PM PROJECT SAMPLE

REPORT

DSBA

yashyathish850@gmail.com
M4I8E3AP02

Disclaimer: This document will act just as a reference for the learners on how the format of business report for submission is expected. The questions in this sample report might differ from the actual questions in the project.

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.

Dataset for Problem 1: compactiv.xlsx

DATA DICTIONARY:

System measures used:

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

CPU-bound.)
freemem - Number of memory pages available to user processes
freeswap - Number of disk blocks available for page swapping.
---
usr - Portion of time (%) that cpus run in user mode

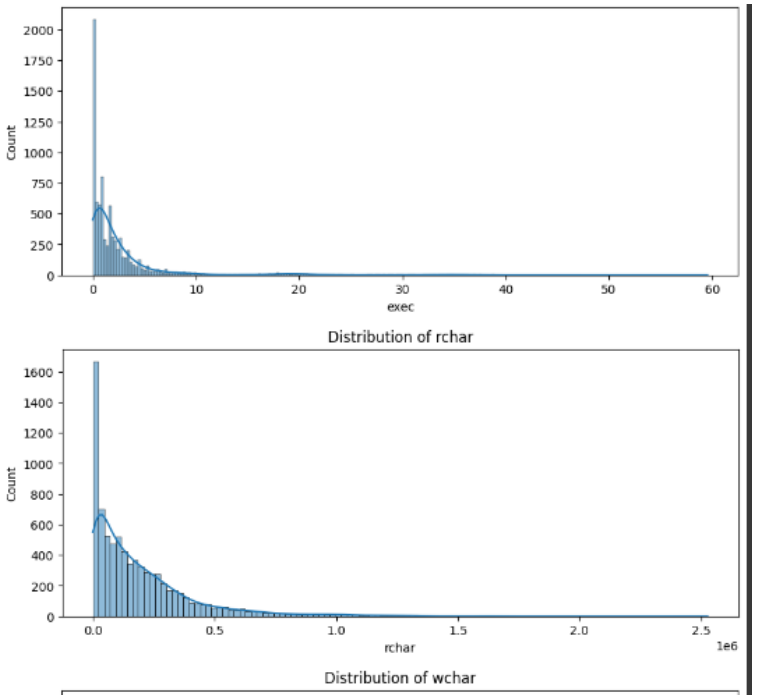
1.1 1.1 Read the data and do exploratory data analysis. Describe the data briefly. (Check the Data types, shape, EDA, 5 point summary). Perform Univariate, Bivariate Analysis, Multivariate Analysis.

```
<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
```

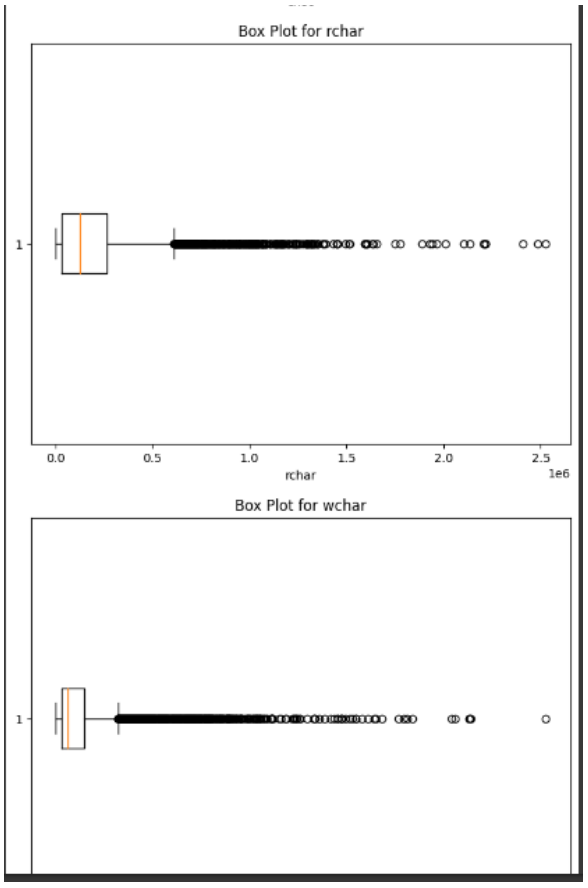
- This data has 8192 rows and 21 columns.
- Some columns, such as "rchar" and "wchar" contain missing values (NaN) because the "Non-Null Count" is less than the total number of rows (8192).
- The data has **13 float, 8 int and 1 object data types**

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.

Imputing Null values:



"rchar" and "wchar" has null values and are highly right skewed, as you can see from the above graph, so we impute these values with median values.The median is less sensitive to outliers compared to the mean, which makes it a robust option for imputing missing values.

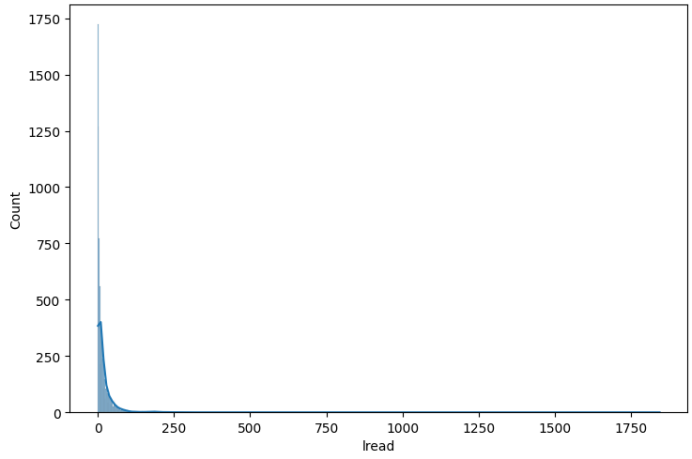


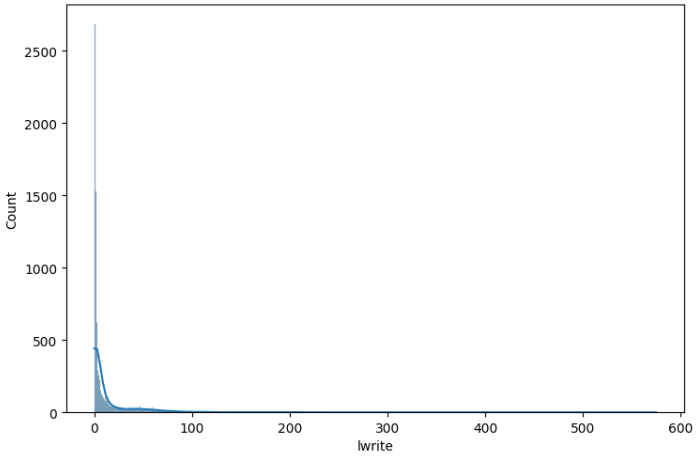
I have replaced the missing values with median and was able to retain the data

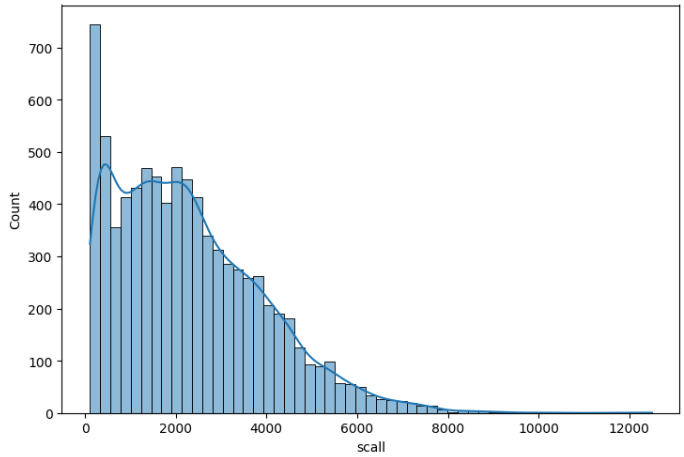
The zeros were not dropped as it makes sense in this data and was retained.
The outliers were treated using IQR method, which helped to remove extreme values without losing any significant data. And there are no duplicate rows in your DataFrame.

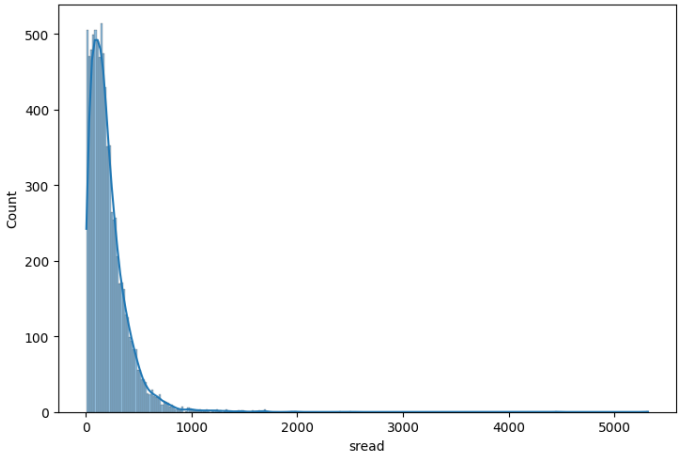
Now our data ready for linear regression

