Predictive Modelling Project

Pavithra Devi
DSBA
Great Learning

Index

Problem 1:

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

Method used for prediction:

Linear Regression:-

What is linear regression? Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Here, our dependent variable is ‘usr’ and all the other variables are independent variables.

Context:

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 Exploratory Data Analysis:

	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	0.20	40671.0	53995.0	0.00	...	0.00	0.0	1.60	2.60	16.00	26.40	CPU_Bound	4670	1730946	95
1	0	0	170	18	21	0.2	0.20	448.0	8385.0	0.00	...	0.00	0.0	0.00	0.00	15.63	16.83	Not_CPU_Bound	7278	1869002	97
2	15	3	2162	159	119	2.0	2.40	NaN	31950.0	0.00	...	0.00	1.2	6.00	9.40	150.20	220.20	Not_CPU_Bound	702	1021237	87
3	0	0	160	12	16	0.2	0.20	NaN	8670.0	0.00	...	0.00	0.0	0.20	0.20	15.60	16.80	Not_CPU_Bound	7248	1863704	98
4	5	1	330	39	38	0.4	0.40	NaN	12185.0	0.00	...	0.00	0.0	1.00	1.20	37.80	47.60	Not_CPU_Bound	633	1760253	90
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8187	16	12	3009	360	244	1.6	5.81	405250.0	85282.0	8.02	...	55.11	0.6	35.87	47.90	139.28	270.74	CPU_Bound	387	986647	80
8188	4	0	1596	170	146	2.4	1.80	89489.0	41764.0	3.80	...	0.20	0.8	3.80	4.40	122.40	212.60	Not_CPU_Bound	263	1055742	90
8189	16	5	3116	289	190	0.6	0.60	325948.0	52640.0	0.40	...	0.00	0.4	28.40	45.20	60.20	219.80	Not_CPU_Bound	400	969106	87
8190	32	45	5180	254	179	1.2	1.20	62571.0	29505.0	1.40	...	18.04	0.4	23.05	24.25	93.19	202.81	CPU_Bound	141	1022458	83
8191	2	0	985	55	46	1.6	4.80	111111.0	22256.0	0.00	...	0.00	0.2	3.40	6.20	91.80	110.00	CPU_Bound	659	1756514	94

8192 rows × 22 columns

Fig(1.1)

Fig(1.1). Dataset Size: Dataset contains 8192 rows and 22 columns.

Observations on the dataset attributes: Fig(1.2)

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

Fig(1.2)

1. lread (Reads between system memory and user memory): - The data suggests a wide range of values, with a maximum value significantly higher than the mean. There are observations with minimum values of 0, indicating periods with no reads.

2. lwrite (Writes between system memory and user memory):- Similar to lread, this attribute also exhibits a wide range of values.Some observations have minimum values of 0, implying no writes.

3. scall (Number of system calls):- The data shows substantial variation, with a relatively high standard deviation.The minimum value is 109, and the maximum value is 12,493.

4. sread (Number of system read calls):- This attribute also demonstrates variability, with some observations having relatively high values.The minimum value is 6, and the maximum value is 5,318.

5. swrite (Number of system write calls):- Similar to sread, there is variability in the number of write calls.Some observations have low values, including 0.
6. fork (Number of system fork calls):- The number of fork calls varies, with a maximum value of 20.12. Some periods have no fork calls (minimum value of 0).

7. exec (Number of system exec calls):- The data shows a range of values with a maximum of 59.56.Some observations have no exec calls (minimum value of 0).

8. rchar (Characters transferred by system read calls):- The data demonstrates significant variability, with a high standard deviation.The minimum value is 278, and the maximum value is 2,526,649.

9. wchar (Characters transferred by system write calls):- Similar to rchar, this attribute exhibits variability. The minimum value is 1,498, and the maximum value is 1,801,623.

10. pgout (Number of page out requests):- The data suggests variability, with some observations having a maximum value of 81.44.There are periods with no page out requests (minimum value of 0).

11. ppgout (Number of pages paged out):- Similar to pgout, there is variability, with some periods having no pages paged out (minimum value of 0).

12. pgfree (Number of pages placed on the free list):- The data demonstrates variation, with some periods having no pages placed on the free list (minimum value of 0).

13. pgscan (Number of pages checked for freeing):- There is variability in the number of pages checked for freeing, including periods with no scans (minimum value of 0).

14. atch (Number of page attaches):- The data shows variability, with some periods having no page attaches (minimum value of 0).

15. pgin (Number of page-in requests):- The number of page-in requests varies, with a maximum value of 141.20.

16. ppgin (Number of pages paged in):- Similar to pgin, this attribute demonstrates variability, with some periods having no pages paged in (minimum value of 0).

17. pflt (Number of page faults caused by protection errors):- The data suggests variability in page faults, with a maximum value of 899.80.

18. vflt (Number of page faults caused by address translation):- Similar to pflt, there is variation in page faults, with a maximum value of 1,365.00.

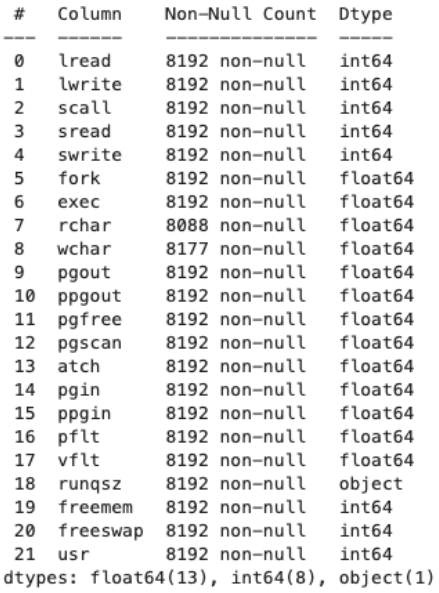
19. freemem (Number of memory pages available to user processes):- The data exhibits variability, with a maximum value of 12,027.

20. freeswap (Number of disk blocks available for page swapping):- This attribute shows variability, with a maximum value of 2,243,187.

21. usr (Portion of time CPUs run in user mode):- The data indicates that CPUs spend a significant portion of time in user mode, with a mean of approximately 83%.

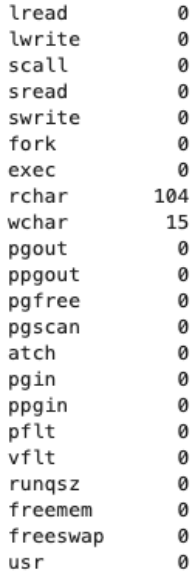
Overall, the dataset contains attributes with varying ranges and degrees of variability. Further analysis, including data visualization, modeling, and hypothesis testing, may be necessary to gain deeper insights and draw meaningful conclusions about the relationships between these attributes and the target variable or other aspects of the data

.



```
 #   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)
```

Fig(1.3)



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

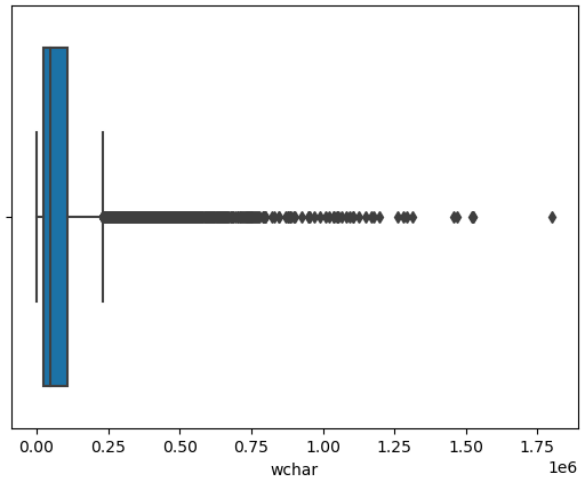
Fig(1.4)

Fig (1.3,1.4). Most of the columns have non-null counts of 8192, indicating that the majority of the data points are complete for many of the columns.

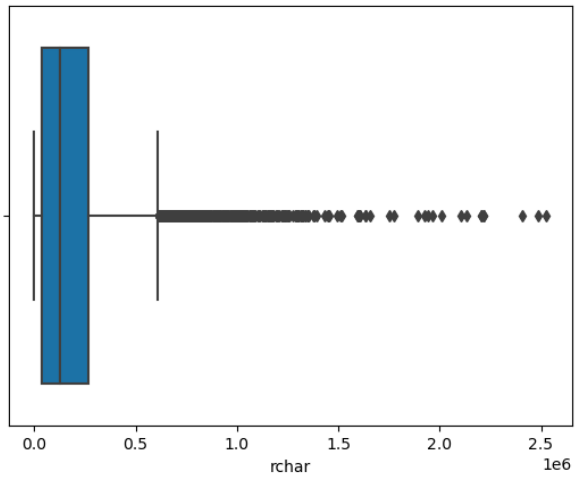
Data Types: - Most columns are of integer (int64) data type, which is appropriate for numeric values. Columns like "fork" and "exec" are of float64 data type, indicating that they may contain decimal values. Column "runqsz" is identified as an object data type.
It appears that some of the attributes in the dataset have **null (missing) values**. Here's a summary of the attributes with null values and the number of missing values for each:

1. rchar: There are 104 missing values in the "rchar" attribute.
2. wchar: There are 15 missing values in the "wchar" attribute.
Fig(1.5),(1.6).As the data is skewed, it is good to consider using the median value for replacing the missing values

- Imputing with the median is a robust approach that replaces missing values with the middle value of the column, which is less sensitive to extreme values than the mean.
- This process ensures that your dataset is ready for further analysis or modeling.



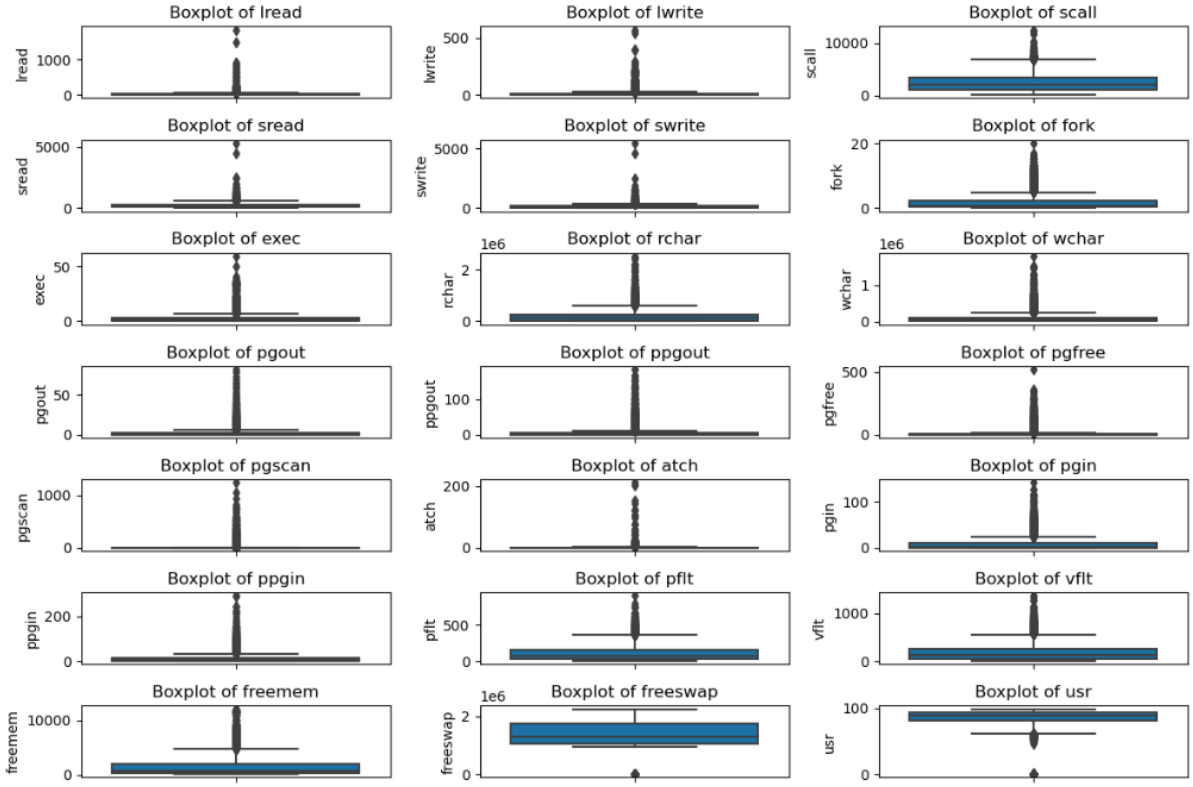
Fig(1.5)



Fig(1.6)

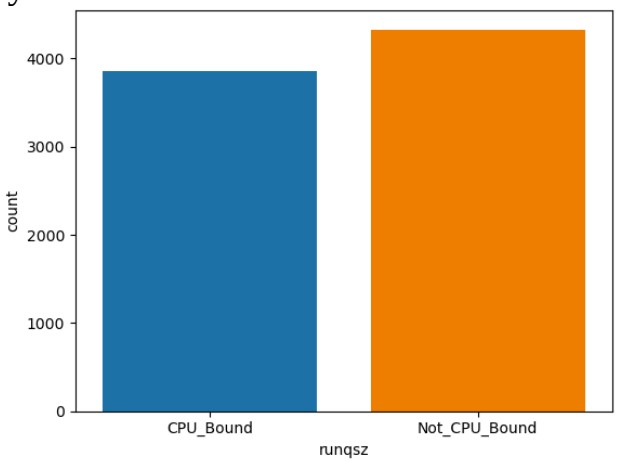
Data Visualization:

Univariate Analysis:



Upon conducting univariate analysis on the dataset, it is evident that the majority of the variables exhibit a right-skewed distribution. (Right-skewness implies that the tail of the distribution extends towards higher values, with relatively few data points having exceptionally high values.)
Additionally, several of these variables exhibit high outliers, indicating the presence of data points that significantly deviate from the typical values in the distribution.

