PREDECTIVE MODELING

BUSINESS REPORT

BY

G S JAIGURURAM

TABLE OF CONTENTS

Problem 1:

The comp-activ database comprises activity measures of computer systems. Data was gathered from a Sun Sparcstation 20/712 with 128 Mbytes of memory, operating in a multi-user university department. Users engaged in diverse tasks, such as internet access, file editing, and CPU-intensive programs.

Being an aspiring data scientist, you aim to establish a linear equation for predicting 'usr' (the percentage of time CPUs operate in user mode). Your goal is to analyze various system attributes to understand their influence on the system's 'usr' mode

.

PROBLEM 1.1 DEFINE THE PROBLEM AND PERFORM EXPLORATORY DATA ANALYSIS:

Problem 1.1.1 Problem definition:

The comp-activ database contains activity measures of computer systems. The data was collected from a Sun Sparcstation 20/712 with 128 Mbytes of memory, operating in a multi-user university department. The goal is to establish a linear equation for predicting 'usr' (the percentage of time CPUs operate in user mode) by analyzing the influence of various system attributes.

Problem 1.1.2 Check shape, Data types, statistical summary:

First, we import all the necessary libraries seaborn, numpy, pandas, sklearn, matplotlib etc. to perform our analysis

Next, we import the data set “compactiv.xlsx”

Data Dictionary:

Column Name	Column Description
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

Shape of the data:

Shape of the dataset is 8192 rows and 22 Columns.

Data Type:

```
Timestamp                    object
lread                        int64
lwrite                       int64
scall                        int64
sread                        int64
swrite                       int64
fork                         float64
exec                         float64
rchar                        float64
wchar                        float64
pgout                        float64
ppgout                       float64
pgfree                       float64
pgscan                       float64
atch                         float64
pgin                         float64
ppgin                        float64
pflt                         float64
vflt                         float64
runqsz                       object
freemem                      int64
freeswap                     int64
usr                          int64
```

There are total of 21 Numerical and 1 Categorical data type are available.

Statistical Summary:

Insights:

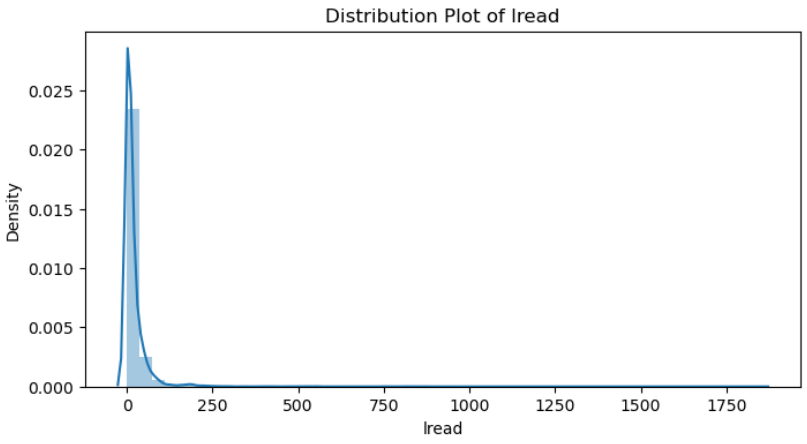
1. The dataset has 8192 rows and 22 columns.
2. The dataset has only one categorical variable and 21 numerical variables
3. The dataset has 13 float type variable, 8 int type variable, 1 object type variable

4. System Memory Reads and Writes: lread (Reads): Mean of 19.56 with a large standard deviation (53.35) indicates high variability in memory reads, ranging from 0 to 1845. lwrite (Writes): Mean of 13.11 and standard deviation of 29.89, also showing significant variability, ranging from 0 to 575.
5. System Calls and Read/Write Calls: scall (System Calls): High mean of 2306.32 and a wide range from 109 to 12493, indicating a high frequency of system calls. sread (System Read Calls): Mean of 210.48 and standard deviation of 198.98, suggesting moderate variability with some extreme values up to 5318. swrite (System Write Calls): Mean of 150.06, indicating fewer write calls compared to read calls, with values up to 5456.
6. Process Management: fork: Mean of 1.88 with a range from 0 to 20.12. exec: Mean of 2.79 and a wide range up to 59.56, indicating variability in process execution calls.
7. Character Transfers: rchar (Read Characters): High mean of 197385.7, indicating substantial data transfer via read calls, with extreme values up to 2526649. wchar (Write Characters): Mean of 95902.99, suggesting considerable data transfer via write calls, with values up to 1801623.
8. Page Management: pgout and ppgout: Low means (2.29 and 5.98, respectively) with some extreme values, indicating occasional high page-out activities. pgfree and pgscan: Low means (11.92 and 21.53, respectively) but high standard deviations, indicating sporadic high values. pgin and ppgin: Mean values (8.28 and 12.39, respectively) with high standard deviations, suggesting occasional high page-in activities.
9. Page Faults: pflt and vflt: Mean values (109.79 and 185.32, respectively) with significant variability, indicating a frequent occurrence of page faults.
10. Memory and Swap Space: freemem: Mean of 1763.46, indicating available memory, with values ranging from 55 to 12027. freeswap: Mean of 1328126, indicating a substantial amount of swap space, with values ranging from 2 to 2243187.
11. CPU User Mode (usr): Mean of 83.97 with a standard deviation of 18.40, suggesting that CPUs spend a significant amount of time in user mode, with values ranging from 0 to 99.

	count	mean	std	min	0.25	0.5	0.75	max
lread	8192	19.55969	53.3538	0	2	7	20	1845
lwrite	8192	13.1062	29.89173	0	0	1	10	575
scall	8192	2306.318	1633.617	109	1012	2051.5	3317.25	12493
sread	8192	210.48	198.9801	6	86	166	279	5318
swrite	8192	150.0582	160.479	7	63	117	185	5456
fork	8192	1.884554	2.479493	0	0.4	0.8	2.2	20.12
exec	8192	2.791998	5.212456	0	0.2	1.2	2.8	59.56
rchar	8088	197385.7	239837.5	278	34091.5	125473.5	267828.8	2526649
wchar	8177	95902.99	140841.7	1498	22916	46619	106101	1801623
pgout	8192	2.285317	5.307038	0	0	0	2.4	81.44
ppgout	8192	5.977229	15.21459	0	0	0	4.2	184.2
pgfree	8192	11.91971	32.36352	0	0	0	5	523
pgscan	8192	21.52685	71.14134	0	0	0	0	1237
atch	8192	1.127505	5.708347	0	0	0	0.6	211.58
pgin	8192	8.27796	13.87498	0	0.6	2.8	9.765	141.2
ppgin	8192	12.38859	22.28132	0	0.6	3.8	13.8	292.61
pflt	8192	109.7938	114.4192	0	25	63.8	159.6	899.8
vflt	8192	185.3158	191.0006	0.2	45.4	120.4	251.8	1365
freemem	8192	1763.456	2482.105	55	231	579	2002.25	12027
freeswap	8192	1328126	422019.4	2	1042624	1289290	1730380	2243187
usr	8192	83.96887	18.40191	0	81	89	94	99

Problem 1.1.3 Univariate analysis:

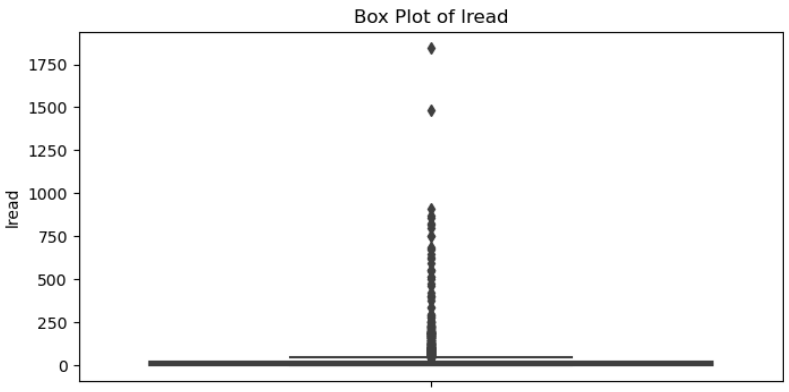
Numerical Data type:



Right-Skewed Distribution: The distribution of "lread" is highly right-skewed. Most of the data points are concentrated at the lower end of the scale, close to zero, while there are fewer data points as the value of "lread" increases.

High Density Near Zero: The peak density is very close to zero, suggesting that the majority of the data points have lower values of "lread." This indicates that in whatever context "lread" is used, lower values are much more common.

Long Tail: The plot has a long tail extending towards the right, indicating the presence of some high values of "lread."



However, these high values are much less frequent compared to the lower values.

Potential Outliers: The presence of the long tail suggests there may be outliers or extreme values far from the mean. These outliers can have significant effects on the mean but less on the median.

Outlier Handling: Depending on the context, it might be important to investigate the high values (outliers) to understand their cause and impact on the analysis.

Central Tendency:

Median: The line inside the box represents the median of the "Iread" values. The median is located close to the lower end of the scale, which aligns with the earlier observation from the density plot that most values are clustered near zero.

Interquartile Range (IQR):

The box represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). Most of the data points (50%) lie within this range.

The IQR is quite narrow, indicating that the middle 50% of the "Iread" values are close to each other and clustered towards the lower end of the scale.

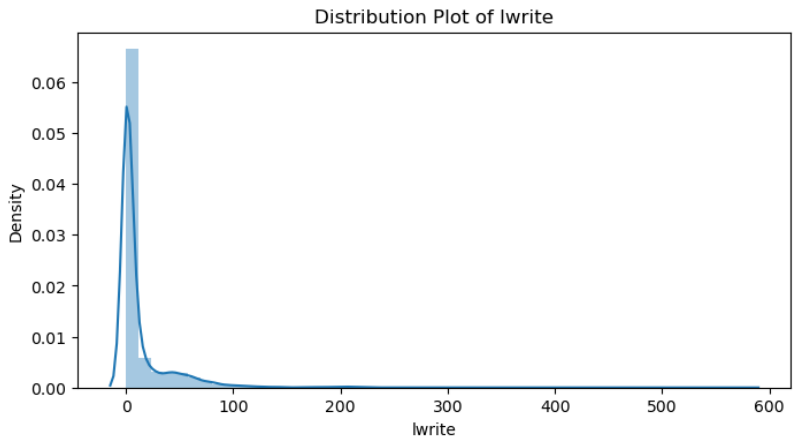
Whiskers and Outliers:

The "whiskers" extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles. The whiskers are very short, suggesting that the majority of the data points are not spread out widely.

There are numerous outliers, represented by points beyond the whiskers. The presence of outliers is significant, particularly the two extreme outliers above 1500. This confirms the observation from the density plot that there are a few very high values of "Iread."

Skewness:

The box plot confirms the right-skewed nature of the distribution. The majority of the data is concentrated near the lower values, with a long tail extending to the right due to the outliers.



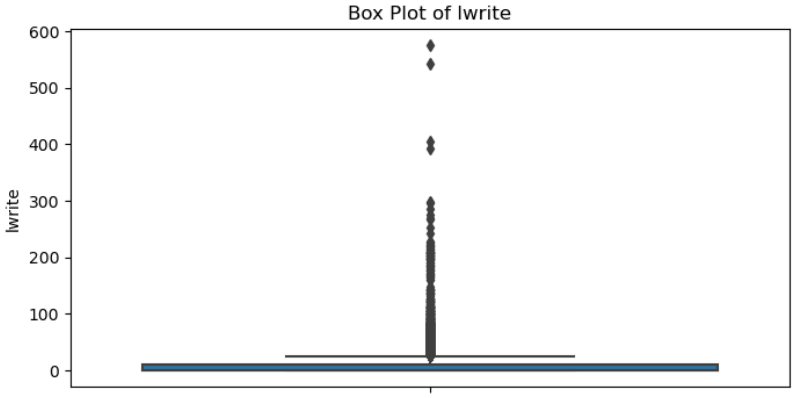


FIG 1

Skewness and Distribution Shape:

The distribution of lwrite (writes) is highly skewed to the right (positive skewness).

Most of the lwrite values are concentrated at the lower end, close to zero.

Density Peaks:

There is a pronounced peak near zero, indicating that a significant number of observations have very low or zero write operations.

. The density decreases sharply after the peak, showing that higher values of lwrite are much less frequent.

. **Long Tail:**

. The plot has a long tail extending towards higher values, up to around 600.

. This indicates the presence of a few observations with very high lwrite values, but they are rare.

. **Implications for Analysis:**

. The high skewness suggests that lwrite may not follow a normal distribution, which could affect certain statistical analyses that assume normality.

. Transformations (e.g., log transformation) might be necessary if normality is a requirement for the analysis.

. The presence of extreme values (outliers) should be considered in further analysis, as they can significantly impact model performance and interpretation.

.

. **Median and Interquartile Range (IQR):**

. The median lwrite value is very close to zero, indicating that at least 50% of the data points have very low write operations.

. The interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3), is also very narrow, suggesting that the majority of lwrite values are clustered close to the lower end.

. **Outliers:**

. There are numerous outliers above the upper whisker. These outliers extend up to approximately 600.

. The presence of these outliers indicates that while most lwrite values are low, there are occasional instances of very high write operations.

. **Whiskers:**

. The whiskers extend to relatively low values compared to the outliers,

reinforcing that the bulk of the data points are at the lower end of the scale.

. **Skewness:**

. The plot confirms the right skewness observed in the density plot. The concentration of values near the lower end with a long tail of high values signifies a skewed distribution.

. **Implications for Data Analysis:**

. The outliers need to be addressed depending on the analysis context. For instance, if these high values are legitimate, they should be retained; otherwise, they might need to be transformed or removed.

. The skewness and presence of outliers suggest that statistical methods that assume normality might not be appropriate without transformation (e.g., log transformation).

. In predictive modeling, the outliers could disproportionately influence the model, and techniques such as robust regression or outlier removal might be necessary.

.

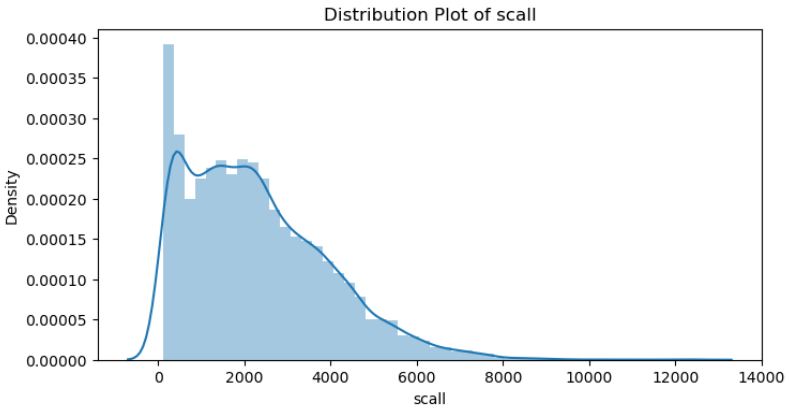


FIG 5

. **Distribution Shape:**

. The distribution of scall (number of system calls per second) shows a right-skewed pattern.

. A large number of system calls occur at the lower end of the scale, with a long tail extending towards higher values.

Central Tendency:

. The peak of the distribution is around a lower scall value, indicating that most system calls per second are clustered around this lower range.

. There is a noticeable drop-off in density as scall values increase.

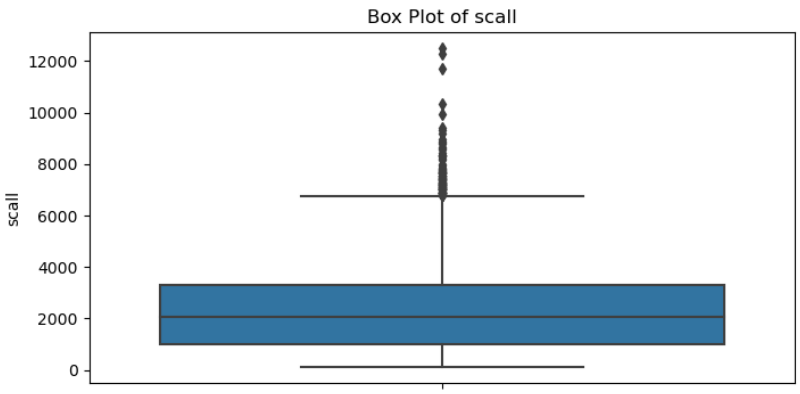
Spread and Variability:

. The distribution has a wide range, with scall values extending from near zero to over 12,000.

. This suggests high variability in the number of system calls per second.

Skewness:

. The right skewness is evident, indicating that while the majority of system call values are low, there are significant instances where system calls per second are quite high.



Central Tendency:

Median: The line inside the box represents the median of the "scall" values. The median is around 2500, indicating that half of the data points are below this value and half are above.

Interquartile Range (IQR):

The box represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). This range spans from approximately 1000 to 4000.

This indicates that the middle 50% of the "scall" values lie within this range.

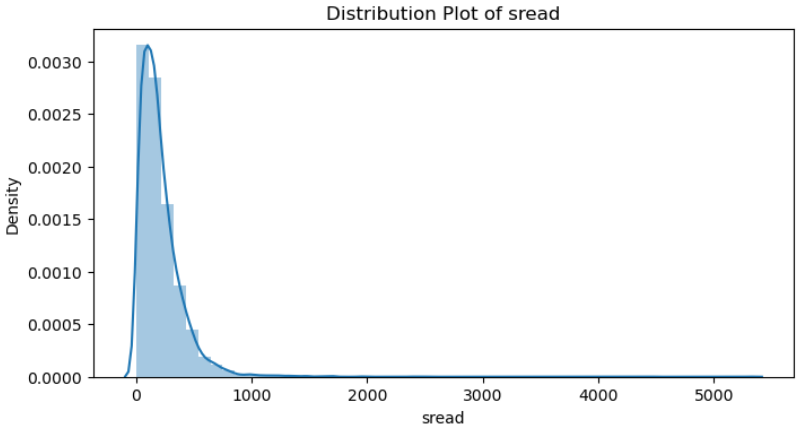
Whiskers and Outliers:

The "whiskers" extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles. The lower whisker extends to 0, while the upper whisker extends to approximately 7500, suggesting a wider range for the upper values.

There are numerous outliers beyond the upper whisker, with values extending up to around 12000. These outliers indicate the presence of some exceptionally high values in the dataset.

Skewness:

The distribution of "scall" appears to be right-skewed due to the presence of higher values and the outliers extending significantly beyond the upper whisker.



Distribution Shape:

The distribution of sread (number of system read calls per second) is right-skewed.

The majority of the values are concentrated at the lower end, with a sharp decline in frequency as sread values increase.

Central Tendency:

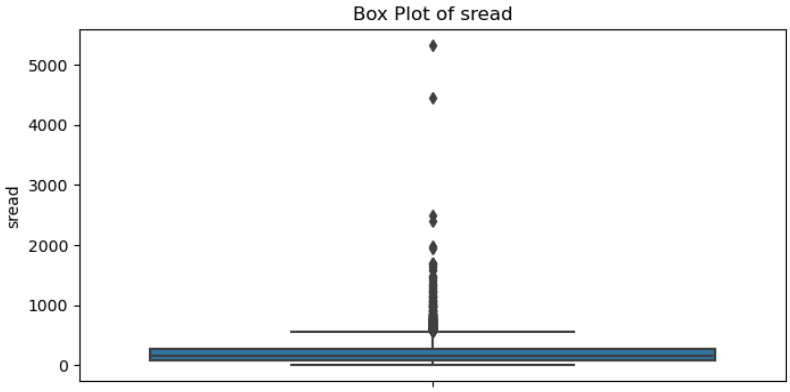
The peak of the distribution is at the lower values of sread, indicating that most system read calls per second are relatively low.

. The mode (most frequent value) appears at the very beginning of the distribution, close to zero.

Spread and Variability:

. The range of sread extends from near zero to over 5000, indicating a wide variability.

. Although the majority of sread values are low, there are some significantly higher



values, suggesting occasional bursts of high system read activity.

. **Skewness:**

. The right skewness indicates that while most sread values are low, there are fewer but significant instances of high sread values.

Central Tendency:

Median: The line inside the box represents the median of the "sread" values. The median is quite close to zero, indicating that half of the data points are below this value and half are above.

Interquartile Range (IQR):

The box represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). This range is also very close to zero, suggesting that the middle 50% of the "sread" values are clustered near zero.

Whiskers and Outliers:

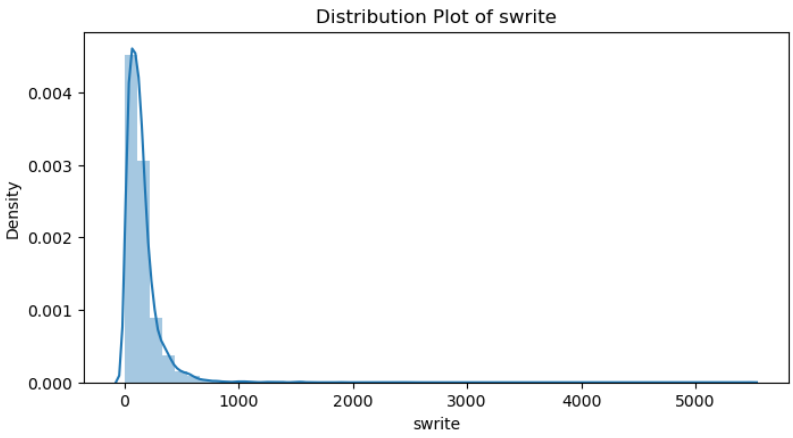
The "whiskers" extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles. The lower whisker extends to zero, while the upper whisker extends to a value significantly higher than the IQR, indicating some spread in the higher values.

There are numerous outliers beyond the upper whisker, with values extending up to around 5500. These outliers indicate the presence of some exceptionally high values in the dataset.

Skewness:

The distribution of "sread" is highly right-skewed due to the presence of higher values and the outliers extending significantly beyond the upper whisker.

.



Distribution Shape:

The distribution of swrite (number of system write calls per second) is right-skewed.

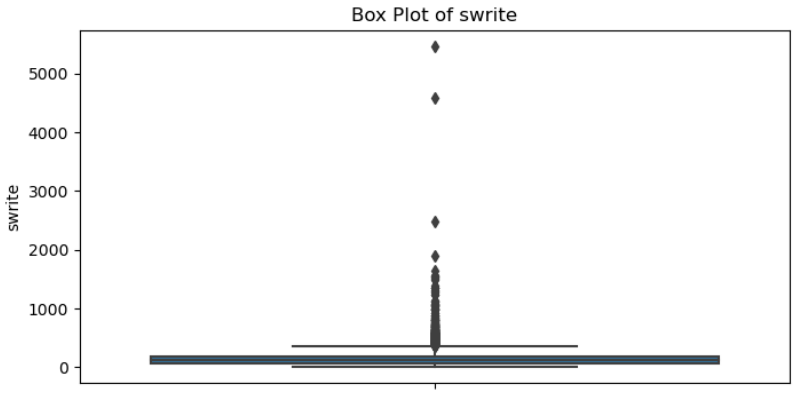
Most values are concentrated on the lower end, with a sharp decline as swrite values increase.

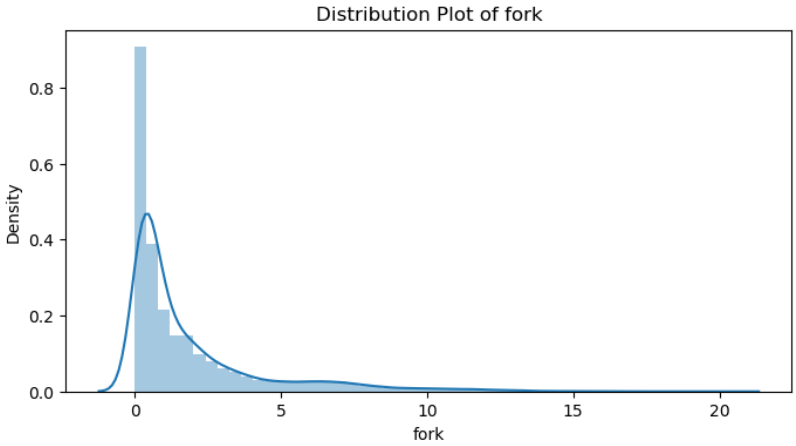
Central Tendency:

The peak of the distribution is at the lower values of swrite, indicating that most system write calls per second are relatively low.

. The mode (most frequent value) appears very close to zero.

. **Spread and Variability:**





. The range of swrite extends from near zero to over 5000, indicating a wide variability.

. While the majority of swrite values are low, there are instances of significantly higher values, suggesting occasional bursts of high system write activity.

. Skewness:

. The right skewness indicates that while most swrite values are low, there are fewer but significant instances of high swrite values.

.

Median (Q2):

. The median swrite value is close to zero, indicating that at least half of the data points have a low number of system write calls per second.

. Interquartile Range (IQR):

. The box itself is very narrow and close to zero, suggesting that the middle 50% of the swrite values are very low and close to each other.

. Whiskers:

. The whiskers extend further than the box but are still relatively short compared to the overall range of swrite values, indicating that most of the data points are close to the lower end of the scale.

. Outliers:

. There are several outliers far beyond the whiskers, with some values reaching over 5000 swrite. This shows that while the majority of the swrite values are low, there are significant instances where the system write calls per second spike to very high levels.

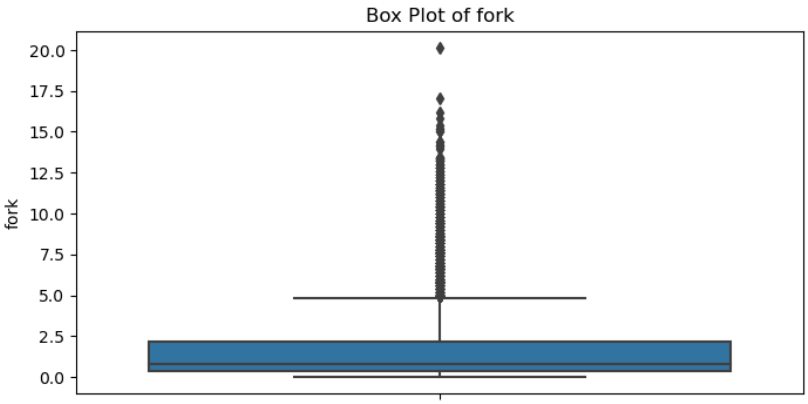
.

Shape of the Distribution:

The histogram is heavily right-skewed (positively skewed). Most of the data points are concentrated on the left side, close to zero.

Center:

The peak of the histogram is at or near zero, indicating that the majority of fork values are very low, with a high density at zero.



Spread:

The histogram shows that while most fork values are close to zero, there are some instances where fork values range up to around 20. However, the frequency of these higher values is very low.

Outliers:

There are no distinct bars far separated from the rest, indicating that there are no significant outliers in the data, but the long tail suggests occasional higher values.

Median (Q2):

. The median fork value is close to zero, indicating that at least half of the data points have a low Number of system fork calls per second.

. Interquartile Range (IQR):

. The box itself is very narrow and close to zero, suggesting that the middle 50% of the fork values are very low and close to each other.

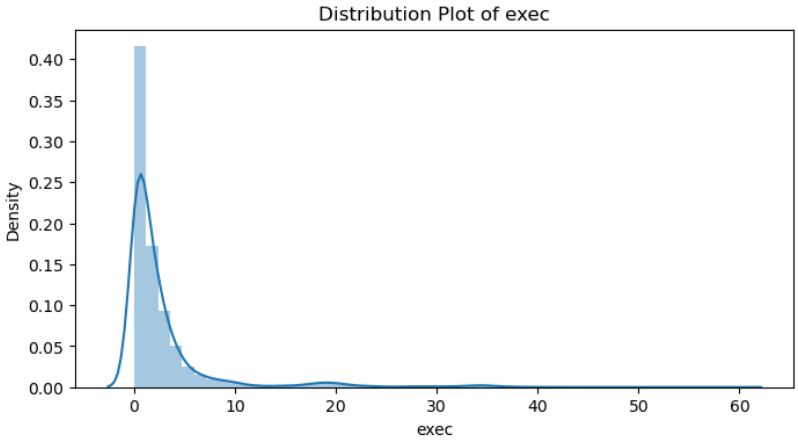
. Whiskers:

. The whiskers extend further than the box but are still relatively short compared to the overall range of fork values, indicating that most of the data points are close to the lower end of the scale.

