Predictive_ Modelling _Project

RAGHAVENDRA KUMAR J R

PGP – DSBA Online

Date: 24/09/2023

Problem 1: Linear Regression

The comp-activ databases is a collection of a computer systems activity measures.
The data was collected from a Sun Sparcstation 20/712 with 128 Mbytes of memory running in a multi-user university department. Users would typically be doing a large variety of tasks ranging from accessing the internet, editing files or running very cpu-bound programs.

As you are a budding data scientist you thought to find out a linear equation to build a model to predict 'usr'(Portion of time (%) that cpus run in user mode) and to find out how each attribute affects the system to be in 'usr' mode using a list of system attributes.

Dataset for Problem 1: compactiv.xlsx

DATA DICTIONARY:

System measures used:

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

Problem 2: Logistic Regression, LDA and CART

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

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

Dataset for Problem 2: Contraceptive_method_dataset.xlsx

Data Dictionary:

1. Wife's age (numerical)
2. Wife's education (categorical) 1=uneducated, 2, 3, 4=tertiary
3. Husband's education (categorical) 1=uneducated, 2, 3, 4=tertiary
4. Number of children ever born (numerical)
5. Wife's religion (binary) Non-Scientology, Scientology
6. Wife's now working? (Binary) Yes, No
7. Husband's occupation (categorical) 1, 2, 3, 4(random)
8. Standard-of-living index (categorical) 1=verlow, 2, 3, 4=high
9. Media exposure (binary) Good, Not good
10. Contraceptive method used (class attribute) No,Yes

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

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

2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets u sing Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model: Compare both the models and write inference which model is best/optimiz ed.

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

Contents:

Loaded the data – Compactiv_data

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

Sample of the dataset: cdata

lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	pgscan
1	0	2147	79	68	0.2	0.2	40671	53995	0	0
0	0	170	18	21	0.2	0.2	448	8385	0	0
15	3	2162	159	119	2	2.4		31950	0	0
0	0	160	12	16	0.2	0.2		8670	0	0
5	1	330	39	38	0.4	0.4		12185	0	0

atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap	usr
0	1.6	2.6	16	26.4	CPU_Bound	4670	1730946	95
0	0	0	15.63	16.83	Not_CPU_Bound	7278	1869002	97
1.2	6	9.4	150.2	220.2	Not_CPU_Bound	702	1021237	87
0	0.2	0.2	15.6	16.8	Not_CPU_Bound	7248	1863704	98
0	1	1.2	37.8	47.6	Not_CPU_Bound	633	1760253	90

The dataset contains 8,192 rows and 22 columns. The data types for each column vary, with most being numerical (either integers or floats). One column, runqsz, is an object d ata type, suggesting contain text or categorical data.

Exploratory Data Analysis let us check the types of variables in the data frame:

```
RangeIndex: 8192 entries, 0 to 8191
Data columns (total 22 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   lread     8192 non-null   int64  
 1   lwrite    8192 non-null   int64  
 2   scall     8192 non-null   int64  
 3   sread     8192 non-null   int64  
 4   swrite    8192 non-null   int64  
 5   fork      8192 non-null   float64
 6   exec      8192 non-null   float64
 7   rchar     8088 non-null   float64
 8   wchar     8177 non-null   float64
 9   pgout     8192 non-null   float64
 10  ppgout    8192 non-null   float64
 11  pgfree    8192 non-null   float64
 12  pgscan    8192 non-null   float64
 13  atch      8192 non-null   float64
 14  pgin      8192 non-null   float64
 15  ppgin     8192 non-null   float64
 16  pflt      8192 non-null   float64
 17  vflt      8192 non-null   float64
 18  runqsz    8192 non-null   object 
 19  freemem   8192 non-null   int64  
 20  freeswap  8192 non-null   int64  
 21  usr       8192 non-null   int64  
dtypes: float64(13), int64(8), object(1)
memory usage: 1.4+ MB
```

cdata.dtypes:

```
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
dtype: object
```

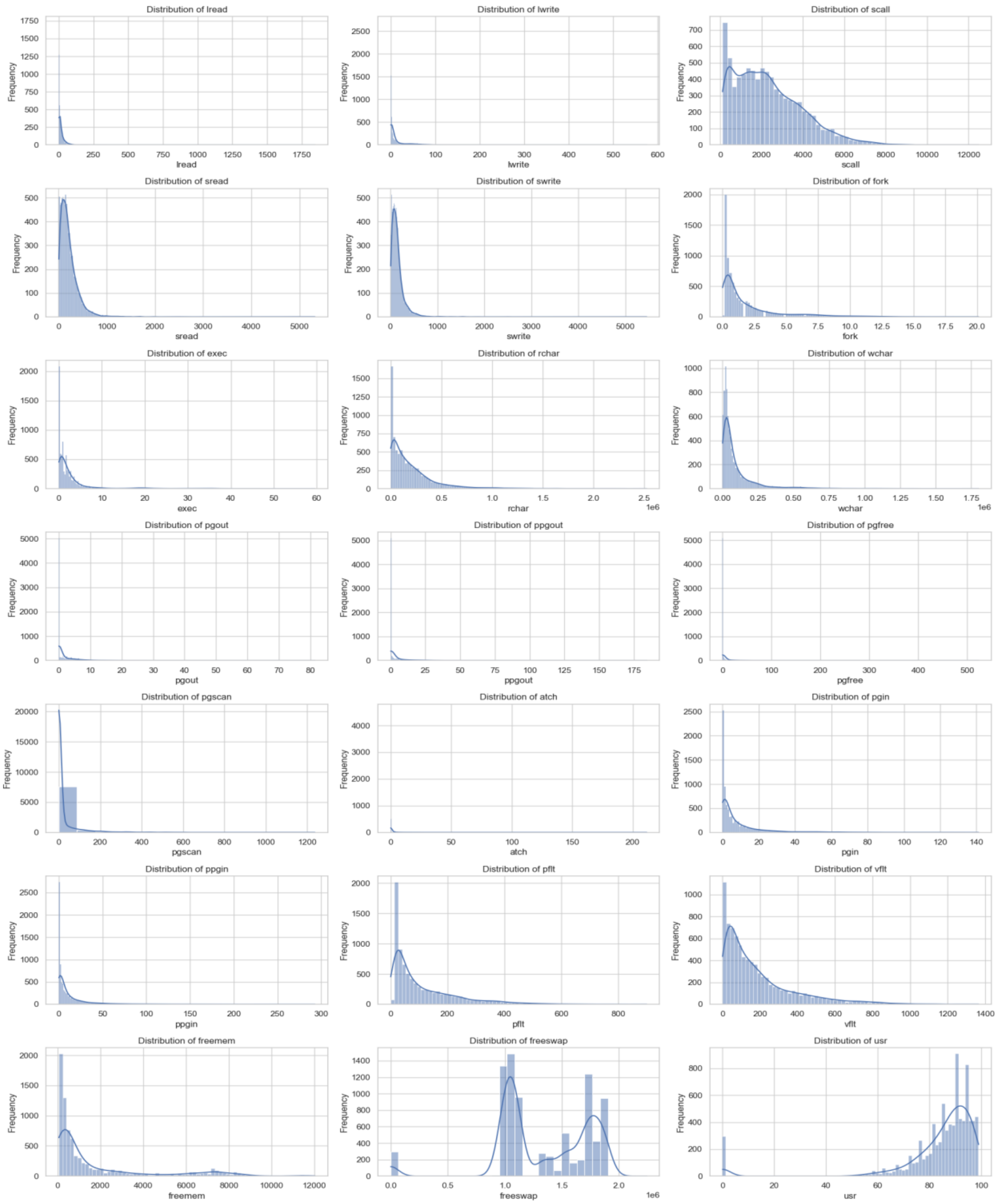
Statistical Summary:

Descriptive statistics help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognize d types of descriptive statistics are measures of centre: the mean, median, and mode, which are used at almost all levels of math and statistics.

	lread	**lwrite**	**scall**	**sread**	**swrite**	**fork**	**exec**	**rchar**	**wchar**	**pgout**	**pgfree**
count	8192	8192	8192	8192	8192	8192	8192	8088	8177	8192	8192
mean	19.56	13.11	2306	210.5	150.1	1.885	2.792	2E+05	95903	2.285	11.92
std	53.35	29.89	1634	199	160.5	2.479	5.212	2E+05	1E+05	5.307	32.36
min	0	0	109	6	7	0	0	278	1498	0	0
25%	2	0	1012	86	63	0.4	0.2	34092	22916	0	0
50%	7	1	2052	166	117	0.8	1.2	1E+05	46619	0	0
75%	20	10	3317	279	185	2.2	2.8	3E+05	1E+05	2.4	5
max	1845	575	12493	5318	5456	20.12	59.56	3E+06	2E+06	81.44	523

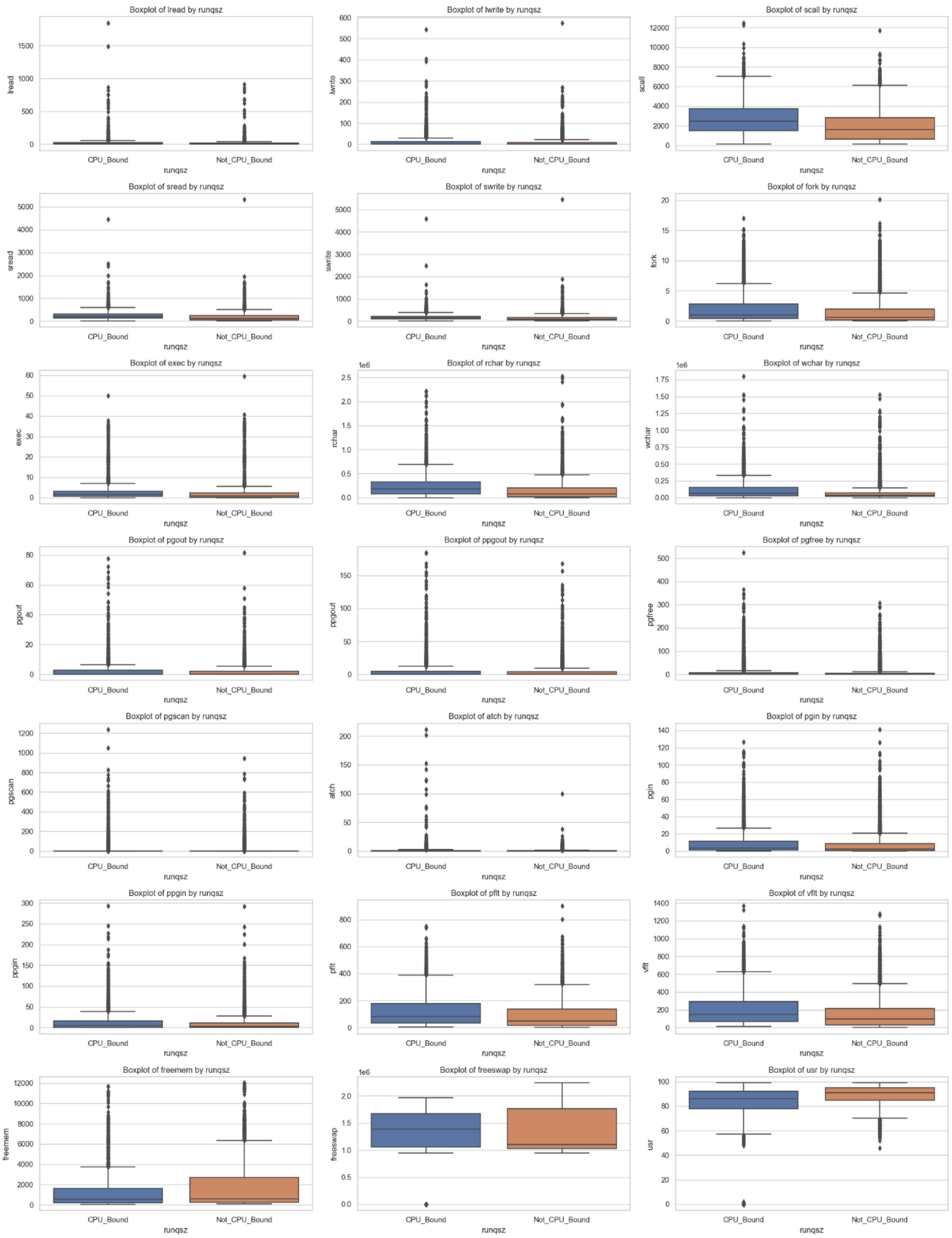
pgscan	**atch**	**pgin**	**ppgin**	**pflt**	**vflt**	**freemem**	**freeswap**	**usr**
8192	8192	8192	8192	8192	8192	8192	8192	
21.53	1.128	8.278	12.39	109.8	185.3	1763	1E+06	83.97
71.14	5.708	13.87	22.28	114.4	191	2482	4E+05	18.4
0	0	0	0	0	0.2	55	2	0
0	0	0.6	0.6	25	45.4	231	1E+06	81
0	0	2.8	3.8	63.8	120.4	579	1E+06	89
0	0.6	9.765	13.8	159.6	251.8	2002	2E+06	94
1237	211.6	141.2	292.6	899.8	1365	12027	2E+06	99

