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.

Here we 5 rows and 22 column are present in our Dataset.

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppg
0	1	0	2147	79	68	0.2	0.2	40671.0	53995.0	0.0	...	0.0	0.0	1.6	2
1	0	0	170	18	21	0.2	0.2	448.0	8385.0	0.0	...	0.0	0.0	0.0	0
2	15	3	2162	159	119	2.0	2.4	NaN	31950.0	0.0	...	0.0	1.2	6.0	9
3	0	0	160	12	16	0.2	0.2	NaN	8670.0	0.0	...	0.0	0.0	0.2	0

There are 7 integer,13 float and 1 object are occurred in the dataset.

```
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
```

Out of 21 column, 17 are high standard deviation so there is a large dispersion and heavy variation occurred in compactiv database.

	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

We can see the null is present in rchar and wchar features.

```
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
```

So we can identify the the rchar features has been 87% null are present in the compactiv. so we can drop the feature.

```
lread         0.000000
lwrite        0.000000
scall         0.000000
sread         0.000000
swrite        0.000000
fork          0.000000
exec          0.000000
rchar        87.394958
wchar        12.605042
pgout         0.000000
ppgout        0.000000
pgfree        0.000000
pgscan        0.000000
atch          0.000000
pgin          0.000000
ppgin         0.000000
```

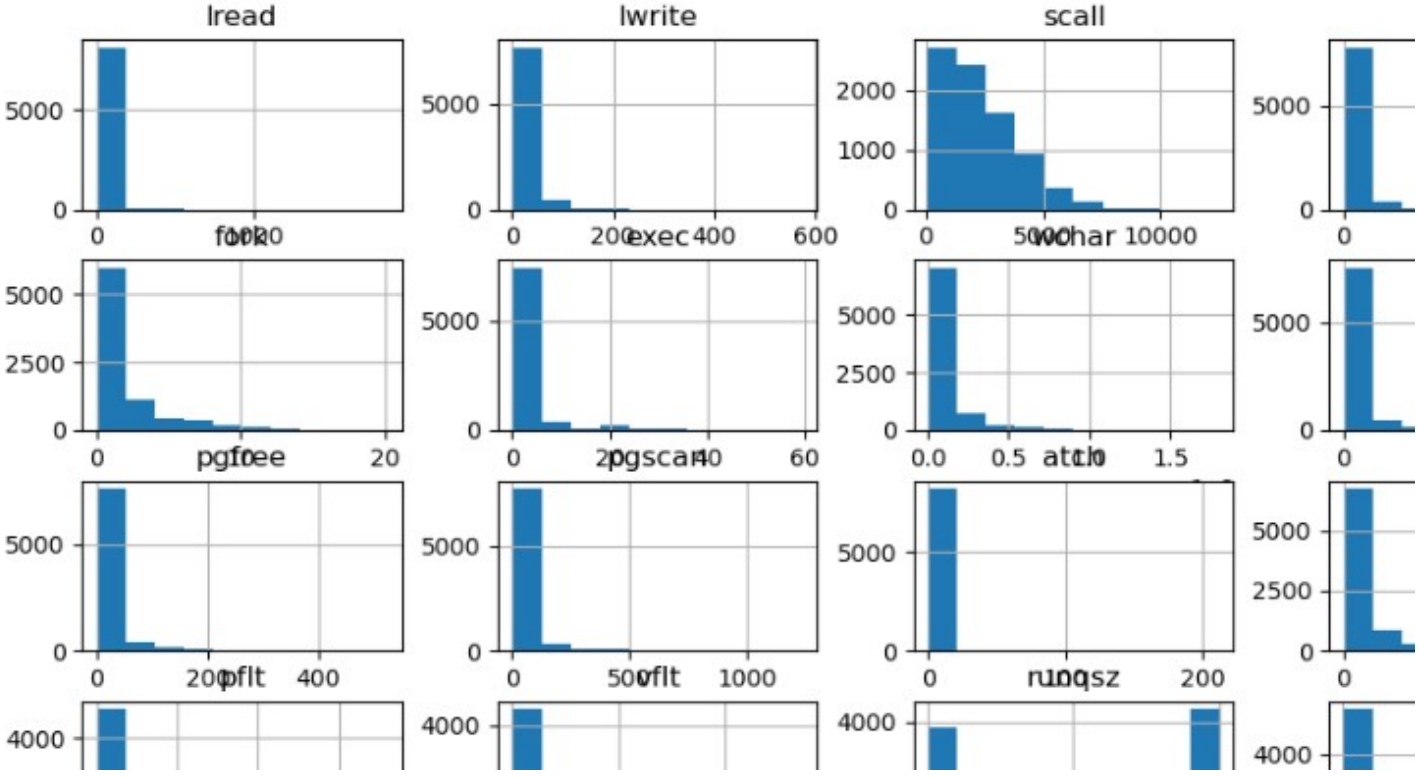
	lread	lwrite	scall	sread	swrite	fork	exec	wchar	pgout	ppgout	...	pgscan	atch	pgin	p
0	1	0	2147	79	68	0.2	0.20	53995.0	0.00	0.00	...	0.00	0.0	1.60	
1	0	0	170	18	21	0.2	0.20	8385.0	0.00	0.00	...	0.00	0.0	0.00	
2	15	3	2162	159	119	2.0	2.40	31950.0	0.00	0.00	...	0.00	1.2	6.00	
3	0	0	160	12	16	0.2	0.20	8670.0	0.00	0.00	...	0.00	0.0	0.20	
4	5	1	330	39	38	0.4	0.40	12185.0	0.00	0.00	...	0.00	0.0	1.00	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
8187	16	12	3009	360	244	1.6	5.81	85282.0	8.02	20.64	...	55.11	0.6	35.87	4
8188	4	0	1596	170	146	2.4	1.80	41764.0	3.80	4.80	...	0.20	0.8	3.80	

Histogram:

In the boxplot, most features are right skewed and except “usr” and “freeswap” features. The runqsz features observed as a slight negative skewed distribution.The scall features are slight normal distribution occurred.

