8
vflt	26.4	16.83	220.2	16.8	47.6
runqsz	CPU_Bound	Not_CPU_Bound	Not_CPU_Bound	Not_CPU_Bound	Not_CPU_Bound
freemem	4670	7278	702	7248	633
freeswap	1730946	1869002	1021237	1863704	1760253
usr	95	97	87	98	90

Table 1.1A
First five rows of the data

Last Five rows of the data

	8187	8188	8189	8190	8191
lread	16	4	16	32	2
lwrite	12	0	5	45	0
scall	3009	1596	3116	5180	985
sread	360	170	289	254	55
swrite	244	146	190	179	46
fork	1.6	2.4	0.6	1.2	1.6
exec	5.81	1.8	0.6	1.2	4.8
rchar	405250.0	89489.0	325948.0	62571.0	111111.0
wchar	85282.0	41764.0	52640.0	29505.0	22256.0
pgout	8.02	3.8	0.4	1.4	0.0
ppgout	20.64	4.8	0.6	1.6	0.0
pgfree	43.69	4.8	0.6	13.03	0.0
pgscan	55.11	0.2	0.0	18.04	0.0
atch	0.6	0.8	0.4	0.4	0.2
pgin	35.87	3.8	28.4	23.05	3.4
ppgin	47.9	4.4	45.2	24.25	6.2
pflt	139.28	122.4	60.2	93.19	91.8
vflt	270.74	212.6	219.8	202.81	110.0
runqsz	CPU_Bound	Not_CPU_Bound	Not_CPU_Bound	CPU_Bound	CPU_Bound
freemem	387	263	400	141	659
freeswap	986647	1055742	969106	1022458	1756514
usr	80	90	87	83	94

Table 1.1B
Last five rows of the data

Data Types of all the features in a dataset

```
<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
```

Table 1.1C
Datatypes in a Dataset

From the above Table 1.1C

1. Float64: - There are 13 float data types.

   (fork, exec, rchar, wchar, pgout, ppgout, pgfree, pgscan, atch, pgin, ppgin, pflt, vflt)

2. Int64: - There are 8 integer data types.

   (lread, lwrite, scall, sread, swrite, freemen, freeswap, usr)

3. Object: - There are 1 object datatypes

   (runqsz)

Shape of Dataset

```
Number of rows in a dataset =  8192
Number of column in a dataset =  22
```

Fig. 1.1A
Shape of Dataset

Descriptive Statistic of Dataset

	count	mean	std	min	25%	50%	75%	max
lread	8192.0	19.560	53.354	0.0	2.0	7.0	20.000	1845.00
lwrite	8192.0	13.106	29.892	0.0	0.0	1.0	10.000	575.00
scall	8192.0	2306.318	1633.617	109.0	1012.0	2051.5	3317.250	12493.00
sread	8192.0	210.480	198.980	6.0	86.0	166.0	279.000	5318.00
swrite	8192.0	150.058	160.479	7.0	63.0	117.0	185.000	5456.00
fork	8192.0	1.885	2.479	0.0	0.4	0.8	2.200	20.12
exec	8192.0	2.792	5.212	0.0	0.2	1.2	2.800	59.56
rchar	8088.0	197385.728	239837.494	278.0	34091.5	125473.5	267828.750	2526649.00
wchar	8177.0	95902.993	140841.708	1498.0	22916.0	46619.0	106101.000	1801623.00
pgout	8192.0	2.285	5.307	0.0	0.0	0.0	2.400	81.44
ppgout	8192.0	5.977	15.215	0.0	0.0	0.0	4.200	184.20
pgfree	8192.0	11.920	32.364	0.0	0.0	0.0	5.000	523.00
pgscan	8192.0	21.527	71.141	0.0	0.0	0.0	0.000	1237.00
atch	8192.0	1.128	5.708	0.0	0.0	0.0	0.600	211.58
pgin	8192.0	8.278	13.875	0.0	0.6	2.8	9.765	141.20
ppgin	8192.0	12.389	22.281	0.0	0.6	3.8	13.800	292.61
pflt	8192.0	109.794	114.419	0.0	25.0	63.8	159.600	899.80
vflt	8192.0	185.316	191.001	0.2	45.4	120.4	251.800	1365.00
freemem	8192.0	1763.456	2482.105	55.0	231.0	579.0	2002.250	12027.00
freeswap	8192.0	1328125.960	422019.427	2.0	1042623.5	1289289.5	1730379.500	2243187.00
usr	8192.0	83.969	18.402	0.0	81.0	89.0	94.000	99.00

Table 1.1D

Descriptive stats of the Dataset

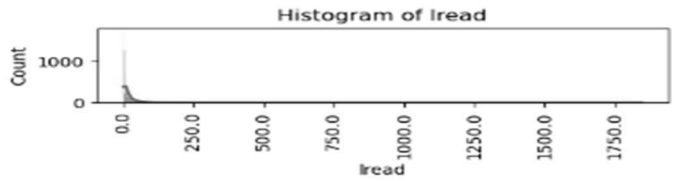
From Table 1.1D we can see the Descriptive stats or 5-point summary of the dataset. The 5-point summary includes min, 25%, 50%, 75% and max. It also includes mean and standard deviation.

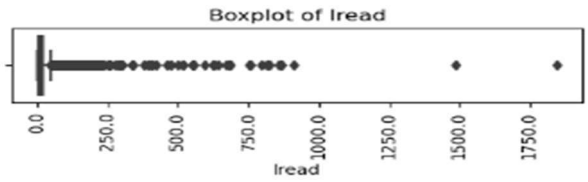
We can see that there is large difference between 75% and maximum which clearly indicates that there are outliers present in our dataset

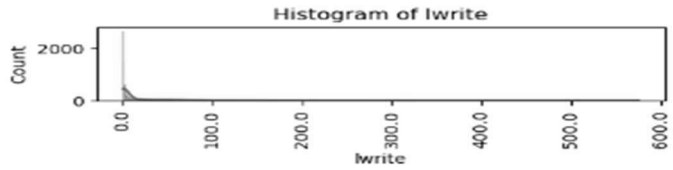
There are some attributes (pgout, ppgout, pgfree, atch) which has 0's as more than 50% of the dataset. Also there is an attribute (pgscan) that has 75% of data as 0's. We need to check these attributes.

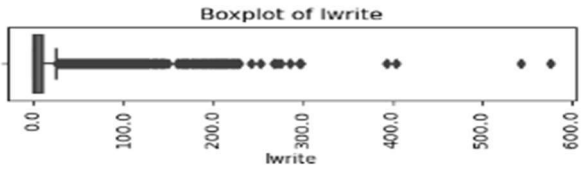
Let's Perform Univariate Analysis of our Dataset

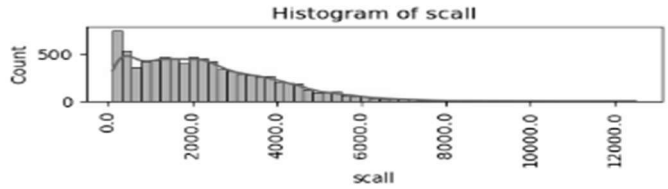
1. Univariate Analysis of Numeric columns