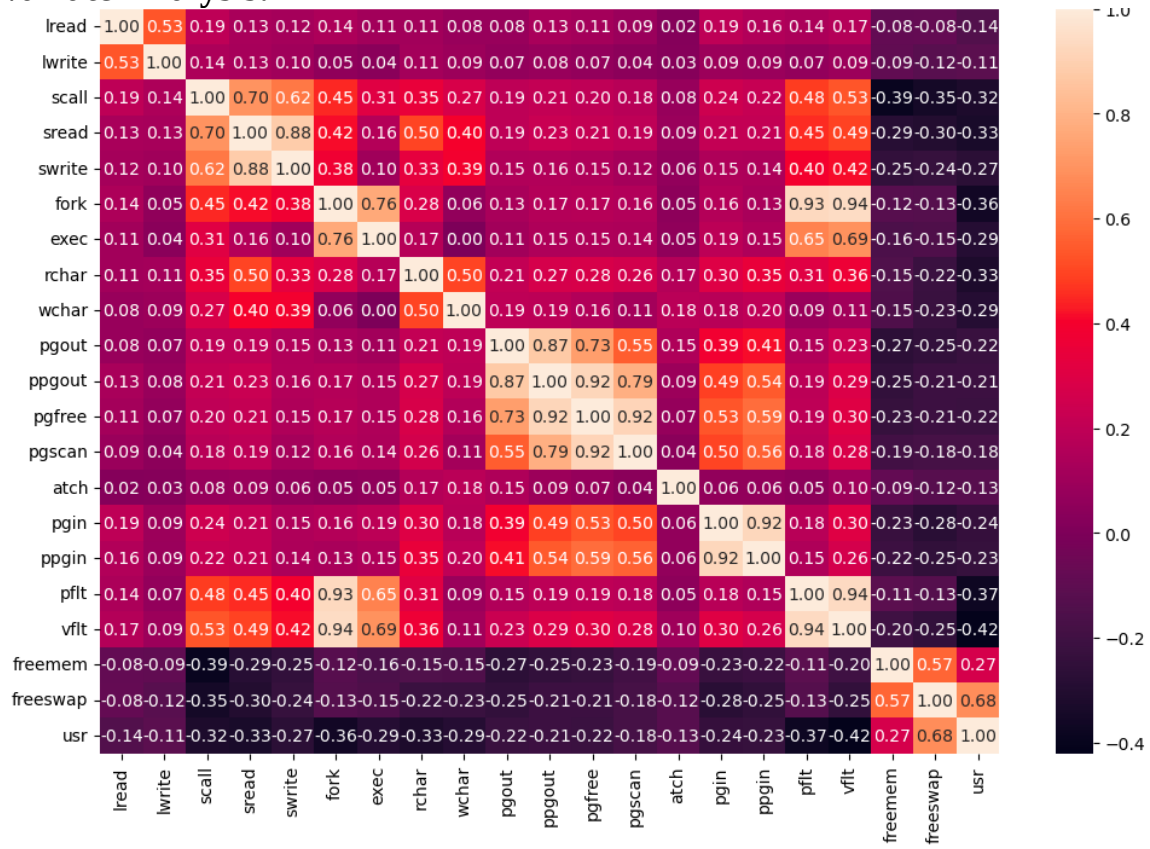
Notably, two variables, namely "usr" and "free swap," do not display right-skewed characteristics. The "usr" and “freeswap’ variable has very few outliers comparatively.



The variable “runqsz” has majority of observations fall into the "Not_CPU_Bound" category, with 4,331 instances.

The "CPU_Bound" category comprises 3,861 observations.

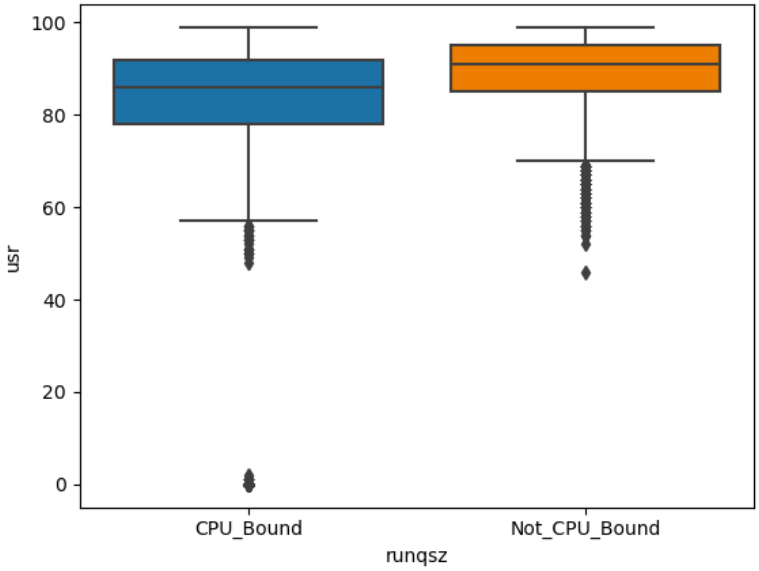
Bivariate Analysis:



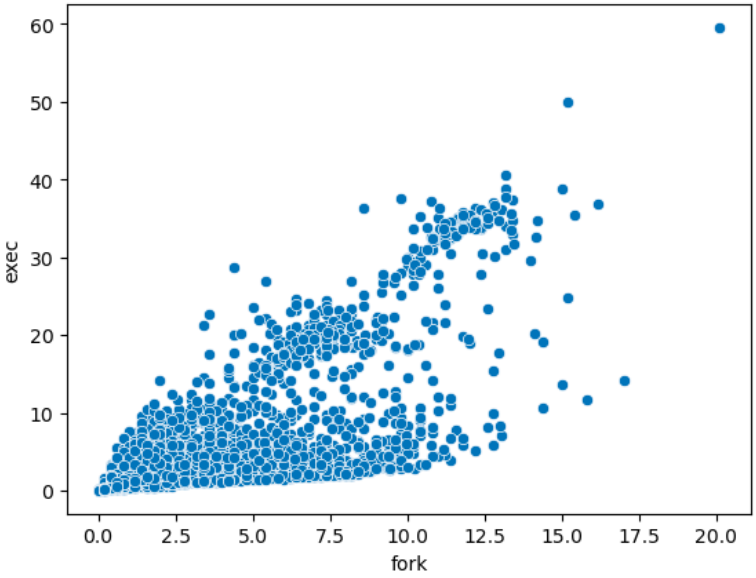
Fig(1.7)

Several variables, including 'fork,' 'pflt,' 'vflt,' 'exec,' 'freeswap,' 'usr,' 'pgscan,' 'pgfree,' 'ppgout,' 'swrite,' 'sread,' and others, exhibit strong correlations among themselves.
While Principal Component Analysis (PCA) could be employed to address multicollinearity, it's important to note that introducing PCA-transformed variables into a Linear Regression model can complicate interpretability. Hence, at present, we have chosen not to utilize PCA in our analysis.

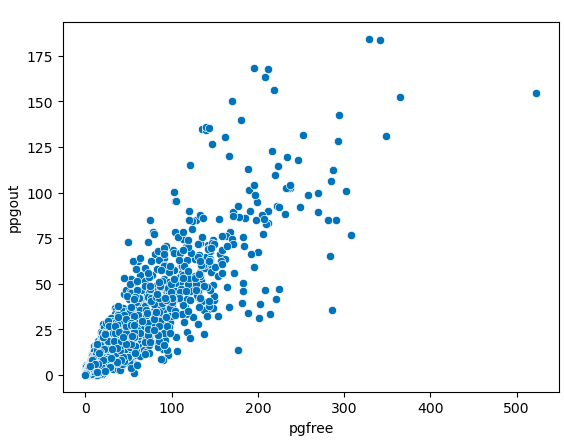
Bivariate Analysis with Categorical Variables:

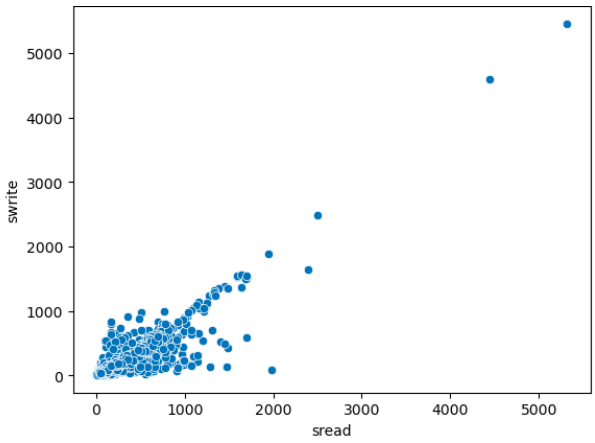


Average portion of time (%) that CPUs run in user mode is higher for the 'Non_CPU_Bound' category (indicating a lower number of kernel threads waiting for a CPU to run) compared to the 'CPU_Bound' category.

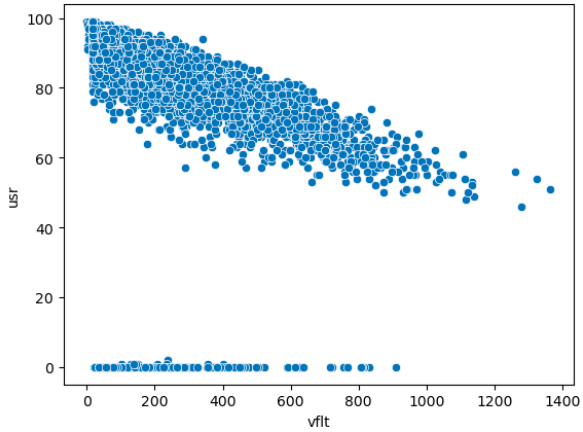


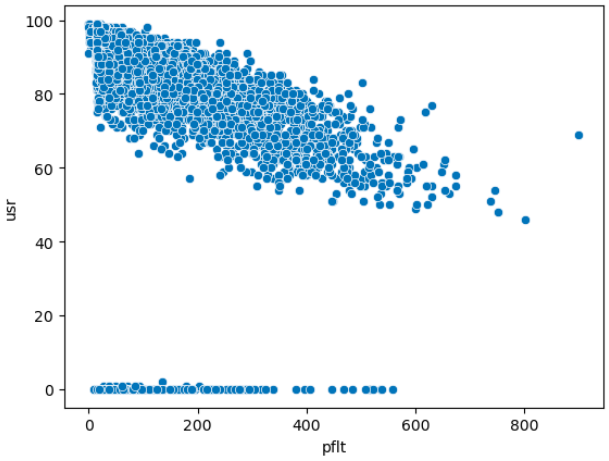
The above graph, represents positive correlation between the variable “fork” and ‘exec’. The higher the number of system fork calls per second, the higher the number of system exec calls per second.



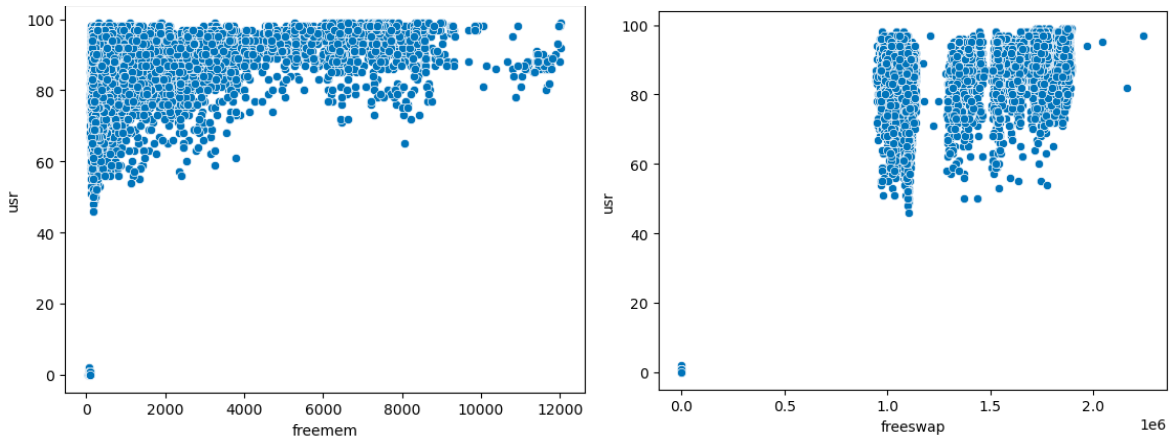


Clearly from the above graphs, it’s clear that the variables possess positive linear relation. The higher the number of system read calls per second, the higher the number of system write calls per second. In the same way, the higher the number of pages, paged out per second, the higher the number of pages per second placed on the free list.



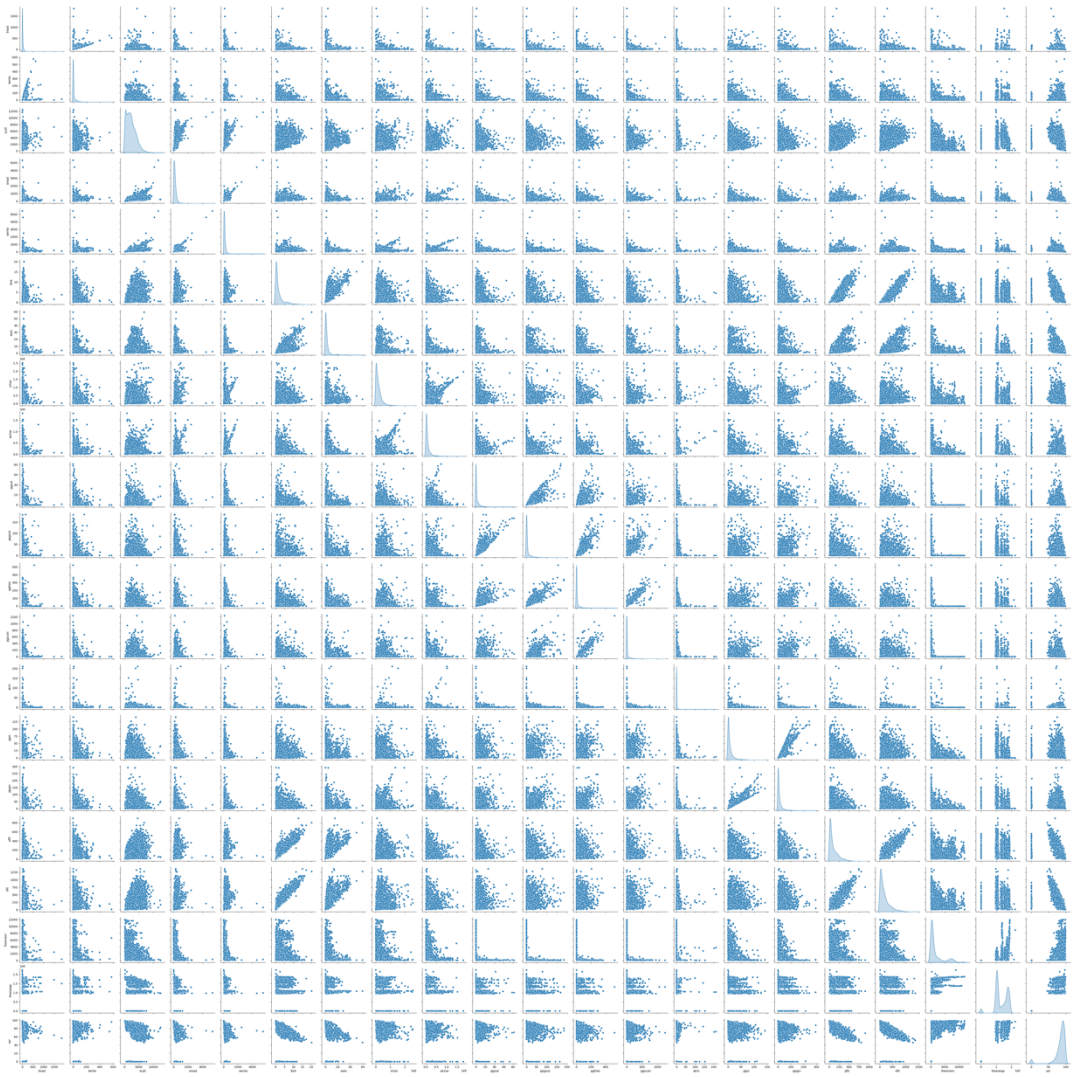


Clearly from the above graphs, it’s clear that the variables possess negative relation. The higher the number of page faults caused by address translation, the lower is the portion of time (%) that cpus run in user mode. The higher the portion of time (%) that cpus run in user mode, the lower is the number of page faults caused by address translation.