. Outliers:

. There are several outliers far beyond the whiskers, with some values reaching over 20 fork. This shows that while the majority of the fork values are low, there are significant instances where the system fork calls per second spike to very high levels.

.



Shape of the Distribution:

The histogram is heavily right-skewed (positively skewed). Most of the data points are concentrated on the left side, close to zero.

Center:

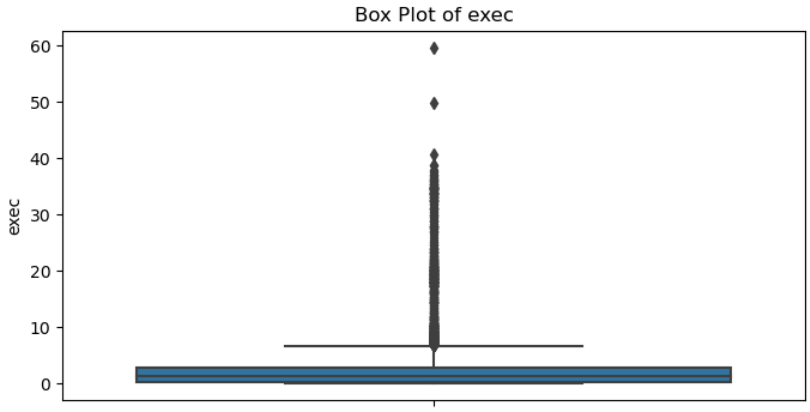
The peak of the histogram is at or near zero, indicating that the majority of exec values are very low, with a high density at zero.

Spread:

The histogram shows that while most exec values are close to zero, there are some instances where exec values range up to around 20. However, the frequency of these higher values is very low.

Outliers:

There are no distinct bars far separated from the rest, indicating that there are no significant outliers in the data, but the long tail suggests occasional higher values.



Median (Q2):

. The median exec value is close to zero, indicating that at least half of the data points have a low Number of system exec calls per second.

. Interquartile Range (IQR):

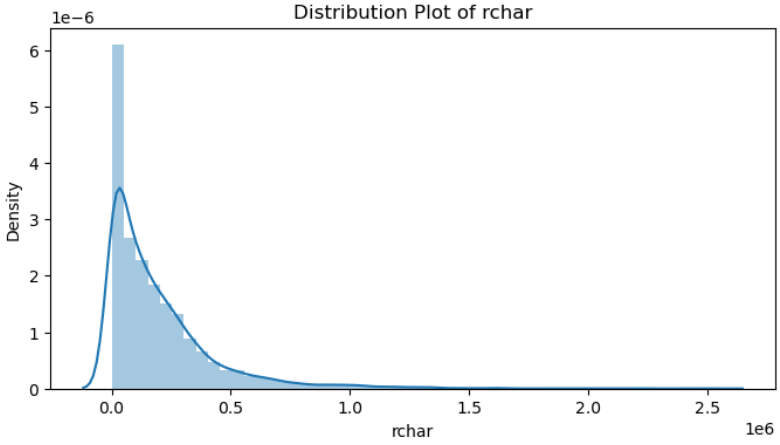
. The box itself is very narrow and close to zero, suggesting that the middle 50% of the exec values are very low and close to each other.

. Whiskers:

. The whiskers extend further than the box but are still relatively short compared to the overall range of exec values, indicating that most of the data points are close to the lower end of the scale.

. Outliers:

. There are several outliers far beyond the whiskers, with some values reaching over 20 exec. This shows that while the majority of the exec values are low, there are significant instances where the system exec calls per second spike to very high levels.

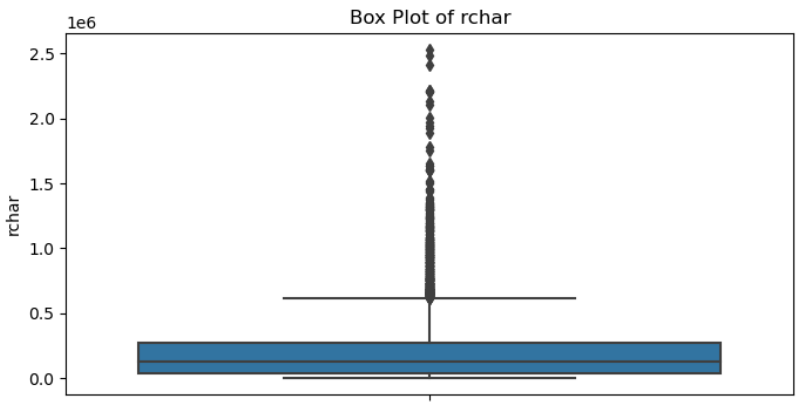


Shape of the Distribution:

The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.

Center:

The peak of the histogram is near zero, indicating that the majority of rchar values are very low.



Spread:

The histogram shows that while most rchar values are close to zero, there are some instances where rchar values extend up to approximately 2.5 million. However, the frequency of these higher values is very low.

Outliers:

There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values.

Median (Q2):

. The median rchar value is close to zero, indicating that at least half of the data points have a low Number of characters transferred per second by system read calls.

. Interquartile Range (IQR):

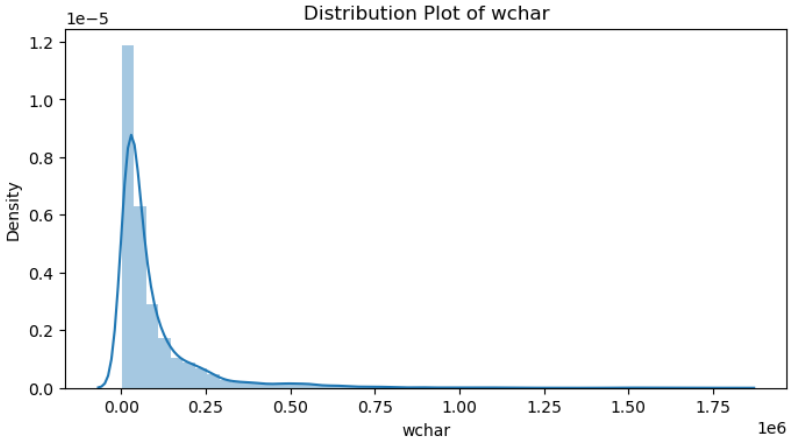
. The box itself is very narrow and close to zero, suggesting that the middle 50% of the rchar values are very low and close to each other.

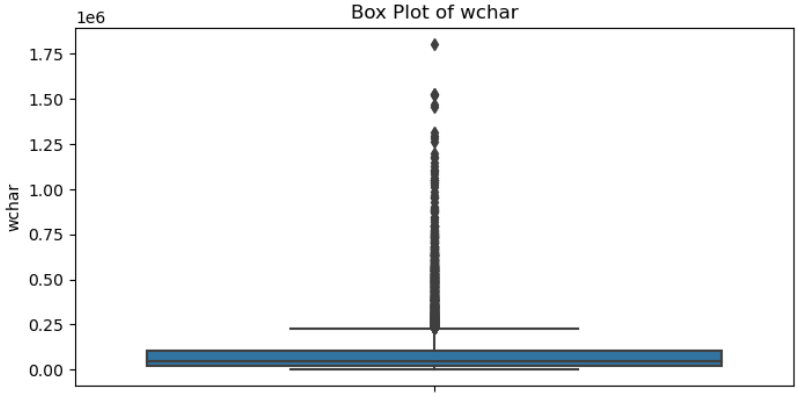
. Whiskers:

. The whiskers extend further than the box but are still relatively short compared to the overall range of rchar values, indicating that most of the data points are close to the lower end of the scale.

. Outliers:

. There are several outliers far beyond the whiskers, with some values reaching over 20 exec. This shows that while the majority of the rchar values are low, there are significant instances where the characters transferred per second by system read calls spike to very high levels.





Shape of the Distribution:

The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.

Center:

The peak of the histogram is near zero, indicating that the majority of wchar values are very low.

Spread:

The histogram shows that while most wchar values are close to zero, there are some instances where wchar values extend up to approximately 0.5 million. However, the frequency of these higher values is very low.

Outliers:

There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values

Median (Q2):

. The median wchar value is close to zero, indicating that at least half of the data points have a low Number of characters transfreed per second by system write calls

. Interquartile Range (IQR):

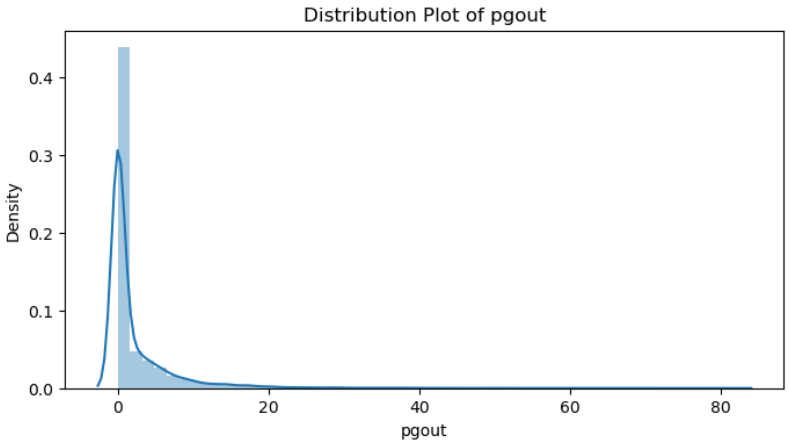
. The box itself is very narrow and close to zero, suggesting that the middle 50% of the wchar values are very low and close to each other.

. Whiskers:

. The whiskers extend further than the box but are still relatively short compared to the overall range of wchar values, indicating that most of the data points are close to the lower end of the scale.

. Outliers:

. There are several outliers far beyond the whiskers, with some values reaching over 20 exec. This shows that while the majority of the wchar values are low, there are significant instances where the characters transferred per second by system write calls spike to very high levels.



Shape of the Distribution:

The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.

Center:

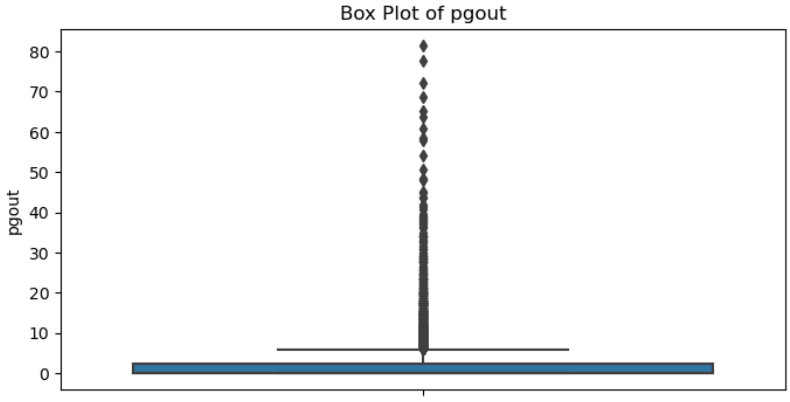
The peak of the histogram is near zero, indicating that the majority of pgout values are very low.

Spread:

The histogram shows that while most pgout values are close to zero. However, the frequency of these higher values is very low.

Outliers:

There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values



Median (Q2):

. The median pgout value is close to zero, indicating that at least half of the data points have a low Number of page out requests per second.

. Interquartile Range (IQR):

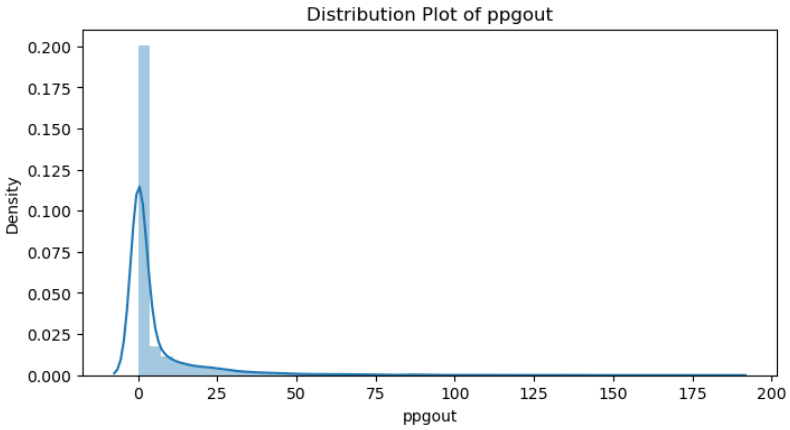
. The box itself is very narrow and close to zero, suggesting that the middle 50% of the pgout values are very low and close to each other.

. Whiskers:

. The whiskers extend further than the box but are still relatively short compared to the overall range of pgout values, indicating that most of the data points are close to the lower end of the scale.

. Outliers:

. There are several outliers far beyond the whiskers, with some values reaching over 20 exec. This shows that while the majority of the pgout values are low, there are significant instances where page out requests per second spike to very high levels.



Shape of the Distribution:

The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.

Center:

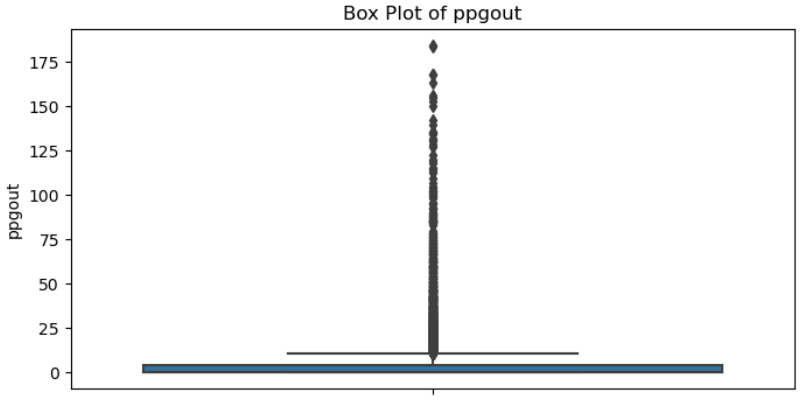
The peak of the histogram is near zero, indicating that the majority of ppgout values are very low.

Spread:

The histogram shows that while most ppgout values are close to zero. However, the frequency of these higher values is very low.

Outliers:

There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values



Median (Q2):

. The median ppgout value is close to zero, indicating that at least half of the data points have a low Number of page, paged out requests per second.

. Interquartile Range (IQR):

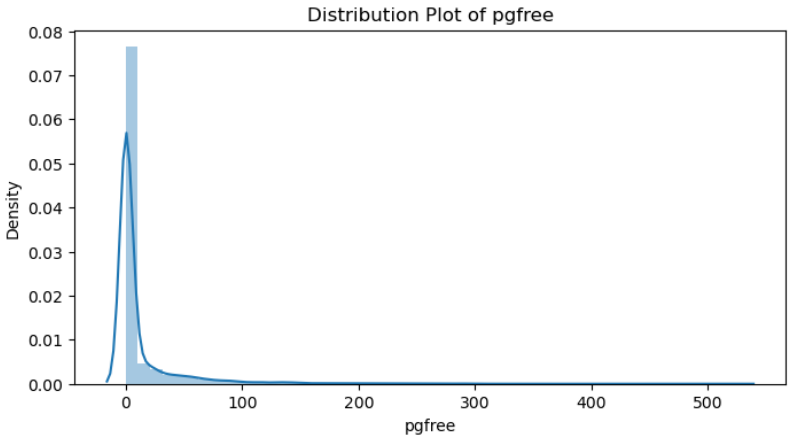
. The box itself is very narrow and close to zero, suggesting that the middle 50% of the ppgout values are very low and close to each other.

. Whiskers:

. The whiskers extend further than the box but are still relatively short compared to the overall range of ppgout values, indicating that most of the data points are close to the lower end of the scale.

. Outliers:

. There are several outliers far beyond the whiskers, with some values reaching over 20 exec. This shows that while the majority of the ppgout values are low, there are significant instances where page, paged out requests per second spike to very high levels.



Shape of the Distribution:

The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.

Center:

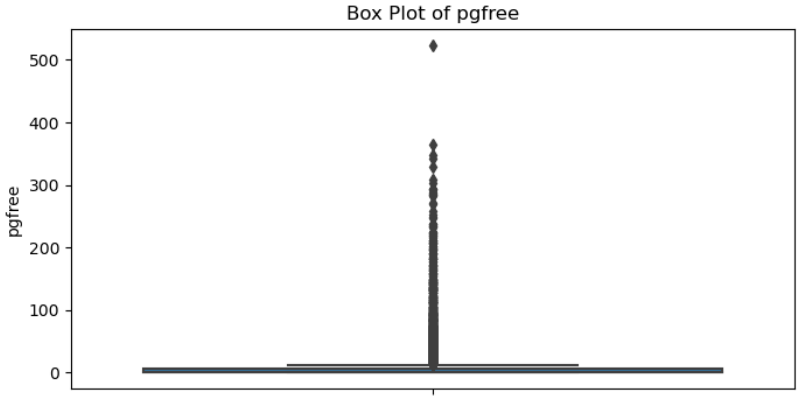
The peak of the histogram is near zero, indicating that the majority of pgfree values are very low.

Spread:

The histogram shows that while most pgfree values are close to zero. However, the frequency of these higher values is very low.

Outliers:

There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values



Median (Q2):

. The median pgfree value is close to zero, indicating that at least half of the data points have a low Number of pages per second placed on the free list..

. Interquartile Range (IQR):

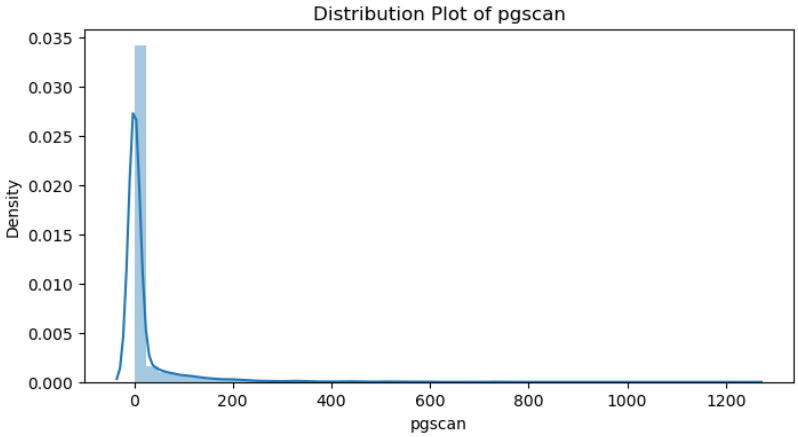
. The box itself is very narrow and close to zero, suggesting that the middle 50% of the pgfree values are very low and close to each other.

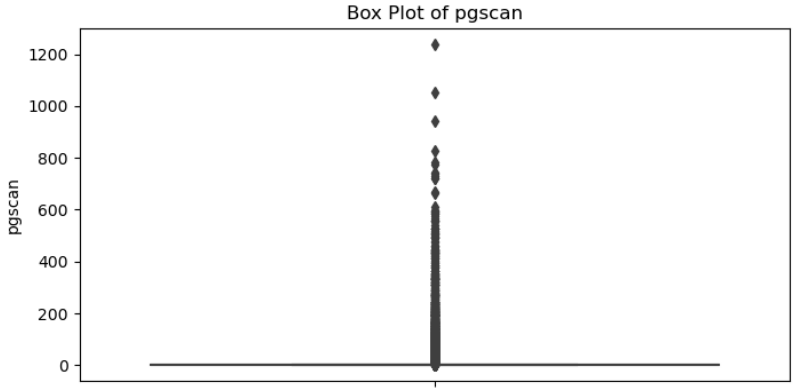
. Whiskers:

. The whiskers extend further than the box but are still relatively short compared to the overall range of pgfree values, indicating that most of the data points are close to the lower end of the scale.

. Outliers:

. There are several outliers far beyond the whiskers, with some values reaching over 20 exec. This shows that while the majority of the pgfree values are low, there are significant instances where pages per second placed on the free list spike to very high levels.





Shape of the Distribution:
The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.

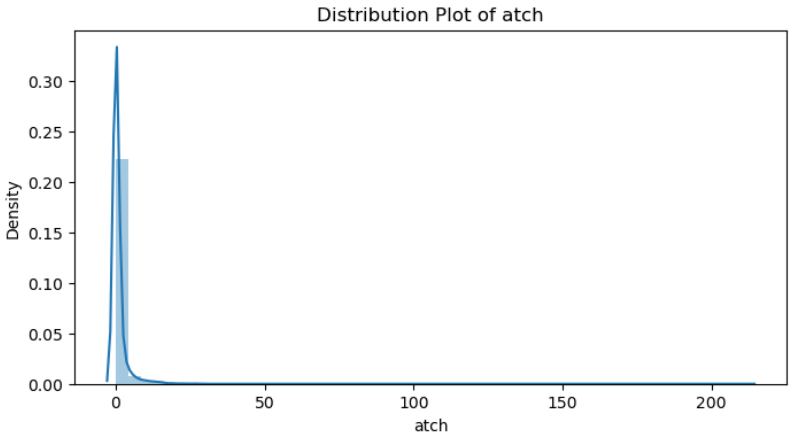
Center:
The peak of the histogram is near zero, indicating that the majority of pgscan values are very low.

Spread:
The histogram shows that while most pgscan values are close to zero. However, the frequency of these higher values is very low.

Outliers:
There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values

Median (Q2):
The line inside the box represents the median (Q2) of the "pgscan" values. The median is very close to zero, indicating that half of the data points are below this value and half are above. This suggests that the majority of the "pgscan" values are clustered around the lower end of the scale.

Interquartile Range (IQR):
The box represents the interquartile range, which is the range between the first quartile (Q1) and the third quartile (Q3). This range is very narrow, suggesting that the middle 50% of the "pgscan" values are closely packed together near the lower end of the scale.
The lower edge of the box (Q1) is slightly above zero, while the upper edge (Q3) is also relatively close to zero, indicating that the interquartile range is small.



Whiskers:

The whiskers extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles.

The lower whisker extends to zero.

The upper whisker extends to a value slightly above 200, indicating that most data points are below this value.

The whiskers encompass the vast majority of the data points, but the presence of a significant number of outliers beyond the upper whisker suggests a wider spread in higher values.

Outliers:

Outliers are represented by points beyond the whiskers. There are numerous outliers extending upwards, with values reaching as high as approximately 1200. These outliers indicate the presence of some exceptionally high values in the dataset.

The outliers form a long tail, which suggests a right-skewed distribution where most data points are clustered near the lower end, but a few data points are much higher.

Shape of the Distribution:

The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.

Center:

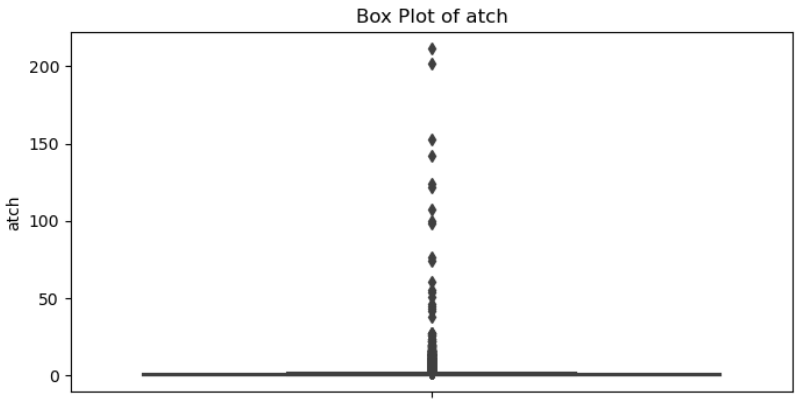
The peak of the histogram is near zero, indicating that the majority of atch values are very low.

Spread:

The histogram shows that while most atch values are close to zero. However, the frequency of these higher values is very low.

Outliers:

There are no distinct bars far separated from the rest, indicating that there are no



significant outliers, but the long tail suggests occasional higher values

Median (Q2):

The line inside the box represents the median (Q2) of the " atch" values. The median is very close to zero, indicating that half of the data points are below this value and half are above. This suggests that the majority of the "pgscan" values are clustered around the lower end of the scale.

Interquartile Range (IQR):

The box represents the interquartile range, which is the range between the first quartile (Q1) and the third quartile (Q3). This range is very narrow, suggesting that the middle 50% of the " atch " values are closely packed together near the lower end of the scale.

The lower edge of the box (Q1) is slightly above zero, while the upper edge (Q3) is also relatively close to zero, indicating that the interquartile range is small.

Whiskers:

The whiskers extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles.

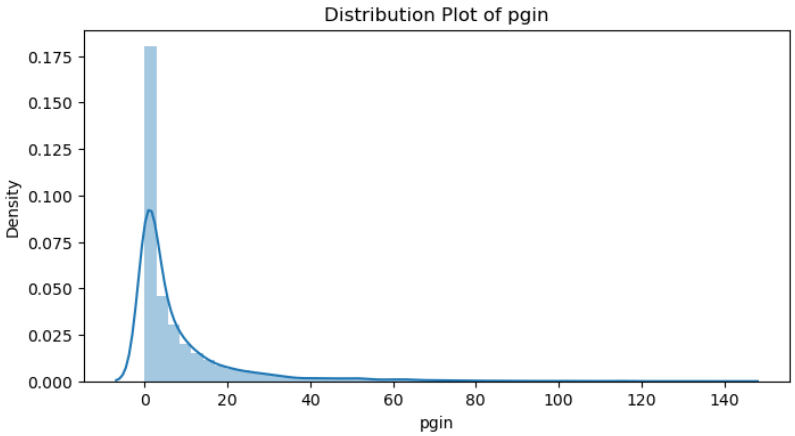
The lower whisker extends to zero.

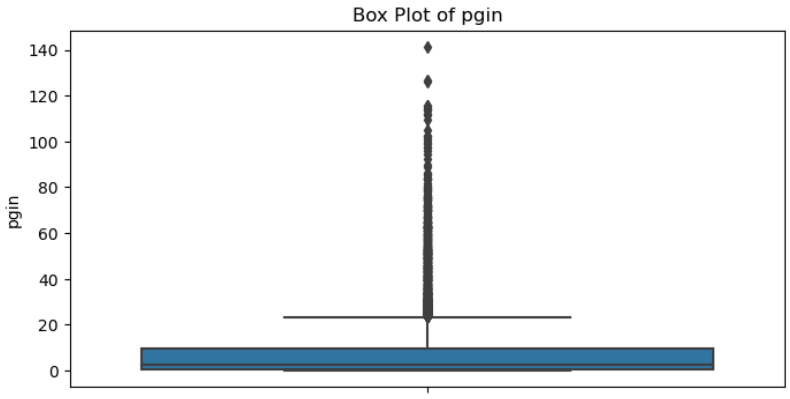
The upper whisker extends to a value slightly above 200, indicating that most data points are below this value.

The whiskers encompass the vast majority of the data points, but the presence of a significant number of outliers beyond the upper whisker suggests a wider spread in higher values.

Outliers:

Outliers are represented by points beyond the whiskers. There are numerous outliers extending upwards, with values reaching as high as approximately 200. These outliers indicate the presence of some exceptionally high values in the dataset.





The outliers form a long tail, which suggests a right-skewed distribution where most data points are clustered near the lower end, but a few data points are much higher.

Shape of the Distribution:

The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.

Center:

The peak of the histogram is near zero, indicating that the majority of pgin values are very low.

Spread:

The histogram shows that while most pgin values are close to zero. However, the frequency of these higher values is very low.

Outliers:

There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values

Median (Q2):

. The median pgin value is close to zero, indicating that at least half of the data points have a low Number of page-in requests per second.