The atch column has a highly positive skewed distribution and lread has a second positive skewness.

```
lread
Skew: 13.9
lwrite
Skew: 5.28
scall
Skew: 0.9
sread
Skew: 5.46
swrite
Skew: 9.61
fork
Skew: 2.25
exec
Skew: 4.07
wchar
Skew: 3.85
pgout
Skew: 5.07
ppgout
Skew: 4.68
pgfree
Skew: 4.77
pgscan
Skew: 5.81
atch
Skew: 21.54
pgin
Skew: 3.24
```



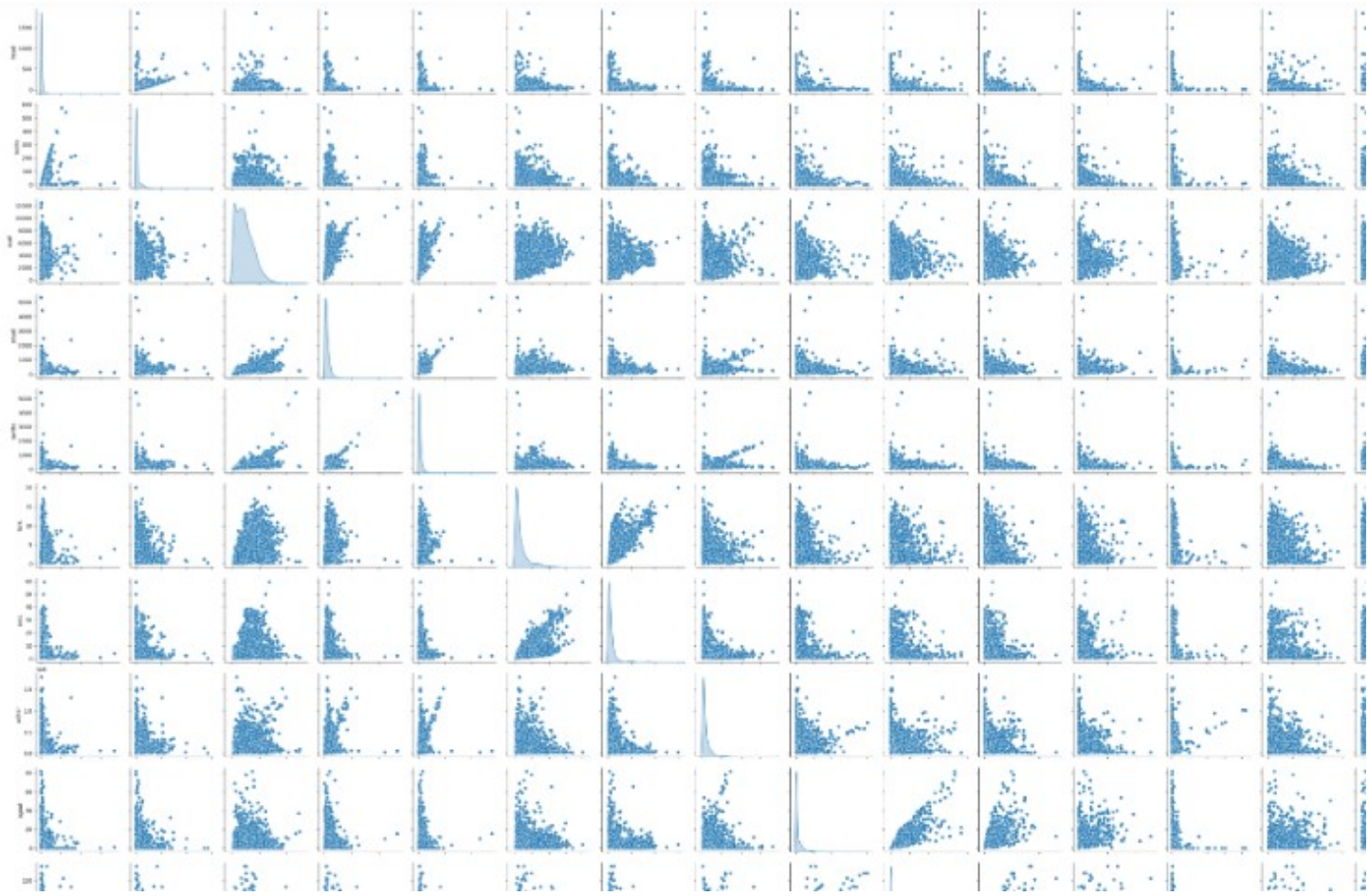
Pairplot features.