Few variables such as “freemem”, “freeswap” etc does not show any relationship between the independent variable. It can be uncovered in the upcoming analysis.

Multivariate Analysis:



1.2 Data Pre-processing:

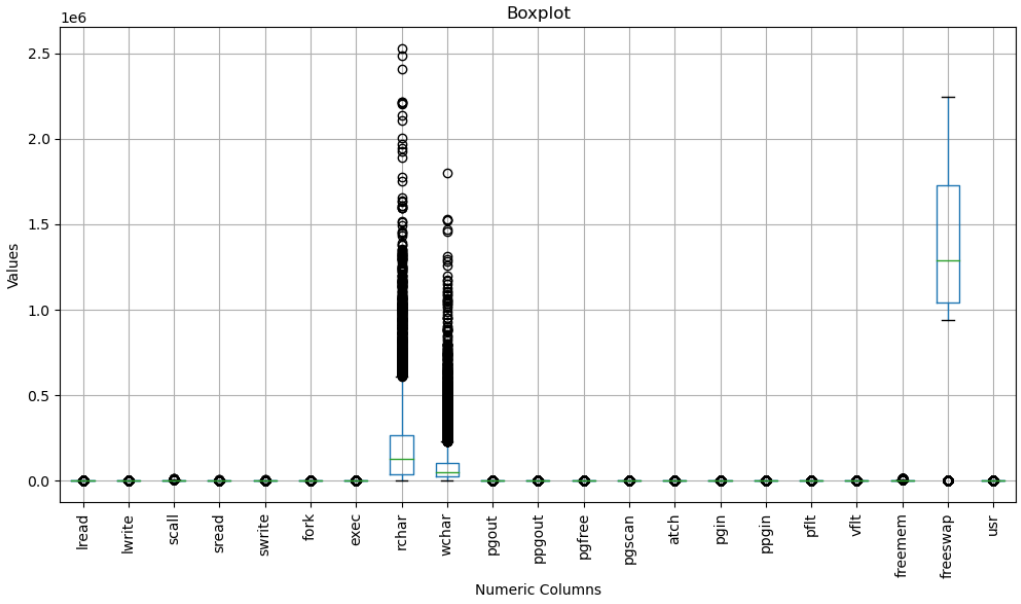
The dataset does not have any **duplicate data**.
Missing values in the ‘rchar’ and ‘wchar’ columns have been imputed with the **median** value. Imputing with the median is a robust approach that replaces missing values with the middle value of the column, which is less sensitive to extreme values than the mean as there are outliers in the dataset.

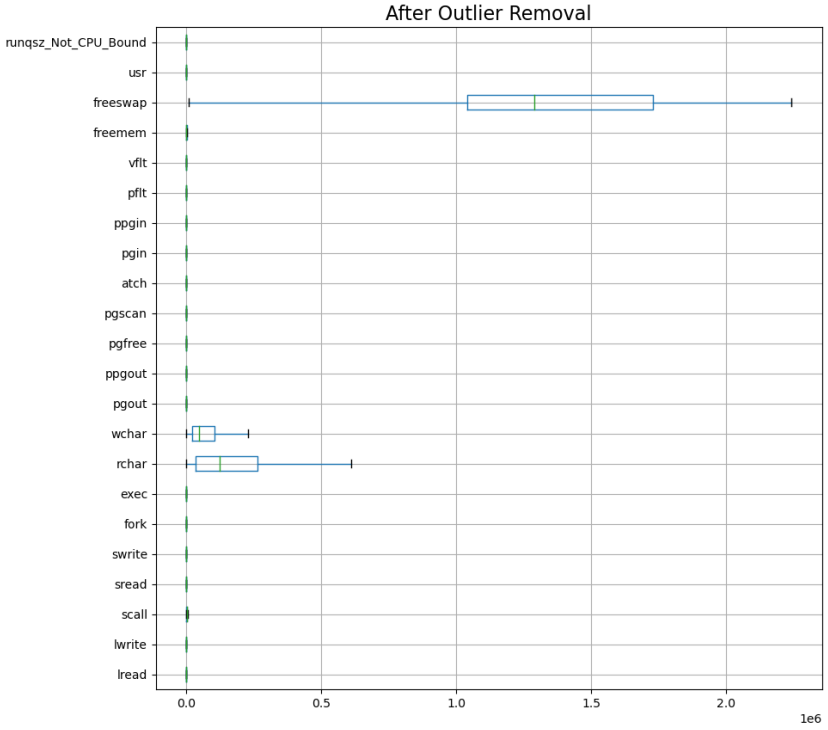
```
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   lread     8192 non-null   float64
 1   lwrite    8192 non-null   float64
 2   scall     8192 non-null   float64
 3   sread     8192 non-null   float64
 4   swrite    8192 non-null   float64
 5   fork      8192 non-null   float64
 6   exec      8192 non-null   float64
 7   rchar     8192 non-null   float64
 8   wchar     8192 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  freemem   8192 non-null   float64
 19  freeswap  8192 non-null   float64
 20  usr       8192 non-null   float64
```

pgout	8192.0	2.285317e+00	5.307038	0.0	0.0	0.0
ppgout	8192.0	5.977229e+00	15.214590	0.0	0.0	0.0
pgfree	8192.0	1.191971e+01	32.363520	0.0	0.0	0.0
pgscan	8192.0	2.152685e+01	71.141340	0.0	0.0	0.0
atch	8192.0	1.127505e+00	5.708347	0.0	0.0	0.0

Few values in the columns are observed to be 0. ‘pgout’ (Number of page out requests per second): A value of zero indicates that during the specific measurement interval, there were no requests to write or swap pages from memory to disk. This might occur when the system is not under memory pressure or when the workload doesn't require paging data out to disk.
Same goes with the metric "lread," which measures reads (transfers per second) between system memory and user memory, can indeed have a minimum value of 0. In fact, it's quite common for metrics like "lread" to have 0 as their minimum values in certain scenarios.
So, we are proceeding with the further steps.

Outliers: Outliers can distort the statistical properties and distribution of the data. Removing or transforming them with the IQR method can improve the overall quality of your dataset.
The IQR method, which involves identifying outliers based on the spread of the data (the difference between the third quartile - Q3 and the first quartile - Q1), is a robust and commonly used technique for handling outliers. It helps strike a balance between retaining valuable data and eliminating extreme values that can distort analysis results.





1.3 Predictive Modelling: Linear Regression

Fig(1.8) represents the ‘runqsz’ column after data-encoding.

lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	freemem	freeswap	usr	runqsz_Not_CPU_Bound
1.0	0.0	2147.0	79.0	68.0	0.2	0.20	40671.0	53995.0	0.00	...	0.00	0.0	1.60	2.60	16.00	26.40	4670.0	1730946.0	95.0	0
0.0	0.0	170.0	18.0	21.0	0.2	0.20	448.0	8385.0	0.00	...	0.00	0.0	0.00	0.00	15.63	16.83	7278.0	1869002.0	97.0	1
15.0	3.0	2162.0	159.0	119.0	2.0	2.40	125473.5	31950.0	0.00	...	0.00	1.2	6.00	9.40	150.20	220.20	702.0	1021237.0	87.0	1
0.0	0.0	160.0	12.0	16.0	0.2	0.20	125473.5	8670.0	0.00	...	0.00	0.0	0.20	0.20	15.60	16.80	7248.0	1863704.0	98.0	1
5.0	1.0	330.0	39.0	38.0	0.4	0.40	125473.5	12185.0	0.00	...	0.00	0.0	1.00	1.20	37.80	47.60	633.0	1760253.0	90.0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
16.0	12.0	3009.0	360.0	244.0	1.6	5.81	405250.0	85282.0	8.02	...	55.11	0.6	35.87	47.90	139.28	270.74	387.0	986647.0	80.0	0
4.0	0.0	1596.0	170.0	146.0	2.4	1.80	89489.0	41764.0	3.80	...	0.20	0.8	3.80	4.40	122.40	212.60	263.0	1055742.0	90.0	1
16.0	5.0	3116.0	289.0	190.0	0.6	0.60	325948.0	52640.0	0.40	...	0.00	0.4	28.40	45.20	60.20	219.80	400.0	969106.0	87.0	1
32.0	45.0	5180.0	254.0	179.0	1.2	1.20	62571.0	29505.0	1.40	...	18.04	0.4	23.05	24.25	93.19	202.81	141.0	1022458.0	83.0	0
2.0	0.0	985.0	55.0	46.0	1.6	4.80	111111.0	22256.0	0.00	...	0.00	0.2	3.40	6.20	91.80	110.00	659.0	1756514.0	94.0	0

Fig(1.8)

For Linear Regression, the data is split in the ratio 70:30.

Using OLS 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:	Fri, 29 Sep 2023	Prob (F-statistic):	0.00
Time:	12:52:51	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	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	-3.035e-14	1.45e-16	-209.209	0.000	-3.06e-14	-3.01e-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
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
runqsz_Not_CPU_Bound	1.6153	0.126	12.819	0.000	1.368	1.862

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.	4.18e+22

Now, let's make some inferences based on this information:

The coefficient for the constant (intercept) is approximately 84.1217.

The adjusted R-squared value of 0.795 suggests that the model with these variables explains approximately 79.5% of the variability in the dependent variable "usr."

Several independent variables have statistically significant coefficients ($p < 0.05$), indicating that they have a significant impact on the dependent variable (usr). These variables include "lread," "lwrite," "scall," "exec," "rchar," "wchar," "pgout," "atch," "pflt," "vflt," "freemem," "freeswap," and "runqsz_Not_CPU_Bound."

For example, the "exec" variable has a negative coefficient of approximately -0.3212, indicating that an increase in "exec" is associated with a decrease in "usr" (holding other variables constant).

In conclusion, this analysis suggests that several independent variables in the model are statistically significant predictors of the dependent variable "usr."

```
VIF values:

const                 29.229332
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
freemem                1.961304
freeswap               1.841239
runqsz_Not_CPU_Bound   1.156815
dtype: float64
```

Here's what the VIF values suggest:

A VIF of 1 indicates no multicollinearity. Generally, VIF values below 5 are considered acceptable, although the threshold can vary depending on the context.

VIF values above 5 or 10 are often considered indicative of problematic multicollinearity.

Based on the provided VIF values:

Some other variables also have high VIF values, such as "ppgout" (VIF = 29.40), "pgfree" (VIF = 16.50), "pflt" (VIF = 12.00), and "vflt" (VIF = 15.97). These high VIF values suggest that these variables may be highly correlated with other independent variables in the model. High multicollinearity can make it challenging to interpret the individual effects of these variables accurately.

On the other hand, several variables have relatively low VIF values, indicating lower multicollinearity. These include "scall," "exec," "rchar," "wchar," "atch," "freemem," "freeswap," and "runqsz_Not_CPU_Bound."

We remove those predictors with multicolinarity due to which there is least impact on the adjusted R2.

After dropping the features causing strong multicollinearity (i.e, ‘ppgout’, ‘vflt’, ‘pgin’, ‘fork’) and the statistically insignificant ones, our model performance hasn't dropped sharply. This shows that these variables did not have much predictive power.

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.795
Model:	OLS	Adj. R-squared:	0.795
Method:	Least Squares	F-statistic:	1389.
Date:	Fri, 29 Sep 2023	Prob (F-statistic):	0.00
Time:	12:52:57	Log-Likelihood:	-16667.
No. Observations:	5734	AIC:	3.337e+04
Df Residuals:	5717	BIC:	3.348e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.0586	0.311	270.133	0.000	83.449	84.669
lread	-0.0664	0.009	-7.412	0.000	-0.084	-0.049
lwrite	0.0506	0.013	3.863	0.000	0.025	0.076
scall	-0.0007	6.24e-05	-10.630	0.000	-0.001	-0.001
sread	2.026e-06	0.001	0.002	0.998	-0.002	0.002
swrite	-0.0058	0.001	-4.137	0.000	-0.009	-0.003
exec	-0.3604	0.049	-7.430	0.000	-0.455	-0.265
rchar	-5.326e-06	4.86e-07	-10.963	0.000	-6.28e-06	-4.37e-06
wchar	-4.846e-06	1.02e-06	-4.738	0.000	-6.85e-06	-2.84e-06
pgout	-0.4193	0.068	-6.178	0.000	-0.552	-0.286
pgfree	0.0397	0.029	1.363	0.173	-0.017	0.097
pgscan	-1.376e-15	1.14e-15	-1.210	0.226	-3.61e-15	8.53e-16
atch	0.5934	0.143	4.163	0.000	0.314	0.873
ppgin	-0.0622	0.007	-9.432	0.000	-0.075	-0.049
pflt	-0.0398	0.001	-37.312	0.000	-0.042	-0.038
freemem	-0.0005	5.07e-05	-9.194	0.000	-0.001	-0.000
freeswap	8.928e-06	1.87e-07	47.823	0.000	8.56e-06	9.29e-06
runqsz_Not_CPU_Bound	1.6043	0.126	12.721	0.000	1.357	1.852

Omnibus:	1049.833	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2217.488
Skew:	-1.075	Prob(JB):	0.00
Kurtosis:	5.158	Cond. No.	1.82e+22

VIF values:

const	28.247467
lread	5.313325
lwrite	4.307722
scall	2.913367
sread	6.373415
swrite	5.389533
exec	2.852464
rchar	2.113026
wchar	1.550562
pgout	6.442834
pgfree	6.136186
pgscan	NaN
atch	1.862881
ppgin	1.558638
pflt	3.463109
freemem	1.957438
freeswap	1.772123
runqsz_Not_CPU_Bound	1.155723

dtype: float64

Each coefficient in the equation signifies the weight or influence of the corresponding independent variable on the predicted 'usr' value.

Each coefficient quantifies the change in the predicted 'usr' value associated with a one-unit change in the corresponding independent variable, assuming all other variables remain constant. For instance, if 'lread' increases by one unit while keeping all other variables constant, the 'usr' value is expected to decrease by approximately 0.0664.

The equation:

usr = 84.05 + -0.066 * (lread) + 0.050 * (lwrite) + -0.0006 * (scall) + 2.0263043186531464e-06 * (sread) + -0.00583 * (swrite) + -0.360 * (exec) + -5.325655903933241e-06 * (rchar) + -4.846280495168696e-06 * (wchar) + -0.4192 * (pgout) + 0.0397* (pgfree) + -1.3758765788197688e-15 * (pgscan) + 0.5933 * (atch) + -0.0622 * (ppgin) + -0.039 * (pflt) + -0.00046 * (freemem) + 8.927920908386326e-06 * (freeswap) + 1.604311* (runqsz_Not_CPU_Bound)

Root Mean Squared Error (RMSE) on training data: 4.4268
Root Mean Squared Error (RMSE) on test data: 4.665

The Root Mean Squared Error (RMSE) is a common metric used to measure the accuracy of a predictive model, typically a regression model. It quantifies the difference between the actual values (or target values) and the predicted values produced by the model

1. RMSE on training data: 4.4268: This value, 4.4268, represents the average error (in the same units as the dependent variable) that the model makes when predicting the training data.