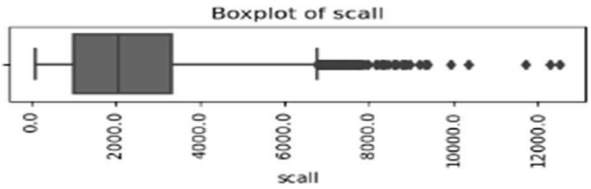


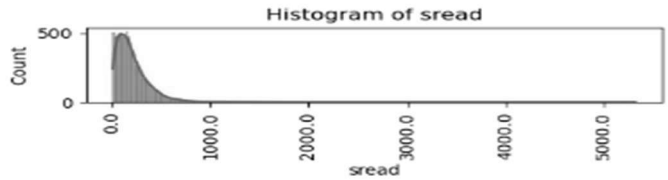


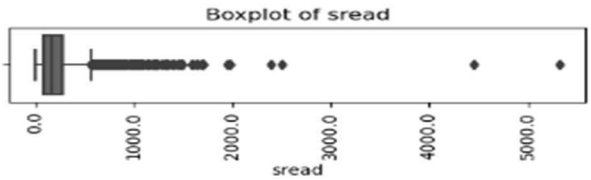


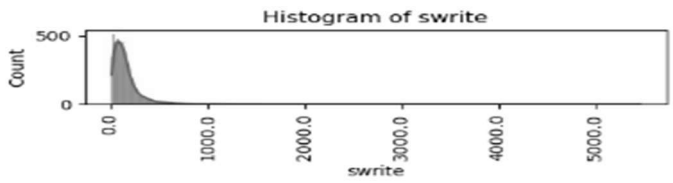


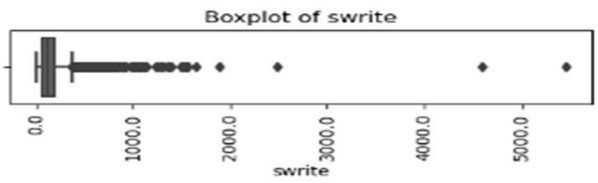


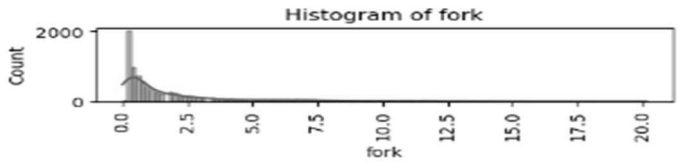


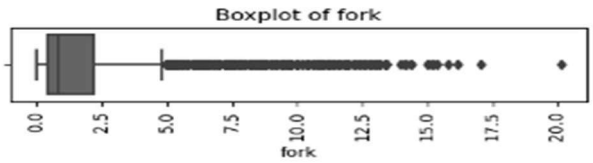


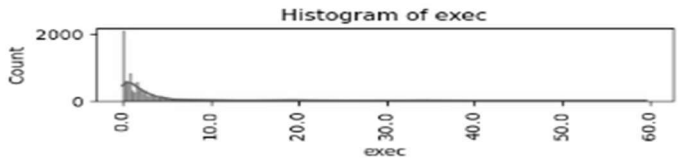


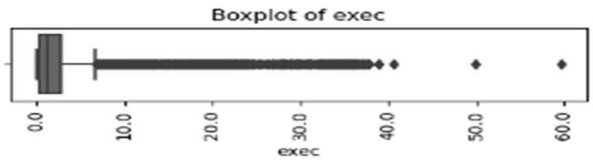


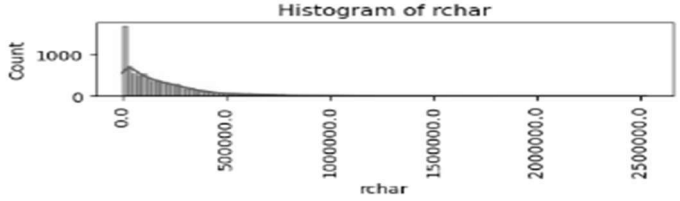


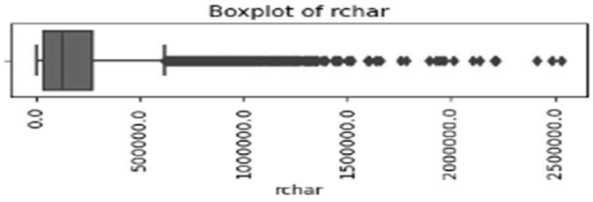


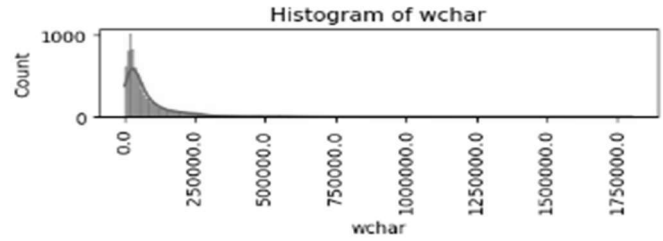


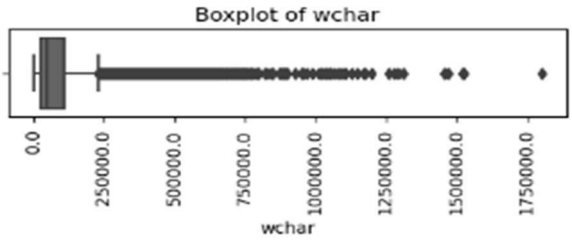


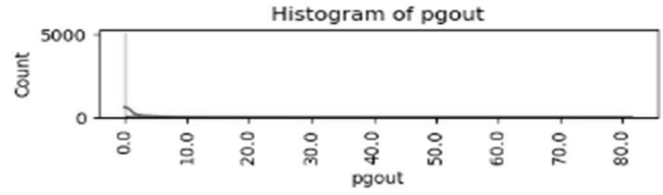


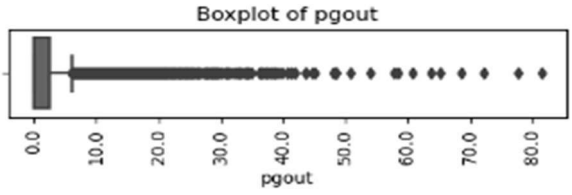


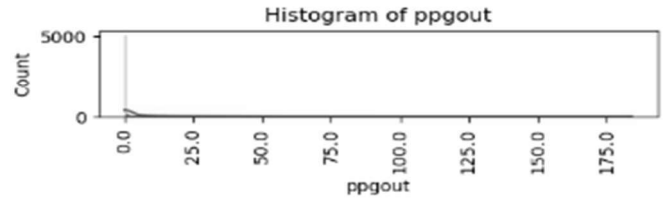


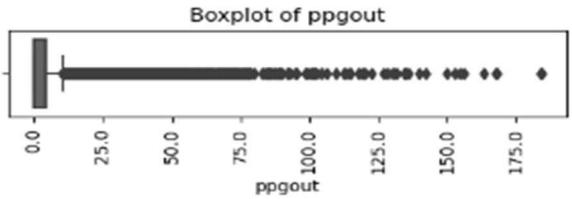

Histogram of wchar
Count
wchar

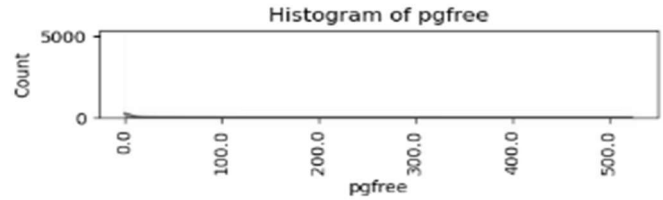

Boxplot of wchar
wchar

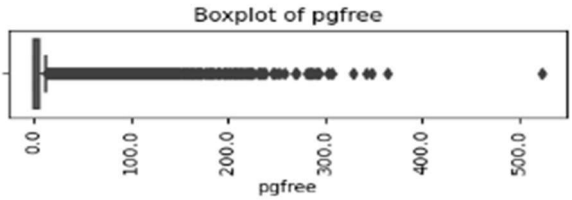

Histogram of pgout
Count
pgout

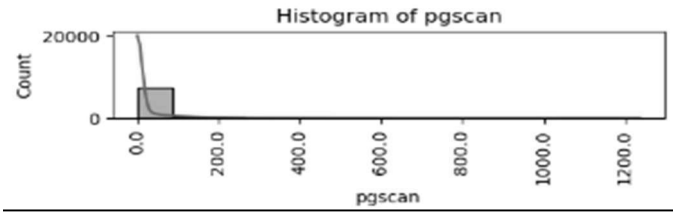

Boxplot of pgout
pgout

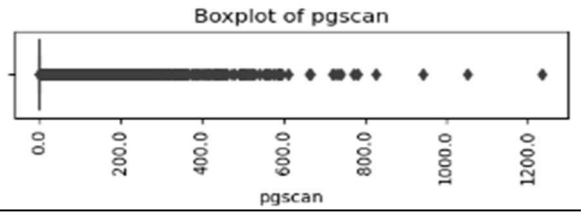

Histogram of ppgout
Count
ppgout

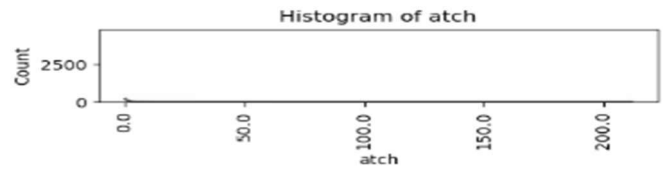

Boxplot of ppgout
ppgout

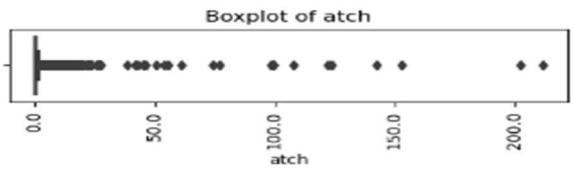

Histogram of pgfree
Count
pgfree

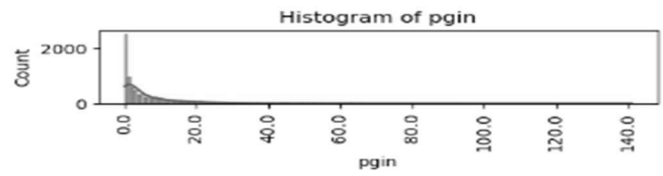

Boxplot of pgfree
pgfree

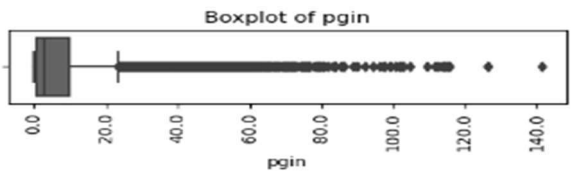

Histogram of pgscan
Count
pgscan

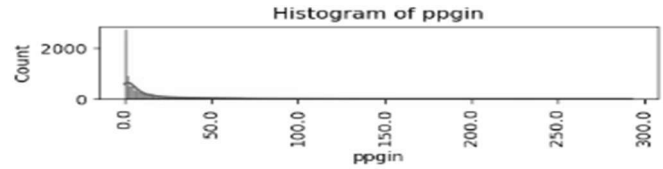

Boxplot of pgscan
pgscan

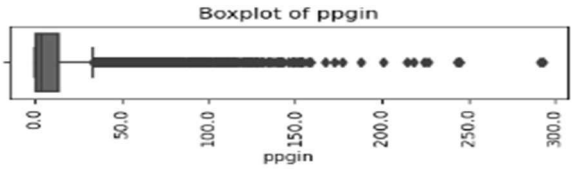

Histogram of atch
Count
atch


Boxplot of atch
atch


Histogram of pgin
Count
pgin


Boxplot of pgin
pgin


Histogram of ppgin
Count
ppgin


Boxplot of ppgin
ppgin

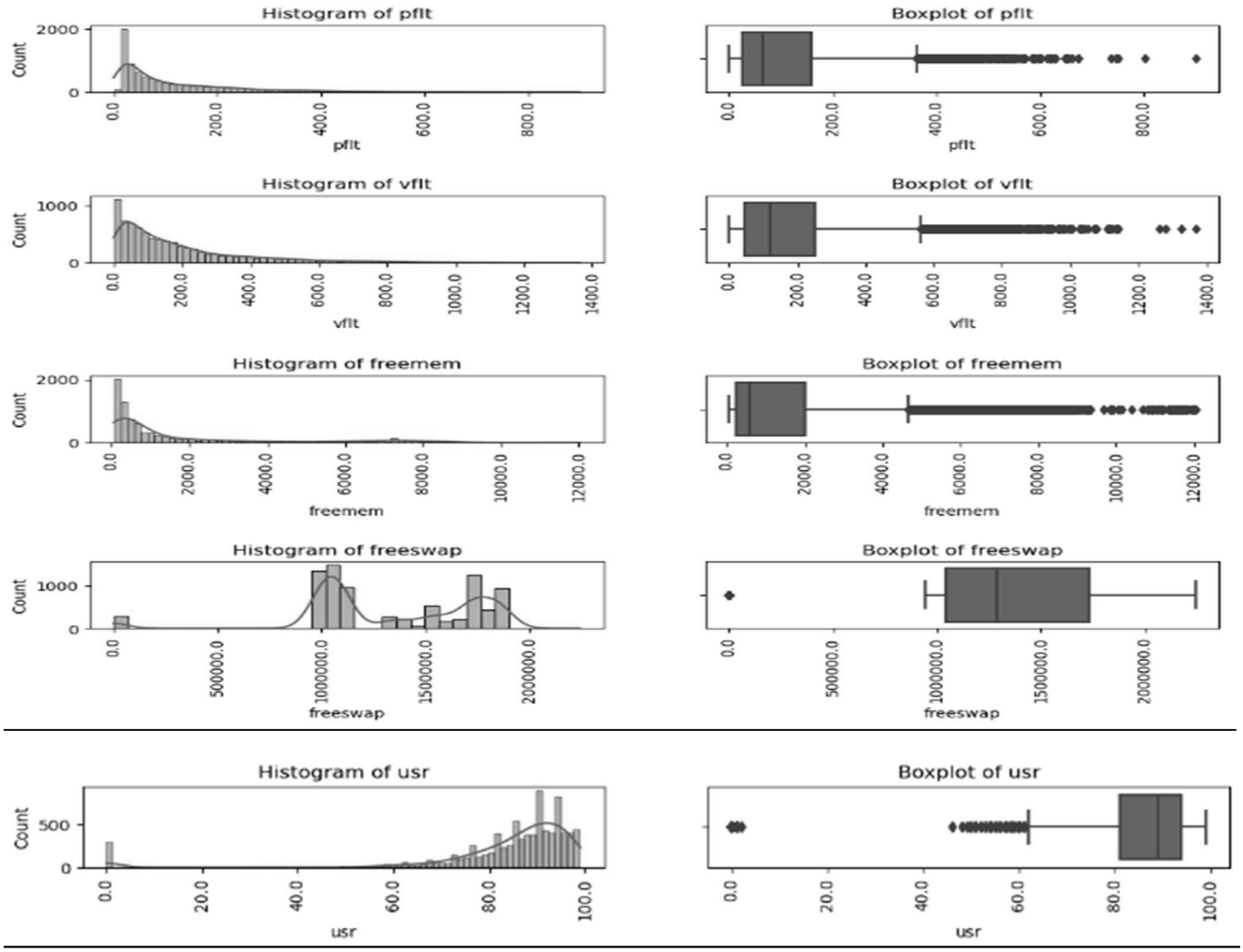


Fig 1.1B
Univariate Analysis of Numeric Columns (Histogram & Boxplot)

Inferences from above Fig 1.1B,

- We have plotted the Histogram and Boxplot for our Numeric dtypes Attributes.
- We can see all the numeric columns has outliers present in it which needs to be treated in further process
- All the columns except ‘usr’ & ‘freeswap’ are right skewed
- ‘usr’ and ‘freeswap’ are left skewed**.**