Univariate analysis:

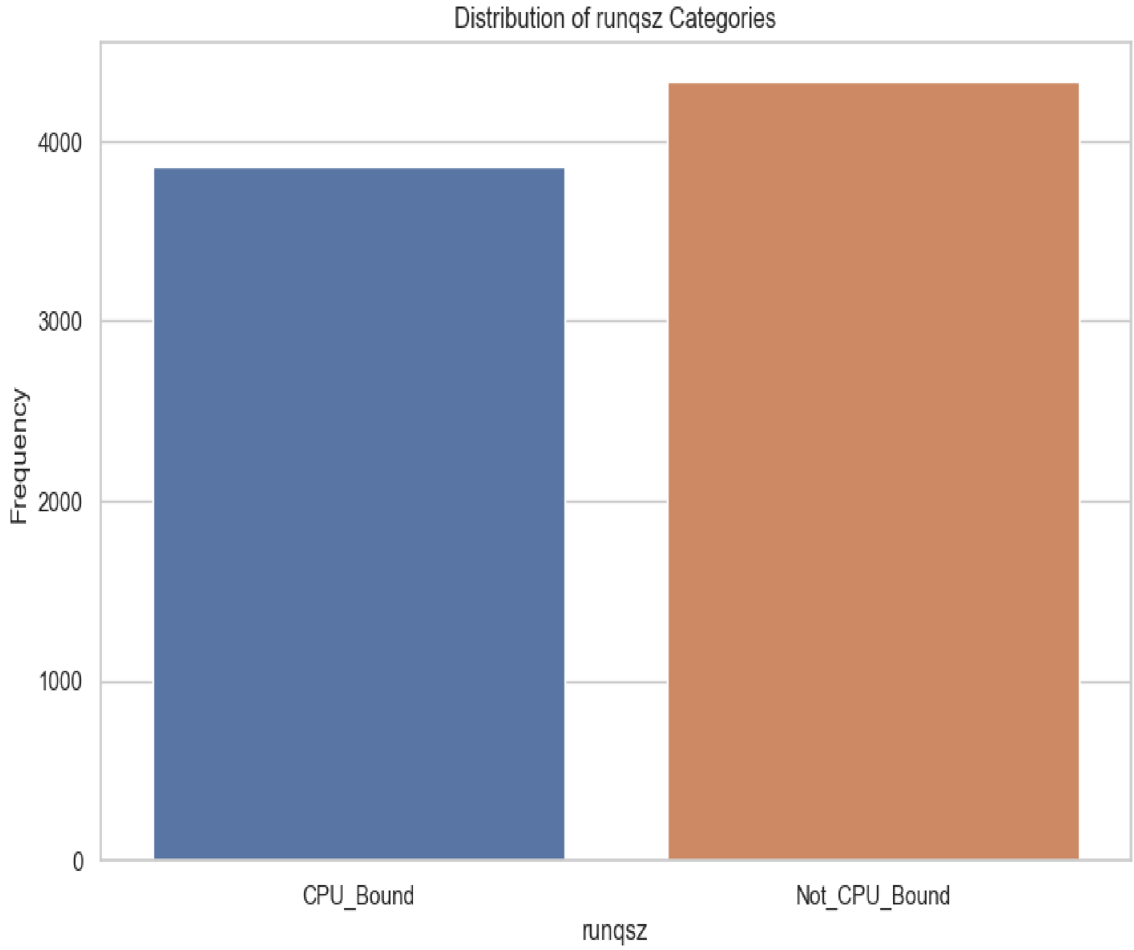


- Some variables like usr, fork, and exec show a somewhat normal distribution, although not perfectly symmetrical.
- Others like lread, lwrite, and swrite are heavily skewed towards the lower end, indicating that most of the values are small.
- Several variables like pgout, ppgout, and pgfree have a lot of zeros, indicating that these events are rare.

Looping through the numerical columns to create box plots

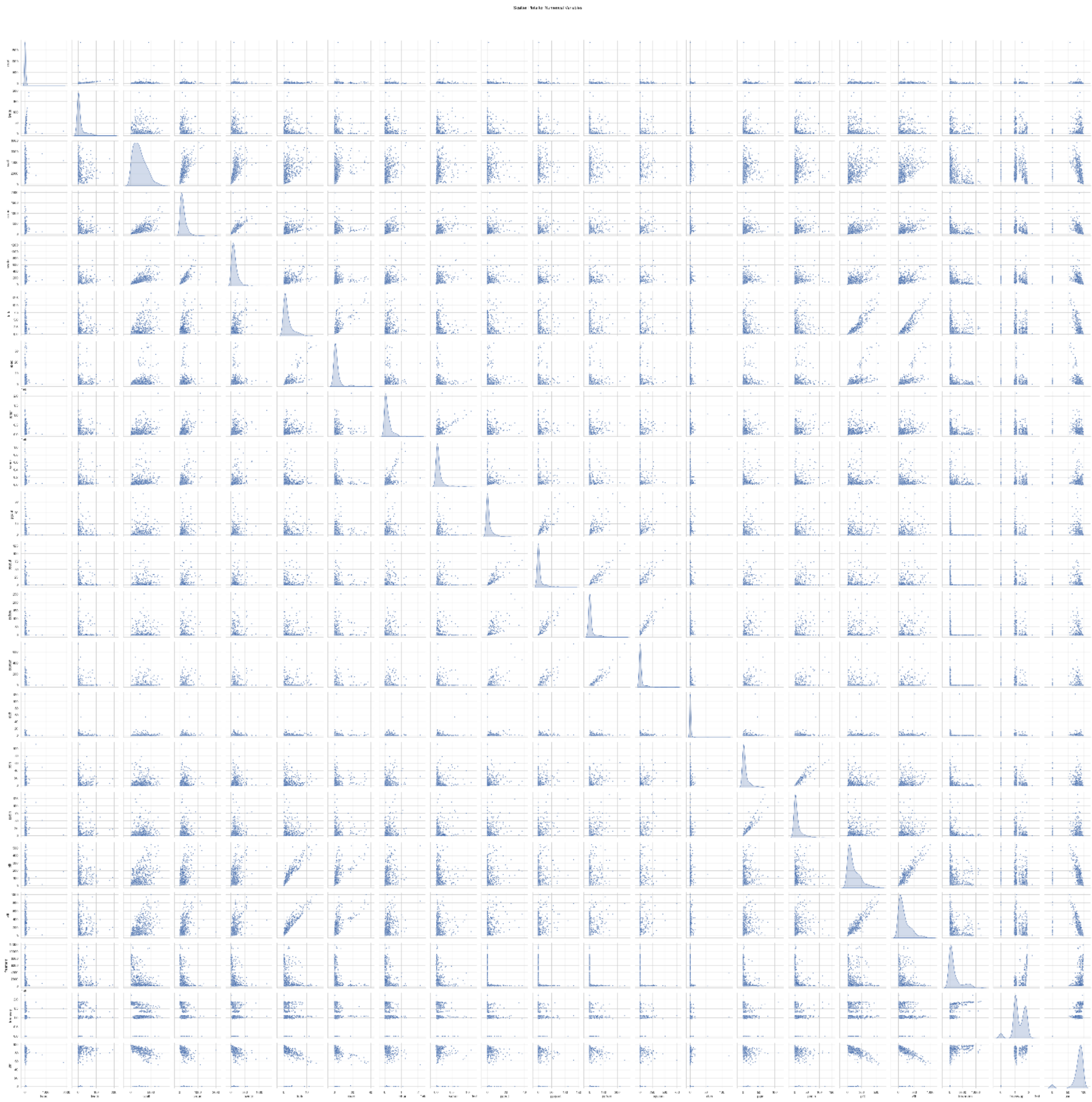


Exploring the distribution of the categorical variable 'runqsz'



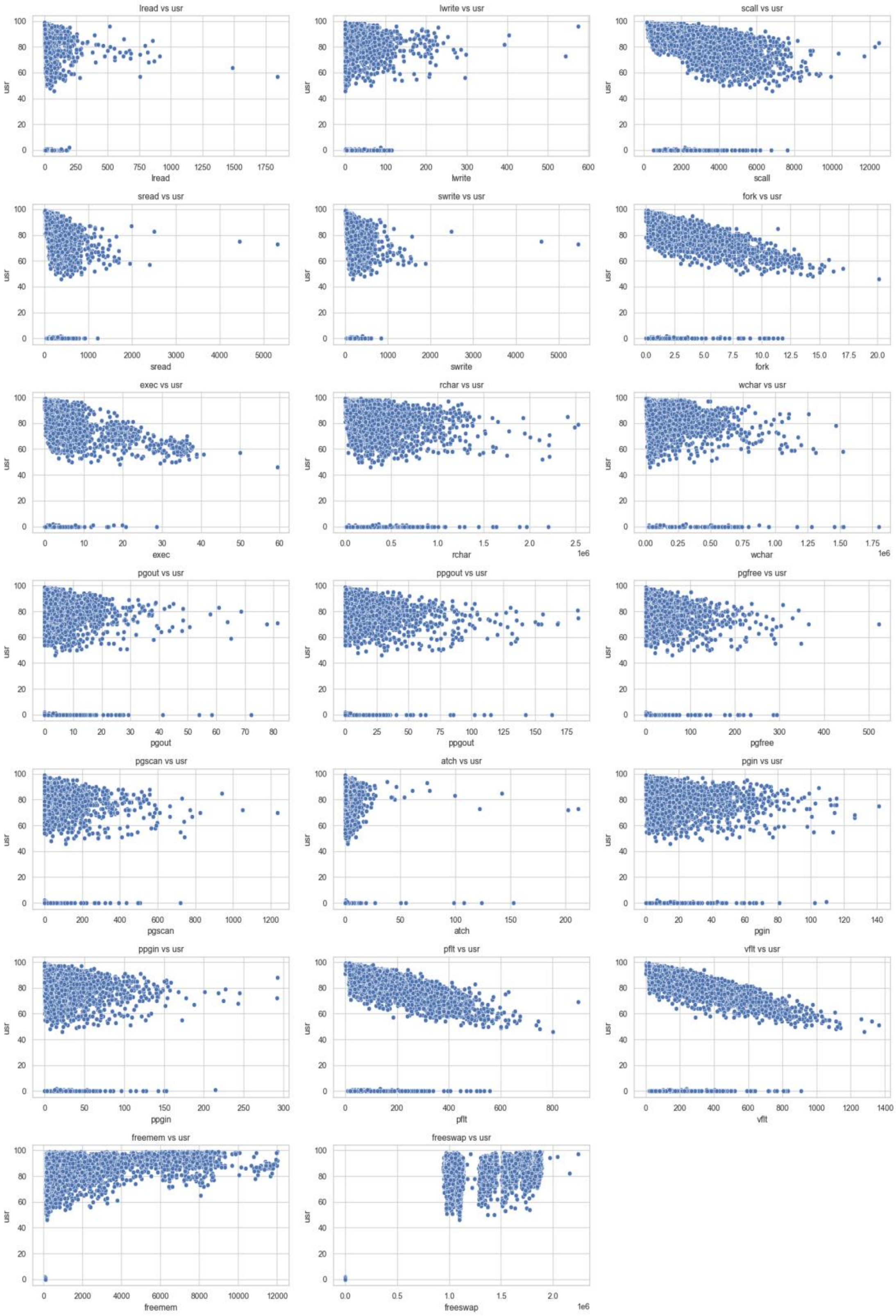
Bivariate Analysis:

Scatter plots for numerical vs. numerical variables (Sample of 500 points for visualization)

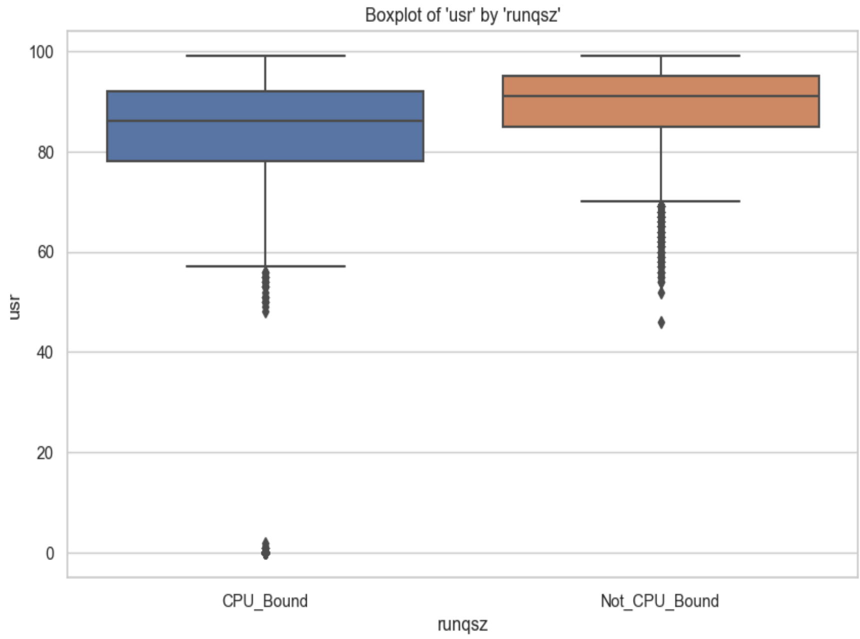


- The scatter plots provide an idea of how each feature correlates with the target v ariable 'usr'. Some features, like freeswap, appear to have a somewhat linear rela tionship with usr, while others don't show a clear pattern.
- Correlation with 'usr'
- The correlation values range from -1 to 1, with values close to 1 indicating a stro ng positive correlation, values close to -1 indicating a strong negative correlation, and values around 0 indicating no correlation.
- freeswap shows the highest positive correlation with usr (0.6790.679).
- vflt shows the highest negative correlation with usr (−0.421−0.421).
- These features could be important predictors for the target variable 'usr'

Performing Bivariate Analysis using scatter plots



Box plot for the relationship between the categorical variable 'runqsz' and 'usr'

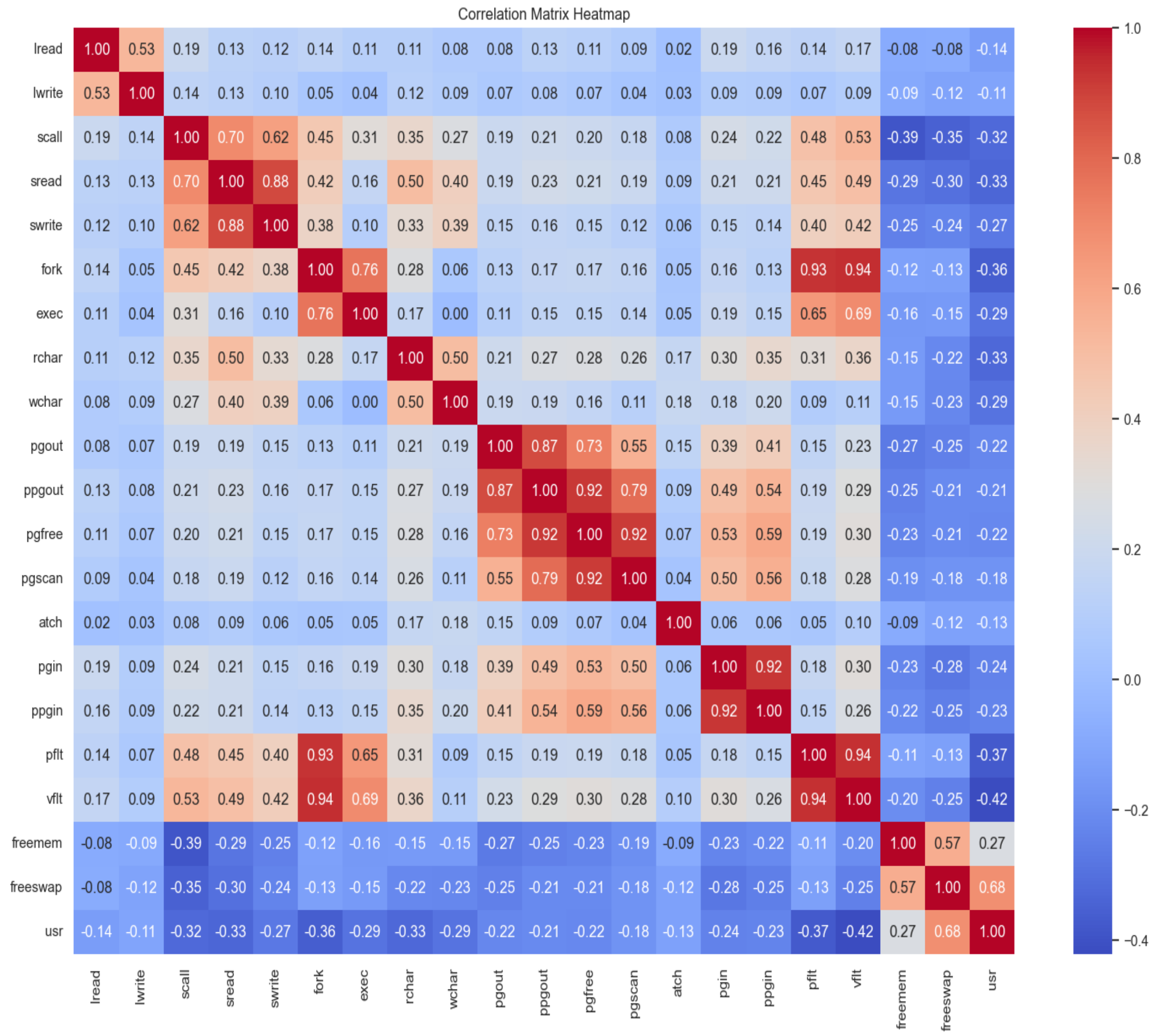


Correlation with 'usr':

usr 1.000000
freeswap 0.678526
freemem 0.270308
lwrite -0.111213
atch -0.125074
lread -0.141394
pgscan -0.181488
ppgout -0.212295
pgfree -0.216278
pgout -0.221877
ppgin -0.233682
pgin -0.241720
swrite -0.272252
exec -0.288526
wchar -0.288974
scall -0.323188
rchar -0.329737
sread -0.332160
fork -0.363277
pflt -0.372495
vflt -0.420685
Name: usr, dtype: float64

Multivariate Analysis:

Performing Multivariate Analysis using a correlation matrix Heatmaps for all numerical variables.



- The heat map provides a visual representation of the correlation matrix among a ll numerical features, including our target variable usr.
- Dark blue and dark red cells indicate strong correlations, either negative or posit ive, respectively.
- Light-coloured cells indicate weak correlations.
- Observations:
- The features free swap and freemem have a strong positive correlation with each other (0.630.63), and both have a positive correlation with usr.
- Features like pgin and ppgin, pflt and vflt, etc., also have strong positive correlati ons among themselves.

1.2 Impute null values if present, also check for the values which are equal to zero. Do they have a ny meaning or do we need to change them or drop them? Check for the possibility of creating new f eatures if required. Also check for outliers and duplicates if there.

Check for missing values in the dataset: We have missing values in the **rchar 104**, and **wchar 15** columns, there are various ways to handle missing data.

```
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
```

Replace the missing values in the rchar and wchar columns with their respective median s.

```
cdata['rchar'].fillna(cdata['rchar'].median(), inplace=True)
cdata['wchar'].fillna(cdata['wchar'].median(), inplace=True)
 cdata.isnull().sum()
```

```
lread         0
lwrite        0
scall         0
sread         0
swrite        0
fork          0
exec          0
rchar         0
wchar         0
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
```

Checking for Zero value counts:

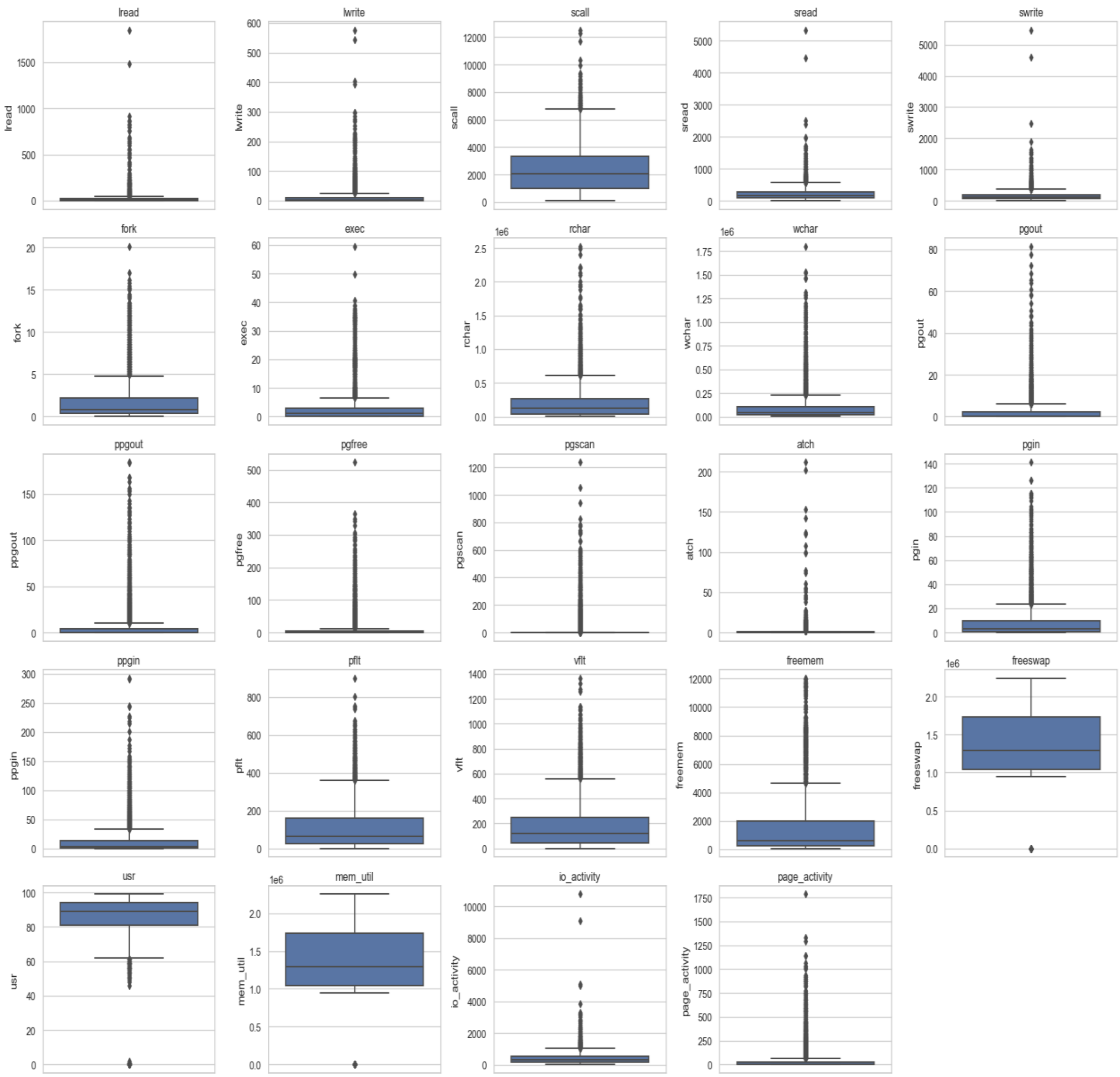
```
lread        675
lwrite      2684
scall          0
sread          0
swrite         0
fork          21
exec          21
rchar          0
wchar          0
pgout       4878
ppgout      4878
pgfree      4869
pgscan      6448
atch        4575
pgin        1220
ppgin       1220
pflt           3
vflt           0
runqsz         0
freemem        0
freeswap       0
usr          283
dtype: int64
```

Zero values could indicate inactivity or lack of certain types of system calls or operation s during the collection interval. But we keeping zeros.