2. RMSE on test data: 4.665: This value, 4.665, represents the average error the model makes when predicting new, unseen data.

The RMSE is a measure of how well the model's predictions match the actual values. A lower RMSE indicates that the model's predictions are closer to the true values, which is generally desirable.

Linear Regression using (sklearn):

```
The coefficient for const is 0.0
The coefficient for lread is -0.06639749302628149
The coefficient for lwrite is 0.05062949916519846
The coefficient for scall is -0.0006633462059539676
The coefficient for sread is 2.026304334694975e-06
The coefficient for swrite is -0.0058346740662969096
The coefficient for exec is -0.36039218359231795
The coefficient for rchar is -5.325655904027929e-06
The coefficient for wchar is -4.846280495177564e-06
The coefficient for pgout is -0.4192847257304126
The coefficient for pgfree is 0.0397383894817412
The coefficient for pgscan is 1.6653345369377348e-16
The coefficient for atch is 0.5933871975567543
The coefficient for ppgin is -0.06223579575800658
The coefficient for pflt is -0.0397694274423693
The coefficient for freemem is -0.00046652122143602855
The coefficient for freeswap is 8.927920908410336e-06
The coefficient for runqsz_Not_CPU_Bound is 1.6043116109101607
```

The intercept for our model is 84.05856335896442
79% of the variation in the usr is explained by the predictors in the model for train set
76% of the variation in the usr is explained by the predictors in the model for test set
Root Mean Squared Error (RMSE) on training data: 4.4268
Root Mean Squared Error (RMSE) on test data: 4.665.

1.4 Inference

1. Optimizing System Performance:

- lread, lwrite, sread, swrite: These system measures have positive coefficients, suggesting that increasing the rates of reads, writes, and data transfers between system and user memory may lead to higher user mode CPU usage (usr). This could imply that optimizing data transfer and I/O operations can improve system performance.

- pgfree: The positive coefficient for "pgfree" indicates that placing more pages on the free list per second is associated with higher usr time. This suggests that maintaining a sufficient number of free pages in memory is essential for user mode performance.

- atch: The positive coefficient for "atch" implies that increasing the number of page attaches per second can contribute to higher usr time. This suggests that efficiently handling page attachments, which help satisfy page faults, is important for user mode performance.

- freeswap: The positive coefficient for "freeswap" suggests that having more disk blocks available for page swapping is associated with higher usr time. Ensuring ample swap space can be beneficial for system responsiveness.

2. Resource Management and Optimization:

- scall, exec, pgout, ppgin, pflt, freemem: These system measures have negative coefficients, indicating that higher values of these metrics are associated with lower usr time. This suggests that optimizing these aspects of system performance, such as reducing system calls, optimizing memory usage, and handling page faults efficiently, can lead to improved usr time.

- runqsz_Not_CPU_Bound: The positive coefficient for "runqsz_Not_CPU_Bound" suggests that when the system is not CPU-bound (i.e., when there are fewer threads waiting for a CPU), usr time tends to be higher. It may be worthwhile to investigate and reduce CPU-bound conditions.

3. Performance Monitoring and Alerts:

- Implement monitoring systems that continuously track these system measures and usr time. Set up alerts to notify administrators when unusual patterns or performance issues are detected.

4. Capacity Planning:

- Use the insights from the model to guide capacity planning efforts. Understanding how various system measures impact usr time can help allocate resources effectively and plan for future hardware upgrades.

5. Documentation and Training:

- Ensure that system administrators and operators understand the relationship between these system measures and usr time. Provide training and documentation on how to interpret and respond to these metrics effectively.

6. Fault Tolerance and Redundancy:

- Implement fault tolerance mechanisms to handle unexpected spikes in usr time. Ensure that critical systems have redundancy to maintain user mode performance even in the face of hardware failures.

7. Continuous Improvement:

- Regularly review and update your monitoring and optimization strategies based on changing system demands and hardware configurations. Continuously seek opportunities to improve system performance and user mode CPU usage.

Problem 2:
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

Method used for prediction:

Logistic Regression :-

Logit classifier is a supervised learning method for classification. It establishes relationship between dependent class variable and independent variables using regression. The dependent variable is categorical

LDA(Linear Discriminant Analysis) :-

LDA is used for classifying observations to a class or category based on predictor(independent variable).
LDA creats a model to predict the classes of the new or future observation.

CART :-

CART is a predictive algorithm. It explains how the target variable's values can be predicted based on other matters. It is a decision tree where each fork is split into a predictor variable and each node has a prediction for the target variable at the end.

Context:

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.

2.1 Exploratory Data Analysis:

	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_
0	24.0	Primary	Secondary	3.0	Scientology	No	2	High	Exposed	
1	45.0	Uneducated	Secondary	10.0	Scientology	No	3	Very High	Exposed	
2	43.0	Primary	Secondary	7.0	Scientology	No	3	Very High	Exposed	
3	42.0	Secondary	Primary	9.0	Scientology	No	3	High	Exposed	
4	36.0	Secondary	Secondary	8.0	Scientology	No	3	Low	Exposed	
...	...	...	...	...	...	...	...	...	...	
1468	33.0	Tertiary	Tertiary	NaN	Scientology	Yes	2	Very High	Exposed	
1469	33.0	Tertiary	Tertiary	NaN	Scientology	No	1	Very High	Exposed	
1470	39.0	Secondary	Secondary	NaN	Scientology	Yes	1	Very High	Exposed	
1471	33.0	Secondary	Secondary	NaN	Scientology	Yes	2	Low	Exposed	
1472	17.0	Secondary	Secondary	1.0	Scientology	No	2	Very High	Exposed	

1473 rows × 10 columns

Fig(2.1). Dataset Size: Dataset contains 1473 rows and 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)
```

The dataset exhibits the following characteristics:

1. Missing Values: There are instances of missing data within the dataset, particularly in the "Number of children ever born" and “Wife_age” attribute. These missing values need to be addressed before proceeding with analysis.

2. Data Types: The dataset contains a mix of data types, including:
Numerical Data: "Wife's age", "Husband's occupation" ,"Number of children ever born" are represented as floating-point numbers.
Categorical Data: Several features, such as "Wife's education," "Husband's education," "Wife's religion," "Wife's employment status,"," "Standard of living index," and "Media exposure," are recorded as categorical variables.

Addressing missing values and appropriately encoding categorical variables are crucial steps in preparing the dataset for further analysis or modelling.

	count	mean	std	min	25%	50%	75%	max
Wife_age	1402.0	32.606277	8.274927	16.0	26.0	32.0	39.0	49.0
No_of_children_born	1452.0	3.254132	2.365212	0.0	1.0	3.0	4.0	16.0
Husband_Occupation	1473.0	2.137814	0.864857	1.0	1.0	2.0	3.0	4.0

Here's a summary of the descriptive statistics of the numerical attributes in the dataset:

1. Wife's Age:
    - Count: 1402 observations out of 1473 observations
    - Mean: 32.61 years
    - Standard Deviation: 8.27 years
    - Minimum Age: 16 years
    - 25th Percentile (Q1): 26 years
    - Median (Q2 or 50th Percentile): 32 years
    - 75th Percentile (Q3): 39 years
    - Maximum Age: 49 years

2. Number of Children Ever Born:
    - Count: 1452 observations out of 1473 observations
    - Mean: 3.25 children
    - Standard Deviation: 2.37 children
    - Minimum: 0 children (indicating no children)
    - 25th Percentile (Q1): 1 child
    - Median (Q2 or 50th Percentile): 3 children
    - 75th Percentile (Q3): 4 children
    - Maximum: 16 children (indicating the highest number of children reported)

3. Husband's Occupation:
    - Count: 1473 observations
    - Mean: 2.14 (rounded to 2 decimal places)
    - Standard Deviation: 0.86 (rounded to 2 decimal places)
    - Minimum: 1 (indicating the lowest category of occupation)
    - 25th Percentile (Q1): 1 (25% of the observations have an occupation code of 1)
    - Median (Q2 or 50th Percentile): 2 (the median occupation category)
    - 75th Percentile (Q3): 3 (75% of the observations have an occupation code of 3)
    - Maximum: 4 (indicating the highest category of occupation)

The data has 80 duplicate records.

To enhance the quality of our analysis, we have identified and removed 80 duplicate records from the dataset. This step was taken to ensure that our analysis is based on unique and non-repetitive data entries.
Fig(2.2) shows the info after removing duplicate records.

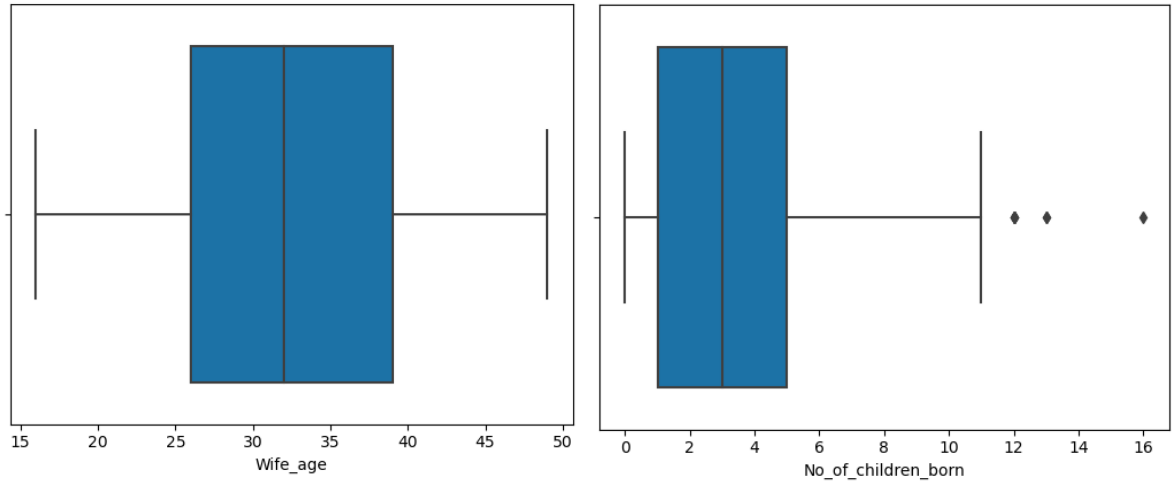
```
Data columns (total 10 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Wife_age                  1326 non-null   float64
 1   Wife_ education           1393 non-null   int64
 2   Husband_education         1393 non-null   int64
 3   No_of_children_born       1372 non-null   float64
 4   Wife_religion             1393 non-null   int64
 5   Wife_Working              1393 non-null   int64
 6   Husband_Occupation        1393 non-null   int64
 7   Standard_of_living_index  1393 non-null   int64
 8   Media_exposure            1393 non-null   int64
 9   Contraceptive_method_used 1393 non-null   int64
dtypes: float64(2), int64(8)
memory usage: 119.7 KB
```

Fig(2.2)

There are 67 **missing values** in the column 'Wife_age' and 21 missing values in the column "No_of_children_born".

```
Wife_age                    67
Wife_ education               0
Husband_education             0
No_of_children_born          21
Wife_religion                 0
Wife_Working                  0
Husband_Occupation            0
Standard_of_living_index      0
Media_exposure                0
Contraceptive_method_used     0
dtype: int64
```

Using Median to impute the missing values. As we do not want the "wife_age" and the "no_of_children_born" feature to have float value.



The median of the feature **"wife_age"** is 32years and **"no_of_children_born"** is 3. Imputing all the missing values with the same.Fig(2.3) shows 0 missing values.