2. Univariate Analysis of Categorical Data

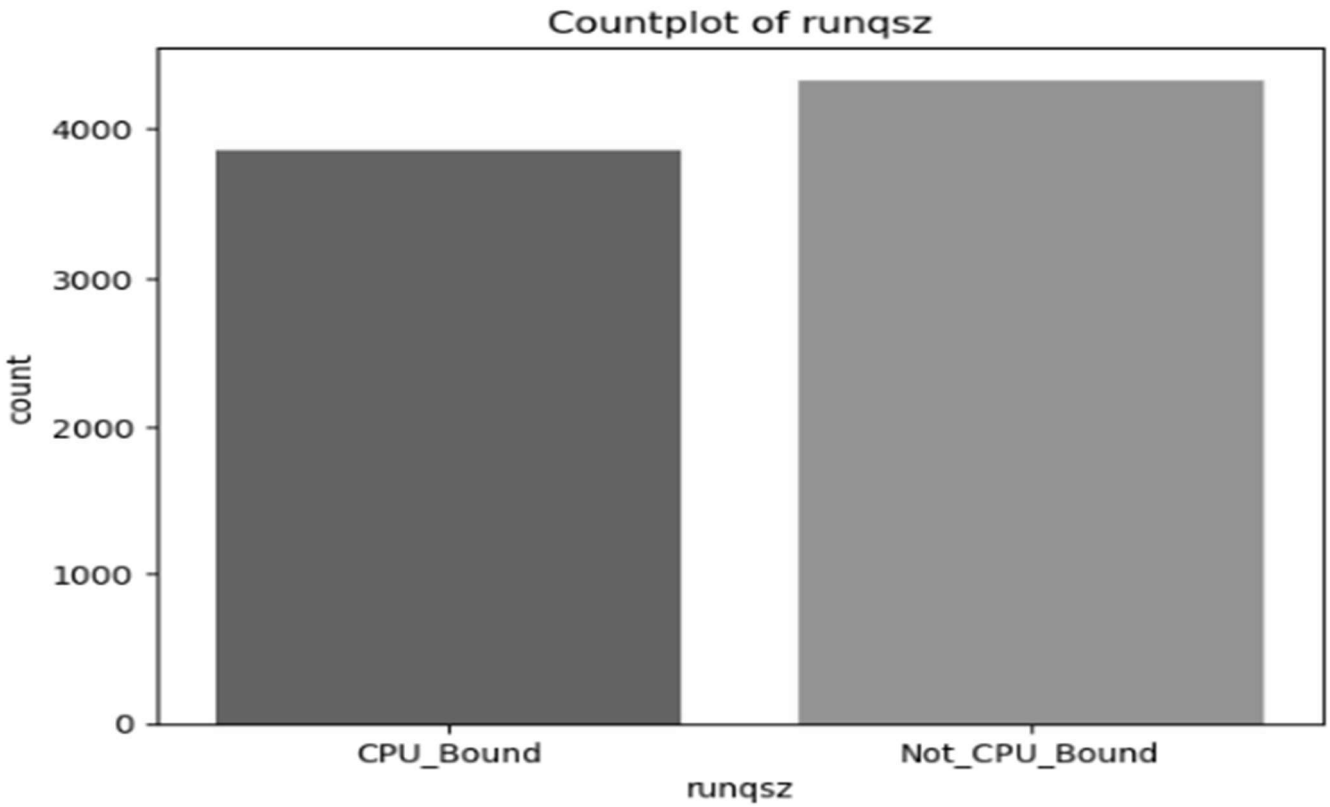


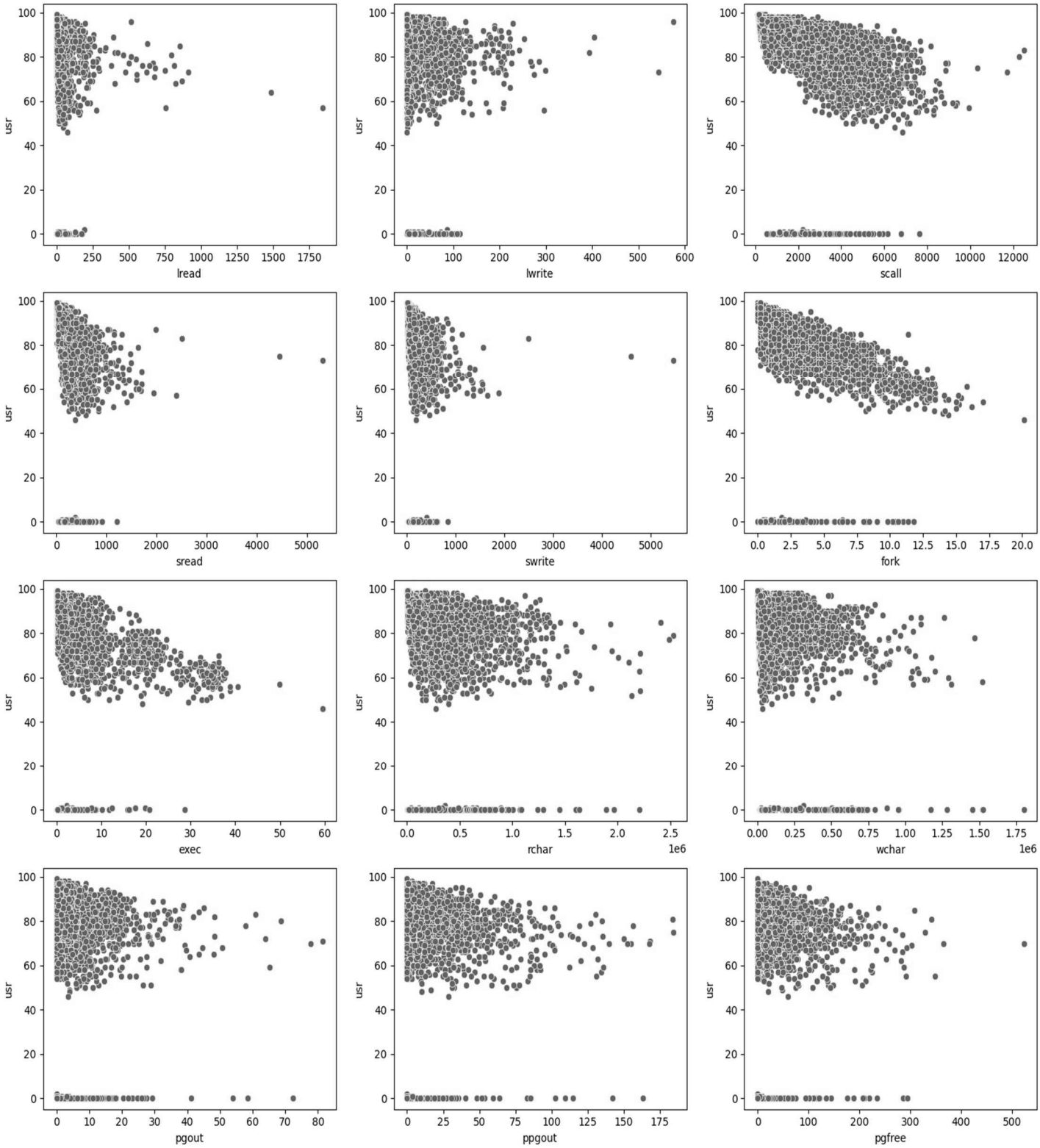
Fig. 1.1C
Univariate Analysis of Categorical Column (Countplot)

Inferences from above Fig 1.1C

- There are two categories in 'runqsz' column i.e., CPU_Bound & Not_CPU_Bound
- There are more number of system which are Not_CPU_Bound than system which are CPU_Bound
- Number of system which are CPU_Bound = 3861
- Number of system which are Not_CPU_Bound = 4331

Bi-Variate & Multi-Variate Analysis of Dataset

1. Bi-Variate analysis between independent variable and target variable (usr)



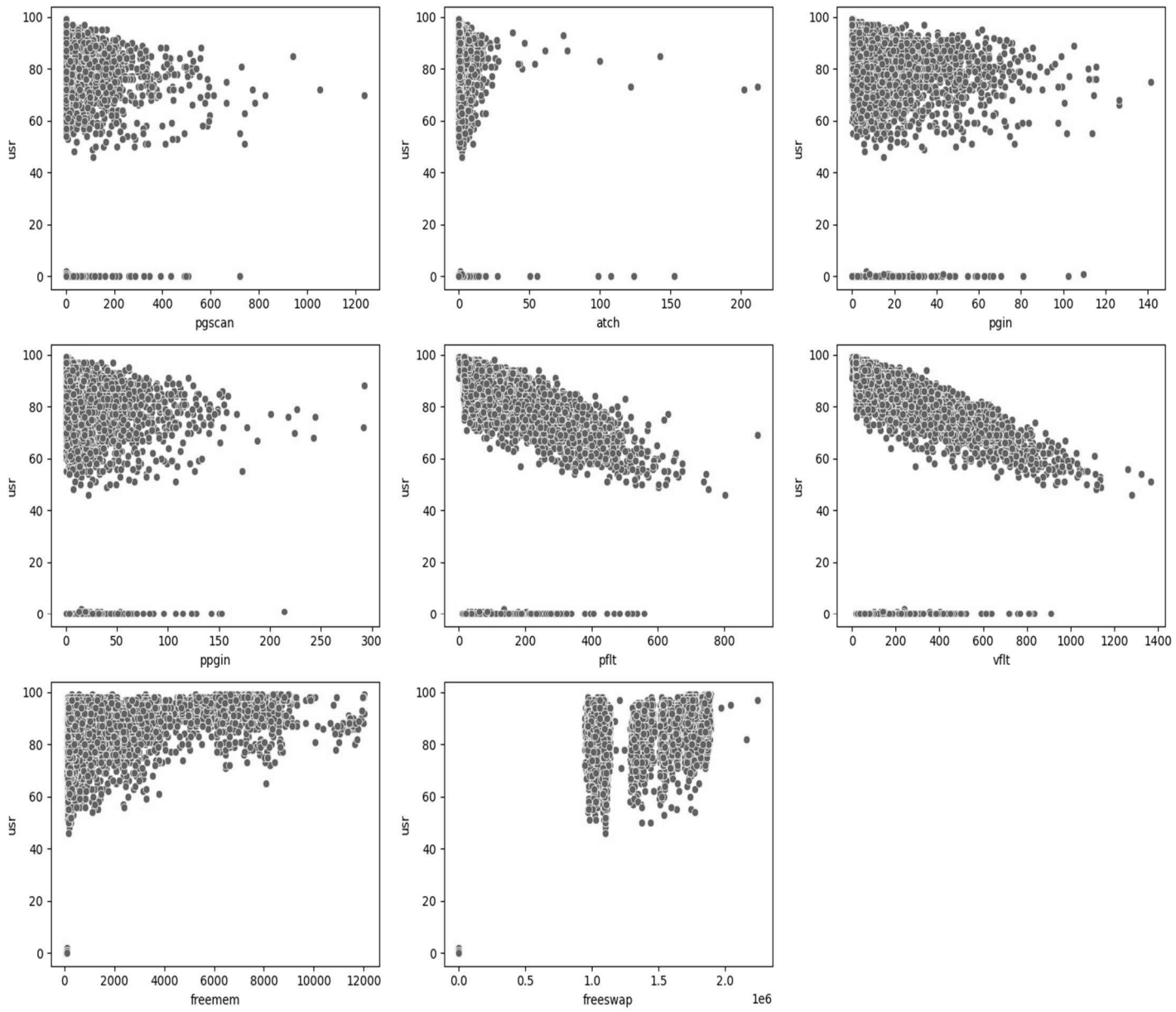


Fig. 1.1D
Bi-variate analysis between Independent & Target Variable

Inferences from above Fig. 1.1D

- We can see for all the independent variables the target variable 'usr' is cluttered between 40 to 100. Also there are few points which are cluttered at 0.
- As 'usr' increases 'lread' ,'lwrite' , 'sread' , 'swrite' , 'atach' doesn't increases significantly.
- As 'usr' increases 'scall' , 'fork' , 'pflt' , 'vflt' , 'exec' decreases significantly
- As 'usr' increases 'freewap' , 'freemen' , 'ppgin' , 'pgin' , 'pgfree' , 'ppgout' , 'pgout' , 'wchar' , 'rchar' increases

