Predictive Modelling Project

Ambrish Verma

PGP-DSBA Online

Date: 11-Nov-2023

Contents

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.

1.1 EDA, data description and Analyses: **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.**

A. Variables present in the dataframe:

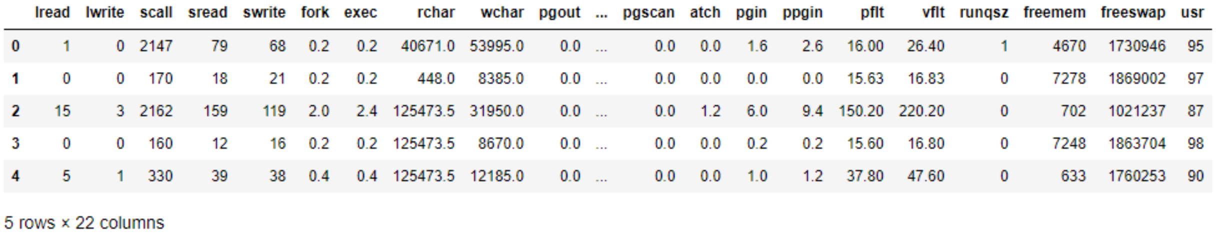
```
#   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   lread     8192 non-null   int64
 1   lwrite    8192 non-null   int64
 2   scall     8192 non-null   int64
 3   sread     8192 non-null   int64
 4   swrite    8192 non-null   int64
 5   fork      8192 non-null   float64
 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
```

There are a total of 22 columns, 8192 entries. Out of all the columns, 21 are numeric and 1 column is of string datatype.

B. Missing values: in the column ‘rchar’, there are 8088 entries which are non-null and the rest are null. In the column ‘wchar’, 8088 entries are non-null, rest are null.

C. The column ‘runqsz’ in the data frame has 2 string values: **Not_CPU_Bound** and **CPU_Bound**. They are converted in the following manner: 'CPU_Bound'=1 and 'Not_CPU_Bound'=0.

D. Null values replaced with median values. Snapshot of data after replacing nulls with median:



	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	1	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	0	7278	1869002	97
2	15	3	2162	159	119	2.0	2.4	125473.5	31950.0	0.0	...	0.0	1.2	6.0	9.4	150.20	220.20	0	702	1021237	87
3	0	0	160	12	16	0.2	0.2	125473.5	8670.0	0.0	...	0.0	0.0	0.2	0.2	15.60	16.80	0	7248	1863704	98
4	5	1	330	39	38	0.4	0.4	125473.5	12185.0	0.0	...	0.0	0.0	1.0	1.2	37.80	47.60	0	633	1760253	90

5 rows × 22 columns

E. Univariate analysis – Boxplot of all the variables in the data frame:

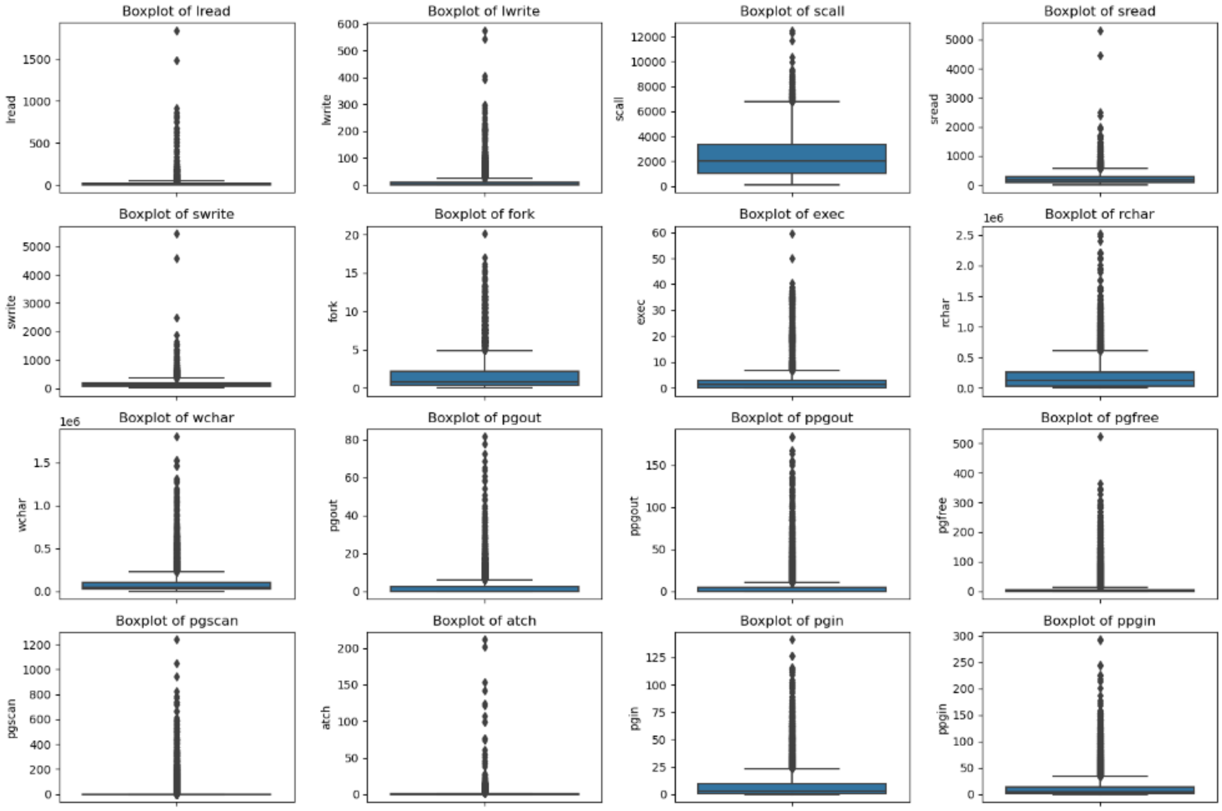


Figure 1

There are a lot of outliers in all the attributes.