```
Wife_age                    0
Wife_ education              0
Husband_education            0
No_of_children_born          0
Wife_religion                0
Wife_Working                 0
Husband_Occupation           0
Standard_of_living_index     0
Media_exposure               0
Contraceptive_method_used    0
```

Fig(2.3)

Prior to conducting any analysis, it is necessary to transform the data by encoding categorical variables, which are represented as strings, into a numerical format.

Fig(2.4) shows the glimpse of the dataset, encoded based on the given Data Dictionary inputs.

	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_inde
0	24.0	2	3	3.0	1	0	2	
1	45.0	1	3	10.0	1	0	3	
2	43.0	2	3	7.0	1	0	3	
3	42.0	3	2	9.0	1	0	3	
4	36.0	3	3	8.0	1	0	3	
...	...	...	...	...	...	...	...	
1468	33.0	4	4	NaN	1	1	2	
1469	33.0	4	4	NaN	1	0	1	
1470	39.0	3	3	NaN	1	1	1	
1471	33.0	3	3	NaN	1	1	2	
1472	17.0	3	3	1.0	1	0	2	

1393 rows × 10 columns

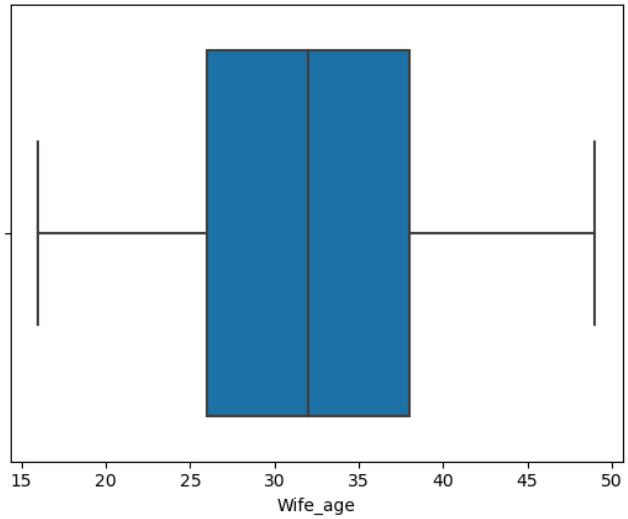
Fig(2.4)

All of the columns are now changed to numerical values.

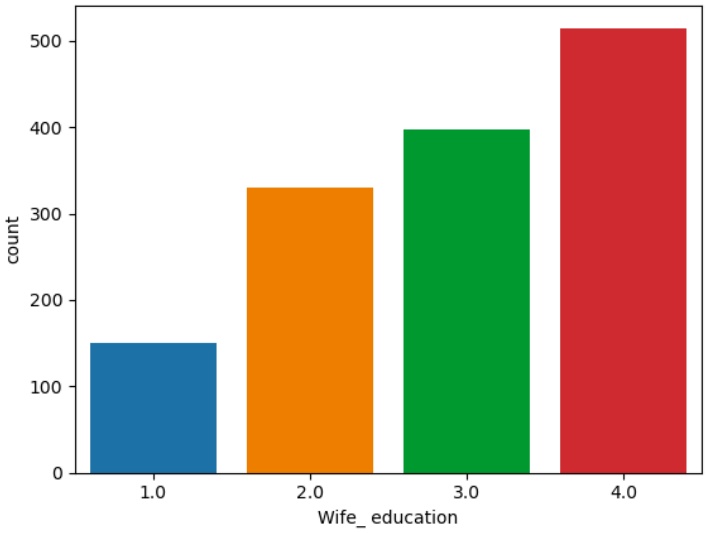
```
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Wife_age                1393 non-null   float64
 1   Wife_ education         1393 non-null   float64
 2   Husband_education       1393 non-null   float64
 3   No_of_children_born     1393 non-null   float64
 4   Wife_religion           1393 non-null   float64
 5   Wife_Working            1393 non-null   float64
 6   Husband_Occupation      1393 non-null   float64
 7   Standard_of_living_index 1393 non-null   float64
 8   Media_exposure          1393 non-null   float64
 9   Contraceptive_method_used 1393 non-null   float64
```

Data Visualization:

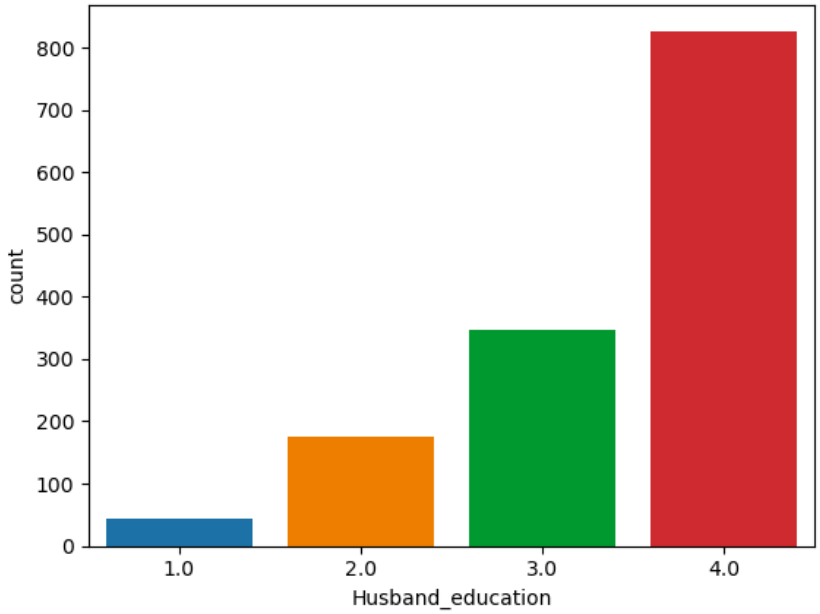
Univariate Analysis:



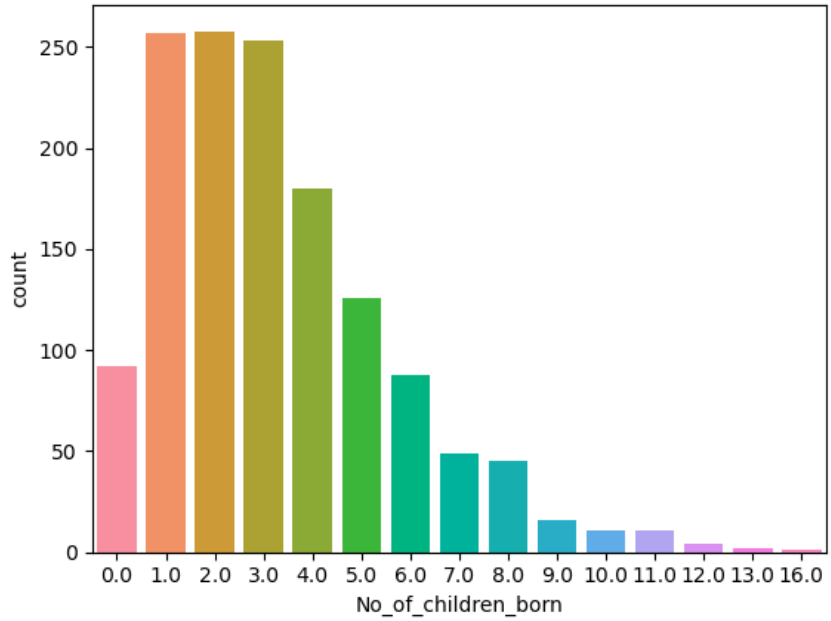
From the graph, we can see that the dataset includes a wide age range of married women. The youngest wife recorded is 16 years old, while the oldest is 49 years old. On average, the age of the women in the dataset is approximately 32 years.



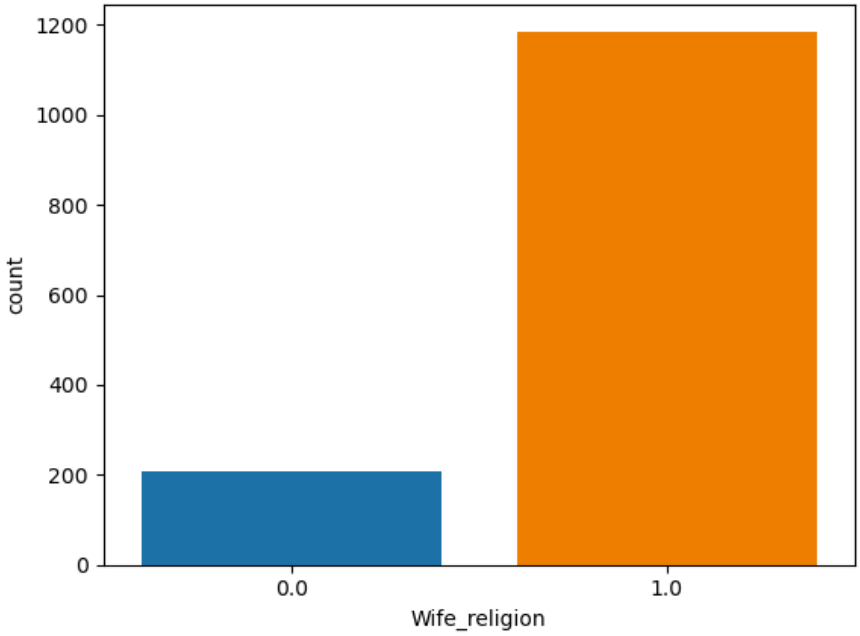
The average education level among these individuals is approximately 2.92, indicating a moderate level of education.(School or Primary level of Education) A significant proportion of wives in the dataset have an education level of 4, as indicated by the high count of 515 individuals in this category. This suggests that a substantial number of wives have attained a relatively higher level of education, which could be an important factor to consider in further analysis or modeling.



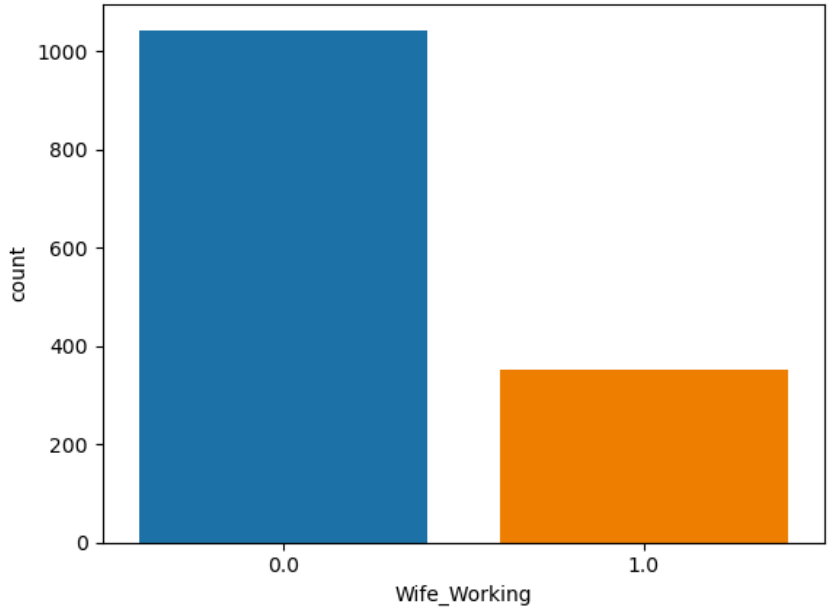
The median education level among husbands is 4(Tertiary Education). This indicates that a substantial number of husbands have achieved a high level of education.



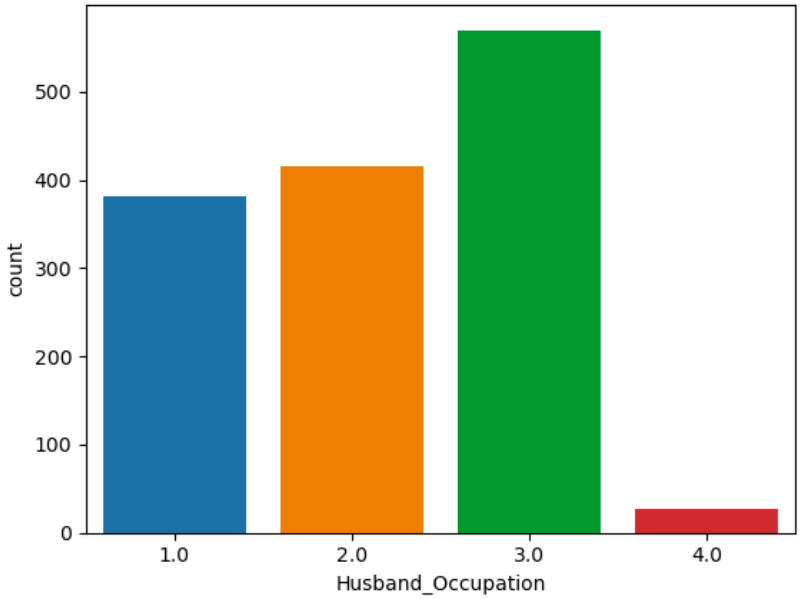
A notable observation regarding the number of children ever born (No_of_children_born) is that a substantial proportion of women in the dataset have reported having either 1, 2, or 3 children. On average, women in the dataset have had approximately 3.29 children, with the range of reported children varying from none to as many as 16.



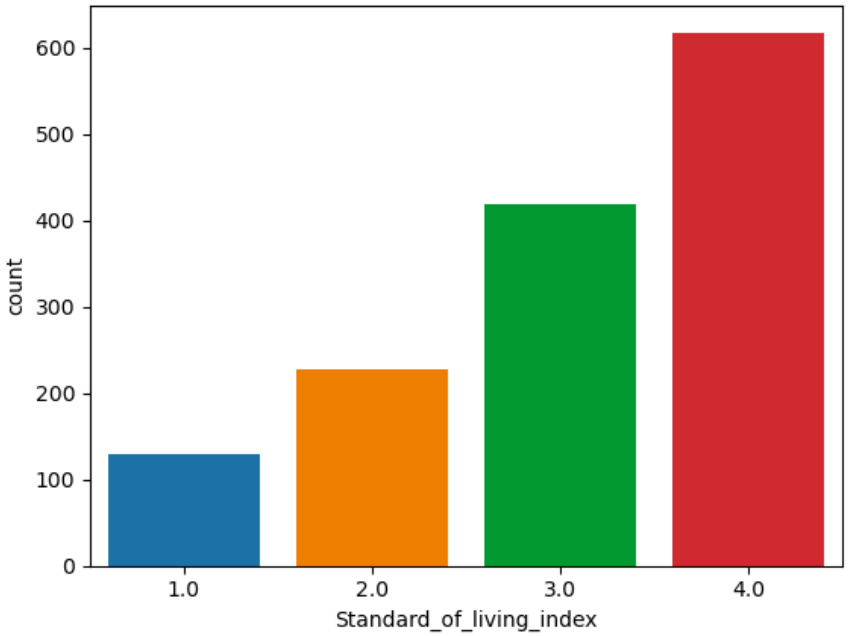
About 85% of the wives belong to the Scientology religion.



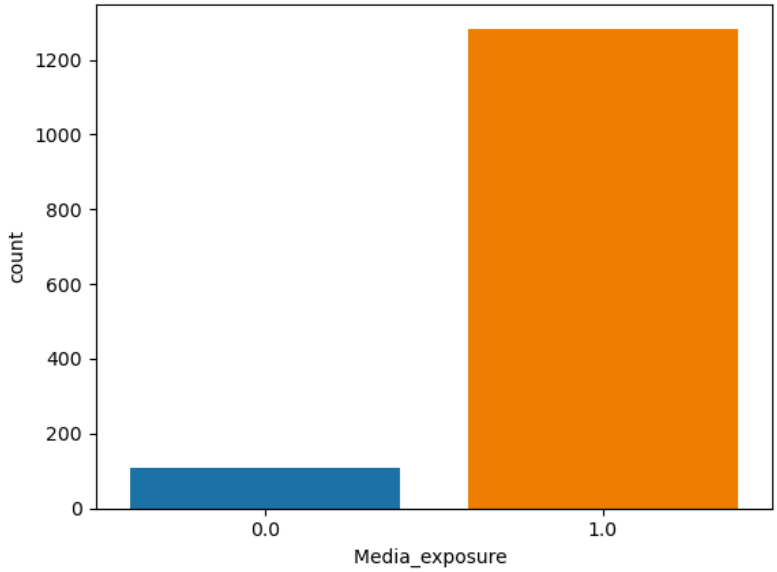
Approximately 25.13% of wives are employed.



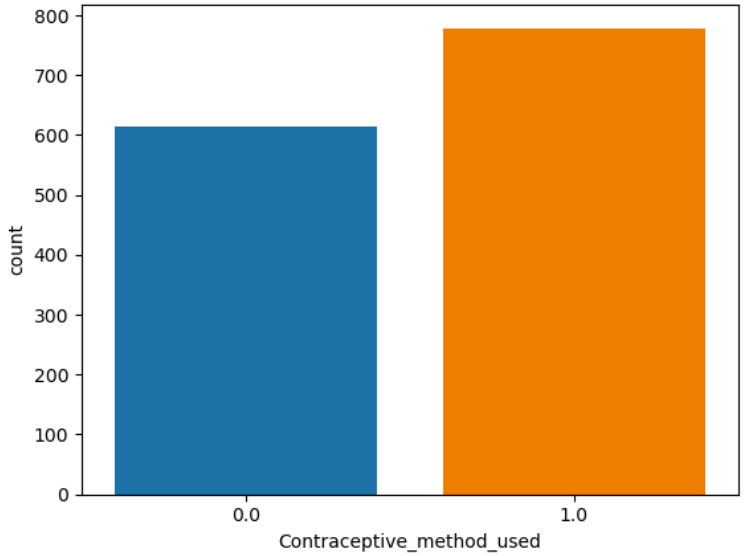
Husbands in the dataset exhibit a range of occupational diversity, spanning occupation codes from 1 to 4. On average, their occupation code is approximately 2.17. Notably, the highest occupation code observed is 3, while the lowest is 1.



A notable observation in the dataset is that a substantial number of wives have a standard of living index of 4, indicating a relatively high standard of living. On average, the standard of living for the wives is approximately 3, reflecting a moderate to high standard across the dataset.

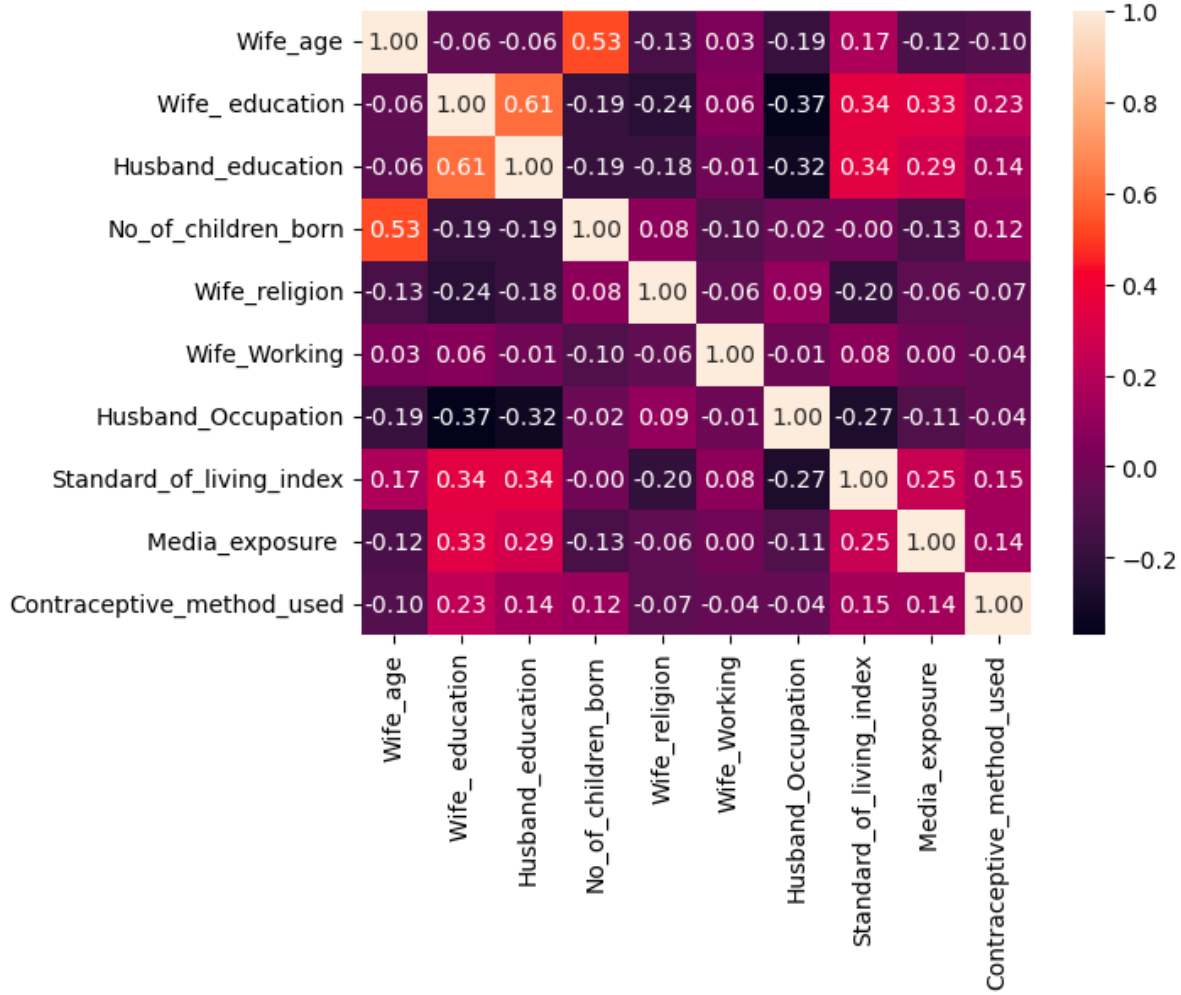


About 92.18% of individuals report having good media exposure.