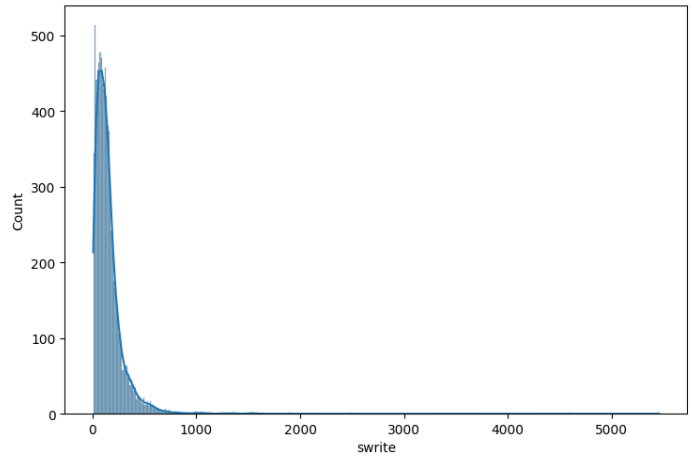
Univariate Analysis-

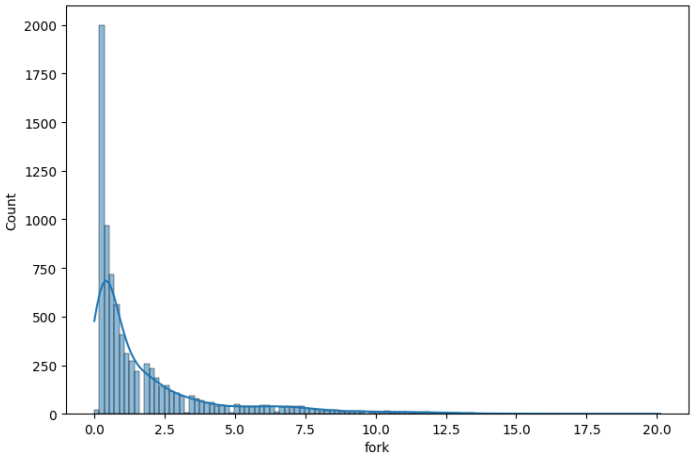


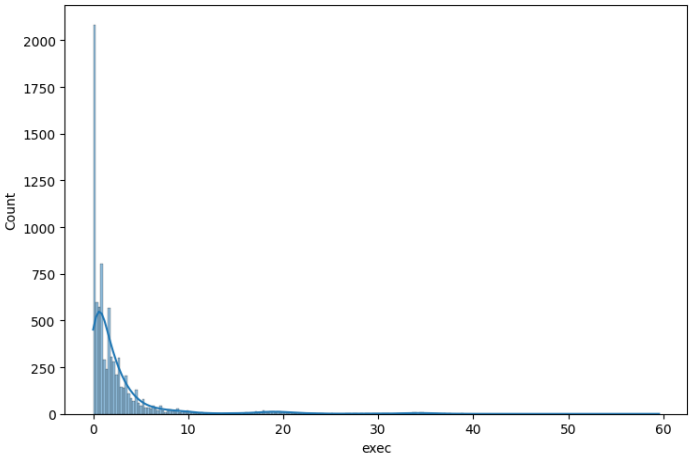


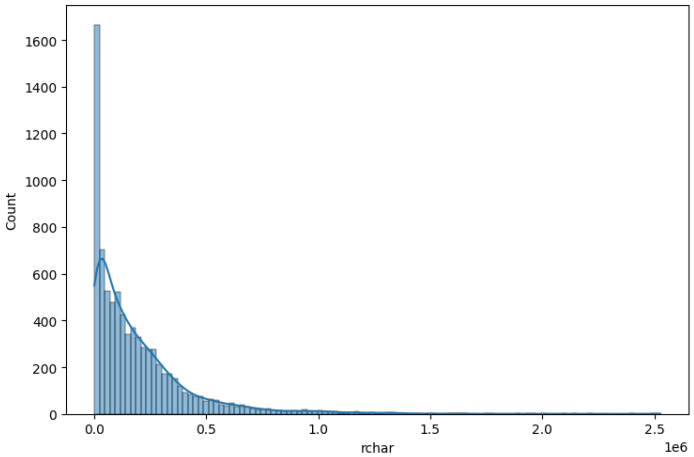


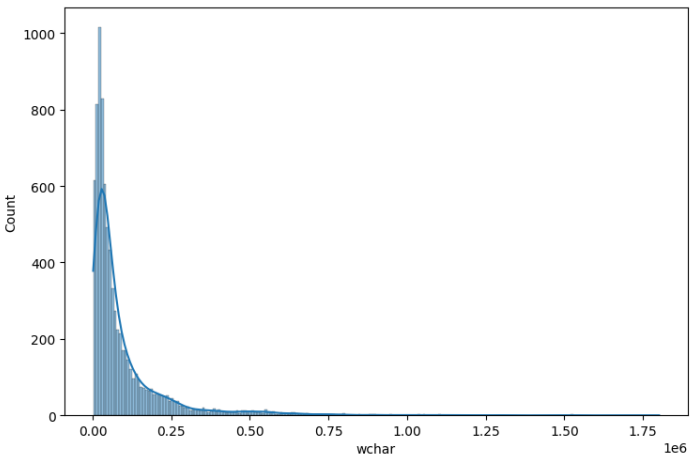


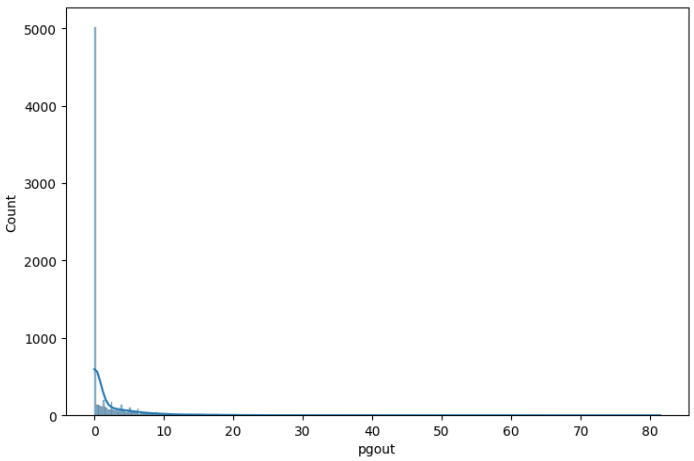

Count
swrite

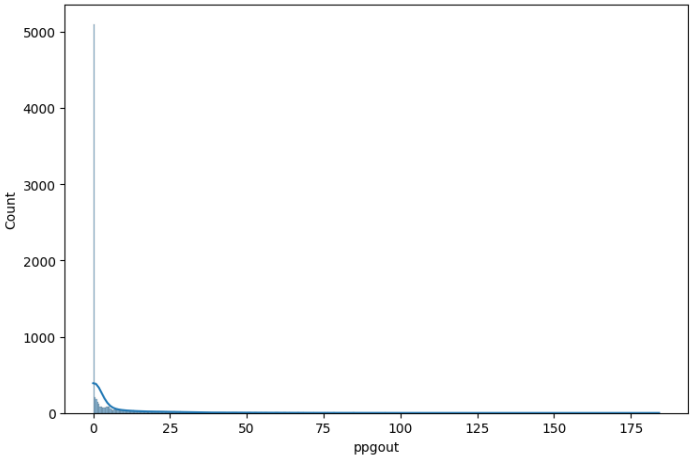

Count
fork

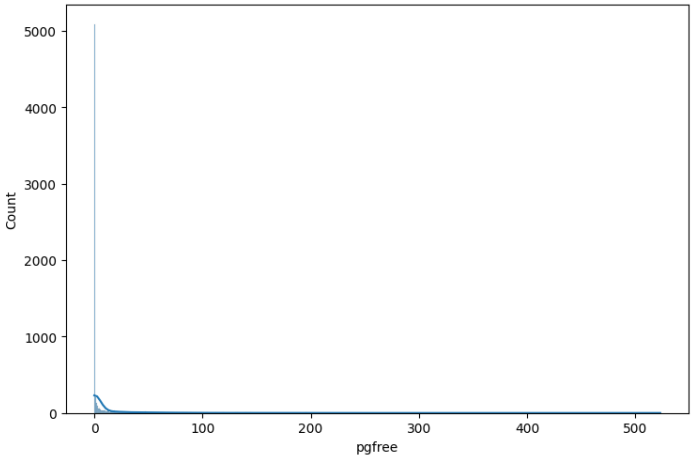

Count
exec

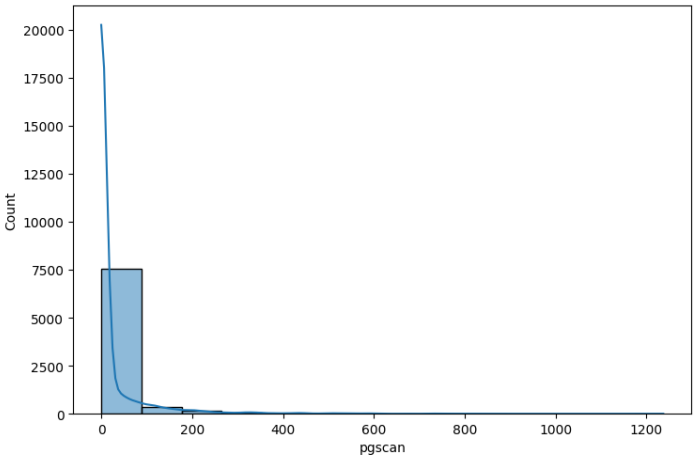

Count
rchar
1e6

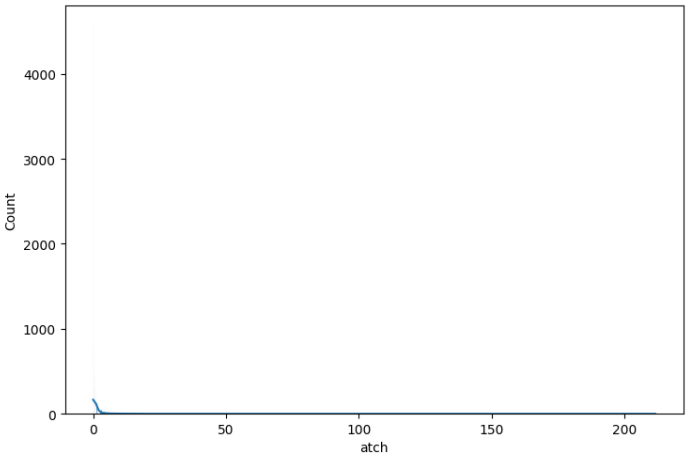

Count
wchar
1e6

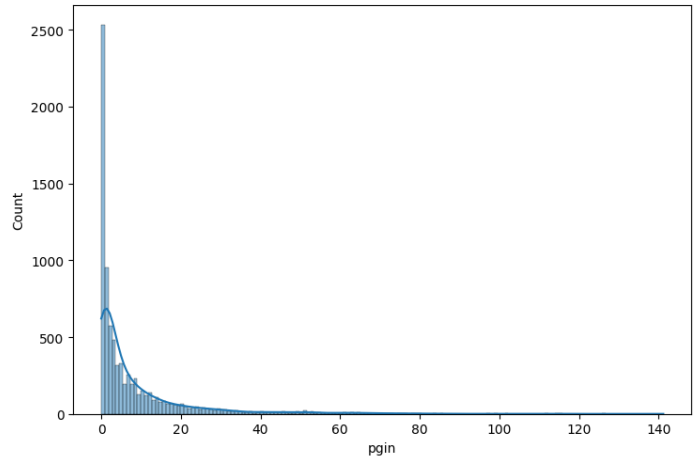

Count
pgout

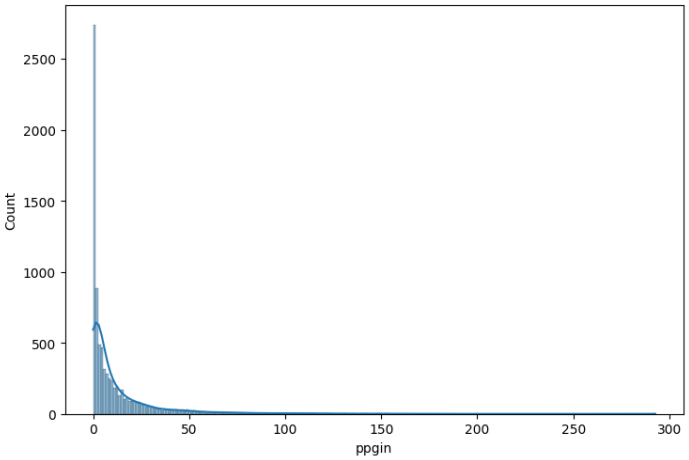

Count
ppgout

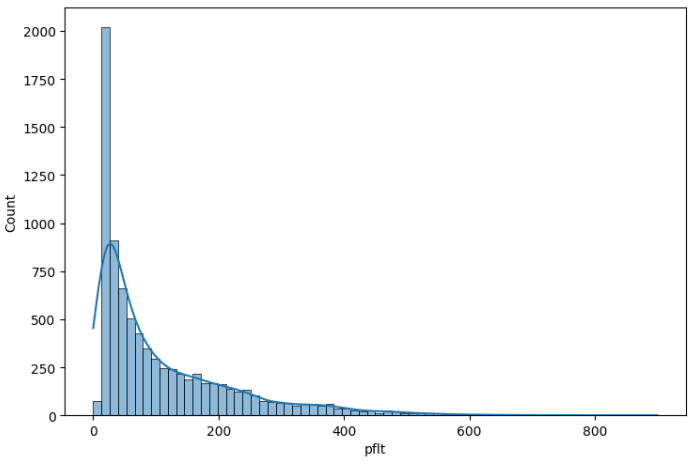

Count
pgfree

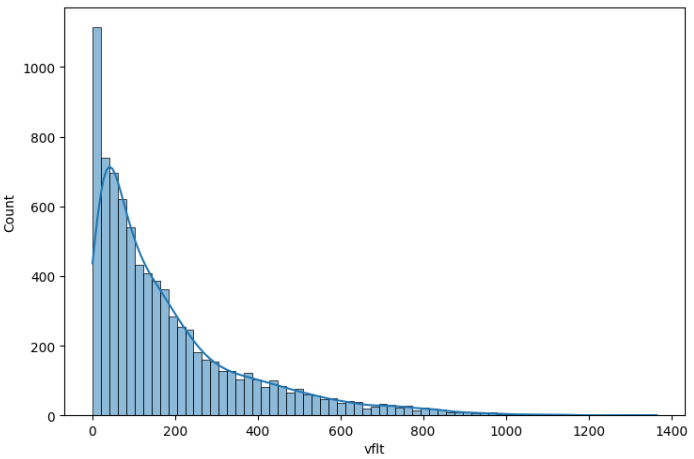

Count
pgscan

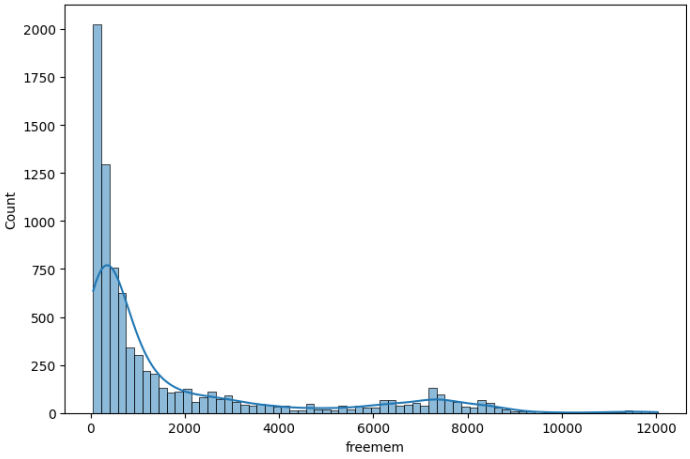

Count
atch

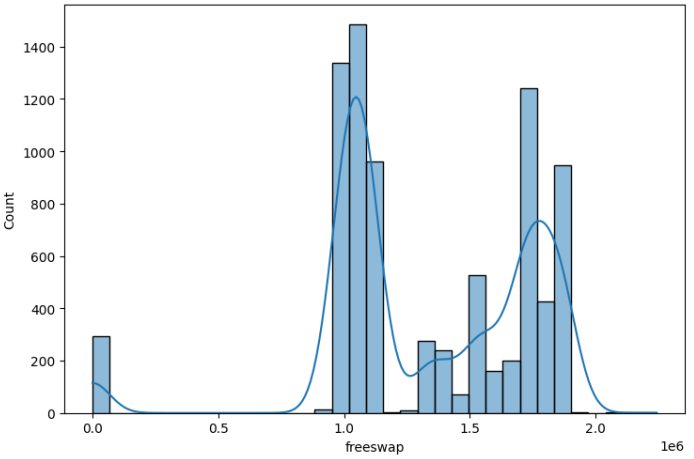

Count
pgin

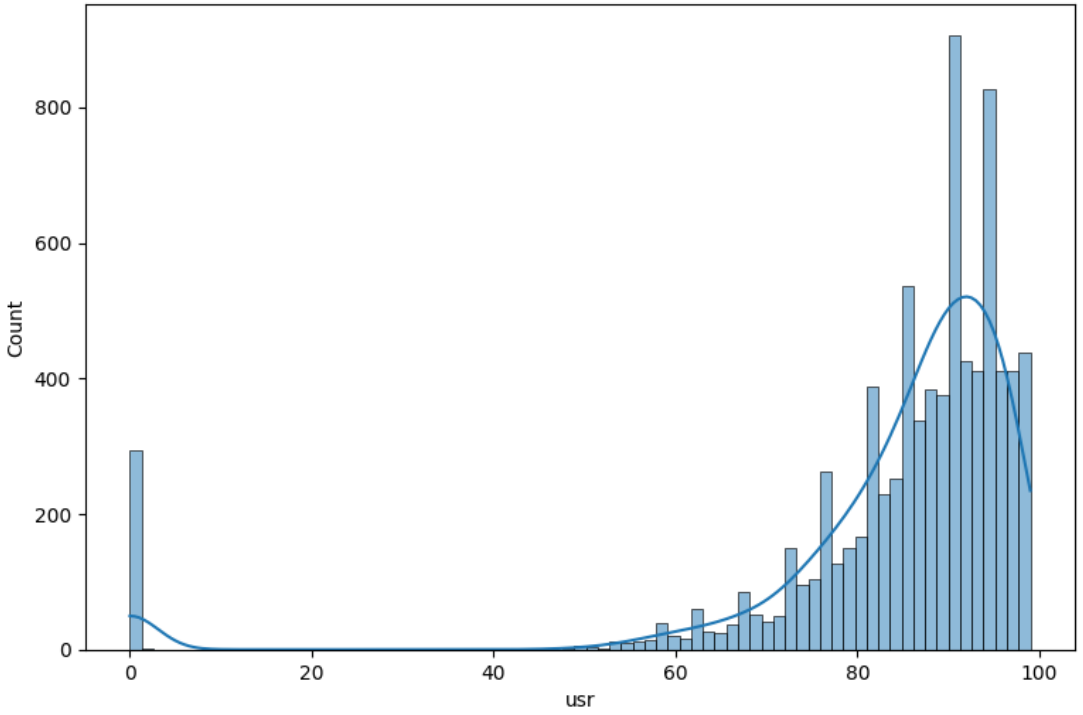

Count
ppgin

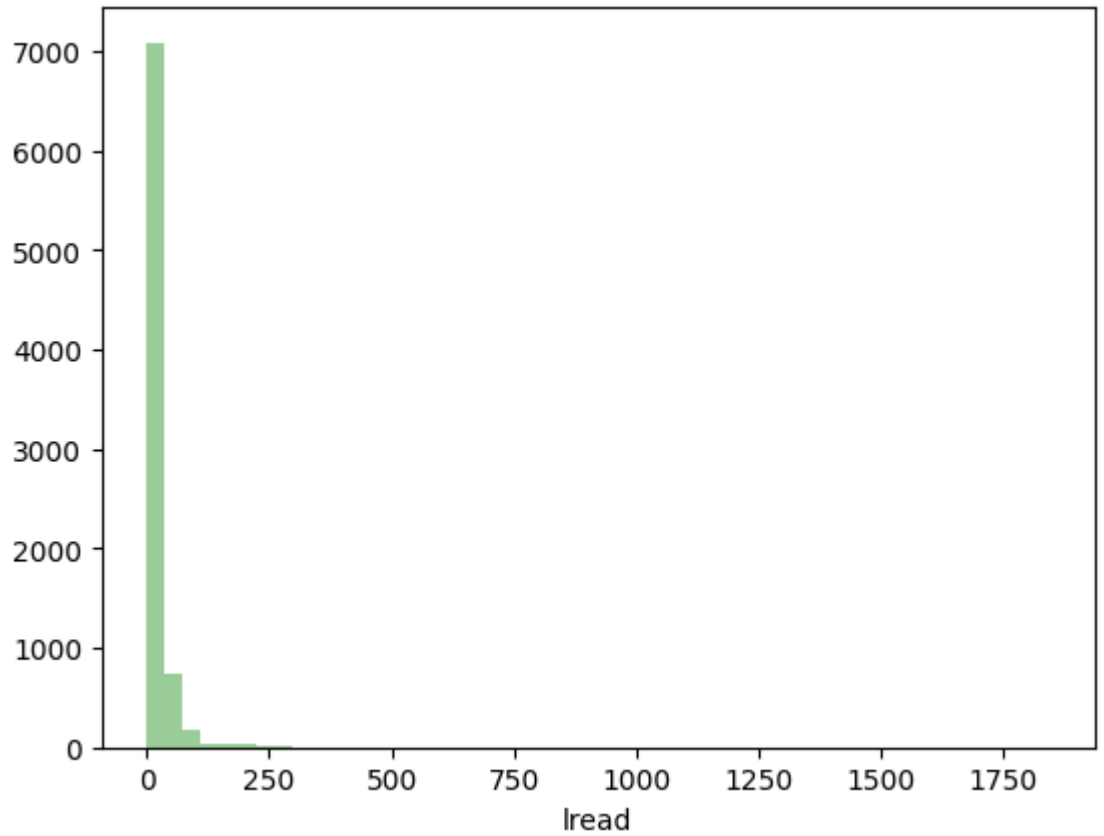

Count
pflt

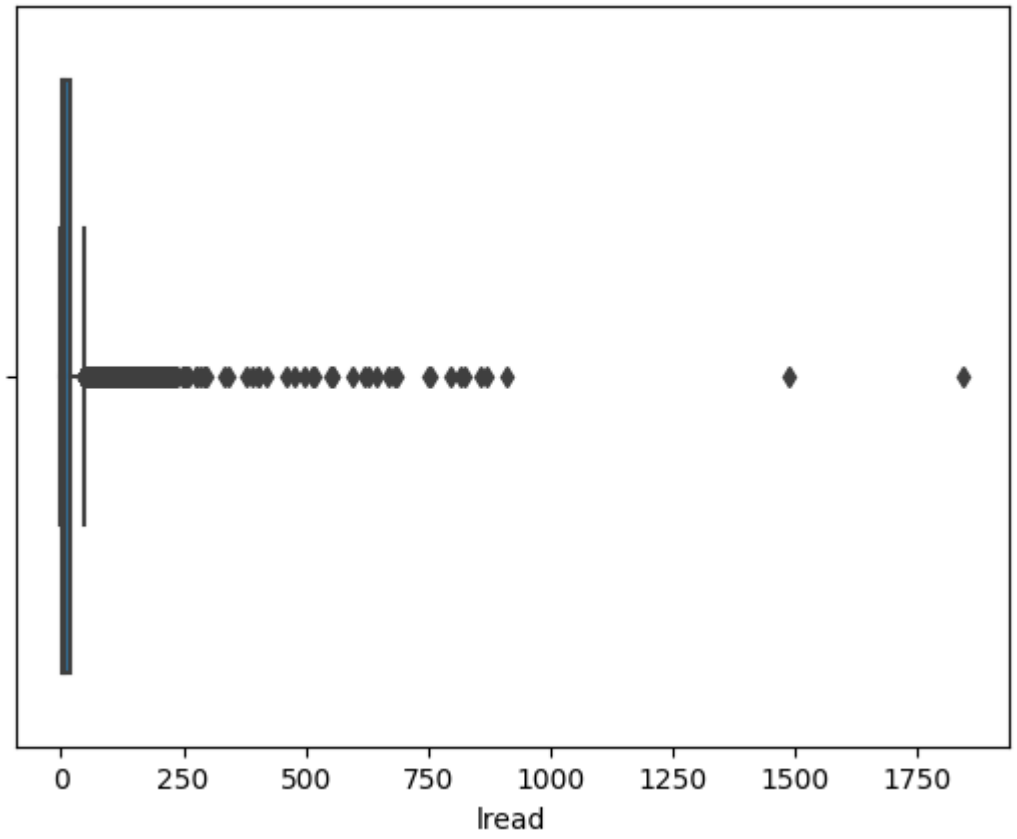

Count
vflt

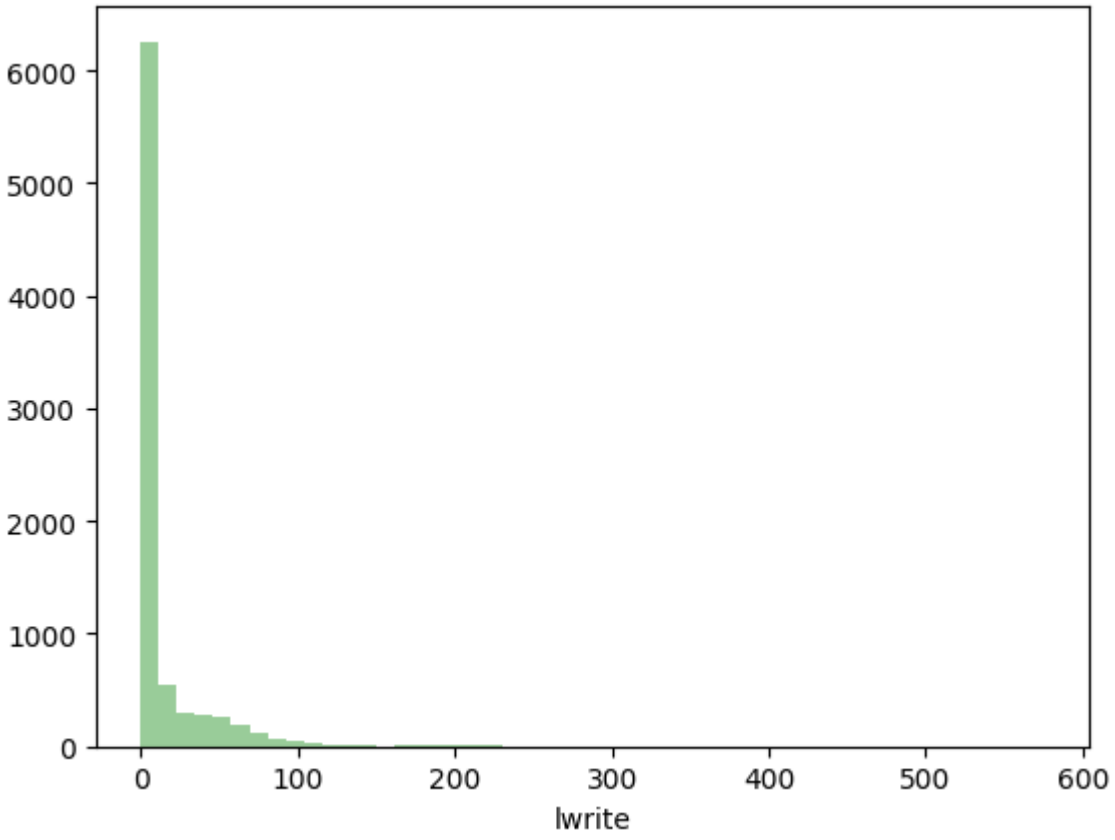

Count
freemem

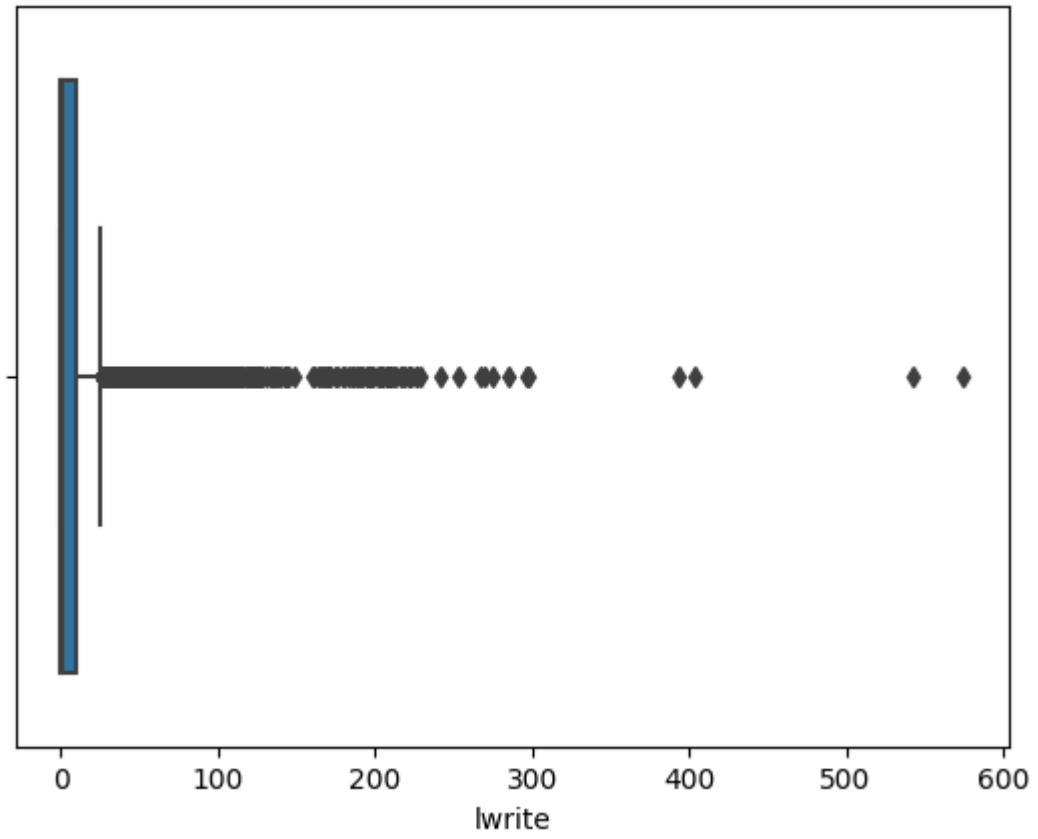

Count
freeswap
1e6

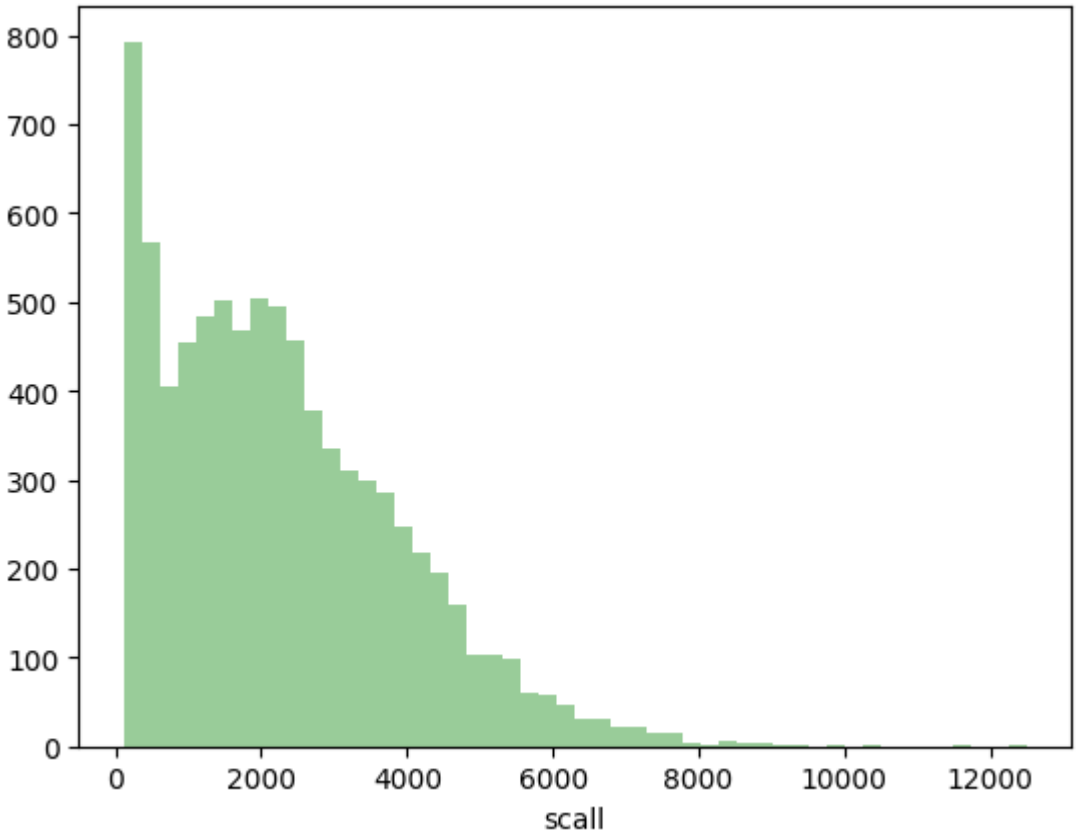

800
600
400
200
0
Count
0
20
40
60
80
100
usr

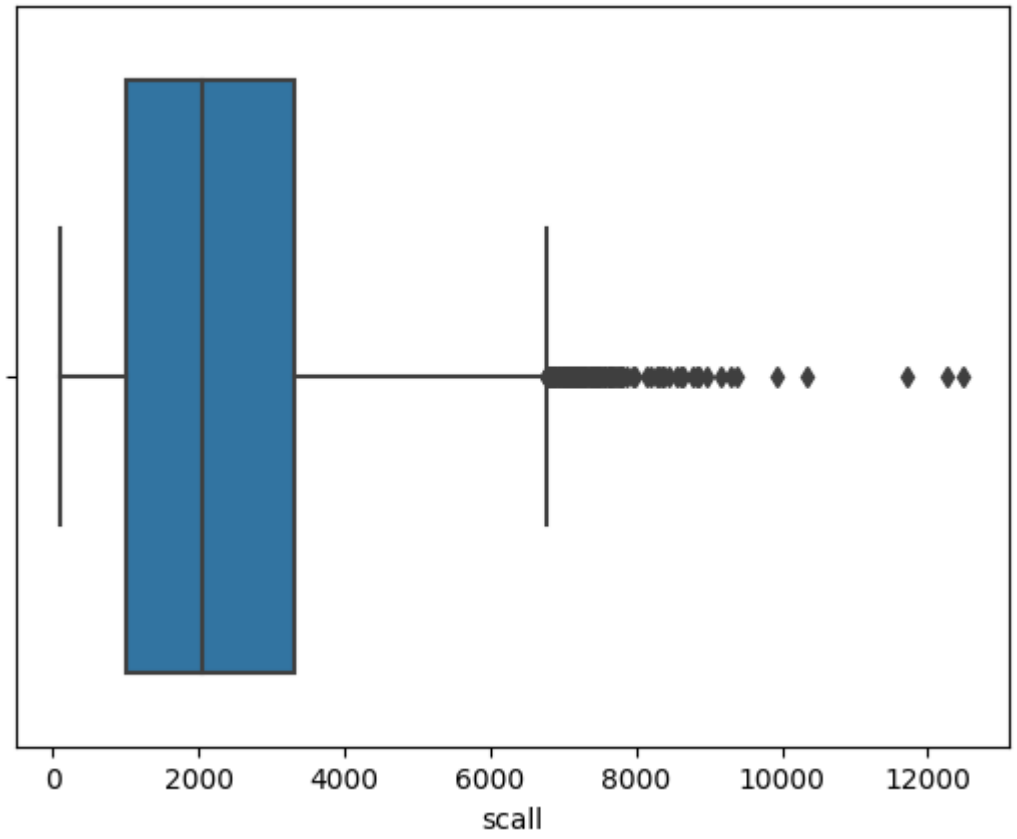

7000
6000
5000
4000
3000
2000
1000
0
0
250
500
750
1000
1250
1500
1750
lread

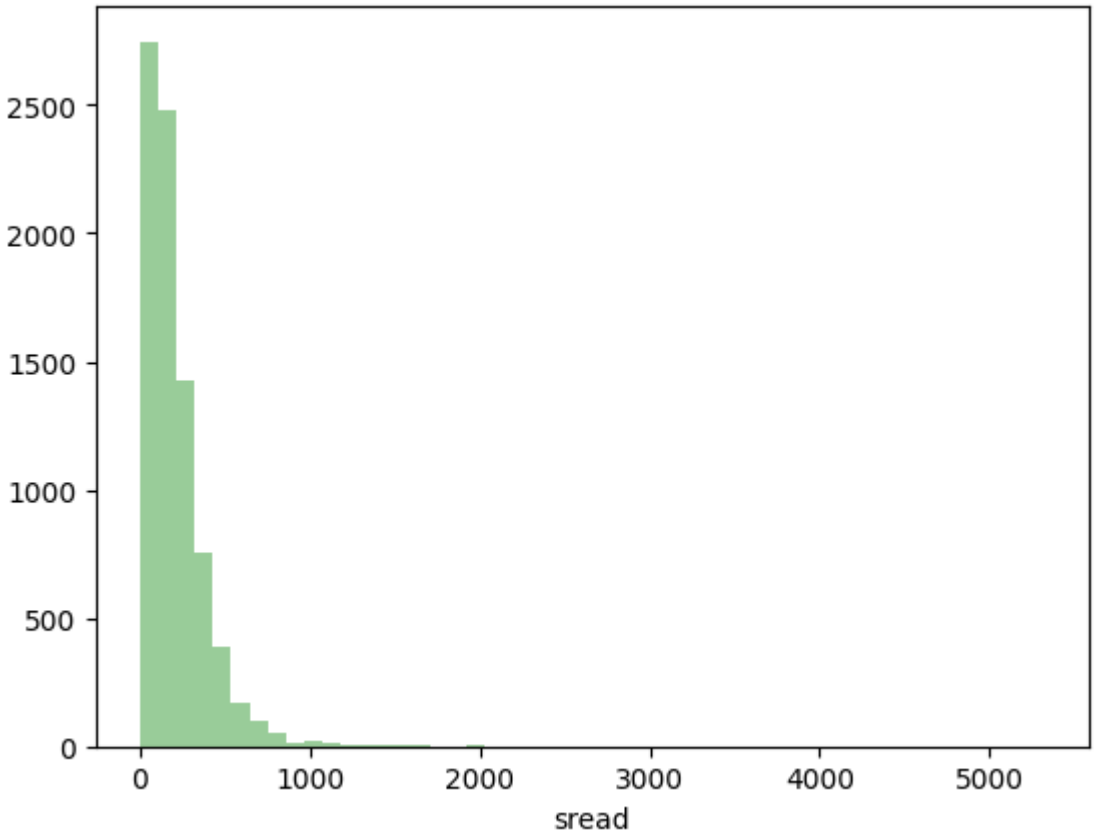

0
250
500
750
1000
1250
1500
1750
lread

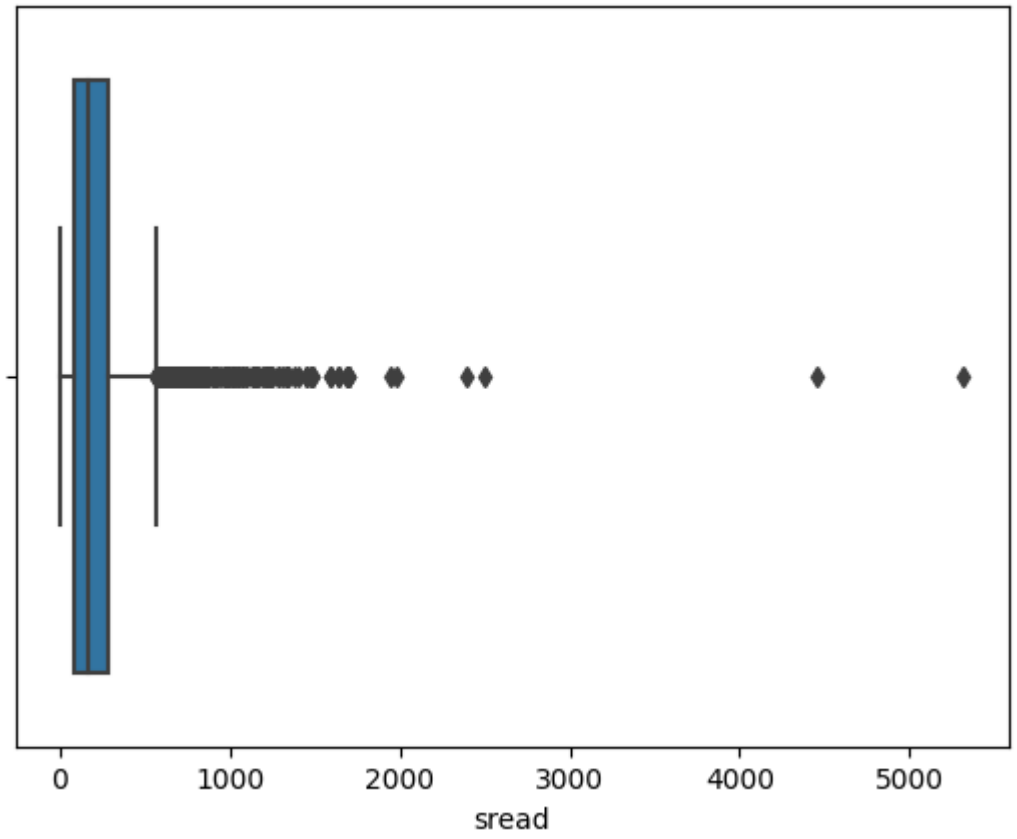

6000
5000
4000
3000
2000
1000
0
0
100
200
300
400
500
600
lwrite


0
100
200
300
400
500
600
lwrite


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

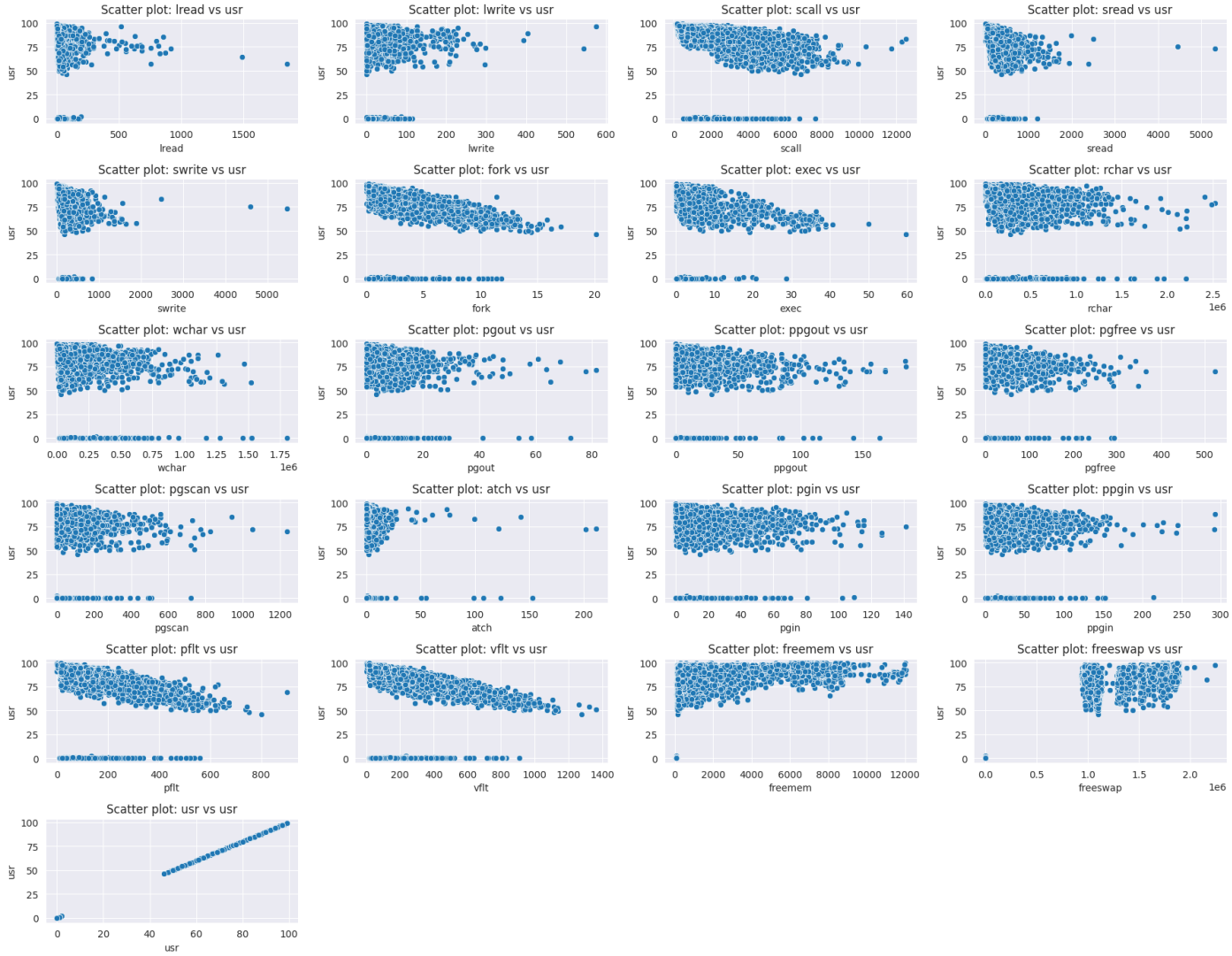

0
2000
4000
6000
8000
10000
12000
scall


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

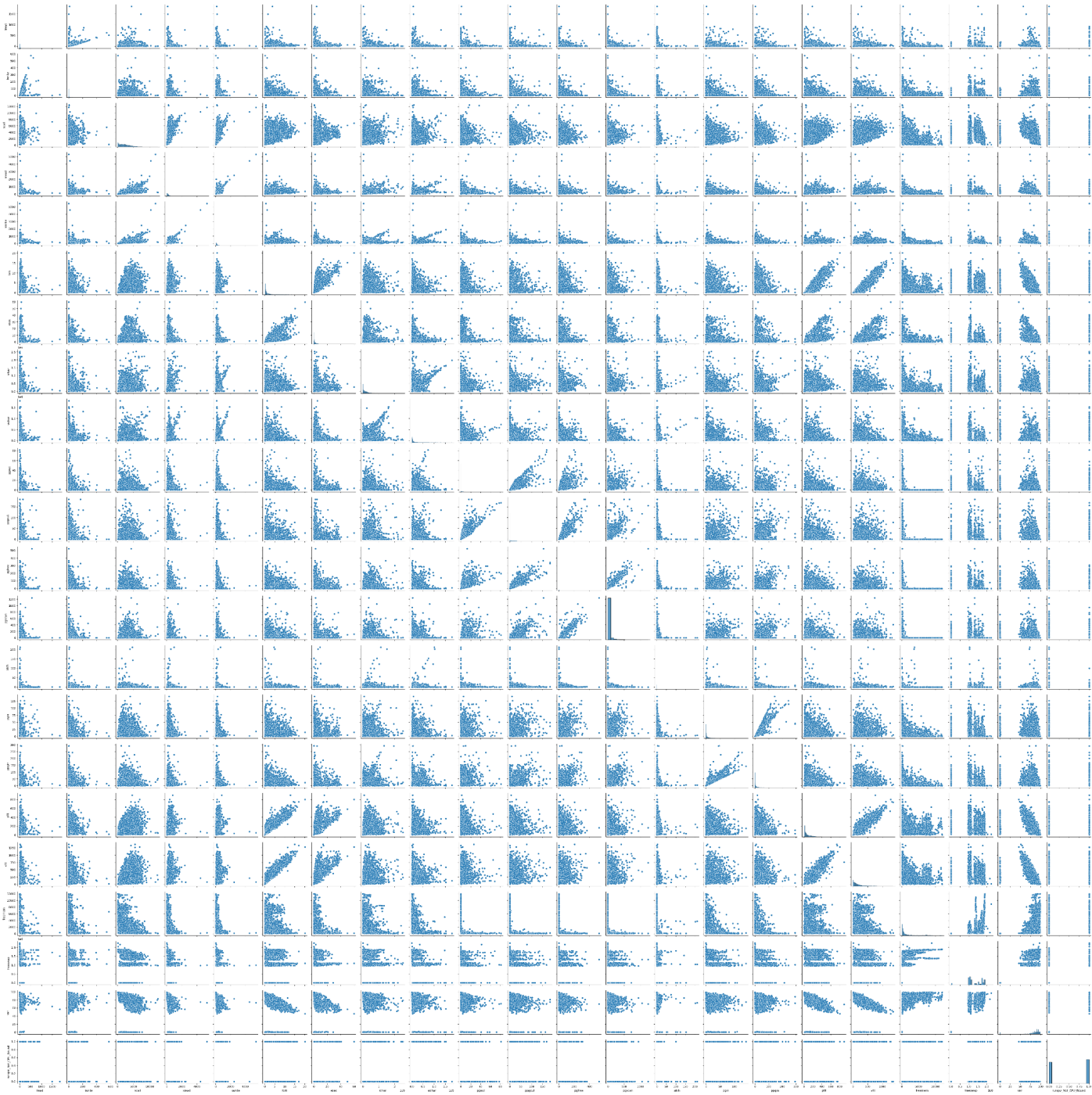


- The dataset does not contain any duplicate records.
- The 'rchar' feature has 104 missing values.
- There are 15 missing records in the 'wchar' feature.
- Most of the features exhibit positive skewness, indicating that their distributions are skewed to the right.
- The presence of long tails on the right side of the distributions suggests the existence of outliers on the higher end of the data.
- Specifically, the 'usr' feature is left-skewed.
- The presence of outliers is quite evident in the dataset's features, and addressing them is essential as regression analysis is sensitive to the influence of outliers.

Bivariate Analysis:



Multivariate Analysis:



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.

Encoding the data:

runqsz_Not_CPU_Bound
0
1
1
1
1
...
0
1
1
0
0

'runqsz' column had categorical values - CPU_Bound and Not_CPU_Bound, which was encoded using one hot encoding method and dropped the 1st row.
'runqsz' is encoded as runqsz_Not_CPU_Bound, 0 indicating CPU_Bound and 1 indicating Not_CPU_Bound.

The dependent and independent values were separated and split into 70:30 using train_test_split method

Linear Regression model using Statsmodel (OLS):
Ordinary Least Square method

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, 22 Oct 2023	Prob (F-statistic):	0.00
Time:	08:32:29	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]
Intercept	84.1217	0.316	266.106	0.000	83.502	84.741
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	3.325e-14	1.46e-16	228.488	0.000	3.3e-14	3.35e-14
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_Not_CPU_Bound	1.6153	0.126	12.819	0.000	1.368	1.862
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.	1.09e+22

Observations:

- Explained Variance: The R-squared value measures how much of the variability in the dependent variable ('usr') can be predicted by the independent variables. In this case, approximately 79.6% of the 'usr' variance is accounted for by these predictors.

- Model Fit: The high adjusted R-squared value of 79.5% indicates that the model fits the data quite well, suggesting that it captures the relationship between the independent variables and 'usr.'

- Intercept: The constant term (intercept) is approximately 84.1217. It represents the expected 'usr' value when all independent variables are zero.

- Significant Predictors: Notably, 'lread,' 'lwrite,' 'scall,' 'exec,' 'rchar,' 'wchar,' 'pgout,' 'atch,' 'pflt,' 'vflt,' 'freemem,' 'freeswap,' and 'runqsz_Not_CPU_Bound' have substantial impacts on 'usr.' This is evident from their very low P-values.

