PM PROJECT BUSINESS REPORT

Contents:

List Of Tables:

List Of Figures:

Problem 1

Problem Definition

Context:

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.

We aim to establish a linear equation for predicting 'usr' (the percentage of time CPUs operate in user mode). Our goal is to analyze various system attributes to understand their influence on the system's 'usr' mode.

Objective:

The primary objective is to establish a linear equation that predicts 'usr' based on the given system measures. This involves:

Understanding the Data: Analyzing the various system measures provided in the dataset to understand their impact on 'usr'.

Handling Missing Data: Ensuring that missing values are properly handled to maintain data integrity.

Feature Selection: Identifying the most relevant system measures that significantly influence 'usr'.

Model Building: Developing a linear regression model to predict 'usr' using the selected features.

Model Evaluation: Evaluating the performance of the model using appropriate metrics to ensure its accuracy and reliability.

Model Interpretation: Interpreting the model coefficients to understand the impact of each feature on 'usr'.

By achieving these objectives, the goal is to provide a robust predictive model that helps in understanding and managing CPU utilization in user mode based on various system activities.

Data Description:

compactiv.xlsx : The data set database comprises activity measures of computer systems.

Data Dictionary:

- 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.

Data Overview:

Load the required packages, set the working directory, and load the data file.

The dataset has 8192 rows and 22 columns. It is always a good practice to view a sample of the rows. A simple way to do that is to use head() function.

lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap	usr
1	0	2147	79	68	0.2	0.2	40671.0	53995.0	0.0	...	0.0	0.0	1.6	2.6	16.00	26.40	CPU_Bound	4670	1730946	95
0	0	170	18	21	0.2	0.2	448.0	8385.0	0.0	...	0.0	0.0	0.0	0.0	15.63	16.83	Not_CPU_Bound	7278	1869002	97
15	3	2162	159	119	2.0	2.4	NaN	31950.0	0.0	...	0.0	1.2	6.0	9.4	150.20	220.20	Not_CPU_Bound	702	1021237	87
0	0	160	12	16	0.2	0.2	NaN	8670.0	0.0	...	0.0	0.0	0.2	0.2	15.60	16.80	Not_CPU_Bound	7248	1863704	98
5	1	330	39	38	0.4	0.4	NaN	12185.0	0.0	...	0.0	0.0	1.0	1.2	37.80	47.60	Not_CPU_Bound	633	1760253	90

Table 1:Top Five rows of dataset

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

Table 2: Basic Information of Dataset

A quick look at the dataset information tells us that there are 1 categorical and 21 numerical variables. There are few missing records present in three variables in rchar and wchar, which will be analyzed in detail in the next section. No duplicate rows were found.

Missing Value Treatment:

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

Table 3:Null Values in Dataframe

Found 104 values missing in rchar and 15 values in wchar.

After treating the missing values using mean.

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

Table 4:Null Values in Dataframe after calculation

Statistical Summary:

	count	mean	std	min	25%	50%	75%	max
lread	8192.0	1.955969e+01	53.353799	0.0	2.00	7.0	20.000	1845.00
lwrite	8192.0	1.310620e+01	29.891726	0.0	0.00	1.0	10.000	575.00
scall	8192.0	2.306318e+03	1633.617322	109.0	1012.00	2051.5	3317.250	12493.00
sread	8192.0	2.104800e+02	198.980146	6.0	86.00	166.0	279.000	5318.00
swrite	8192.0	1.500582e+02	160.478980	7.0	63.00	117.0	185.000	5456.00
fork	8192.0	1.884554e+00	2.479493	0.0	0.40	0.8	2.200	20.12
exec	8192.0	2.791998e+00	5.212456	0.0	0.20	1.2	2.800	59.56
rchar	8192.0	1.973857e+05	238310.037735	278.0	34860.50	127825.0	265394.750	2526649.00
wchar	8192.0	9.590299e+04	140712.688639	1498.0	22977.75	46653.0	106037.000	1801623.00
pgout	8192.0	2.285317e+00	5.307038	0.0	0.00	0.0	2.400	81.44
ppgout	8192.0	5.977229e+00	15.214590	0.0	0.00	0.0	4.200	184.20
pgfree	8192.0	1.191971e+01	32.363520	0.0	0.00	0.0	5.000	523.00
pgscan	8192.0	2.152685e+01	71.141340	0.0	0.00	0.0	0.000	1237.00
atch	8192.0	1.127505e+00	5.708347	0.0	0.00	0.0	0.600	211.58
pgin	8192.0	8.277960e+00	13.874978	0.0	0.60	2.8	9.765	141.20
ppgin	8192.0	1.238859e+01	22.281318	0.0	0.60	3.8	13.800	292.61
pflt	8192.0	1.097938e+02	114.419221	0.0	25.00	63.8	159.600	899.80
vflt	8192.0	1.853158e+02	191.000603	0.2	45.40	120.4	251.800	1365.00
freemem	8192.0	1.763456e+03	2482.104511	55.0	231.00	579.0	2002.250	12027.00
freeswap	8192.0	1.328126e+06	422019.426957	2.0	1042623.50	1289289.5	1730379.500	2243187.00
usr	8192.0	8.396887e+01	18.401905	0.0	81.00	89.0	94.000	99.00

Table 5: Numerical summarization of the dataframe

Uni-variate Analysis:

For performing Univariate analysis we will take a look at the Boxplots and Histograms to get better understanding of the distributions.