F. Pair plot of all the attributes:

Since the attributes: pgout, ppgout, pgfree, pgscan, atch have the median value as 0, indicating that most of the values of these attributes = 0 . Hence, the pairplot is not going to give much information. So, dropping these attributes and then plotting the rest.

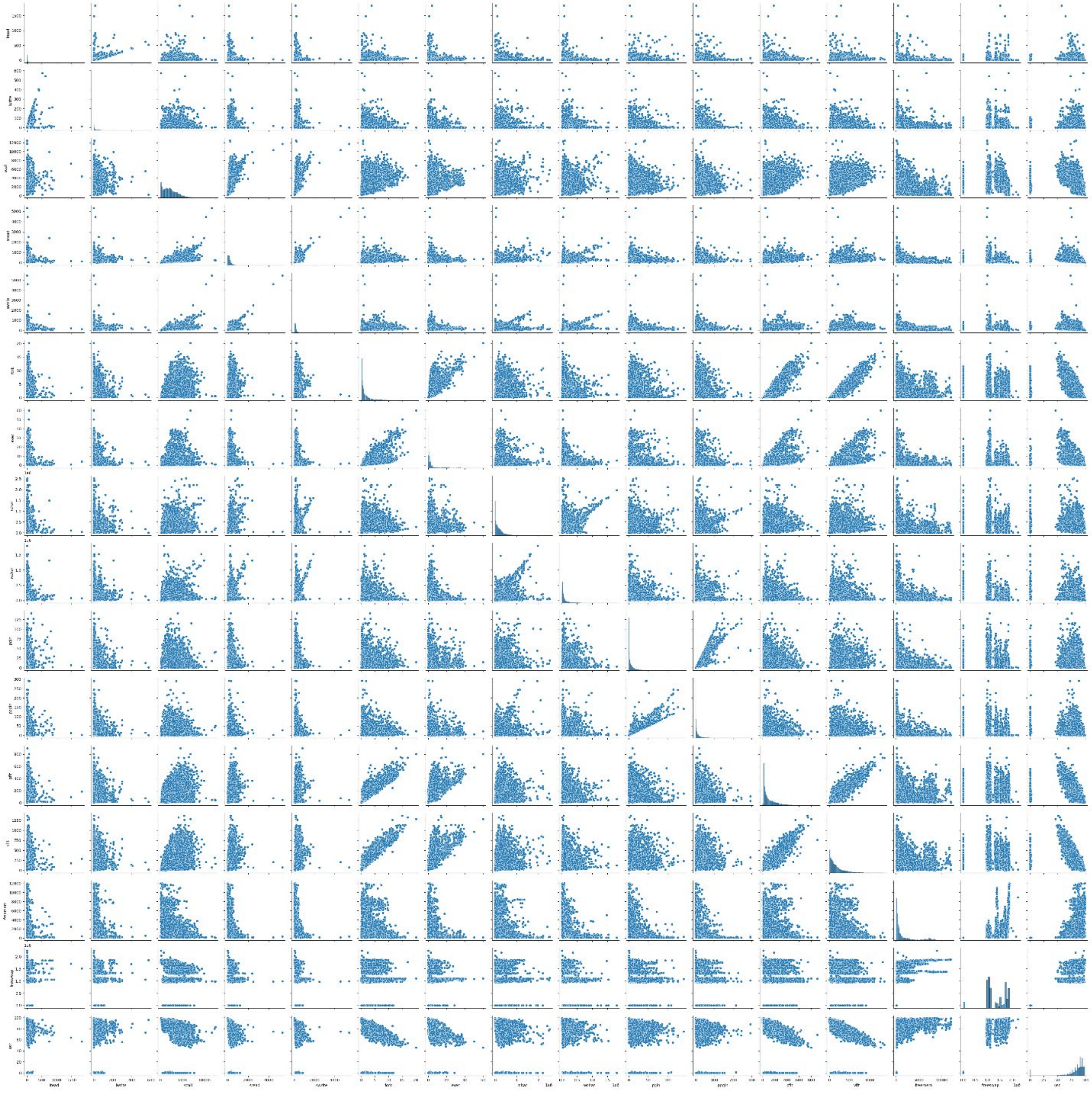
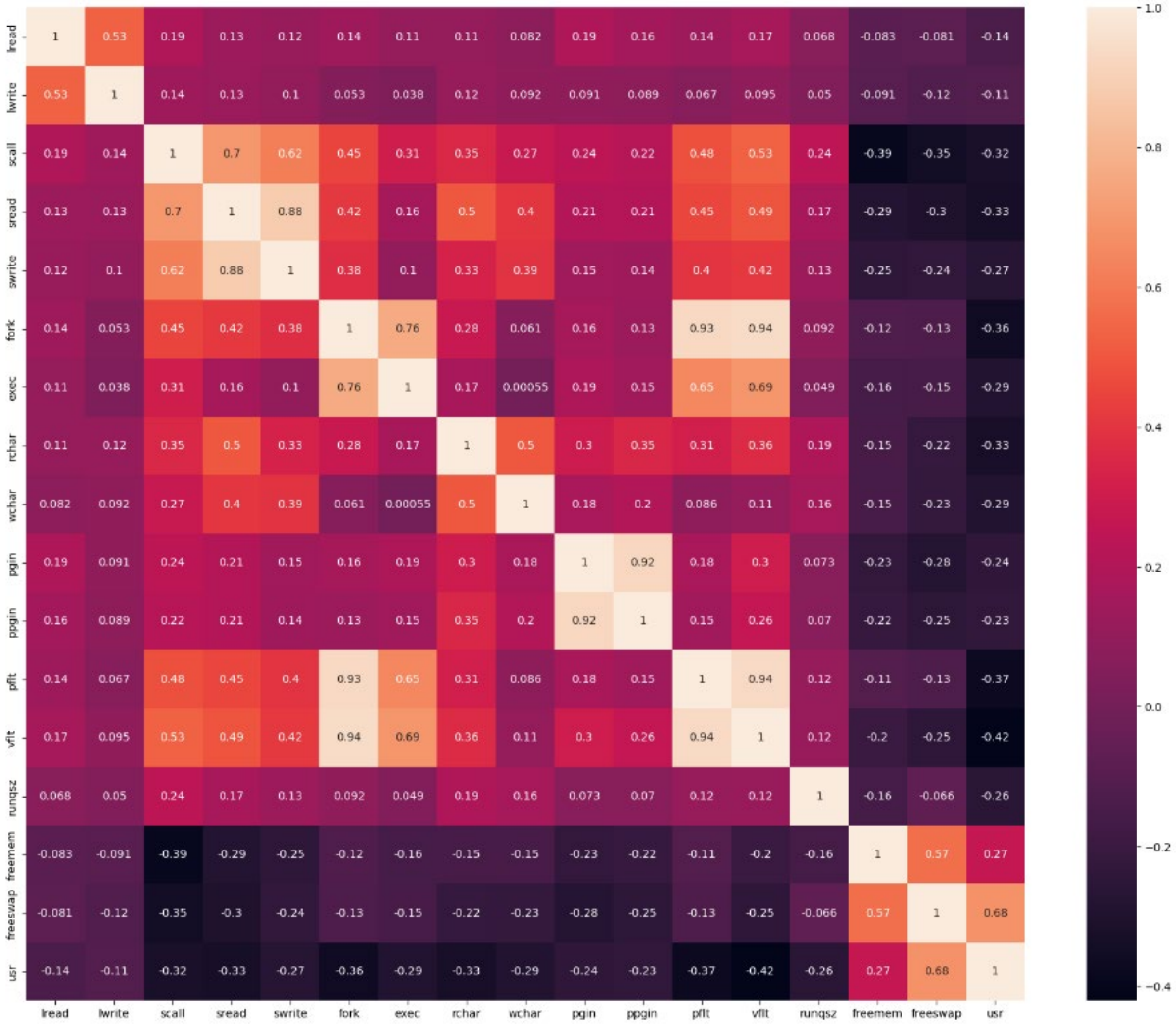