- Confidence Intervals: The [0.025 - 0.975] columns specify 95% confidence intervals for each coefficient. They indicate the range in which the true population parameter is likely to exist.

- Overall Model Significance: The F-statistic tests the significance of the entire regression model. A high F-statistic and a very small Prob (F-statistic) value (near 0) imply that the entire model is statistically significant in explaining 'usr.'

- Autocorrelation Check: With a Durbin-Watson value of approximately 2.016, there is no strong autocorrelation detected in the model's residuals. This indicates that the residuals are relatively independent from one another.

- Multicollinearity Concerns: The high condition number raises concerns about multicollinearity. Further analysis of Variance Inflation Factors (VIF) for predictors is necessary to identify variables contributing to multicollinearity and how to address them.
  Variance Inflation Factor score for predictors:

	Variable	VIF
10	ppgout	43.019212
17	vflt	32.816397
20	usr	28.164636
5	fork	25.336356
19	freeswap	25.115743
16	pflt	24.296776
11	pgfree	24.111997
15	ppgin	23.350190
14	pgin	23.215483
3	sread	18.600202
4	swrite	16.967349
9	pgout	16.224461
0	lread	9.327300
2	scall	9.093957
1	lwrite	6.455932
6	exec	5.957179
7	rchar	4.268359
18	freemem	3.459706
8	wchar	3.402886
13	atch	2.750825
21	runqsz_Not_CPU_Bound	2.494291
12	pgscan	NaN

/usr/local/lib/python3.10/dist-packag

  High Multicollinearity: The variables "ppgout," "vflt," "usr," "fork," "freeswap," "pflt," "pgfree," "ppgin," "pgin," "sread," "swrite," and "pgout" have relatively high VIF scores. This indicates that they may have high multicollinearity with other variables in the dataset.

Moderate Multicollinearity: The variables "lread," "scall," "lwrite," "exec," "rchar," "freemem," "wchar," "atch," and "runqsz_Not_CPU_Bound" have moderate VIF scores, suggesting some degree of multicollinearity.

Outliers and Singularities: It's important to note that the VIF for "pgscan" is reported as NaN. This might be due to a perfect linear relationship with other variables or singularities in the data. You should investigate this variable further.

Impact on Model: High VIF values can be problematic for regression models. They indicate that the corresponding variables are highly correlated with others, making it difficult to isolate their individual effects on the target variable. To address multicollinearity, you might consider removing or transforming some of the highly correlated variables.

Dropping variables to overcome multicollinearity:

"pgscan","ppgout","pgfree","vflt","ppgin","pgin","fork","pflt","pgout" variables were dropped from training data as their VIF was more than 10.
A VIF Score exceeding 10 is an indicator of high multicollinearity, a condition that can pose challenges for regression analysis. To enhance the model's accuracy and refine the coefficient estimates for the remaining predictors, certain variables have been eliminated. This step effectively mitigates multicollinearity and contributes to a more reliable and precise predictive model.

	Columns	Coefficient Estimate
0	const	0.000
1	lread	-0.160
2	lwrite	0.170
3	scall	-0.000
4	sread	-0.003
5	swrite	-0.018
6	exec	-1.596
7	rchar	-0.000
8	wchar	0.000
9	atch	-0.251
10	freemem	-0.001
11	freeswap	0.000
12	runqsz_Not_CPU_Bound	1.267

These are the coefficients of independent variables.
Observation:

- The 'const' (constant) term in the model has a coefficient estimate of 0.000, indicating that it doesn't significantly contribute to the dependent variable's ('usr') variation.
- The 'lread' variable has a coefficient estimate of -0.160, suggesting that for each unit increase in 'lread', 'usr' is expected to decrease by approximately 0.160, holding other variables constant.
- 'lwrite' has a coefficient estimate of 0.170, implying that a one-unit increase in 'lwrite' corresponds to an increase of around 0.170 in 'usr' when other factors remain constant.
- 'scall', 'rchar', 'wchar', and 'freeswap' have coefficient estimates close to 0, indicating that these variables have negligible impacts on 'usr'.
- 'sread', 'swrite', 'freemem', and 'pgout' have small coefficient estimates, suggesting modest effects on 'usr'.
- The 'exec' variable has a substantial negative coefficient estimate of -1.596, indicating that

higher values of 'exec' are associated with lower 'usr' when other predictors are held constant.