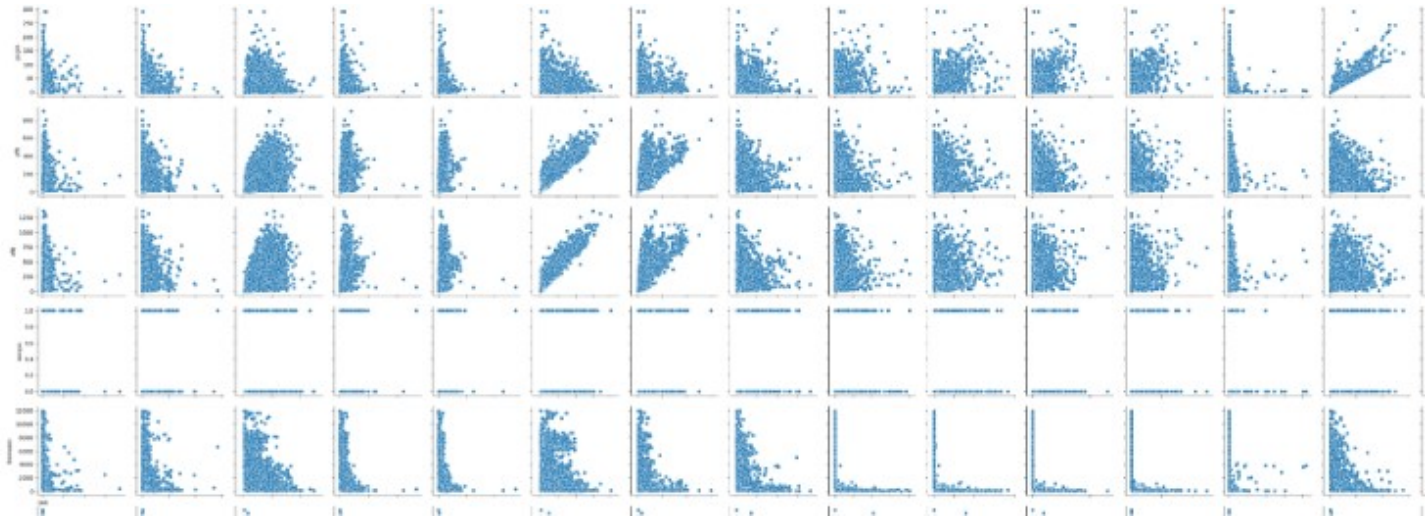
Strongly Correlation:

In this pair plot , the “lread”vs “lwrite” features are positive correlation and sread vs swrite column has positive correlation.the fork vs exec feature has positive correlation and strongly to each other.pgout vs ppgout has strongly correlated.

Weak Correlation:

The **pairplot** clearly say the USR column has some cloudy(non-linear) and few has negative correlated with the respective columns.





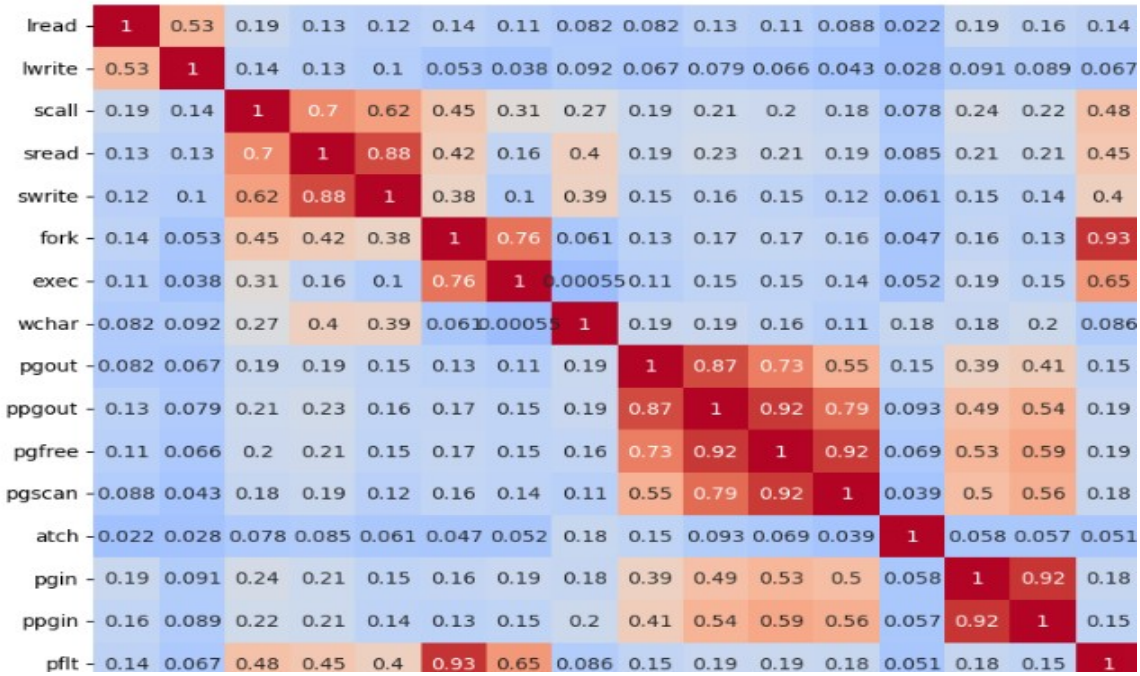
Heatmap – Correlation:

Strong Features:

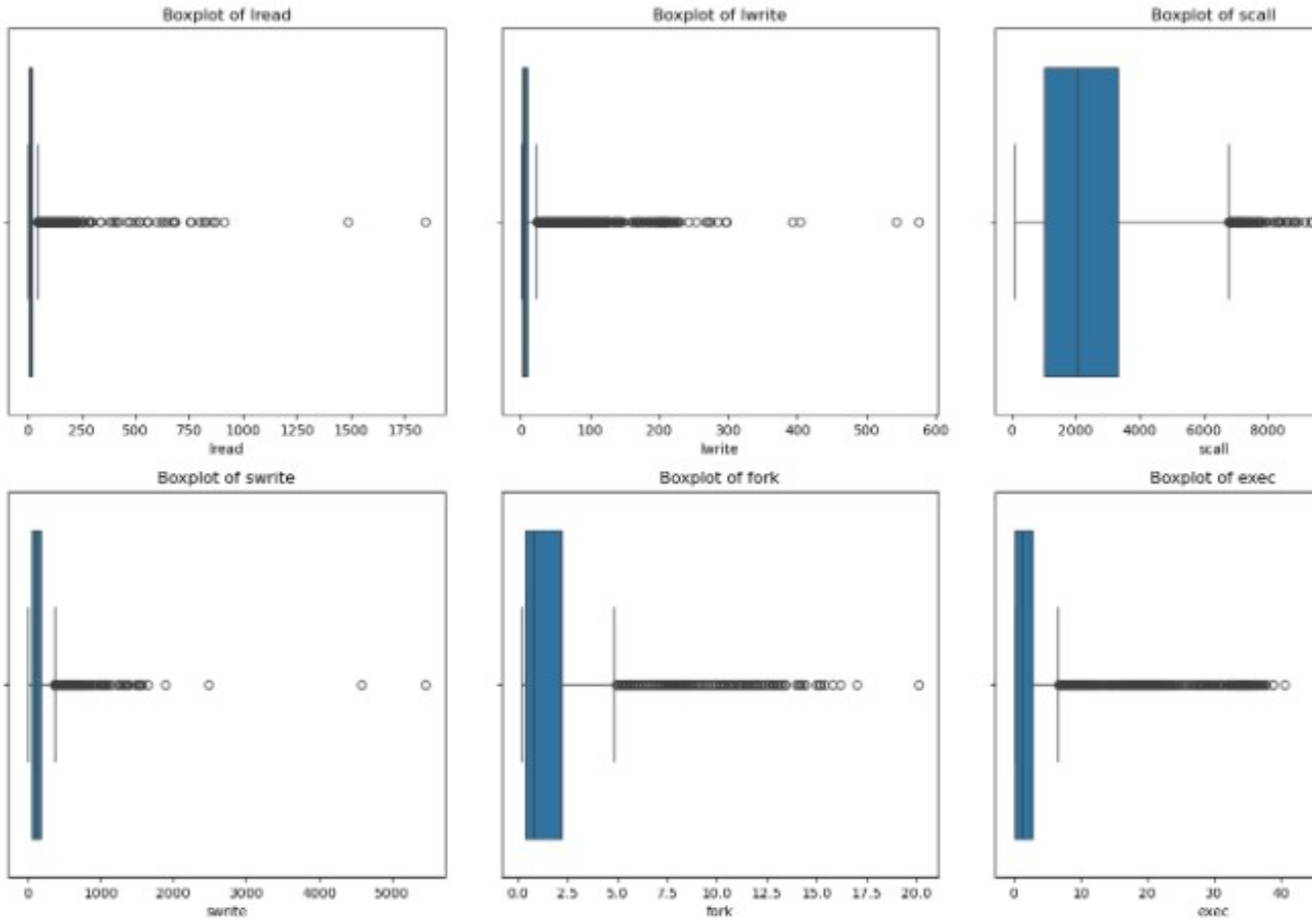
Fork vs Vflt and **fork vs pflt,pflt vs vflt** are the strong correlation and the 2nd strong correlation is **ppgout vs pgfree** and **pgfree vs pgscan**.