Figure 2

The correlations between attributes are better illustrated by the heatmap as shown below:



	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap	usr
lread	1	0.53	0.19	0.13	0.12	0.14	0.11	0.11	0.082	0.19	0.16	0.14	0.17	0.068	-0.083	-0.081	-0.14
lwrite	0.53	1	0.14	0.13	0.1	0.053	0.038	0.12	0.092	0.091	0.089	0.067	0.095	0.05	-0.091	-0.12	-0.11
scall	0.19	0.14	1	0.7	0.62	0.45	0.31	0.35	0.27	0.24	0.22	0.48	0.53	0.24	-0.39	-0.35	-0.32
sread	0.13	0.13	0.7	1	0.88	0.42	0.16	0.5	0.4	0.21	0.21	0.45	0.49	0.17	-0.29	-0.3	-0.33
swrite	0.12	0.1	0.62	0.88	1	0.38	0.1	0.33	0.39	0.15	0.14	0.4	0.42	0.13	-0.25	-0.24	-0.27
fork	0.14	0.053	0.45	0.42	0.38	1	0.76	0.28	0.061	0.16	0.13	0.93	0.94	0.092	-0.12	-0.13	-0.36
exec	0.11	0.038	0.31	0.16	0.1	0.76	1	0.17	0.00055	0.19	0.15	0.65	0.69	0.049	-0.16	-0.15	-0.29
rchar	0.11	0.12	0.35	0.5	0.33	0.28	0.17	1	0.5	0.3	0.35	0.31	0.36	0.19	-0.15	-0.22	-0.33
wchar	0.082	0.092	0.27	0.4	0.39	0.061	0.00055	0.5	1	0.18	0.2	0.086	0.11	0.16	-0.15	-0.23	-0.29
pgin	0.19	0.091	0.24	0.21	0.15	0.16	0.19	0.3	0.18	1	0.92	0.18	0.3	0.073	-0.23	-0.28	-0.24
ppgin	0.16	0.089	0.22	0.21	0.14	0.13	0.15	0.35	0.2	0.92	1	0.15	0.26	0.07	-0.22	-0.25	-0.23
pflt	0.14	0.067	0.48	0.45	0.4	0.93	0.65	0.31	0.086	0.18	0.15	1	0.94	0.12	-0.11	-0.13	-0.37
vflt	0.17	0.095	0.53	0.49	0.42	0.94	0.69	0.36	0.11	0.3	0.26	0.94	1	0.12	-0.2	-0.25	-0.42
runqsz	0.068	0.05	0.24	0.17	0.13	0.092	0.049	0.19	0.16	0.073	0.07	0.12	0.12	1	-0.16	-0.066	-0.26
freemem	-0.083	-0.091	-0.39	-0.29	-0.25	-0.12	-0.16	-0.15	-0.15	-0.23	-0.22	-0.11	-0.2	-0.16	1	0.57	0.27
freeswap	-0.081	-0.12	-0.35	-0.3	-0.24	-0.13	-0.15	-0.22	-0.23	-0.28	-0.25	-0.13	-0.25	-0.066	0.57	1	0.68
usr	-0.14	-0.11	-0.32	-0.33	-0.27	-0.36	-0.29	-0.33	-0.29	-0.24	-0.23	-0.37	-0.42	-0.26	0.27	0.68	1