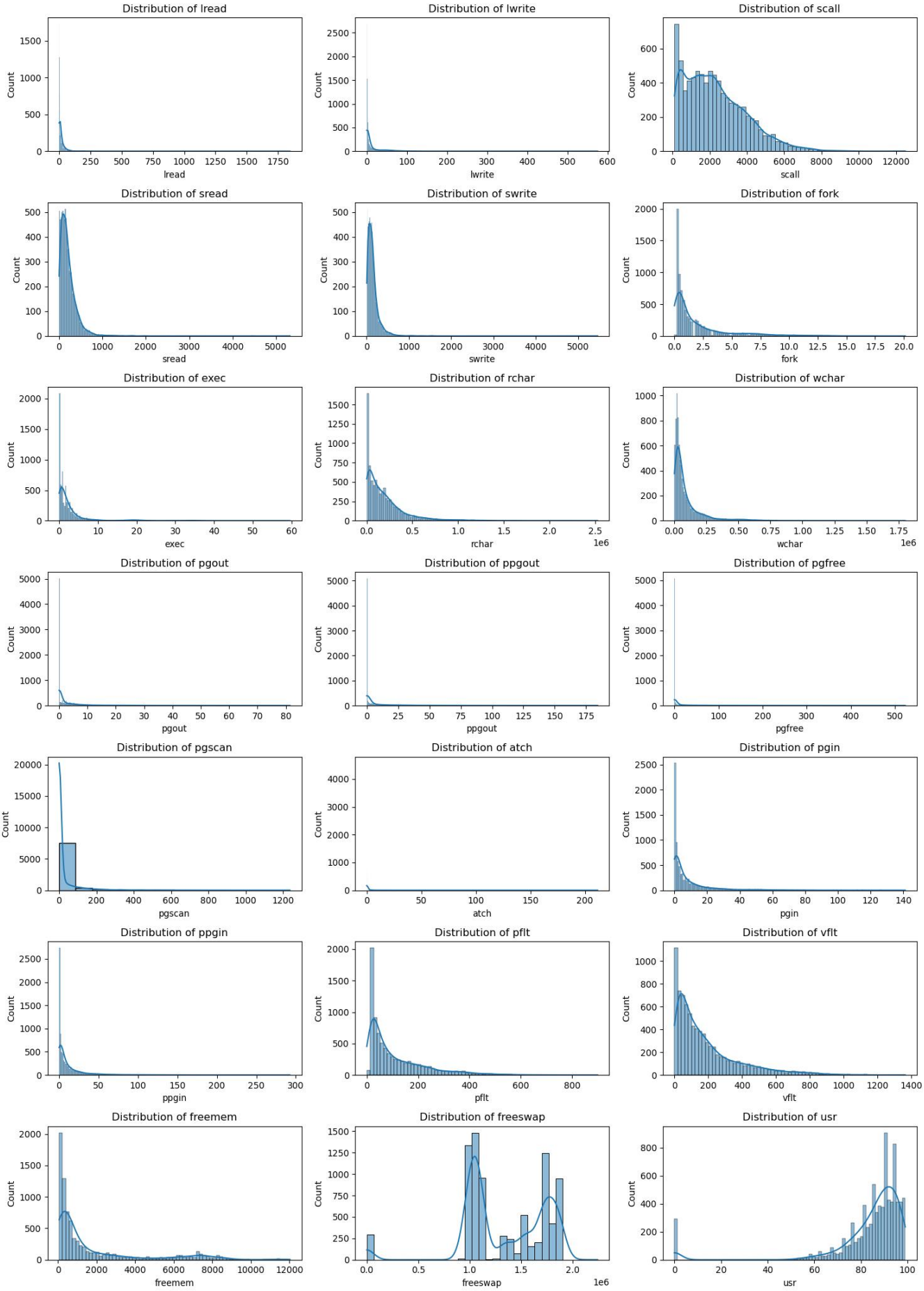

Figure 1 : Histogram of Numerical values

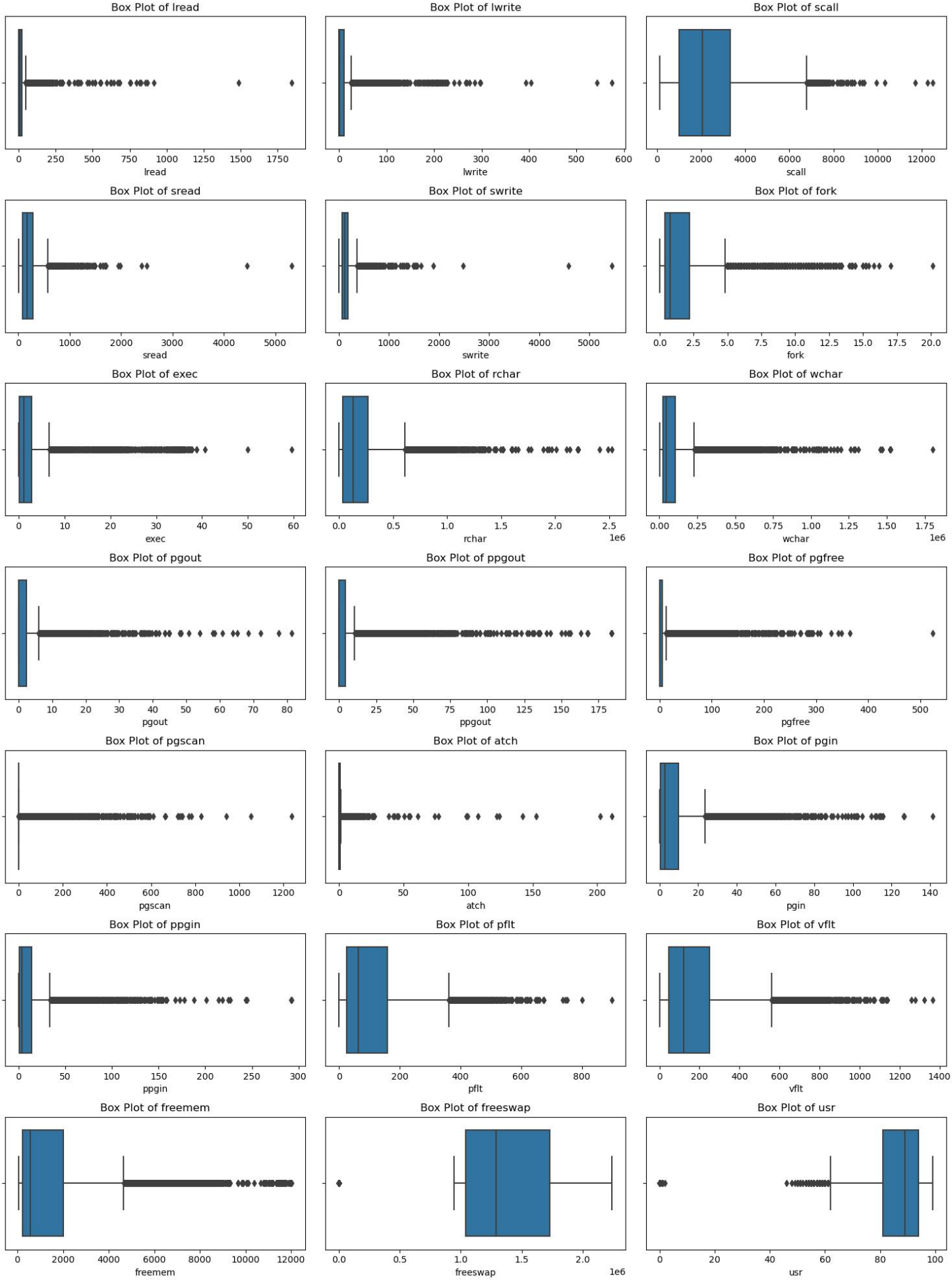

Figure 2 : Box plot of Numerical values

Observations :

- Variables like rchar, wchar, scall, sread, and swrite have very high maximum values, suggesting occasional spikes in activity.
- freemem and freeswap show that memory availability and disk blocks for swapping can vary widely, affecting system performance.
- Found outliers for all numerical variables.

Multivariate Analysis:

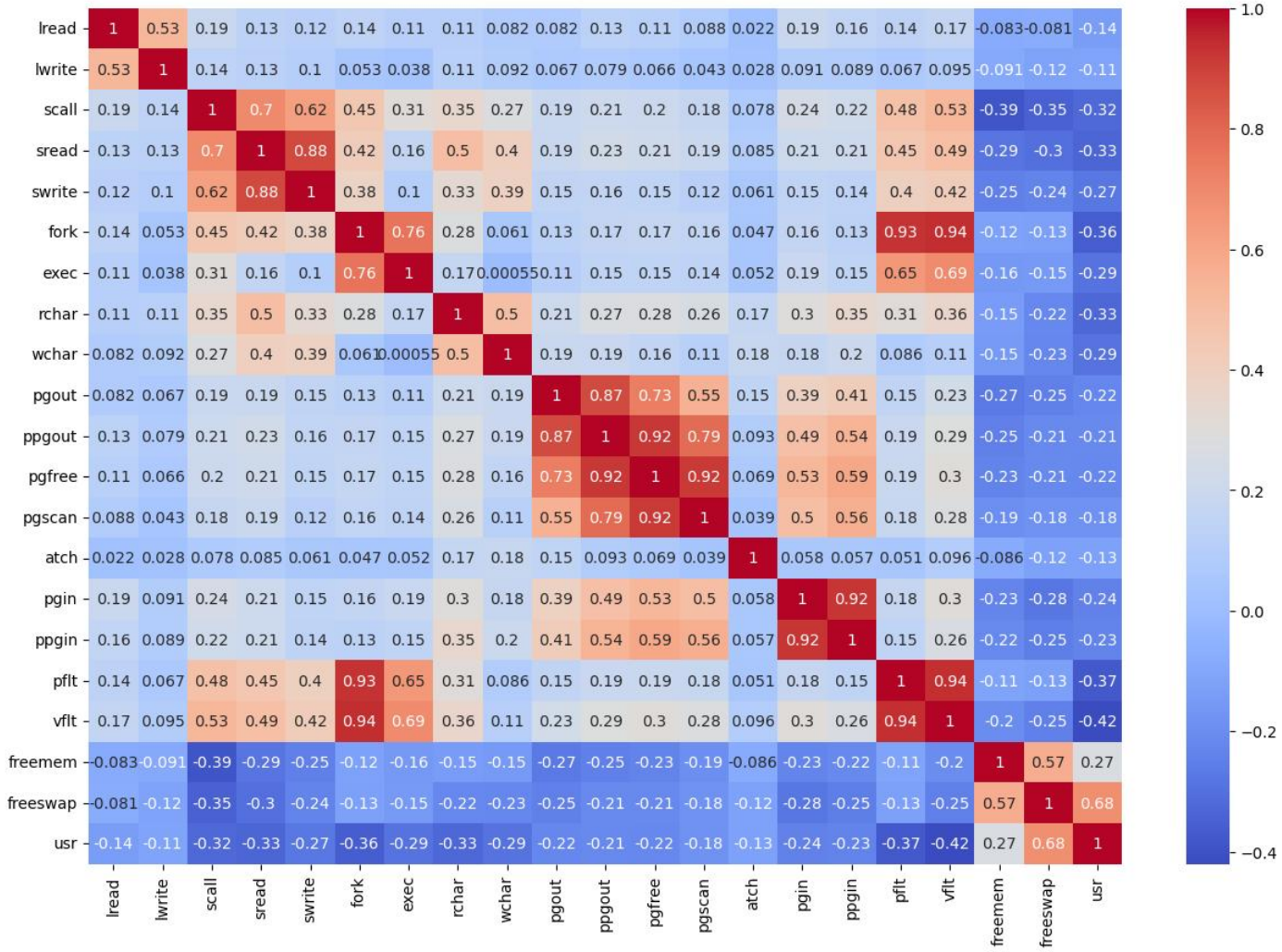

Figure 3 : Heat Map for Correlation

Observations :

- There High Positive Correlations:
  - sread and swrite: 0.881
  - fork and exec: 0.764
  - pflt and vflt: 0.935
  - pgfree and ppgout: 0.918
  - pgfree and pgscan: 0.915

- pgout and ppgout: 0.872
- pgout and pgfree: 0.730
- pgin and ppgin: 0.924
- usr and freeswap: 0.679

- High Negative Correlations:

  - usr with sread,swrite,fork,rchar,wchar,pgout,ppgout,pflt,vflt.
  - freemem with scall,sread,swrite,pgin,ppgin,pflt,vflt,freeswap.is a very strong positive correlation between available impressions and matched queries.