2. **Bi-Variate Analysis between Categorical columns and target column**

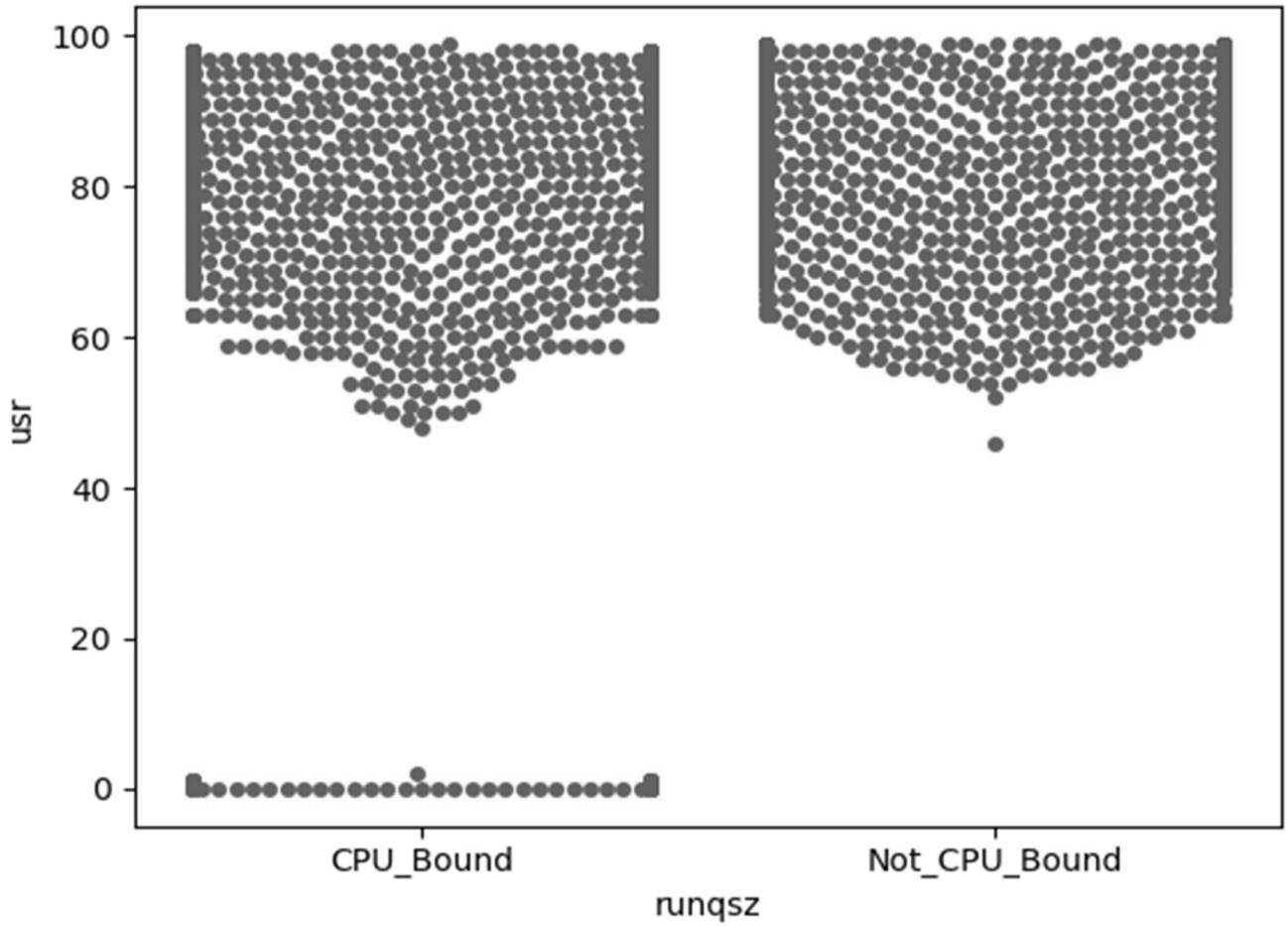


Fig. 1.1E
Swarm Plot between Categorical and Target column

Inferences from above Fig. 1.1E: -

- We can see that for both categories CPU_Bound and Not_CPU_Bound the 'usr' values are cluttered between 50 & 100
- For system which are CPU_Bound there are 'usr' values 0's also

Multi-Variate Analysis

1. Heatmap

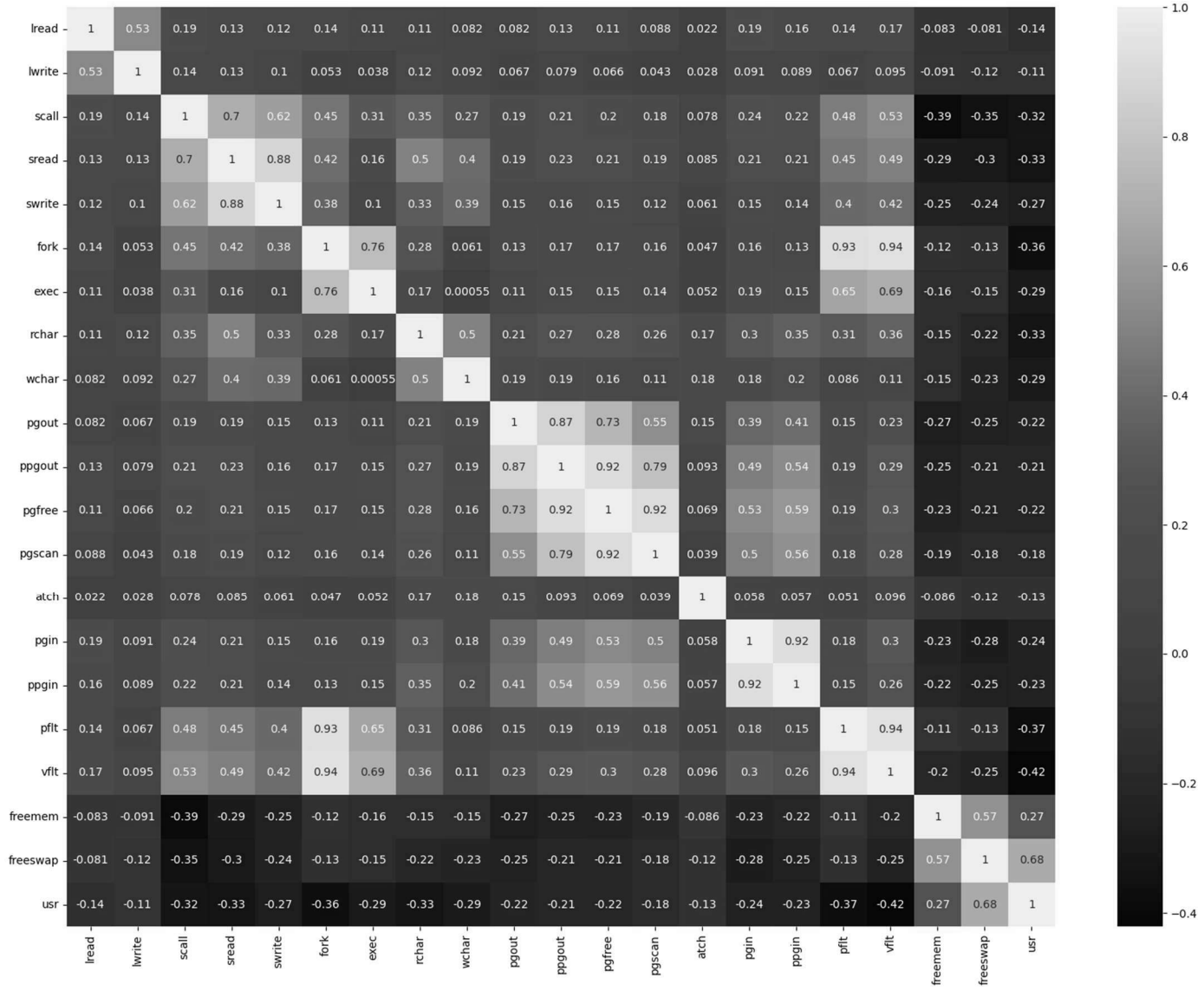


Fig. 1.1E
Multi-Variate Analysis (Heatmap)

Inferences from above Fig 1.1E

- Correlation value greater than 0.85 is considered as very strong correlation
- 'fork' has very strong correlation with 'vflt' , 'pflt' as its correlation value is greater than 0.85
- 'pgout' has strong correlation with 'ppgout'
- 'ppgout' has strong correlation with 'pgfree'
- 'pgfree' has strong correlation with 'pgscan'

2. Pair Plot

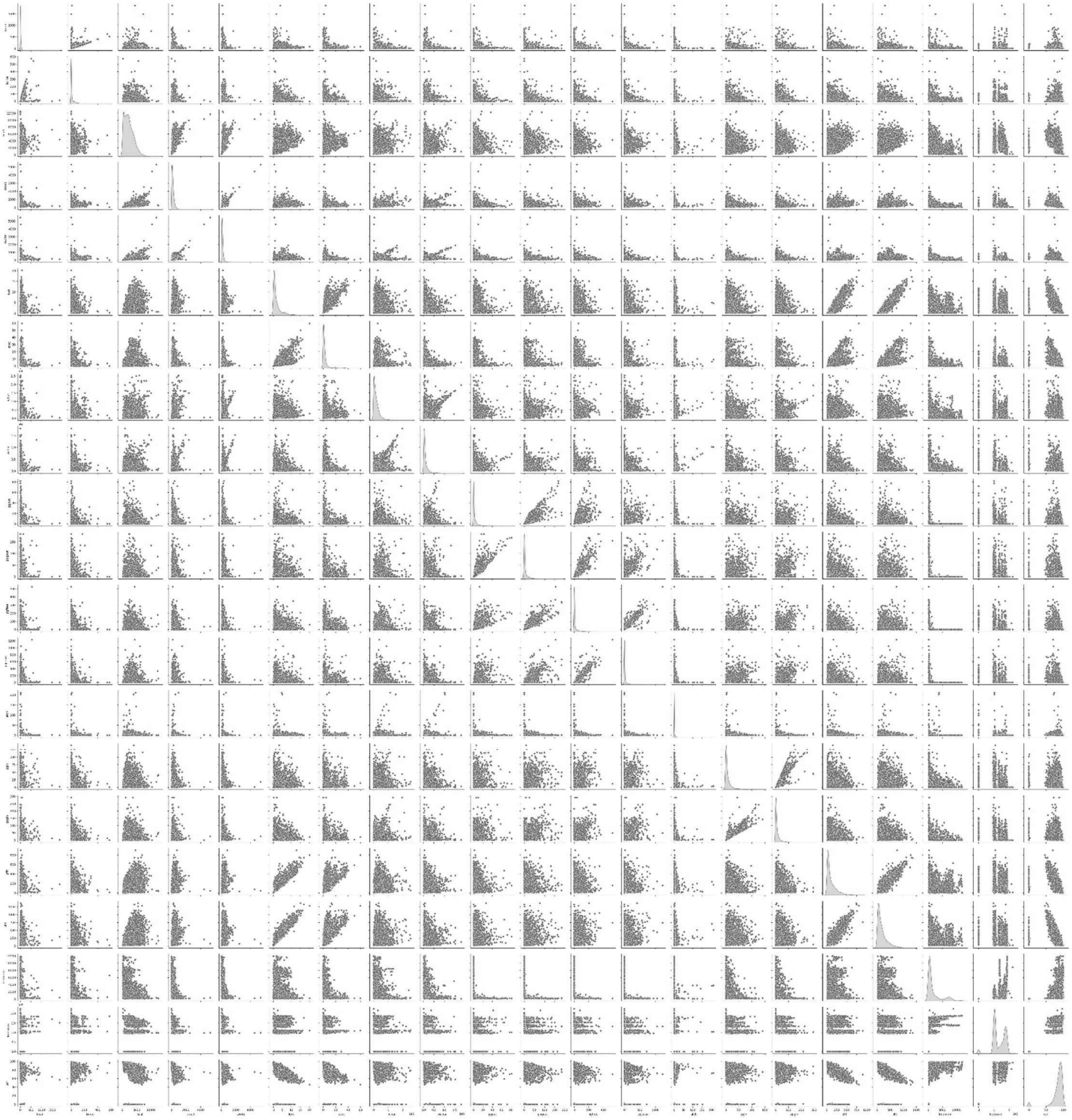


Fig. 1.1E
Multi-Variate Analysis (Pair plot)

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.

Solution: -

A. Null values in a dataset

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

Fig. 1.1A
Null Entries in a dataset (before imputation)

From above Fig.1.1A, we can see that there are null entries in two attributes: -

1. rchar = 104 null entries
2. wchar = 15 null entries

We need to fill these null values with either mean or median. Since it is a continuous variable we have imputed these null values with median.

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

Fig. 1.1B
Null Entries in a dataset (after imputation)

From above Fig., 1.1B, we can see there are no null values present in our dataset.

B. Check for the values which are equal to zero

```
Number of 0 in lread =  675
Number of 0 in lwrite =  2684
Number of 0 in scall =  0
Number of 0 in sread =  0
Number of 0 in swrite =  0
Number of 0 in fork =  21
Number of 0 in exec =  21
Number of 0 in rchar =  0
Number of 0 in wchar =  0
Number of 0 in pgout =  4878
Number of 0 in ppgout =  4878
Number of 0 in pgfree =  4869
Number of 0 in pgscan =  6448
Number of 0 in atch =  4575
Number of 0 in pgin =  1220
Number of 0 in ppgin =  1220
Number of 0 in pflt =  3
Number of 0 in vflt =  0
Number of 0 in runqsz =  0
Number of 0 in freemem =  0
Number of 0 in freeswap =  0
Number of 0 in usr =  283
```

Fig 1.1C
Number of 0's in a column

```
Percentage of 0 in lread =  8.24
Percentage of 0 in lwrite =  32.76
Percentage of 0 in scall =  0.0
Percentage of 0 in sread =  0.0
Percentage of 0 in swrite =  0.0
Percentage of 0 in fork =  0.26
Percentage of 0 in exec =  0.26
Percentage of 0 in rchar =  0.0
Percentage of 0 in wchar =  0.0
Percentage of 0 in pgout =  59.55
Percentage of 0 in ppgout =  59.55
Percentage of 0 in pgfree =  59.44
Percentage of 0 in pgscan =  78.71
Percentage of 0 in atch =  55.85
Percentage of 0 in pgin =  14.89
Percentage of 0 in ppgin =  14.89
Percentage of 0 in pflt =  0.04
Percentage of 0 in vflt =  0.0
Percentage of 0 in runqsz =  0.0
Percentage of 0 in freemem =  0.0
Percentage of 0 in freeswap =  0.0
Percentage of 0 in usr =  3.45
```

Fig1.1D
Percentage of 0's in a column

From above figures we can there are large number of 0's present in our dataset.