Figure 3

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.

In the columns: rchar and wchar, there were null values present. For both these attributes, the nulls were replaced with their respective median values. By doing this, the linear regression result will have a better outcome.

The attributes: 'lwrite','fork', 'exec','pgout','ppgout','pgfree','pgscan','atch','pgin','ppgin' do not seem to have any significant correlation with 'usr' column. This is proved true in the pairplot and the correlation matrix as shown below.

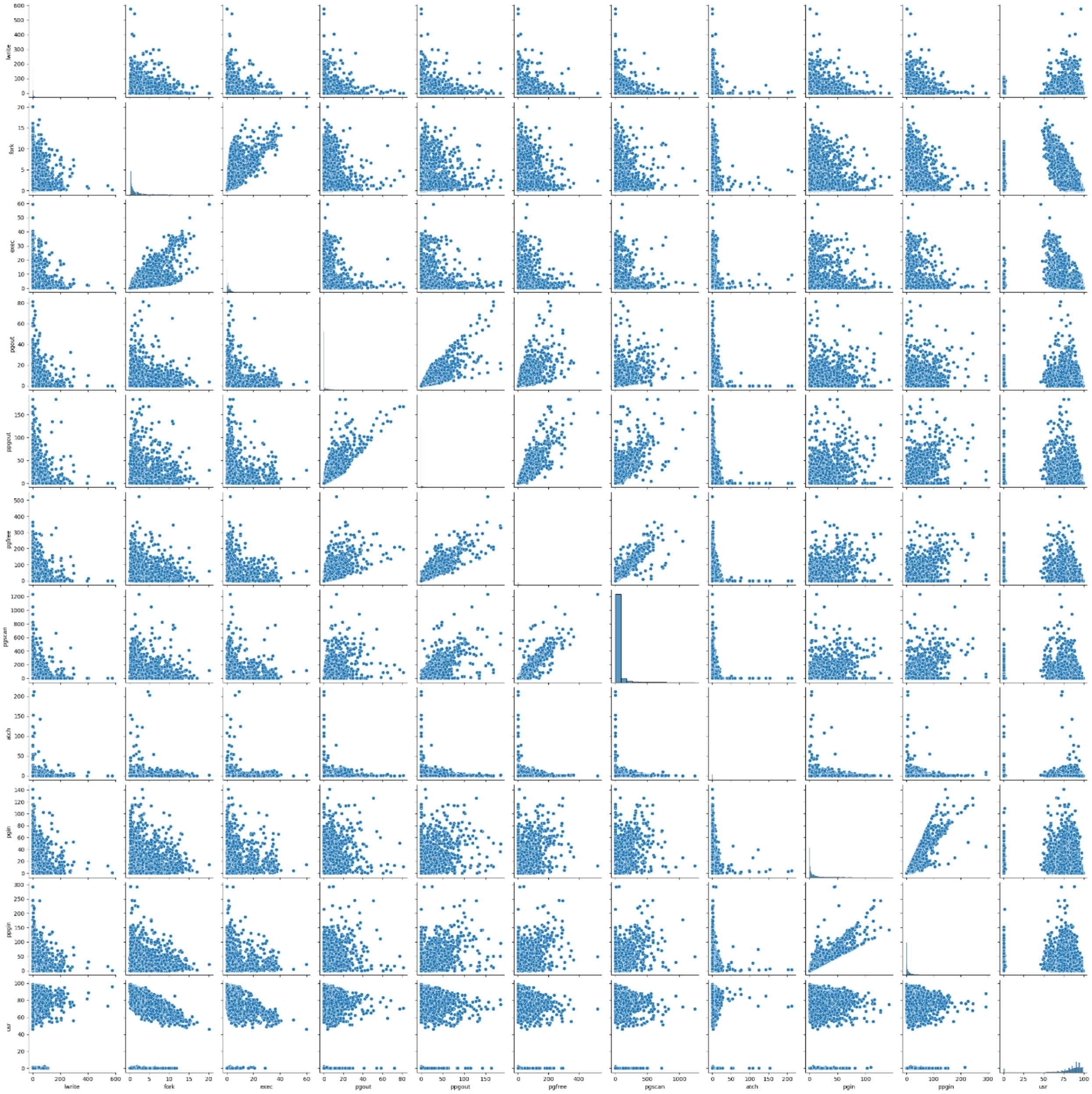
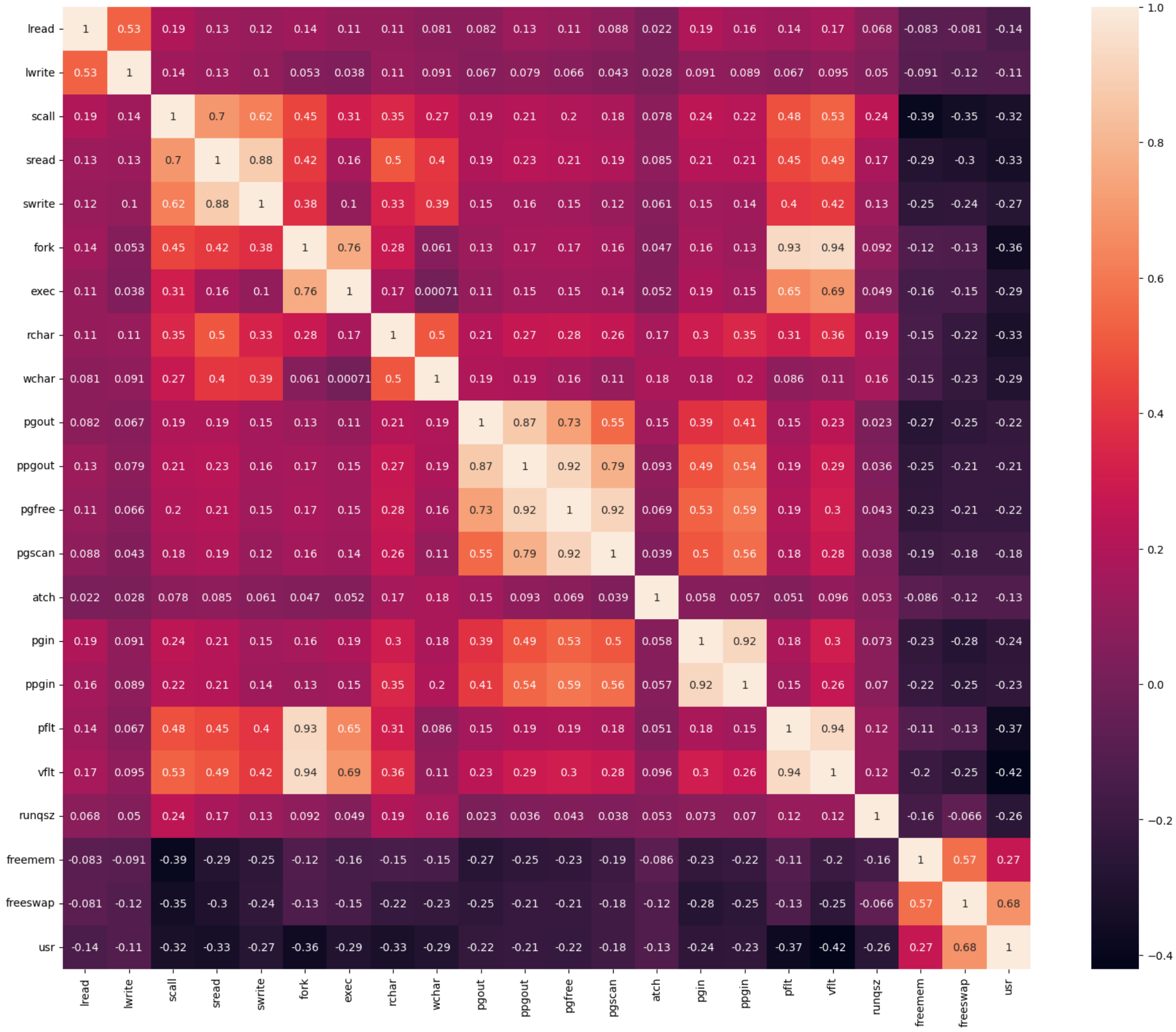