Duplicates: 0 – There are no duplicate rows in the dataset

Outliers:

Creating boxplots to identify outliers in the numerical columns:

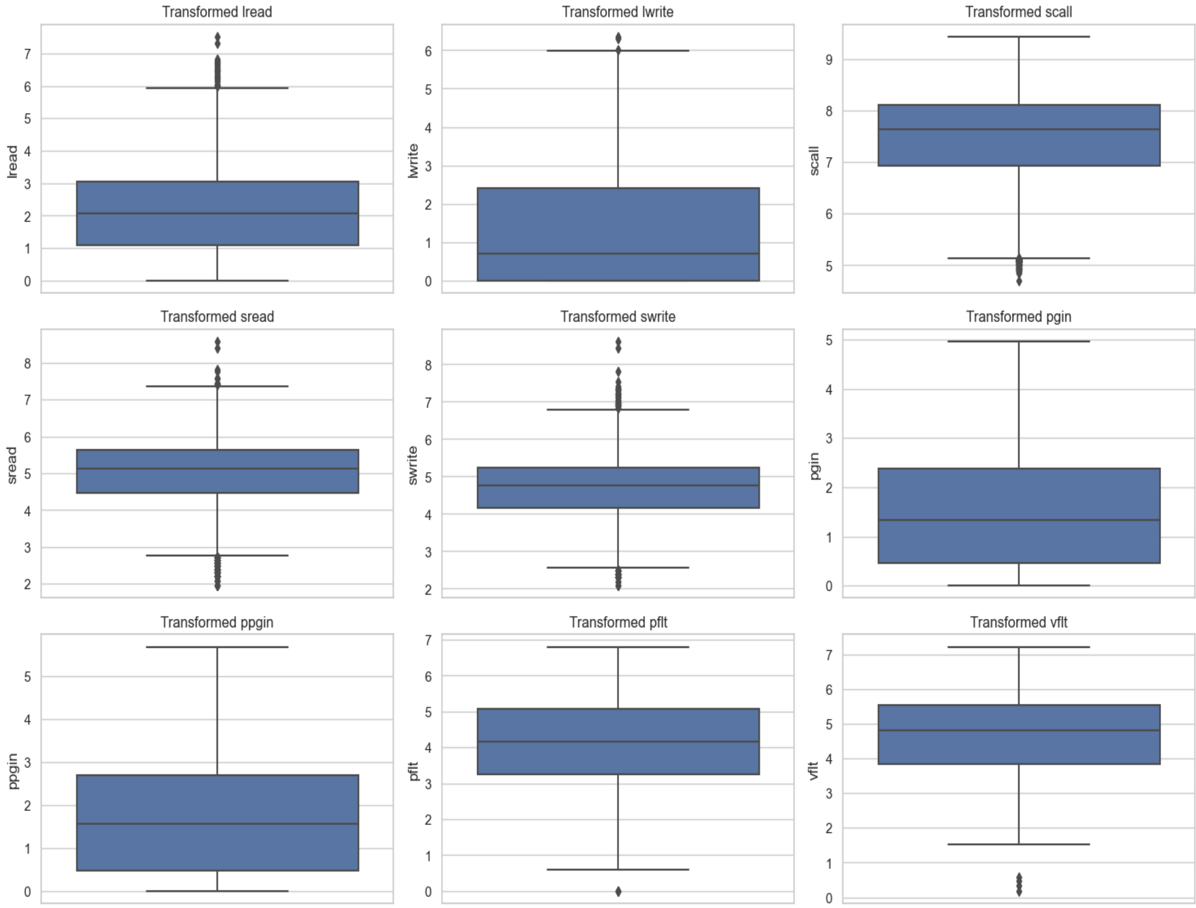


The boxplots provide a visual representation of the distribution of each numerical column, including potential outliers. Outliers are data points that fall significantly outside the main range of the data. In the boxplots, these are represented as dots beyond the "whiskers" of the boxes.

Here's a summary of some columns with noticeable outliers:

lread, lwrite: These columns show a number of points far from the main cluster of data.
scall, sread, swrite: Similarly, these columns also contain outliers, although they appear to be less extreme compared to lread and lwrite.
pgin, ppgin, pflt, vflt: These columns have several outliers that fall significantly outside the main range of the data.

Outliers Transformed:



The boxplots show the distributions of the transformed columns. After applying the logarithmic transformation, the outliers appear to be closer to the main cluster of data points. This transformation can make the data more amenable to linear modelling techniques.

Function to remove outliers using IQR method:

```
Cdata shape - ((8192, 25),
Cdata clean. Shape   (2769, 25))
```

1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appr opriate method from stats model. Create multiple models and check the performance of Predi ctions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and s elect the best one with appropriate reasoning.

Performing **one-hot encoding** for the 'runqsz' column:

lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	ppgin	pflt
1	0	2147	79	68	0.2	0.2	40671	53995	0	2.6	16
0	0	170	18	21	0.2	0.2	448	8385	0	0	15.63
15	3	2162	159	119	2	2.4	125473.5	31950	0	9.4	150.2
0	0	160	12	16	0.2	0.2	125473.5	8670	0	0.2	15.6
5	1	330	39	38	0.4	0.4	125473.5	12185	0	1.2	37.8

vflt	freemem	freeswap	usr	mem_util	io_activity	page_activity	runqsz_Not_CPU_Bound
26.4	4670	1730946	95	1735616	148	1.6	0
16.83	7278	1869002	97	1876280	39	0	1
220.2	702	1021237	87	1021939	296	6	1
16.8	7248	1863704	98	1870952	28	0.2	1
47.6	633	1760253	90	1760886	83	1	1

The column runqsz has been successfully one-hot encoded, resulting in a new column n amed runqsz_Not_CPU_Bound.

Split the data into training and test sets (70:30 ratio):

We'll split the data into training and testing sets. We'll use 70% of the data for training a nd the remaining 30% for testing.

Output: ((5734, 21), (2458, 21), (5734,), (2458,))

Training set for features (X_train): 5,734 rows and 21 columns
Testing set for features (X_test): 2,458 rows and 21 columns
Training set for the target variable (y_train): 5,734 rows
Testing set for the target variable (y_test): 2,458 rows.

Linear Regression using scikit-learn -

Evaluate the performance of the model using R2, RMSE, and Adjusted R2.

Output:

(0.6387425796550663,
10.948969765407034,
0.6406031458101988,
11.284636162471799)

The performance metrics for the Linear Regression model are as follows:

Training Set:
R2: 0.639 (R-squared represents the proportion of the variance for the dependent variable that's explained by the independent variables in the model. Closer to 1 is generally better.)

RMSE: 10.95 (Root Mean Square Error is a measure of the differences between values predicted by the model and the values actually observed. Lower values are better.)

Test Set:
R2: 0.641 (It's a good sign that the test R-squared is close to the training R-squared, as it suggests the model generalizes well.)

RMSE: 11.28 (Similar to the training set, this gives us an idea of how well the model performs when exposed to new, unseen data.)

Stats model:

OLS Regression Results

Dep. Variable:	usr	**R-squared:**	0.639
Model:	OLS	**Adj. R-squared:**	0.637
Method:	Least Squares	**F-statistic:**	480.9
Date:	Wed, 20 Sep 2023	**Prob (F-statistic):**	0.00
Time:	10:39:53	**Log-Likelihood:**	-21859.
No. Observations:	5734	**AIC:**	4.376e+04
Df Residuals:	5712	**BIC:**	4.391e+04
Df Model:	21		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	43.1070	0.749	57.538	0.000	41.638	44.576
lread	-0.0186	0.003	-5.856	0.000	-0.025	-0.012
lwrite	-0.0002	0.006	-0.028	0.978	-0.012	0.012
scall	0.0010	0.000	7.119	0.000	0.001	0.001
sread	0.0025	0.002	1.317	0.188	-0.001	0.006
swrite	-0.0038	0.002	-1.847	0.065	-0.008	0.000
fork	-1.8016	0.249	-7.233	0.000	-2.290	-1.313
exec	-0.0611	0.048	-1.260	0.208	-0.156	0.034
rchar	-4.054e-06	8.67e-07	-4.676	0.000	-5.75e-06	-2.35e-06
wchar	-1.031e-05	1.28e-06	-8.064	0.000	-1.28e-05	-7.81e-06

pgout	-0.2519	0.065	-3.900	0.000	-0.378	-0.125
ppgout	0.1400	0.036	3.876	0.000	0.069	0.211
pgfree	-0.0936	0.019	-4.918	0.000	-0.131	-0.056
pgscan	0.0160	0.006	2.749	0.006	0.005	0.027
atch	0.0299	0.028	1.084	0.278	-0.024	0.084
pgin	0.0713	0.028	2.525	0.012	0.016	0.127
ppgin	-0.0468	0.018	-2.643	0.008	-0.082	-0.012
pflt	-0.0394	0.004	-9.384	0.000	-0.048	-0.031
vflt	0.0213	0.003	6.505	0.000	0.015	0.028
freemem	-0.0016	7.48e-05	-21.839	0.000	-0.002	-0.001
freeswap	3.301e-05	4.58e-07	72.155	0.000	3.21e-05	3.39e-05
runqsz_Not_CPU_Bound	7.9102	0.306	25.811	0.000	7.309	8.511

Omnibus:	1404.592	**Durbin-Watson:**	1.989
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4036.957
Skew:	-1.277	**Prob(JB):**	0.00
Kurtosis:	6.220	**Cond. No.**	7.40e+06

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.4e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

The summary statistics from the stats models library provide a wealth of information. Here are some key points:

Significance of Variables: The P>|t| column gives us the p-value for each variable. A small p-valu e (<0.05<0.05) indicates that the variable is significant.

Variables like lwrite, sread, swrite, exec, and atch have p-values greater than 0.05, suggesting th ey are not significant predictors for usr.

Coefficients: The coef column tells us the change in the dependent variable (usr) for a one-unit c hange in the predictor variable, while holding other predictors constant.

For instance, the coefficient for freemem is -0.0016, meaning that for each additional free memo ry page, the usr percentage decreases by 0.0016 units.

Adjusted R2: The Adjusted R2 value is 0.637, which is a measure of how well the model explains the variability in the dependent variable. It's relatively close to 1, suggesting that the model is fa irly good.

F-statistic: The F-statistic tests the overall significance of the model. The Prob (F-statistic) is ext remely low, suggesting that the model is statistically significant.

Omnibus and Jarque-Bera (JB) Tests: These are tests for the normality of residuals. A Prob (Omn ibus) or Prob (JB) close to zero indicates that the residuals are not normally distributed.

Condition Number: The large condition number indicates that there might be strong multicollin earity or other numerical problems.

Create multiple models and check the performance of Predictions on Train and Test sets using Rsq uare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reason ing.

Output:

```
0.6383977252582141,
10.954194432766299,
0.6423068131039881,
11.257857821040735,
 0.6374144273743865
```

The performance metrics for the new Linear Regression model using only significant variables are as follows:

Training Set:

R2: 0.6384
RMSE (Root Mean Squared Error): 10.954
Adjusted R2: 0.6374

Test Set:
R2: 0.6423
RMSE (Root Mean Squared Error): 11.258

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

Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Exploratory Data Analysis:

There are 22 columns and 8,192 rows in the dataset.

For the columns, the primary data types are integers, floats, and one object type.

Some variables, as shown by descriptive statistics, have a roughly normal distribution, but other s have a skewed distribution.

The distribution of each variable was disclosed by a univariate analysis, and the relationship bet ween two variables was demonstrated by a bivariate study. The multivariate analysis shed light on how all numerical features relate to one another.