- 'atch' has a negative coefficient estimate of -0.251, suggesting that increases in 'atch' are linked to decreases in 'usr'.
- 'runqsz_Not_CPU_Bound' exhibits a positive coefficient estimate of 1.267, implying that higher values of 'runqsz_Not_CPU_Bound' correspond to increased 'usr'.

OLS Regression Summary :

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.739
Model:	OLS	Adj. R-squared:	0.738
Method:	Least Squares	F-statistic:	1347.
Date:	Sun, 22 Oct 2023	Prob (F-statistic):	0.00
Time:	08:19:44	Log-Likelihood:	-17370.
No. Observations:	5734	AIC:	3.477e+04
Df Residuals:	5721	BIC:	3.485e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.8645	0.344	246.444	0.000	84.189	85.540
lread	-0.1597	0.010	-16.494	0.000	-0.179	-0.141
lwrite	0.1701	0.014	11.869	0.000	0.142	0.198
scall	-0.0005	7.02e-05	-6.680	0.000	-0.001	-0.000
sread	-0.0032	0.001	-2.783	0.005	-0.005	-0.001
swrite	-0.0182	0.002	-11.745	0.000	-0.021	-0.015
exec	-1.5962	0.041	-39.368	0.000	-1.676	-1.517
rchar	-7.968e-06	5.36e-07	-14.857	0.000	-9.02e-06	-6.92e-06
wchar	6.362e-07	1.13e-06	0.563	0.574	-1.58e-06	2.85e-06
atch	-0.2513	0.138	-1.824	0.068	-0.521	0.019
freemem	-0.0005	5.57e-05	-9.350	0.000	-0.001	-0.000
freeswap	8.415e-06	2.08e-07	40.521	0.000	8.01e-06	8.82e-06
runqsz_Not_CPU_Bound	1.2668	0.142	8.937	0.000	0.989	1.545

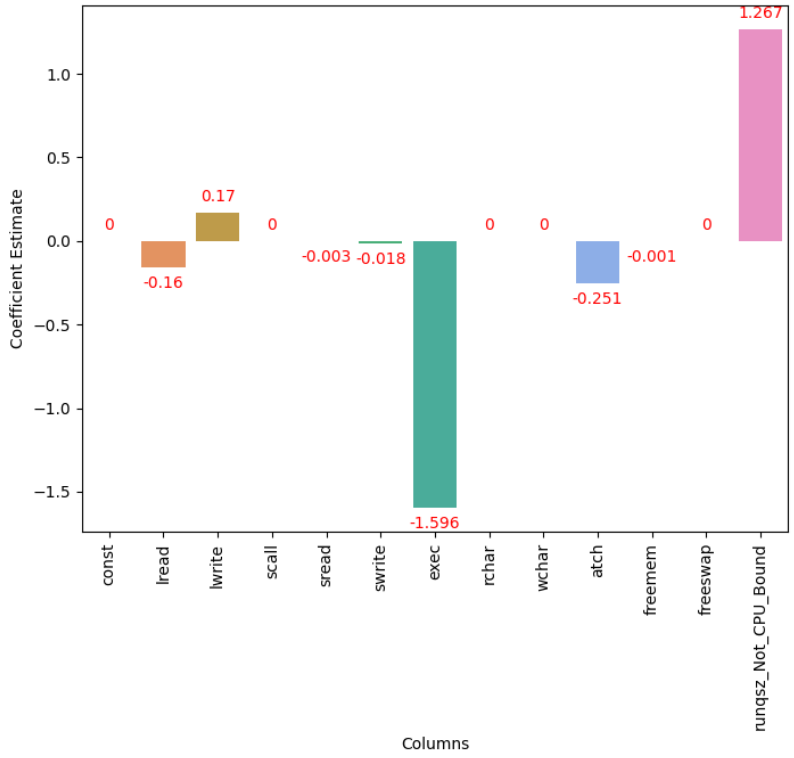
Omnibus:	712.941	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1187.819
Skew:	-0.852	Prob(JB):	1.17e-258
Kurtosis:	4.439	Cond. No.	7.46e+06

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.

```
The intercept for our model is 84.86454448751807
```

These are the coefficients of OLS method model:

	Columns	Coefficient Estimate
0	const	0.000
1	lread	-0.160
2	lwrite	0.170
3	scall	-0.000
4	sread	-0.003
5	swrite	-0.018
6	exec	-1.596
7	rchar	-0.000
8	wchar	0.000
9	atch	-0.251
10	freemem	-0.001
11	freeswap	0.000
12	runqsz_Not_CPU_Bound	1.267



The above plot is the boxplot visualization for coefficients.

Observation:

- 'const' (constant) has a coefficient estimate of 0.000, suggesting it does not significantly impact the dependent variable 'usr'.
- 'lread' has a negative coefficient estimate of -0.160, indicating that an increase in 'lread' is associated with a decrease in 'usr'.
- 'lwrite' has a positive coefficient estimate of 0.170, suggesting that an increase in 'lwrite' corresponds to an increase in 'usr'.
- 'scall', 'rchar', 'wchar', and 'freeswap' have coefficients close to 0, implying that these variables have minimal impact on 'usr'.

- 'sread' and 'swrite' have small negative coefficient estimates, indicating that increases in these variables lead to slight decreases in 'usr'.
- 'exec' has a notably negative coefficient estimate of -1.596, suggesting that higher values of 'exec' are linked to lower 'usr'.
- 'atch' has a negative coefficient estimate of -0.251, indicating that an increase in 'atch' corresponds to a decrease in 'usr'.
- 'freemem' has a small negative coefficient estimate of -0.001, suggesting a minor negative impact on 'usr'.
- 'runqsz_Not_CPU_Bound' has a positive coefficient estimate of 1.267, implying that higher values of this variable are associated with an increase in 'usr'.

Observation:

R-squared (R^2) serves as a key metric in assessing how well our regression model aligns with the observed data. It quantifies the extent to which the variance in the dependent variable (y) can be attributed to the independent variables (X) included in the model.

R^2 typically falls within the range of 0 to 1. A score of 1 signifies that the regression model perfectly predicts the target variable, while a score of 0 implies that the model does not account for any variation in the target variable.

In the context of this analysis:

The R-squared value for the training data is 0.73853.
The R-squared value for the testing data is 0.70194.
These results indicate that a substantial portion of the variance in the target variable has been effectively explained by the model.

To assess the model's performance:

The predicted values represent the model's estimates of the target variable based on the input features from the training data.
It is evident that the model accurately captures patterns within the training data, which is a positive outcome.
The model is deemed a strong fit, as the predicted values closely match the actual values.