lwrite
0
100
200
300
400
500
600
fork
0
5
10
15
20
exec
0
10
20
30
40
50
60
pgout
0
20
40
60
80
ppgout
0
50
100
150
pgfree
0
100
200
300
400
500
pgscan
0
200
400
600
800
1000
1200
atch
0
50
100
150
200
pgin
0
20
40
60
80
100
120
140
ppgin
0
50
100
150
200
250
300
usr
0
20
40
60
80
100
0
200
400
600
lwrite
0
5
10
15
20
fork
0
20
40
60
exec
0
20
40
60
80
pgout
0
50
100
150
ppgout
0
200
400
pgfree
0
500
1000
pgscan
0
50
100
150
200
atch
0
50
100
pgin
0
100
200
300
ppgin
0
25
50
75
100
usr

Figure 4



	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	ppgout	pgfree	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap	usr
lread	1	0.53	0.19	0.13	0.12	0.14	0.11	0.11	0.081	0.082	0.13	0.11	0.088	0.022	0.19	0.16	0.14	0.17	0.068	-0.083	-0.081	-0.14
lwrite	0.53	1	0.14	0.13	0.1	0.053	0.038	0.11	0.091	0.067	0.079	0.066	0.043	0.028	0.091	0.089	0.067	0.095	0.05	-0.091	-0.12	-0.11
scall	0.19	0.14	1	0.7	0.62	0.45	0.31	0.35	0.27	0.19	0.21	0.2	0.18	0.078	0.24	0.22	0.48	0.53	0.24	-0.39	-0.35	-0.32
sread	0.13	0.13	0.7	1	0.88	0.42	0.16	0.5	0.4	0.19	0.23	0.21	0.19	0.085	0.21	0.21	0.45	0.49	0.17	-0.29	-0.3	-0.33
swrite	0.12	0.1	0.62	0.88	1	0.38	0.1	0.33	0.39	0.15	0.16	0.15	0.12	0.061	0.15	0.14	0.4	0.42	0.13	-0.25	-0.24	-0.27
fork	0.14	0.053	0.45	0.42	0.38	1	0.76	0.28	0.061	0.13	0.17	0.17	0.16	0.047	0.16	0.13	0.93	0.94	0.092	-0.12	-0.13	-0.36
exec	0.11	0.038	0.31	0.16	0.1	0.76	1	0.17	0.00071	0.11	0.15	0.15	0.14	0.052	0.19	0.15	0.65	0.69	0.049	-0.16	-0.15	-0.29
rchar	0.11	0.11	0.35	0.5	0.33	0.28	0.17	1	0.5	0.21	0.27	0.28	0.26	0.17	0.3	0.35	0.31	0.36	0.19	-0.15	-0.22	-0.33
wchar	0.081	0.091	0.27	0.4	0.39	0.061	0.00071	0.5	1	0.19	0.19	0.16	0.11	0.18	0.18	0.2	0.086	0.11	0.16	-0.15	-0.23	-0.29
pgout	0.082	0.067	0.19	0.19	0.15	0.13	0.11	0.21	0.19	1	0.87	0.73	0.55	0.15	0.39	0.41	0.15	0.23	0.023	-0.27	-0.25	-0.22
ppgout	0.13	0.079	0.21	0.23	0.16	0.17	0.15	0.27	0.19	0.87	1	0.92	0.79	0.093	0.49	0.54	0.19	0.29	0.036	-0.25	-0.21	-0.21
pgfree	0.11	0.066	0.2	0.21	0.15	0.17	0.15	0.28	0.16	0.73	0.92	1	0.92	0.069	0.53	0.59	0.19	0.3	0.043	-0.23	-0.21	-0.22
pgscan	0.088	0.043	0.18	0.19	0.12	0.16	0.14	0.26	0.11	0.55	0.79	0.92	1	0.039	0.5	0.56	0.18	0.28	0.038	-0.19	-0.18	-0.18
atch	0.022	0.028	0.078	0.085	0.061	0.047	0.052	0.17	0.18	0.15	0.093	0.069	0.039	1	0.058	0.057	0.051	0.096	0.053	-0.086	-0.12	-0.13
pgin	0.19	0.091	0.24	0.21	0.15	0.16	0.19	0.3	0.18	0.39	0.49	0.53	0.5	0.058	1	0.92	0.18	0.3	0.073	-0.23	-0.28	-0.24
ppgin	0.16	0.089	0.22	0.21	0.14	0.13	0.15	0.35	0.2	0.41	0.54	0.59	0.56	0.057	0.92	1	0.15	0.26	0.07	-0.22	-0.25	-0.23
pflt	0.14	0.067	0.48	0.45	0.4	0.93	0.65	0.31	0.086	0.15	0.19	0.19	0.18	0.051	0.18	0.15	1	0.94	0.12	-0.11	-0.13	-0.37
vflt	0.17	0.095	0.53	0.49	0.42	0.94	0.69	0.36	0.11	0.23	0.29	0.3	0.28	0.096	0.3	0.26	0.94	1	0.12	-0.2	-0.25	-0.42
runqsz	0.068	0.05	0.24	0.17	0.13	0.092	0.049	0.19	0.16	0.023	0.036	0.043	0.038	0.053	0.073	0.07	0.12	0.12	1	-0.16	-0.066	-0.26
freemem	-0.083	-0.091	-0.39	-0.29	-0.25	-0.12	-0.16	-0.15	-0.15	-0.27	-0.25	-0.23	-0.19	-0.086	-0.23	-0.22	-0.11	-0.2	-0.16	1	0.57	0.27
freeswap	-0.081	-0.12	-0.35	-0.3	-0.24	-0.13	-0.15	-0.22	-0.23	-0.25	-0.21	-0.21	-0.18	-0.12	-0.28	-0.25	-0.13	-0.25	-0.066	0.57	1	0.68
usr	-0.14	-0.11	-0.32	-0.33	-0.27	-0.36	-0.29	-0.33	-0.29	-0.22	-0.21	-0.22	-0.18	-0.13	-0.24	-0.23	-0.37	-0.42	-0.26	0.27	0.68	1

Figure 5

When checked for duplicates, there was no row found to be a duplicate of any other existing row.

There were significant number of outliers present in the data as shown in Figure 1.