We can keep the 0's in our dataset for further analysis as we have some features in our dataset which can be 0 if the system stays Idle. Also 'pgscan' column has 78.71% of data points as 0.

C. Duplicated data in a dataset

```
Number of duplicated data in a dataset =  0
```

D. Check for Outliers

For Numeric Variables data outliers can be checked by plotting BoxPlot

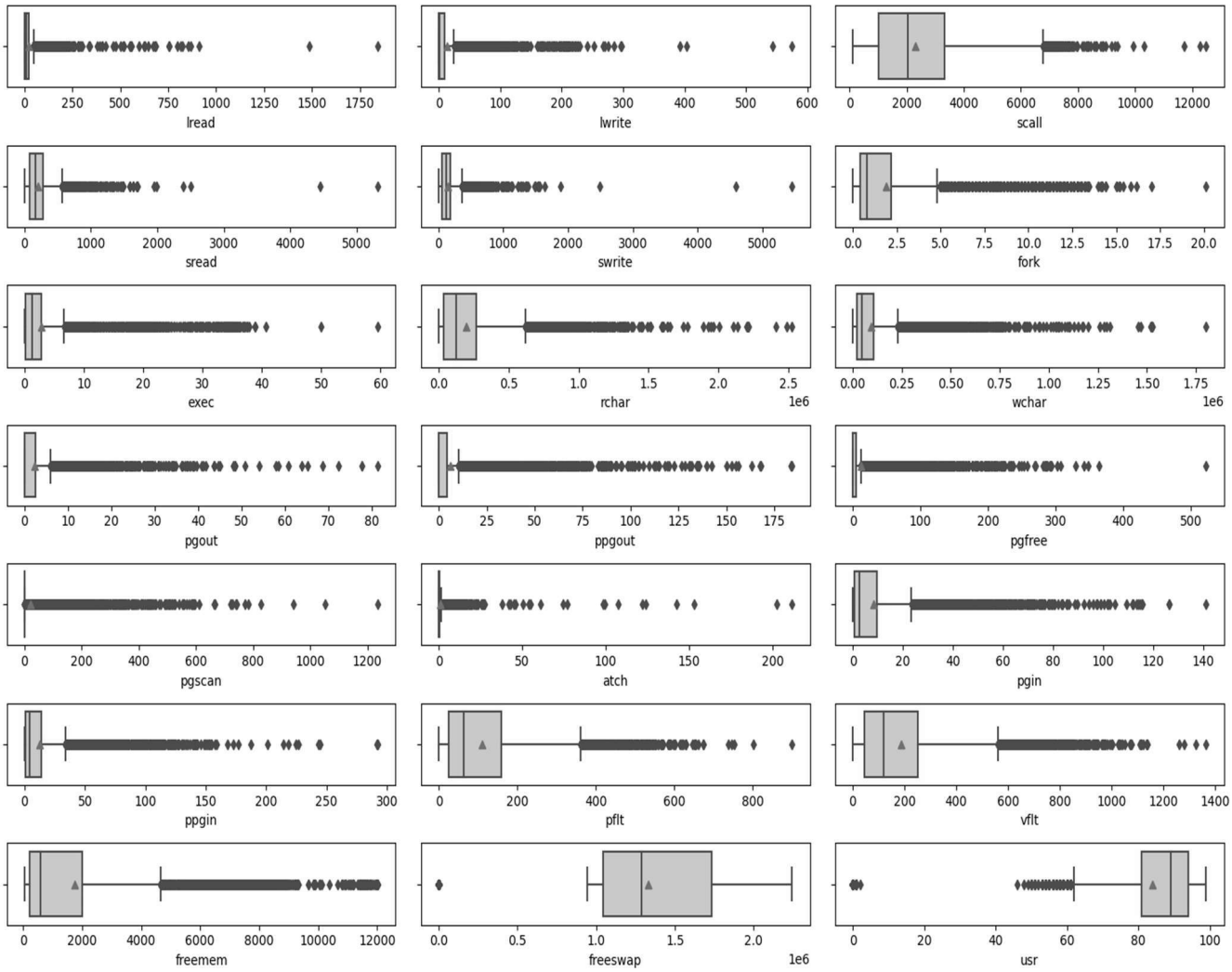


Fig 1.2A
Boxplot for Numeric Variables before Treating Outliers

From Fig 1.2A we can see that outliers are present in all the numeric columns. We need to treat these outliers.

For treating outliers, we have two options i.e.,

1. Inter Quartile Range (IQR) method
2. Z-score method (+ - 3rd standard deviation)

Inter Quartile Range is used when we have skewed distribution. In IQR method all the data points which are more than Upper Fence are brought to Upper Fence and all the data points which are less than Lowe Fence are brought to Lower Fence.

Q1 = 1st Quartile (25%)	Q3 = 3rd Quartile (75%)
IQR = Q3 – Q1	
Upper Fence = Q3 + 1.5 * IQR	Lower Fence = Q1 – 1.5 * IQR

Z-Score (+ - 3rd standard deviation) method is used when we have normal distribution. In Z-score method all the data points which are above 3rd Standard Deviation (upper side) and less than 3rd standard deviation (lower side) are treated as outlier.

Since all the numeric columns for the given dataset are either Right skewed or Left skewed. So for treating outliers of given dataset we will use IQR method.

In order to treat outliers of numeric columns we have created a user defined function remove_outliers . This user defined function takes input as column and return us its Upper Fence and Lower Fence.

We have created a user defined function **remove_outlier ( ).** This function will take input as numeric column of dataset and return us its Upper Limit and Lower Limit

```
Lower Fence and Upper Fence of lread  =  -25.0 & 47.0
Lower Fence and Upper Fence of lwrite  =  -15.0 & 25.0
Lower Fence and Upper Fence of scall  =  -2445.875 & 6775.125
Lower Fence and Upper Fence of sread  =  -203.5 & 568.5
Lower Fence and Upper Fence of swrite  =  -120.0 & 368.0
Lower Fence and Upper Fence of fork  =  -2.3 & 4.9
Lower Fence and Upper Fence of exec  =  -3.7 & 6.7
Lower Fence and Upper Fence of rchar  =  -310940.875 & 611196.125
Lower Fence and Upper Fence of wchar  =  -101611.125 & 230625.875
Lower Fence and Upper Fence of pgout  =  -3.6 & 6.0
Lower Fence and Upper Fence of ppgout  =  -6.3 & 10.5
Lower Fence and Upper Fence of pgfree  =  -7.5 & 12.5
Lower Fence and Upper Fence of pgscan  =  0.0 & 0.0
Lower Fence and Upper Fence of atch  =  -0.9 & 1.5
Lower Fence and Upper Fence of pgin  =  -13.148 & 23.513
Lower Fence and Upper Fence of ppgin  =  -19.2 & 33.6
Lower Fence and Upper Fence of pflt  =  -176.9 & 361.5
Lower Fence and Upper Fence of vflt  =  -264.2 & 561.4
Lower Fence and Upper Fence of freemem  =  -2425.875 & 4659.125
Lower Fence and Upper Fence of freeswap  =  10989.5 & 2762013.5
Lower Fence and Upper Fence of usr  =  61.5 & 113.5
```

Fig. 1.2B
Upper Fence and Lower Fence of all the Numeric Columns

For treating outliers we have used np.where ( ) function which replace all the values which are greater than UL will be replaced with upper limit and values less than LL will be replaced with lower limit.

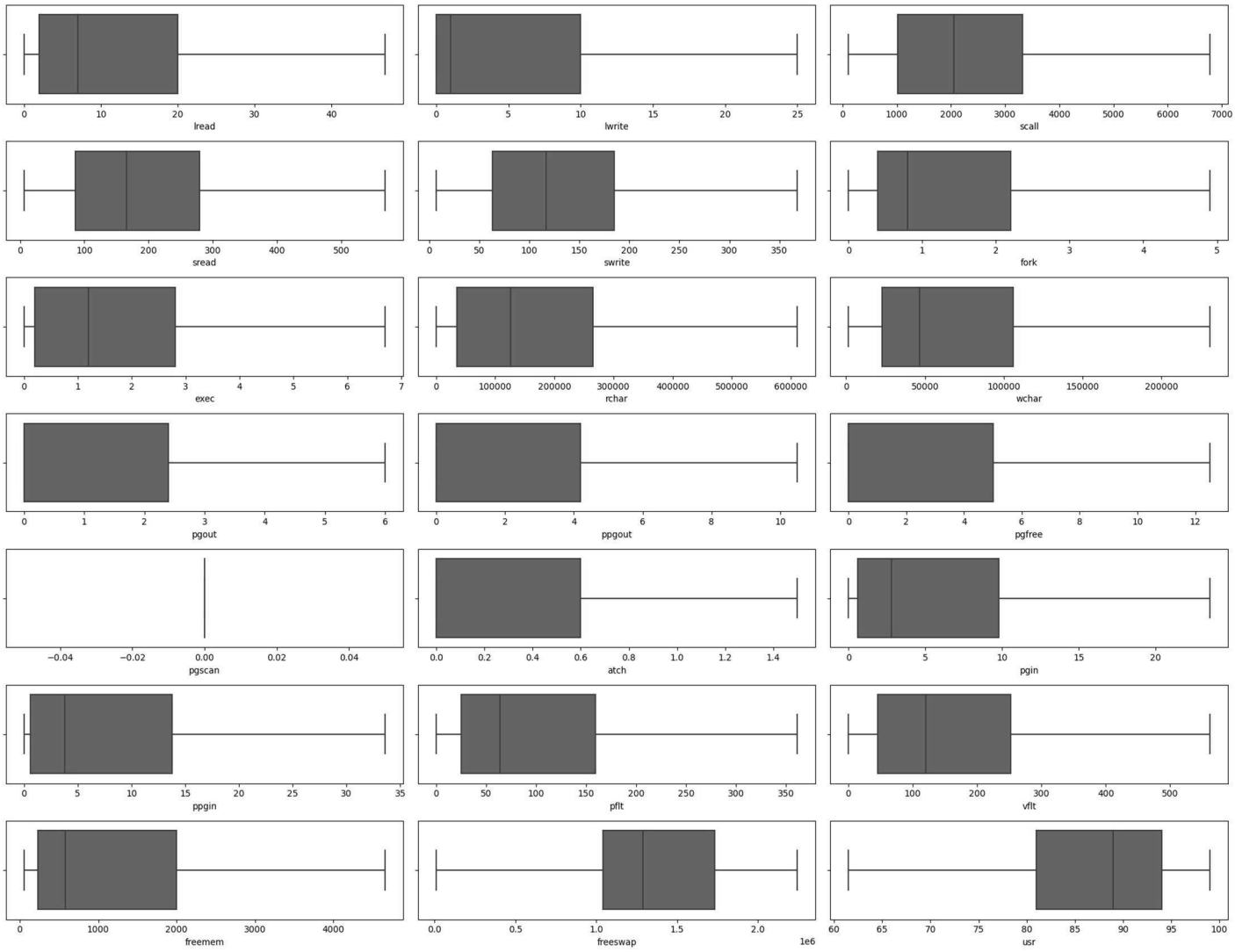


Fig. 1.2C

Boxplot of Numeric columns after Treating Outliers

Inferences from Fig 1.2C

- We can see all the outliers have been treated
- For 'pgscan' column we can see all the data has become zero which is because in 'pgscan' column more than 75% of data points are 0's

For categorical column

There are two categories in our 'runqsz' columns.

```
Not_CPU_Bound     4331
CPU_Bound         3861
Name: runqsz, dtype: int64
```

For converting categorical column, we have used dummies function from pandas and dropped the first dummy.