. Interquartile Range (IQR):

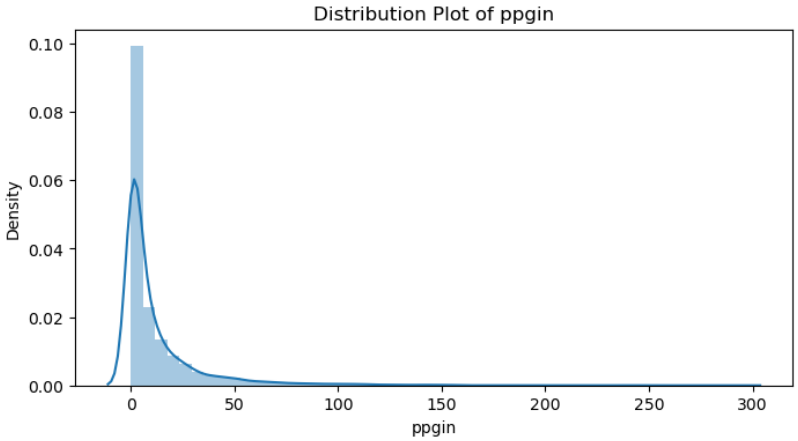
. The box itself is very narrow and close to zero, suggesting that the middle 50% of the pgin values are very low and close to each other.

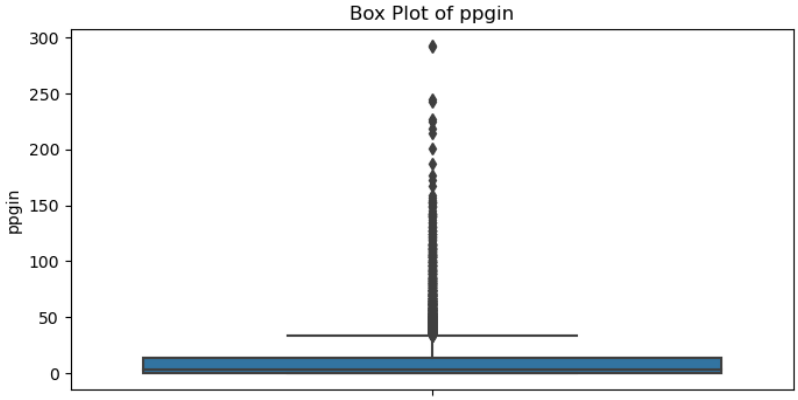
. Whiskers:

. The whiskers extend further than the box but are still relatively short compared to the overall range of pgin values, indicating that most of the data points are close to the lower end of the scale.

. Outliers:

. There are several outliers far beyond the whiskers, with some values reaching over 20 exec. This shows that while the majority of the pgin values are low, there are significant instances where page-in





requests per second spike to very high levels.

Shape of the Distribution:

The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.

Center:

The peak of the histogram is near zero, indicating that the majority of ppgin values are very low.

Spread:

The histogram shows that while most ppgin values are close to zero. However, the frequency of these higher values is very low.

Outliers:

There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values

Median (Q2):

. The median ppgin value is close to zero, indicating that at least half of the data points have a low Number of pages paged in per second.

. Interquartile Range (IQR):

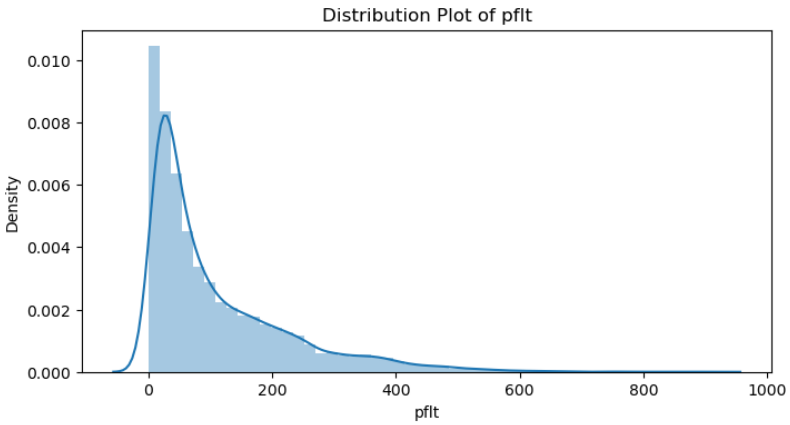
. The box itself is very narrow and close to zero, suggesting that the middle 50% of the ppgin values are very low and close to each other.

. Whiskers:

. The whiskers extend further than the box but are still relatively short compared to the overall range of ppgin values, indicating that most of the data points are close to the lower end of the scale.

. Outliers:

. There are several outliers far beyond the whiskers, with some values reaching over 20 exec. This shows that while the majority of the ppgin values are low, there are significant instances where pages paged in per second spike to very high levels.



Shape of the Distribution:

The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.

Center:

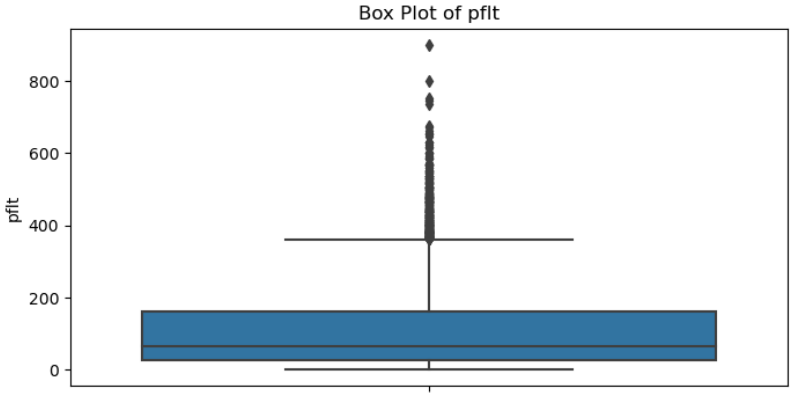
The peak of the histogram is near zero, indicating that the majority of pfit values are very low.

Spread:

The histogram shows that while most pfit values are close to zero. However, the frequency of these higher values is very low.

Outliers:

There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values



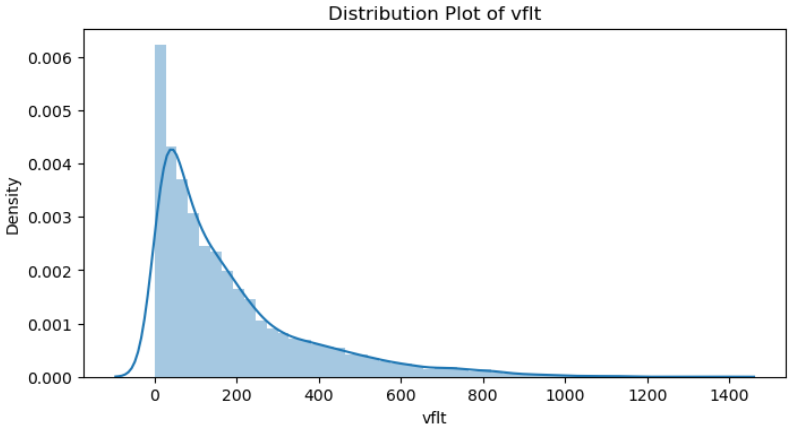
. **Central Tendency**:
Median: The line inside the box represents the median of the " pfit " values. The median is around 50, indicating that half of the data points are below this value and half are above.

. **Interquartile Range (IQR)**:
The box represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). This range spans from approximately 10 to 200.

This indicates that the middle 50% of the " pfit " values lie within this range.

. **Whiskers and Outliers**:
The "whiskers" extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles. The lower whisker extends to 0, while the upper whisker extends to approximately 1000, suggesting a wider range for the upper values.

. **Skewness**:
The distribution of "pfit" appears to be right-skewed due to the presence of higher values and the outliers extending significantly beyond the upper whisker.



Shape of the Distribution:

The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.

Center:

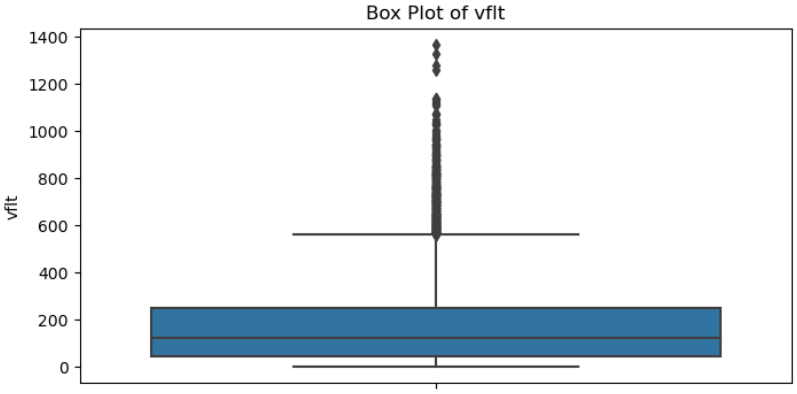
The peak of the histogram is near zero, indicating that the majority of vfit values are very low.

Spread:

The histogram shows that while most vfit values are close to zero. However, the frequency of these higher values is very low.

Outliers:

. There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values



. **Central Tendency**:

Median: The line inside the box represents the median of the " vfit " values. The median is around 100, indicating that half of the data points are below this value and half are above.

. **Interquartile Range (IQR)**:

The box represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). This range spans from approximately 10 to 220.

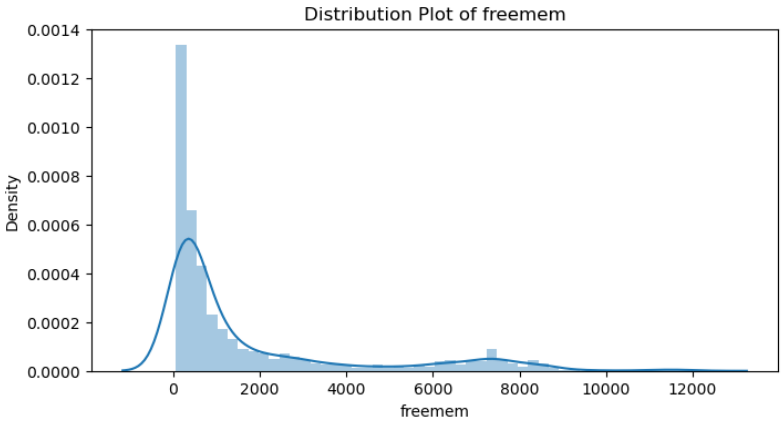
This indicates that the middle 50% of the " vfit " values lie within this range.

. **Whiskers and Outliers**:

The "whiskers" extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles. The lower whisker extends to 0, while the upper whisker extends to approximately 1350, suggesting a wider range for the upper values.

. **Skewness**:

The distribution of "vfit" appears to be right-skewed due to the presence of higher values and the outliers extending significantly beyond the upper whisker.



Shape of the Distribution:

The histogram is heavily right-skewed (positively skewed). Most of the data points are concentrated on the left side, close to zero.

Center:

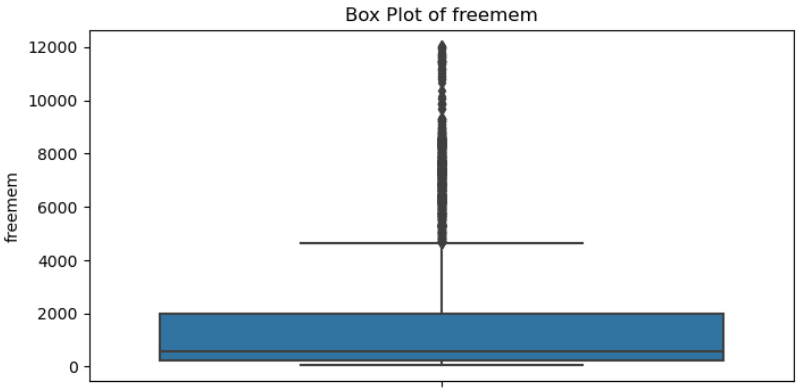
The peak of the histogram is at or near zero, indicating that the majority of freemem values are very low, with a high density at zero.

Spread:

The histogram shows that while most freemem values are close to zero, there are some instances where freemem values range up to around 7000. However, the frequency of these higher values is very low.

Outliers:

. There are no distinct bars far separated from the rest, indicating that there are no significant outliers in the data, but the long tail suggests occasional higher values.



. **Central Tendency**:

Median: The line inside the box represents the median of the "freemem" values. The median is around 100, indicating that half of the data points are below this value and half are above.

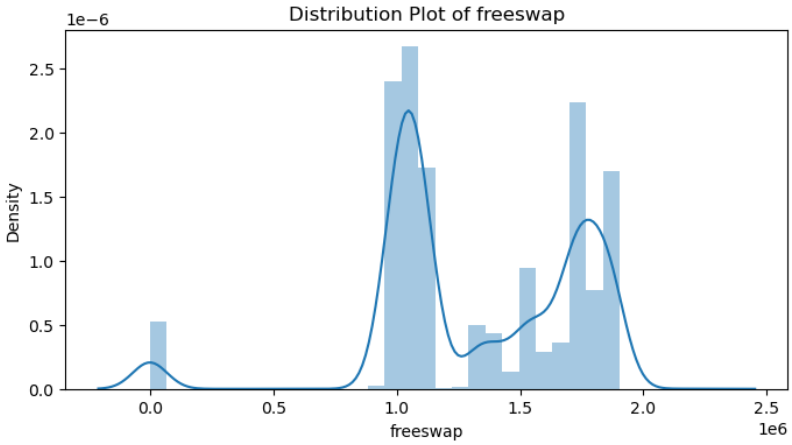
. **Interquartile Range (IQR)**:

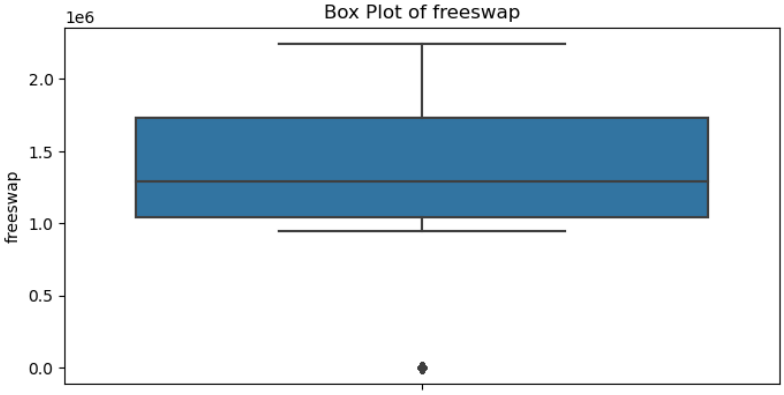
The box represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). This range spans from approximately 10 to 2000.

This indicates that the middle 50% of the "freemem" values lie within this range.

. **Whiskers and Outliers**:

The "whiskers" extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles. The lower whisker extends to 0, while the upper whisker extends to approximately





12000, suggesting a wider range for the upper values.

. **Skewness**:

The distribution of "freemem" appears to be right-skewed due to the presence of higher values and the outliers extending significantly beyond the upper whisker.

Shape of the Distribution:

. The distribution of freeswap is multimodal, showing multiple peaks. This suggests the presence of several clusters within the data.

.

. **Modes:**

. There are distinct peaks around 0, 1 million, and 2 million. These modes indicate that freeswap values tend to cluster around these points.

.

. **Spread:**

. The spread is quite wide, ranging from 0 to approximately 2.5 million. This indicates a significant variation in the freeswap values.

.

. **Density:**

. The density plot shows that the highest density is around 1 million, indicating that this is where freeswap values are most concentrated.

. There is also a notable density peak around 2 million, indicating another significant cluster of freeswap values.

.

. **Central Tendency:**

. The median value (the line inside the box) is around 1.5 million, indicating that half of the freeswap values are below this point and half are above.

.

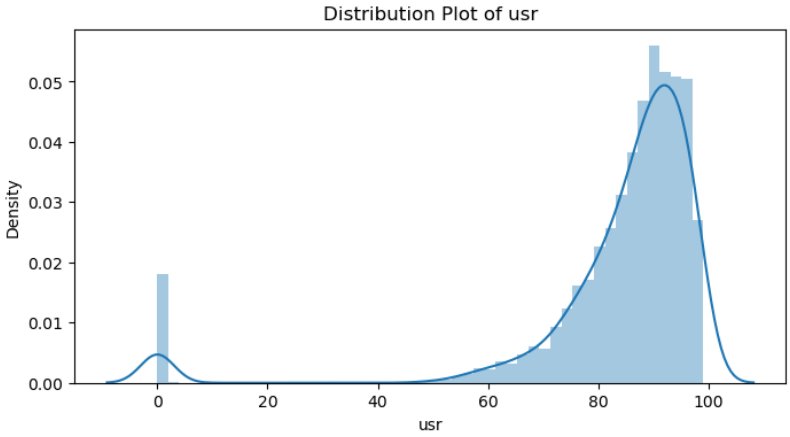
. **Interquartile Range (IQR):**

. The box represents the interquartile range (IQR), which contains the middle 50% of the data. For freeswap, the IQR extends from about 1 million to 2 million.

. This indicates that the majority of the freeswap values are concentrated between 1 million and 2 million.

.

. **Whiskers:**



. The whiskers extend from the quartiles to the minimum and maximum values within 1.5 times the IQR.
. The upper whisker extends slightly above 2 million, and the lower whisker extends to just above 0.5 million, indicating that most freeswap values fall within this range.

.

. Outliers:

. There is one clear outlier below 0.5 million, indicating that there are instances where the freeswap value is significantly lower than the rest of the data.
. Outliers are data points that fall outside the whiskers and can indicate unusual conditions or rare events.

.

. Distribution Shape:

. The distribution of usr shows a clear right-skewed pattern, with a concentration of values towards the higher end of the range.
. There is a significant peak around 80-100, indicating that a large number of observations fall within this range.

.

. Modes:

. There are two distinct peaks in the distribution. The first peak is around 0, and the second, more prominent peak is around 80-100.
. The peak around 0 suggests a subset of observations where usr is very low, possibly indicating periods of low activity or idleness.

.

. Density:

. The density reaches its highest point between 80 and 100, suggesting that this range is where the majority of usr values lie.
. The presence of the peak at 0 suggests a bimodal distribution, where the system alternates between low and high usage states.

.

. Range and Spread:

. The usr values span from 0 to slightly over 100, indicating a wide range of usage values.
. The right tail of the distribution is long, indicating a few observations with exceptionally high usr values.

.

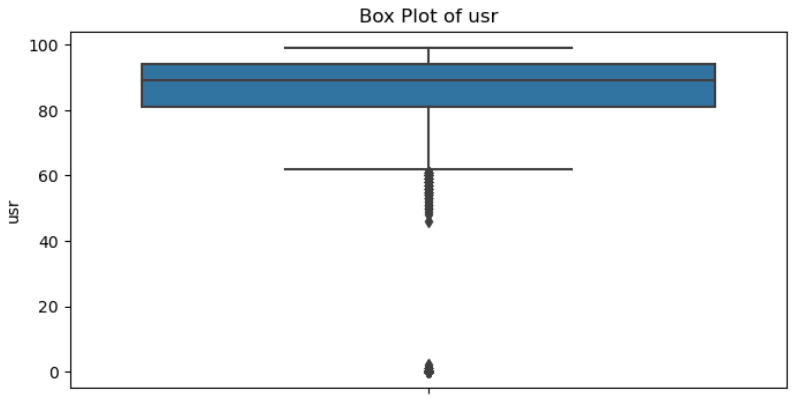


FIG 2

. **Median (Q2)**

. The line inside the box represents the median (Q2) of the "usr" values. The median is approximately 90, indicating that half of the data points are below this value and half are above.

.

. **Interquartile Range (IQR)**

. The box represents the interquartile range, which is the range between the first quartile (Q1) and the third quartile (Q3). This range spans from approximately 80 to 95.

. Q1 (lower quartile) is around 80.

. Q3 (upper quartile) is around 95.

. This indicates that the middle 50% of the "usr" values lie within this range.

.

. **Whiskers**

. The whiskers extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles.

. The lower whisker extends to around 60, indicating the smallest non-outlier value.

. The upper whisker extends to around 100, indicating the largest non-outlier value.

. The whiskers encompass most of the data points, but there are several outliers beyond the whiskers.

.

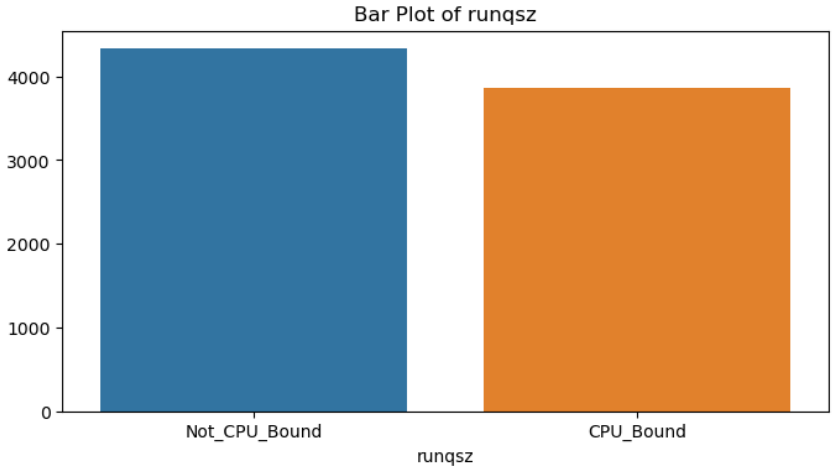
. **Outliers**

. Outliers are represented by points beyond the whiskers. There are several outliers below the lower whisker, with values extending down to around 0.

. These outliers indicate the presence of some exceptionally low values in the dataset and suggest a left-skewed distribution for these outlier values.

.

Categorical Data Type:



Comparison of Categories:

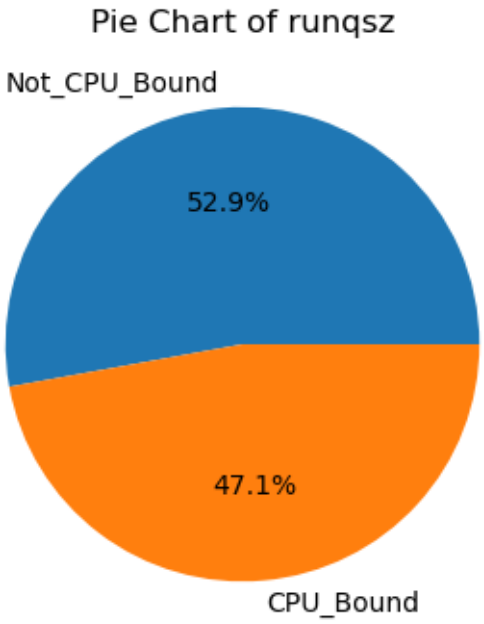
Not_CPU_Bound: The bar representing the "Not_CPU_Bound" category has a height of approximately 4500, indicating that the "runqsz" value for this category is around 4500.

CPU_Bound: The bar representing the "CPU_Bound" category has a height of approximately 4000, indicating that the "runqsz" value for this category is around 4000.

Relative Difference:

The "Not_CPU_Bound" category has a higher "runqsz" value compared to the "CPU_Bound" category. There is a difference of roughly 500 units between the two categories.

This suggests that the "Not_CPU_Bound" category has a larger queue size (runqsz) compared to the "CPU_Bound" category.



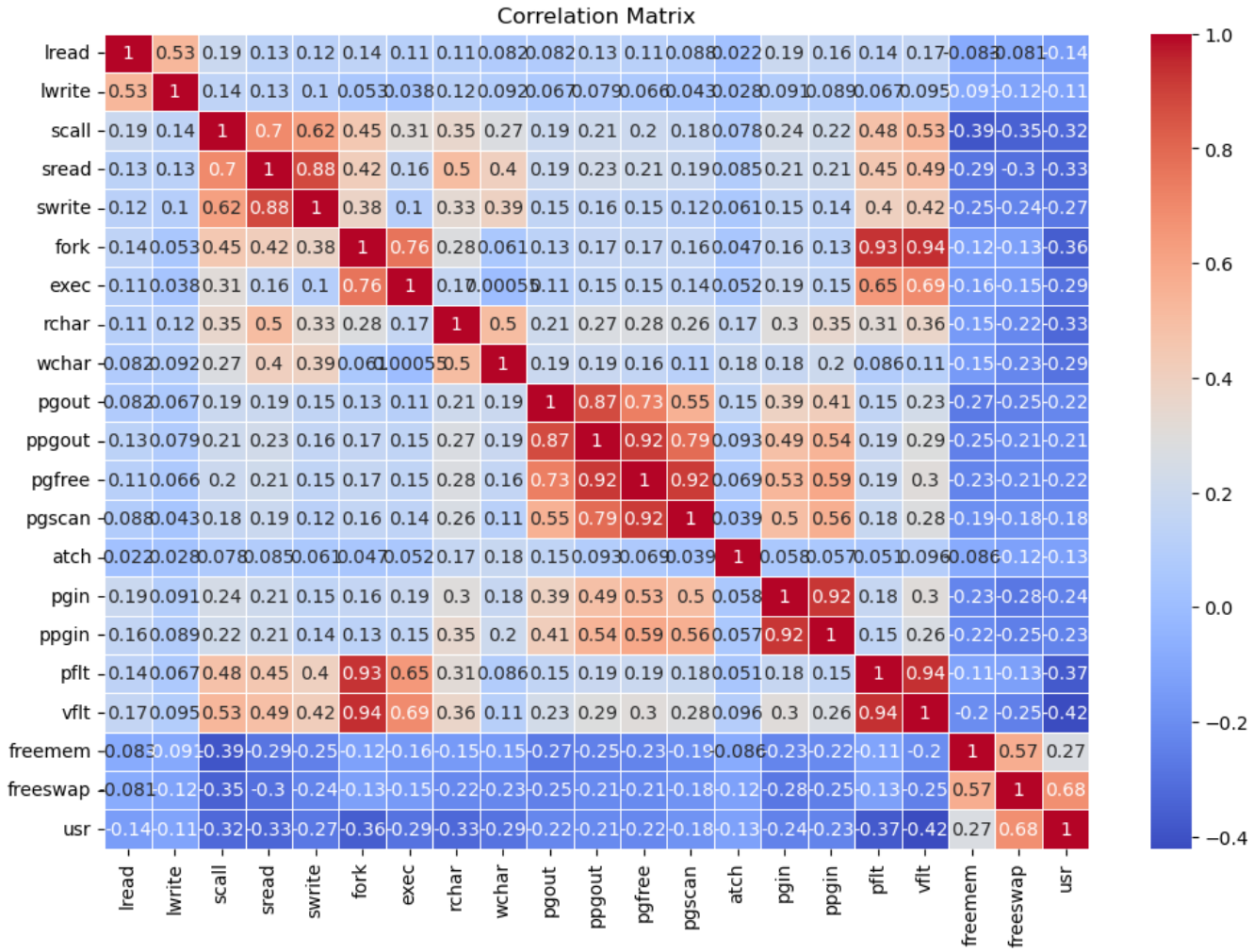
Categories Proportions:
Not_CPU_Bound category covers 52.9% of whole data and remaining 47.1% was covered by CPU_Bound category.

Not_CPU_Bound has largest slices.

This means that Process run queue size was mostly less than 2 for more than 50% of population.