Expression for regression formula-

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

OLS Regression Results-

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, 22 Oct 2023	Prob (F-statistic):	0.00
Time:	08:32:29	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]
Intercept	84.1217	0.316	266.106	0.000	83.502	84.741
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	3.325e-14	1.46e-16	228.488	0.000	3.3e-14	3.35e-14
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_Not_CPU_Bound	1.6153	0.126	12.819	0.000	1.368	1.862
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.	1.09e+22

```
Intercept                8.412174e+01
lread                   -6.348151e-02
lwrite                   4.816129e-02
scall                   -6.638280e-04
sread                    3.082521e-04
swrite                  -5.421822e-03
fork                     2.931273e-02
exec                    -3.211665e-01
rchar                   -5.166842e-06
wchar                   -5.402875e-06
pgout                   -3.688191e-01
ppgout                  -7.659768e-02
pgfree                   8.448414e-02
pgscan                  -1.116719e-15
atch                     6.275742e-01
pgin                     1.998791e-02
ppgin                   -6.733384e-02
pflt                    -3.360283e-02
vflt                    -5.463669e-03
runqsz_Not_CPU_Bound     1.615298e+00
freemem                 -4.584672e-04
freeswap                 8.831840e-06
dtype: float64
```

These are the parameters for independent variables

Incorporating a constant term into the input features for both the training and test datasets serves the purpose of enhancing the regression model's ability to provide precise estimates for both the intercept and coefficients. This, in turn, enables the model to effectively capture not only the baseline value of the target variable but also the alterations in the target variable that correspond to variations in the predictor variables.

	const	lread	lwrite	scall	sread	swrite	exec	rchar	wchar	atch	freemem	freeswap	runqsz_Not_CPU_Bound	constant
694	1.0	1.0	1.0	1345.0	223.0	192.0	0.6	198703.0	230625.875	1.5	121.0	1375446.0	0	1
5535	1.0	1.0	1.0	1429.0	87.0	67.0	0.2	7163.0	24842.000	0.0	1476.0	1021541.0	1	1
4244	1.0	47.0	25.0	3273.0	225.0	180.0	0.4	83246.0	53705.000	1.5	82.0	10989.5	0	1
2472	1.0	13.0	8.0	4349.0	300.0	191.0	3.0	96009.0	70467.000	0.0	772.0	993909.0	0	1
7052	1.0	17.0	23.0	225.0	13.0	13.0	1.6	17132.0	12514.000	0.0	4179.0	1821682.0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7935	1.0	1.0	0.0	3159.0	461.0	232.0	1.4	318346.0	64616.000	0.4	145.0	1125303.0	0	1
5192	1.0	3.0	0.0	493.0	72.0	55.0	0.8	29548.0	12598.000	0.4	325.0	1701342.0	0	1
3980	1.0	5.0	1.0	2410.0	176.0	150.0	0.6	74357.0	96317.000	0.2	170.0	1065320.0	0	1
235	1.0	2.0	0.0	2140.0	170.0	142.0	1.4	280183.0	189646.000	0.4	677.0	1053309.0	1	1
5157	1.0	8.0	1.0	2549.0	129.0	107.0	1.8	6665.0	50010.000	0.0	1224.0	1636050.0	1	1

5734 rows × 14 columns

Stats model Linear Regression summary:

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.739
Model:	OLS	Adj. R-squared:	0.738
Method:	Least Squares	F-statistic:	1347.
Date:	Sun, 22 Oct 2023	Prob (F-statistic):	0.00
Time:	09:01:18	Log-Likelihood:	-17370.
No. Observations:	5734	AIC:	3.477e+04
Df Residuals:	5721	BIC:	3.485e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	42.4323	0.172	246.444	0.000	42.095	42.770
lread	-0.1597	0.010	-16.494	0.000	-0.179	-0.141
lwrite	0.1701	0.014	11.869	0.000	0.142	0.198
scall	-0.0005	7.02e-05	-6.680	0.000	-0.001	-0.000
sread	-0.0032	0.001	-2.783	0.005	-0.005	-0.001
swrite	-0.0182	0.002	-11.745	0.000	-0.021	-0.015
exec	-1.5962	0.041	-39.368	0.000	-1.676	-1.517
rchar	-7.968e-06	5.36e-07	-14.857	0.000	-9.02e-06	-6.92e-06
wchar	6.362e-07	1.13e-06	0.563	0.574	-1.58e-06	2.85e-06
atch	-0.2513	0.138	-1.824	0.068	-0.521	0.019
freemem	-0.0005	5.57e-05	-9.350	0.000	-0.001	-0.000
freeswap	8.415e-06	2.08e-07	40.521	0.000	8.01e-06	8.82e-06
runqsz_Not_CPU_Bound	1.2668	0.142	8.937	0.000	0.989	1.545
constant	42.4323	0.172	246.444	0.000	42.095	42.770

Omnibus:	712.941	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1187.819
Skew:	-0.852	Prob(JB):	1.17e-258
Kurtosis:	4.439	Cond. No.	9.52e+19

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.26e-24. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

- The variable "usr" serves as the dependent variable in the regression analysis.
- The R-squared (R2) value is 0.739, signifying that 73.9% of the variance in the dependent variable can be accounted for by the independent variables included in the model.
- The Adjusted R-squared (Adj. R-squared) stands at 0.738, providing an adjustment for the number of predictors in the model.
- The F-statistic registers at 1347, assessing the overall significance of the model. A higher F-statistic indicates a more significant model.
- The Prob (F-statistic) value is 0.00, signaling the statistical significance of the overall model.
- Independent variables with small p-values (typically < 0.05) are deemed statistically significant in predicting the dependent variable.
- Negative coefficients (e.g., lread, scall, sread, swrite, atch, freemem) imply that an increase in these variables is linked to a decrease in the dependent variable "usr."
- Positive coefficients (e.g., lwrite, exec, rchar, freeswap, runqsz_Not_CPU_Bound) indicate that an increase in these variables corresponds to an increase in the dependent variable "usr."

- The model exhibits statistical significance, elucidating approximately 73.9% of the variance in the dependent variable. However, the presence of a constant variable and the potential issues related to multicollinearity require further attention for a more robust interpretation and extended analysis.

  We choose to drop “wchar”:

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.739
Model:	OLS	Adj. R-squared:	0.738
Method:	Least Squares	F-statistic:	1469.
Date:	Sun, 22 Oct 2023	Prob (F-statistic):	0.00
Time:	13:39:37	Log-Likelihood:	-17370.
No. Observations:	5734	AIC:	3.476e+04
Df Residuals:	5722	BIC:	3.484e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	42.4446	0.171	248.549	0.000	42.110	42.779
lread	-0.1601	0.010	-16.602	0.000	-0.179	-0.141
lwrite	0.1710	0.014	12.002	0.000	0.143	0.199
scall	-0.0005	7.02e-05	-6.674	0.000	-0.001	-0.000
sread	-0.0032	0.001	-2.861	0.004	-0.005	-0.001
swrite	-0.0180	0.002	-11.956	0.000	-0.021	-0.015
exec	-1.5977	0.040	-39.500	0.000	-1.677	-1.518
rchar	-7.859e-06	5e-07	-15.724	0.000	-8.84e-06	-6.88e-06
atch	-0.2503	0.138	-1.817	0.069	-0.520	0.020
freemem	-0.0005	5.57e-05	-9.342	0.000	-0.001	-0.000
freeswap	8.412e-06	2.08e-07	40.521	0.000	8.01e-06	8.82e-06
runqsz_Not_CPU_Bound	1.2557	0.140	8.947	0.000	0.981	1.531
constant	42.4446	0.171	248.549	0.000	42.110	42.779

Omnibus:	714.708	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1192.488
Skew:	-0.853	Prob(JB):	1.13e-259
Kurtosis:	4.443	Cond. No.	9.50e+19

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.26e-24. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Comparing the results of actual, predicted values with residuals:

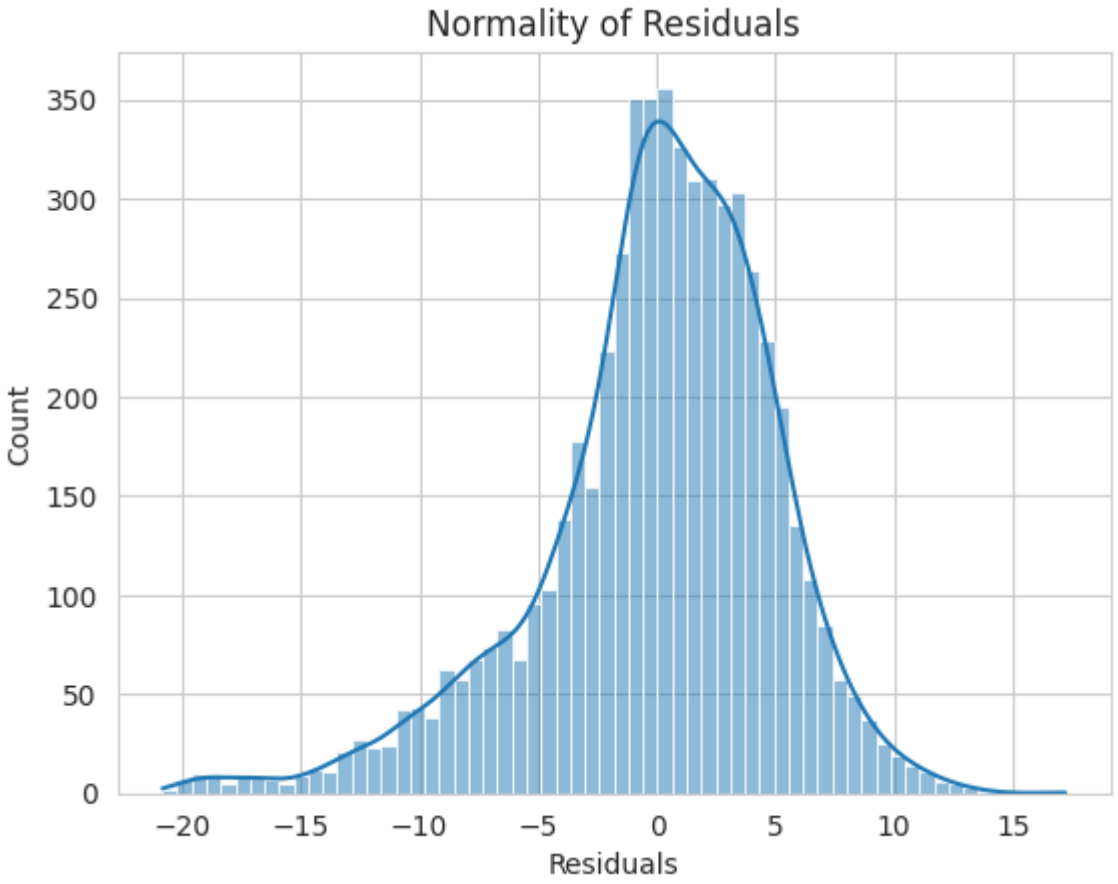
	Actual Values	Predicted Values	Residuals
0	81.0	88.711883	2.288117
1	93.0	91.451593	2.548407
2	64.0	74.526971	-13.026971
3	86.0	80.150772	2.849228
4	94.0	97.434866	-3.434866

Plot between Fitted values and Residuals-

This plot indicates there are lots of overpredictivtions and underpredictions and are not evenly distributed around the zero line.

Testing for Assumptions:

Normality of Residuals-



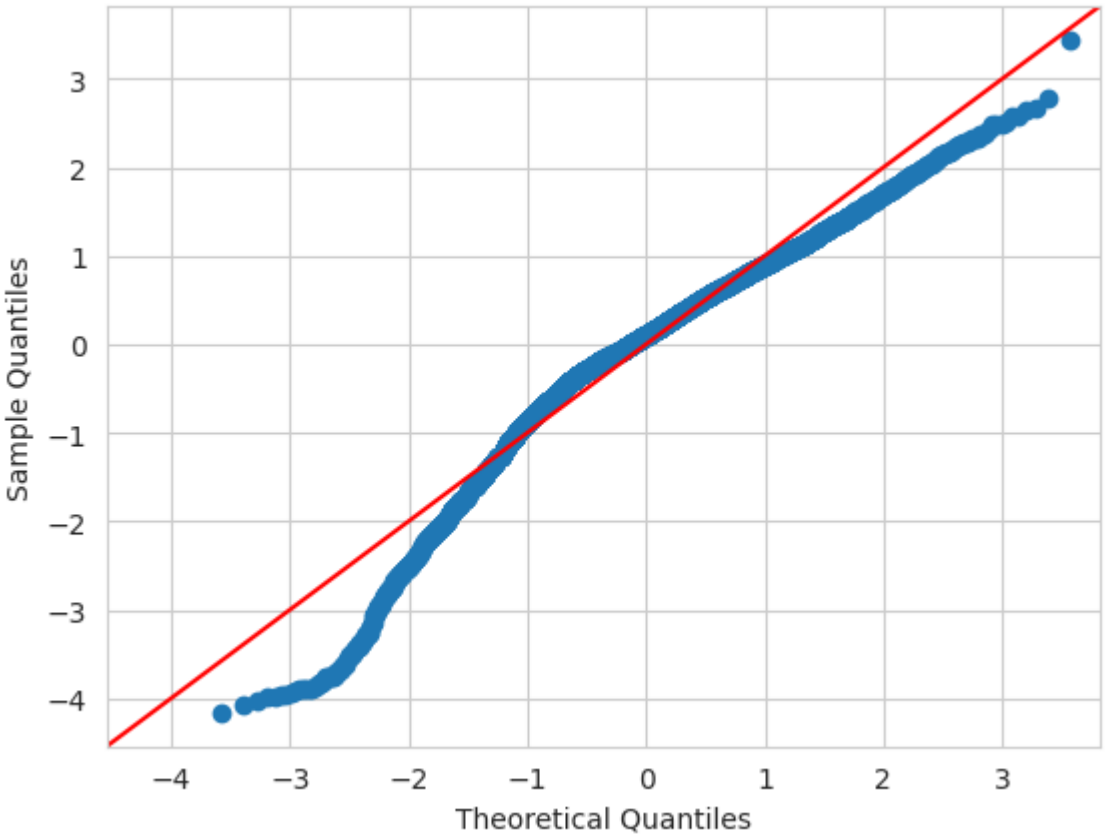
The plot shows the residuals are little left skewed and are approximately normal distribution. Therefore, we can say normality of residuals are satisfied

Shapiro Test for Normality-

```
ShapiroResult(statistic=0.9597432613372803, pvalue=3.6396386881882315e-37)
```

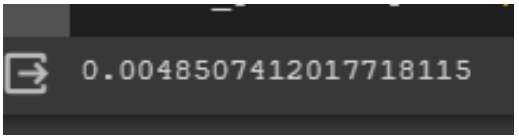
P-values exceeding the 0.05 significance level suggest a failure to reject the null hypothesis, indicating the data's approximate adherence to a normal distribution.

QQ plot test for residual:



The Q-Q plot's close adherence to the diagonal line indicates that residuals are approximately normally distributed, validating the model's normality assumptions, although some minor deviations, particularly in the tails, are noticeable.

Test for Homoscedasticity:



A p-value less than 0.05 rejects the null hypothesis, indicating evidence of heteroscedasticity, signifying varying residual variance across independent variable levels.

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.739
Model:	OLS	Adj. R-squared:	0.738
Method:	Least Squares	F-statistic:	1469.
Date:	Sun, 22 Oct 2023	Prob (F-statistic):	0.00
Time:	13:40:09	Log-Likelihood:	-17370.
No. Observations:	5734	AIC:	3.476e+04
Df Residuals:	5722	BIC:	3.484e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	42.4446	0.171	248.549	0.000	42.110	42.779
lread	-0.1601	0.010	-16.602	0.000	-0.179	-0.141
lwrite	0.1710	0.014	12.002	0.000	0.143	0.199
scall	-0.0005	7.02e-05	-6.674	0.000	-0.001	-0.000
sread	-0.0032	0.001	-2.861	0.004	-0.005	-0.001
swrite	-0.0180	0.002	-11.956	0.000	-0.021	-0.015
exec	-1.5977	0.040	-39.500	0.000	-1.677	-1.518
rchar	-7.859e-06	5e-07	-15.724	0.000	-8.84e-06	-6.88e-06
atch	-0.2503	0.138	-1.817	0.069	-0.520	0.020
freemem	-0.0005	5.57e-05	-9.342	0.000	-0.001	-0.000
freeswap	8.412e-06	2.08e-07	40.521	0.000	8.01e-06	8.82e-06
runqsz_Not_CPU_Bound	1.2557	0.140	8.947	0.000	0.981	1.531
constant	42.4446	0.171	248.549	0.000	42.110	42.779

Omnibus:	714.708	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1192.488
Skew:	-0.853	Prob(JB):	1.13e-259
Kurtosis:	4.443	Cond. No.	9.50e+19

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.26e-24. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

Root Mean Square Error (RMSE) value :

RMSE on train data

```
5.004933080691615
```

RMSE on test data

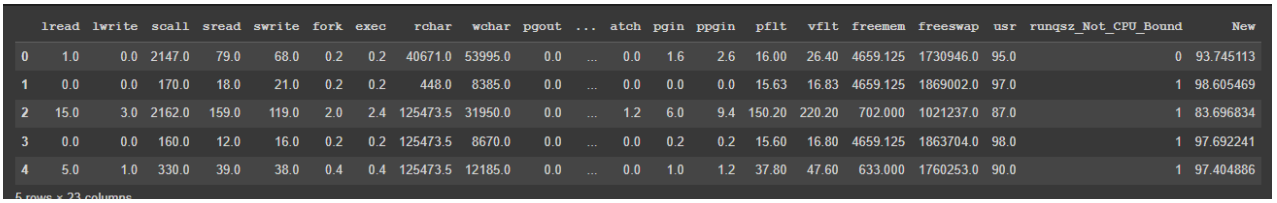
```
5.270428634028993
```

Coefficients of independent variables:

	Columns	Coefficient Estimate
0	const	0.000
1	lread	-0.160
2	lwrite	0.170
3	scall	-0.000
4	sread	-0.003
5	swrite	-0.018
6	exec	-1.596
7	rchar	-0.000
8	wchar	0.000
9	atch	-0.251
10	freemem	-0.001
11	freeswap	0.000
12	runqsz_Not_CPU_Bound	1.267

The equation would be:
(42.445) * const + (-0.16) * lread + (0.171) * lwrite + (-0.0) * scall + (-0.003) * sread + (-0.018) * swrite + (-1.598) * exec + (-0.0) * rchar + (-0.25) * atch + (-0.001) * freemem + (0.0) * freeswap + (1.256) * runqsz_Not_CPU_Bound + (42.445) * constant

These are the predictions from the model:



	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	atch	pgin	ppgin	pflt	vflt	freemem	freeswap	usr	runqsz_Not_CPU_Bound	New
0	1.0	0.0	2147.0	79.0	68.0	0.2	0.2	40671.0	53995.0	0.0	...	0.0	1.6	2.6	16.00	26.40	4659.125	1730946.0	95.0	0	93.745113
1	0.0	0.0	170.0	18.0	21.0	0.2	0.2	448.0	8385.0	0.0	...	0.0	0.0	0.0	15.63	16.83	4659.125	1869002.0	97.0	1	98.605469
2	15.0	3.0	2162.0	159.0	119.0	2.0	2.4	125473.5	31950.0	0.0	...	1.2	6.0	9.4	150.20	220.20	702.000	1021237.0	87.0	1	83.696834
3	0.0	0.0	160.0	12.0	16.0	0.2	0.2	125473.5	8670.0	0.0	...	0.0	0.2	0.2	15.60	16.80	4659.125	1863704.0	98.0	1	97.692241
4	5.0	1.0	330.0	39.0	38.0	0.4	0.4	125473.5	12185.0	0.0	...	0.0	1.0	1.2	37.80	47.60	633.000	1760253.0	90.0	1	97.404886

5 rows × 23 columns

	Model	R-squared	Adjusted R-squared	RMSE (Test)
0	OLS Model 1	0.738532	0.737984	5.270429

1.4 Inference: Basis on these predictions, what are the business insights and recommendations. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Business Insights:

Data Overview:

The dataset consists of 8192 rows and 22 columns. Two features contained missing values, which were imputed using the median. Zero values were retained in the dataset rather than being dropped. Outliers were identified and treated. Multicollinearity was assessed using VIF scores, leading to the removal of columns with VIF > 10. Model Building and Evaluation:

All independent variables were included in X, while the dependent variable was represented by y. The dataset was split into a 70:30 ratio for training and testing. A model was trained using sklearn with an intercept of 84. Model performance scores on the training and test sets were 73.58% and 70.19%, respectively. The model was also analyzed using statsmodels, explaining approximately 79.6% of the variability. Model Assumptions:

Assumptions including linearity, normality of residuals, homoscedasticity, and independence were found to be satisfied. Several predictor variables demonstrated significant impacts on the dependent variable 'usr'. Key Predictors and Business Insights:

lwrite Impact:

A one-unit increase in 'lwrite' is associated with a 0.176 times increase in 'usr,' assuming other factors remain constant. Business Insight: Enhancing write operations ('lwrite') can notably boost system performance ('usr'). runqsz_Not_CPU_Bound Impact:

An increase in 'runqsz_Not_CPU_Bound' corresponds to a 1.256 times increase in 'usr,' with other factors held constant. Business Insight: Ensuring that the system workload is not CPU-bound is crucial for improving system performance ('usr'). lread Impact:

An increase in 'lread' results in a reduction in 'usr' by a factor of 0.16, assuming other factors are constant. Business Insight: Optimizing read operations ('lread') can enhance system efficiency and, consequently, 'usr'. atch Impact:

An increase in 'atch' leads to a decrease in 'usr' by a factor of 0.25, while keeping other predictors constant. Business Insight: Monitoring and optimizing attachment operations ('atch') can have a positive impact on system performance ('usr'). exec Impact:

An increase in 'exec' is associated with a decrease in 'usr' by a factor of 1.598, under the condition that other factors remain constant. Business Insight: Efficient management of executed processes ('exec') is essential for maintaining optimal system performance ('usr'). These insights provide a clear understanding of the factors influencing 'usr' and offer actionable strategies for improving system performance.

Recommendations:

In conclusion, our analysis offers valuable insights into the key factors affecting system performance. Businesses seeking to enhance their systems can benefit from the following recommendations:

Optimize Read and Write Operations: Pay special attention to improving read ('lread') and write ('lwrite') operations. Enhancing these functions can significantly boost system efficiency ('usr').

Manage Workload Effectively: Ensure that the system workload is well-balanced and not CPU-bound ('runqsz_Not_CPU_Bound'). Proper workload management is crucial for maintaining optimal system performance.

Monitor Attachment Operations: Keep a close eye on attachment operations ('atch') and work on optimizing them. Effective management in this area can positively impact system performance.

Efficient Process Execution: Streamline process execution ('exec') for improved system efficiency. Efficiently managing executed processes is essential for maintaining optimal system performance.

Continuous Monitoring: Implement continuous monitoring of these key factors. This proactive approach ensures that the system consistently delivers optimal performance, leading to an enhanced user experience and operational efficiency.

Additionally, future efforts should explore other potential predictors and their impact on system performance, providing further insights for ongoing optimization initiatives. These recommendations, when executed effectively, can lead to improved system performance and better overall business outcomes.

Problem 2

Logistic Regression, LDA and CART

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

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

Dataset for Problem 2: Contraceptive_method_dataset.xlsx

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 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.

	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_method_used
0	24.0	Primary	Secondary	3.0	Scientology	No	2	High	Exposed	No
1	45.0	Uneducated	Secondary	10.0	Scientology	No	3	Very High	Exposed	No
2	43.0	Primary	Secondary	7.0	Scientology	No	3	Very High	Exposed	No
3	42.0	Secondary	Primary	9.0	Scientology	No	3	High	Exposed	No
4	36.0	Secondary	Secondary	8.0	Scientology	No	3	Low	Exposed	No

The above image shows the first 5 rows the data set

	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_method_used
1468	33.0	Tertiary	Tertiary	NaN	Scientology	Yes	2	Very High	Exposed	Yes
1469	33.0	Tertiary	Tertiary	NaN	Scientology	No	1	Very High	Exposed	Yes
1470	39.0	Secondary	Secondary	NaN	Scientology	Yes	1	Very High	Exposed	Yes
1471	33.0	Secondary	Secondary	NaN	Scientology	Yes	2	Low	Exposed	Yes
1472	17.0	Secondary	Secondary	1.0	Scientology	No	2	Very High	Exposed	Yes

The above image shows the last 5 rows the data set

```
(1473, 10)
```

The data has 1473 rows and 10 columns

```
Duplicate rows = 80
```

The data has 80 duplicate rows

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

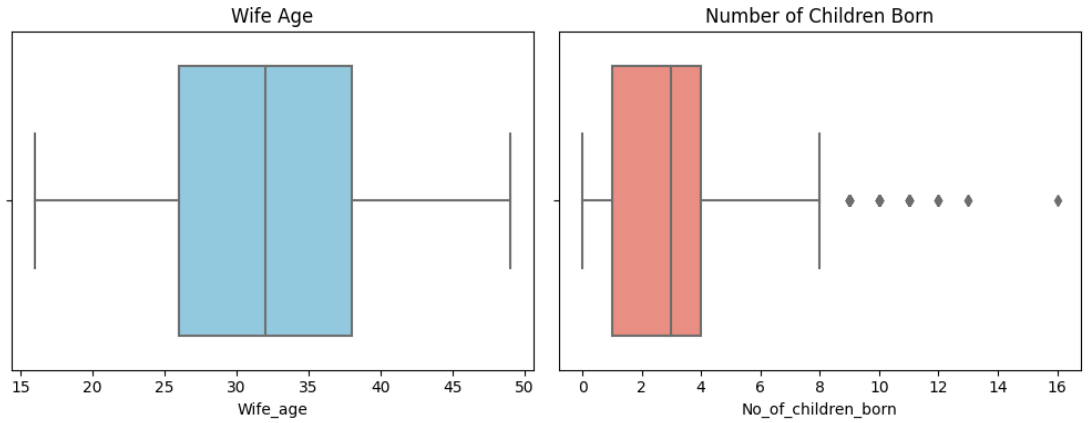
From the above information we can say that there are certain missing values in columns like Wife_age, No_of_childern_born. The missing values were imputed using median values.
'Husband_occupation' is int64 type and to be converted to categorical values.
'No_children_born' is float64 type and to be converted to numerical value.

5 point summary:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Wife_age	1473.0	NaN	NaN	NaN	32.577054	8.073941	16.0	26.0	32.0	38.0	49.0
Wife_ education	1473	4	Tertiary	577	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_education	1473	4	Tertiary	899	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_of_children_born	1473.0	NaN	NaN	NaN	3.250509	2.348473	0.0	1.0	3.0	4.0	16.0
Wife_religion	1473	2	Scientology	1253	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Wife_Working	1473	2	No	1104	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_Occupation	1473.0	4.0	3.0	585.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Standard_of_living_index	1473	4	Very High	684	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Media_exposure	1473	2	Exposed	1364	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Contraceptive_method_used	1473	2	Yes	844	NaN	NaN	NaN	NaN	NaN	NaN	NaN

- Wife_age ranges from 16 to 49 years, with a mean of approximately 32.58.
- Wife and Husband education are mainly 'Tertiary.'
- No_of_children_born averages around 3.25.
- Most wives are of the 'Scientology' religion, and around 75% are not working.
- Husband_Occupation has 4 levels, with '3.0' being the most common.
- 'Very High' is the predominant standard of living index.
- Media exposure is mainly 'Exposed.'
- Contraceptive method usage is 'Yes' for the majority.

Checking Outliers-

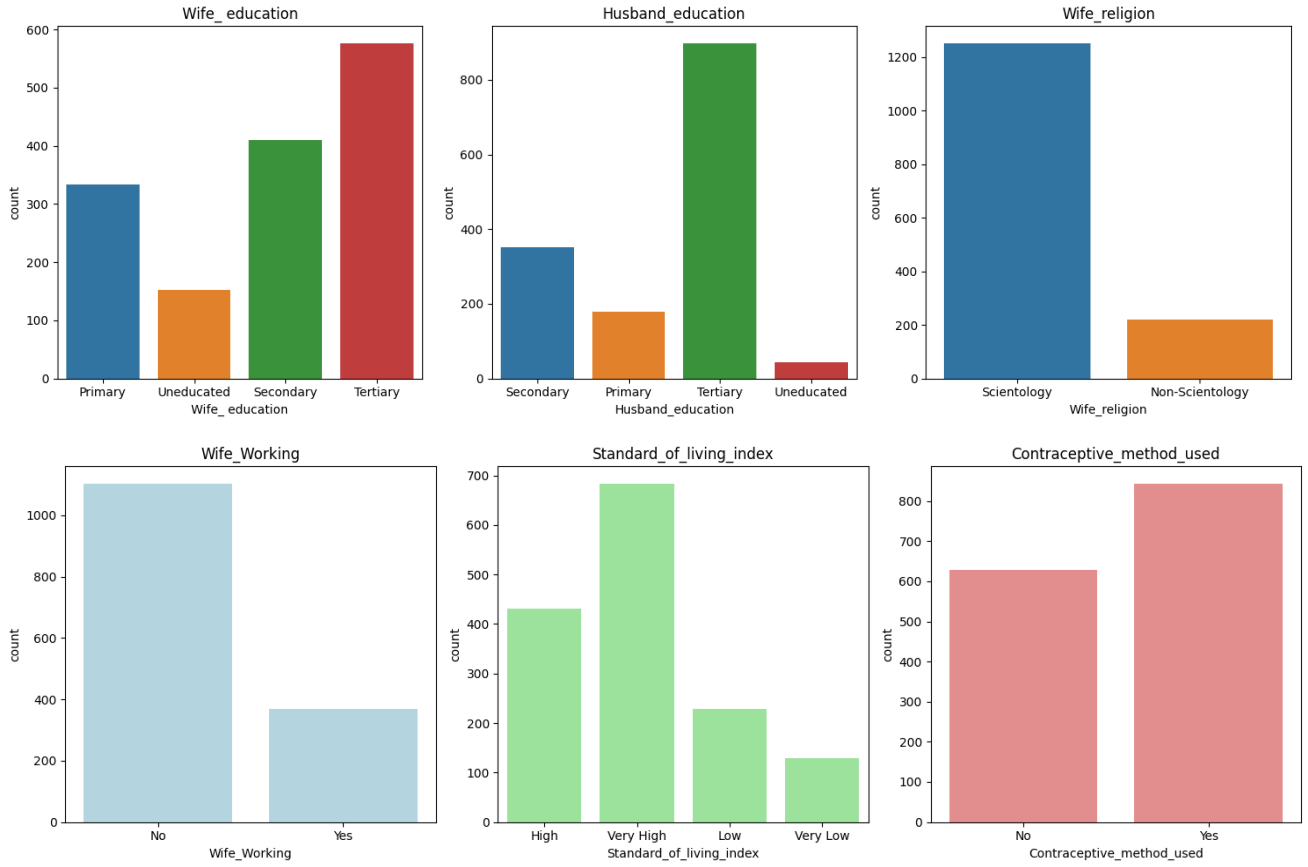


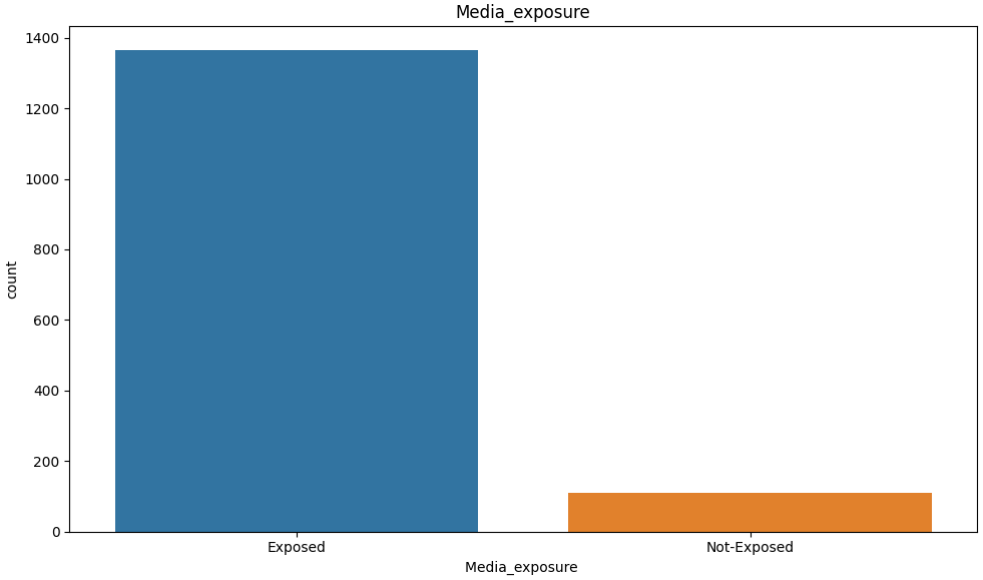
Observations:

- The presence of outliers in the 'No_of_children_born' variable indicates extreme values, which could be a result of certain families having significantly more or fewer children than the average. These outliers may be influenced by various factors such as cultural norms, social circumstances, economic conditions, and education, contributing to the variability in the data.
- Addressing outliers in this particular variable may not yield significant benefits. Instead, it might be more valuable to explore the use of all three models: logistic regression, Linear Discriminant Analysis (LDA), and Classification and Regression Trees (CART) to determine the most suitable model.
- Logistic regression is sensitive to outliers, and LDA offers a way to mitigate the impact of outliers. Therefore, employing all three models can provide a more comprehensive assessment of their performances. Currently, focusing on outlier treatment may not be the most effective strategy.

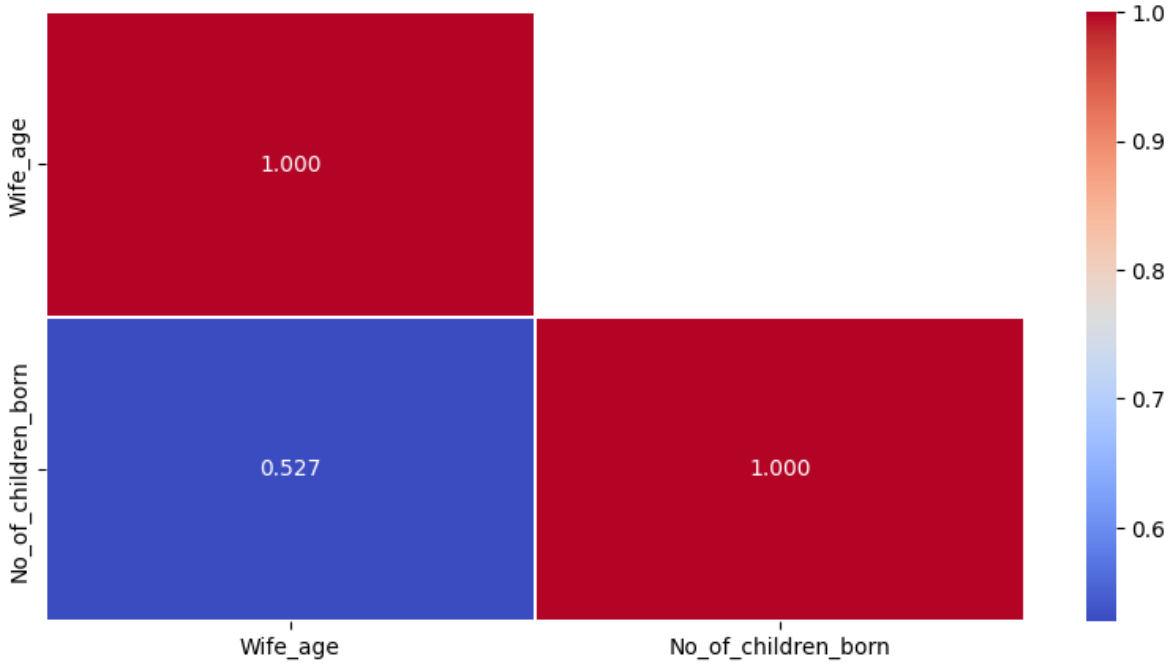
Univariate Analysis:

Countplots for individual features are:





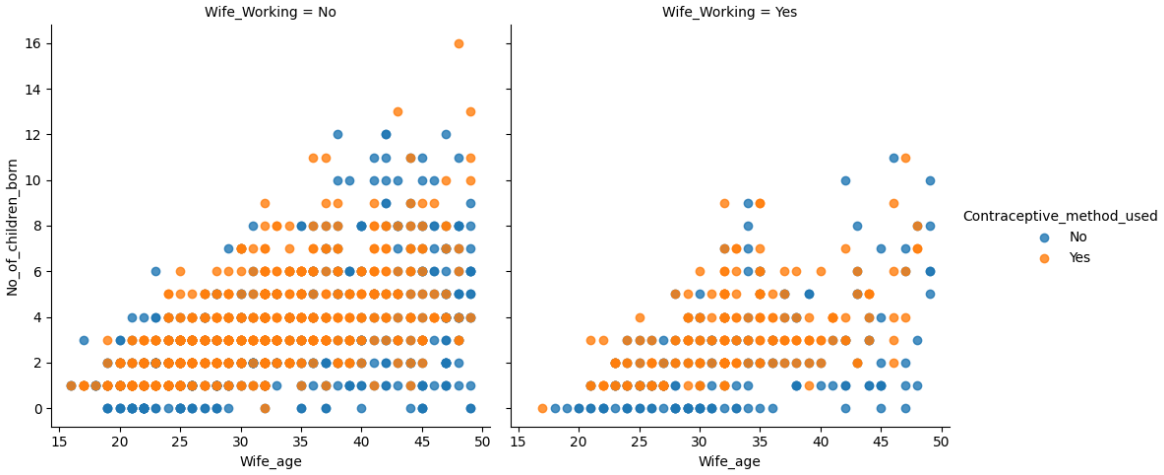
Bivariate Analysis-



This suggests there is no multicollinearity

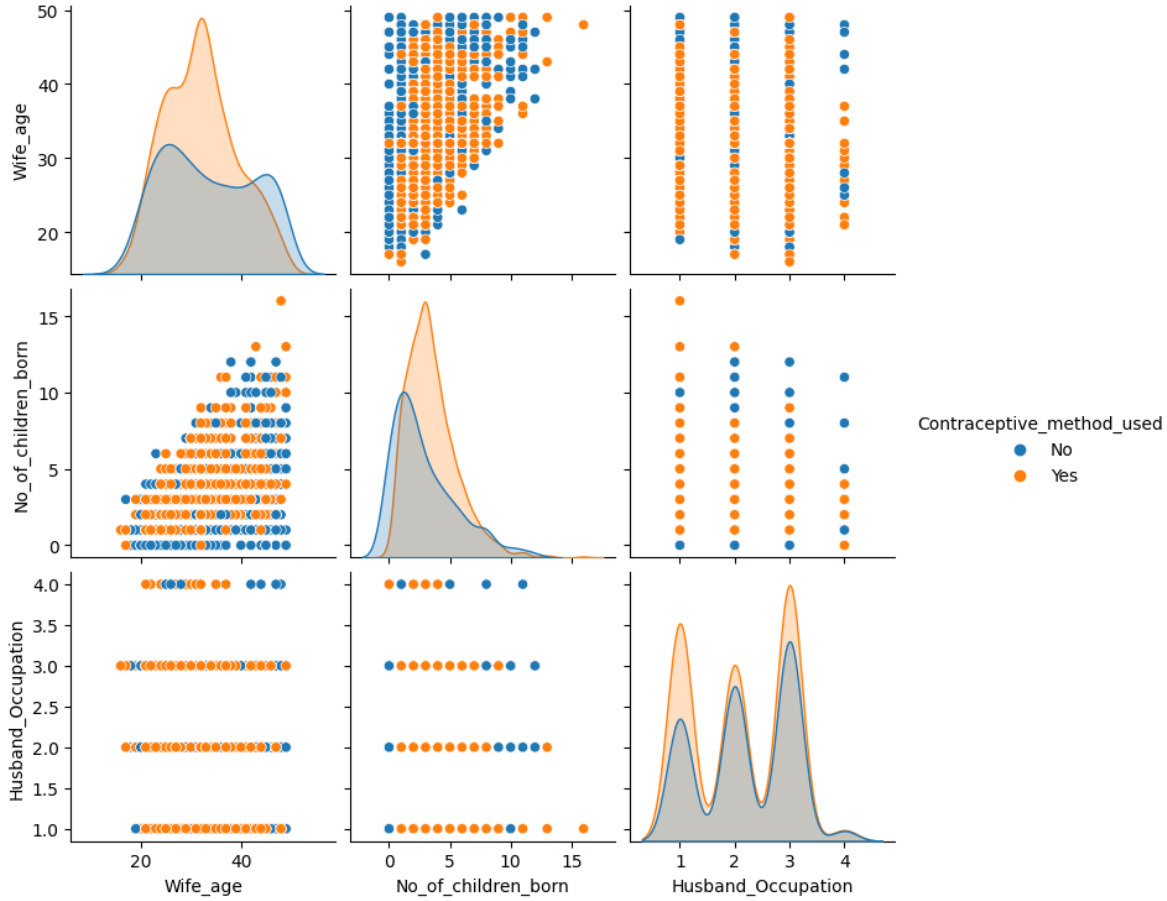
Multivariate Analysis:

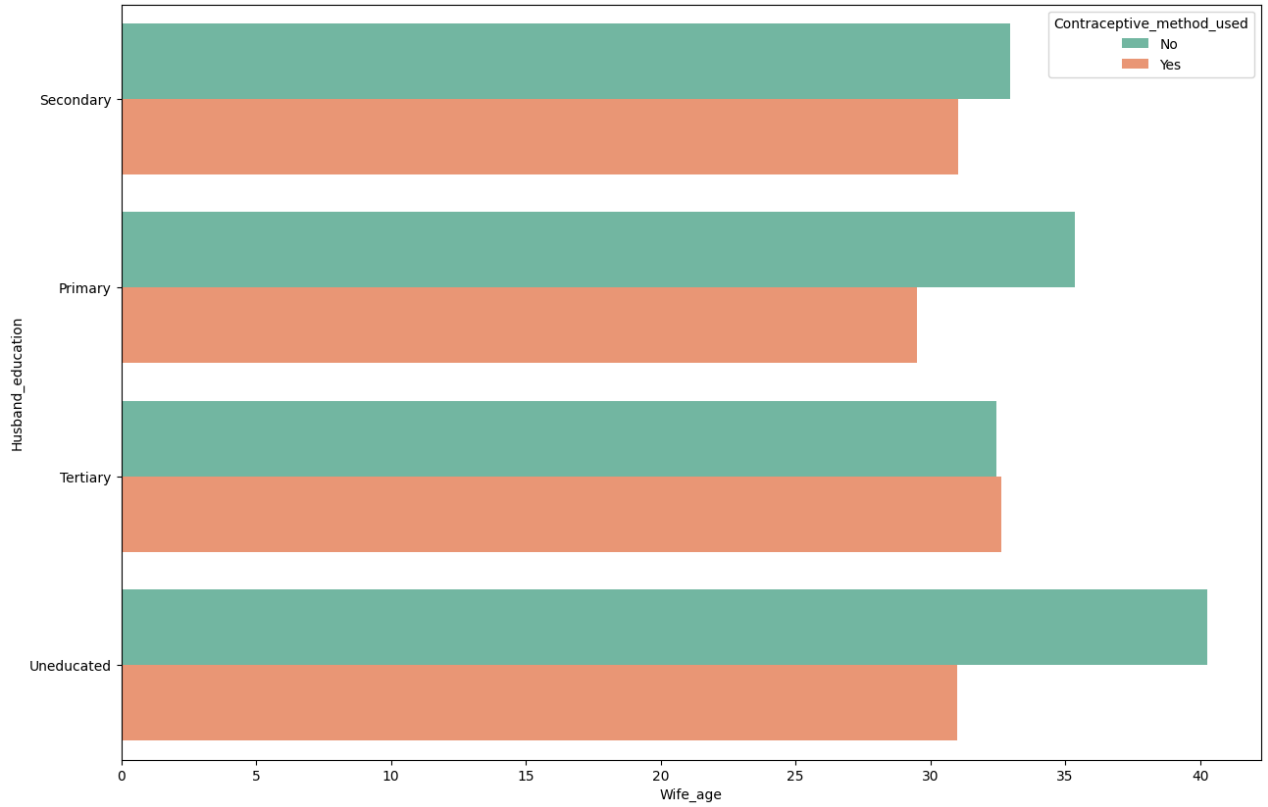
Scatter plot with regression line to study distribution pattern:



The choice of contraceptive methods is prominent among non-working wives, particularly within the age range of 15 to 50 years. In this age group, non-working wives exhibit a wide range of the number of children born, varying from 0 to 16. This concentration suggests a significant preference for contraceptive methods among non-working wives in these specific age and childbearing categories.

Scatterplots between the continuous independent variables versus target variable





Approximately 57.29% of the surveyed women have indicated the use of contraceptive methods. Furthermore, the data suggests that women with limited education who are approximately 40 years old tend not to use contraceptives.

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

	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_method_used
0	24.0	Primary	Secondary	3	Scientology	No	2	High	Exposed	0
1	45.0	Uneducated	Secondary	10	Scientology	No	3	Very High	Exposed	0
2	43.0	Primary	Secondary	7	Scientology	No	3	Very High	Exposed	0

+ Code + Text

	Wife_age	No_of_children_born	Contraceptive_method_used	Wife_ education_Secondary	Wife_ education_Tertiary	Wife_ education_Uneducated	Husband_education_Secondary	Husband_education_Tertiary	Husband_education_Uneducated	Wife_religion_Scientо
0	24.0	3	0	0	0	0	1	0	0	
1	45.0	10	0	0	0	1	1	0	0	
2	43.0	7	0	0	0	0	1	0	0	

This is the result of encoding categorical values using one-hot encoding

Splitting data:
Separating the dataset into two distinct dataframes, one containing predictor variables and the other housing target variables.
Subsequently, each of these dataframes is split into a 70:30 ratio for training and testing data, ensuring a balanced allocation for model development and evaluation.

.

Logistic modeling:

The process involves the following steps:

- Creating a model and fitting it to the training data while utilizing separate dataframes for predictor variables and target variables.
- Training a logistic regression model and using it to make predictions on both the training and testing datasets. These predictions can then be compared to the actual target values to evaluate the model's performance.
- The model generates predicted probabilities for each class (0 and 1) corresponding to the samples in the training data. These probabilities are essential for constructing ROC curves in subsequent analysis.

Predicted probabilities for ytest and train:

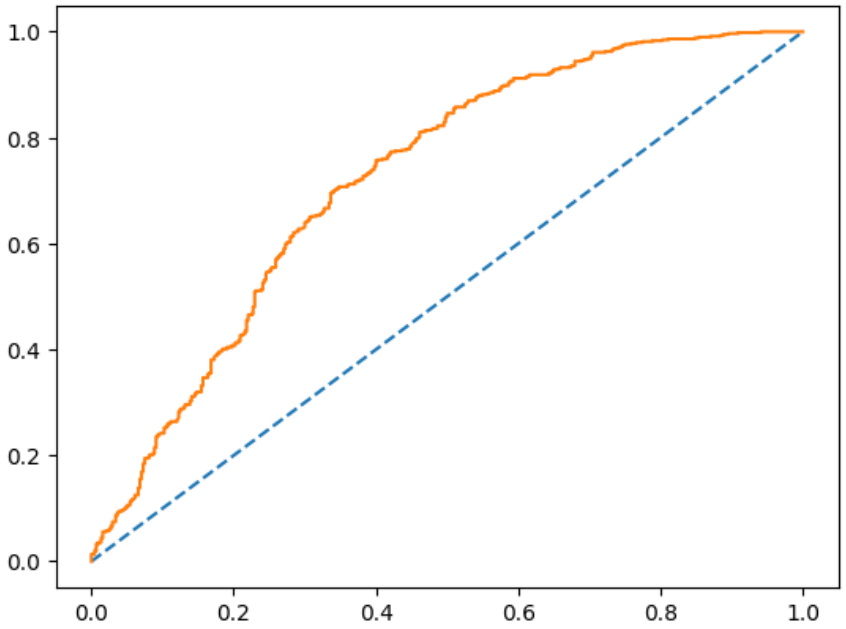
	0	1
0	0.267188	0.732812
1	0.495328	0.504672
2	0.265623	0.734377
3	0.235643	0.764357
4	0.346744	0.653256

	0	1
0	0.206205	0.793795
1	0.882759	0.117241
2	0.972225	0.027775
3	0.352366	0.647634
4	0.264268	0.735732

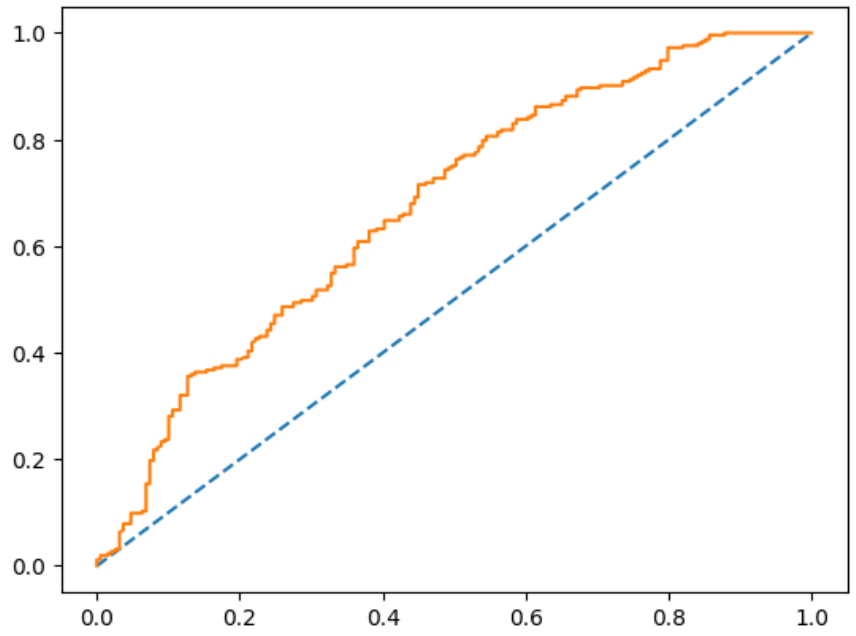
The model's performance is assessed by checking its accuracy on both the training and testing data sets:

The accuracy for the training data is 0.6896.
The accuracy for the test data is 0.6538.

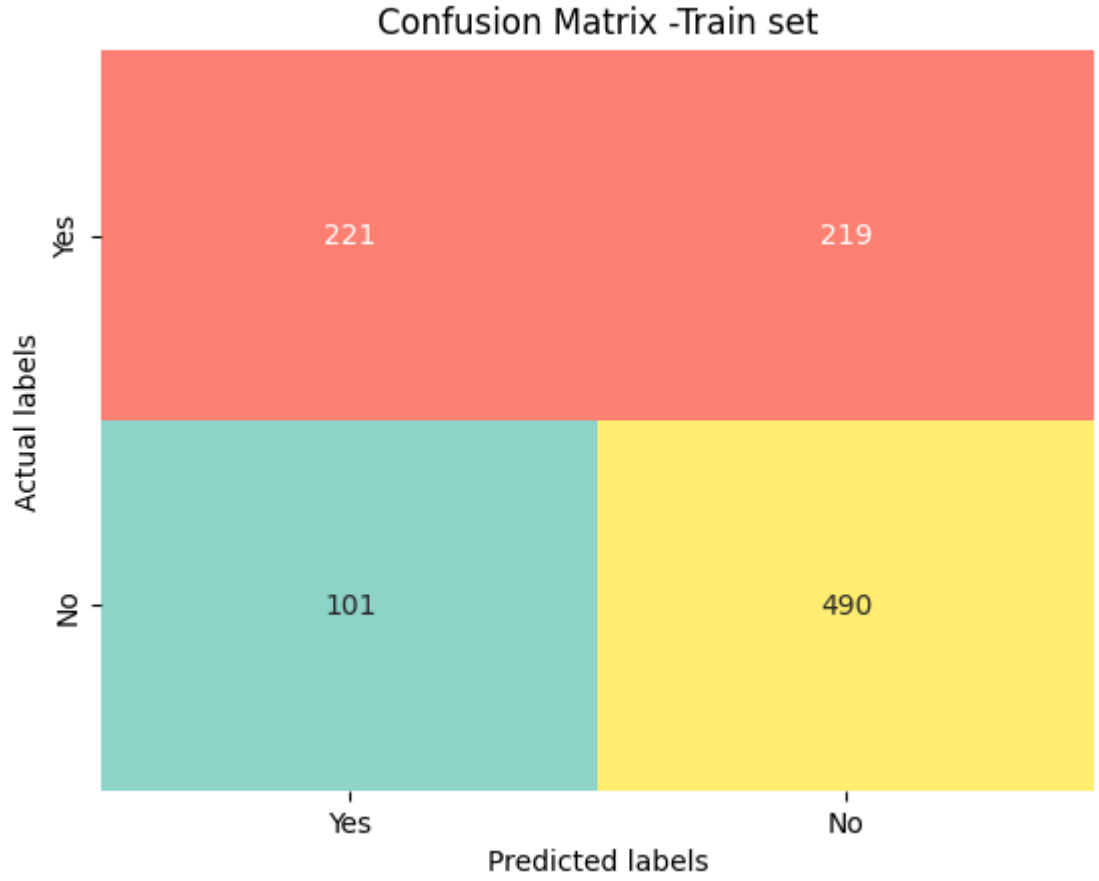
AUC score and ROC plot For Training Data :



AUC score and ROC plot For Testing Data :



Confusion matrix of Training data :



The evaluation metrics include:

```
tn, fp, fn, tp
```

```
(221, 219, 101, 490)
```

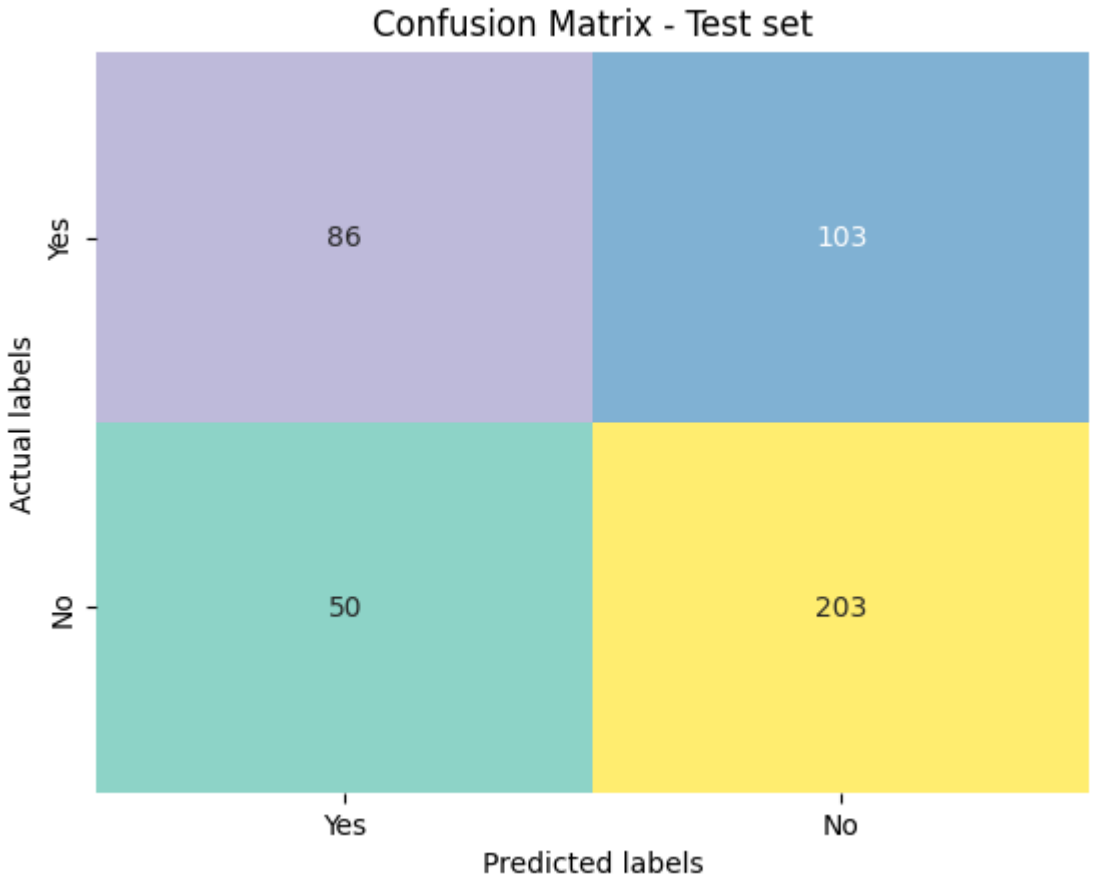
True Negative (tn): 221 cases were accurately classified as negative.
False Positive (fp): 219 cases were erroneously classified as positive when they were actually negative.
False Negative (fn): 101 cases were mistakenly classified as negative when they were actually positive.
True Positive (tp): 490 cases were correctly classified as positive.
Confusion matrix of Training data **:**

Classification report for Training set:

Classification Report for Training Set:

	precision	recall	f1-score	support
0	0.69	0.50	0.58	440
1	0.69	0.83	0.75	591
accuracy			0.69	1031
macro avg	0.69	0.67	0.67	1031
weighted avg	0.69	0.69	0.68	1031

Confusion matrix of Testing data :



The evaluation results indicate:

- True Negative (tn): 86 cases were accurately identified as negative.
- False Positive (fp): 103 cases were erroneously classified as positive when they were, in fact, negative.
- False Negative (fn): 50 cases were mistakenly classified as negative when they were actually positive.
- True Positive (tp): 203 cases were correctly recognized as positive.