After treatment, following are the boxplots with outliers treated:

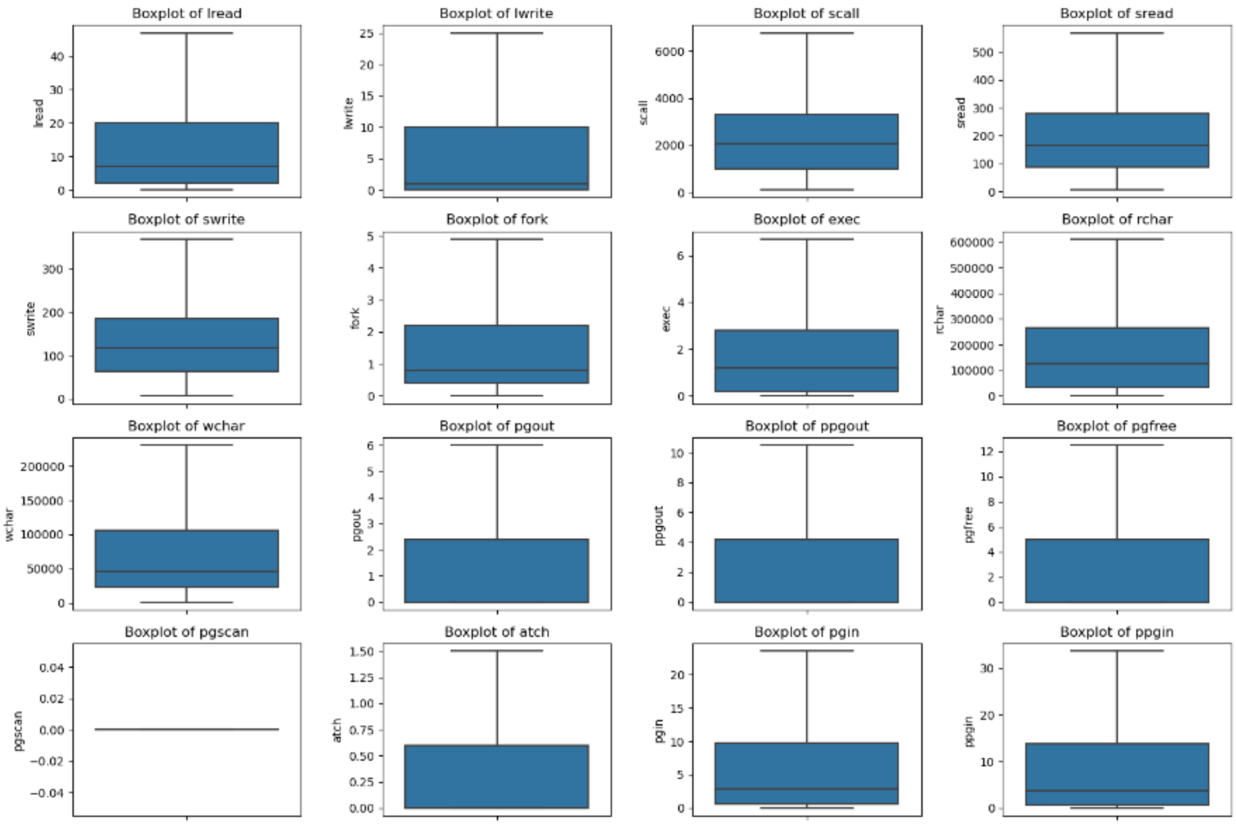


Figure 6

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.

VIF values calculated are listed below:

VIF_Values:

const	**25.663697**
lread	**5.350560**
lwrite	**4.328397**
scall	**2.960609**
sread	**6.420172**

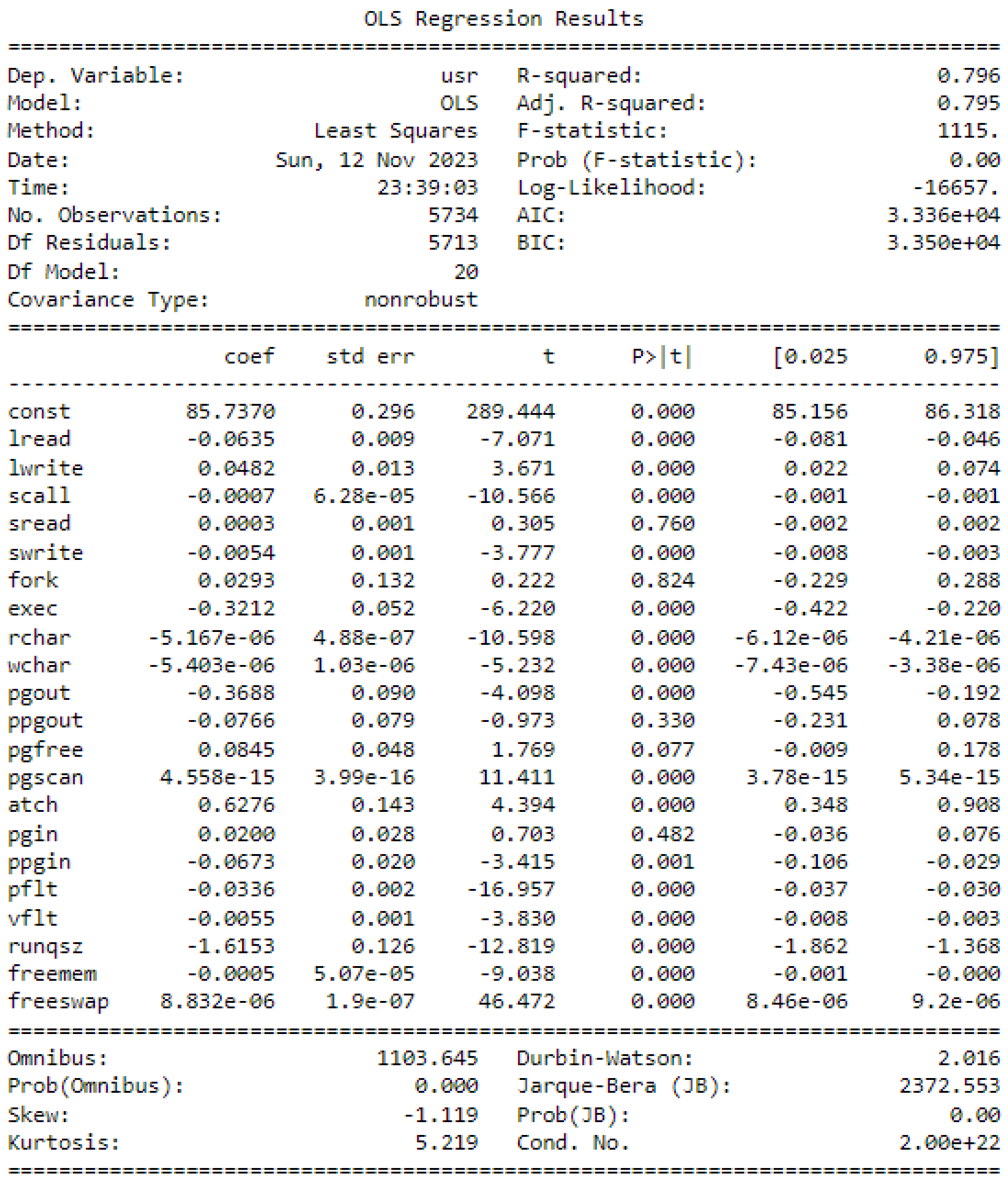
swrite	5.597135
fork	13.035359
exec	3.241417
rchar	2.133616
wchar	1.584381
pgout	11.360363
ppgout	29.404223
pgfree	16.496748
pgscan	NaN
atch	1.875901
pgin	13.809339
ppgin	13.951855
pflt	12.001460
vflt	15.971049
runqsz	1.156815
freemem	1.961304
freeswap	1.841239