On average, 55.92% of individuals report using contraceptive methods.

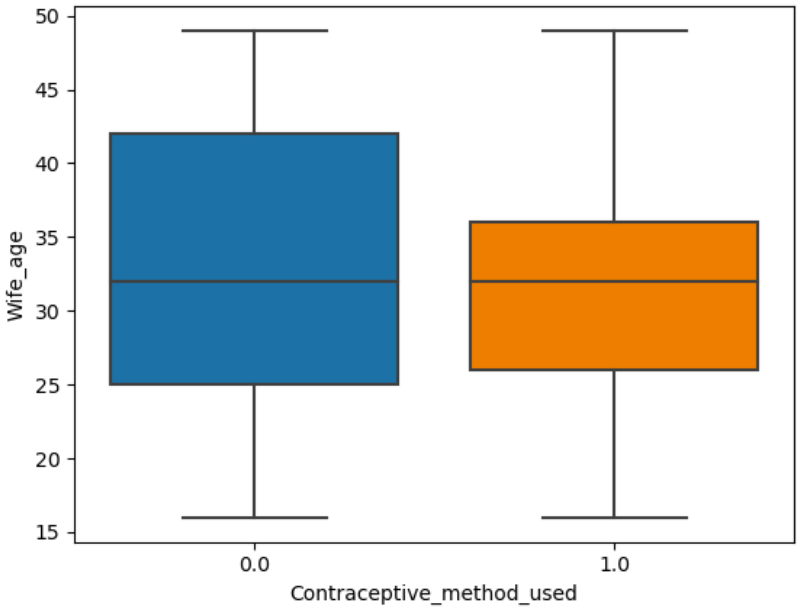
Bi-variate Analysis:



The heatmap visualization reveals strong correlations between certain pairs of variables in the dataset. Specifically, there are notable associations between:

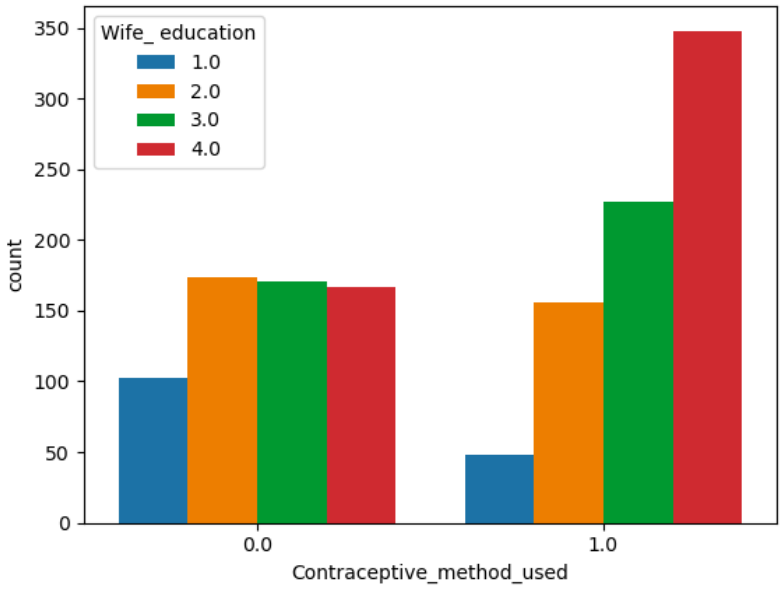
- Wife's age and the number of children ever born.
- Husband's education level and wife's education level.

These correlations indicate that as wife's age increases, the number of children born tends to rise. Additionally, there is a connection between the education levels of husbands and wives, suggesting that they may be related factors within the dataset.

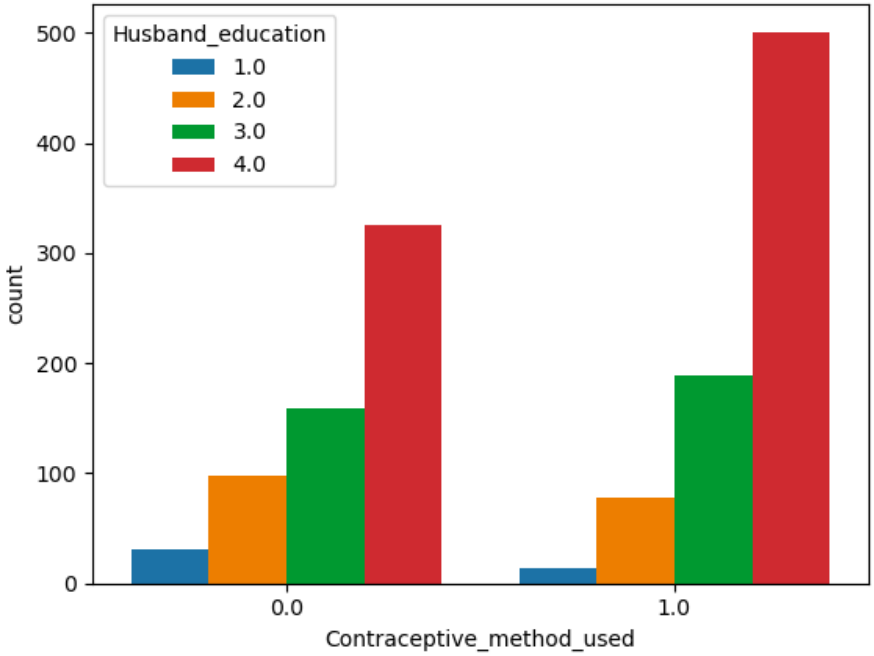


The median age of women who use contraceptives is approximately the same as that of women who do not use contraceptives.

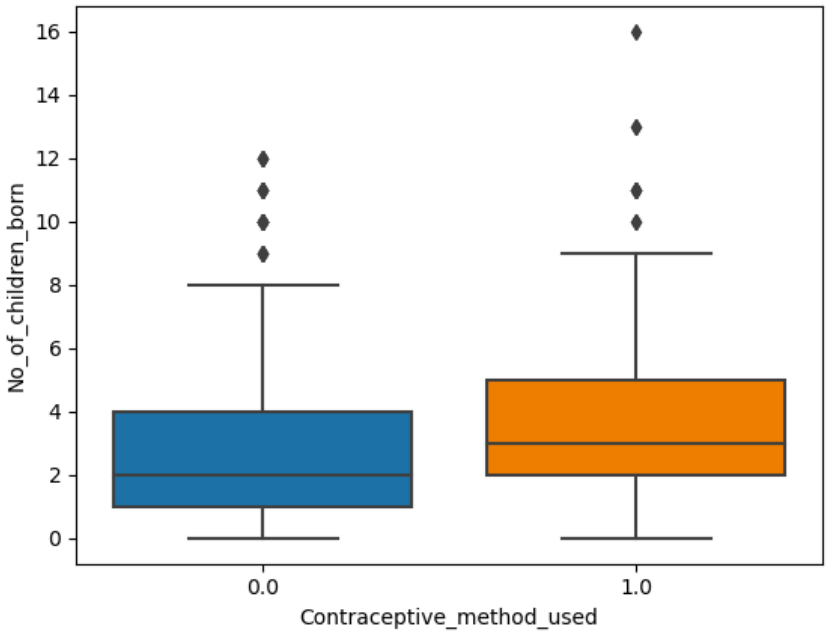
Additionally, it is noteworthy that, on average, about 55% of individuals in the dataset report using contraceptive fall under the age group of 26 to 36 years.



An interesting observation is that an equal number of women from various educational backgrounds do not use contraceptives. However, there is a notable contrast in contraceptive usage among women with different education levels. A substantial number of women with a level 4 education background are using contraceptives, indicating a potential correlation between higher education and contraceptive usage.

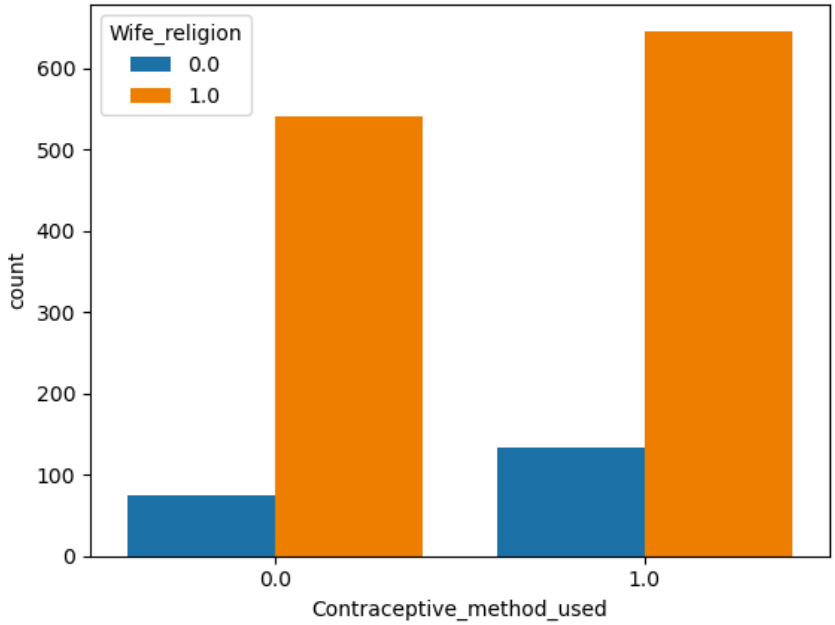


Specifically, as the education level of husbands increases, there is a higher probability that their wives are using contraceptives.

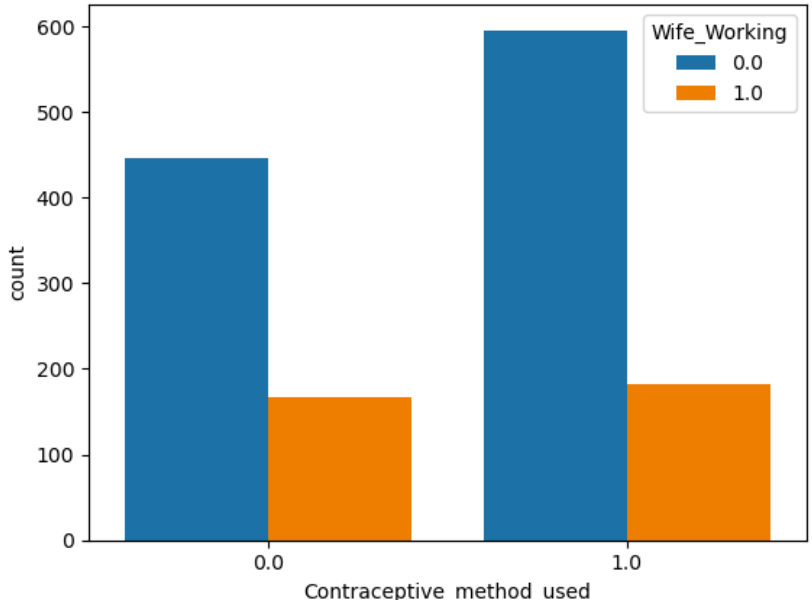


There is a noticeable trend in the dataset regarding contraceptive usage and the average number of children among women:
- Women who are using contraceptives tend to have, on average, 2 to 3 children.
- In contrast, women who are not using contraceptives have an average of around 2 children.

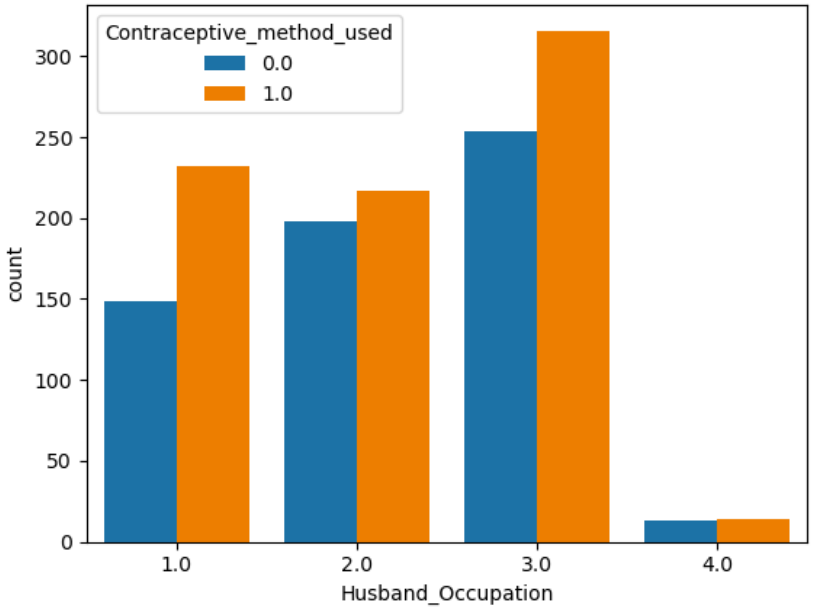
This observation suggests that women with fewer children are more likely to use contraceptives.



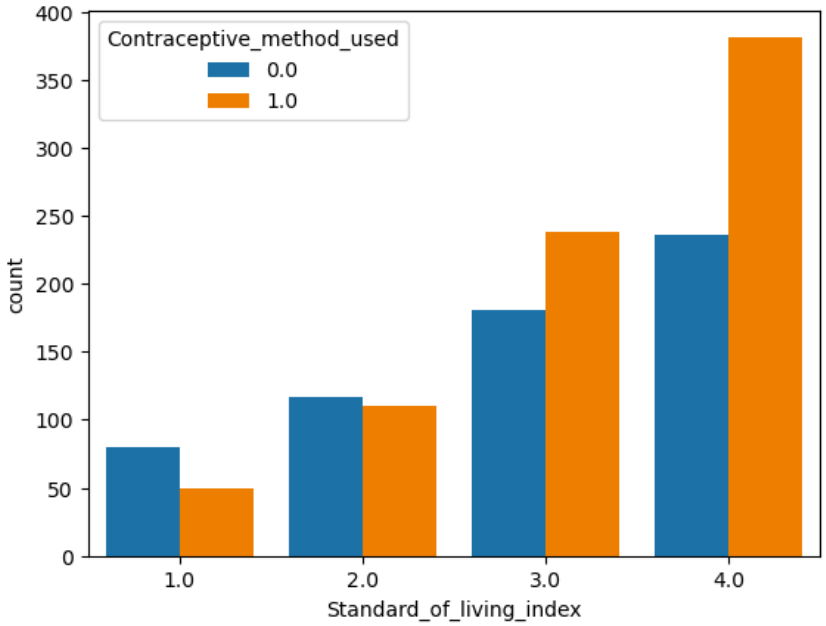
A notable observation is that a higher usage of contraceptives is associated with the religion "Scientology" among the wives in the dataset.



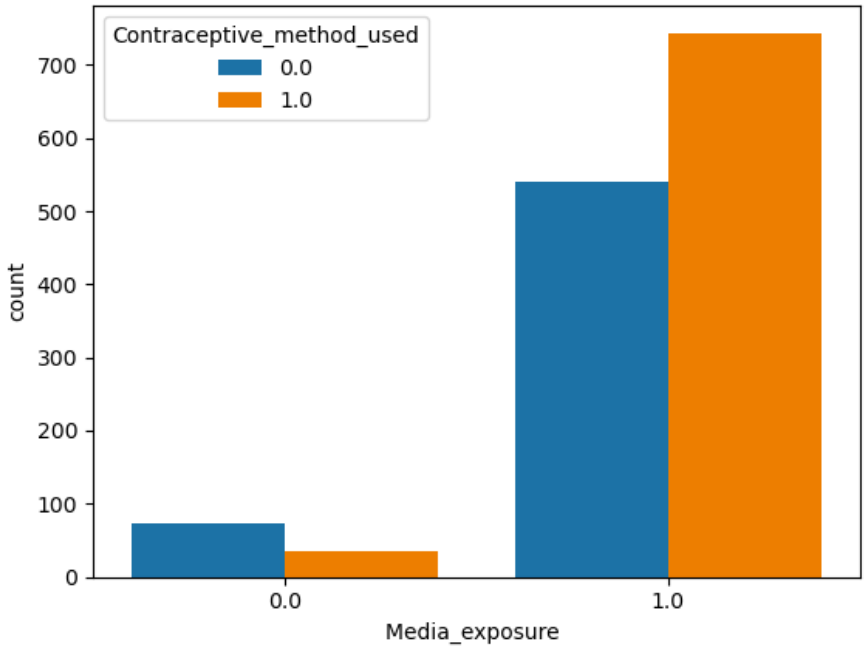
It's interesting to observe that contraceptive usage is not significantly influenced by the employment status of women in the dataset. Contrary to some expectations, both working and non-working women appear to use contraceptives at a similar rate. This suggests that employment status may not be a primary factor influencing contraceptive choices among the women in this study.



There is a pattern where a higher usage of contraceptives is observed among women whose husbands have an occupation code of 3(Secondary). This indicates that women whose husbands fall into this specific occupational category are more likely to use contraceptives compared to women whose husbands have different occupations.

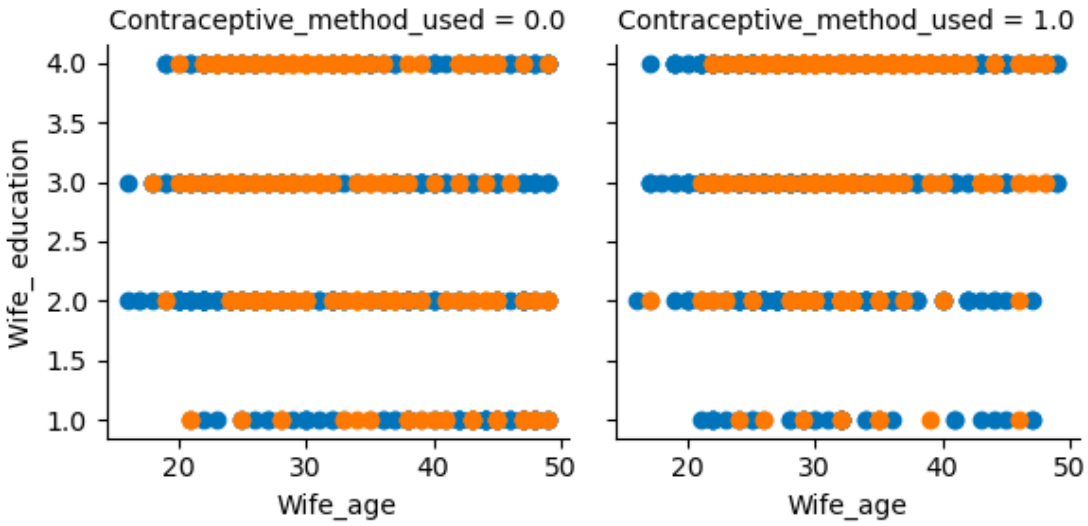


The graph illustrates a positive correlation between the standard of living and contraceptive usage among women in the dataset. As the standard of living index increases, there is a corresponding increase in the usage of contraceptives.



It's evident from the data that women with good media exposure tend to have a higher rate of contraceptive usage.

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
0.0
1.0

2.2.3 Predictive Modelling:

Fig(2.4) shows the glimpse of the dataset, encoded based on the given Data Dictionary inputs.

	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_inde
0	24.0	2	3	3.0	1	0	2	
1	45.0	1	3	10.0	1	0	3	
2	43.0	2	3	7.0	1	0	3	
3	42.0	3	2	9.0	1	0	3	
4	36.0	3	3	8.0	1	0	3	
...	...	...	...	...	...	...	...	
1468	33.0	4	4	NaN	1	1	2	
1469	33.0	4	4	NaN	1	0	1	
1470	39.0	3	3	NaN	1	1	1	
1471	33.0	3	3	NaN	1	1	2	
1472	17.0	3	3	1.0	1	0	2	

1393 rows × 10 columns

Fig(2.4)

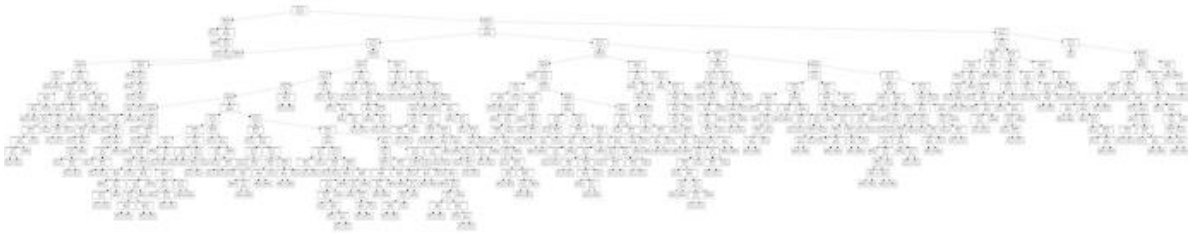
All of the columns are now changed to numerical values.