Sample dataset after converting categorical column into numeric

	0	1	2	3	4
lread	1.0	0.00	15.0	0.0	5.0
lwrite	0.0	0.00	3.0	0.0	1.0
scall	2147.0	170.00	2162.0	160.0	330.0
sread	79.0	18.00	159.0	12.0	39.0
swrite	68.0	21.00	119.0	16.0	38.0
fork	0.2	0.20	2.0	0.2	0.4
exec	0.2	0.20	2.4	0.2	0.4
rchar	40671.0	448.00	125473.5	125473.5	125473.5
wchar	40671.0	448.00	125473.5	125473.5	125473.5
pgout	0.0	0.00	0.0	0.0	0.0
ppgout	0.0	0.00	0.0	0.0	0.0
pgfree	0.0	0.00	0.0	0.0	0.0
pgscan	0.0	0.00	0.0	0.0	0.0
atch	0.0	0.00	1.2	0.0	0.0
pgin	1.6	0.00	6.0	0.2	1.0
ppgin	2.6	0.00	9.4	0.2	1.2
pflt	16.0	15.63	150.2	15.6	37.8
vflt	26.4	16.83	220.2	16.8	47.6
freemem	4670.0	7278.00	702.0	7248.0	633.0
freeswap	1730946.0	1869002.00	1021237.0	1863704.0	1760253.0
usr	95.0	97.00	87.0	98.0	90.0
runqsz_Not_CPU_Bound	0.0	1.00	1.0	1.0	1.0

Table 1.2A
Sample dataset after converting categorical column into numeric

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

Solution: -

Let's separate our dataset as Independent variable (X) and target variable (Y)

After separating the dataset, we need to split our dataset into train and test in order to perform Linear Regression

For splitting of dataset we have imported **train_test_split** from **sklearn.model_selection**.

First we will conduct Linear Regression using scikit i.e., sklearn.linear_model import LinearRegression

We have created model using both dataset i.e.,

1. Dataset with outlier
2. Dataset without outliers

After looking into the model score and model performance. We will consider one having better result and proceed.

1. **<u>Results after creating Model using Dataset with Outliers.</u>**

   A. Coefficient of the linear model for each variables

Coefficient of the linear model(with outliers) for each variables: -

	Coefficient Values
lread	-0.020132
lwrite	0.003576
scall	0.001004
sread	0.000644
swrite	-0.005854
fork	-1.719612
exec	-0.101220
rchar	-0.000004
wchar	-0.000004
pgout	-0.180252
ppgout	0.083714
pgfree	-0.070620
pgscan	0.011557
atch	-0.103213
pgin	0.089589
ppgin	-0.061296
pflt	-0.043168
vflt	0.025220
freemem	-0.001645
freeswap	0.000033
runqsz_Not_CPU_Bound	7.999830

Fig 1.3A

Coefficient values of variables

   B. Intercept of the Linear model

Intercept/Constant for the linear model(with outliers) = 43.524224487037976

C. Model score and Performance (Root Mean Squared Error) for test and train data

```
Model score for Train dataset (with outliers) =  0.6376409985472127
Model score for Test dataset (with outliers) =  0.6295035442082677
Root Mean Squared Error of Train dataset(with outliers) =  10.891625247561016
Root Mean Squared Error of Test dataset (with outliers) =  11.620918673000842
```

Fig. 1.3B

Model score and Performance of train and test data (with outliers)

2. <u>Results after creating Model using Dataset without Outliers.</u>

A. Coefficient of the linear model for each variables

Coefficient of the linear model(without outliers) for each variables:-

	Coefficient Values
lread	-0.0544
lwrite	0.0499
scall	-0.0006
sread	0.0015
swrite	-0.0058
fork	-0.0306
exec	-0.3024
rchar	-0.0000
wchar	-0.0000
pgout	-0.4776
ppgout	0.0117
pgfree	0.0564
pgscan	-0.0000
atch	0.6480
pgin	-0.0003
ppgin	-0.0560
pflt	-0.0328
vflt	-0.0059
freemem	-0.0005
freeswap	0.0000
runqsz_Not_CPU_Bound	1.7584

Fig 1.3C

Coefficient values of variables

B. Intercept of the Linear model

```
Intercept/constant of linear model (without outliers) =  83.33997607746511
```

C. Model score and Performance (Root Mean Squared Error) for test and train data

```
Model score of Train dataset(without outliers) =  0.7883733273531547
Model score for Test dataset(without outliers) =  0.7876764031972878
Root Mean Squared Error of Train dataset(without outliers) =  4.483
Root Mean Squared Error of Test dataset(without outliers) =  4.493
```

Fig. 1.3C

Model score and Performance of train and test data (without outliers)

Comparison between model with Outliers and model without outliers		
	Model with Outliers	**Model Without outliers**
Model Score Train Dataset	0.6376	0.78837
Model Score Test Dataset	0.6295	0.78767
Model Performance/ RSME of Train Dataset	10.891	4.483
Model Performance/ RSME of Test Dataset	11.620	4.493

Table 1.3A

Comparison between model with and without Outliers

Inference from above table 1.3A

- Model score for Train data of model without outliers is greater than model with outliers
- Model score for Test data of model without outliers is greater than model with outliers
- We know that higher the model score better is our model, hence model without outliers are better than with outliers
- Model performance/RSME of train data of model without outlier is smaller than model with outliers
- Model performance/RSME of test data of model without outlier is smaller than model with outliers
- We know that lower the RSME score better the model performance, hence model without outlier is better than model with outliers

So, Final Linear Equation after creating model using sklearn without Outliers is

```
Equation:
 USR = 83.34 +( -0.054 )* lread +( 0.05 )* lwrite +( -0.001 )* scall +( 0.001 )* sread +( -0.006 )* swrite +( -0.031 )* fork +(
-0.302 )* exec +( -0.0 )* rchar +( -0.0 )* wchar +( -0.478 )* pgout +( 0.012 )* ppgout +( 0.056 )* pgfree +( -0.0 )* pgscan
( 0.648 )* atch +( -0.0 )* pgin +( -0.056 )* ppgin +( -0.033 )* pflt +( -0.006 )* vflt +( -0.0 )* freemem +( 0.0 )* freeswap +(
1.758 )* runqsz_Not_CPU_Bound +
```

 # Checking Multicollinearity using Variance Inflation Factor (VIF)

Variance Inflation Factor (VIF) is one of the methods to check if independent variables have correlation between them. If they are correlated, then it is not ideal for linear regression models as they inflate the standard errors which in turn affects the regression parameters. As a result, the regression model becomes non-reliable and lacks interpretability.

lread	9.121587
lwrite	6.277178
scall	8.982271
sread	18.305496
swrite	16.594722
fork	25.117476
exec	5.926613
rchar	4.248851
wchar	3.345895
pgout	16.132951
ppgout	40.657129
pgfree	22.454330
pgscan	NaN
atch	2.808673
pgin	22.886080
ppgin	23.022093
pflt	23.215699
vflt	32.833721
freemem	3.405678
freeswap	7.106312
runqsz_Not_CPU_Bound	2.175886

Table 1.3B
VIF score of independent variables

- If VIF values are equal to 1, that means there is no multicollinearity
- If VIF values are more than 5 and less than 10, that means there is moderate multicollinearity
- If VIF values more than 10, that means there is high multicollinearity
- We can see there are some variable whose VIF value is greater than 10 which indicates there is a high multicollinearity
- For ‘atch’ we have VIF value = NaN as after treating outliers all the data points in that column has become 0.

Create model using Statsmodel (OLS)

For creating model using statsmodel we need to import statsmodel.api

Result of statsmodel(OLS)

Model No 01

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.788
Model:	OLS	Adj. R-squared:	0.788
Method:	Least Squares	F-statistic:	1064.
Date:	Thu, 02 Nov 2023	Prob (F-statistic):	0.00
Time:	20:42:38	Log-Likelihood:	-16740.
No. Observations:	5734	AIC:	3.352e+04
Df Residuals:	5713	BIC:	3.366e+04
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	83.3400	0.315	264.931	0.000	82.723	83.957
lread	-0.0544	0.009	-6.127	0.000	-0.072	-0.037
lwrite	0.0499	0.013	3.815	0.000	0.024	0.076
scall	-0.0006	6.45e-05	-9.884	0.000	-0.001	-0.001
sread	0.0015	0.001	1.426	0.154	-0.001	0.003
swrite	-0.0058	0.001	-4.008	0.000	-0.009	-0.003
fork	-0.0306	0.134	-0.228	0.819	-0.293	0.232
exec	-0.3024	0.052	-5.846	0.000	-0.404	-0.201
rchar	-5.297e-06	4.9e-07	-10.818	0.000	-6.26e-06	-4.34e-06
wchar	-5.062e-06	1.06e-06	-4.786	0.000	-7.14e-06	-2.99e-06
pgout	-0.4776	0.091	-5.259	0.000	-0.656	-0.300
ppgout	0.0117	0.079	0.148	0.882	-0.143	0.167
pgfree	0.0564	0.047	1.190	0.234	-0.037	0.149
pgscan	-2.955e-14	1.37e-16	-216.384	0.000	-2.98e-14	-2.93e-14
atch	0.6480	0.145	4.472	0.000	0.364	0.932
pgin	-0.0003	0.029	-0.010	0.992	-0.056	0.056
ppgin	-0.0560	0.020	-2.826	0.005	-0.095	-0.017
pflt	-0.0328	0.002	-16.839	0.000	-0.037	-0.029
vflt	-0.0059	0.001	-4.127	0.000	-0.009	-0.003
freemem	-0.0005	5.19e-05	-9.202	0.000	-0.001	-0.000
freeswap	9.181e-06	1.9e-07	48.236	0.000	8.81e-06	9.55e-06
runqsz_Not_CPU_Bound	1.7584	0.127	13.832	0.000	1.509	2.008

Omnibus:	1021.607	Durbin-Watson:	2.011
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2058.687
Skew:	-1.069	Prob(JB):	0.00
Kurtosis:	5.011	Cond. No.	2.61e+22

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.66e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Table 1.3C

Summary of Statsmodel

Inferences from Table 1.3C: -

- **Value of R2 and Adj. R2 are same i.e., 0.788**
- **Constant/Intercept of the model = 83.340**
- **Cond. No. tells us about the multicollinearity so higher the number, high is the chance of multicollinearity. Since Cond. No. = 2.61e+22 means there is high multicollinearity in dataset**
- **Let's calculate the VIF value and check for multicollinearity**

 # Calculate VIF values

```
VIF values of const 28.122
VIF values of lread 5.108
VIF values of lwrite 4.17
VIF values of scall 2.988
VIF values of sread 6.472
VIF values of swrite 5.568
VIF values of fork 12.992
VIF values of exec 3.133
VIF values of rchar 2.098
VIF values of wchar 1.596
VIF values of pgout 11.412
VIF values of ppgout 29.047
VIF values of pgfree 15.977
VIF values of pgscan nan
VIF values of atch 1.91
VIF values of pgin 13.65
VIF values of ppgin 13.855
VIF values of pflt 11.162
VIF values of vflt 15.348
VIF values of freemem 1.963
VIF values of freeswap 1.824
VIF values of runqsz_Not_CPU_Bound 1.143
```

Fig 1.3D VIF Values

- We can see in above fig 1.3D VIF values is nan for pgscan which is due to only 0's in column
- So 1st we will drop 'pgscan' column and again create the model and calculate VIF value

MODEL NO: - 02 (Dropping 'pgscan' and creating a model)

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.788
Model:	OLS	Adj. R-squared:	0.788
Method:	Least Squares	F-statistic:	1064.
Date:	Thu, 02 Nov 2023	Prob (F-statistic):	0.00
Time:	21:07:47	Log-Likelihood:	-16740.
No. Observations:	5734	AIC:	3.352e+04
Df Residuals:	5713	BIC:	3.366e+04
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	83.3400	0.315	264.931	0.000	82.723	83.957
lread	-0.0544	0.009	-6.127	0.000	-0.072	-0.037
lwrite	0.0499	0.013	3.815	0.000	0.024	0.076
scall	-0.0006	6.45e-05	-9.884	0.000	-0.001	-0.001
sread	0.0015	0.001	1.426	0.154	-0.001	0.003
swrite	-0.0058	0.001	-4.008	0.000	-0.009	-0.003
fork	-0.0306	0.134	-0.228	0.819	-0.293	0.232
exec	-0.3024	0.052	-5.846	0.000	-0.404	-0.201
rchar	-5.297e-06	4.9e-07	-10.818	0.000	-6.26e-06	-4.34e-06
wchar	-5.062e-06	1.06e-06	-4.786	0.000	-7.14e-06	-2.99e-06
pgout	-0.4776	0.091	-5.259	0.000	-0.656	-0.300
ppgout	0.0117	0.079	0.148	0.882	-0.143	0.167
pgfree	0.0564	0.047	1.190	0.234	-0.037	0.149
atch	0.6480	0.145	4.472	0.000	0.364	0.932
pgin	-0.0003	0.029	-0.010	0.992	-0.056	0.056
ppgin	-0.0560	0.020	-2.826	0.005	-0.095	-0.017
pflt	-0.0328	0.002	-16.839	0.000	-0.037	-0.029
vflt	-0.0059	0.001	-4.127	0.000	-0.009	-0.003
freemem	-0.0005	5.19e-05	-9.202	0.000	-0.001	-0.000
freeswap	9.181e-06	1.9e-07	48.236	0.000	8.81e-06	9.55e-06
runqsz_Not_CPU_Bound	1.7584	0.127	13.832	0.000	1.509	2.008