Problem 1.1.4 Multivariate analysis:

Heat Map:

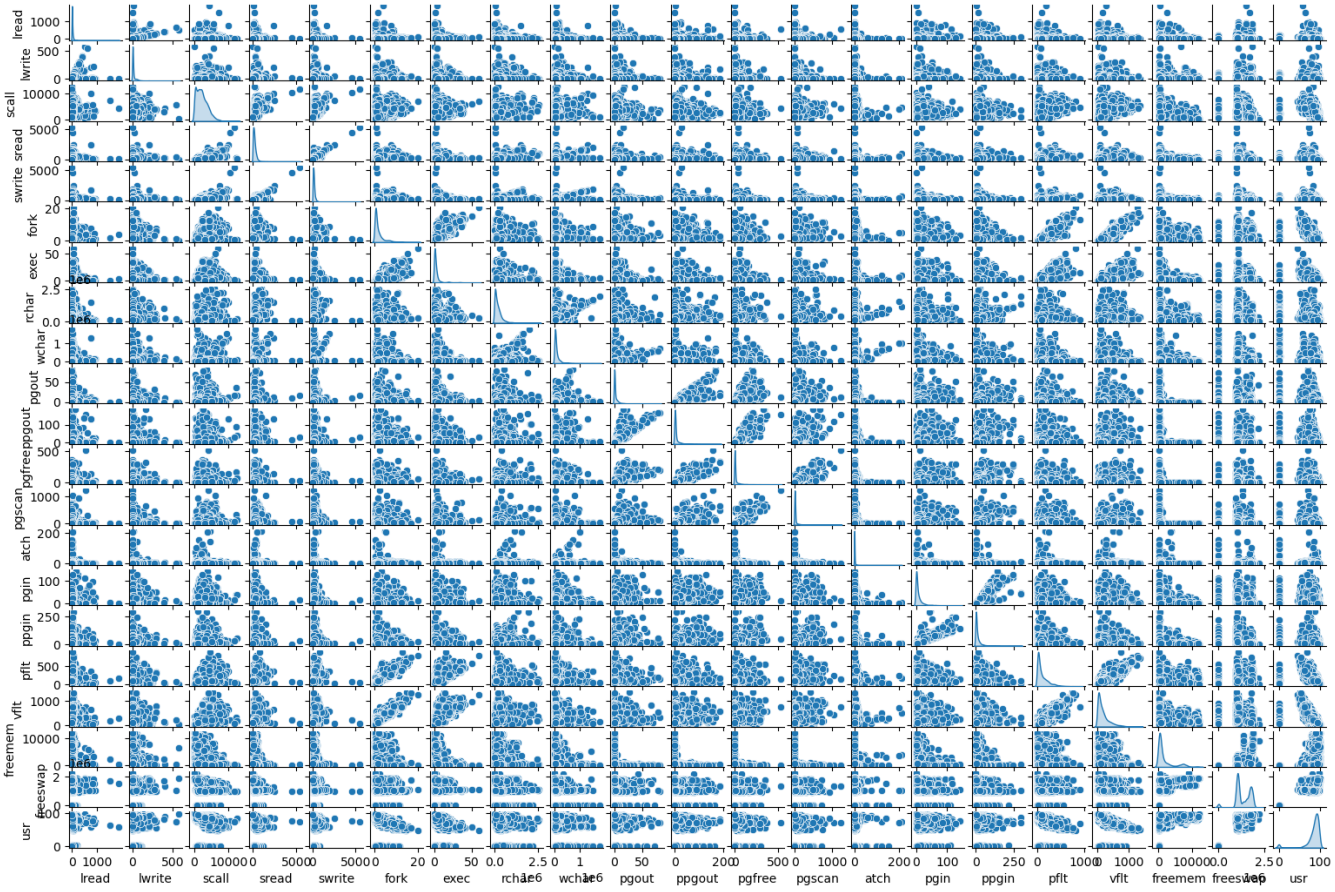


Insights:

- Right-Skewed Distributions: Most variables have right-skewed distributions, indicating that most values are clustered near zero with a few higher values.
- Correlations: Strong positive correlations are observed between several pairs of variables, such as swrite and sread, fork and exec, pgfree and pgscan, and pflt and vflt. These correlations suggest that these variables tend to increase together.
- Negative Correlations: A few negative correlations are observed, such as between lread and lwrite, and freemem and usr. These correlations suggest that these variables tend to move in opposite directions.
- Outliers: Several scatterplots show the presence of outliers, particularly in variables like lread, pgfree, and pgscan. These outliers can significantly impact the analysis and should be investigated further.

- Implications: The strong positive relationships can be leveraged in predictive modeling, while understanding these relationships can help in optimizing resource allocation and performance tuning.

Pair Plot:



General Observations

1. **Distribution of Individual Variables**:
    - The diagonal of the plot shows the distribution of each variable. Most variables appear to have skewed distributions, with some showing right skewness (e.g., lread, sread, pgfree).

Key Relationships

1. **swrite and sread**:
    - The scatterplot shows a strong positive linear relationship, confirming the high correlation observed earlier. This suggests that as swrite increases, sread also increases in a linear fashion.
2. **fork and exec**:
    - There is a clear positive linear relationship between fork and exec, indicating that these variables tend to increase together. This relationship is strong and linear.
3. **pgfree and pgscan**:

- The scatterplot shows a positive linear relationship, indicating that higher values of pgfree are associated with higher values of pgscan.

4. **pflt and vflt**:
    - There is a very strong positive linear relationship between pflt and vflt, suggesting that these two variables are highly likely to increase together.

Other Notable Relationships

1. **scall and sread**:
    - A positive relationship is observed, suggesting that higher values of scall are associated with higher values of sread.
2. **pgin and pgpin**:
    - A strong positive relationship is evident, indicating that these variables tend to increase together.
3. **freemem and usr**:
    - There appears to be a negative relationship, suggesting that higher usr values are associated with lower freemem values.

Outliers and Variability

1. **Presence of Outliers**:
    - Several scatterplots show the presence of outliers, particularly in variables like lread, pgfree, and pgscan. These outliers can significantly impact the analysis and should be investigated further.
2. **Variability**:
    - The scatterplots show varying degrees of spread, with some relationships being more tightly clustered (e.g., swrite and sread) and others showing more variability (e.g., pgin and pgpin).

Summary

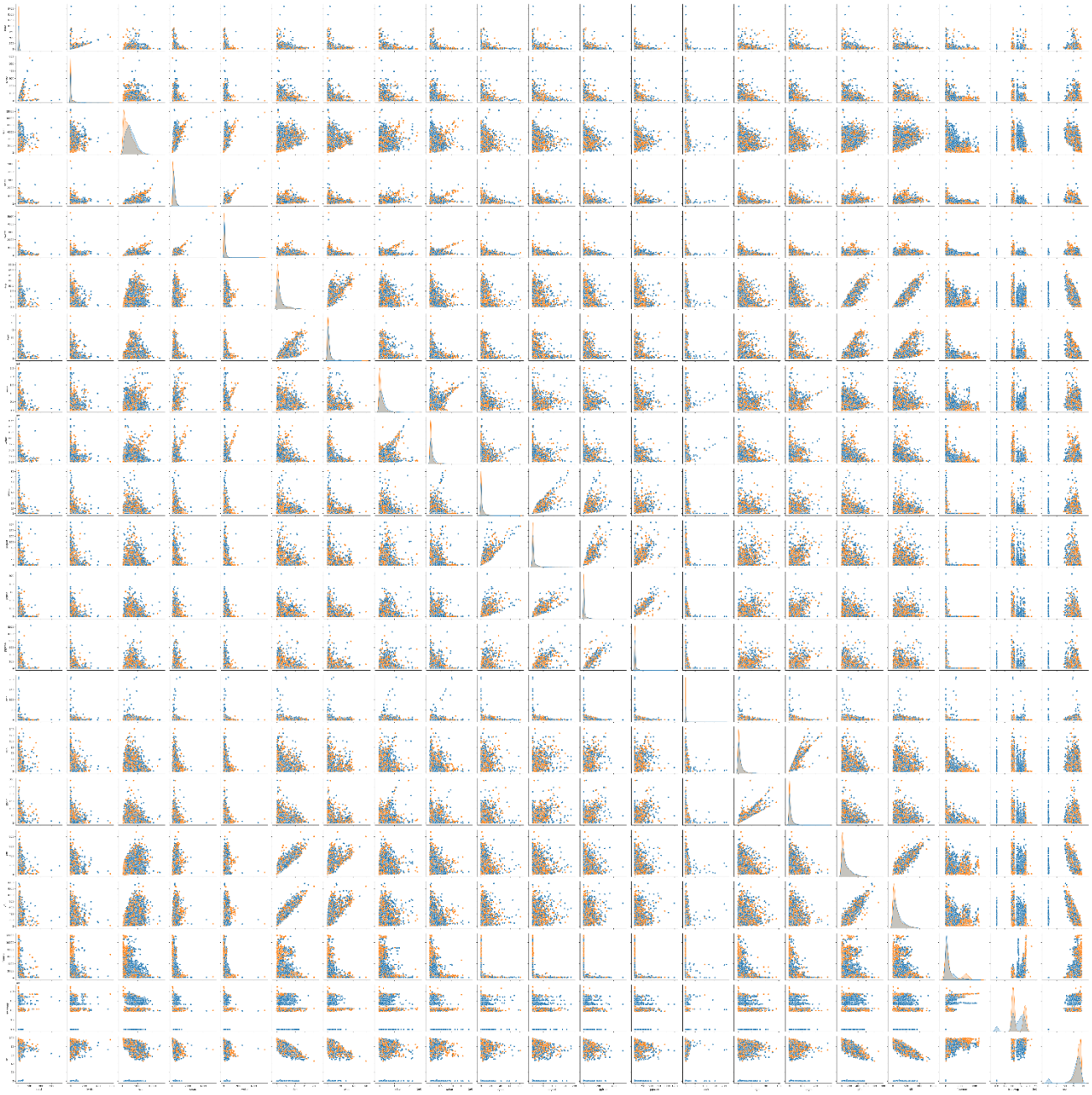
- **Strong Positive Relationships**: **swrite** with **sread**, **fork** with **exec**, **pgfree** with **pgscan**, and **pflt** with **vflt**.
- **Moderate Positive Relationships**: **scall** with **sread**, **pgin** with **pgpin**.
- **Negative Relationships**: **freemem** with **usr**.
- **Outliers**: Multiple scatterplots indicate the presence of outliers, particularly in variables like **lread**, **pgfree**, and **pgscan**.

Implications

- **Predictive Modeling**: The strong positive relationships can be leveraged in predictive modeling, where one variable can be used to predict the other.
- **Resource Allocation**: Understanding these relationships can help in optimizing resource allocation and performance tuning, particularly for variables like **freemem** and **usr**.

- **Further Analysis**: The presence of outliers and skewed distributions suggests the need for further analysis, possibly including data transformation and outlier treatment.

Problem 1.1.5 Use appropriate visualizations to identify the patterns and insights:



Insights

- Strong Positive Relationships: swrite with sread, fork with exec, pgfree with pgscan, pflt with vflt.
- Moderate Positive Relationships: scall with sread, pgin with pgpin.
- Negative Relationships: freemem with usr.
- Outliers: Multiple scatterplots indicate the presence of outliers, particularly in variables like lread, pgfree, and pgscan.
- Distributions: The distributions of variables vary, with some showing skewness.

USE APPROPRIATE VISUALIZATIONS TO IDENTIFY THE PATTERNS AND INSIGHTS

- Above we have Histogram & box plot for Individual Numerical data and Bar plot & pie chart for Individual categorical data.
- For Comparison of multple numerical data we have heatmap & pairplot.
- For Comparison between multiple numerical and categorical data we have pairplot with digonal showing the categorical data distribution

Problem 1.1.6 Key meaningful observations on individual variables and the relationship between variables:

lread:

- Right-Skewed Distributions: Most variables have right-skewed distributions, indicating that most values are clustered near zero with a few higher values.
- High Density Near Zero: The majority of data points have lower values, indicating that lower values are much more common.
- Long Tail: The plots have long tails extending towards higher values, indicating the presence of some high values, but these are much less frequent compared to the lower values.
- Potential Outliers: The presence of the long tail suggests there may be outliers or extreme values far from the mean, which can have significant effects on the mean but less on the median.
- Central Tendency: The median is located close to the lower end of the scale, and the interquartile range (IQR) is quite narrow, indicating that the middle 50% of the data points are close to each other and clustered towards the lower end of the scale.
- Whiskers and Outliers: The whiskers are very short, suggesting that the majority of the data points are not spread out widely. There are numerous outliers beyond the whiskers, particularly in variables like "lread" and "lwrite".
- Skewness: The distributions are right-skewed due to the presence of higher values and the outliers extending significantly beyond the upper whisker.
- Implications: The high skewness suggests that these variables may not follow a normal distribution, which could affect certain statistical analyses that assume normality. Transformations (e.g., log transformation) might be necessary if normality is a requirement for the analysis. The presence of outliers should be considered in further analysis, as they can significantly impact model performance and interpretation.

lwrite:

- Right-Skewed Distributions: Most variables have right-skewed distributions, indicating that most values are clustered near zero with a few higher values.
- High Density Near Zero: The majority of data points have lower values, indicating that lower values are much more common.
- Long Tail: The plots have long tails extending towards higher values, indicating the presence of some high values, but these are much less frequent compared to the lower values.
- Potential Outliers: The presence of the long tail suggests there may be outliers or extreme values far from the mean. These outliers can have significant effects on the mean but less on the median.
- Central Tendency: The median is located close to the lower end of the scale, and the interquartile range (IQR) is quite narrow, indicating that the middle 50% of the data points are close to each other and clustered towards the lower end.
- Whiskers and Outliers: The whiskers are very short, suggesting that the majority of the data points are not spread out widely. There are numerous outliers beyond the whiskers, particularly in variables like "lread" and "lwrite".

- Skewness: The distributions are right-skewed due to the presence of higher values and the outliers extending significantly beyond the upper whisker.
- Implications for Analysis: The high skewness suggests that these variables may not follow a normal distribution, which could affect certain statistical analyses that assume normality. Transformations (e.g., log transformation) might be necessary if normality is a requirement for the analysis. The presence of extreme values (outliers) should be considered in further analysis, as they can significantly impact model performance and interpretation.

scall:

- Right-Skewed Distributions: Most variables have right-skewed distributions, indicating that most values are clustered near zero with a few higher values.
- High Density Near Zero: The majority of data points have lower values, indicating that lower values are much more common.
- Long Tail: The plots have long tails extending towards higher values, indicating the presence of some high values, but these are much less frequent compared to the lower values.
- Potential Outliers: The presence of the long tail suggests there may be outliers or extreme values far from the mean. These outliers can have significant effects on the mean but less on the median.
- Central Tendency: The median is located close to the lower end of the scale, and the interquartile range (IQR) is quite narrow, indicating that the middle 50% of the data points are close to each other and clustered towards the lower end.
- Whiiskers and Outliers: The whiskers are very short, suggesting that the majority of the data points are not spread out widely. There are numerous outliers beyond the whiskers, particularly in variables like "lread" and "lwrite".
- Skewness: The distributions are right-skewed due to the presence of higher values and the outliers extending significantly beyond the upper whisker.
- Implications for Analysis: The high skewness suggests that these variables may not follow a normal distribution, which could affect certain statistical analyses that assume normality. Transformations (e.g., log transformation) might be necessary if normality is a requirement for the analysis. The presence of extreme values (outliers) should be considered in further analysis, as they can significantly impact model performance and interpretation.

sread:

- Distribution Shape: The distribution of sread (number of system read calls per second) is right-skewed. The majority of the values are concentrated at the lower end, with a sharp decline in frequency as sread values increase.
- Central Tendency: The peak of the distribution is at the lower values of sread, indicating that most system read calls per second are relatively low. The mode (most frequent value) appears at the very beginning of the distribution, close to zero.
- Spread and Variability: The range of sread extends from near zero to over 5000, indicating a wide variability. Although the majority of sread values are low, there are some significantly higher values, suggesting occasional bursts of high system read activity.
- Skewness: The right skewness indicates that while most sread values are low, there are fewer but significant instances of high sread values.
- Central Tendency: The line inside the box represents the median of the "sread" values. The median is quite close to zero, indicating that half of the data points are below this value and half are above.
- Interquartile Range (IQR): The box represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). This range is also very close to zero, suggesting that the middle 50% of the "sread" values are clustered near zero.
- Whiskers and Outliers: The "whiskers" extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles. The lower whisker extends to zero, while the upper whisker extends to a value significantly higher than the IQR, indicating some spread in the higher values. here are numerous outliers beyond the upper whisker, with values extending up to around 5500. These outliers indicate the presence of some exceptionally high values in the dataset.

- Skewness: he distribution of "sread" is highly right-skewed due to the presence of higher values and the outliers extending significantly beyond the upper whisker.

swrite:

- Distribution Shape: The distribution of swrite is right-skewed, with most values concentrated on the lower end and a sharp decline as swrite values increase.
- Central Tendency: The peak of the distribution is at the lower values of swrite, indicating that most system write calls per second are relatively low. The mode (most frequent value) appears very close to zero.
- Spread and Variability: The range of swrite extends from near zero to over 5000, indicating a wide variability. While the majority of swrite values are low, there are instances of significantly higher values, suggesting occasional bursts of high system write activity.
- Skewness: The right skewness indicates that while most swrite values are low, there are fewer but significant instances of high swrite values.
- Median and Interquartile Range (IQR): The median swrite value is close to zero, indicating that at least half of the data points have a low number of system write calls per second. The interquartile range (IQR) is very narrow and close to zero, suggesting that the middle 50% of the swrite values are very low and close to each other.
- Whiskers: The whiskers extend further than the box but are still relatively short compared to the overall range of swrite values, indicating that most of the data points are close to the lower end of the scale.
- Outliers: There are several outliers far beyond the whiskers, with some values reaching over 5000 swrite. This shows that while the majority of the swrite values are low, there are significant instances where the system write calls per second spike to very high levels.

fork:

- Shape of the Distribution: The histogram is heavily right-skewed (positively skewed). Most of the data points are concentrated on the left side, close to zero.
- Center: The peak of the histogram is at or near zero, indicating that the majority of fork values are very low, with a high density at zero.
- Spread: The histogram shows that while most fork values are close to zero, there are some instances where fork values range up to around 20. However, the frequency of these higher values is very low.
- Outliers: There are no distinct bars far separated from the rest, indicating that there are no significant outliers in the data, but the long tail suggests occasional higher values.
- Median (Q2): The median fork value is close to zero, indicating that at least half of the data points have a low Number of system fork calls per second.
- Interquartile Range (IQR): The box itself is very narrow and close to zero, suggesting that the middle 50% of the fork values are very low and close to each other.
- Whiskers: The whiskers extend further than the box but are still relatively short compared to the overall range of fork values, indicating that most of the data points are close to the lower end of the scale.
- Outliers: There are several outliers far beyond the whiskers, with some values reaching over 20 fork. This shows that while the majority of the fork values are low, there are significant instances where the system fork calls per second spike to very high levels.

exec:

- Shape of the Distribution: The histogram is heavily right-skewed (positively skewed). Most of the data points are concentrated on the left side, close to zero.
- Center: The peak of the histogram is at or near zero, indicating that the majority of exec values are very low, with a high density at zero.
- Spread: The histogram shows that while most exec values are close to zero, there are some instances where exec values range up to around 20. However, the frequency of these higher values is very low.
- Outliers: There are no distinct bars far separated from the rest, indicating that there are no significant outliers in the data, but the long tail suggests occasional higher values.

- Median (Q2): The median exec value is close to zero, indicating that at least half of the data points have a low Number of system exec calls per second.
- Interquartile Range (IQR): The box itself is very narrow and close to zero, suggesting that the middle 50% of the exec values are very low and close to each other.
- Whiskers: The whiskers extend further than the box but are still relatively short compared to the overall range of exec values, indicating that most of the data points are close to the lower end of the scale.
- Outliers: There are several outliers far beyond the whiskers, with some values reaching over 20 exec. This shows that while the majority of the exec values are low, there are significant instances where the system exec calls per second spike to very high levels.

rchar:

- Shape of the Distribution: The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.
- Center: The peak of the histogram is near zero, indicating that the majority of rchar values are very low.
- Spread: The histogram shows that while most rchar values are close to zero, there are some instances where rchar values extend up to approximately 2.5 million. However, the frequency of these higher values is very low.
- Outliers:There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values.
- Median (Q2): The median rchar value is close to zero, indicating that at least half of the data points have a low Number of characters transferred per second by system read calls.
- Interquartile Range (IQR): The box itself is very narrow and close to zero, suggesting that the middle 50% of the rchar values are very low and close to each other.
- Whiskers: The whiskers extend further than the box but are still relatively short compared to the overall range of rchar values, indicating that most of the data points are close to the lower end of the scale.
- Outliers: There are several outliers far beyond the whiskers, with some values reaching over 20 exec. This shows that while the majority of the rchar values are low, there are significant instances where the characters transferred per second by system read calls spike to very high levels.

wchar:

- Shape of the Distribution: The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.
- Center: The peak of the histogram is near zero, indicating that the majority of wchar values are very low.
- Spread: he histogram shows that while most wchar values are close to zero, there are some instances where wchar values extend up to approximately 0.5 million. However, the frequency of these higher values is very low.
- Outliers: There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values
- Median (Q2): The median wchar value is close to zero, indicating that at least half of the data points have a low Number of characters transfreed per second by system write calls
- Interquartile Range (IQR): The box itself is very narrow and close to zero, suggesting that the middle 50% of the wchar values are very low and close to each other.
- Whiskers: The whiskers extend further than the box but are still relatively short compared to the overall range of wchar values, indicating that most of the data points are close to the lower end of the scale.
- Outliers: There are several outliers far beyond the whiskers, with some values reaching over 20 exec. This shows that while the majority of the wchar values are low, there are significant instances where the characters transferred per second by system write calls spike to very high levels.

pgout:

- Shape of the Distribution:The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.
- Center: The peak of the histogram is near zero, indicating that the majority of pgout values are very low.
- Spread: The histogram shows that while most pgout values are close to zero. However, the frequency of these higher values is very low.
- Outliers: There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values
- Median (Q2): The median pgout value is close to zero, indicating that at least half of the data points have a low Number of page out requests per second.
- Interquartile Range (IQR): The box itself is very narrow and close to zero, suggesting that the middle 50% of the pgout values are very low and close to each other.
- Whiskers: The whiskers extend further than the box but are still relatively short compared to the overall range of pgout values, indicating that most of the data points are close to the lower end of the scale.
- Outliers: There are several outliers far beyond the whiskers, with some values reaching over 20 exec. This shows that while the majority of the pgout values are low, there are significant instances where page out requests per second spike to very high levels.

ppgout:

- Shape of the Distribution: The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.
- Center: The peak of the histogram is near zero, indicating that the majority of ppgout values are very low.
- Spread: The histogram shows that while most ppgout values are close to zero. However, the frequency of these higher values is very low.
- Outliers: There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values
- Median (Q2): The median ppgout value is close to zero, indicating that at least half of the data points have a low Number of page, paged out requests per second.
- Interquartile Range (IQR): The box itself is very narrow and close to zero, suggesting that the middle 50% of the ppgout values are very low and close to each other.
- Whiskers: The whiskers extend further than the box but are still relatively short compared to the overall range of ppgout values, indicating that most of the data points are close to the lower end of the scale.
- Outliers: There are several outliers far beyond the whiskers, with some values reaching over 20 exec. This shows that while the majority of the ppgout values are low, there are significant instances where page, paged out requests per second spike to very high levels.

pgfree:

- Shape of the Distribution: The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.
- Center: The peak of the histogram is near zero, indicating that the majority of pgfree values are very low.
- Spread: The histogram shows that while most pgfree values are close to zero. However, the frequency of these higher values is very low.
- Outliers: There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values
- Median (Q2): The median pgfree value is close to zero, indicating that at least half of the data points have a low Number of pages per second placed on the free list..
- Interquartile Range (IQR): The box itself is very narrow and close to zero, suggesting that the middle 50% of the pgfree values are very low and close to each other.

- Whiskers: The whiskers extend further than the box but are still relatively short compared to the overall range of pgfree values, indicating that most of the data points are close to the lower end of the scale.
- Outliers: There are several outliers far beyond the whiskers, with some values reaching over 20 exec. This shows that while the majority of the pgfree values are low, there are significant instances where pages per second placed on the free list spike to very high levels.

pgscan:

- Shape of the Distribution: The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.
- Center: The peak of the histogram is near zero, indicating that the majority of pgscan values are very low.
- Spread: The histogram shows that while most pgscan values are close to zero. However, the frequency of these higher values is very low.
- Outliers: There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values
- Median (Q2): The line inside the box represents the median (Q2) of the "pgscan" values. The median is very close to zero, indicating that half of the data points are below this value and half are above. This suggests that the majority of the "pgscan" values are clustered around the lower end of the scale.
- Interquartile Range (IQR): The box represents the interquartile range, which is the range between the first quartile (Q1) and the third quartile (Q3). This range is very narrow, suggesting that the middle 50% of the "pgscan" values are closely packed together near the lower end of the scale. The lower edge of the box (Q1) is slightly above zero, while the upper edge (Q3) is also relatively close to zero, indicating that the interquartile range is small.
- Whiskers: The whiskers extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles. The lower whisker extends to zero. The upper whisker extends to a value slightly above 200, indicating that most data points are below this value. The whiskers encompass the vast majority of the data points, but the presence of a significant number of outliers beyond the upper whisker suggests a wider spread in higher values.
- Outliers: Outliers are represented by points beyond the whiskers. There are numerous outliers extending upwards, with values reaching as high as approximately 1200. These outliers indicate the presence of some exceptionally high values in the dataset. The outliers form a long tail, which suggests a right-skewed distribution where most data points are clustered near the lower end, but a few data points are much higher.

atch:

- Shape of the Distribution: The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.
- Center: The peak of the histogram is near zero, indicating that the majority of atch values are very low.
- Spread: The histogram shows that while most atch values are close to zero. However, the frequency of these higher values is very low.
- Outliers: There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values
- Median (Q2): The line inside the box represents the median (Q2) of the " atch" values. The median is very close to zero, indicating that half of the data points are below this value and half are above. This suggests that the majority of the "pgscan" values are clustered around the lower end of the scale.
- Interquartile Range (IQR): The box represents the interquartile range, which is the range between the first quartile (Q1) and the third quartile (Q3). This range is very narrow, suggesting that the middle 50% of the " atch " values are closely packed together near the lower end of the scale.The lower edge of the box (Q1) is slightly above zero, while the upper edge (Q3) is also relatively close to zero, indicating that the interquartile range is small.
- Whiskers: The whiskers extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles. The lower whisker extends to zero. The upper whisker extends to a value slightly above 200, indicating that most data points are below this value. The whiskers encompass the

vast majority of the data points, but the presence of a significant number of outliers beyond the upper whisker suggests a wider spread in higher values.

- Outliers: Outliers are represented by points beyond the whiskers. There are numerous outliers extending upwards, with values reaching as high as approximately 200. These outliers indicate the presence of some exceptionally high values in the dataset. The outliers form a long tail, which suggests a right-skewed distribution where most data points are clustered near the lower end, but a few data points are much higher.

pgin:

- Shape of the Distribution: The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.
- Center: The peak of the histogram is near zero, indicating that the majority of pgin values are very low.
- Spread: The histogram shows that while most pgin values are close to zero. However, the frequency of these higher values is very low.
- Outliers: There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values
- Median (Q2): The median pgin value is close to zero, indicating that at least half of the data points have a low Number of page-in requests per second.
- Interquartile Range (IQR): The box itself is very narrow and close to zero, suggesting that the middle 50% of the pgin values are very low and close to each other.
- Whiskers: The whiskers extend further than the box but are still relatively short compared to the overall range of pgin values, indicating that most of the data points are close to the lower end of the scale.
- Outliers: There are several outliers far beyond the whiskers, with some values reaching over 20 exec. This shows that while the majority of the pgin values are low, there are significant instances where page-in requests per second spike to very high levels.

ppgin:

- Shape of the Distribution: The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.
- Center: The peak of the histogram is near zero, indicating that the majority of ppgin values are very low.
- Spread: The histogram shows that while most ppgin values are close to zero. However, the frequency of these higher values is very low.
- Outliers: There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values
- Median (Q2): The median ppgin value is close to zero, indicating that at least half of the data points have a low Number of pages paged in per second.
- Interquartile Range (IQR): The box itself is very narrow and close to zero, suggesting that the middle 50% of the ppgin values are very low and close to each other.
- Whiskers: The whiskers extend further than the box but are still relatively short compared to the overall range of ppgin values, indicating that most of the data points are close to the lower end of the scale.
- Outliers: There are several outliers far beyond the whiskers, with some values reaching over 20 exec. This shows that while the majority of the ppgin values are low, there are significant instances where pages paged in per second spike to very high levels.

pflt:

- Shape of the Distribution: The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.
- Center: he peak of the histogram is near zero, indicating that the majority of pfit values are very low.
- Spread: The histogram shows that while most pfit values are close to zero. However, the frequency of these higher values is very low.