Data Pre-processing :

1. Missing Value Treatment:

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	

Table 6 : Missing values before treatment

Found 104 values missing in rchar and 15 values in wchar.

After treating the missing values using mean.

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	

Table 7:Null Values in Dataframe after calculation

2. Outlier Detection:

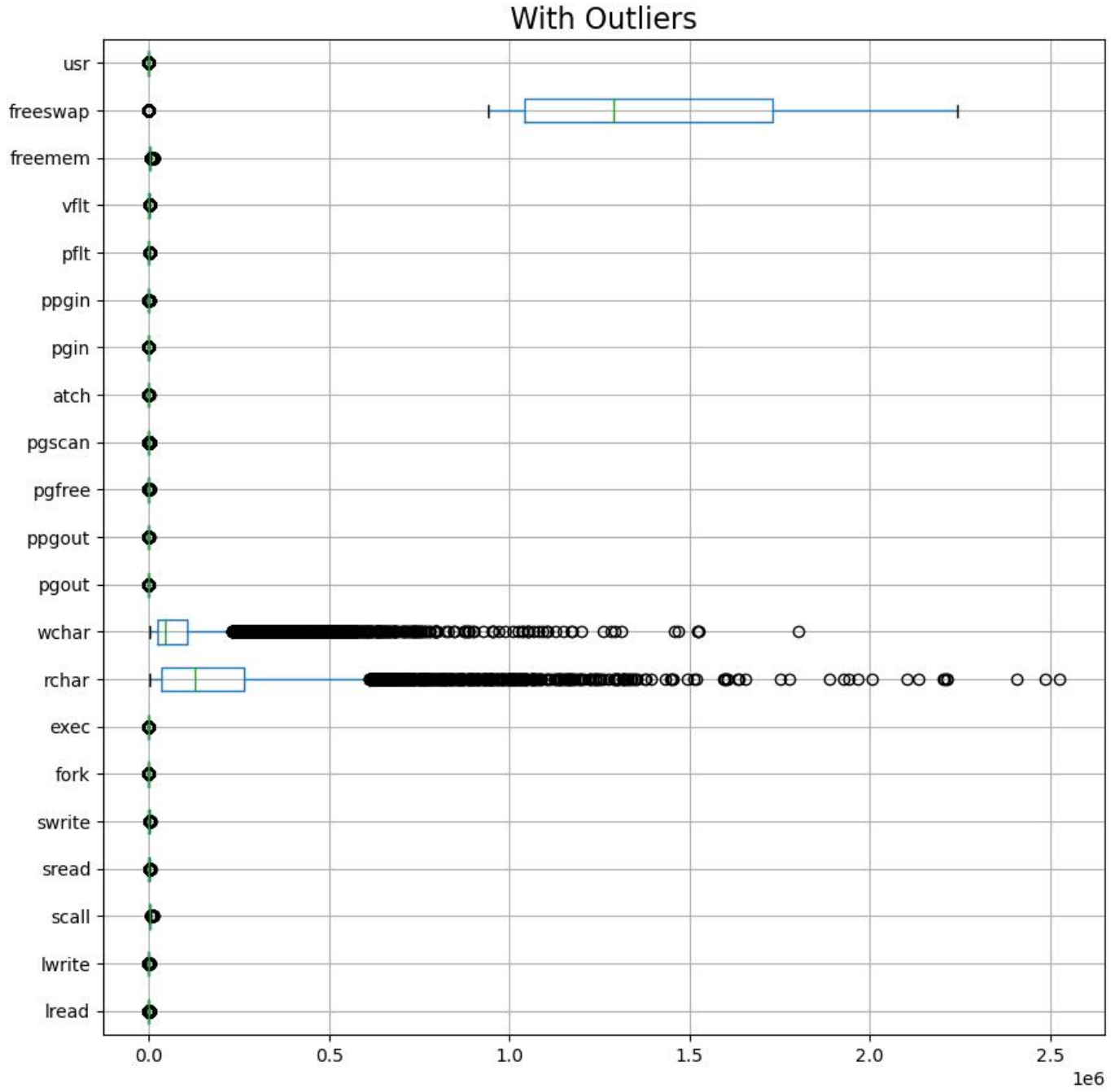

Figure 4 : Box Plot for Outliers check

- For each continuous variable, we calculated the first quartile (Q1) and third quartile (Q3), which represent the 25th and 75th percentiles of the data, respectively.

- We then determined the Interquartile Range (IQR) by subtracting Q1 from Q3. The IQR is a measure of statistical dispersion, or how spread out the data is.

- We calculated the lower boundary as Q1 minus 1.5 times the IQR and the upper boundary as Q3 plus 1.5 times the IQR. These boundaries are commonly used to identify potential outliers in the data.

- Values below the lower boundary were replaced with the lower boundary, and values above the upper boundary were replaced with the upper boundary.

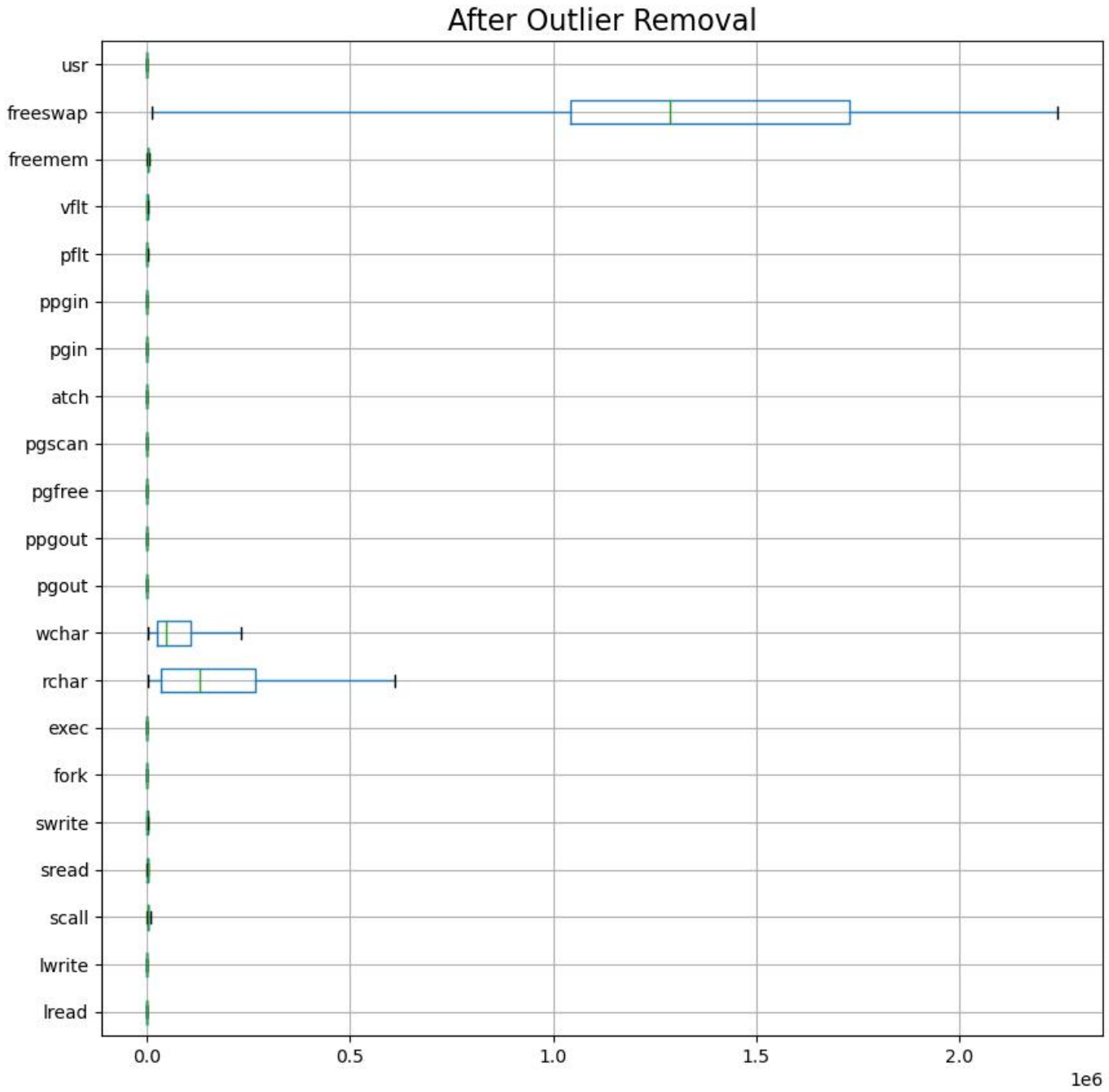

Figure 5 : Box Plot for Outliers check

The variable "pgscan" was excluded from the analysis because it contained zero values across all 8,192 observations, rendering it uninformative for the purposes of the study.