The attribute : ppgout has the highest VIF. Hence dropping it and re- doing the linear regression model.

After dropping ppgout, the R-squared and Adjusted R-Squared values are:

```
R-squared: 0.796
 Adjusted R-squared: 0.795
```

The linear regression model result summary :



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.796
Model:                            OLS   Adj. R-squared:                  0.795
Method:                 Least Squares   F-statistic:                     1115.
Date:                Sun, 12 Nov 2023   Prob (F-statistic):               0.00
Time:                        23:39:03   Log-Likelihood:                -16657.
No. Observations:                5734   AIC:                         3.336e+04
Df Residuals:                    5713   BIC:                         3.350e+04
Df Model:                          20                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         85.7370      0.296    289.444      0.000      85.156      86.318
lread         -0.0635      0.009     -7.071      0.000      -0.081      -0.046
lwrite         0.0482      0.013      3.671      0.000       0.022       0.074
scall         -0.0007   6.28e-05    -10.566      0.000      -0.001      -0.001
sread          0.0003      0.001      0.305      0.760      -0.002       0.002
swrite        -0.0054      0.001     -3.777      0.000      -0.008      -0.003
fork           0.0293      0.132      0.222      0.824      -0.229       0.288
exec          -0.3212      0.052     -6.220      0.000      -0.422      -0.220
rchar      -5.167e-06   4.88e-07    -10.598      0.000   -6.12e-06   -4.21e-06
wchar      -5.403e-06   1.03e-06     -5.232      0.000   -7.43e-06   -3.38e-06
pgout         -0.3688      0.090     -4.098      0.000      -0.545      -0.192
ppgout        -0.0766      0.079     -0.973      0.330      -0.231       0.078
pgfree         0.0845      0.048      1.769      0.077      -0.009       0.178
pgscan      4.558e-15   3.99e-16     11.411      0.000    3.78e-15    5.34e-15
atch           0.6276      0.143      4.394      0.000       0.348       0.908
pgin           0.0200      0.028      0.703      0.482      -0.036       0.076
ppgin         -0.0673      0.020     -3.415      0.001      -0.106      -0.029
pflt          -0.0336      0.002    -16.957      0.000      -0.037      -0.030
vflt          -0.0055      0.001     -3.830      0.000      -0.008      -0.003
runqsz        -1.6153      0.126    -12.819      0.000      -1.862      -1.368
freemem       -0.0005   5.07e-05     -9.038      0.000      -0.001      -0.000
freeswap    8.832e-06    1.9e-07     46.472      0.000    8.46e-06     9.2e-06
==============================================================================
Omnibus:                     1103.645   Durbin-Watson:                   2.016
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2372.553
Skew:                          -1.119   Prob(JB):                         0.00
Kurtosis:                       5.219   Cond. No.                     2.00e+22
==============================================================================
```

Figure 7

Problem 2: Logistic Regression, LDA and CART

Summary:

Data Dictionary:

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

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

Below is the data type and the number of rows present in the dataset:

```
#   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Wife_age                   1402 non-null   float64
 1   Wife_ education            1473 non-null   object
 2   Husband_education          1473 non-null   object
 3   No_of_children_born        1452 non-null   float64
 4   Wife_religion              1473 non-null   object
 5   Wife_Working               1473 non-null   object
 6   Husband_Occupation         1473 non-null   int64
 7   Standard_of_living_index   1473 non-null   object
 8   Media_exposure             1473 non-null   object
 9   Contraceptive_method_used  1473 non-null   object
```

There are a total of 80 duplicated rows present in the dataset.

Wife_education, Husband_education, Wife_religion, Wife_Working, Standard_of_living_index, Media_exposure, Contraceptive_method_used are all categorical variables and can be converted to dummy variables using various encoding techniques.