- Outliers: There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values
- Central Tendency: The line inside the box represents the median of the " pfit " values. The median is around 50, indicating that half of the data points are below this value and half are above.
- Interquartile Range (IQR): The box represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). This range spans from approximately 10 to 200. This indicates that the middle 50% of the " pfit " values lie within this range.
- Whiskers and Outliers: The "whiskers" extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles. The lower whisker extends to 0, while the upper whisker extends to approximately 1000, suggesting a wider range for the upper values.
- Skewness: The distribution of "pfit" appears to be right-skewed due to the presence of higher values and the outliers extending significantly beyond the upper whisker.

vflt:

- Shape of the Distribution: The histogram is right-skewed (positively skewed). Most data points are concentrated on the left side, close to zero.
- Center: The peak of the histogram is near zero, indicating that the majority of vfit values are very low.
- Spread: The histogram shows that while most vfit values are close to zero. However, the frequency of these higher values is very low.
- Outliers: There are no distinct bars far separated from the rest, indicating that there are no significant outliers, but the long tail suggests occasional higher values
- Central Tendency: Median: The line inside the box represents the median of the " vfit " values. The median is around 100, indicating that half of the data points are below this value and half are above.
- Interquartile Range (IQR): The box represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). This range spans from approximately 10 to 220. This indicates that the middle 50% of the " vfit " values lie within this range.
- Whiskers and Outliers: The "whiskers" extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles. The lower whisker extends to 0, while the upper whisker extends to approximately 1350, suggesting a wider range for the upper values.
- Skewness: The distribution of "vfit" appears to be right-skewed due to the presence of higher values and the outliers extending significantly beyond the upper whisker.

freemem:

- Shape of the Distribution: The histogram is heavily right-skewed (positively skewed). Most of the data points are concentrated on the left side, close to zero.
- Center: The peak of the histogram is at or near zero, indicating that the majority of freemem values are very low, with a high density at zero.
- Spread: The histogram shows that while most freemem values are close to zero, there are some instances where freemem values range up to around 7000. However, the frequency of these higher values is very low.
- Outliers: There are no distinct bars far separated from the rest, indicating that there are no significant outliers in the data, but the long tail suggests occasional higher values.
- Central Tendency: The line inside the box represents the median of the "freemem" values. The median is around 100, indicating that half of the data points are below this value and half are above.
- Interquartile Range (IQR): The box represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). This range spans from approximately 10 to 2000. This indicates that the middle 50% of the "freemem" values lie within this range.
- Whiskers and Outliers: The "whiskers" extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles. The lower whisker extends to 0, while the upper whisker extends to approximately 12000, suggesting a wider range for the upper values.
- Skewness: The distribution of "freemem" appears to be right-skewed due to the presence of higher values and the outliers extending significantly beyond the upper whisker.

freeswap:

- Shape of the Distribution: The distribution of freeswap is multimodal, showing multiple peaks. This suggests the presence of several clusters within the data.
- Modes: There are distinct peaks around 0, 1 million, and 2 million. These modes indicate that freeswap values tend to cluster around these points.
- Spread: The spread is quite wide, ranging from 0 to approximately 2.5 million. This indicates a significant variation in the freeswap values.
- Density: The density plot shows that the highest density is around 1 million, indicating that this is where freeswap values are most concentrated. There is also a notable density peak around 2 million, indicating another significant cluster of freeswap values.
- Central Tendency: The median value (the line inside the box) is around 1.5 million, indicating that half of the freeswap values are below this point and half are above.
- Interquartile Range (IQR): The box represents the interquartile range (IQR), which contains the middle 50% of the data. For freeswap, the IQR extends from about 1 million to 2 million. This indicates that the majority of the freeswap values are concentrated between 1 million and 2 million.
- Whiskers: The whiskers extend from the quartiles to the minimum and maximum values within 1.5 times the IQR. The upper whisker extends slightly above 2 million, and the lower whisker extends to just above 0.5 million, indicating that most freeswap values fall within this range.
- Outliers: There is one clear outlier below 0.5 million, indicating that there are instances where the freeswap value is significantly lower than the rest of the data. Outliers are data points that fall outside the whiskers and can indicate unusual conditions or rare events.

usr:

- Distribution Shape: The distribution of usr shows a clear right-skewed pattern, with a concentration of values towards the higher end of the range. There is a significant peak around 80-100, indicating that a large number of observations fall within this range.
- Modes: There are two distinct peaks in the distribution. The first peak is around 0, and the second, more prominent peak is around 80-100. The peak around 0 suggests a subset of observations where usr is very low, possibly indicating periods of low activity or idleness.
- Density: The density reaches its highest point between 80 and 100, suggesting that this range is where the majority of usr values lie. The presence of the peak at 0 suggests a bimodal distribution, where the system alternates between low and high usage states.
- Range and Spread: The usr values span from 0 to slightly over 100, indicating a wide range of usage values. The right tail of the distribution is long, indicating a few observations with exceptionally high usr values.
- Median (Q2): The line inside the box represents the median (Q2) of the "usr" values. The median is approximately 90, indicating that half of the data points are below this value and half are above.
- Interquartile Range (IQR): The box represents the interquartile range, which is the range between the first quartile (Q1) and the third quartile (Q3). This range spans from approximately 80 to 95. Q1 (lower quartile) is around 80. Q3 (upper quartile) is around 95. This indicates that the middle 50% of the "usr" values lie within this range.
- Whiskers: The whiskers extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles. The lower whisker extends to around 60, indicating the smallest non-outlier value. The upper whisker extends to around 100, indicating the largest non-outlier value. The whiskers encompass most of the data points, but there are several outliers beyond the whiskers.
- Outliers: Outliers are represented by points beyond the whiskers. There are several outliers below the lower whisker, with values extending down to around 0. These outliers indicate the presence of some exceptionally low values in the dataset and suggest a left-skewed distribution for these outlier values.

runqsz:

- Comparison of Categories:
  - Not_CPU_Bound: The bar representing the "Not_CPU_Bound" category has a height of approximately 4500, indicating that the "runqsz" value for this category is around 4500.
  - CPU_Bound: The bar representing the "CPU_Bound" category has a height of approximately 4000, indicating that the "runqsz" value for this category is around 4000.
- Relative Difference: The "Not_CPU_Bound" category has a higher "runqsz" value compared to the "CPU_Bound" category. There is a difference of roughly 500 units between the two categories. This suggests that the "Not_CPU_Bound" category has a larger queue size (runqsz) compared to the "CPU_Bound" category.

Positive Correlations:

- Strong Positive Relationships: Some pairs of variables show strong positive linear relationships, indicating that as one variable increases, the other also increases. For example:
  - swrite and sread
  - fork and exec
  - pgfree and pgscan
  - pflt and vflt
- These relationships are indicated by a clear upward trend in the scatterplots.

Moderate Positive Relationships:

- Other pairs show moderate positive relationships, suggesting a less pronounced but still positive association. For example:
  - scall and sread
  - pgin and pgpin

Negative Correlations:

- Notable Negative Relationships: Some scatterplots indicate negative relationships, where an increase in one variable corresponds to a decrease in another. For example:
  - freemem and usr
- These relationships are indicated by a downward trend in the scatterplots.

Outliers and Variability

- Presence of Outliers: Several scatterplots reveal the presence of outliers, which can significantly affect the analysis. For example, variables like lread, pgfree, and pgscan show points that deviate markedly from the general trend.
- Variability: The scatterplots exhibit varying degrees of spread. Some relationships are tightly clustered, indicating consistent patterns, while others show more variability.

Specific Observations

- swrite and sread: There is a strong positive linear relationship, indicating that higher swrite values are strongly associated with higher sread values.
- fork and exec: A clear positive linear relationship is observed, suggesting that these variables typically increase together.
- pgfree and pgscan: The scatterplot shows a strong positive relationship, indicating that higher pgfree values are associated with higher pgscan values.
- pflt and vflt: There is a very strong positive relationship, suggesting that these two variables are highly correlated.

Predictive Modeling:

- The strong positive relationships can be leveraged for predictive modeling. For instance, knowing the value of swrite can help predict sread.

Resource Management:

- Understanding these relationships can aid in resource management and optimization. For example, the relationship between freemem and usr can help in performance tuning.

Further Analysis:

- The presence of outliers and skewed distributions suggests the need for further analysis, including potential data transformations and outlier treatment.

Summary

- Strong Positive Relationships: swrite with sread, fork with exec, pgfree with pgscan, pflt with vflt.
- Moderate Positive Relationships: scall with sread, pgin with pgpin.
- Negative Relationships: freemem with usr.
- Outliers: Multiple scatterplots indicate the presence of outliers, particularly in variables like lread, pgfree, and pgscan.
- Distributions: The distributions of variables vary, with some showing skewness.

PROBLEM 1.2 DATA PREPROCESSING:

Problem 1.2.1 Missing value check and treatment:

There are Few variables that has data missing.

```
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
dtype: int64
```

Missing Value Treatment by Imputing missing values with the mean of the column.

Problem 1.2.2 Outlier Treatment:

We Checked Outlier and following observations are made:

1. Outliers were identified in lread', lwrite', sread', swrite', fork', exec', rchar', wchar', pgout', ppgout', pgfree', pgscan', atch', pgin', ppgin', pflt', vflt', freemem', freeswap' and usr' columns, indicating the presence of values that are significantly higher or lower than the rest of the data points.

2. Given the high number of outliers in the dataset, it is recommended to treat outliers before proceeding with Mean Value.

Problem 1.2.3 Feature Engineering:

We are going to create new features using the existing features as described below:

total_reads_writes = lread + lwrite

total_char_transferred = rchar + wchar

total_page_faults = pflt + vflt

Problem 1.2.4 Encode the data:

Encoding of categorical data like 'runqsz' so that we can convert it to numerical data type. Runqsz feature has two data CPU_Bound and Not_CPU_Bound so we are converting

CPU_Bound values to '0' and Not_CPU_Bound values to '1'.

Problem 1.2.5 Train-test split:

We are going to split data into train set and test set using train_test_split function from scikit-learn. We are split data into 75 – 25 splits where 75% of entire data will be train dataset and 25% will be test dataset.

Shape values of train – test dataset are:

(X_train.shape, X_test.shape, y_train.shape, y_test.shape)

```
(6144, 24) (2048, 24) (6144, 1) (2048, 1)
```

PROBLEM 1.3 MODEL BUILDING - LINEAR REGRESSION:

Problem 1.3.1 Apply linear Regression using Sklearn:

We used the scikit-learn library to apply a Linear Regression model to the data.

The Linear Regression model was trained on the training data using the fit() method.

The trained model was then used to make predictions on the test data using the predict() method.

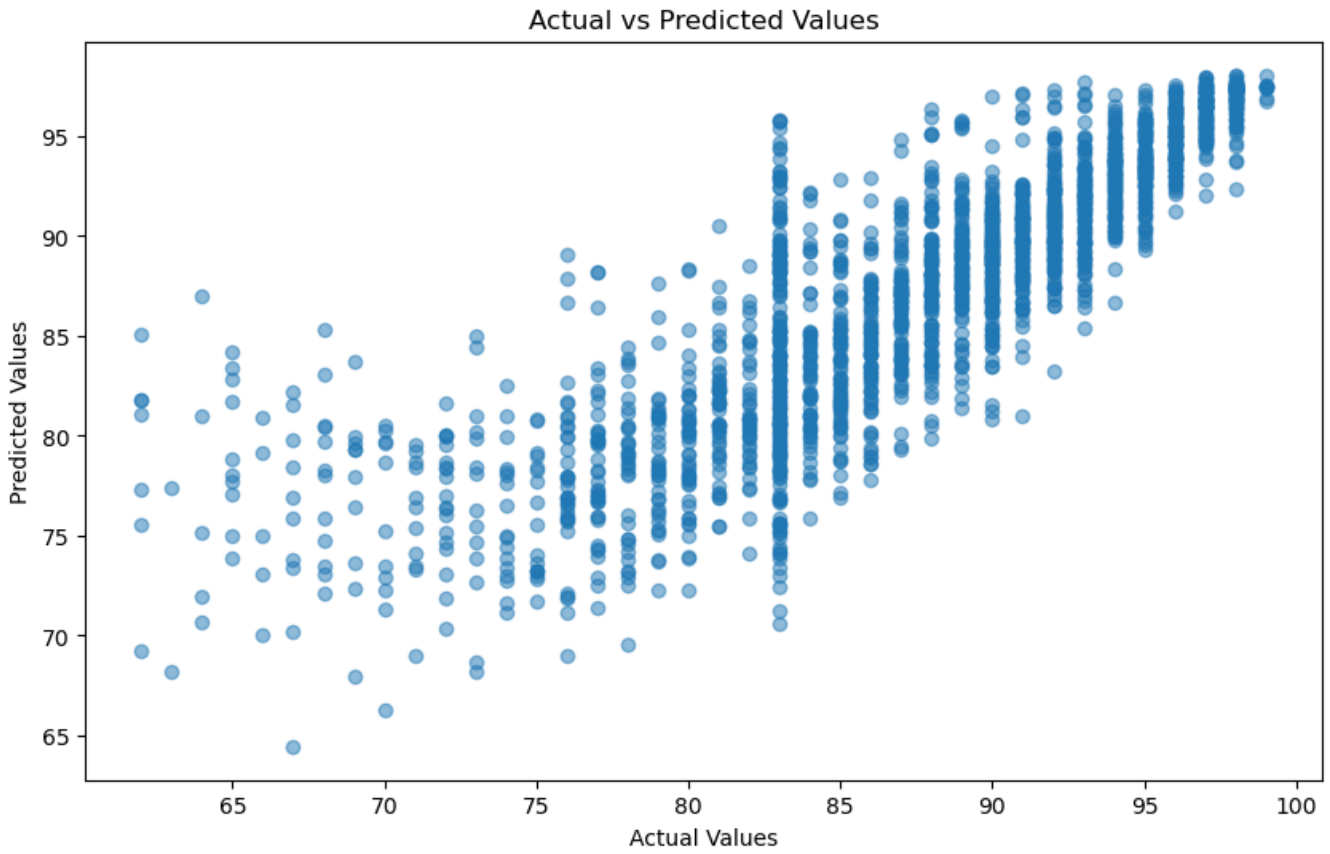
Results:

Mean Squared Error: 15.920332879634783

Root Mean Squared Error: 3.9900291828049057

R^2 Score: 0.7235863960696005

We draw a plot between actual and predicted values:



Insights

This scatter plot shows the relationship between actual values and predicted values from a regression model. Analyzing this plot can provide insights into the performance of the model. Here are some key points:

- Positive Correlation: The plot shows a generally positive correlation between actual and predicted values, indicating that as the actual values increase, the predicted values also tend to increase. This is a good sign that the model is capturing the general trend in the data.

Key Insights

- Accuracy and Linearity: A perfect prediction would result in all points lying on the 45-degree line where predicted values equal actual values. The scatter plot shows that while many points are close to this line, there is some spread, indicating prediction errors. The spread around the line suggests that the model has some degree of error, but the overall alignment with the 45-degree line suggests reasonable performance.
- Bias: The plot shows that for lower actual values (below 75), the model tends to over-predict, as predicted values are generally higher than actual values. For higher actual values (above 85), the

model generally under-predicts, as predicted values are lower than actual values. This indicates a potential bias in the model, where it tends to predict values closer to the mean rather than the extremes.

- Variance: The variance in predictions appears to be higher for mid-range actual values (around 80 to 85), as indicated by the wider spread of points. This suggests that the model is less consistent in this range.
- Outliers: There are a few points that deviate significantly from the general trend. These outliers could be due to various factors such as noise in the data, errors in the model, or special cases that the model fails to capture.

Implications

- Model Improvement: The model might benefit from further tuning to reduce bias and variance. Techniques such as regularization, ensembling, or feature engineering could help improve performance.
- Error Analysis: Conducting a more detailed error analysis to understand why certain predictions are off can provide insights into potential improvements. Identifying patterns in the residuals (differences between actual and predicted values) can help diagnose issues with the model.
- Validation: Further validation using different datasets or cross-validation techniques can provide a more comprehensive assessment of the model's performance.

Overall, the scatter plot provides valuable insights into the performance of the regression model, highlighting areas where the model performs well and areas where there is room for improvement.

Problem 1.3.2 Using Statsmodels Perform checks for significant variables using the appropriate method:

To check for significant variables using Statsmodels, we can perform a t-test or an F-test.

First model OLS Results,

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.747
Model:                            OLS   Adj. R-squared:                  0.746
Method:                 Least Squares   F-statistic:                     786.3
Date:                Thu, 20 Jun 2024   Prob (F-statistic):               0.00
Time:                        17:34:53   Log-Likelihood:                -17255.
No. Observations:                6144   AIC:                         3.456e+04
Df Residuals:                    6120   BIC:                         3.472e+04
Df Model:                          23
Covariance Type:            nonrobust
==============================================================================
==========
                             coef    std err          t      P>|t|      [0.025
     0.975]
------------------------------------------------------------------------------
----------
const                     99.3047      0.380    261.478      0.000      98.560
    100.049
lread                     -0.0723      0.009     -8.303      0.000      -0.089
     -0.055
lwrite                     0.0650      0.011      6.117      0.000       0.044
      0.086
```

```
scall                       -0.0015   5.25e-05    -28.579      0.000      -0.002
     -0.001
sread                       -0.0055      0.001     -7.182      0.000      -0.007
     -0.004
swrite                      -0.0025      0.001     -2.047      0.041      -0.005
     -0.000
fork                         0.4460      0.078      5.727      0.000       0.293
      0.599
exec                        -0.3462      0.051     -6.812      0.000      -0.446
     -0.247
rchar                        2.2880      0.319      7.180      0.000       1.663
      2.913
wchar                        2.2880      0.319      7.180      0.000       1.663
      2.913
pgout                       -0.0593      0.076     -0.777      0.437      -0.209
      0.090
ppgout                       0.0003      0.067      0.004      0.997      -0.130
      0.131
pgfree                      -0.1635      0.044     -3.735      0.000      -0.249
     -0.078
pgscan                       0.0088      0.013      0.668      0.504      -0.017
      0.034
atch                         0.2260      0.156      1.450      0.147      -0.080
      0.532
pgin                        -0.0677      0.020     -3.405      0.001      -0.107
     -0.029
ppgin                       -0.0519      0.014     -3.840      0.000      -0.078
     -0.025
pflt                         0.5220      0.104      5.032      0.000       0.319
      0.725
vflt                         0.5413      0.104      5.210      0.000       0.338
      0.745
freemem                      0.0002   7.06e-05      2.869      0.004    6.41e-05
      0.000
freeswap                -3.046e-07   2.12e-07     -1.437      0.151    -7.2e-07
   1.11e-07
total_reads_writes          -0.0073      0.003     -2.411      0.016      -0.013
     -0.001
total_char_transferred      -2.2880      0.319     -7.180      0.000      -2.913
     -1.663
total_page_faults           -0.5503      0.104     -5.298      0.000      -0.754
     -0.347
runqsz_Not_CPU_Bound        -0.1173      0.113     -1.035      0.300      -0.339
      0.105
==============================================================================
Omnibus:                     1715.787   Durbin-Watson:                   1.972
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6061.320
Skew:                          -1.377   Prob(JB):                         0.00
Kurtosis:                       7.012   Cond. No.                     1.51e+20
==============================================================================
```

VIF Details,

```
VIF values:

const                                54.82632
lread                                    inf
lwrite                                   inf
scall                                2.40349
sread                                3.57516
swrite                               3.33773
fork                                 2.40396
exec                                 1.94643
rchar                    732471273866.87744
wchar                     98285731095.02081
pgout                                4.00538
ppgout                              10.03682
pgfree                              14.27698
pgscan                               4.76870
atch                                 1.51888
pgin                                 4.22526
ppgin                                4.14457
pflt                             28799.53712
vflt                             71110.09419
freemem                              1.68737
freeswap                             1.91978
total_reads_writes                       inf
total_char_transferred   1038054541286.27319
total_page_faults                175007.14901
runqsz_Not_CPU_Bound                 1.21674
```

A few predictor VIF > 5 therefore there is some multicolinearity in the data. So We remove those predictors one by one with multicolinarity due to which there is least impact on the adjusted R2.

After doing to trail test we found that feature “total_reads_writes” has no significant changes in R-Squared values, so we decide to remove it.

After Removing “total_reads_writes”

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.747
Model:                            OLS   Adj. R-squared:                  0.746
Method:                 Least Squares   F-statistic:                     786.3
Date:                Thu, 20 Jun 2024   Prob (F-statistic):               0.00
Time:                        17:34:54   Log-Likelihood:                -17255.
No. Observations:                6144   AIC:                         3.456e+04
Df Residuals:                    6120   BIC:                         3.472e+04
Df Model:                          23
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         99.3047      0.380    261.478      0.000      98.560     100.049
```

```
lread                        -0.0796      0.008    -10.543      0.000      -0.094
     -0.065
lwrite                        0.0577      0.013      4.448      0.000       0.032
      0.083
scall                        -0.0015   5.25e-05    -28.579      0.000      -0.002
     -0.001
sread                        -0.0055      0.001     -7.182      0.000      -0.007
     -0.004
swrite                       -0.0025      0.001     -2.047      0.041      -0.005
     -0.000
fork                          0.4460      0.078      5.727      0.000       0.293
      0.599
exec                         -0.3462      0.051     -6.812      0.000      -0.446
     -0.247
rchar                         2.2880      0.319      7.180      0.000       1.663
      2.913
wchar                         2.2880      0.319      7.180      0.000       1.663
      2.913
pgout                        -0.0593      0.076     -0.777      0.437      -0.209
      0.090
ppgout                        0.0003      0.067      0.004      0.997      -0.130
      0.131
pgfree                       -0.1635      0.044     -3.735      0.000      -0.249
     -0.078
pgscan                        0.0088      0.013      0.668      0.504      -0.017
      0.034
atch                          0.2260      0.156      1.450      0.147      -0.080
      0.532
pgin                         -0.0677      0.020     -3.405      0.001      -0.107
     -0.029
ppgin                        -0.0519      0.014     -3.840      0.000      -0.078
     -0.025
pflt                          0.5220      0.104      5.032      0.000       0.319
      0.725
vflt                          0.5413      0.104      5.210      0.000       0.338
      0.745
freemem                       0.0002   7.06e-05      2.869      0.004    6.41e-05
      0.000
freeswap                -3.046e-07   2.12e-07     -1.437      0.151    -7.2e-07
   1.11e-07
total_char_transferred       -2.2880      0.319     -7.180      0.000      -2.913
     -1.663
total_page_faults            -0.5503      0.104     -5.298      0.000      -0.754
     -0.347
runqsz_Not_CPU_Bound         -0.1173      0.113     -1.035      0.300      -0.339
      0.105
==============================================================================
Omnibus:                     1715.787   Durbin-Watson:                   1.972
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6061.320
Skew:                          -1.377   Prob(JB):                         0.00
Kurtosis:                       7.012   Cond. No.                     1.55e+07
==============================================================================
```

```
VIF values:

const                                54.82632
lread                                 2.62789
lwrite                                2.27383
scall                                 2.40349
sread                                 3.57516
swrite                                3.33773
fork                                  2.40396
exec                                  1.94643
rchar                    732471273866.87744
wchar                     98285731095.02081
pgout                                 4.00538
ppgout                               10.03682
pgfree                               14.27698
pgscan                                4.76870
atch                                  1.51888
pgin                                  4.22526
ppgin                                 4.14457
pflt                              28799.53712
vflt                              71110.09419
freemem                               1.68737
freeswap                              1.91978
total_char_transferred   1038054541286.27319
total_page_faults                175007.14901
runqsz_Not_CPU_Bound                  1.21674
```

As We can see that after removal of feature “total_read_write” few features VIF values has changed.

Similarily we started doing Trail for other feature have VIF values > 5 to reduce multicollinearity.

After Removing above said feature

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.733
Model:                            OLS   Adj. R-squared:                  0.733
Method:                 Least Squares   F-statistic:                     1053.
Date:                Thu, 20 Jun 2024   Prob (F-statistic):               0.00
Time:                        17:35:01   Log-Likelihood:                -17418.
No. Observations:                6144   AIC:                         3.487e+04
Df Residuals:                    6127   BIC:                         3.498e+04
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
=========
                             coef    std err          t      P>|t|      [0.025
   0.975]
------------------------------------------------------------------------------
---------
const                     98.4700      0.385    255.540      0.000      97.715
   99.225
lread                     -0.0813      0.008    -10.500      0.000      -0.096
   -0.066
lwrite                     0.0505      0.013      3.797      0.000       0.024
    0.077
```

```
scall                   -0.0016   5.22e-05    -30.960      0.000      -0.002
   -0.002
swrite                  -0.0090      0.001     -8.857      0.000      -0.011
   -0.007
fork                     0.2900      0.078      3.727      0.000       0.137
    0.443
exec                    -0.4590      0.052     -8.887      0.000      -0.560
   -0.358
rchar                -7.569e-06   4.64e-07    -16.303      0.000   -8.48e-06
-6.66e-06
wchar                 3.176e-07   1.22e-06      0.261      0.794   -2.07e-06
  2.7e-06
pgout                   -0.2441      0.055     -4.454      0.000      -0.352
   -0.137
pgscan                  -0.0467      0.008     -5.654      0.000      -0.063
   -0.031
atch                     0.0462      0.157      0.295      0.768      -0.261
    0.353
pgin                    -0.1546      0.012    -13.045      0.000      -0.178
   -0.131
pflt                    -0.0362      0.001    -36.242      0.000      -0.038
   -0.034
freemem                  0.0002   7.19e-05      3.128      0.002     8.4e-05
    0.000
freeswap              7.728e-09   2.16e-07      0.036      0.971   -4.15e-07
  4.3e-07
runqsz_Not_CPU_Bound    -0.0658      0.116     -0.569      0.570      -0.293
    0.161
==============================================================================
Omnibus:                     1632.235   Durbin-Watson:                   1.991
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5511.921
Skew:                          -1.323   Prob(JB):                         0.00
Kurtosis:                       6.811   Cond. No.                     1.06e+07
==============================================================================


VIF values:

const                   53.58174
lread                    2.62091
lwrite                   2.26531
scall                    2.25371
swrite                   2.22694
fork                     2.27876
exec                     1.90857
rchar                    1.47575
wchar                    1.35738
pgout                    1.96523
pgscan                   1.79842
atch                     1.45643
pgin                     1.42350
pflt                     2.52990
freemem                  1.66375
freeswap                 1.88646
runqsz_Not_CPU_Bound     1.20613
```

Now We have to remove the non-significant predictor variables (p-Value > 0.05) one by one.

Final OLs results are:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.733
Model:                            OLS   Adj. R-squared:                  0.733
Method:                 Least Squares   F-statistic:                     1405.
Date:                Thu, 20 Jun 2024   Prob (F-statistic):               0.00
Time:                        17:35:56   Log-Likelihood:                -17418.
No. Observations:                6144   AIC:                         3.486e+04
Df Residuals:                    6131   BIC:                         3.495e+04
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         98.4314      0.148    667.213      0.000      98.142      98.721
lread         -0.0810      0.008    -10.527      0.000      -0.096      -0.066
lwrite         0.0506      0.013      3.829      0.000       0.025       0.077
scall         -0.0016   5.04e-05    -31.992      0.000      -0.002      -0.002
swrite        -0.0089      0.001     -9.109      0.000      -0.011      -0.007
fork           0.2911      0.078      3.750      0.000       0.139       0.443
exec          -0.4588      0.052     -8.908      0.000      -0.560      -0.358
rchar      -7.479e-06   4.42e-07    -16.927      0.000   -8.35e-06   -6.61e-06
pgout         -0.2395      0.052     -4.564      0.000      -0.342      -0.137
pgscan        -0.0466      0.008     -5.662      0.000      -0.063      -0.030
pgin          -0.1546      0.012    -13.441      0.000      -0.177      -0.132
pflt          -0.0362      0.001    -36.389      0.000      -0.038      -0.034
freemem        0.0002   6.62e-05      3.448      0.001    9.85e-05       0.000
==============================================================================
Omnibus:                     1631.511   Durbin-Watson:                   1.991
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5513.469
Skew:                          -1.323   Prob(JB):                         0.00
Kurtosis:                       6.813   Cond. No.                     5.86e+05
==============================================================================
```

After dropping the features causing strong multicollinearity and the statistically insignificant ones, our model performance hasn't dropped much . This shows that these variables did not have much predictive power.