Train-test split:

- The dataset (df) is split into predictor variables (features) and the target variable (usr). The predictor variables are stored in the X dataframe, while the target variable (usr) is stored in the y dataframe.

- The data is then divided into training and testing sets using the “train_test_split” function from “sklearn.model_selection”.

- “X_train” and “y_train” contain 70% of the data and will be used to train the model.

- “X_test” and “y_test” contain 30% of the data and will be used to evaluate the model’s performance.

- The random_state parameter is set to 1 to ensure that the split is reproducible.

Model Building - Linear regression:

- The sm.add_constant(X_train) and sm.add_constant(X_test) functions are used to add a constant term (intercept) to both the training and test data. This constant term represents the bias term in a linear regression model.
- The sm.OLS(y_train, X_train) function from the statsmodels library is used to create an Ordinary Least Squares (OLS) regression model, where y_train is the target variable and X_train is the set of predictor variables.
- The .fit() method is called on the model to estimate the coefficients of the regression equation.

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.796
Model:	OLS	Adj. R-squared:	0.795
Method:	Least Squares	F-statistic:	1116.
Date:	Sun, 04 Aug 2024	Prob (F-statistic):	0.00
Time:	13:31:43	Log-Likelihood:	-16656.
No. Observations:	5734	AIC:	3.335e+04
Df Residuals:	5713	BIC:	3.349e+04
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.1314	0.316	266.122	0.000	83.512	84.751
lread	-0.0634	0.009	-7.064	0.000	-0.081	-0.046
lwrite	0.0480	0.013	3.660	0.000	0.022	0.074
scall	-0.0007	6.28e-05	-10.576	0.000	-0.001	-0.001
sread	0.0003	0.001	0.336	0.737	-0.002	0.002
swrite	-0.0055	0.001	-3.805	0.000	-0.008	-0.003
fork	0.0296	0.132	0.225	0.822	-0.229	0.288
exec	-0.3211	0.052	-6.219	0.000	-0.422	-0.220
rchar	-5.212e-06	4.87e-07	-10.696	0.000	-6.17e-06	-4.26e-06
wchar	-5.346e-06	1.03e-06	-5.179	0.000	-7.37e-06	-3.32e-06
pgout	-0.3669	0.090	-4.077	0.000	-0.543	-0.190
ppgout	-0.0786	0.079	-0.999	0.318	-0.233	0.076
pgfree	0.0853	0.048	1.786	0.074	-0.008	0.179
atch	0.6304	0.143	4.414	0.000	0.350	0.910
pgin	0.0198	0.028	0.695	0.487	-0.036	0.076
ppgin	-0.0672	0.020	-3.406	0.001	-0.106	-0.029
pflt	-0.0336	0.002	-16.954	0.000	-0.037	-0.030
vflt	-0.0055	0.001	-3.831	0.000	-0.008	-0.003
freemem	-0.0005	5.07e-05	-9.022	0.000	-0.001	-0.000
freeswap	8.829e-06	1.9e-07	46.463	0.000	8.46e-06	9.2e-06
runqsz_Not_CPU_Bound	1.6137	0.126	12.807	0.000	1.367	1.861

Omnibus:	1102.551	Durbin-Watson:	2.016
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2367.549
Skew:	-1.118	Prob(JB):	0.00
Kurtosis:	5.216	Cond. No.	7.74e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.74e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 8: Summary of the regression model

The R-squared value of 0.796 indicates that approximately 79.6% of the variability in the target variable (usr) is explained by the predictor variables included in the model. The Adjusted R-squared, which accounts for the number of predictors, is very close at 0.795, indicating a strong overall model fit.

The condition number of the model is 7.74e+06, which is relatively large. This might indicate the presence of multicollinearity, meaning that some of the predictor variables are highly correlated with each other.

1) **VIF of the predictors:**

VIF values:

const	29.239441
lread	5.350218
lwrite	4.328116
scall	2.960756
sread	6.415575
swrite	5.594100
fork	13.035282
exec	3.241124
rchar	2.129470
wchar	1.583532
pgout	11.359771
ppgout	29.404123
pgfree	16.497072
atch	1.876238
pgin	13.809962
ppgin	13.951564
pflt	12.001532
vflt	15.968862
freemem	1.961657
freeswap	1.841358
runqsz_Not_CPU_Bound	1.157096

dtype: float64

Table 9: VIF of the predictors

As few predictors have VIF values > 3 therefore there is some multicolinearity in the data.

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

After removing lwrite,sread,swrite,fork,pgfree,pgin,vflt variables from the predictors, we got VIF values below 3.

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.794
Model:	OLS	Adj. R-squared:	0.793
Method:	Least Squares	F-statistic:	1835.
Date:	Sun, 04 Aug 2024	Prob (F-statistic):	0.00
Time:	13:31:45	Log-Likelihood:	-16690.
No. Observations:	5734	AIC:	3.341e+04
Df Residuals:	5721	BIC:	3.349e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	83.8573	0.308	272.402	0.000	83.254	84.461
lread	-0.0367	0.004	-8.252	0.000	-0.045	-0.028
scall	-0.0009	4.84e-05	-17.948	0.000	-0.001	-0.001
exec	-0.2882	0.046	-6.247	0.000	-0.379	-0.198
rchar	-5.429e-06	4.37e-07	-12.436	0.000	-6.28e-06	-4.57e-06
wchar	-6.11e-06	9.79e-07	-6.243	0.000	-8.03e-06	-4.19e-06
pgout	-0.3457	0.038	-9.008	0.000	-0.421	-0.270
atch	0.6211	0.143	4.345	0.000	0.341	0.901
ppgin	-0.0651	0.006	-10.074	0.000	-0.078	-0.052
pflt	-0.0429	0.001	-44.743	0.000	-0.045	-0.041
freemem	-0.0004	5.05e-05	-8.660	0.000	-0.001	-0.000
freeswap	9.001e-06	1.86e-07	48.339	0.000	8.64e-06	9.37e-06
runqsz_Not_CPU_Bound	1.6356	0.126	12.959	0.000	1.388	1.883

Omnibus:	999.129	Durbin-Watson:	2.011
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2053.646
Skew:	-1.039	Prob(JB):	0.00
Kurtosis:	5.068	Cond. No.	7.50e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.5e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 10: Summary of the regression model after removing variables

VIF values:

const	27.444949
lread	1.299323
scall	1.736362
exec	2.562518
rchar	1.692081
wchar	1.408728
pgout	2.044806
atch	1.859928
ppgin	1.483532
pflt	2.778197
freemem	1.924179
freeswap	1.750119
runqsz_Not_CPU_Bound	1.149187

dtype: float64

Table 11: VIF of predictors below 3

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

Testing the Assumptions of Linear Regression

Before relying on the results of a linear regression model, it's crucial to ensure that certain assumptions hold true. These assumptions include linearity, independence, homoscedasticity, normality of error terms, and the absence of multicollinearity. Below is a detailed explanation of the code used to test these assumptions.

1. Linearity and Independence of Predictors:

- A new DataFrame (df_pred) that contains the actual target values (Actual Values), the predicted values from the linear regression model (Fitted Values), and the residuals (Residuals), which are the differences between the actual and predicted values.
- This data is used to analyze the model's performance and check assumptions. Residuals should ideally be randomly distributed with no discernible pattern, which would indicate that the model assumptions are met.