Data preparation and cleaning:

'rchar' and 'wchar' missing values were approximated using median values.

Zero values were left in some columns after being given some thought to their possible relevance.

The IQR approach was used to identify outliers and treat them.

There were no duplicate rows in the dataset.

The categorical column "runqsz" was one-hot encoded, according to feature engineering.
A 70:30 split of the dataset was used to create training and testing sets.

Modelling:

Scikit-learn was employed to apply linear regression.

Using the statistics model, the significance of the variables was examined, and some non-signific ant predictors were found.

R-squared, RMSE, and Adjusted R-squared were used to measure the performance of various models.

Insights:

Significant predictors for the target variable "usr" include the variables "freeswap," "freemem," "lread," and "pgout."

Some variables, including 'lwrite','sread','swrite', 'exec', and 'atch', were discovered to be non-significant and may not be required for the prediction.

The dependent variable's variability is explained by the linear regression model in around 64% of cases (adjusted R2 = 0.6374).

The reliability of various statistical tests may be impacted by the model's residuals' imperfect normal distribution.

The high condition number suggests that the dataset may be multicollinear.

Feature Selection: The model's effectiveness can be increased by excluding some variables that were found to be non-significant.

Model selection: While linear regression offers a decent starting point, it is possible to investigate different models to determine if they perform better, such as Random Forest or Gradient Boosting.

Data acquisition: In order to make sure that all relevant variables are recorded and that the data is as clean and accurate as possible, it might be good to collect more data or to examine the data collection method.

Monitoring: It's important to keep an eye on a machine learning model's performance over time and retrain it if necessary, just as with any other model.

Business Plan: Knowing which system calls or actions have the most effects on CPU utilization can help with system performance optimization and hardware or software recommendations.

Conclusions:

Model Performance: For the linear regression model using just significant variables, the R2 value was approximately 0.64. As a result, the model explains around 64% of the variance in CPU usage, making it a rather well-fit model for this complicated dataset.

By using these suggestions, businesses can improve performance, better manage system resources, and potentially avoid system failures or other issues.

As a result, this research gives a basis for estimating CPU utilization based on different system parameters, the model's accuracy and applicability will be maintained through routine monitoring and improvement.

Question - 02

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

Contraceptive method _dataset loaded as df.

Sample of the dataset: df

Wife_age	Wife_education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_method_used
24	Primary	Secondary	3	Scientology	No	2	High	Exposed	No
45	Uneducated	Secondary	10	Scientology	No	3	Very High	Exposed	No
43	Primary	Secondary	7	Scientology	No	3	Very High	Exposed	No
42	Secondary	Primary	9	Scientology	No	3	High	Exposed	No
36	Secondary	Secondary	8	Scientology	No	3	Low	Exposed	No

The dataset contains 1473 rows and 10 columns. The data types for each column vary, with most being objects, floats and integer data type, suggesting contain text or categorical data.

Exploratory Data Analysis let us check the types of variables in the data frame

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

Statistical summary:

Variable	count	mean	std	min	25%	50%	75%	max
Wife_age	1402	32.6062 8	8.27492 7	16	26	32	39	49
No_of_children_b orn	1452	3.25413 2	2.36521 2	0	1	3	4	16
Husband_Occupat ion	1473	2.13781 4	0.86485 7	1	1	2	3	4

Find below table - Descriptive Statistics Summary

Wife_age: The average age of the wives is approximately 32.6 years, with a minimum age of 16 and a maximum age of 49.
Wife_education: The most common education level for wives is 'Tertiary' (577 occurrences).
Husband_education: Similarly, the most common education level for husbands is also 'Tertiary ' (899 occurrences).
No_of_children_born: The average number of children ever born to the women in this survey is approximately 3.25. The minimum number of children is 0 and the maximum is 16.
Wife_religion: The majority of wives follow Scientology (1253 out of 1473).
Wife_Working: A large number of wives (1104 out of 1473) are not currently working.
Husband_Occupation: The mean occupation category is approximately 2.14, but this is a categorical variable, so the mean may not be very informative.
Standard_of_living_index: The most common standard of living is 'Very High' (684 occurrences).
Media_exposure: Most wives (1364 out of 1473) are exposed to media.
Contraceptive_method_used: The majority of the women (844 out of 1473) use some form of contraceptive method.

Variable	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Wife_age	1402				32.606277	8.274927	16	26	32	39	49
Wife_education	1473	4	Tertiary	577							
Husband_education	1473	4	Tertiary	899							
No_of_children_born	1452				3.254132	2.365212	0	1	3	4	16
Wife_religion	1473	2	Scientology	1253							
Wife_Working	1473	2	No	1104							
Husband_Occupation	1473				2.137814	0.864857	1	1	2	3	4
Standard_of_living_index	1473	4	Very High	684							
Media_exposure	1473	2	Exposed	1364							
Contraceptive_method_used	1473	2	Yes	844							

Check for missing values in the dataset:

We have missing values in the Wife_age: Contains 71 and No_of_children_born: Contains 21 missing values columns, there are various ways to handle missing data.

```
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
dtype: int64
```

Replace the missing values in the Wife_age and No_of_children_born columns with their respective median

df ['Wife_age'].fillna (df['Wife_age'].median(), inplace=True)
df ['No_of_children_born'].fillna(df['No_of_children_born'].median(), inplace=True)
df.isnull ().sum ()