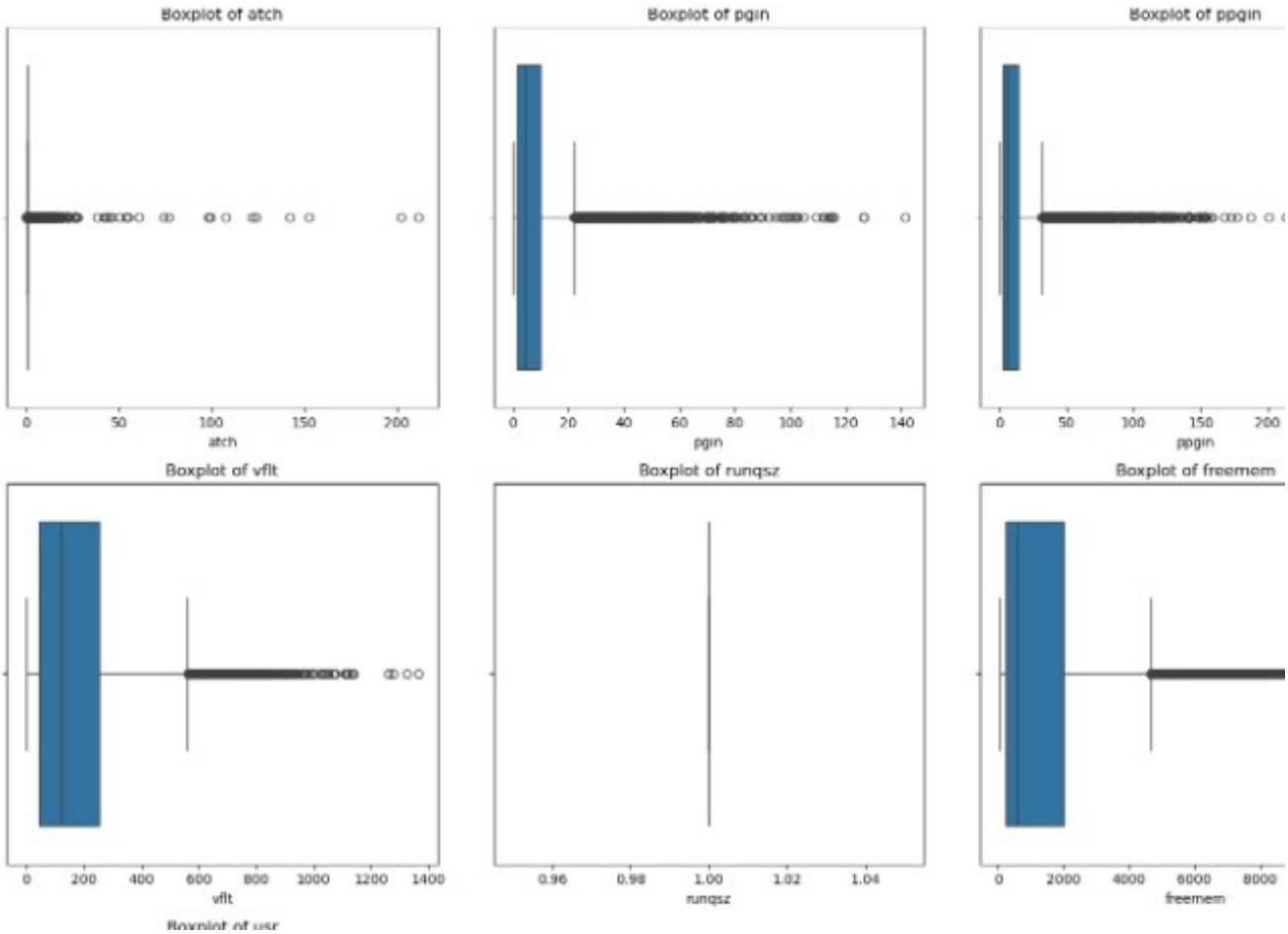
The page fault caused by protection errors and address translation are strongly correlated with each other because those are all faults.

Sread vs Swrite are the strong correlation and ppgout vs pgout, ppgout vs pgfree column,pgscan vs pgfree, ppgin vs pgin are also having the Strong correlation.

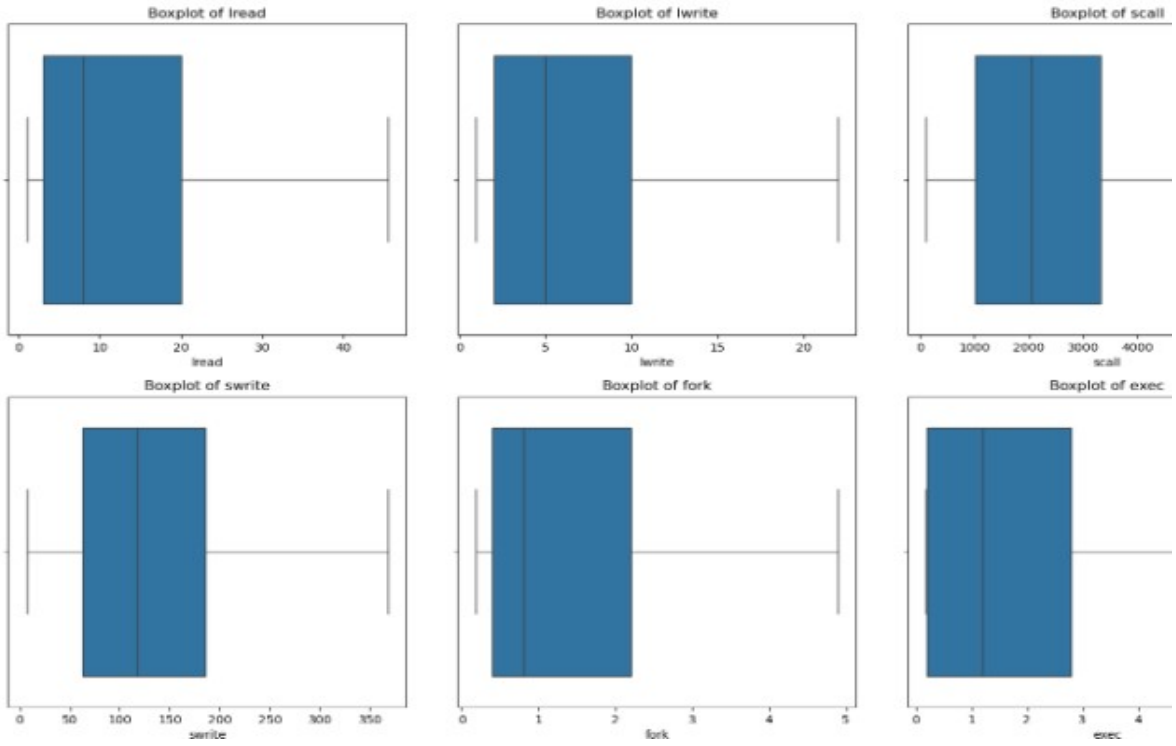


Outlier Detection and removal process.