	Actual Values	Fitted Values	Residuals
0	91.0	91.353579	-0.353579
1	94.0	91.736987	2.263013
2	61.5	74.662706	-13.162706
3	83.0	80.659585	2.340415
4	94.0	97.764015	-3.764015

Table 12: Top 5 rows of df_pred

2. Residual Plot:

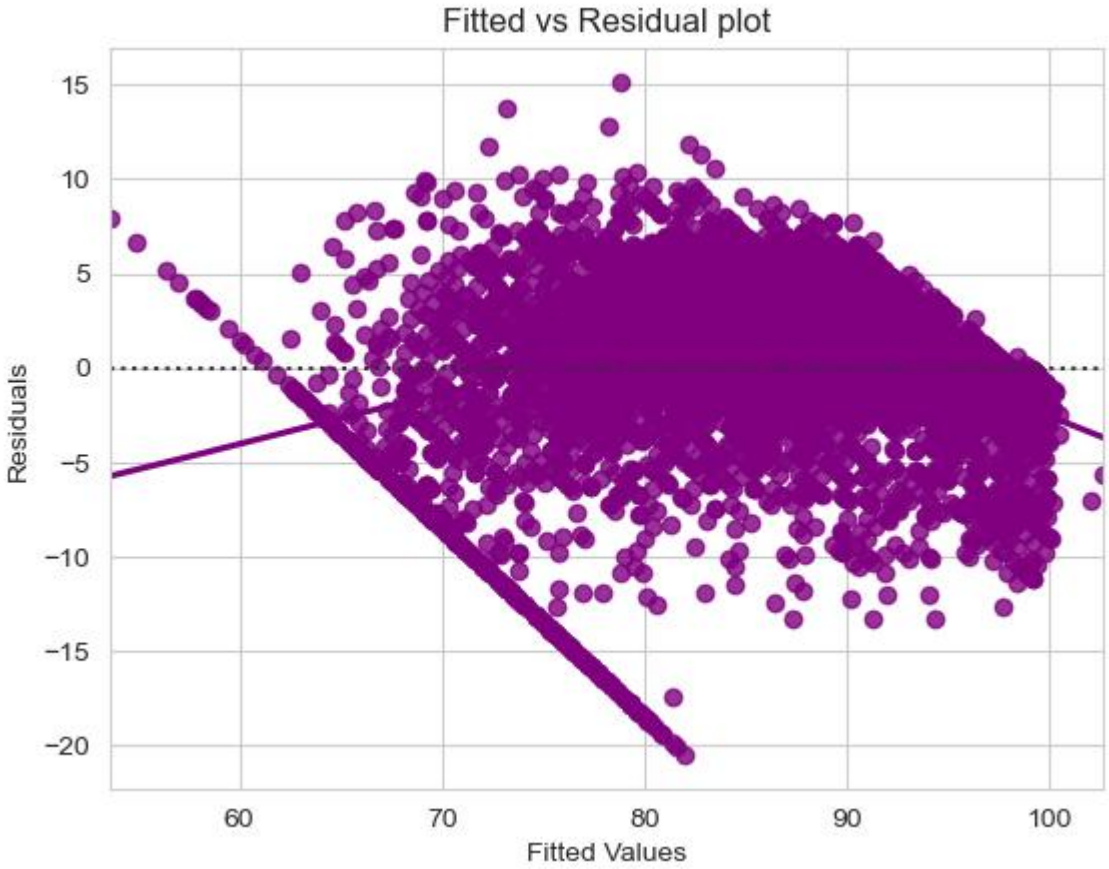

Figure 6 : Fitted vs Residual plot

The residual plot visualizes the relationship between the fitted values (predictions) and the residuals.

3. Test for Normality of Error Terms:

Shapiro-Wilk test and Homoscedasticity test was conducted on df_pred.

The Shapiro-Wilk test checks whether the residuals follow a normal distribution. The null hypothesis is that the residuals are normally distributed.

A p-value(7.862685683326549e-42) greater than 0.05 suggests that the residuals are normally distributed, supporting the assumption of normality.

The low p-value(0.0017459395679956555) indicates that there is evidence of heteroscedasticity in the residuals. This means that the variance of residuals changes with the level of the independent variables, which violates the homoscedasticity assumption of linear regression.

LINEAR REGRESSION EQUATION:

OLS Regression Results

Dep. Variable:	usr	**R-squared:**	0.794
Model:	OLS	**Adj. R-squared:**	0.793
Method:	Least Squares	**F-statistic:**	1835.
Date:	Sun, 04 Aug 2024	**Prob (F-statistic):**	0.00
Time:	13:31:56	**Log-Likelihood:**	-16690.
No. Observations:	5734	**AIC:**	3.341e+04
Df Residuals:	5721	**BIC:**	3.349e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	83.8573	0.308	272.402	0.000	83.254	84.461
lread	-0.0367	0.004	-8.252	0.000	-0.045	-0.028
scall	-0.0009	4.84e-05	-17.948	0.000	-0.001	-0.001
exec	-0.2882	0.046	-6.247	0.000	-0.379	-0.198
rchar	-5.429e-06	4.37e-07	-12.436	0.000	-6.28e-06	-4.57e-06
wchar	-6.11e-06	9.79e-07	-6.243	0.000	-8.03e-06	-4.19e-06
pgout	-0.3457	0.038	-9.008	0.000	-0.421	-0.270
atch	0.6211	0.143	4.345	0.000	0.341	0.901
ppgin	-0.0651	0.006	-10.074	0.000	-0.078	-0.052
pflt	-0.0429	0.001	-44.743	0.000	-0.045	-0.041
freemem	-0.0004	5.05e-05	-8.660	0.000	-0.001	-0.000
freeswap	9.001e-06	1.86e-07	48.339	0.000	8.64e-06	9.37e-06
runqsz_Not_CPU_Bound	1.6356	0.126	12.959	0.000	1.388	1.883

Omnibus:	999.129	**Durbin-Watson:**	2.011
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2053.646
Skew:	-1.039	**Prob(JB):**	0.00
Kurtosis:	5.068	**Cond. No.**	7.50e+06

Table 13: OLS REGRESSION RESULT

log_price = 83.85732586544668 + -0.036690269583115435 * (lread) + -0.0008678489465544963 * (scall) + -0.2882180586865333 * (exec) + -5.428931310443116e-06 * (rchar) + -6.110130803347406e-06 * (wchar) + -0.3456626619874853 * (pgout) + 0.6210731289413629 * (atch) + -0.06509379773683682 * (ppgin) + -0.04287084853575193 * (pflt) + -0.0004372778354186676 * (freemem) + 9.000993477470678e-06 * (freeswap) + 1.6355878445677583 * (runqsz_Not_CPU_Bound)

TAKEAWAYS:

- A positive coefficient indicates that an increase in the atch value is associated with an increase in the usr value. Specifically, for each unit increase in atch, the usr value is expected to increase by approximately 0.6211 units. This suggests that atch has a significant positive impact on usr, meaning that higher atch values contribute positively to the dependent variable.

- A negative coefficient signifies that an increase in pgout is associated with a decrease in usr. Specifically, for each unit increase in pgout, the usr value decreases by approximately 0.3457 units. This suggests that pgout has a significant negative impact on usr.

Conclusions and Recommendations:

- Since “atch” positively impacts “usr”, strategies to enhance this variable should be considered. This could involve optimizing processes or resources linked to “atch”.
- Given its negative impact, investigate ways to reduce “pgout”. This might involve improving system efficiency or managing resources better to minimize page output.
- Utilize the insights from the regression analysis to make informed decisions. Enhance metrics with positive coefficients while addressing issues related to those with negative coefficients.

Problem 2

Problem Definition

Context:

The Republic of Indonesia Ministry of Health is conducting an analysis on contraceptive use among married women. The dataset includes information on various demographic and socio-economic factors that may influence a woman's decision to use contraceptive methods. Understanding these factors is crucial for formulating effective public health policies and educational programs to increase the prevalence of contraceptive use, which can have significant implications for family planning and population control.

Objective:

The primary objective of this analysis is to develop a predictive model that can accurately determine whether a married woman in Indonesia is likely to use a contraceptive method based on her demographic and socio-economic attributes. By analyzing the data, we aim to identify key factors that influence contraceptive use and provide actionable insights for policymakers to design targeted interventions.

The model will be trained using the following features:

Wife's age (numerical)
Wife's education level (categorical: uneducated, tertiary)
Husband's education level (categorical: uneducated, tertiary)
Number of children ever born (numerical)
Wife's religion (binary: Non-Scientology, Scientology)
Wife's employment status (binary: Yes, No)
Husband's occupation (categorical)
Standard of living index (categorical: very low, high)
Media exposure (binary: Good, Not good)
The target variable is:

Contraceptive method used (binary: No, Yes)
The successful model will help in predicting contraceptive use and could potentially be used to guide policy decisions and health programs aimed at improving family planning services in Indonesia.

Data Description:

Contraceptive_method_dataset.xlsx : The data set database comprises activity measures of computer systems.

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=very low, 2, 3, 4=high
9. Media exposure (binary) Good, Not good
10. Contraceptive method used (class attribute) No,Yes

Data Overview:

Load the required packages, set the working directory, and load the data file.

The dataset has 1473 rows and 10 columns. It is always a good practice to view a sample of the rows. A simple way to do that is to use head() function.

Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_
Primary	Secondary	3.0	Scientology	No	2	High	Exposed	
Uneducated	Secondary	10.0	Scientology	No	3	Very High	Exposed	
Primary	Secondary	7.0	Scientology	No	3	Very High	Exposed	
Secondary	Primary	9.0	Scientology	No	3	High	Exposed	
Secondary	Secondary	8.0	Scientology	No	3	Low	Exposed	

Table 14: Top Five rows of dataset

```
<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
```

Table 15: Basic Information of Dataset

A quick look at the dataset information tells us that there are 7 categorical and 3 numerical variables. There are 71 missing values are present in "wife_age" and 21 in "no_of_children_ born", which will be analyzed in detail in the next section.