Insight from above summary

R-squared:

- The R-squared value remains high at 0.733, indicating that the model still explains a significant portion of the variation in the dependent variable.

F-statistic:

- The F-statistic remains high at 1405, indicating that the model is statistically significant.

P-values:

- All p-values are below the 0.05 significance level, indicating that all variables in the model are statistically significant.

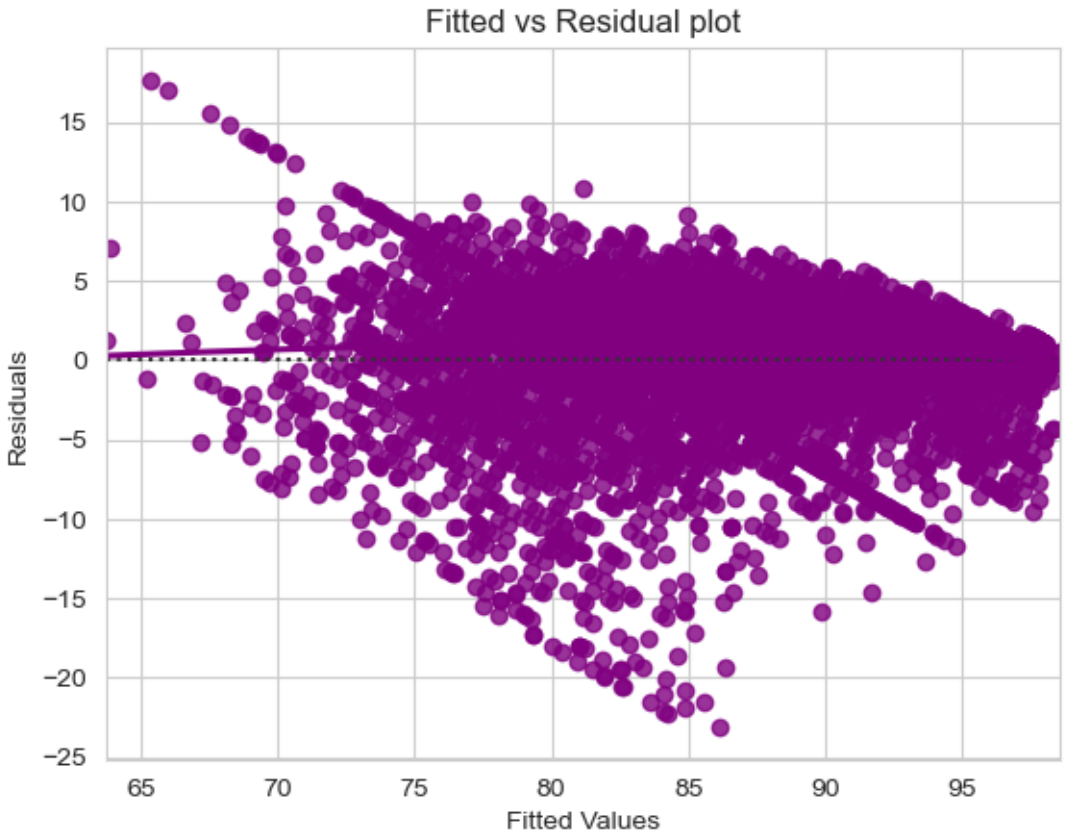
Now We have to check test Assumption for Linear regression model.

Testing the Assumptions of Linear Regression

- For Linear Regression, we need to check if the following assumptions hold:-
  - Linearity
  - Independence
  - Homoscedasticity
  - Normality of error terms
  - No strong Multicollinearity

Linearity and Independence of predictors

To check this assumption, we have to plot a chart between Fitted and residuals



No pattern in the data thus the assumption of linearity and independence of predictors satisfied

Test for Normality

We check the p-Values, so we used shapiro test to find the p-Values.

```
ShapiroResult(statistic=0.9090408086776733, pvalue=0.0)
```

Since p-value < 0.05, the residuals are not normal as per shapiro test.

Test for Homoscedasticity

To check this we have to do Goldfeld-Quandt test and check p-Value.

```
p-Values = 0.7969925454967073
```

Since p-value > 0.05 we can say that the residuals are homoscedastic.

Conclusion:

The model built olsres_26 satisfies all assumptions of Linear Regression

Final Model OLS Reslts:

OLS Regression Results

Dep. Variable:	usr	**R-squared:**	0.733
Model:	OLS	**Adj. R-squared:**	0.733
Method:	Least Squares	**F-statistic:**	1405.
Date:	Thu, 20 Jun 2024	**Prob (F-statistic):**	0.00
Time:	18:16:43	**Log-Likelihood:**	-17418.
No. Observations:	6144	**AIC:**	3.486e+04
Df Residuals:	6131	**BIC:**	3.495e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	98.4314	0.148	667.213	0.000	98.142	98.721
lread	-0.0810	0.008	-10.527	0.000	-0.096	-0.066
lwrite	0.0506	0.013	3.829	0.000	0.025	0.077
scall	-0.0016	5.04e-05	-31.992	0.000	-0.002	-0.002

swrite	-0.0089	0.001	-9.109	0.000	-0.011	-0.007
fork	0.2911	0.078	3.750	0.000	0.139	0.443
exec	-0.4588	0.052	-8.908	0.000	-0.560	-0.358
rchar	-7.479e-06	4.42e-07	-16.927	0.000	-8.35e-06	-6.61e-06
pgout	-0.2395	0.052	-4.564	0.000	-0.342	-0.137
pgscan	-0.0466	0.008	-5.662	0.000	-0.063	-0.030
pgin	-0.1546	0.012	-13.441	0.000	-0.177	-0.132
pflt	-0.0362	0.001	-36.389	0.000	-0.038	-0.034
freemem	0.0002	6.62e-05	3.448	0.001	9.85e-05	0.000

Omnibus:	1631.511	**Durbin-Watson:**	1.991
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	5513.469
Skew:	-1.323	**Prob(JB):**	0.00
Kurtosis:	6.813	**Cond. No.**	5.86e+05

Problem 1.3.3 Creating Multiple Models and Evaluate:

We Created Ridge model and lasso model too and checked there “Mean Squared Error”, “Root Mean Squared Error” & “R^2 Score”.

Results:

```
Ridge Model - Mean Squared Error: 15.921242830946197
Ridge Model - Root Mean Squared Error: 3.9901432093279805
Ridge Model - R^2 Score: 0.7235705972214677
Lasso Model - Mean Squared Error: 16.163813082274423
Lasso Model - Root Mean Squared Error: 4.0204244903087565
Lasso Model - R^2 Score: 0.7193590196192363
```

We Check the performance of models using Rsquare, RMSE & Adj Rsquare values, OLS Model seems to best fit as it has **high R^2 Score value, Low Mean Squared Error & Low Root Mean Squared Error value**.

PROBLEM 1.4 BUSINESS INSIGHTS & RECOMMENDATIONS:

Problem 1.4.1 Comment on the Linear Regression equation from the final model and impact of relevant variables (atleast 2) as per the equation:

Equation as per Final model:

```
usr = 98.43144987636954 + -0.08102029433914065 * ( lread ) +  0.050602638437615
366 * ( lwrite ) +  -0.001612059404412203 * ( scall ) +  -0.00887608566641272 *
 ( swrite ) +  0.291132984369165 * ( fork ) +  -0.4588275777138781 * ( exec ) +
  -7.479133181490279e-06 * ( rchar ) +  -0.23947322253190773 * ( pgout ) +  -0.
046576205725542524 * ( pgscan ) +  -0.15457470899017925 * ( pgin ) +  -0.036200
43358461722 * ( pflt ) +  0.0002283688923334947 * ( freemem )
```

Insights

- lread: Negative coefficient (-0.08102029433914065) suggests that an increase in lread (reads between system and user memory) will decrease the usr value.
- lwrite: Positive coefficient (0.050602638437615366) suggests that an increase in lwrite (writes between system and user memory) will increase the usr value.
- scall: Negative coefficient (-0.001612059404412203) suggests that an increase in scall (system calls) will decrease the usr value.
- swrite: Negative coefficient (-0.00887608566641272) suggests that an increase in swrite (system write calls) will decrease the usr value.
- fork: Positive coefficient (0.291132984369165) suggests that an increase in fork (system fork calls) will increase the usr value.
- exec: Negative coefficient (-0.4588275777138781) suggests that an increase in exec (system exec calls) will decrease the usr value.
- rchar: Negative coefficient (-7.479133181490279e-06) suggests that an increase in rchar (characters transferred by system read calls) will decrease the usr value.
- pgout: Negative coefficient (-0.23947322253190773) suggests that an increase in pgout (page out requests) will decrease the usr value.
- pgscan: Negative coefficient (-0.046576205725542524) suggests that an increase in pgscan (pages checked for freeing) will decrease the usr value.
- pgin: Negative coefficient (-0.15457470899017925) suggests that an increase in pgin (page-in requests) will decrease the usr value.
- pflt: Negative coefficient (-0.03620043358461722) suggests that an increase in pflt (page faults caused by protection errors) will decrease the usr value.
- freemem: Positive coefficient (0.0002283688923334947) suggests that an increase in freemem (free memory pages) will increase the usr value.

Problem 1.4.2 Conclude with the key takeaways (actionable insights and recommendations) for the business:

Business Insights:

Optimization of Memory Read/Write Operations:

- Decrease in lread is associated with an increase in usr. This suggests that reducing the number of reads between system memory and user memory can free up CPU time for user processes.
- Increase in lwrite positively affects usr, indicating that optimizing write operations might improve CPU efficiency in user mode.

System Calls and Fork Operations:

- A decrease in system calls (scall) is associated with an increase in usr. Streamlining and reducing unnecessary system calls can improve CPU efficiency.
- Increasing fork operations positively impacts usr. Optimizing processes that require forking might enhance CPU performance in user mode.

Page Faults and Memory Management:

Decrease in pgout and pgin is associated with better CPU performance in user mode. Efficient memory management that reduces paging activities can lead to better CPU utilization.

Increase in available memory (freemem) has a positive, though minor, impact on usr. Ensuring sufficient free memory can help maintain better CPU performance in user mode.

System Exec Calls:

A decrease in exec calls correlates with better CPU performance in user mode. Optimizing the execution of programs and reducing the frequency of exec calls can improve CPU efficiency.

Recommendations:

Memory Operation Optimization:

- Reduce Read Operations: Implement caching mechanisms or optimize memory access patterns to reduce the frequency of read operations.
- Optimize Write Operations: Ensure that write operations are efficient and minimize delays, possibly through improved buffering or batching strategies.

System Call Reduction:

- Streamline System Calls: Audit and optimize system calls to remove unnecessary operations. This can involve code reviews and optimizing system-level interactions.
- Optimize Fork Operations: Ensure that processes that require forking are optimized, and use efficient process management techniques.

Efficient Memory Management:

- Reduce Paging: Implement better memory management strategies to reduce paging activities. This can involve optimizing application memory usage and ensuring adequate physical memory is available.
- Increase Free Memory: Ensure that there is sufficient free memory to avoid excessive paging and to maintain efficient CPU performance.

Execution Optimization:

- Minimize Exec Calls: Reduce the frequency of exec calls by optimizing application workflows and ensuring efficient execution paths.
- By implementing these recommendations, the business can improve CPU efficiency in user mode, leading to better overall system performance and potentially reduced operational costs.

Problem 2:

In your role as a statistician at the Republic of Indonesia Ministry of Health, you have been entrusted with a dataset containing information from a Contraceptive Prevalence Survey. This dataset encompasses data from 1473 married females who were either not pregnant or were uncertain of their pregnancy status during the survey. Your task involves predicting whether these women opt for a contraceptive method of choice. This prediction will be based on a comprehensive analysis of their demographic and socio-economic attributes.

PROBLEM 2.1 DEFINE THE PROBLEM AND PERFORM EXPLORATORY DATA ANALYSIS:

Problem 2.1.1 Problem Definition:

The task is to Predict Contraceptive Method Choice among Married Women in Indonesia based on Demographic and Socio-economic Factors

Problem 2.1.2 Check shape, Data types, statistical summary:

First, we import all the necessary libraries seaborn, numpy, pandas, sklearn, matplotlib etc. to perform our analysis

Next, we import the data set “Contraceptive_method_dataset”

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

Shape of the data:

Shape of the dataset is 1473 rows and 10 Columns.

Data Type:

Data columns (total 61 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
```

There are 3 Numerical Data and 7 Categorical Data.

Statistical Summary:

```
: data.describe().T
```

:

	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

Key Insights from the Data

- The dataset has 1473 rows and 10 columns.
- The dataset has 7 categorical variable and 3 numerical variable.
- The dataset has 2 float type variable, 1 int type variable, 7 object type variable.
- Missing Values: The data has some missing values, particularly in the "Wife_age" and "No_of_children_born" columns, which have 1402 and 1452 non-null values, respectively. This may need to be addressed during data preprocessing.

Observations:

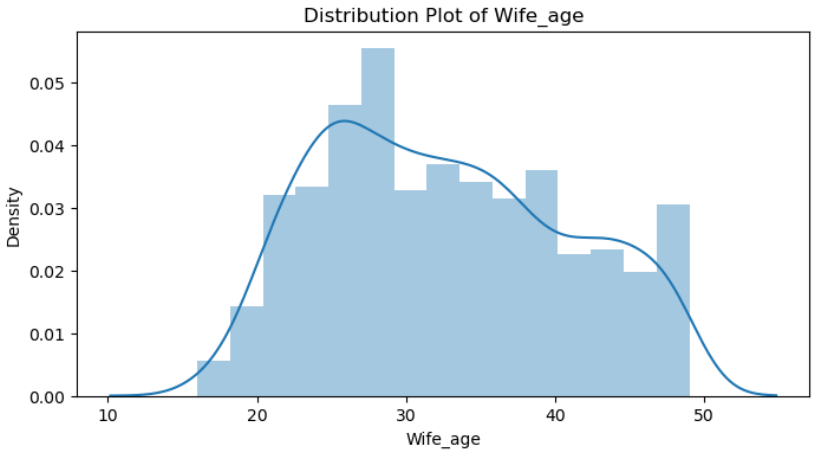
- Demographic Characteristics: The average age of the women in the dataset is 32.6 years, with a standard deviation of 8.3 years. The average number of children ever born is 3.25, with a standard deviation of 2.37. The husband's occupation is categorized into 4 levels, with the majority of husbands falling into the middle categories (2 and 3).
- Socio-economic Factors: The wife's and husband's education levels are categorical variables with 4 levels each, ranging from "uneducated" to "tertiary". The standard of living index is also a

categorical variable with 4 levels, ranging from "very low" to "high". The media exposure variable is a binary variable, indicating whether the wife has "Good" or "Not good" media exposure.

- Target Variable: The contraceptive method used by the women is the target variable, which is a categorical variable with two values: "No" and "Yes". The distribution of the target variable is not provided in the data summary, so it's unclear if the dataset is balanced or if there is a significant imbalance between the two classes.

Problem 2.1.3 Univariate analysis:

Numerical Data type:

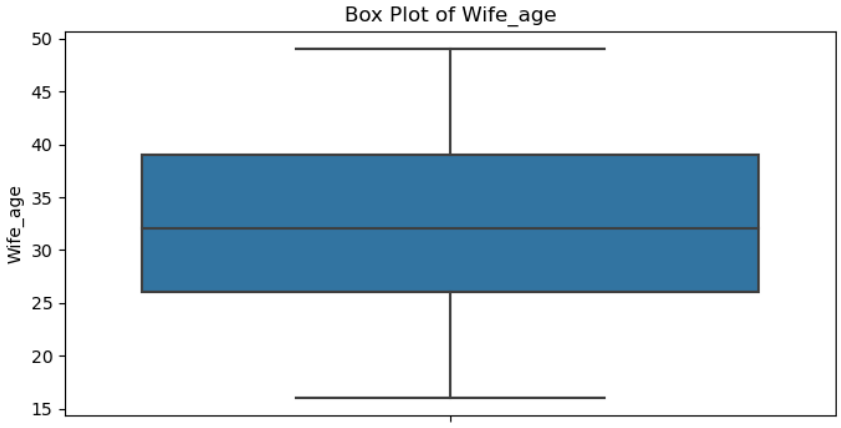


Shape of Distribution: The distribution appears to be approximately unimodal with a peak around the mid-20s. There are some fluctuations, but the overall shape suggests a moderately normal distribution with slight irregularities.

Central Tendency: The highest density is observed around the age of 25-30. This indicates that the most common ages in the dataset fall within this range.

Spread of Data: The ages range from about 15 to 50, indicating a wide spread of data points. The density drops off significantly beyond 50, suggesting fewer older individuals in the dataset.

Symmetry and Skewness: The distribution is roughly symmetric, but with slight skewness towards the right (older ages). This is indicated by the longer tail on the right side compared to the left.



Median: The median age (the line inside the box) is approximately 30. This means that half of the ages are below 30 and half are above 30.

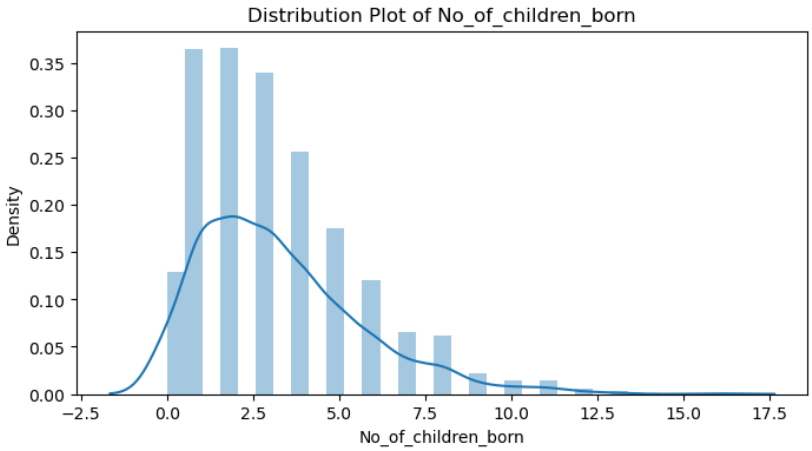
Interquartile Range (IQR):
The box represents the interquartile range (IQR), which is the range between the first quartile (Q1, 25th percentile) and the third quartile (Q3, 75th percentile).

Q1 is around 25, and Q3 is around 38. This means that the middle 50% of the ages fall between 25 and 38.

Range of Data:
The "whiskers" extend from the box to the minimum and maximum values within 1.5 times the IQR from the quartiles.

The minimum age is around 15, and the maximum age is around 50.

Symmetry:
. The box plot appears fairly symmetric, indicating that the ages are evenly distributed around the median.

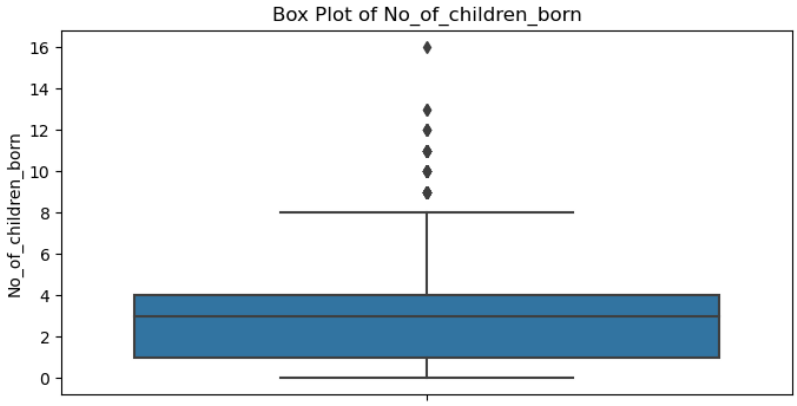


Right-Skewed Distribution: The distribution is right-skewed, indicating that most families have a smaller number of children, with fewer families having a larger number of children. This is typical in many populations where having a smaller number of children is more common.

Central Tendency:

The peak of the density plot (mode) is around 2 children, suggesting that this is the most common number of children born to families in the dataset.

The mean (average) number of children born is likely to be higher



than the mode due to the right skewness.

Spread and Variability: The number of children born ranges from 0 to approximately 17 There is a gradual decline in the density as the number of children increases, indicating that families with more than 3-4 children become progressively less common.

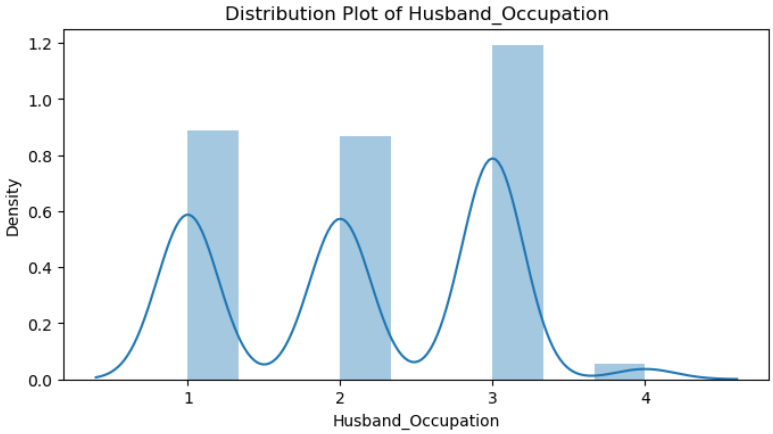
Outliers: There are a few extreme values on the right side of the distribution (e.g., families with more than 10 children), which are considered outliers. These outliers can affect the mean and standard deviation of the dataset.

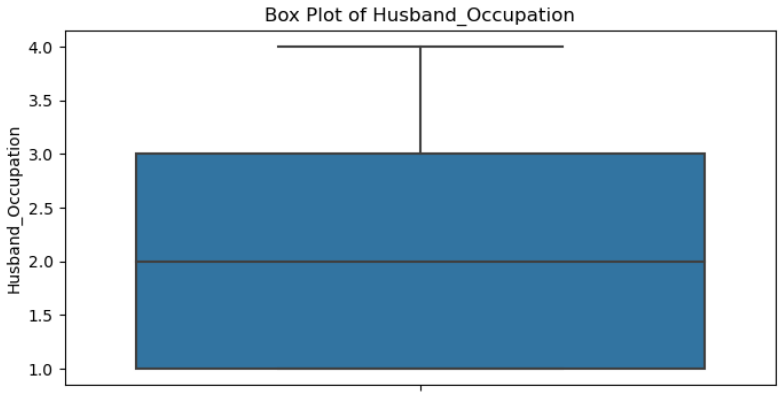
Central Tendency: The median number of children born (the line inside the box) is approximately 3. This indicates that 50% of the families have 3 or fewer children, and the other 50% have more than 3 children.

Interquartile Range (IQR): The box represents the interquartile range (IQR), which is the range between the first quartile (Q1, 25th percentile) and the third quartile (Q3, 75th percentile).

Whiskers and Range: The whiskers extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles.

Outliers: The points above the upper whisker are outliers, indicating families with a significantly higher number of children compared to the rest of the population.





Most Common Occupation (Mode): The highest density is observed at Husband_Occupation = 3, indicating that this is the most common occupation category among the data.

Other Common Occupations: The occupation categories 1 and 2 also have significant densities, making them the next most common categories after 3.

Least Common Occupation: The occupation category 4 has the lowest density, indicating that it is the least common among the data.

Median Occupation: The median value (the line inside the box) is 2. This means that half of the occupation categories are below 2 and half are above 2.

Interquartile Range (IQR): The box represents the interquartile range (IQR), which is the range between the first quartile (Q1, 25th percentile) and the third quartile (Q3, 75th percentile).

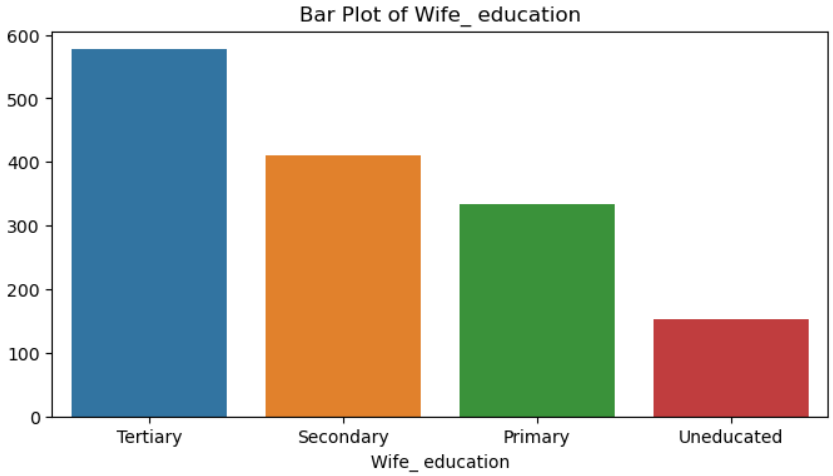
Q1 is 1.5, and Q3 is 3. This means that the middle 50% of the occupation categories fall between 1.5 and 3.

Range of Data:
The "whiskers" extend from the box to the minimum and maximum values within 1.5 times the IQR from the quartiles.

The minimum occupation category is 1, and the maximum category is 4.

Symmetry:
The box plot appears fairly symmetric, indicating that the occupation categories are evenly distributed around the median.

Categorical data:

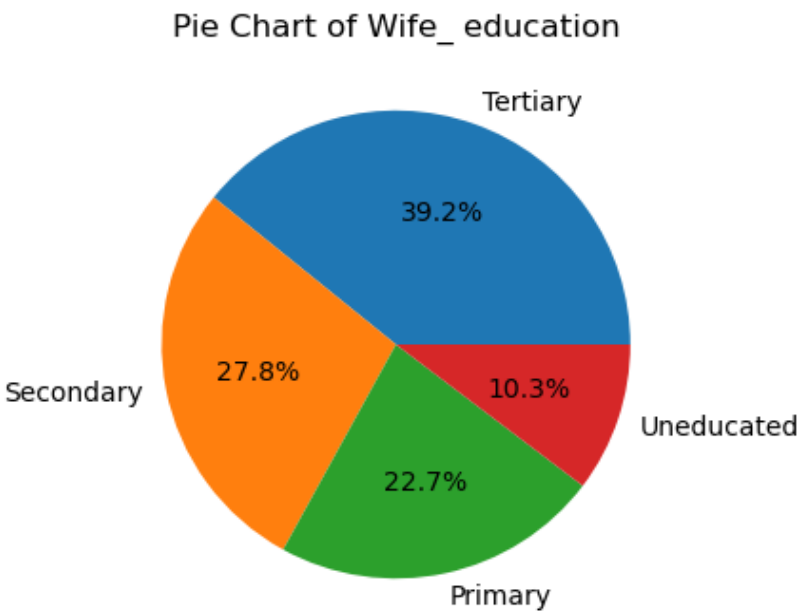


Tertiary Education: The highest number of wives have tertiary education, with around 550 individuals. This indicates that a significant portion of the population has pursued higher education.

Secondary Education: The second largest group is those with secondary education, with around 450 individuals. This shows that a substantial number of wives have completed secondary school.

Primary Education: The number of wives with primary education is around 300. This suggests that fewer wives have only primary education compared to those with secondary and tertiary education.

Uneducated: The smallest group is the uneducated category, with around 150 individuals. This indicates that a relatively small portion of the population has not received any formal education.

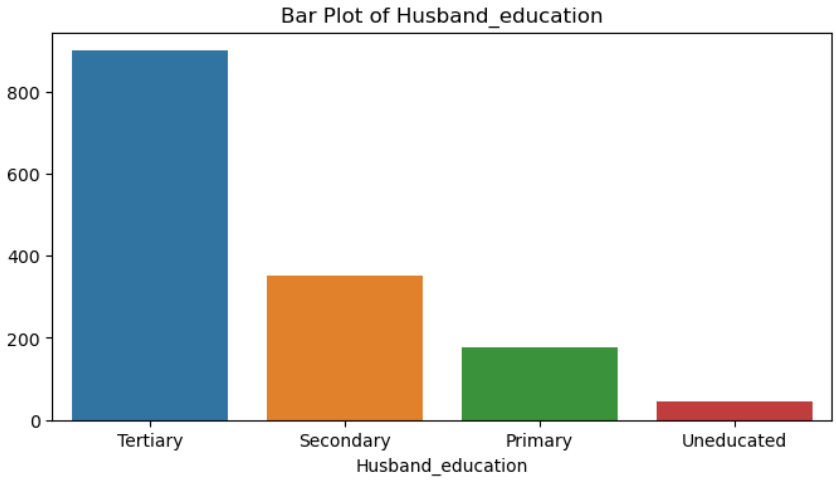


Tertiary Education (39.2%): The largest segment of the pie chart is composed of wives with tertiary education, indicating that a significant portion of the population has pursued higher education.

Secondary Education (27.8%): The second-largest segment represents wives with secondary education, showing that a substantial number of wives have completed secondary school.

Primary Education (22.7%): This segment indicates that a considerable portion of the population has primary education.

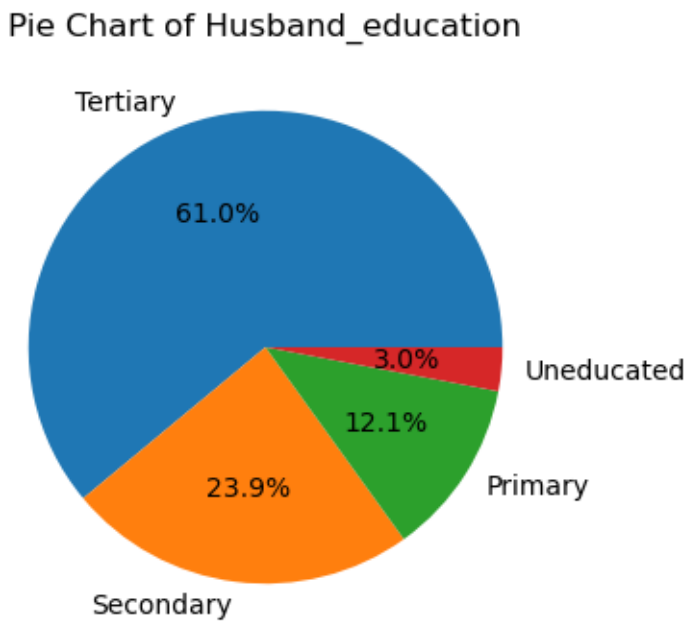
Uneducated (10.3%): The smallest segment represents uneducated wives, suggesting that a relatively small portion of the population has not received any formal education.



Tertiary Education: The highest number of husbands have tertiary education, with above 850 individuals. This indicates that a significant portion of the population has pursued higher education.

Secondary Education: The second largest group is those with secondary education, with around 450 individuals. This shows that a substantial number of husbands have completed secondary school.

Primary Education: The number of husbands with primary education is around 200. This



suggests that fewer husbands have only primary education compared to those with secondary and tertiary education.

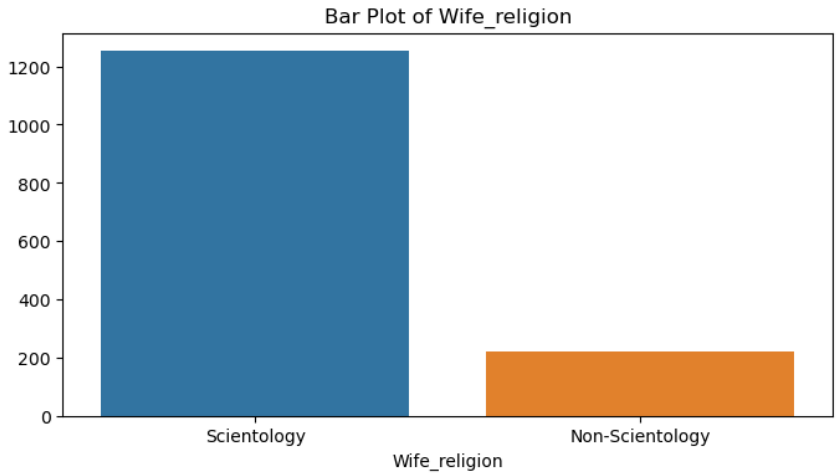
Uneducated: The smallest group is the uneducated category, with around 100 individuals. This indicates that a relatively small portion of the population has not received any formal education.

Tertiary Education (61.0%): The largest segment of the pie chart is composed of husbands with tertiary education, indicating that a significant portion of the population has pursued higher education.

Secondary Education (23.9%): The second-largest segment represents husbands with secondary education, showing that a substantial number of husbands have completed secondary school.

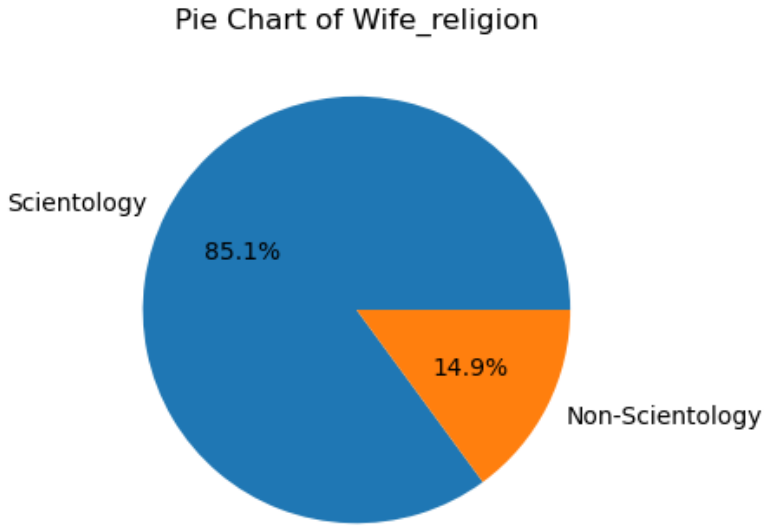
Primary Education (12.1%): This segment indicates that a considerable portion of the population has primary education.

Uneducated (3.0%): The smallest segment represents uneducated husbands, suggesting that a relatively small portion of the population has not received any formal education.



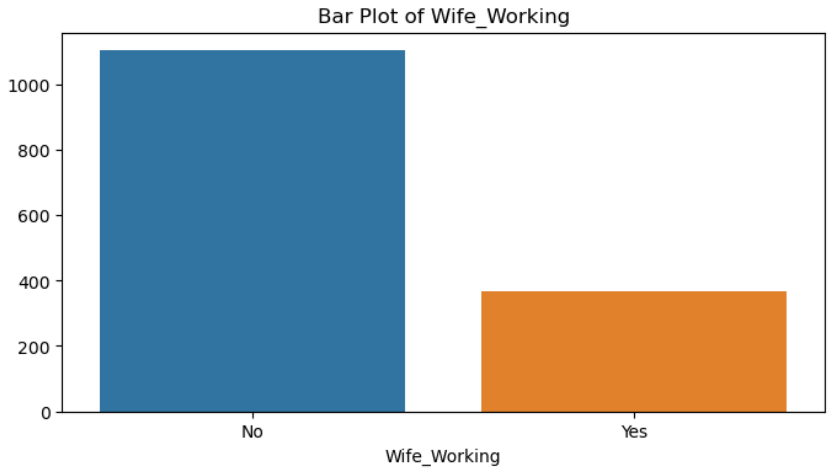
Scientology: The largest number of wives belong to this religion with above 1200 Individuals. It Indicates that large portion of population belong to Scientology religion.

Non-Scientology: Around 200 individuals belong to this religion, indicating that small portion of population belongs to this religion.



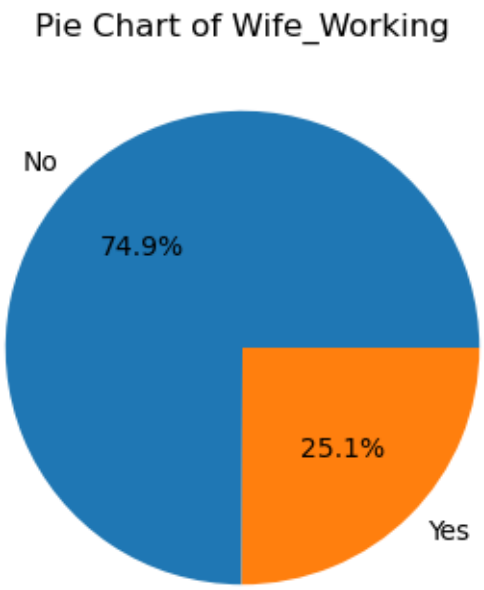
Scientology (85.1%): Pie chart indicates that 85% of population belongs to this religion. It show the Dominance of the religion, as majority of wives belong to this religion.

Non-Scientology (14.9%): Only 15% of population belongs to this religion. It has very least dominance over wives.



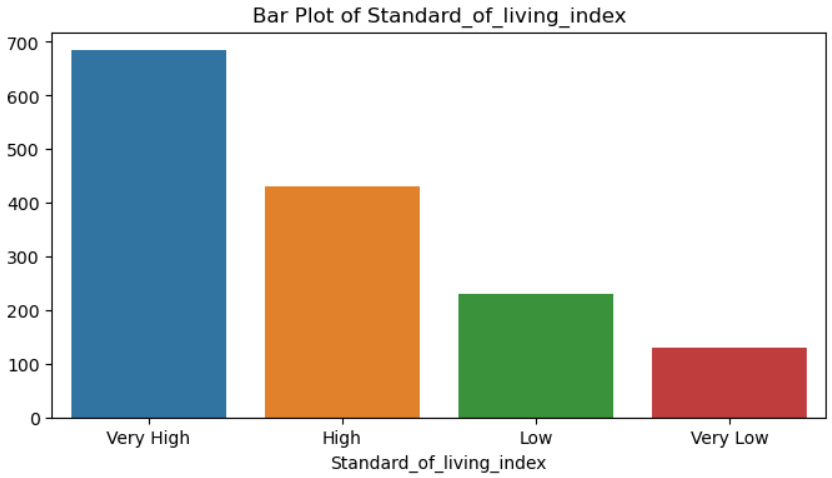
Working Wives (Yes): Around 400 wives are working out of the entire population, indicate that number of working wives are less as compare to Non-working wives.

Non-Working Wives (No): Majority of wives are non-working, with above 1000 individuals from entire population are not working.



Working Wives (25.1%): Only 25% of the entire population of wives are working, indicate that number of wives prefer to work is less.

Non-Working Wives (No): 75 % of entire population of wives are not working. Indicates that majority of wives prefer non-working.

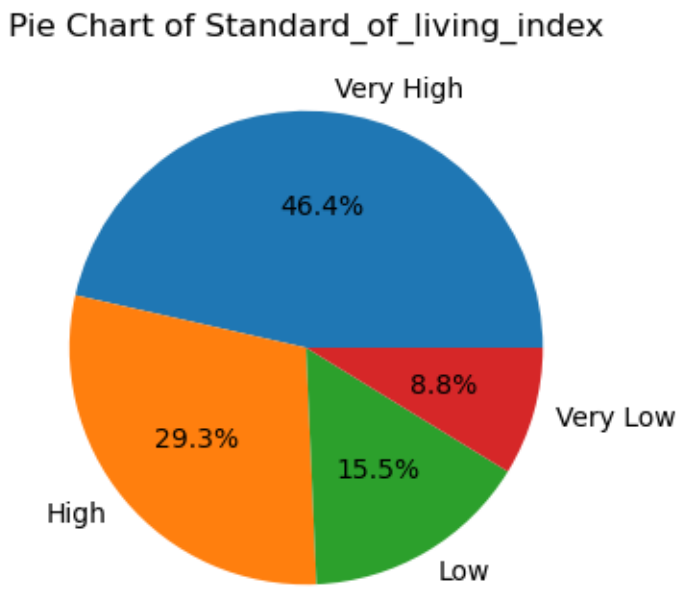


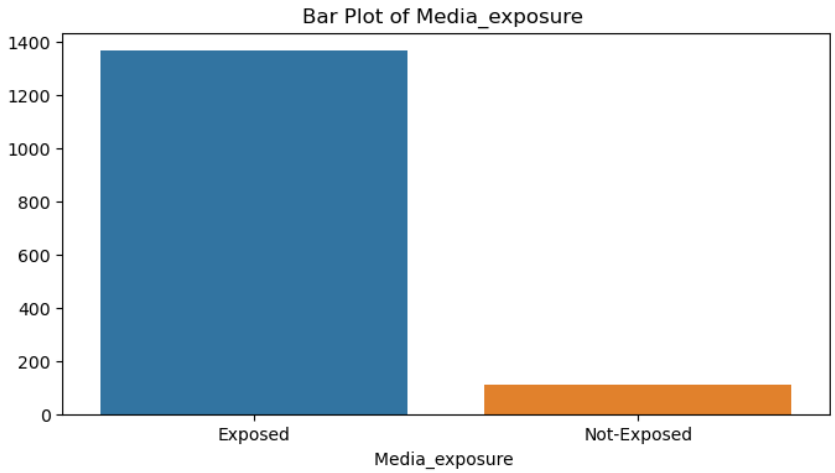
Very High: The largest group, with around 700 individuals, has a very high standard of living.

High: The second-largest group, with around 450 individuals, has a high standard of living.

Low: This group has around 200 individuals with a low standard of living.

Very Low: The smallest group, with around 100 individuals, has a very low standard of living.





Very High (46.4%): The largest group, which cover 46% of entire population and has a very high standard of living.

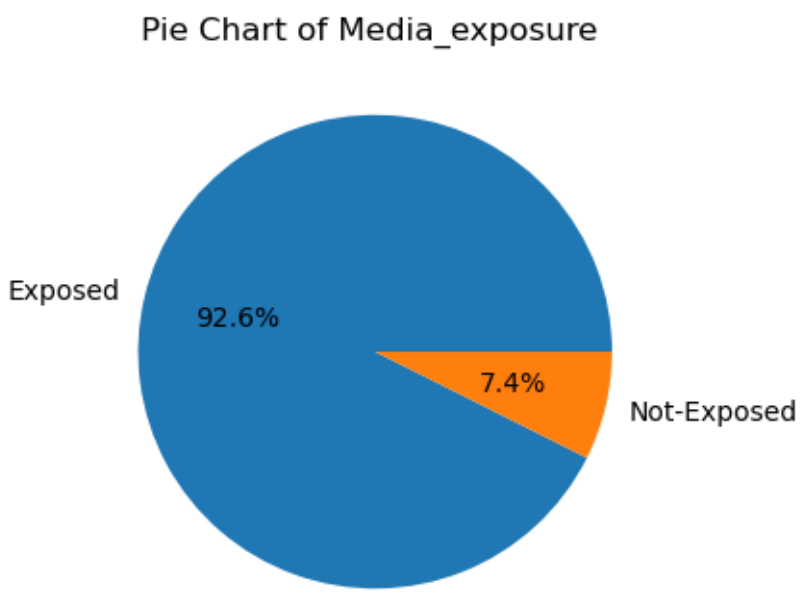
High (29.3%): The second-largest group, with 29% of entire population, has a high standard of living.

Low (15.5%): This group covers around 15% of entire population, which has a low standard of living.

Very Low (8.8%): Least percentage of population has Very low standard of living.

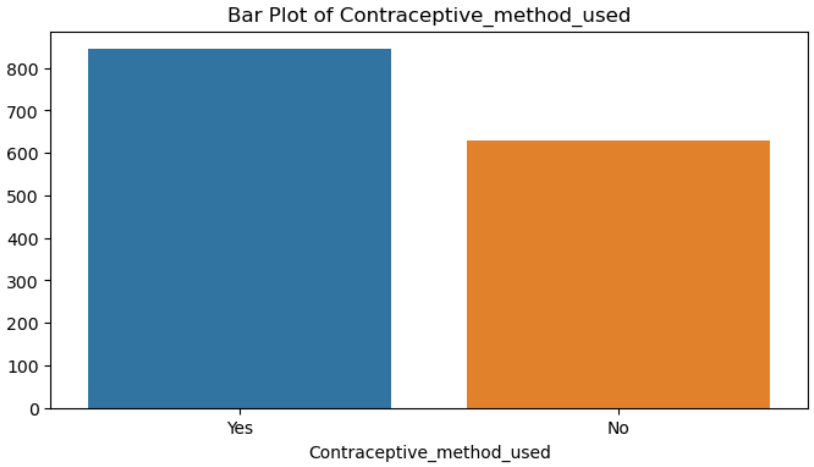
Exposed: Around 1300 individuals are exposed to media, indicate that largest group of population was aware of media.

Non-Exposed: A very least number of individuals, around 100 are not exposed to media, indicate that very least group of population are not introduce to media.



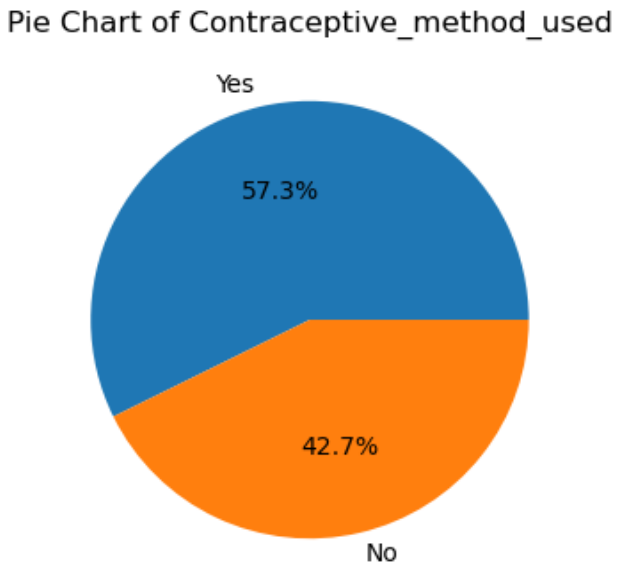
Exposed (92.6%): Largest portion of population are aware of media.

Non-Exposed (7.4%): A very least portion of population, are not exposed to media,



Contraceptive method used (Yes):
More than 800 individuals are using the Contraceptive method.

Contraceptive method used (No):
Around 600 Individuals are not using the Contraceptive method.

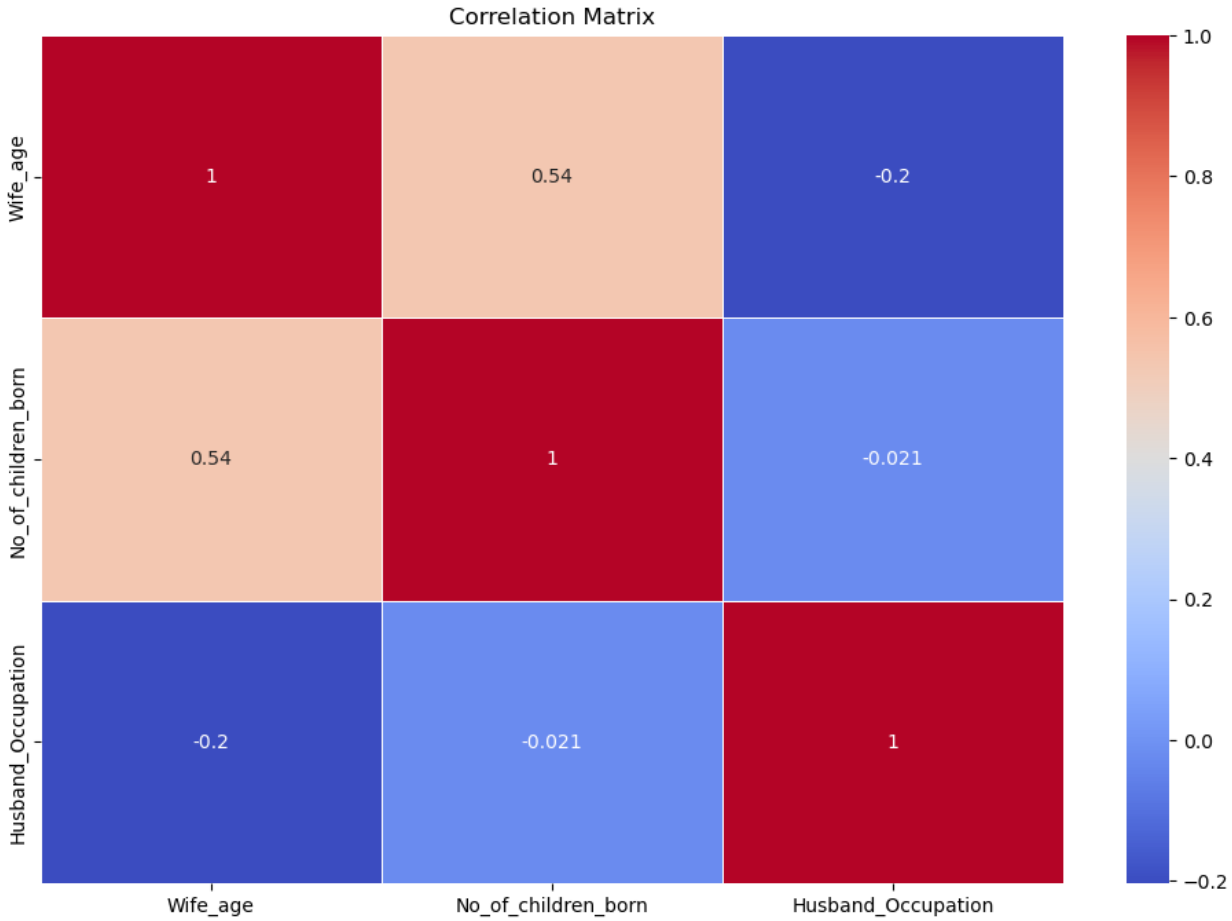


Contraceptive method used (Yes):
57% of population are using the contraceptive method, indicates that more then half of population prefer using contraceptive method.

Contraceptive method used (No):
43% of population are not using the contraceptive method.

Problem 2.1.4 Multivariate analysis:

Heat Map

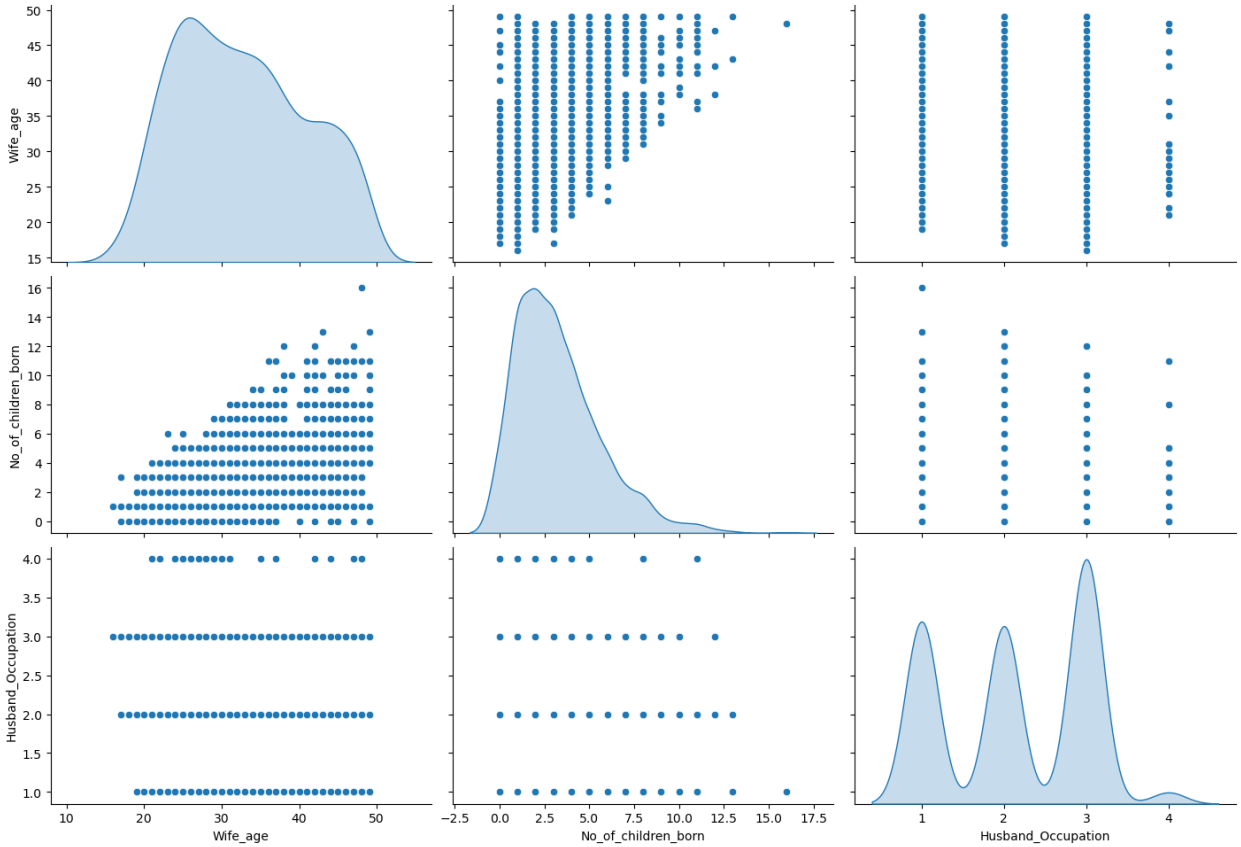


Insights: -

- **Self-Correlation:** Each variable has a perfect correlation (1) with itself, represented by the dark red squares along the diagonal.
- **Moderate Positive Correlation:** Wife_age and No_of_children_born have a moderate positive correlation (0.54).
- **Weak Negative Correlation:** Wife_age and Husband_Occupation show a slight negative correlation (-0.2).
- **Very Weak Negative Correlation:** No_of_children_born and Husband_Occupation have a very weak negative correlation (-0.021).
- **Implications**
- **Family Planning:** The moderate positive correlation between Wife_age and No_of_children_born can be useful in understanding family planning trends and targeting relevant age groups.
- **Sociological Studies:** The weak negative correlation between Wife_age and Husband_Occupation might warrant further investigation to understand any underlying sociological factors.

- **Policy Making:** Policymakers can use these correlations to design programs that address the needs of different demographic groups based on age, occupation, and family size.

Pair Plot:



Insights

Distribution of Variables:

1. Wife_age: The distribution is unimodal with a peak around age 30-35, showing that most wives are in this age range.

2. No_of_children_born: The distribution is right-skewed, with most families having between 0 and 5 children.

2. Husband_Occupation: The distribution is multimodal with peaks at 1, 2, and 3, indicating these are the most common occupation categories.

Relationships Between Variables:

1. Wife_age vs. No_of_children_born:

- There is a positive trend, indicating that as the age of the wife increases, the number of children born tends to increase.
- The scatter plot shows a clear upward trend, supporting a moderate positive correlation.

2. Wife_age vs. Husband_Occupation:

- The scatter plot shows a weak relationship with no clear pattern, indicating that the age of the wife does not strongly correlate with the husband's occupation category.

3. No_of_children_born vs. Husband_Occupation:

- The scatter plot also shows a weak relationship with points spread evenly across different occupation categories, indicating no strong correlation between the number of children and the husband's occupation.

Summary

- **Wife_age**: Most wives are around 30-35 years old.
- **No_of_children_born**: Most families have between 0 and 5 children, with a right-skewed distribution.
- **Husband_Occupation**: The most common occupation categories are 1, 2, and 3.
- **Pairwise Relationships**
- **Wife_age and No_of_children_born**: Moderate positive correlation; more children tend to be born to older wives.
- **Wife_age and Husband_Occupation**: Weak relationship; no clear pattern.
- **No_of_children_born and Husband_Occupation**: Weak relationship; no clear pattern.

Implications

Family Planning:

The positive correlation between Wife_age and No_of_children_born can help in understanding family planning trends and targeting relevant age groups.

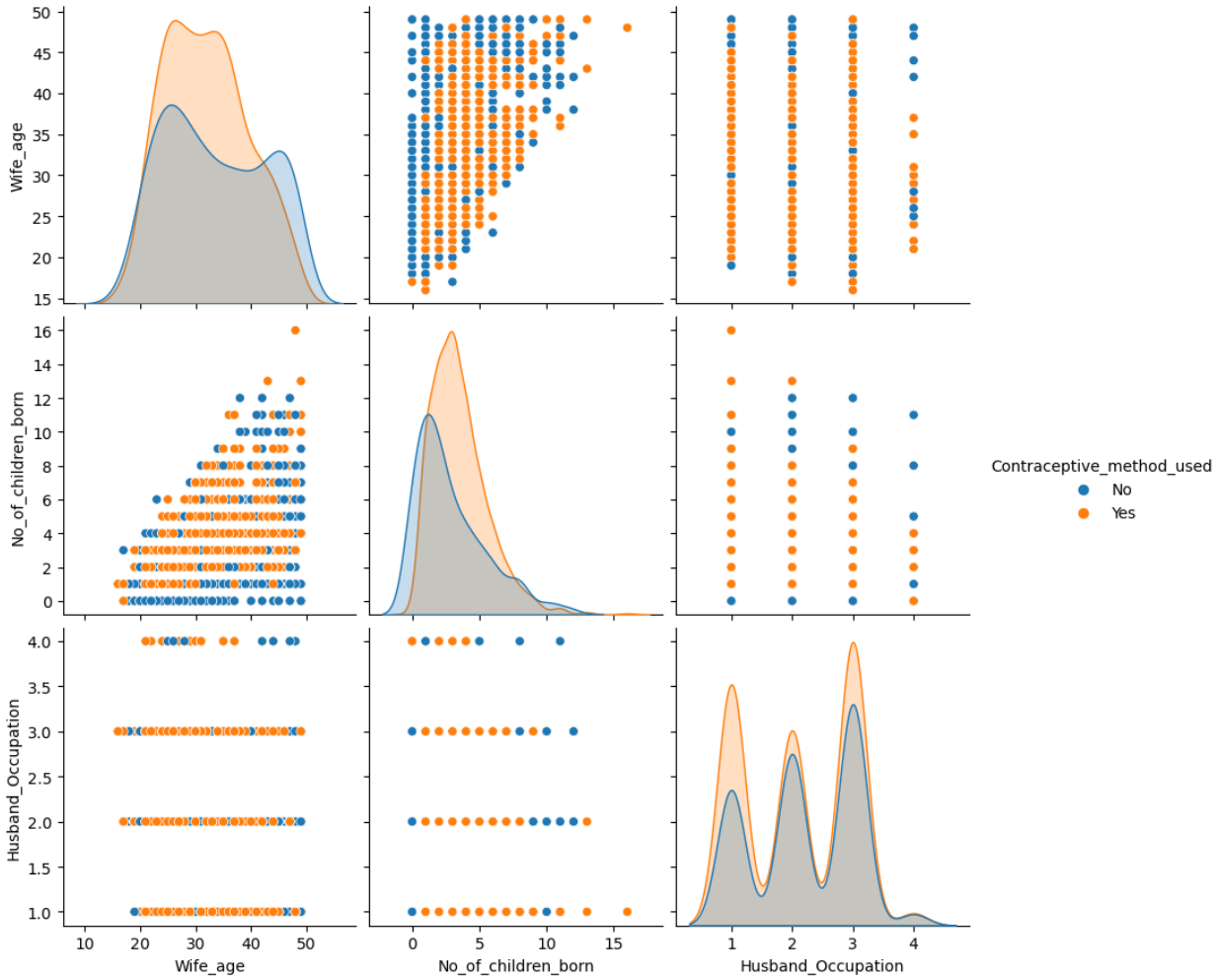
Sociological Studies:

The weak relationships between occupation and the other variables suggest that factors other than age and number of children might influence occupation.

Policy Making:

Policymakers can use these insights to design programs addressing the needs of different demographic groups based on age and family size.

Problem 2.1.5 Use appropriate visualizations to identify the patterns and insights:



Insights

Wife's Age Distribution:

The KDE plot shows that the age distribution of wives using and not using contraceptive methods is quite similar, with peaks around the mid-20s to early 30s. However, there seems to be a higher density of contraceptive users in the younger age range.

Number of Children Born:

There is a clear trend where the number of children increases with the wife's age. Families with fewer children tend to use contraceptive methods more, as seen by the higher density of orange dots in the lower range of No_of_children_born.

Husband's Occupation:

The distribution of husbands' occupations shows distinct clusters at 1, 2, 3, and 4. There is a notable distribution of contraceptive method usage across different occupations, with no single occupation showing a clear preference.

Wife's Age vs. Number of Children Born:

The scatter plot shows an increasing trend of the number of children with the wife's age. Younger wives (age < 30) tend to have fewer children and are more likely to use contraceptive methods (orange dots).

Wife's Age vs. Husband's Occupation:

No clear trend is visible between the wife's age and the husband's occupation.

Number of Children Born vs. Husband's Occupation:

The scatter plot shows no strong relationship between the number of children born and the husband's occupation. Contraceptive usage appears evenly distributed across different occupations.

Problem 2.1.6 Key meaningful observations on individual variables and the relationship between variables:

Insights:

Wife_age:

- Distribution: The distribution of ages is approximately unimodal, with a peak around the mid-20s to early 30s, indicating that this is the most common age range in the dataset.
- Central Tendency: The median age is around 30 years old, meaning half the individuals are below 30 and half are above.
- Spread and Range: The ages range from around 15 to 50 years old, with a wide spread of data points, though the density drops off significantly beyond 50 years old.
- Symmetry and Skewness: The distribution is roughly symmetric, with a slight positive skew (longer right tail) towards the older ages.
- Interquartile Range (IQR): The middle 50% of the ages fall between 25 and 38 years old, as indicated by the IQR.

No_of_childern_born:

- Right-Skewed Distribution of Number of Children Born: The distribution of the number of children born is right-skewed, indicating that most families have a smaller number of children, with fewer families having a larger number of children.
- Central Tendency and Variability: The most common number of children born (mode) is around 2, but the average (mean) is likely higher due to the right-skewed distribution. The number of children born ranges from 0 to around 17, with a gradual decline in frequency as the number of children increases. There are a few outlier families with more than 10 children, which can affect the mean and standard deviation.
- Median and Interquartile Range: The median number of children born is approximately 3, indicating that 50% of the families have 3 or fewer children, and the other 50% have more than 3 children. The interquartile range (IQR) shows the spread of the middle 50% of the data, providing insights into the variability of the number of children born.
- Implications for Policy and Research: These insights into the distribution and variability of the number of children born can inform family planning policies, target interventions, and guide further sociological research on factors influencing family size.

Husband_occupation:

- Moderate Positive Correlation: There is a moderate positive correlation between Wife_age and No_of_children_born, indicating that as the wife's age increases, the number of children born tends to increase.
- Weak Relationships: There are weak relationships between Wife_age and Husband_Occupation, as well as between No_of_children_born and Husband_Occupation, suggesting that factors other than age and number of children might influence occupation.
- Most Common Occupation: The occupation category 3 has the highest density, making it the most common occupation among the data.
- Median Occupation: The median occupation category is 2, meaning half of the occupations are below 2 and half are above 2.
- Interquartile Range: The middle 50% of the occupation categories fall between 1.5 and 3.
- Range of Data: The minimum occupation category is 1, and the maximum is 4.

Wife_education:

- Educational Attainment Distribution: The majority of wives (39.2%) have tertiary education, followed by secondary education (27.8%), primary education (22.7%), and the smallest group being uneducated (10.3%). This indicates that a significant portion of the population has pursued higher education, while a substantial number have also completed secondary school.
- Positive Correlation between Wife's Age and Number of Children: There is a moderate positive correlation between the wife's age and the number of children born. This suggests that as the wife's age increases, the number of children born tends to increase as well.
- Weak Relationships between Variables: The relationships between the wife's age and the husband's occupation, as well as the number of children born and the husband's occupation, are weak and do not show clear patterns. This implies that factors other than age and family size may be more influential in determining the husband's occupation.
- Implications for Policymaking and Research: The insights on educational attainment and the relationship between age and family size can inform family planning policies and programs. The weak relationships between variables suggest the need for further sociological research to understand the underlying factors that influence occupation and family dynamics.

Husband_education:

- Educational Attainment Distribution: The majority of wives (39.2%) have tertiary education, followed by secondary education (27.8%), primary education (22.7%), and the smallest group being uneducated (10.3%). This indicates that a significant portion of the population has pursued higher education, while a substantial number have also completed secondary school.
- Positive Correlation between Wife's Age and Number of Children: There is a moderate positive correlation between the wife's age and the number of children born. This suggests that as the wife's age increases, the number of children born tends to increase as well.
- Weak Relationships between Variables: The relationships between the wife's age and the husband's occupation, as well as the number of children born and the husband's occupation, are weak and do not show clear patterns. This implies that factors other than age and family size may be more influential in determining the husband's occupation.
- Implications for Policymaking and Research: The insights on educational attainment and the relationship between age and family size can inform family planning policies and programs. The weak relationships between variables suggest the need for further sociological research to understand the underlying factors that influence occupation and family dynamics.

Wife_religion:

- Dominance of Scientology Religion: The majority of wives (85.1%) belong to the Scientology religion, indicating it is the dominant religion in the population. The number of Scientology adherents (over 1,200 individuals) is significantly higher compared to the non-Scientology group (around 200 individuals).
- Minimal Presence of Non-Scientology Religion: Only 14.9% of the population belongs to the non-Scientology religion, suggesting it has a relatively small presence compared to Scientology.
- Potential Implications for Policy and Sociological Studies: The stark difference in the religious composition of the population may have implications for policymakers and sociologists in terms of understanding the cultural, social, and demographic dynamics of the community. Further investigation into the factors contributing to the dominance of Scientology and the relatively smaller presence of the non-Scientology religion could provide valuable insights for sociological research and policy development.

Wife_working:

- Majority of Wives are Non-Working: The data shows that the majority of wives, over 1,000 individuals, are non-working. Only around 400 wives out of the entire population are working, indicating that the number of working wives is significantly lower than non-working wives.
- Low Percentage of Working Wives: The data suggests that only 25.1% of the entire population of wives are working. This means that the vast majority, around 75% of wives, are non-working.
- Preference for Non-Working Roles: The data implies that the majority of wives, around 75% of the population, prefer to take on non-working roles rather than working roles. This suggests a cultural or societal preference for non-working wives in the given context.

Standards_of_living_index:

- Majority have Very High or High Standard of Living: The data shows that the largest group, comprising around 46% of the population, has a very high standard of living. The second-largest group, around 29% of the population, has a high standard of living.
- Smaller Groups have Lower Standard of Living: A smaller group, around 15% of the population, has a low standard of living. The smallest group, around 9% of the population, has a very low standard of living.
- Significant Disparity in Living Standards: The data reveals a significant disparity in the standard of living among the different groups, with the majority enjoying a very high or high standard of living, while smaller groups struggle with low or very low standards of living.
- Potential for Targeted Interventions: These insights can inform policymakers and stakeholders to develop targeted interventions and programs to address the needs of the groups with lower standards of living, with the goal of reducing the overall disparity and improving the well-being of the entire population.

Media_exposure:

- Majority of the Population is Exposed to Media: he data shows that around 1,300 individuals are exposed to media, which represents the largest group of the population. he exposed group accounts for 92.6% of the total population, indicating that the vast majority of the population is aware of and has access to media.
- nority of the Population is Not Exposed to Media:Only around 100 individuals are not exposed to media, representing the smallest group of the population. The non-exposed group accounts for 7.4% of the total population, suggesting that a very small portion of the population is not introduced to or aware of media.
- Significant Disparity in Media Exposure: The data reveals a significant disparity between the exposed and non-exposed groups, with the exposed group being the dominant majority. This suggests that

there may be barriers or challenges in reaching and engaging the non-exposed population, which could have implications for media accessibility, information dissemination, and social inclusion.

Contraceptive_method_used:

- Contraceptive Usage: More than 800 individuals (57% of the population) are using the contraceptive method. Around 600 individuals (43% of the population) are not using the contraceptive method.
- Contraceptive Preference: The majority of the population (57%) prefer using the contraceptive method, indicating a higher adoption rate.
- Contraceptive Non-Usage: A significant portion of the population (43%) are not using the contraceptive method, suggesting a need to understand the reasons behind this and address any barriers to access or acceptance.

Relationships Between Variables:

Wife's Age and Number of Children Born:

The scatter plot shows an increasing trend, indicating that as the wife's age increases, the number of children born tends to increase. Younger wives (age < 30) tend to have fewer children and are more likely to use contraceptive methods.

Wife's Age and Husband's Occupation:

The scatter plot shows no clear trend between the wife's age and the husband's occupation, suggesting a weak relationship.

Number of Children Born and Husband's Occupation:

The scatter plot shows no strong relationship between the number of children born and the husband's occupation.Contraceptive usage appears to be evenly distributed across different occupation categories.

Summary:

- There is a moderate positive correlation between the wife's age and the number of children born, with younger wives tending to have fewer children and being more likely to use contraceptives.
- The wife's age and the husband's occupation show a weak relationship, with no clear pattern.
- The number of children born and the husband's occupation also exhibit a weak relationship, with contraceptive usage being evenly distributed across different occupation categories.

Implications:

Family Planning:

The positive correlation between Wife_age and No_of_children_born can help in understanding family planning trends and targeting relevant age groups.

Sociological Studies:

The weak relationships between occupation and the other variables suggest that factors other than age and number of children might influence occupation.

Policy Making:

Policymakers can use these insights to design programs addressing the needs of different demographic groups based on age and family size.

PROBLEM 2.2 DATA PROCESSING:

Problem 2.2.1 Missing value check and treatment:

There are two variables that has data missing.