Boxplot of lread
0
250
500
750
1000
1250
1500
1750
lread
Boxplot of lwrite
0
100
200
300
400
500
600
lwrite
Boxplot of scall
0
2000
4000
6000
8000
scall
Boxplot of swrite
0
1000
2000
3000
4000
5000
swrite
Boxplot of fork
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
fork
Boxplot of exec
0
10
20
30
40
exec



Removal of outlier



Removal of outlier process based on some reasons

- **Impact on Model Performance:** Ouliers are significantly affect the model performance, particularly those sensitive to extreme values,especially linear regression algorithm.It may lead skewed coefficients and less accurate predictions.
- Outlier can introduce heteroscedasticity it leads to the violation of this affecting the precision of coefficient estimates.
- Outlier are affected the model and it can give a non-linearity it leads to poor fitting model and biased coefficients and can give a inaccurate prediction and outlier can make the model more challenging to interpretability.

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.

Linear Regression:

It' s shows the linear relationship between linear relationship between independent variable and dependent variable. The mathematical formula for linear regression formula is **y=a+bx+cx^2.**

Trained Coefficient:

These are all the coefficients for all features.

```
The coefficient for lread is -0.03342162975755825
The coefficient for lwrite is 0.018458229716677235
The coefficient for scall is -0.001313750034605298
The coefficient for sread is -0.0021431027513734288
The coefficient for swrite is -0.004288498652823567
The coefficient for fork is 0.3089989670478895
The coefficient for exec is -0.4543536118883841
The coefficient for wchar is -7.3290963214992365e-06
The coefficient for pgout is 5.551115123125783e-16
The coefficient for ppgout is -2.192690473634684e-15
The coefficient for pgfree is 2.7755575615628914e-17
The coefficient for pgscan is -1.6653345369377348e-16
The coefficient for atch is -2.7755575615628957e-17
```

Interception value:

The interception values is between the mean of response variable and the product of the slope and the mean of explanatory variable.

99.71387625526799

OLS.Summary:

OLS Regression Results

Dep. Variable:	usr	R-squared:	
Model:	OLS	Adj. R-squared:	
Method:	Least Squares	F-statistic:	
Date:	Wed, 17 Jan 2024	Prob (F-statistic):	
Time:	22:24:41	Log-Likelihood:	
No. Observations:	5734	AIC:	
Df Residuals:	5719	BIC:	
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	
lread	-0.0334	0.006	-5.833	0.000	
lwrite	0.0185	0.010	1.824	0.068	
scall	-0.0013	4.35e-05	-30.233	0.000	
pgscan	1.6392	0.003	483.760	0.000	
atch	0.0226	4.68e-05	483.760	0.000	
pgin	-0.0610	0.021	-2.944	0.003	-
ppgin	-0.0710	0.014	-5.023	0.000	-
pflt	-0.0214	0.001	-15.434	0.000	-
vflt	-0.0154	0.001	-15.673	0.000	-
freemem	0.0004	3.31e-05	13.133	0.000	
freeswap	-1.213e-06	1.31e-07	-9.233	0.000	-1.4

Omnibus:	816.260	Durbin-Watson:	
Prob(Omnibus):	0.000	Jarque-Bera (JB):	
Skew:	-0.423	Prob(JB):	
Kurtosis:	8.238	Cond. No.	

P-Value:

The p-value is the statistical number to conclude the relationship between response variable(independent variable) and predictor variable(dependent variable).

Low P-Value:

lread,scall,sread,swrite,fork,exec,wchar,pgout,ppgout,pgfree,pgscan,atch,pgin,ppgin,pflt,vflt,freemem,freeswap

The coefficients are not equal to zero.
High P-Value: write is the only feature and we cannot conclude that explanatory variable(lwrite) directly affect the predict variable(usr).

VIF Factors:

It measures the amount of multicollinearity in regression analysis. VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable.

```
VIF values:

lread        4.021769
lwrite       3.339191
scall        2.902280
sread        5.185813
swrite       5.166778
fork        12.791100
exec         3.207212
wchar        1.305614
pgout        0.000022
ppgout       0.000003
pgfree       0.000122
pgscan       0.000000
atch         0.000005
```

We can drop the column as per the VIF above 5, so the fork, pgin, ppgin, pflt and vflt has been dropped and check the summary once again.