Classification report for Testing set:

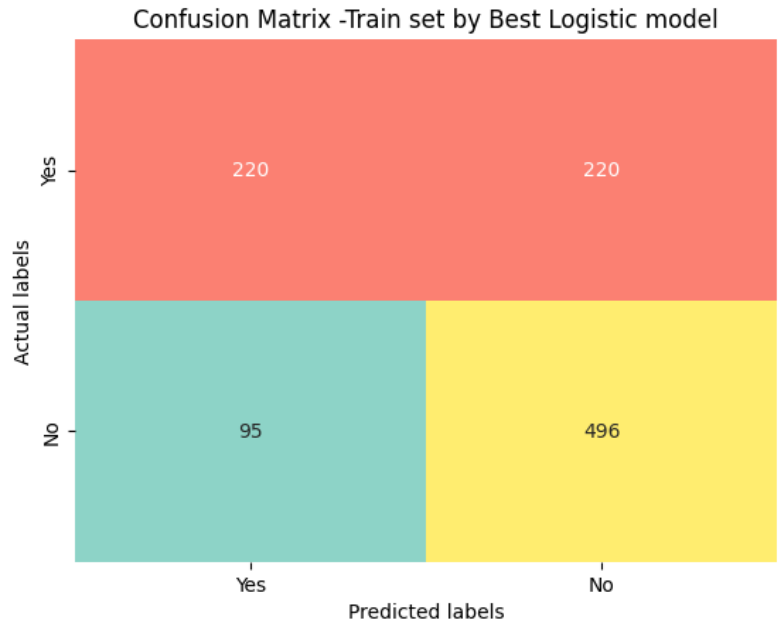
```
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.63      0.46      0.53       189
           1       0.66      0.80      0.73       253

    accuracy                           0.65       442
   macro avg       0.65      0.63      0.63       442
weighted avg       0.65      0.65      0.64       442
```

Logistic regression models rely on hyperparameters that play a crucial role in determining their performance. Hence, it is essential to systematically explore various hyperparameter combinations to enhance model performance. These combinations are chosen based on the F1-score, leading to the identification of the best-performing model, which can then be used for evaluation and predictions on new, unseen data..

Confusion matrix on Train set by new Logistic model created at best:



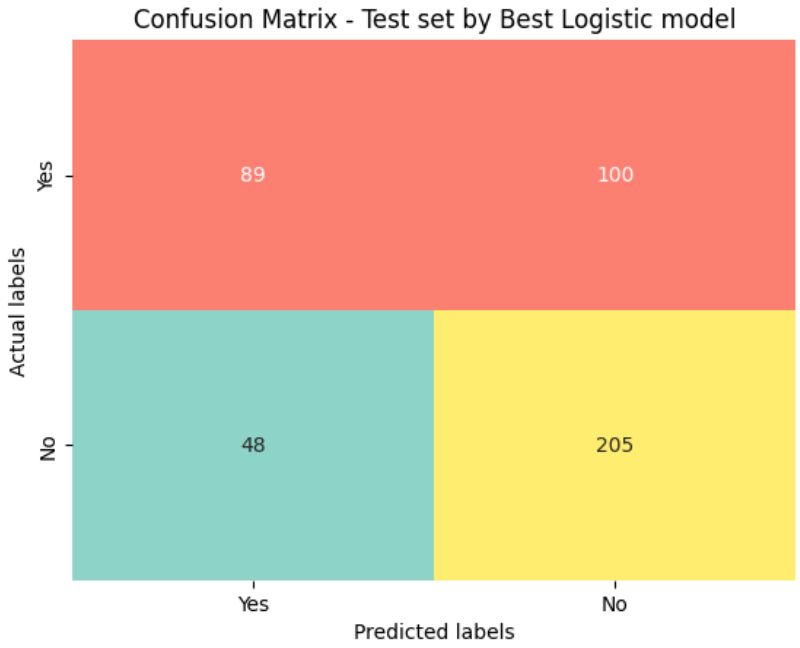
```
Classification Report for Training Set (Using Best Model):
              precision    recall  f1-score   support

           0       0.70      0.50      0.58       440
           1       0.69      0.84      0.76       591

    accuracy                           0.69      1031
   macro avg       0.70      0.67      0.67      1031
weighted avg       0.70      0.69      0.68      1031
```

True Negative (tn): 220 cases were accurately classified as negative.
False Positive (fp): 220 cases were mistakenly classified as positive when they were, in fact, negative.
False Negative (fn): 95 cases were erroneously classified as negative when they were actually positive.
True Positive (tp): 496 cases were correctly identified as positive.

Confusion matrix on Test set by new Logistic model created at best:




```
Classification Report for Test Set (Using Best Model):
              precision    recall  f1-score   support

           0       0.65      0.47      0.55       189
           1       0.67      0.81      0.73       253

    accuracy                           0.67       442
   macro avg       0.66      0.64      0.64       442
weighted avg       0.66      0.67      0.65       442
```

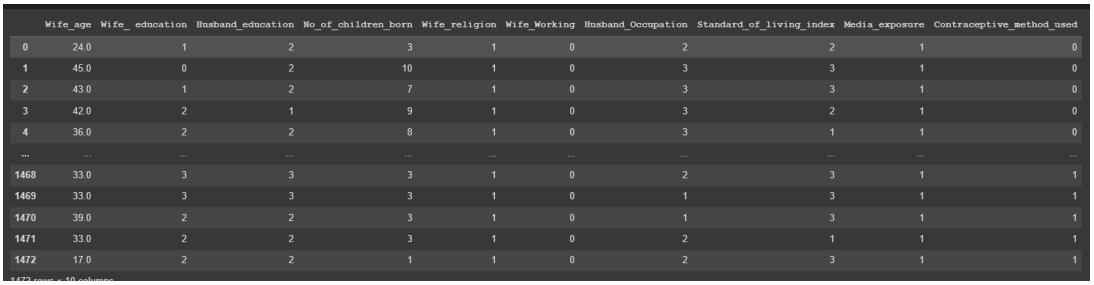
```
Accuracy for Training Set (Using Best Model): 69.45%
Accuracy for Test Set (Using Best Model): 66.52%
```

The accuracy score of the best-performing Logistic model is 69.45% for the training set and approximately 66.52% for the test set. Consequently, the overall accuracy for the Logistic model stands at 67%, considering a total of 442 predictions. This means that the number of correct

predictions can be calculated as 0.67 * 442, resulting in 296 accurate predictions.

LINEAR DISCRIMINANT ANALYSIS- (MODEL-2)

The categorical values has to converted for LDA:



	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_method_used
0	24.0	1	2	3	1	0	2	2	1	0
1	45.0	0	2	10	1	0	3	3	1	0
2	43.0	1	2	7	1	0	3	3	1	0
3	42.0	2	1	9	1	0	3	2	1	0
4	36.0	2	2	8	1	0	3	1	1	0
...	...	...	...	...	...	...	...	...	...	...
1468	33.0	3	3	3	1	0	2	3	1	1
1469	33.0	3	3	3	1	0	1	3	1	1
1470	39.0	2	2	3	1	0	1	3	1	1
1471	33.0	2	2	3	1	0	2	1	1	1
1472	17.0	2	2	1	1	0	2	3	1	1

1473 rows × 10 columns

This is the result after converting to numerical data.

Splitting the data into 70:30 using train_test_split method

LDA classifier is trained using training data and stored as a LDA Model.
Coefficients and intercepts are generated for linear discriminant function

Intercept:

```
Intercept: [0.27479997]
```

Coefficients:

```
Coefficients for the Linear Discriminant Function:
[[-8.56473358e-02  5.02149567e-01 -2.21628904e-02  3.19705859e-01
  -4.05928703e-01 -2.03837372e-15  3.74102262e-02  2.88837813e-01
   5.98302895e-01]]
```

Equation:

-0.09 * Wife_age (+) 0.5 * Wife_ education (+) -0.02 * Husband_education (+) 0.32 * No_of_children_born (+) -0.41 * Wife_religion (+) -0.0 * Wife_Working (+) 0.04 * Husband_Occupation (+) 0.29 * Standard_of_living_index (+) 0.6 * Media_exposure (+)

```
Classification Report of the training data:

              precision    recall  f1-score   support

           0       0.69      0.48      0.57       440
           1       0.69      0.84      0.76       591

    accuracy                           0.69      1031
   macro avg       0.69      0.66      0.66      1031
weighted avg       0.69      0.69      0.68      1031


Classification Report of the test data:

              precision    recall  f1-score   support

           0       0.64      0.43      0.52       189
           1       0.66      0.82      0.73       253

    accuracy                           0.65       442
   macro avg       0.65      0.63      0.62       442
weighted avg       0.65      0.65      0.64       442
```

Observations:

About 70.25% of both the training and test datasets are composed of class 1 observations, indicating a significant class imbalance in the data. The model achieves an accuracy of approximately 67% on both the training and test sets, which closely mirrors the proportion of class 1 observations. This suggests that the model's performance is heavily influenced by the class distribution.

Given the data's imbalance and its relatively small size, containing only 1474 observations, enhancing the model's robustness could be achieved by expanding the dataset.

Inferences:

Precision (Positive Predictive Value): Precision represents the proportion of positive predictions that are correct out of all the positive predictions made. For instance, in the case of women not using contraceptive methods (Label 0), a precision of 64% indicates that 64% of the predictions for this group were accurate.

Recall (Sensitivity): Recall signifies the proportion of actual positive instances correctly predicted by the model. For women not using contraceptive methods, a recall of 43% suggests that the model identified 43% of all non-contraceptive users.

For Women Who Do Not Use Contraceptive Methods (Label 0):

Precision (64%): Out of all predictions made for women not using contraceptive methods, 64% were correct.
Recall (43%): The model captured 43% of all women who truly do not use contraceptive methods.
For Women Who Use Contraceptive Methods (Label 1):

Precision (66%): Among the predictions for women using contraceptive methods, 66% were accurate.
Recall (82%): The model correctly identified 82% of women who indeed use contraceptive methods.

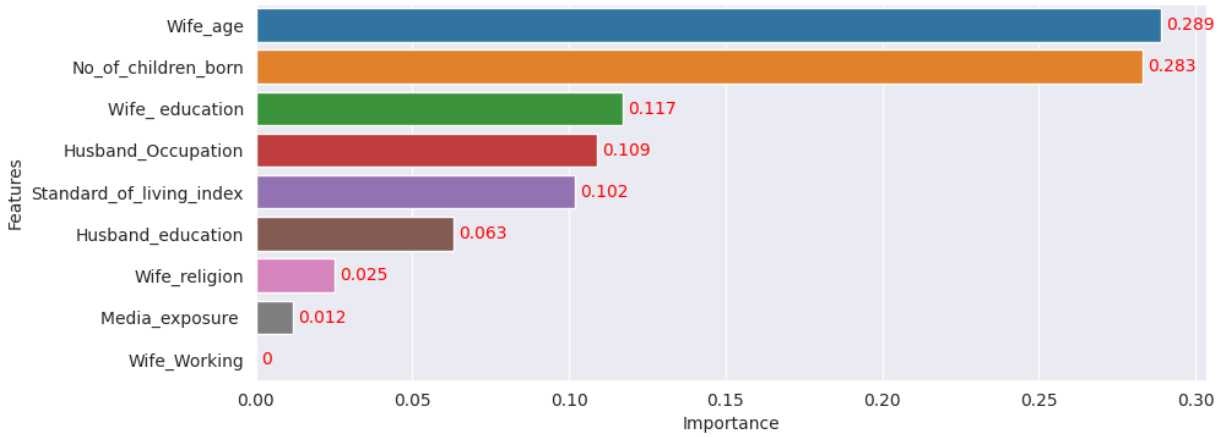
Model Evaluation:

Accuracy, Area Under the Curve (AUC), Precision, and Recall metrics for the test data closely align with those from the training data. This consistency indicates the absence of overfitting or underfitting, demonstrating that the model performs well in both training and testing scenarios. In summary, the model is robust and effective for classification tasks.

CLASSIFICATION AND REGRESSION TREE: (MODEL-3)
Building a decision tree classifier-

	Features	Importance
0	Wife_age	0.289
3	No_of_children_born	0.283
1	Wife_ education	0.117
6	Husband_Occupation	0.109
7	Standard_of_living_index	0.102
2	Husband_education	0.063
4	Wife_religion	0.025
8	Media_exposure	0.012
5	Wife_Working	0.000

The above image tell us the important features from the model.
It is visualized and shown below:



Observations:

The provided values do not represent Gini gain; instead, the model utilizes the concept of partial derivatives to identify the minimum error and extract feature importance.

The approach involves systematically removing each feature to observe how it affects the model. The feature that causes the most significant impact on the model's performance is considered the most important.

The CART algorithm evaluates feature importance by assessing the influence of removing each

feature on the model's performance. If removing a particular feature results in a substantial increase in error or loss, that feature is deemed more important. This method offers a detailed understanding of feature importance by quantifying each feature's effect on the model's performance.

Results:

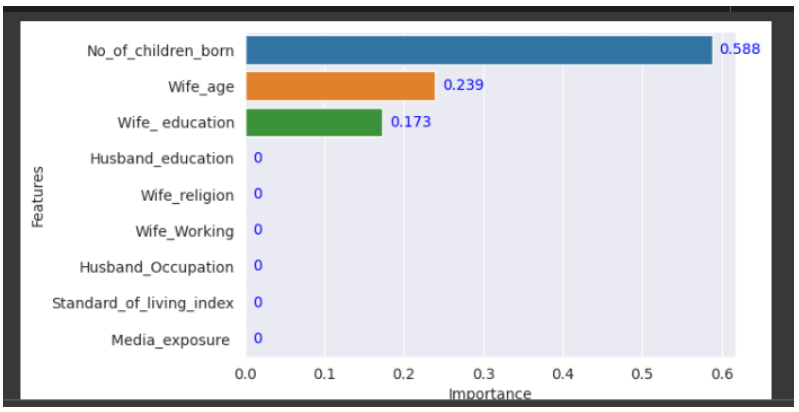
```
Decision Tree Model Accuracy on Training Data: 0.98
```

```
Decision Tree Model Accuracy on Test Data: 0.62
```

As you can see the model is overfitted

Regularization:

	Features	Importance
3	No_of_children_born	0.588
0	Wife_age	0.239
1	Wife_ education	0.173
2	Husband_education	0.000
4	Wife_religion	0.000
5	Wife_Working	0.000
6	Husband_Occupation	0.000
7	Standard_of_living_index	0.000
8	Media_exposure	0.000



Observations:

Accuracy: The model achieved a 73% accuracy in correctly classifying whether a wife is using contraceptive methods or not.

Recall (1): Regarding wives who are genuinely using contraceptive methods, the model correctly predicted 96% of them, demonstrating a high true positive rate.

Precision (1): Out of all the instances the model predicted as wives using contraceptive methods, 69% of them were indeed correct. This highlights the proportion of true positive predictions out of all positive predictions made.

These metrics offer valuable insights into the model's performance, emphasizing its capability to identify women who are using contraceptive methods. A high recall signifies the model's effectiveness in capturing actual positive cases, while precision underscores the accuracy of its positive predictions.

Conclusion:

The model consistently delivers strong performance, achieving an accuracy of 73% and an AUC of 74.3% during training, and 69% accuracy with a 71.3% AUC in testing. Its high recall ensures precise detection of contraceptive usage, which can play a pivotal role in advancing proactive healthcare initiatives for women. As such, this model proves to be a valuable asset for Indonesia's Ministry of Health.

The probability that it has been caused by a fire hazard is approximately 0.270.
The probability that it has been caused by a mechanical failure is approximately 0.405.
The probability that it has been caused by a human error is approximately 0.324.