```
Wife_age                    71
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
```

Missing Value Treatment by Imputing missing values with the mean of the column.

Problem 2.2.2 Outlier Treatment:

We Checked Outlier and following observations are made:

1. Outliers were identified in `No_of_children_born` columns only, indicating the presence of values that are significantly higher or lower than the rest of the data points.
2. Given the high number of outliers in the dataset, it is recommended to treat outliers before proceeding with Mean Value.

Problem 2.2.3 Feature Engineering:

- We can create new feature like wife age group & husband age group.
- We can create Bin based on No_of_children_born.

But i don't think it is necessary of doing so.

Problem 2.2.4 Encode the data:

Encoding of categorical data like 'Wife_education', 'Husband_education', 'Wife_religion', 'Wife_Working', 'Standard_of_living_index', 'Media_exposure ' & 'Contraceptive_method_used' so that we can convert it to numerical data type.

Problem 2.2.5 Train-test split:

We are going to split data into train set and test set using train_test_split function from scikit-learn. We are split data into 75 – 25 splits where 75% of entire data will be train dataset and 25% will be test dataset.

Shape values of train – test dataset are:

(X_train.shape, X_test.shape, y_train.shape, y_test.shape)

```
(1104, 15) (369, 15) (1104, 1) (369, 1)
```

PROBLEM 2.3 MODEL BUILDING - LINEAR REGRESSION:

Problem 2.3.1 Build a Logistic Regression model:

We used the scikit-learn library to apply a Linear Regression model to the data.

Classification Report:

```
Logistic Regression - Train Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.51      0.58       468
           1       0.70      0.82      0.75       636

    accuracy                           0.69      1104
   macro avg       0.69      0.67      0.67      1104
weighted avg       0.69      0.69      0.68      1104


Logistic Regression - Test Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.48      0.58       161
           1       0.68      0.87      0.76       208

    accuracy                           0.70       369
   macro avg       0.71      0.67      0.67       369
weighted avg       0.70      0.70      0.68       369
```

Problem 2.3.2 Build a Linear Discriminant Analysis model:

We require library such as “sklearn.discriminant_analysis.LinearDiscriminantAnalysis” for the LDA model

Classification report:

```
LDA - Train Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.50      0.58       468
           1       0.69      0.82      0.75       636

    accuracy                           0.69      1104
   macro avg       0.68      0.66      0.66      1104
weighted avg       0.69      0.69      0.68      1104


LDA - Test Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.46      0.56       161
           1       0.67      0.87      0.76       208

    accuracy                           0.69       369
   macro avg       0.70      0.66      0.66       369
weighted avg       0.70      0.69      0.67       369
```

Problem 2.3.3 Build a CART model:

We require library such as “DecisionTreeClassifier” from sklearn.tree module for the CART model

Classification report:

```
CART - Train Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.99      0.98       468
           1       0.99      0.97      0.98       636

    accuracy                           0.98      1104
   macro avg       0.98      0.98      0.98      1104
weighted avg       0.98      0.98      0.98      1104


CART - Test Classification Report:
              precision    recall  f1-score   support

           0       0.61      0.58      0.59       161
           1       0.69      0.71      0.70       208

    accuracy                           0.65       369
   macro avg       0.65      0.65      0.65       369
weighted avg       0.65      0.65      0.65       369
```

Problem 2.3.4 Prune the CART model by finding the best hyperparameters using GridSearch:

We require libraries such as “DecisionTreeClassifier” from sklearn.tree module & “GridSearchCV” from sklearn.model_selection Module to Prune the CART model.

Classification report:

```
Pruned CART - Train Classification Report:
               precision    recall  f1-score   support

           0       0.84      0.55      0.66       468
           1       0.74      0.93      0.82       636

    accuracy                           0.77      1104
   macro avg       0.79      0.74      0.74      1104
weighted avg       0.78      0.77      0.75      1104


Pruned CART - Test Classification Report:
               precision    recall  f1-score   support

           0       0.74      0.47      0.58       161
           1       0.68      0.87      0.76       208

    accuracy                           0.70       369
   macro avg       0.71      0.67      0.67       369
weighted avg       0.71      0.70      0.68       369
```

Problem 2.3.5 Check the performance of the models across train and test set using different metrics:

Model Evalution

We are going to use metrics like accuracy, precision, recall, and F1-score to evaluate model performance on both training and testing sets.

Logistic Regression

Train Classification Report: precision recall f1-score support 0 0.68 0.51 0.58 468 1 0.70 0.82 0.75 636 accuracy 0.69 1104 macro avg 0.69 0.67 0.67 1104 weighted avg 0.69 0.69 0.68 1104

Test Classification Report: precision recall f1-score support 0 0.73 0.48 0.58 161 1 0.68 0.87 0.76 208 accuracy 0.70 369 macro avg 0.71 0.67 0.67 369 weighted avg 0.70 0.70 0.68 369

LDA

Train Classification Report: precision recall f1-score support 0 0.68 0.50 0.58 468 1 0.69 0.82 0.75 636 accuracy 0.69 1104 macro avg 0.68 0.66 0.66 1104 weighted avg 0.69 0.69 0.68 1104

Test Classification Report: precision recall f1-score support 0 0.73 0.46 0.56 161 1 0.67 0.87 0.76 208 accuracy 0.69 369 macro avg 0.70 0.66 0.66 369 weighted avg 0.70 0.69 0.67 369

CART

Train Classification Report: precision recall f1-score support 0 0.97 0.99 0.98 468 1 0.99 0.97 0.98 636 accuracy 0.98 1104 macro avg 0.98 0.98 0.98 1104 weighted avg 0.98 0.98 0.98 1104

Test Classification Report: precision recall f1-score support 0 0.61 0.58 0.59 161 1 0.69 0.71 0.70 208 accuracy 0.65 369 macro avg 0.65 0.65 0.65 369 weighted avg 0.65 0.65 0.65 369

Pruned CART

Train Classification Report: precision recall f1-score support 0 0.84 0.55 0.66 468 1 0.74 0.93 0.82 636 accuracy 0.77 1104 macro avg 0.79 0.74 0.74 1104 weighted avg 0.78 0.77 0.75 1104

Test Classification Report: precision recall f1-score support 0 0.74 0.47 0.58 161 1 0.68 0.87 0.76 208 accuracy 0.70 369 macro avg 0.71 0.67 0.67 369 weighted avg 0.71 0.70 0.68 369

Problem 2.3.6 Compare the performance of all the models built and choose the best one with proper rationale:

Metrics Summary

Logistic Regression

- Train Accuracy: 0.69
- Test Accuracy: 0.70
- Train Precision: 0.68 (class 0), 0.70 (class 1)
- Test Precision: 0.73 (class 0), 0.68 (class 1)
- Train Recall: 0.51 (class 0), 0.82 (class 1)
- Test Recall: 0.48 (class 0), 0.87 (class 1)
- Train F1-Score: 0.58 (class 0), 0.75 (class 1)
- Test F1-Score: 0.58 (class 0), 0.76 (class 1)

Linear Discriminant Analysis (LDA)

- Train Accuracy: 0.69
- Test Accuracy: 0.69
- Train Precision: 0.68 (class 0), 0.69 (class 1)
- Test Precision: 0.73 (class 0), 0.67 (class 1)
- Train Recall: 0.50 (class 0), 0.82 (class 1)
- Test Recall: 0.46 (class 0), 0.87 (class 1)
- Train F1-Score: 0.58 (class 0), 0.75 (class 1)
- Test F1-Score: 0.56 (class 0), 0.76 (class 1)

CART

- Train Accuracy: 0.98
- Test Accuracy: 0.65
- Train Precision: 0.97 (class 0), 0.99 (class 1)
- Test Precision: 0.61 (class 0), 0.69 (class 1)
- Train Recall: 0.99 (class 0), 0.97 (class 1)
- Test Recall: 0.58 (class 0), 0.71 (class 1)
- Train F1-Score: 0.98 (class 0), 0.98 (class 1)
- Test F1-Score: 0.59 (class 0), 0.70 (class 1)

Pruned CART

- Train Accuracy: 0.77
- Test Accuracy: 0.70
- Train Precision: 0.84 (class 0), 0.74 (class 1)
- Test Precision: 0.74 (class 0), 0.68 (class 1)
- Train Recall: 0.55 (class 0), 0.93 (class 1)
- Test Recall: 0.47 (class 0), 0.87 (class 1)
- Train F1-Score: 0.66 (class 0), 0.82 (class 1)
- Test F1-Score: 0.58 (class 0), 0.76 (class 1)

Comparison and Rationale

Accuracy:

- Logistic Regression and Pruned CART have the highest test accuracy (0.70), followed closely by LDA (0.69).
- CART has a lower test accuracy (0.65) despite having an extremely high train accuracy (0.98), indicating overfitting.

Precision, Recall, and F1-Score:

- For class 1 (target class), Pruned CART and Logistic Regression perform similarly well in terms of test F1-score (0.76), indicating a balance between precision and recall.
- For class 0, Pruned CART has slightly better test precision (0.74) compared to Logistic Regression (0.73) and LDA (0.73), but Logistic Regression has higher recall (0.48 vs. 0.47 for Pruned CART).

Overfitting:

- CART shows clear signs of overfitting with a huge gap between train and test performance.
- Pruned CART addresses overfitting to some extent, maintaining a balance between train and test performance.

Conclusion:

Best Model: Pruned CART

Rationale: Pruned CART provides a good balance of performance metrics, effectively addressing the overfitting issue observed with the unpruned CART model. It achieves the highest test accuracy along with Logistic Regression and has competitive precision, recall, and F1-scores. Additionally, decision trees are interpretable and can provide valuable insights into the feature importance.

Alternative: Logistic Regression

Rationale: Logistic Regression also performs well, with similar accuracy and F1-scores to Pruned CART. It's a simpler model that is easier to interpret and can be a good choice if interpretability and simplicity are priorities.

PROBLEM 2.4 BUSINESS INSIGHTS & RECOMMENDATIONS:

Problem 2.4.1 Comment on the importance of features based on the best model:

We need to retrieve feature and its importance values using “feature_importances” Function.

	Feature	Importance
1	No_of_children_born	0.373766
0	Wife_age	0.328257
4	Wife_ education_Tertiary	0.127555
12	Standard_of_living_index_Very High	0.036042
2	Husband_Occupation	0.024444
9	Wife_religion_Scientology	0.023753
3	Wife_ education_Secondary	0.020272
6	Husband_education_Secondary	0.015999
10	Wife_Working_Yes	0.014812
13	Standard_of_living_index_Very Low	0.012924
14	Media_exposure _Not-Exposed	0.011924
7	Husband_education_Tertiary	0.007976
8	Husband_education_Uneducated	0.002274
5	Wife_ education_Uneducated	0.000000
11	Standard_of_living_index_Low	0.000000

Analysis of Feature Importance

Number of Children Born (0.373766):

- Importance: Highest importance in predicting contraceptive use.
- Insight: Families with more children are more likely to use contraceptives. This suggests that educational campaigns about family planning should target families with fewer children, highlighting the benefits of contraceptives.

Wife's Age (0.328257):

- Importance: Second highest importance.
- Insight: Younger or older women might have different contraceptive needs. Tailoring messages to different age groups can improve the adoption of contraceptive methods.

Wife's Education - Tertiary (0.127555):

- Importance: Significant importance.
- Insight: Higher education levels in women correlate with contraceptive use. Programs to promote female education can indirectly increase contraceptive use.

Standard of Living - Very High (0.036042):

- Importance: Moderate importance.
- Insight: Higher standard of living leads to higher contraceptive use. This implies that improving the overall standard of living can have a positive impact on contraceptive adoption.

Husband's Occupation (0.024444):

- Importance: Low importance.
- Insight: While not highly important, the husband's occupation can still influence contraceptive use. Understanding the occupations that are less likely to use contraceptives can help in targeted interventions.

Problem 2.4.2 Conclude with the key takeaways (actionable insights and recommendations) for the business:

Based on the feature importance and the insights drawn, here are the actionable recommendations:

Focus on Families with More Children:

Develop targeted campaigns emphasizing the benefits of contraceptives for families with multiple children. Highlighting how contraceptives can help in better managing family size and improving quality of life can be effective.

Tailor Messaging to Different Age Groups:

Create age-specific educational materials and programs. Younger women might be more interested in the health benefits and future family planning, while older women might focus on economic stability and existing family management.

Promote Female Education:

Increase efforts to promote tertiary education for women. Education has a strong correlation with contraceptive use, so investing in female education can have long-term benefits for family planning.

Improve Standard of Living:

Work on policies and programs that enhance the standard of living. Better economic conditions and access to resources will likely lead to increased contraceptive use.

Understand Occupational Influence:

Study the specific occupations of husbands that show lower contraceptive use. Develop occupation-specific interventions to address misconceptions or barriers related to contraceptive use in those occupations.

Address Religious and Cultural Factors:

Since religion and cultural factors have some influence, work with community leaders to promote contraceptive use within cultural and religious contexts. Ensuring that the messages are culturally sensitive will enhance acceptance.

Enhance Media Exposure:

Media exposure is relatively less important but still plays a role. Increase the accessibility and quality of information about contraceptives through various media channels. Utilize social media, TV, and radio to spread awareness.

Conclusion:

By focusing on these key areas, the Republic of Indonesia Ministry of Health can effectively promote the use of contraceptives among married women. Targeted educational campaigns, improving socioeconomic conditions, and leveraging cultural contexts can lead to better adoption rates and improved family planning outcomes.

==========Thanks You!==========