Statistical Summary:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Wife_age	1402.0	NaN	NaN	NaN	32.606277	8.274927	16.0	26.0	32.0	39.0	49.0
Wife_ education	1473	4	Tertiary	577	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_education	1473	4	Tertiary	899	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_of_children_born	1452.0	NaN	NaN	NaN	3.254132	2.365212	0.0	1.0	3.0	4.0	16.0
Wife_religion	1473	2	Scientology	1253	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Wife_Working	1473	2	No	1104	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_Occupation	1473.0	NaN	NaN	NaN	2.137814	0.864857	1.0	1.0	2.0	3.0	4.0
Standard_of_living_index	1473	4	Very High	684	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Media_exposure	1473	2	Exposed	1364	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Contraceptive_method_used	1473	2	Yes	844	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 16： Numerical summarization of the dataframe

- The high levels of education among both wives and husbands suggest that educational p rograms and information dissemination about contraceptive options might be effective.
- Dominant presence of Scientology, it may be useful to develop culturally sensitive educ ational materials and outreach strategies that respect and address religious beliefs.
- Many women are not working could influence their access to healthcare and contracept ive services.

Uni-variate Analysis:

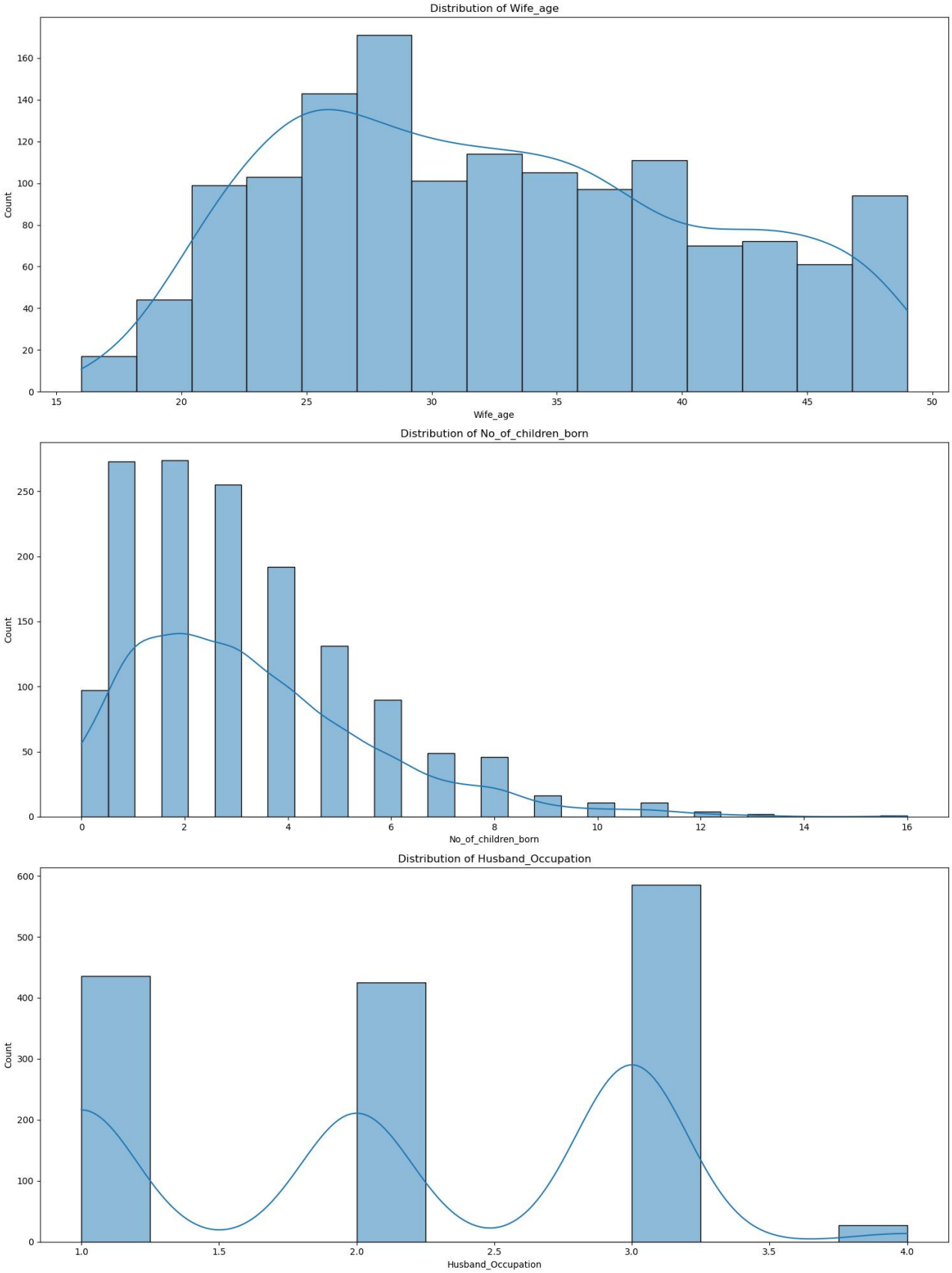

Figure 7 : Histogram of Numerical variables

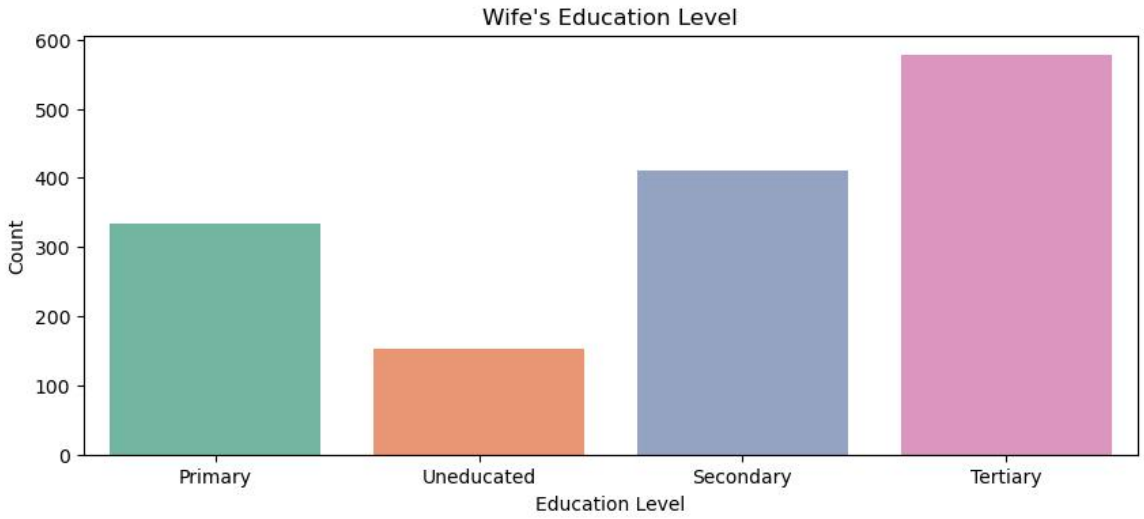

Figure 8 : Wife's Education Level

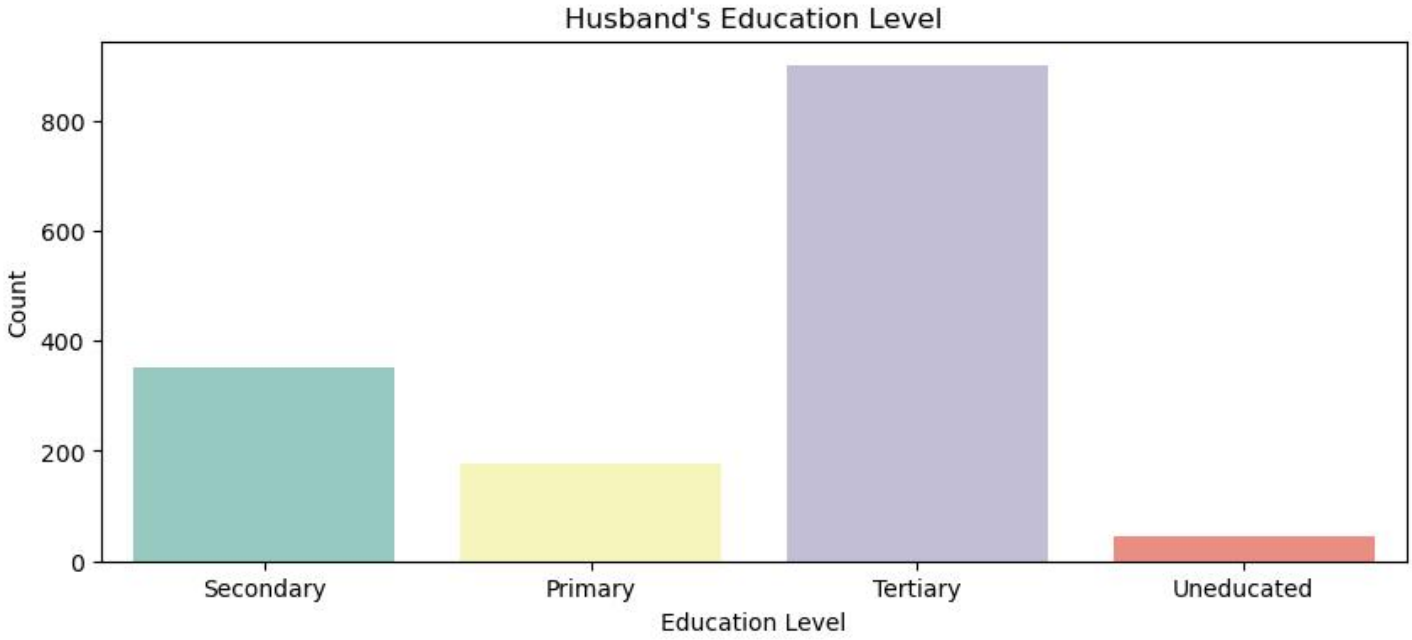

Figure 9 : Husband's Education Level

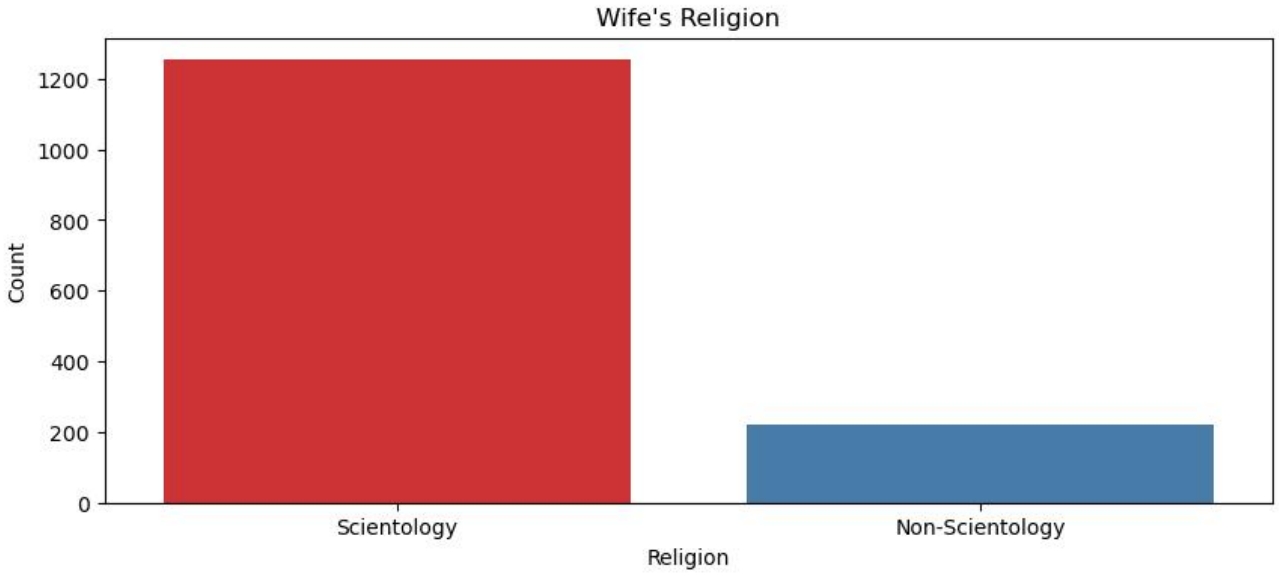