```
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Wife_age                 1393 non-null   float64
 1   Wife_ education          1393 non-null   float64
 2   Husband_education        1393 non-null   float64
 3   No_of_children_born      1393 non-null   float64
 4   Wife_religion            1393 non-null   float64
 5   Wife_Working             1393 non-null   float64
 6   Husband_Occupation       1393 non-null   float64
 7   Standard_of_living_index 1393 non-null   float64
 8   Media_exposure           1393 non-null   float64
 9   Contraceptive_method_used 1393 non-null  float64
```

For Modelling, the data is split in the ratio 70:30.

CART:

Tree Construction (Using GINI Index):



	Imp
Wife_age	0.284674
Wife_ education	0.109656
Husband_education	0.049222
No_of_children_born	0.250483
Wife_religion	0.037642
Wife_Working	0.049698
Husband_Occupation	0.094572
Standard_of_living_index	0.101042
Media_exposure	0.023012

These values indicate the importance of each feature in making predictions. Here's what these importances suggest:

- Wife_age (0.285): Wife's age has the highest importance, suggesting it strongly influences the model's predictions. It may have a significant impact on whether a woman uses contraception or not.

- Wife_education (0.110): Wife's education level is the second most important feature. This indicates that the education level of the wife plays a notable role in predicting contraceptive usage.

- Husband_education (0.049): Husband's education level is also a relevant feature but has less importance than wife's education.

- No_of_children_born (0.250): The number of children ever born is highly important, indicating that family size is a significant factor in contraceptive use.

- Wife_religion (0.038): Wife's religion has some importance, suggesting that religious affiliation may influence contraceptive choices.

-Wife_Working (0.050): Wife's employment status is a moderate influencer, implying that it has a moderate effect on contraceptive use.

- Husband_Occupation (0.095): Husband's occupation is fairly important, indicating that the type of work the husband is engaged in can impact contraceptive decisions.

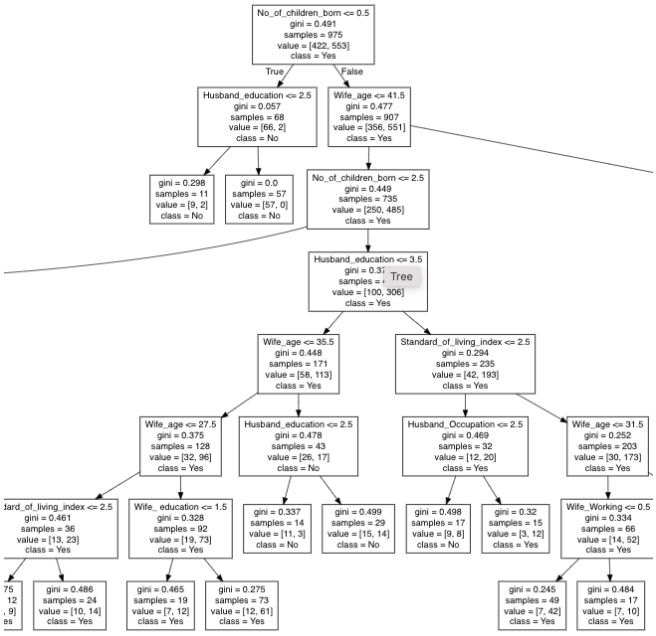
- Standard_of_living_index (0.101: The standard of living index is quite important, suggesting that the economic status of the family is a significant predictor of contraceptive usage.

- Media_exposure (0.023): Media exposure has the lowest importance, but it still contributes to predicting contraceptive use.

These feature importances provide insights into which variables have the most influence on the model's predictions. It can guide further analysis and decision-making when considering the factors that affect contraceptive usage among married women in the dataset.

Since, the tree has overgrown, pruning is required. **Pruning** involves removing branches from the tree that do not significantly improve predictive accuracy. This helps avoid overfitting. Lets regularize the above data.

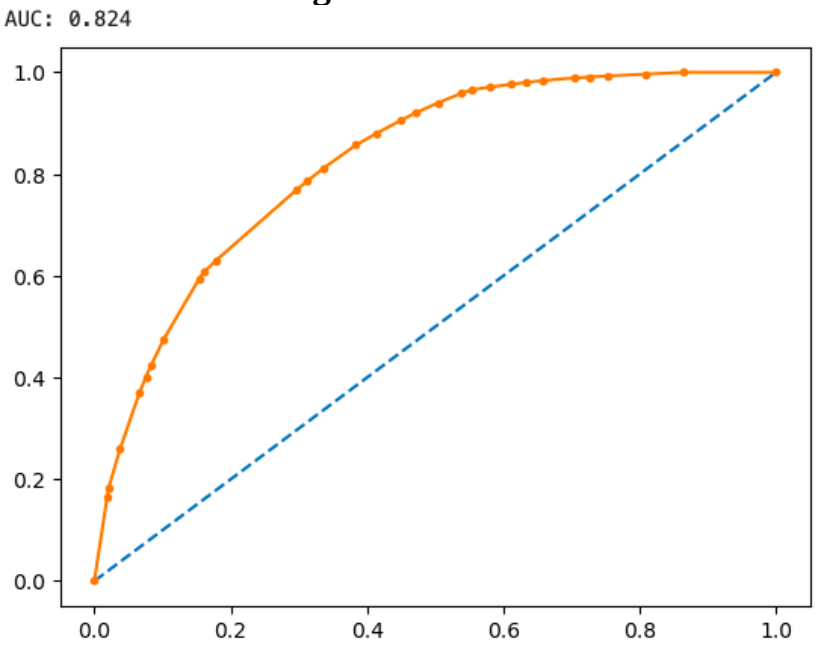
Fig(2.3,2.4) shows the tree after pruning.



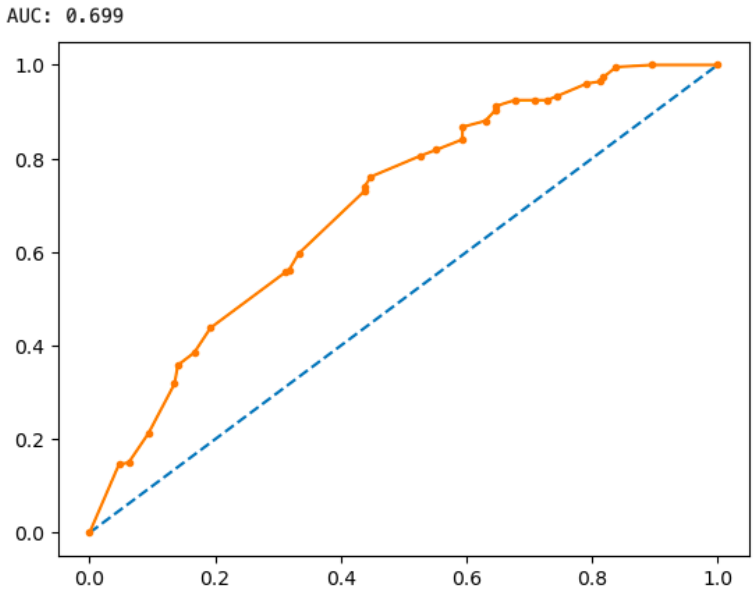
Fig(2.3)

Fig(2.4)

AUC for the training data:- 0.824



AUC for the test data:- 0.69.9



:

AUC: 0.824 for Training Dataset: This AUC value for the training dataset is relatively high (0.824), indicating that the model performs well in discriminating between the two classes (e.g., contraceptive users and non-users) within the data it was trained on. It suggests that the model has learned the relationships between features and the target variable effectively during training.

AUC for Test Data: The AUC value for the test dataset is 0.699. While this value is lower than the training AUC, it still suggests that the model has reasonable predictive power on unseen data. An AUC of 0.699 indicates that the model performs better than random chance but may benefit from further refinement.

Classification report for training data:
Accuracy:0.75

	precision	recall	f1-score	support
0.0	0.77	0.62	0.68	422
1.0	0.75	0.86	0.80	553
accuracy			0.75	975
macro avg	0.76	0.74	0.74	975
weighted avg	0.75	0.75	0.75	975

Classification report for test data:
Accuracy:0.65

	precision	recall	f1-score	support
0.0	0.67	0.47	0.56	192
1.0	0.64	0.81	0.72	226
accuracy			0.65	418
macro avg	0.66	0.64	0.64	418
weighted avg	0.66	0.65	0.64	418

- In the training dataset, the model has a relatively good ability to identify contraceptive users (class 1), with a high F1-score of 0.80. However, it performs slightly less well in identifying non-users (class 0) with an F1-score of 0.68.

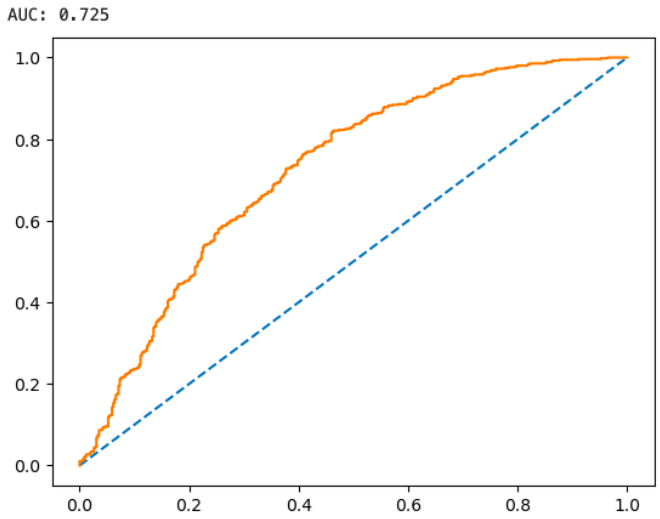
- In the test dataset, the model's performance is similar, but the F1-scores for both classes are slightly lower, indicating some drop in predictive performance on unseen data.

- The model exhibits better recall for contraceptive users (class 1) compared to non-users (class 0), suggesting it is better at capturing true positive cases of contraceptive usage.

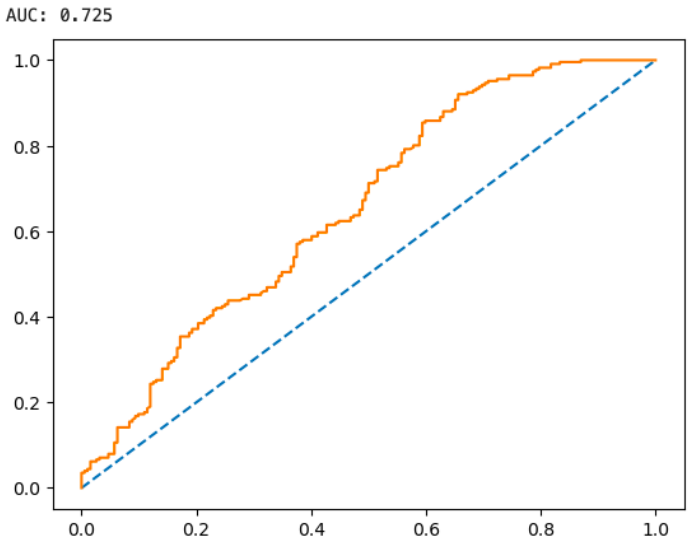
Overall, the model provides reasonable predictive performance.

Logistic Regression:

AUC for the training data:- 0.725



AUC for the test data:- 0.725



The logistic regression model achieves an AUC of 0.725 for both the training and test datasets, indicating that it has reasonable discriminatory power in distinguishing between contraceptive users and non-users.

Classification report for training data:
Accuracy:0.70

	precision	recall	f1-score	support
0.0	0.69	0.54	0.61	422
1.0	0.70	0.81	0.75	553
accuracy			0.70	975
macro avg	0.69	0.68	0.68	975
weighted avg	0.69	0.70	0.69	975

Classification report for training data:
Accuracy:0.62

	precision	recall	f1-score	support
0.0	0.62	0.44	0.52	192
1.0	0.62	0.77	0.68	226
accuracy			0.62	418
macro avg	0.62	0.60	0.60	418
weighted avg	0.62	0.62	0.61	418

- In the training dataset, the logistic regression model achieved an accuracy of 0.70. It has a better ability to identify contraceptive users (class 1), with a higher F1-score of 0.75 compared to non-users (class 0) with an F1-score of 0.61.

- In the test dataset, the model's performance is similar, but the F1-scores for both classes are slightly lower, indicating some drop in predictive performance on unseen data. The overall accuracy on the test dataset is 0.62.

- This logistic regression model provides reasonable predictive performance.

GridSearchCV is a valuable tool for finding the best hyperparameters for a machine learning model. It systematically explores the hyperparameter space to identify the configuration that results in the best model performance.

Best estimator:
LogisticRegression(max_iter=10000, n_jobs=2, penalty='none', solver='sag')

Accuracy after using GridSearchCV for training and testing dataset:

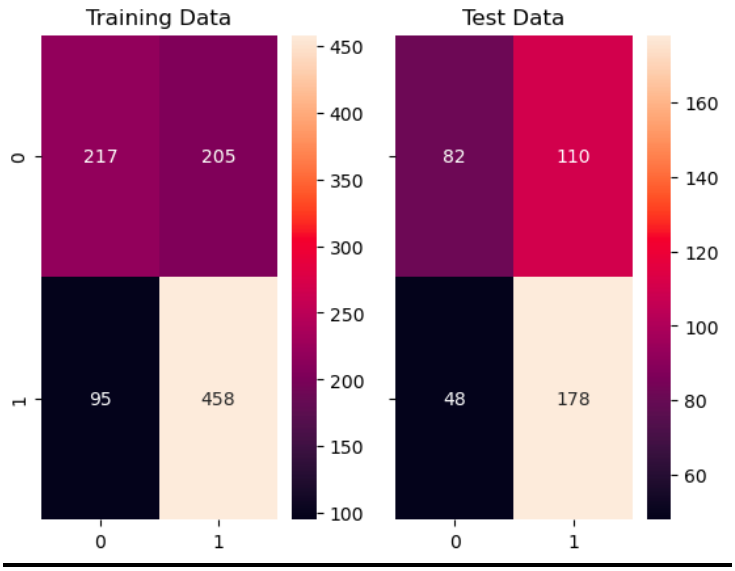
```
best_model.score(X_train, train_labels)
```