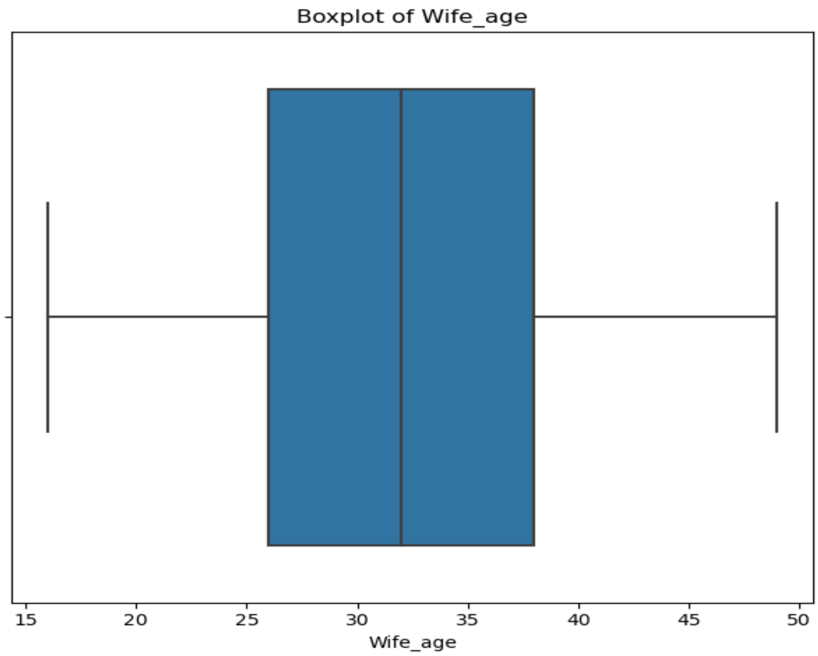
```
Wife_age                     0
Wife_ education              0
Husband_education            0
No_of_children_born          0
Wife_religion                0
Wife_Working                 0
Husband_Occupation           0
Standard_of_living_index     0
Media_exposure               0
Contraceptive_method_used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

Duplicate check:

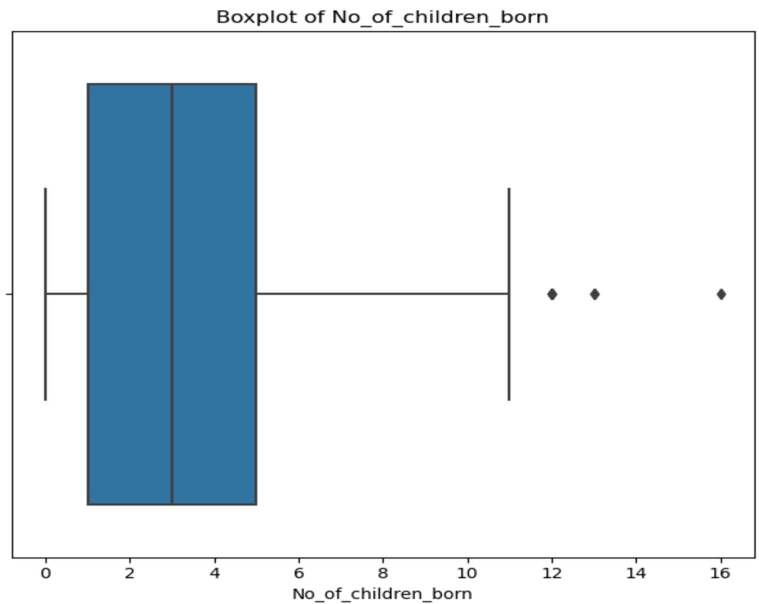
There are 85 duplicate rows in the dataset. We should consider removing these duplicat es to improve the quality of our analysis.

After dropping Duplicates: 0 – There are no duplicate rows in the dataset.

Outlier's analysis for Wife_age and No_of_children_born:

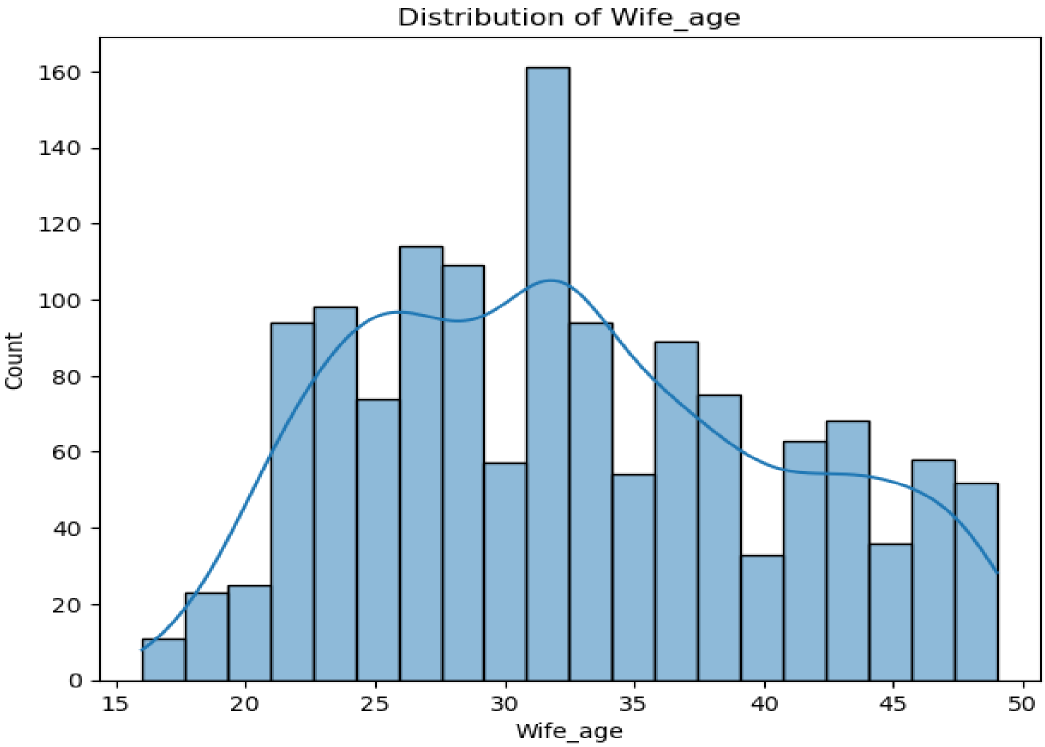


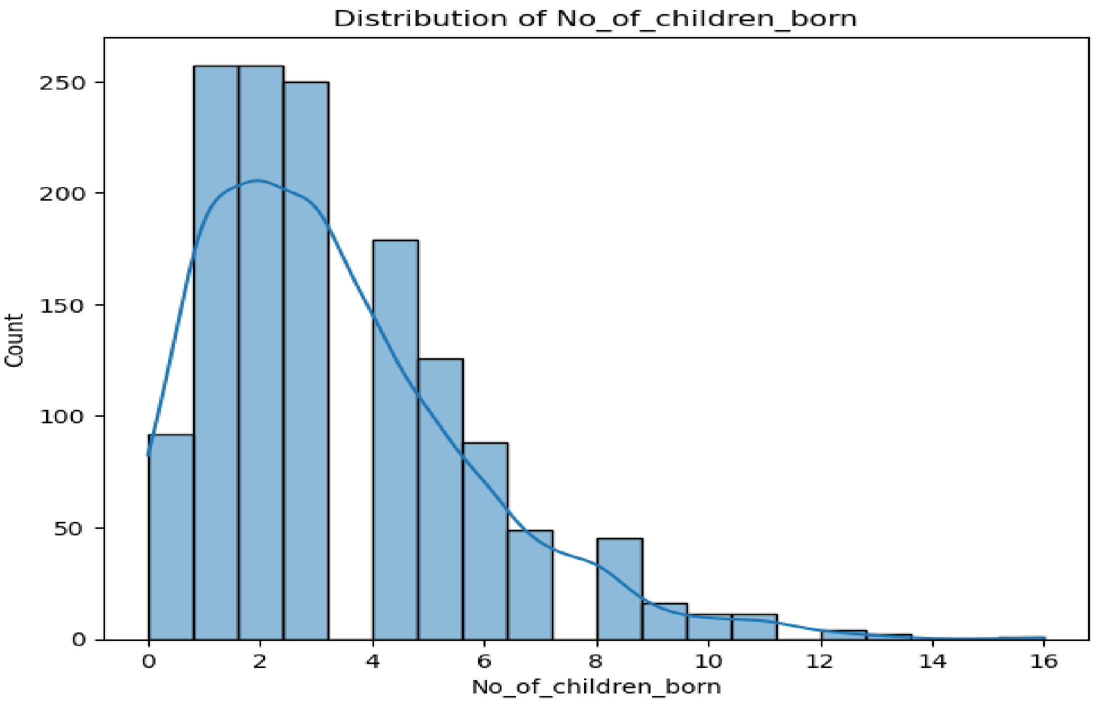
Wife_age: There don't appear to be any outliers in the "Wife's age" variable. The data seems to be fairly well-distributed.

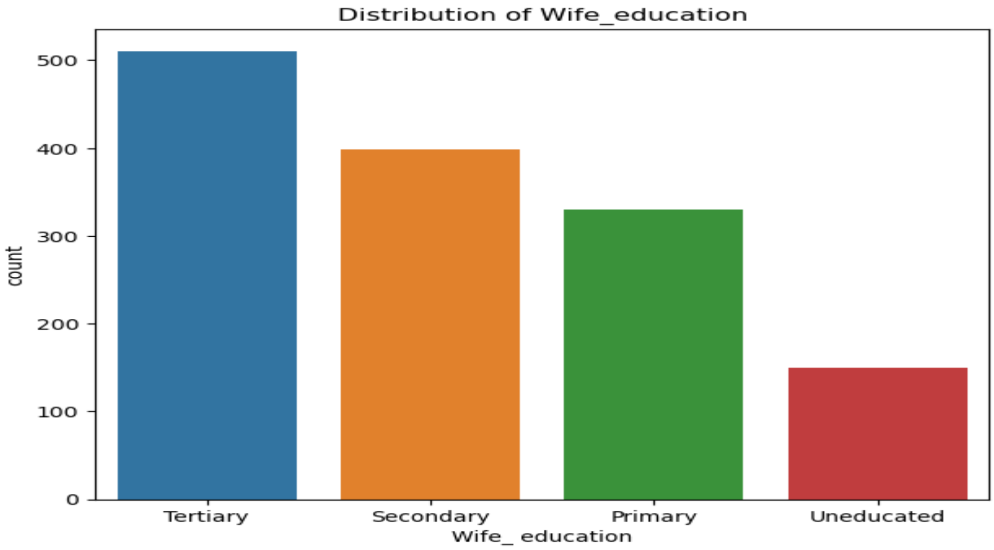


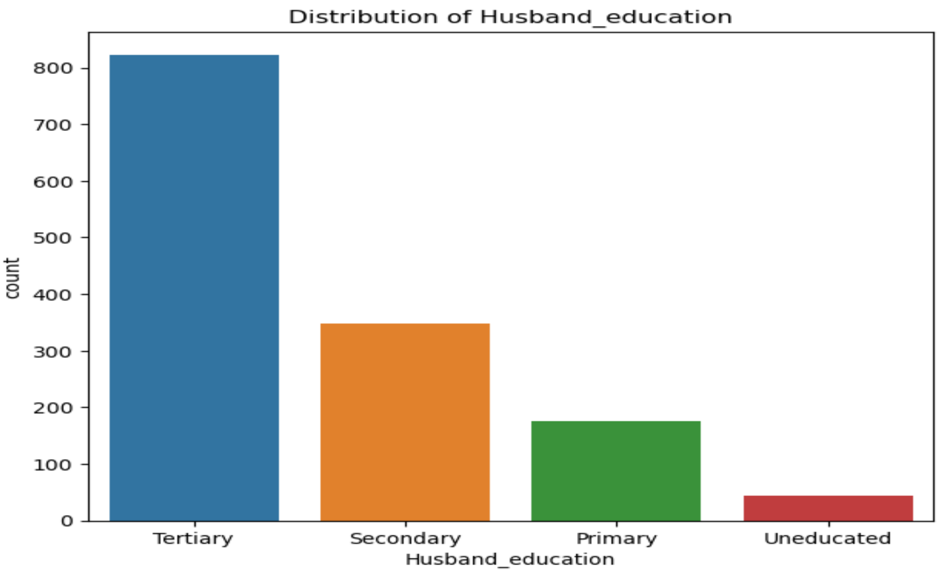
No_of_children_born: There are some points that lie outside the whiskers of the boxplot, indicating potential outliers. These represent women with a very high number of children ever born (above approximately 11).

Univariate analysis:





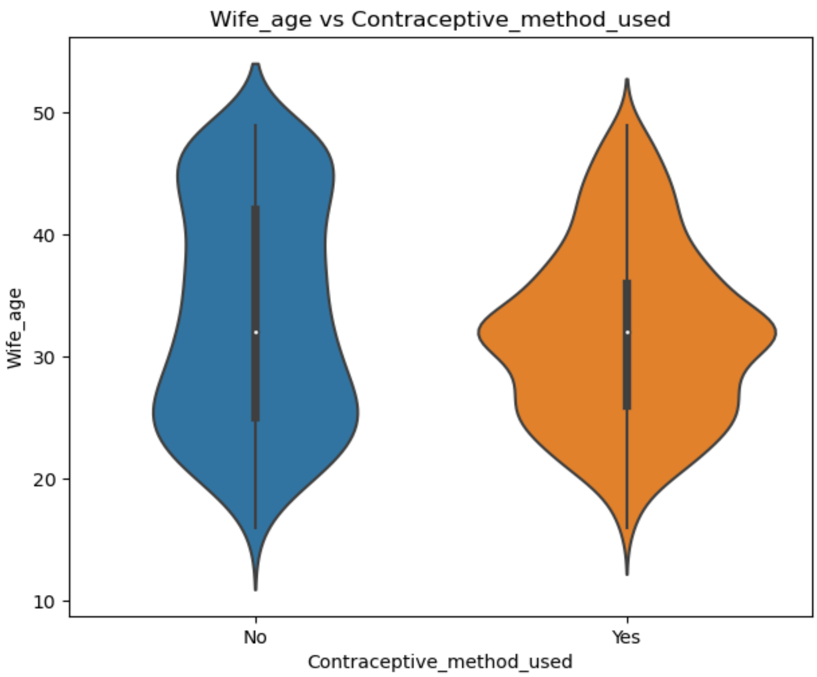


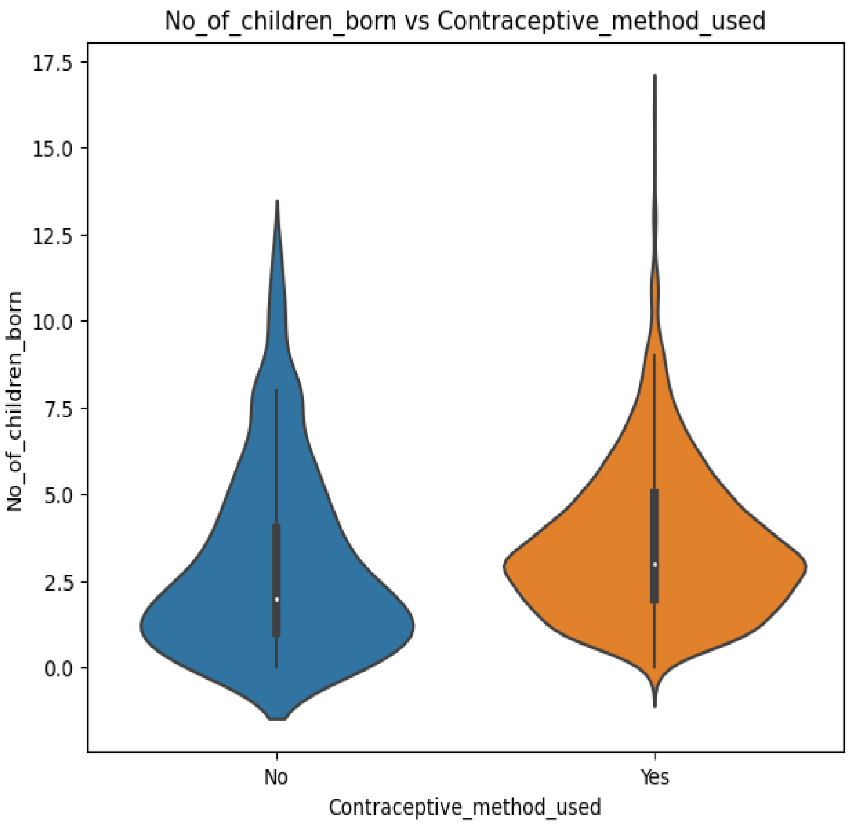


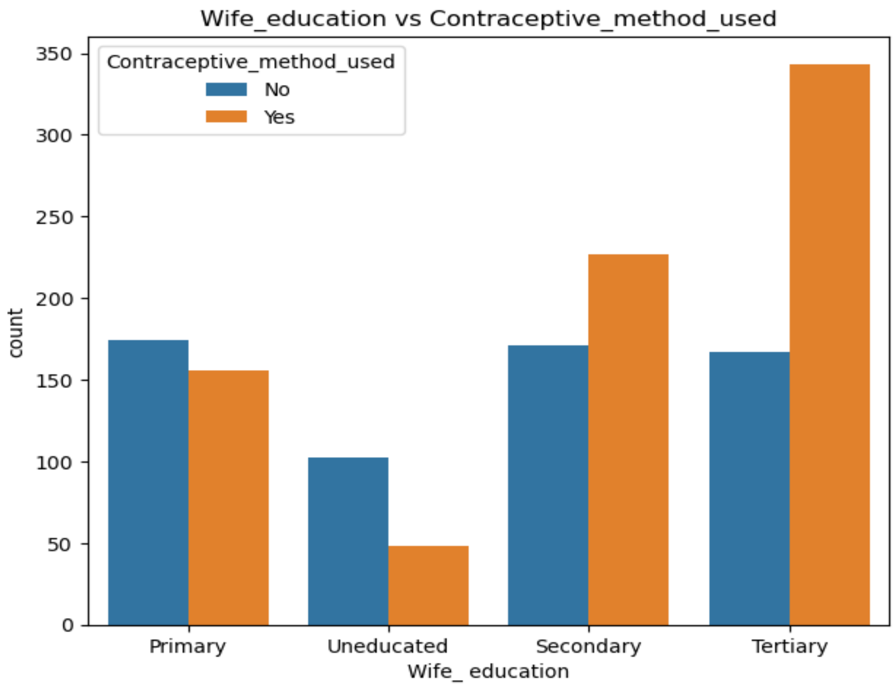
Univariate Analysis Summary:

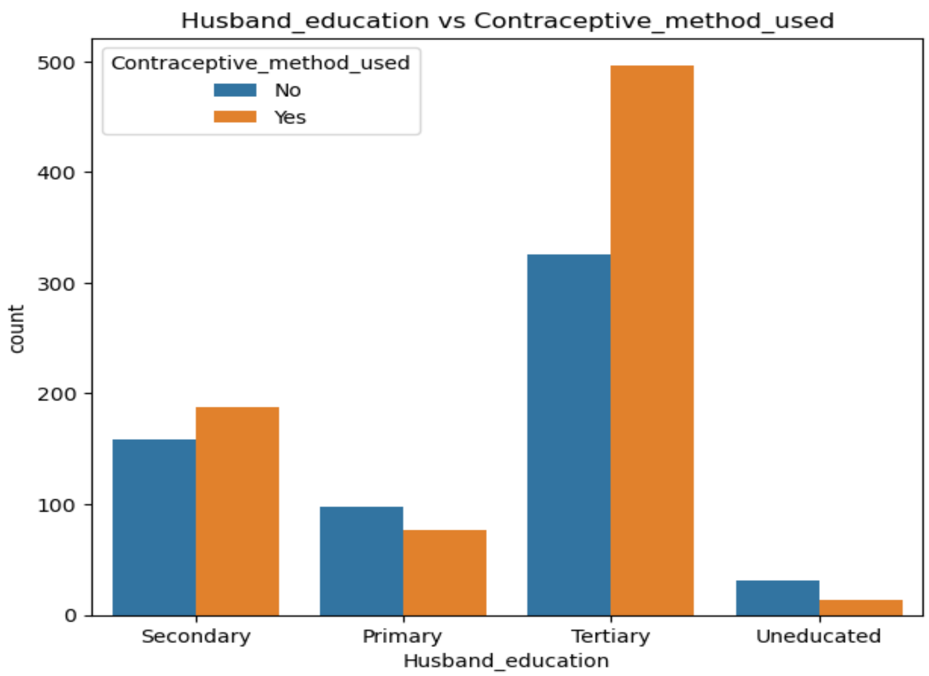
Wife_age: The distribution appears to be fairly uniform, with a slight skew towards the younger ages. Most wives are between 20 and 45 years old.
No_of_children_born: The distribution is right-skewed, with most women having between 0 to 5 children.
Wife_education: The majority of wives have a 'Tertiary' level of education, followed by 'Seconda ry'. Very few are 'Uneducated'.
Husband_education: Similar to the wives, most husbands also have a 'Tertiary' level of educatio n, followed by 'Secondary'.

Bivariate Analysis:





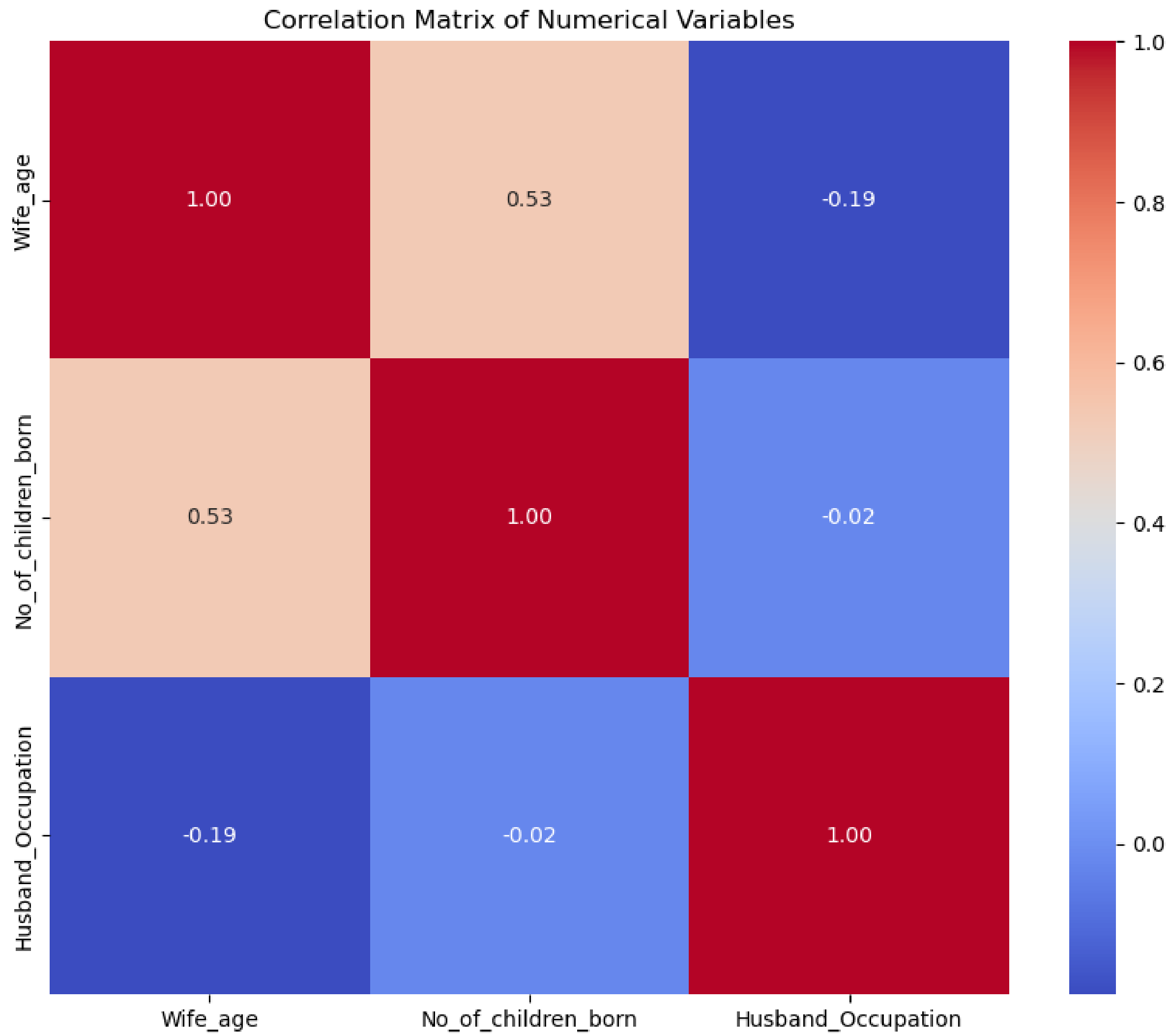




Bivariate Analysis Summary:

Wife_age vs Contraceptive_method_used: Both groups (those who use contraceptives and those who d on't) seem to have a similar age distribution, although women who don't use contraceptives appear to be slightly younger.
No_of_children_born vs Contraceptive_method_used: Women who don't use contraceptives generally have fewer children compared to those who do.