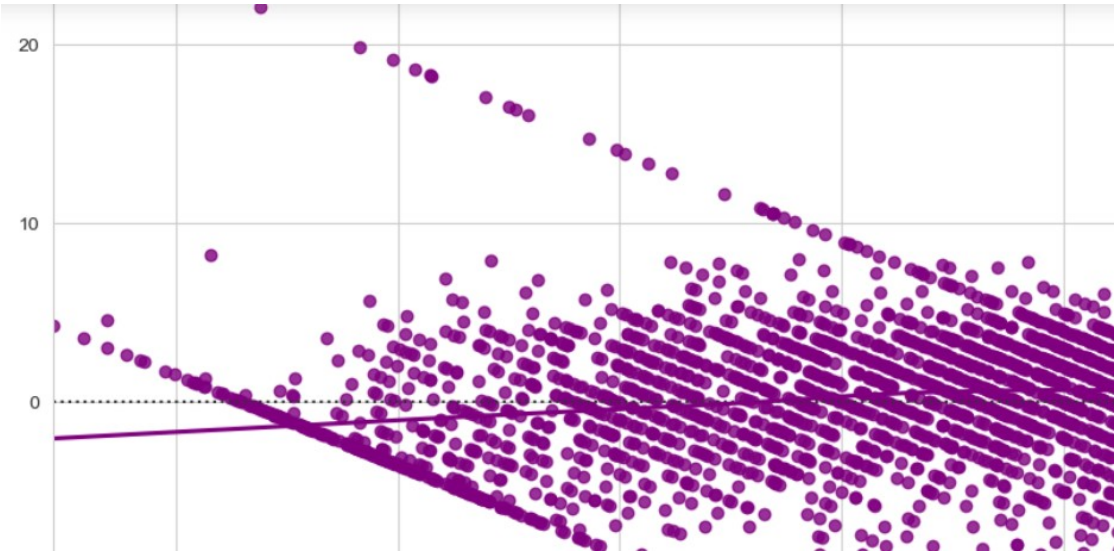
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.861
Model:                            OLS   Adj. R-squared:                  0.861
Method:                 Least Squares   F-statistic:                     2728.
Date:                Fri, 19 Jan 2024   Prob (F-statistic):               0.00
Time:                        14:24:42   Log-Likelihood:                -14605.
No. Observations:                5734   AIC:                         2.924e+04
Df Residuals:                    5720   BIC:                         2.933e+04
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
lread         -0.0247      0.003     -7.749      0.000      -0.031      -0.018
scall         -0.0013   4.34e-05    -30.431      0.000      -0.001      -0.001
```

pgscan	1.6403	0.003	491.678	0.000	1.634	1.647
atch	0.0226	4.6e-05	491.678	0.000	0.023	0.023
pgin	-0.0612	0.021	-2.953	0.003	-0.102	-0.021
ppgin	-0.0720	0.014	-5.092	0.000	-0.100	-0.044
pflt	-0.0215	0.001	-15.525	0.000	-0.024	-0.019
vflt	-0.0154	0.001	-15.724	0.000	-0.017	-0.014
freemem	0.0004	3.31e-05	13.100	0.000	0.000	0.000
freeswap	-1.21e-06	1.31e-07	-9.205	0.000	-1.47e-06	-9.52e-07

Omnibus:	817.370	Durbin-Watson:	1.977
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6686.361
Skew:	-0.427	Prob(JB):	0.00
Kurtosis:	8.221	Cond. No.	7.47e+21

Assumption of Linear regression:

The assumption of linearity is satisfied as per the residuals.



The residuals are not normal as per the Shapiro test.