```
0.6923076923076923
```

```
best_model.score(X_test, test_labels)
```

```
0.6196172248803827
```

LDA:



Let's interpret these confusion matrices:

Training Data Confusion Matrix:
- True Positives (TP): 458
- True Negatives (TN): 217
- False Positives (FP): 205
- False Negatives (FN): 95

Test Data Confusion Matrix:
- True Positives (TP): 178
- True Negatives (TN): 82
- False Positives (FP): 110
- False Negatives (FN): 48

Now, let's break down the key metrics based on these confusion matrices:

The classification reports below provides a detailed summary of the model's performance on both the training and test data. Let's analyze the key metrics:

Here's the interpretation:

Training Data:
- Precision (Positive Class): This metric indicates that when the model predicts the positive class, it is correct approximately 69% of the time.
- Recall (Positive Class): This metric shows that the model correctly identifies around 83% of the actual positive instances.
- F1-Score (Positive Class): The F1-Score combines precision and recall into a single metric and provides a balanced assessment of the model's performance for the positive class. In this case, it's 0.75.
- Accuracy: The overall accuracy of the model on the training data is 69%.

Test Data:
- Precision (Positive Class): When the model predicts the positive class on the test data, it is correct approximately 62% of the time.
- Recall (Positive Class): The model correctly identifies around 79% of the actual positive instances in the test data.
- F1-Score (Positive Class): The F1-Score for the positive class on the test data is 0.69.
- Accuracy: The overall accuracy of the model on the test data is 62%.

In summary, the model performs reasonably well on both the training and test datasets, with a better F1-Score for the positive class on the training data

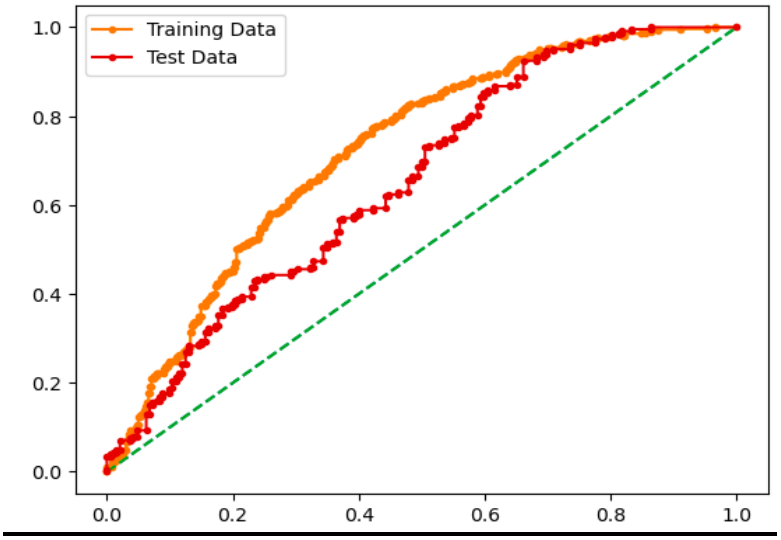
Classification Report of the training data:

	precision	recall	f1-score	support
0.0	0.70	0.51	0.59	422
1.0	0.69	0.83	0.75	553
accuracy			0.69	975
macro avg	0.69	0.67	0.67	975
weighted avg	0.69	0.69	0.68	975

Classification Report of the test data:

	precision	recall	f1-score	support
0.0	0.63	0.43	0.51	192
1.0	0.62	0.79	0.69	226
accuracy			0.62	418
macro avg	0.62	0.61	0.60	418
weighted avg	0.62	0.62	0.61	418

```
AUC for the Training Data: 0.724
AUC for the Test Data: 0.656
```



Training Data AUC (0.724): This value indicates that the LDA model has reasonably good discriminatory power on the training data. It can effectively distinguish between the positive and negative classes, with an AUC of 0.724.

Test Data AUC (0.656): The AUC for the test data, while slightly lower than the training data, still suggests that the model performs reasonably well in discriminating between the classes on unseen data. An AUC of 0.656 is a respectable performance metric.

```
clf.intercept_
```

```
array([-0.90675224])
```

```
clf.coef_
```

```
array([[-0.08383057,  0.48973534,  0.07542596,  0.35443563, -0.42671887,
        -0.02218792,  0.17313293,  0.29931418,  0.1827471 ]])
```

```
X.columns
```

```
Index(['Wife_age', 'Wife_ education', 'Husband_education',
       'No_of_children_born', 'Wife_religion', 'Wife_Working',
       'Husband_Occupation', 'Standard_of_living_index', 'Media_exposure '],
      dtype='object')
```

The predictor 'Wife_education' has the largest magnitude thus this helps in classifying the best.

2.4 Insights and Recommendations:

Based on the analysis and predictions from the classification model, here are some insights and potential recommendations:

Insights:

Average Education Level: The average education level of wives in the dataset is approximately 2.92, indicating that, on average, wives have received a moderate level of education.

Standard of Living: A significant proportion of wives in the dataset have a standard of living index of 4, suggesting a relatively high standard of living for many of them.

Children: Many women in the dataset have 1, 2, or 3 children on average, with some having as few as none and others as many as 16 children.

Husband's Occupation: Husbands in the dataset have diverse occupations, ranging from 1 to 4.

Media Exposure: A majority of wives in the dataset have good media exposure.

Religion: Most wives in the dataset follow Scientology religions.

Working Women: It's observed that both working and non-working women use contraceptives, with similar usage rates.

Husband's Education: There is a positive correlation between higher husband education levels and contraceptive usage.

Recommendations:

1.Education and Awareness: Promote education and awareness programs, especially among women with lower education levels, to help them make informed decisions about family planning and contraceptive use.

2. Family Planning Services: Ensure that family planning services and contraceptives are accessible and affordable to women with varying socio-economic backgrounds.

3. Husband's Involvement: Encourage husbands to be involved in family planning decisions and to support their wives in making choices about contraceptive use.

4. Media Campaigns:Utilize media exposure to disseminate information about family planning and contraceptive methods to reach a wider audience.

5. Religion and Culture: Respect and consider the cultural and religious beliefs of the population when implementing family planning programs.

6. Occupation-Based Outreach: Tailor outreach and education programs to the diverse occupations of husbands to effectively convey family planning information.

7. Standard of Living: Recognize that women with a higher standard of living may have different family planning needs and preferences.

8. Children: Provide support and resources for women with varying numbers of children to ensure they have access to suitable family planning options.

9. Continuous Monitoring: Continuously monitor and evaluate the effectiveness of family planning programs to make data-driven adjustments and improvements.

These insights and recommendations can help inform policies and programs aimed at promoting family planning and contraceptive use among married women, taking into account their demographic and socio-economic characteristics.