Omnibus:	1021.607	Durbin-Watson:	2.011
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2058.687
Skew:	-1.069	Prob(JB):	0.00
Kurtosis:	5.011	Cond. No.	7.55e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.55e+06. This might indicate that there are strong multicollinearity or other numerical problems.

```
Model 2 Variables VIF values
const 28.122
lread 5.108
lwrite 4.17
scall 2.988
sread 6.472
swrite 5.568
fork 12.992
exec 3.133
rchar 2.098
wchar 1.596
pgout 11.412
ppgout 29.047
pgfree 15.977
atch 1.91
pgin 13.65
ppgin 13.855
pflt 11.162
vflt 15.348
freemem 1.963
freeswap 1.824
runqsz_Not_CPU_Bound 1.143
```

Table 1.3E

Model No. 02 Summary & VIF values of variables

Inferences from Table 1.3E

- We can see there is no drop in R2 and adjusted R2 values
- Next Highest VIF value is of 'ppgout' i.e., 29.047
- So we will drop 'ppgout' from previous model dataset and create a model/calculate VIF

Model No 03 (Dropping 'ppgout' and creating a model)

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.788
Model:	OLS	Adj. R-squared:	0.788
Method:	Least Squares	F-statistic:	1064.
Date:	Thu, 02 Nov 2023	Prob (F-statistic):	0.00
Time:	21:21:26	Log-Likelihood:	-16740.
No. Observations:	5734	AIC:	3.352e+04
Df Residuals:	5713	BIC:	3.366e+04
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	83.3400	0.315	264.931	0.000	82.723	83.957
lread	-0.0544	0.009	-6.127	0.000	-0.072	-0.037
lwrite	0.0499	0.013	3.815	0.000	0.024	0.076
scall	-0.0006	6.45e-05	-9.884	0.000	-0.001	-0.001
sread	0.0015	0.001	1.426	0.154	-0.001	0.003
swrite	-0.0058	0.001	-4.008	0.000	-0.009	-0.003
fork	-0.0306	0.134	-0.228	0.819	-0.293	0.232
exec	-0.3024	0.052	-5.846	0.000	-0.404	-0.201
rchar	-5.297e-06	4.9e-07	-10.818	0.000	-6.26e-06	-4.34e-06
wchar	-5.062e-06	1.06e-06	-4.786	0.000	-7.14e-06	-2.99e-06
pgout	-0.4776	0.091	-5.259	0.000	-0.656	-0.300
ppgout	0.0117	0.079	0.148	0.882	-0.143	0.167
pgfree	0.0564	0.047	1.190	0.234	-0.037	0.149
pgscan	-2.955e-14	1.37e-16	-216.384	0.000	-2.98e-14	-2.93e-14
atch	0.6480	0.145	4.472	0.000	0.364	0.932
pgin	-0.0003	0.029	-0.010	0.992	-0.056	0.056
ppgin	-0.0560	0.020	-2.826	0.005	-0.095	-0.017
pflt	-0.0328	0.002	-16.839	0.000	-0.037	-0.029
vflt	-0.0059	0.001	-4.127	0.000	-0.009	-0.003
freemem	-0.0005	5.19e-05	-9.202	0.000	-0.001	-0.000
freeswap	9.181e-06	1.9e-07	48.236	0.000	8.81e-06	9.55e-06
runqsz_Not_CPU_Bound	1.7584	0.127	13.832	0.000	1.509	2.008

Omnibus:	1021.607	Durbin-Watson:	2.011
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2058.687
Skew:	-1.069	Prob(JB):	0.00
Kurtosis:	5.011	Cond. No.	2.61e+22

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.66e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Model no. 03 VIF values

const 27.966
lread 5.105
lwrite 4.169
scall 2.988
sread 6.472
swrite 5.568
fork 12.988
exec 3.133
rchar 2.097
wchar 1.591
pgout 6.316
pgfree 5.99
atch 1.908
pgin 13.636
ppgin 13.826
pflt 11.162
vflt 15.343
freemem 1.961
freeswap 1.822
runqsz_Not_CPU_Bound 1.143

Table 1.4F

Model No. 03 Summary & VIF values of variables

Inferences from Table 1.4F: -

- We can see there is no drop in R2 and adjusted R2 values
- Next Highest VIF value is of 'vflt' i.e., 15.343
- So we will drop 'vflt' from previous model dataset and create a model/calculate VIF

Model No 04: - (Dropping 'vflt' and creating a model)

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.788
Model:	OLS	Adj. R-squared:	0.787
Method:	Least Squares	F-statistic:	1178.
Date:	Thu, 02 Nov 2023	Prob (F-statistic):	0.00
Time:	21:28:36	Log-Likelihood:	-16748.
No. Observations:	5734	AIC:	3.353e+04
Df Residuals:	5715	BIC:	3.366e+04
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	83.1811	0.312	266.745	0.000	82.570	83.792
lread	-0.0563	0.009	-6.345	0.000	-0.074	-0.039
lwrite	0.0508	0.013	3.874	0.000	0.025	0.076
scall	-0.0007	6.45e-05	-10.099	0.000	-0.001	-0.001
sread	0.0011	0.001	1.074	0.283	-0.001	0.003
swrite	-0.0057	0.001	-3.931	0.000	-0.009	-0.003
fork	-0.3024	0.117	-2.592	0.010	-0.531	-0.074
exec	-0.3116	0.052	-6.021	0.000	-0.413	-0.210
rchar	-5.417e-06	4.89e-07	-11.069	0.000	-6.38e-06	-4.46e-06
wchar	-4.57e-06	1.05e-06	-4.348	0.000	-6.63e-06	-2.51e-06
pgout	-0.4617	0.068	-6.827	0.000	-0.594	-0.329
pgfree	0.0543	0.029	1.873	0.061	-0.003	0.111
atch	0.6031	0.145	4.170	0.000	0.320	0.887
pgin	-0.0130	0.028	-0.455	0.649	-0.069	0.043
ppgin	-0.0566	0.020	-2.858	0.004	-0.095	-0.018
pflt	-0.0366	0.002	-21.179	0.000	-0.040	-0.033
freemem	-0.0005	5.19e-05	-9.315	0.000	-0.001	-0.000
freeswap	9.32e-06	1.88e-07	49.707	0.000	8.95e-06	9.69e-06
runqsz_Not_CPU_Bound	1.7557	0.127	13.795	0.000	1.506	2.005

Omnibus:	982.221	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1936.799
Skew:	-1.041	Prob(JB):	0.00
Kurtosis:	4.942	Cond. No.	7.47e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.47e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Model no. 04 VIF values
const 27.562
lread 5.091
lwrite 4.168
scall 2.98
sread 6.424
swrite 5.566
fork 9.847
exec 3.127
rchar 2.09
wchar 1.572
pgout 6.312
pgfree 5.966
atch 1.898
pgin 13.482
ppgin 13.825
pflt 8.737
freemem 1.959
freeswap 1.765
runqsz_Not_CPU_Bound 1.143

Table 1.3G

Model No. 04 Summary & VIF values of variables

Inferences from Table 1.4G: -

- We can see there is no drop in R2 whereas Adj. R2 value dropped by 0.001 which acceptable.
- Next Highest VIF value is of 'ppgin' i.e., 13.825
- So we will drop 'ppgin' from previous model dataset and create a model/calculate VIF

Model No 05: - (Dropping 'ppgin' and creating a model)

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.787
Model:	OLS	Adj. R-squared:	0.787
Method:	Least Squares	F-statistic:	1246.
Date:	Thu, 02 Nov 2023	Prob (F-statistic):	0.00
Time:	21:39:10	Log-Likelihood:	-16752.
No. Observations:	5734	AIC:	3.354e+04
Df Residuals:	5716	BIC:	3.366e+04
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	83.2131	0.312	266.853	0.000	82.602	83.824
lread	-0.0580	0.009	-6.542	0.000	-0.075	-0.041
lwrite	0.0524	0.013	3.998	0.000	0.027	0.078
scall	-0.0006	6.45e-05	-10.044	0.000	-0.001	-0.001
sread	0.0011	0.001	1.113	0.266	-0.001	0.003
swrite	-0.0057	0.001	-3.975	0.000	-0.009	-0.003
fork	-0.2896	0.117	-2.482	0.013	-0.518	-0.061
exec	-0.3114	0.052	-6.014	0.000	-0.413	-0.210
rchar	-5.577e-06	4.86e-07	-11.464	0.000	-6.53e-06	-4.62e-06
wchar	-4.571e-06	1.05e-06	-4.347	0.000	-6.63e-06	-2.51e-06
pgout	-0.4533	0.068	-6.705	0.000	-0.586	-0.321
pgfree	0.0445	0.029	1.544	0.123	-0.012	0.101
atch	0.6099	0.145	4.215	0.000	0.326	0.894
pgin	-0.0896	0.010	-9.362	0.000	-0.108	-0.071
pflt	-0.0367	0.002	-21.197	0.000	-0.040	-0.033
freemem	-0.0005	5.19e-05	-9.367	0.000	-0.001	-0.000
freeswap	9.306e-06	1.88e-07	49.618	0.000	8.94e-06	9.67e-06
runqsz_Not_CPU_Bound	1.7521	0.127	13.758	0.000	1.502	2.002

Omnibus:	977.800	Durbin-Watson:	2.006
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1922.242
Skew:	-1.038	Prob(JB):	0.00
Kurtosis:	4.933	Cond. No.	7.46e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.46e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Model no. 05 VIF values
const 27.527
lread 5.069
lwrite 4.16
scall 2.979
sread 6.423
swrite 5.565
fork 9.832
exec 3.127
rchar 2.063
wchar 1.572
pgout 6.3
pgfree 5.882
atch 1.897
pgin 1.521
pflt 8.736
freemem 1.958
freeswap 1.764
runqsz_Not_CPU_Bound 1.143

Table 1.3H

Model No. 05 Summary & VIF values of variables

Inferences from Table 1.3H: -

- We can see R2 dropped by 0.001 whereas Adj. R2 value didn't drop which acceptable.
- Next Highest VIF value is of 'fork' i.e.,9.832
- So we will drop 'fork' from previous model dataset and create a model/calculate VIF

Model No 06: - (Dropping 'fork' and creating a model)

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.787
Model:	OLS	Adj. R-squared:	0.787
Method:	Least Squares	F-statistic:	1322.
Date:	Thu, 02 Nov 2023	Prob (F-statistic):	0.00
Time:	21:52:06	Log-Likelihood:	-16755.
No. Observations:	5734	AIC:	3.354e+04
Df Residuals:	5717	BIC:	3.366e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	83.2656	0.311	267.516	0.000	82.655	83.876
lread	-0.0594	0.009	-6.719	0.000	-0.077	-0.042
lwrite	0.0550	0.013	4.208	0.000	0.029	0.081
scall	-0.0006	6.4e-05	-9.808	0.000	-0.001	-0.001
sread	0.0012	0.001	1.124	0.261	-0.001	0.003
swrite	-0.0065	0.001	-4.577	0.000	-0.009	-0.004
exec	-0.3569	0.048	-7.364	0.000	-0.452	-0.262
rchar	-5.618e-06	4.86e-07	-11.550	0.000	-6.57e-06	-4.66e-06
wchar	-4.324e-06	1.05e-06	-4.128	0.000	-6.38e-06	-2.27e-06
pgout	-0.4504	0.068	-6.659	0.000	-0.583	-0.318
pgfree	0.0431	0.029	1.495	0.135	-0.013	0.100
atch	0.6166	0.145	4.260	0.000	0.333	0.900
pgin	-0.0885	0.010	-9.249	0.000	-0.107	-0.070
pflt	-0.0400	0.001	-37.282	0.000	-0.042	-0.038
freemem	-0.0005	5.2e-05	-9.405	0.000	-0.001	-0.000
freeswap	9.289e-06	1.88e-07	49.538	0.000	8.92e-06	9.66e-06
runqsz_Not_CPU_Bound	1.7519	0.127	13.751	0.000	1.502	2.002