```
ShapiroResult(statistic=0.9398130178451538, pvalue=2.788583944006386e-43)
```

Homoscedasticity :

0.2319

Because the residuals is greater than above p-value.

R-Squared and Adjusted R-Squared variable:

It is a statistical measure used in regression analysis to assess the goodness of fit of a model.

R-Squared Value for Train:

It Represents the proportion of the variance in the dependent variable explained by the independent variable for train and test.

0.86

R-Squared Values for Test:

0.84

RMSE For Train:
Root mean Squared Error:

It is used metric in regression analysis to evaluate the performance of a predictive model and it's calculated by taking square root of MSE

3.0890

RMSE for Test:
3.14

Coefficient for Significant Variables:

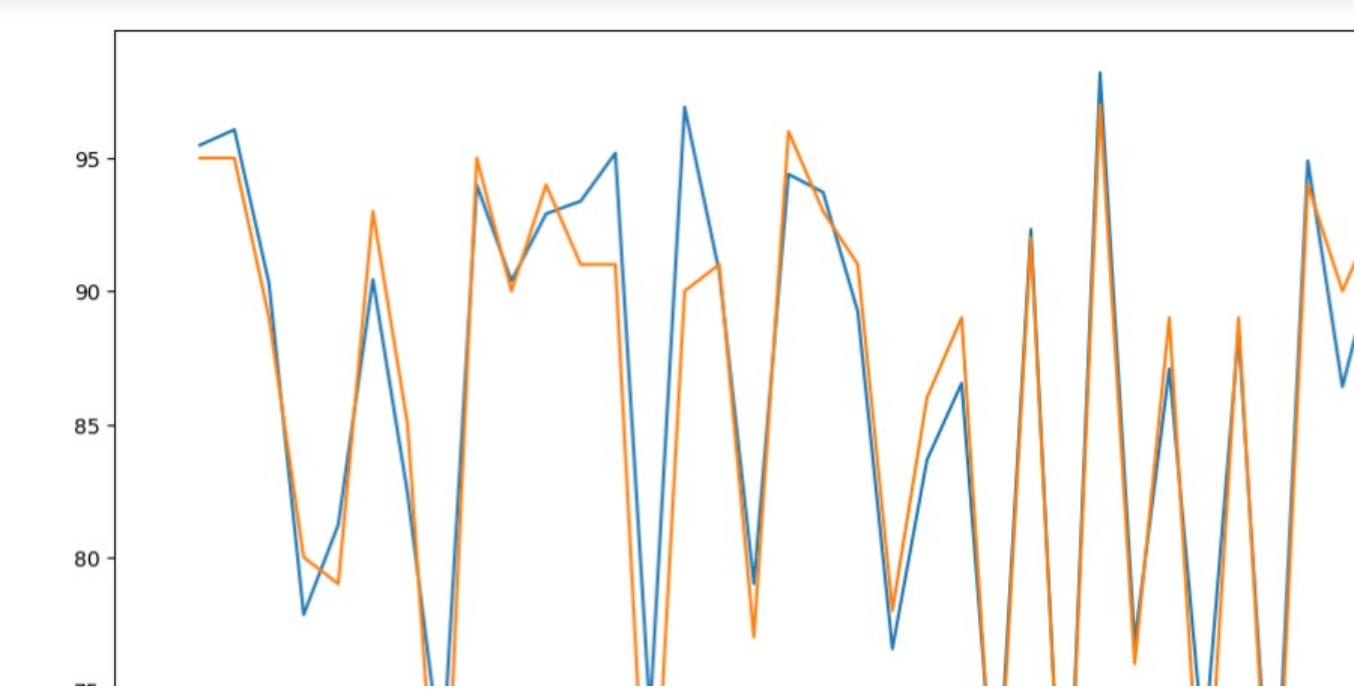
	Variable	Coefficient
0	lread	-3.342163e-02
1	lwrite	1.845823e-02
2	scall	-1.313750e-03
3	sread	-2.143103e-03
4	swrite	-4.288499e-03
5	fork	3.089990e-01
6	exec	-4.543536e-01
7	wchar	-7.329096e-06
8	pgout	5.551115e-16
9	ppgout	-2.192690e-15
10	pgfree	2.775558e-17
11	pgscan	-1.665335e-16

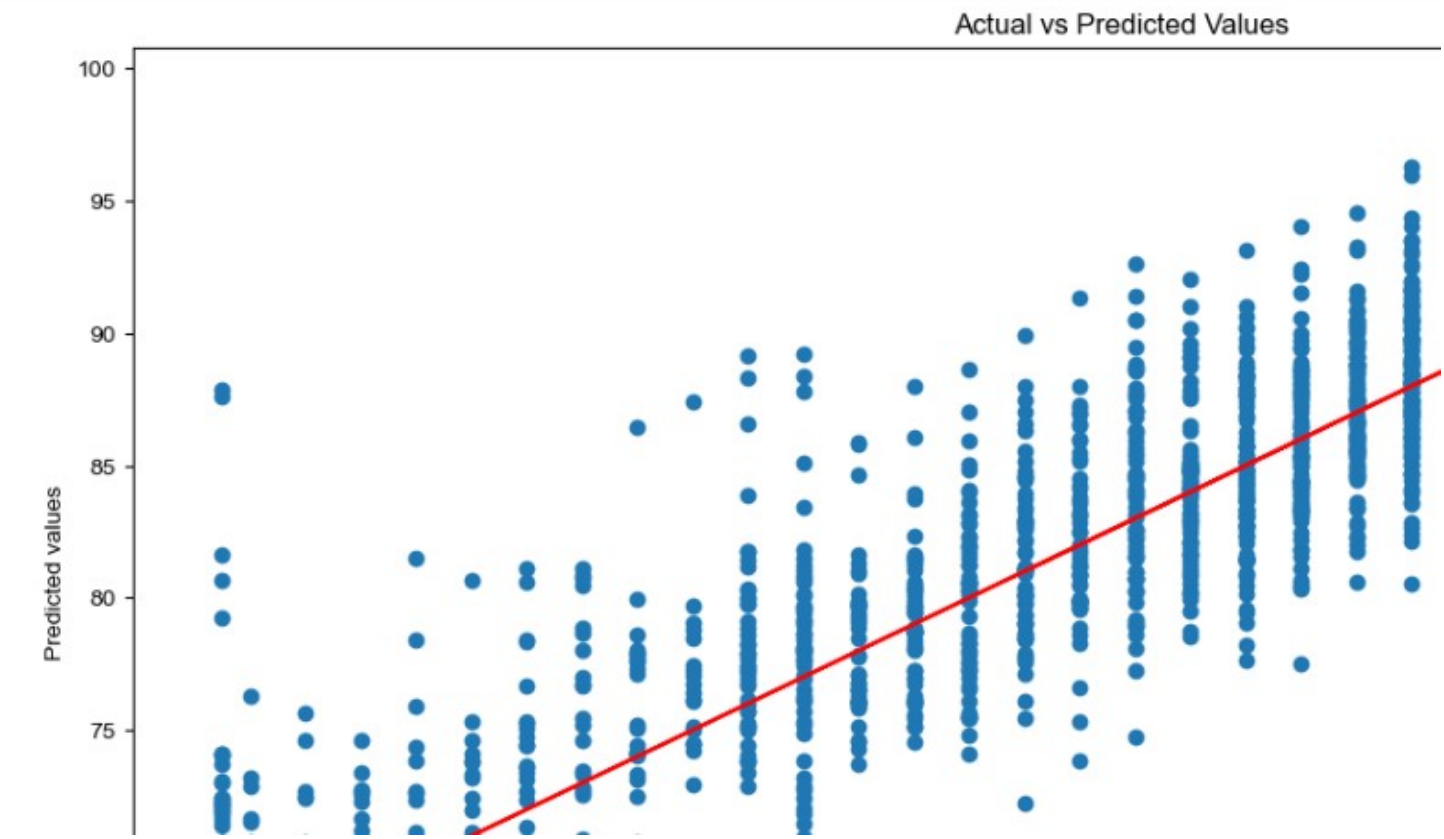
Train Data set:

MSE: 9.542146008328025

MAE: 2.2192887224943303

Actual and Predicted Curve :





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

As per the VIF, usr is affected with fork,pgin,ppgin,vflt,pflt (multicollinearity) we removed the features and linearty and homoscedasticity are satisfied because there is no pattern and p-value is low when compared to homoscedasticity value.
The above actual and predicted value are almost same they are not overlap as much.ofcourse some area might overlap in the plotted graph.

There is positive reaction for actual and predicted values.