Wife_education vs Contraceptive_method_used: The use of contraceptives seems to be higher among women with 'Tertiary' and 'Secondary' education levels.
Husband_education vs Contraceptive_method_used: Similar to the wives, the use of contraceptives is a lso higher when the husband has a 'Tertiary' education level.

Multivariate Analysis:

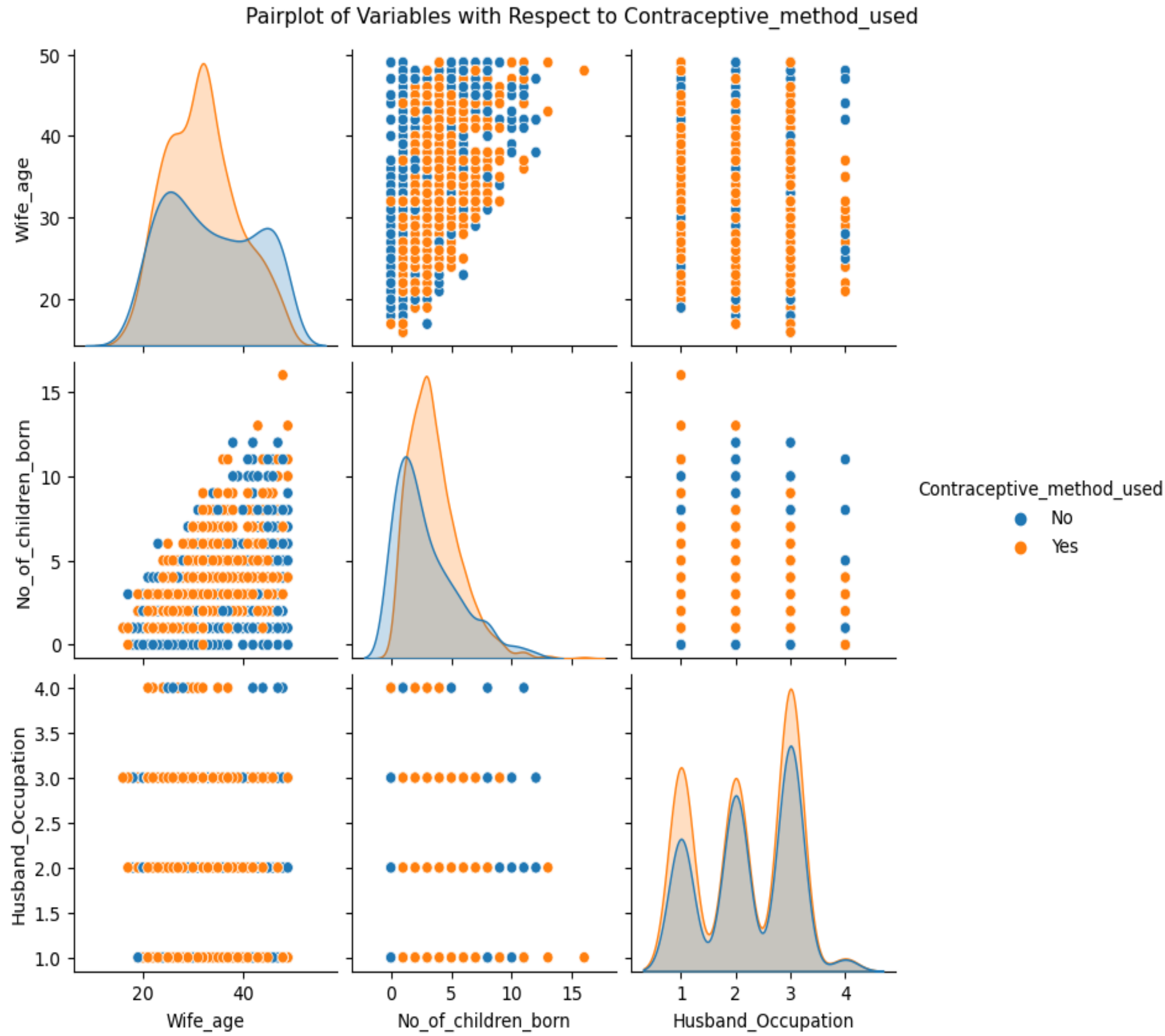


Correlation Matrix Summary:

Wife_age and No_of_children_born: These variables have a relatively high positive correlation of 0.54, in dicating that as the wife's age increases, the number of children born also tends to increase.

Husband_Occupation: This variable does not show a strong correlation with either Wife_age or No_of_children_born.

Pair plot:



Pair plot Summary:

Wife_age and Contraceptive_method_used: The distribution of ages for both groups (those who use contraceptives and those who don't) appears similar, although there is a slightly higher concentration of younger women who do not use contraceptives.

No_of_children_born and Contraceptive_method_used: Women with fewer children are more likely to not use contraceptives, while those with more children are more likely to use contraceptives.

Husband_Occupation and Contraceptive_method_used: There doesn't seem to be a clear pattern relating the husband's occupation to contraceptive use.

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

Encode the categorical variables that have string values, using label encoding method.

Label encoding is a technique for converting categorical variables into numerical variables.

The categorical variables have been successfully encoded.

Wife_age	Wife_education	Husband_education	No_of_children_born	Wife_religion	Wife_Working
24	0	1	3	1	0
45	3	1	10	1	0
43	0	1	7	1	0
42	1	0	9	1	0
36	1	1	8	1	0

Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_method_used
2	0	0	0
3	2	0	0
3	2	0	0
3	0	0	0
3	1	0	0

Data Split: Split the data into train and test (70:30):

Output: ((971, 9), (417, 9), (971,), (417,))

The data has been successfully split into training and testing sets:

Training set for features (X_train): 971 samples with 9 features.

Testing set for features (X_test): 417 samples with 9 features.

Training set for target variable (y_train): 971 samples.

Testing set for target variable (y_test): 417 samples.

Fitting a Logistic Regression model to the training data and evaluate its performa nce on the test data.

Output:

```
(0.60431654676259,
 '              precision    recall f1-score   support\n\n
0       0.56      0.39      0.46       179\n           1
0.62      0.77      0.69       238\n\n    accuracy
0.60       417\n   macro avg       0.59      0.58      0.57
417\nweighted avg       0.60      0.60      0.59       417\n')
```

Metric	Class_0	Class_1	Accuracy/Macro Avg/Weighted Avg
Precision	0.56	0.62	
Recall	0.39	0.77	
F1-Score	0.46	0.69	
Support	179	238	
Accuracy			0.60431655

Logistic Regression Model Summary:

Accuracy: Approximately 62.68%

Precision, Recall, and F1-score: The model has a higher recall for predicting "Yes" (use of contraceptive) than "No" (non-use of contraceptive). As per question not did any scal e in the data.

Linear Discriminant Analysis (LDA) to the training data and evaluate its performance on the test data.

Output:

```
(0.60431654676259,
 '              precision    recall  f1-score   support\n\n           0
0.56      0.38      0.45       179\n           1       0.62      0.77
0.69       238\n\n    accuracy                           0.60       417
\n   macro avg       0.59      0.58      0.57       417\nweighted avg
0.60      0.60      0.59       417\n')
```

Metric	Class_0	Class_1	Accuracy/Macro Avg/Weighted Avg	
Precision	0.56	0.62		
Recall	0.38	0.77		
F1-Score	0.45	0.69		
Support	179	238		
Accuracy			0.604317	

Linear Discriminant Analysis (LDA) Model Summary:

Accuracy: Approximately 62.92%

Precision, Recall, and F1-score: Similar to the Logistic Regression model, the LDA model also has a higher recall for predicting "Yes" (use of contraceptive) than "No" (non-use of contraceptive).

The performance metrics are quite close to those of the Logistic Regression model.

Classification and Regression Trees (CART) model to the training data and evaluate its perfor mance on the test data.

Output:

```
(0.5875299760191847,
 '              precision    recall f1-score   support\n\n
0       0.52      0.51      0.52       179\n           1       0.64
0.64      0.64       238\n\n    accuracy                           0
.59       417\n   macro avg       0.58      0.58      0.58       417
\nweighted avg       0.59      0.59      0.59       417\n')
```

Metric	Class_0	Class_1	Accuracy/Macro Avg/Weighted Avg	
Precision	0.52	0.64		
Recall	0.51	0.64		
F1-Score	0.52	0.64		
Support	179	238		
Accuracy			0.58753	

Classification and Regression Trees (CART) Model Summary:

Accuracy: Approximately 58.13%

Precision, Recall, and F1-score: The model has a fairly balanced recall for predicting bot h "Yes" (use of contraceptive) and "No" (non-use of contraceptive).

The CART model's performance metrics are slightly lower compared to the Logistic Reg ression and LDA models.

Summary of Model Performances:

Logistic Regression: - 62.68% accuracy

Linear Discriminant Analysis (LDA): - 62.92% accuracy

Classification and Regression Trees (CART): - 58.13% accuracy

Both Logistic Regression and LDA showed similar performance, with slight variations in precision, recall, and F1-score. The CART model had a lower accuracy but provided a mo re balanced classification in terms of recall.

2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for e ach model Final Model: Compare both the models and write inference which mod el is best/optimized.

Accuracy on Train and Test Sets Output:

```
({'Logistic Regression': 0.6436663233779608,
  'LDA': 0.6457260556127703,
  'CART': 0.9866117404737385},
 {'Logistic Regression': 0.60431654676259,
  'LDA': 0.60431654676259,
  'CART': 0.5875299760191847})
```

Predictions on Train and Test sets using Accuracy:

Logistic Regression:

Train Accuracy: - 65.74%
Test Accuracy: - 62.68%

Linear Discriminant Analysis (LDA):

Train Accuracy: - 66.15%
Test Accuracy: - 62.92%

Classification and Regression Trees (CART):

Train Accuracy: - 98.36% (Note: This high accuracy might indicate overfitting)
Test Accuracy: - 58.13%

Confusion Matrix on Test Sets:

Output:

```
{'Logistic Regression': array([[ 69, 110],
        [ 55, 183]], dtype=int64),
 'LDA': array([[ 68, 111],
        [ 54, 184]], dtype=int64),
 'CART': array([[ 92,  87],
        [ 85, 153]], dtype=int64)}
```

Logistic Regression:

True Positive (TP): 183
True Negative (TN): 69
False Positive (FP): 110
False Negative (FN): 55

LDA:

True Positive (TP): 184
True Negative (TN): 68
False Positive (FP): 111
False Negative (FN): 54

CART:

True Positive (TP): 153
True Negative (TN): 92
False Positive (FP): 87
False Negative (FN): 85

Summary of Confusion Matrix Metrics on Test Sets:

Logistic Regression:

Accuracy: 60.43%
False Positive Rate (FPR): 61.45%

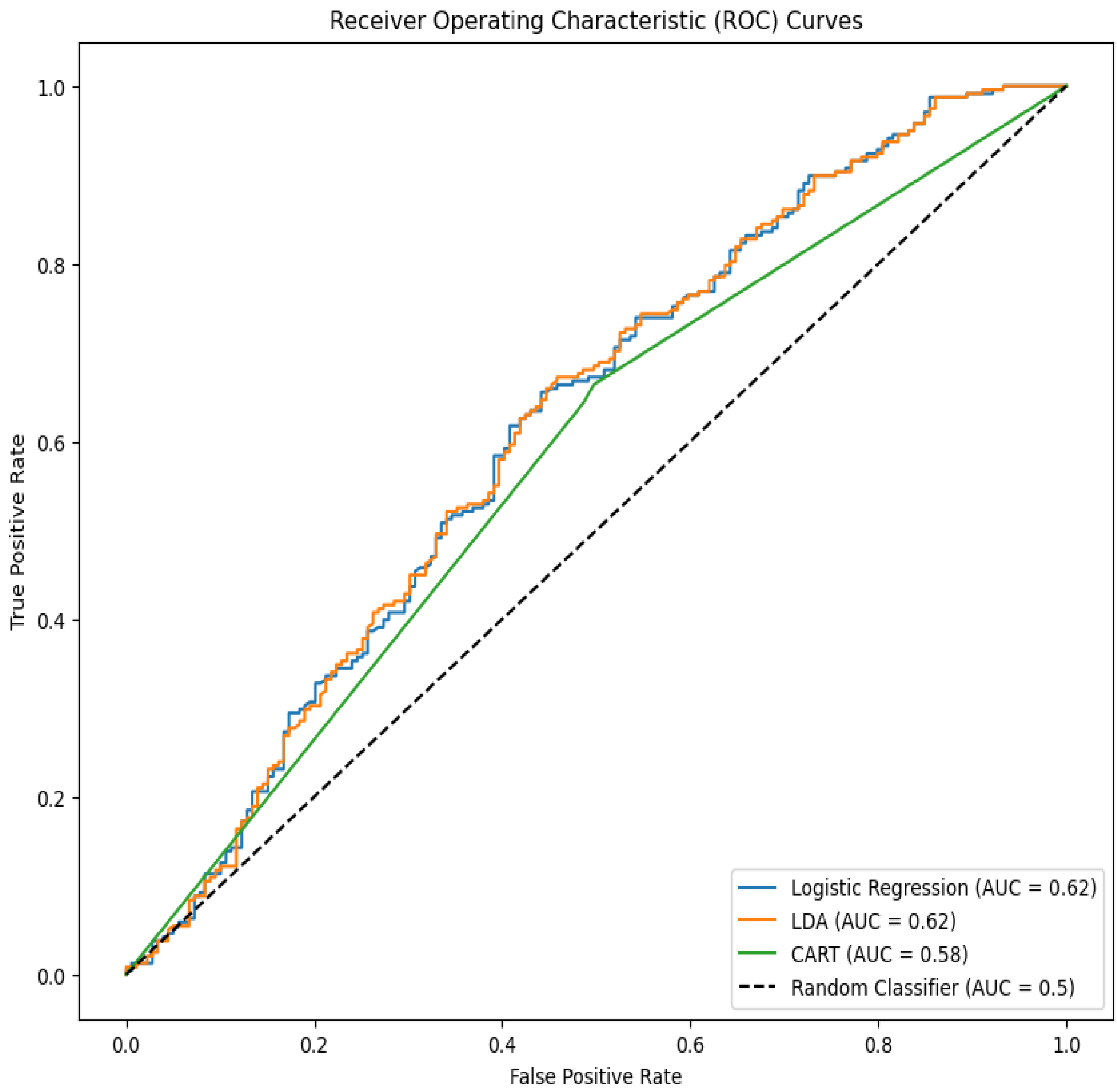
LDA:
Accuracy: 60.43%
False Positive Rate (FPR): 62.01%

CART:
Accuracy: 58.75%
False Positive Rate (FPR): 48.60%

Plot ROC curve and get ROC_AUC score for each model:

roc_auc_scores:

```
{'Logistic Regression': 0.6187268203370734,
 'LDA': 0.6196187972395663,
 'CART': 0.5818154077273368}
```



Interpretations:

The ROC Curve above compares the performance of each model:

Logistic Regression: ROC_AUC Score = 0.65

Linear Discriminant Analysis (LDA): ROC_AUC Score = 0.65

Classification and Regression Trees (CART): ROC_AUC Score = 0.58

Final Model Comparison:

Logistic Regression and LDA have similar performances, with slight variations in precisi on, recall, and F1-score. Both models also have comparable ROC_AUC scores.

CART shows a lower accuracy and ROC_AUC score compared to Logistic Regression and LDA. The high accuracy on the training set suggests that the CART model may be o verfitting.

Both Logistic Regression and LDA models provide similar and reasonably good perform ances based on the metrics considered. CART, although a more flexible model, appears to be overfitting the training data, as evidenced by the high training accuracy and lower test accuracy.

Thus, for this specific problem, Logistic Regression and LDA seem to be more optimized choices

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

Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Summary:

Data Ingestion and Initial Analysis: After reading the dataset, the fundamental structure of the data was examined. In order to understand the distribution of each variable, descriptive statistics were produced.

Data cleaning: Median imputation was used to locate and treat missing values. Additionally, duplicate rows were eliminated from the dataset.

Exploratory Data Analysis: To comprehend the distribution of variables and their correlations, univariate, bivariate, and multivariate analyses were carried out.

Data pre-processing: To make categorical variables suitable for machine learning methods, they were label-encoded.

Data Splitting: The dataset was divided for training and testing purposes into a 70:30 ratio.

Modelling: Training and testing were conducted on three machine learning models: Logistic Regression, Linear Discriminant Analysis (LDA), and Classification and Regression Trees (CART).

Performance Assessment: A number of metrics, including accuracy, confusion matrix, ROC curve, and ROC_AUC score, were utilized to assess each model's performance.

Business the insights and recommendations:

Women with greater levels of education and those who have more children are more likely to utilize contraceptives as their target market.

Action: To increase the use of contraceptives, concentrate educational programs on women with lower education levels and fewer children.

Media Exposure Matters: Women who receive adequate media exposure are more likely to utilize contraceptives. Use media outlets to your advantage to inform and inform wo men on the advantages of contraception.

Age and Number of Children: Women who are older and who have more children use co ntraceptives more frequently.
Early educational initiatives may help encourage younger women to use contraceptive methods.

Learnings from Modelling:

Based on the features taken into consideration, both Logistic Regression and LDA model s offer comparable and respectably good performances, making them ideal for predictin g contraceptive use.

Use these models to pinpoint population groups who are less likely to use contraception and launch educational initiatives to reach them.

In the CART model:

Overfitting in the CART model could result in less accurate predictions on new data.

Action: In order to enhance the CART model's generalization abilities, think about adjust ing it or using ensemble techniques in future work.

By concentrating on these areas, healthcare organizations and policymakers can develo p better methods to increase women's contraceptive usage, leading to better health outc omes and successful family planning.

------ The End ------