Omnibus:	964.218	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1903.267
Skew:	-1.023	Prob(JB):	0.00
Kurtosis:	4.944	Cond. No.	7.44e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.44e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Model no. 06 VIF values

const 27.4
lread 5.046
lwrite 4.133
scall 2.934
sread 6.423
swrite 5.332
exec 2.736
rchar 2.06
wchar 1.558
pgout 6.298
pgfree 5.88
atch 1.897
pgin 1.518
pflt 3.363
freemem 1.958
freeswap 1.762
runqsz_Not_CPU_Bound 1.143

Table 1.3I
Model No. 06 Summary & VIF values of variables

Inferences from Table 1.3I: -

- We can see R2 values and Adj. R2 value didn't drop.
- Next Highest VIF value is of 'sread' i.e., 6.432
- So we will drop 'sread' from previous model dataset and create a model/calculate VIF

Model No 07: - (Dropping 'sread' and creating a model)

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.787
Model:	OLS	Adj. R-squared:	0.787
Method:	Least Squares	F-statistic:	1410.
Date:	Thu, 02 Nov 2023	Prob (F-statistic):	0.00
Time:	22:05:34	Log-Likelihood:	-16756.
No. Observations:	5734	AIC:	3.354e+04
Df Residuals:	5718	BIC:	3.365e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	83.2853	0.311	268.001	0.000	82.676	83.895
lread	-0.0599	0.009	-6.785	0.000	-0.077	-0.043
lwrite	0.0558	0.013	4.283	0.000	0.030	0.081
scall	-0.0006	6.08e-05	-9.955	0.000	-0.001	-0.000
swrite	-0.0054	0.001	-5.127	0.000	-0.007	-0.003
exec	-0.3592	0.048	-7.418	0.000	-0.454	-0.264
rchar	-5.373e-06	4.35e-07	-12.356	0.000	-6.23e-06	-4.52e-06
wchar	-4.433e-06	1.04e-06	-4.251	0.000	-6.48e-06	-2.39e-06
pgout	-0.4512	0.068	-6.671	0.000	-0.584	-0.319
pgfree	0.0440	0.029	1.528	0.126	-0.012	0.101
atch	0.6117	0.145	4.228	0.000	0.328	0.895
pgin	-0.0886	0.010	-9.268	0.000	-0.107	-0.070
pflt	-0.0399	0.001	-37.309	0.000	-0.042	-0.038
freemem	-0.0005	5.2e-05	-9.394	0.000	-0.001	-0.000
freeswap	9.267e-06	1.87e-07	49.687	0.000	8.9e-06	9.63e-06
runqsz_Not_CPU_Bound	1.7557	0.127	13.785	0.000	1.506	2.005

Omnibus:	967.583	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1913.037
Skew:	-1.026	Prob(JB):	0.00
Kurtosis:	4.949	Cond. No.	7.43e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.43e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Model no. 07 VIF values

const	27.313
lread	5.033
lwrite	4.119
scall	2.647
swrite	2.975
exec	2.731
rchar	1.646
wchar	1.544
pgout	6.298
pgfree	5.875
atch	1.895
pgin	1.517
pflt	3.342
freemem	1.958
freeswap	1.743
runqsz_Not_CPU_Bound	1.142

dtype: float64

Table 1.3J

Model No. 07 Summary & VIF values of variables

Inferences from Table 1.3J: -

- We can see R2 values and Adj. R2 value didn't drop.
- Next Highest VIF value is of 'pgout' i.e., 6.298
- So we will drop 'pgout' from previous model dataset and create a model/calculate VIF

Model No 08: - (Dropping 'pgout' and creating a model)

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.786
Model:	OLS	Adj. R-squared:	0.785
Method:	Least Squares	F-statistic:	1496.
Date:	Thu, 02 Nov 2023	Prob (F-statistic):	0.00
Time:	22:11:38	Log-Likelihood:	-16778.
No. Observations:	5734	AIC:	3.359e+04
Df Residuals:	5719	BIC:	3.369e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	83.2332	0.312	266.904	0.000	82.622	83.845
lread	-0.0584	0.009	-6.584	0.000	-0.076	-0.041
lwrite	0.0537	0.013	4.103	0.000	0.028	0.079
scall	-0.0006	6.1e-05	-9.997	0.000	-0.001	-0.000
swrite	-0.0053	0.001	-4.981	0.000	-0.007	-0.003
exec	-0.3453	0.049	-7.112	0.000	-0.441	-0.250
rchar	-5.28e-06	4.36e-07	-12.103	0.000	-6.14e-06	-4.42e-06
wchar	-4.925e-06	1.04e-06	-4.717	0.000	-6.97e-06	-2.88e-06
pgfree	-0.1140	0.016	-6.915	0.000	-0.146	-0.082
atch	0.3284	0.139	2.365	0.018	0.056	0.601
pgin	-0.0897	0.010	-9.350	0.000	-0.109	-0.071
pflt	-0.0402	0.001	-37.418	0.000	-0.042	-0.038
freemem	-0.0005	5.21e-05	-9.068	0.000	-0.001	-0.000
freeswap	9.296e-06	1.87e-07	49.665	0.000	8.93e-06	9.66e-06
runqsz_Not_CPU_Bound	1.7108	0.128	13.401	0.000	1.461	1.961

Omnibus:	976.368	Durbin-Watson:	2.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1935.543
Skew:	-1.033	Prob(JB):	0.00
Kurtosis:	4.958	Cond. No.	7.43e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.43e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Model No. 08 VIF values

const	27.296
lread	5.030
lwrite	4.116
scall	2.647
swrite	2.974
exec	2.726
rchar	1.645
wchar	1.536
pgfree	1.907
atch	1.732
pgin	1.517
pflt	3.338
freemem	1.954
freeswap	1.742
runqsz_Not_CPU_Bound	1.139

dtype: float64

Table 1.3K
Model No. 08 Summary & VIF values of variables

Inferences from Table 1.3K: -

- We can see R2 dropped to 0.786 and Adj. R2 value dropped to 0.785, which is acceptable.
- Now there is no column/features which has VIF values greater than 5. So now our dataset is free from multicollinearity
- Also there is no such column/variable whose p-value is greater than 0.05.

Final Equation after removing the columns/variables which are causing Multicollinearity: -

Equation after dropping varibales have correlation\Multicollinearity :

USR = 83.233 +(-0.058)* lread +(0.054)* lwrite +(-0.001)* scall +(-0.005)* swrite +(-0.345)* exec +(-0.0)* rchar + (-0.0)* wchar +(-0.114)* pgfree +(0.328)* atch +(-0.09)* pgin +(-0.04)* pflt +(-0.0)* freemem +(0.0)* freeswap +(1.711)* runqsz_Not_CPU_Bound

Checking the performance of Predictions on Train sets using Rsquare & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.

	R2 Value	Adjusted R2 value
Model No. 2	0.788	0.788
Model No. 3	0.788	0.788
Model No. 4	0.788	0.788
Model No. 5	0.788	0.787
Model No. 6	0.787	0.787
Model No. 7	0.787	0.787
Model No. 8	0.786	0.785

Table 1.3L
Performance Comparison of Models (Statsmodel)

- From above Table 1.3L we can see that after dropping 7 variables our model performance i.e., R2 value and Adj. R2 values dropped by only 0.002 & 0.003 respectively
- So we will select Model No. 8 after performing Linear regression using Statsmodel

 # Lets create a model using variables of Model No. 08 using sklearn and calculate performance of Predictions on Train and Test sets using Model score and RMSE value.

```
Model score for Train dataset (after dropping variables causing Multicollinearity) =  0.7855
Model score for Test dataset (after dropping variables causing Multicollinearity) =  0.7846
Root Mean Squared Error of Train dataset(with outliers) =  4.514
Root Mean Squared Error of Test dataset (with outliers) =  4.526
```

	Model with Multicollinearity	**Model without Multicollinearity**
Model score (Train Data)	0.7884	0.7855
Model Score (Test Data)	0.7881	0.7846
RSME value of Train Data	4.483	4.514
RSME value of Test Data	4.493	4.526

Table 1.3M
Comparison between model with Multicollinearity and without Multicollinearity

Inferences from above Table 1.3M

- We can see that the model score for train and test data doesn't vary much even after dropping variables causing multicollinearity
- Also RSME value of Train and Test data doesn't increases (very slight i.e., 0.031).

The best Model is the final model i.e Model no. 8, i.e., model without multicolinearity
Hence the best model and its equation after creating model using Statsmodel =

```
Equation after dropping varibales have correlation\Multicollinearity :
 USR = 83.233 +( -0.058 )* lread +( 0.054 )* lwrite +( -0.001 )* scall +( -0.005 )* swrite +( -0.345 )* exec +( -0.0 )* rchar +
( -0.0 )* wchar +( -0.114 )* pgfree +( 0.328 )* atch +( -0.09 )* pgin +( -0.04 )* pflt +( -0.0 )* freemem +( 0.0 )* freeswap +(
1.711 )* runqsz_Not_CPU_Bound
```

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

The comp-activ databases is a collection of a computer systems activity measures.

Inference from dataset and business insights: -

- The dataset has 8192 rows and 21 columns
- Out of 21 columns: - 13 columns with float64 dtypes, 8 columns with int64 dtypes and 1 column with object dtypes
- 'pgscan' column has more than 75% data points as 0's
- Dataset has 104 null entries in 'rchar' column and 14 'wchar' column which has been imputed by Median.
- We can see all the numeric column except usr' & 'freeswap' are right skewed whereas 'usr' and 'freeswap' is left skewed.
- All the numeric column has outliers in them which has been treated with Upper Fence and Lower Fence.
- 'usr' is our target variable.
- We checked for the multicollinearity of the independent variables and found that some of the variables has VIF values greater than 10, which means there is strong correlation/Multicollinearity between some independent variables.

- First we created model using Sklearn library for both datasets (with outliers) and dataset (without outliers)
    - Model score for dataset (with outliers) for Train datasets = 0.6376 and Test datasets = 0.6295
    - Model score for dataset (without outliers) for Train datasets = 0.788 and Test datasets = 0.787
    - RMSE value for dataset (with outliers) for Train datasets = 10.89 and Test datasets = 11.62
    - RMSE value for dataset (without outliers) for Train datasets = 4.48 and Test datasets = 4.49
    - On observing model score and performance for both Model (with outliers) and Model (without outliers), we see that the model behaves better when created without outliers

- Later we created model with datasets (without outliers) using statsmodel
    - After creating model using statsmodel we get R2 and Adj. R2 as 0.788
    - On calculating VIF values for independent variables we found that 'pgscan' has VIF values as nan, which can be dropped 1st. Again create a model and calculate VIF
    - After creating Model no 02 and calculating VIF value, we find that R2 and Adj. R2 value didn't drop On calculating VIF value we find that vif value of 'ppgout' is highest i.e., 29.047
    - We will drop this variable also and again create a model and calculate VIF value. We will continue dropping of variable until all independent variables has VIF values less than 5
    - So after dropping and creating model on basis of VIF values we were able to drop 7 variables i.e., 'pgscan' , 'ppgout' ,'vflt' , 'ppgin' , 'fork' , 'sread' and 'pgout' whereas our R2 and Adjusted R2 values dopped from 0.788 to 0.786 .

- So creating Linear Regression model using Statsmodel we were able to drop 7 variables and our R2 and Adj R2 dropped by slightest i.e., from 0.788 & 0.788 to 0.786 & 0.785

- Insights from the Final Linear Regression Equation Model (Without Multicollinearity): -

    - When lread increases by 1 unit, usr decreases by 0.058 keeping all other predictors variable constant
    - When lwrite increases by 1 unit, usr increase by 0.05 keeping all other predictors variable constant
    - When scall increases by 1 unit, usr decreases by 0.001 keeping all other predictors variable constant
    - When swrite increases by 1 unit, usr decreases by 0.005 keeping all other predictors variable constant
    - When exec increases by 1 unit, usr decreases by 0.345 keeping all other predictors variable constant
    - When pgree increases by 1 unit, usr decreases by 0.114 keeping all other predictors variable constant
    - When atch increases by 1 unit, usr increase by 0.328 keeping all other predictors variable constant
    - When pgin increases by 1 unit, usr decreases by 0.009 keeping all other predictors variable constant
    - When pflt increases by 1 unit, usr decreases by 0.04 keeping all other predictors variable constant
    - When runqz_NOT_CPU_BOUND increases by 1 unit, usr increases by 1.711 keeping all other predictors variable constant