Figure 10 : Wife's Religion

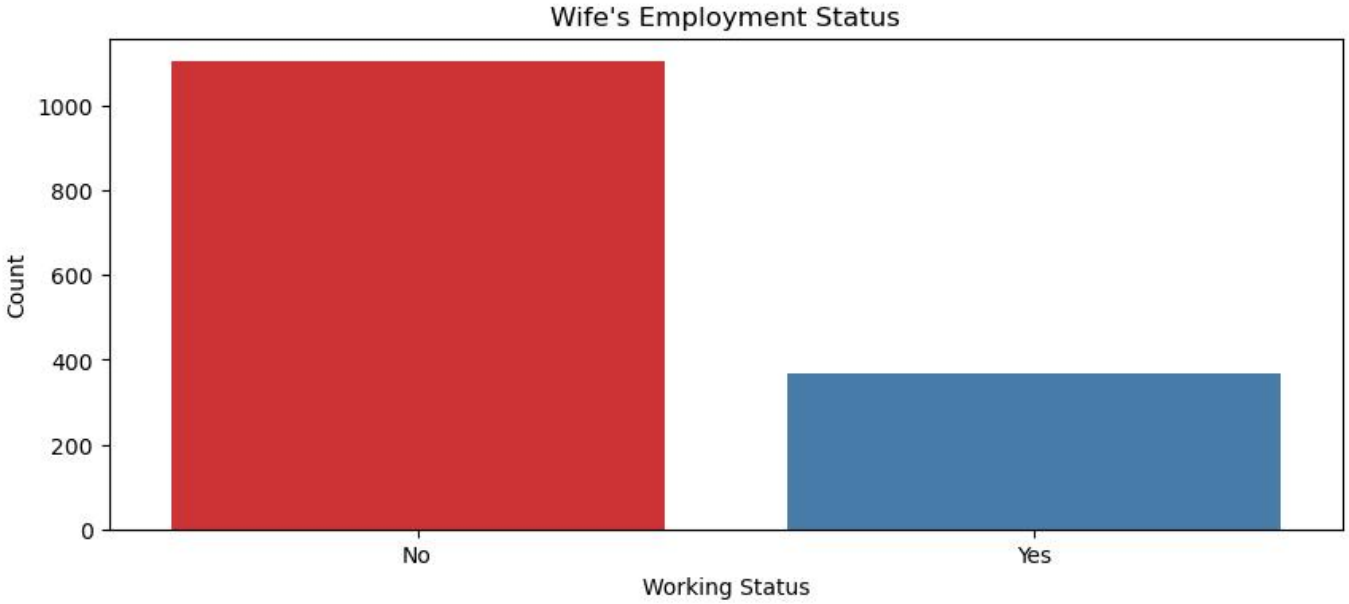

Figure 11 : Wife's Employment Status

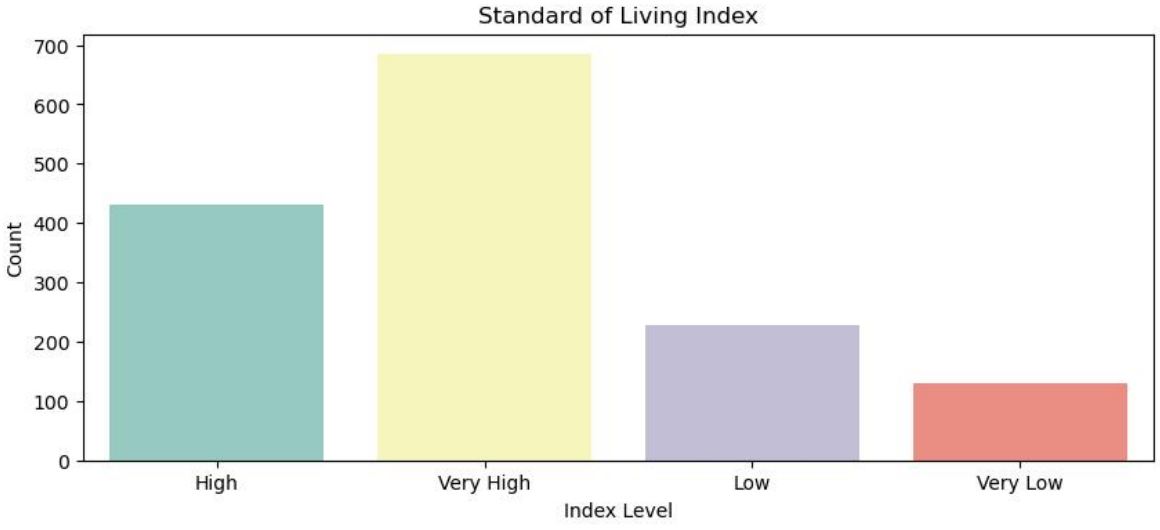

Figure 12 : Standard of Living Index

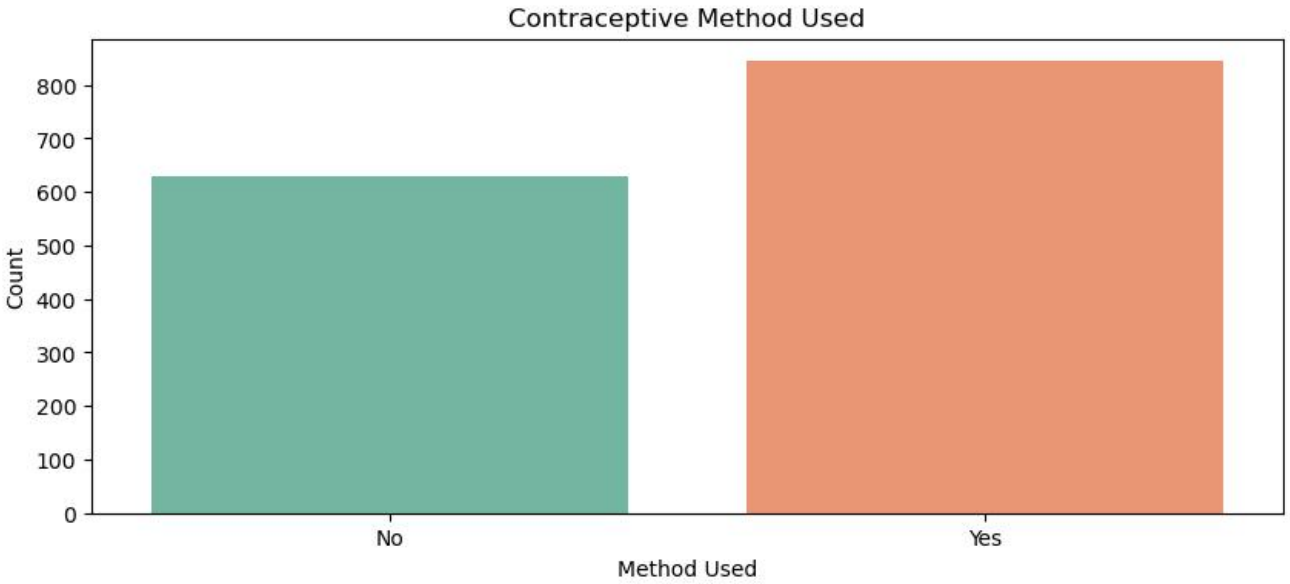

Figure 13 : Contraceptive Method Used

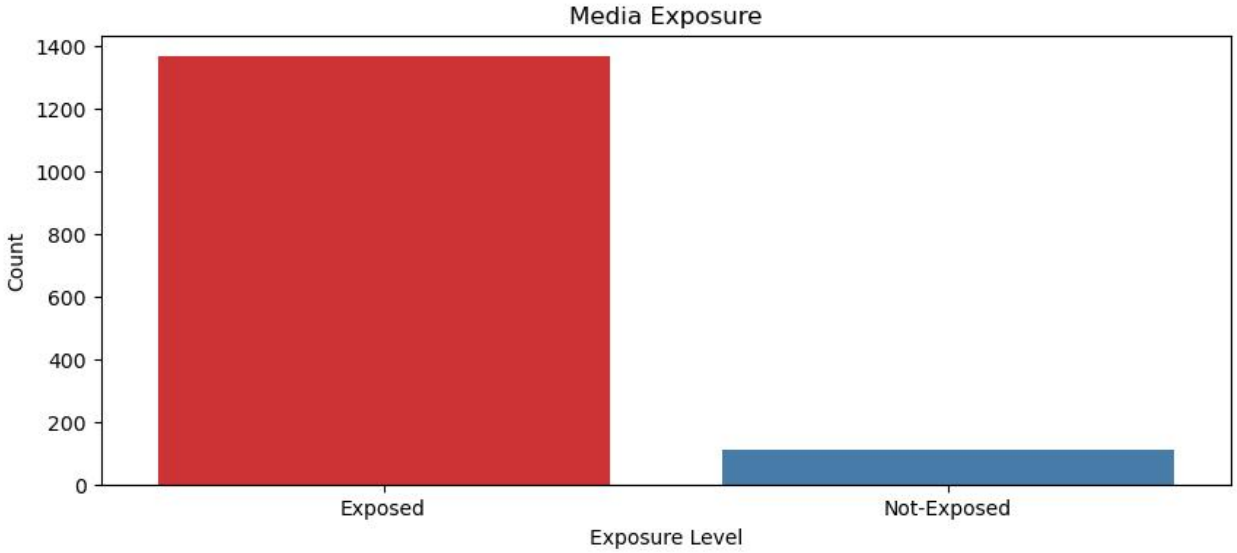

Figure 14 : Media Exposure

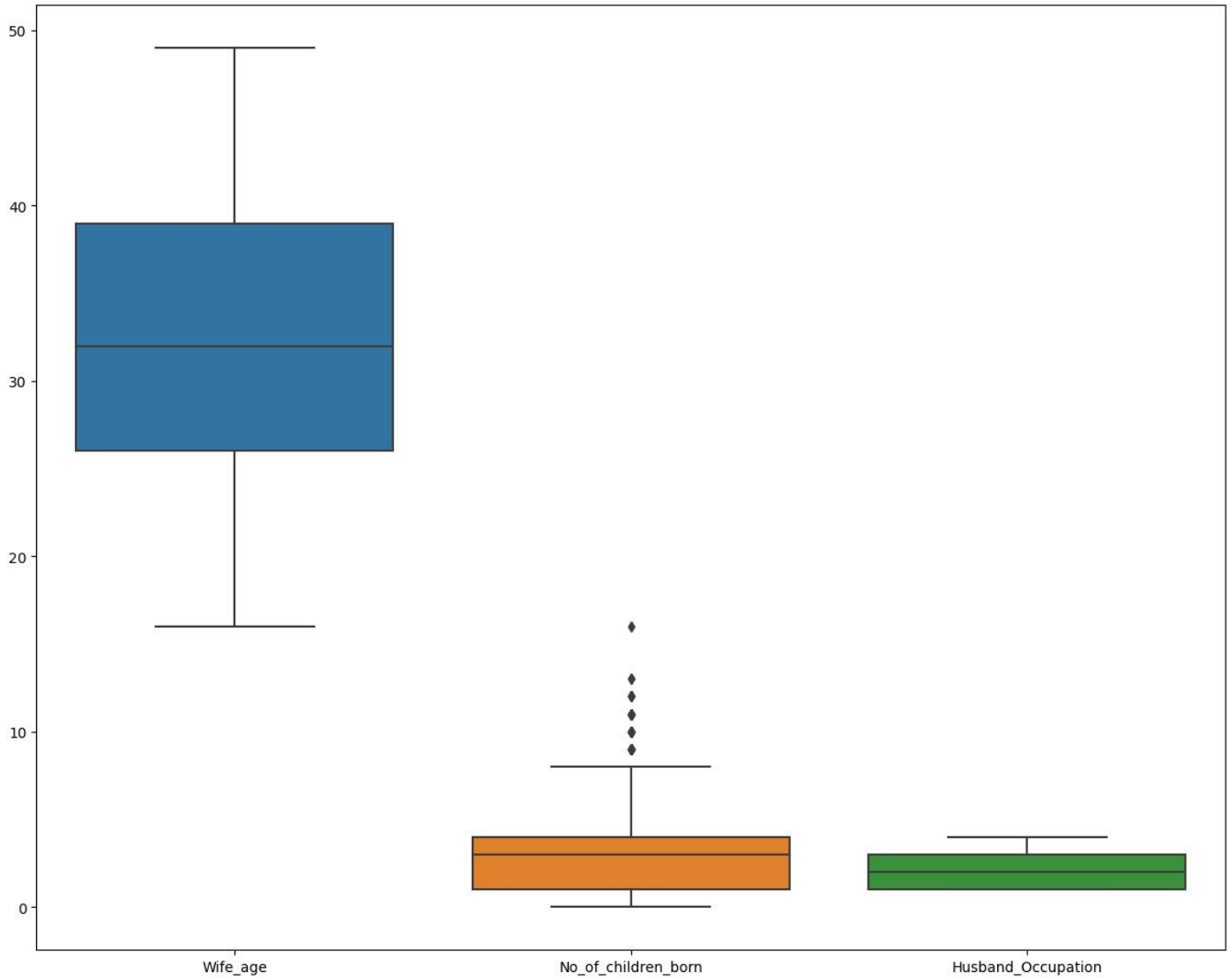

Figure 15 : Box Plot of Numerical variables

Observations :

- The education levels of both wives and husbands seem relatively high, with 'Tertiary' bei ng the most common level of education.
- A significant portion of the wives are not working, and many families have a high standa rd of living.
- The majority of the respondents are using contraceptive methods, which aligns with the objective of predicting contraceptive use.

Multivariate Analysis:

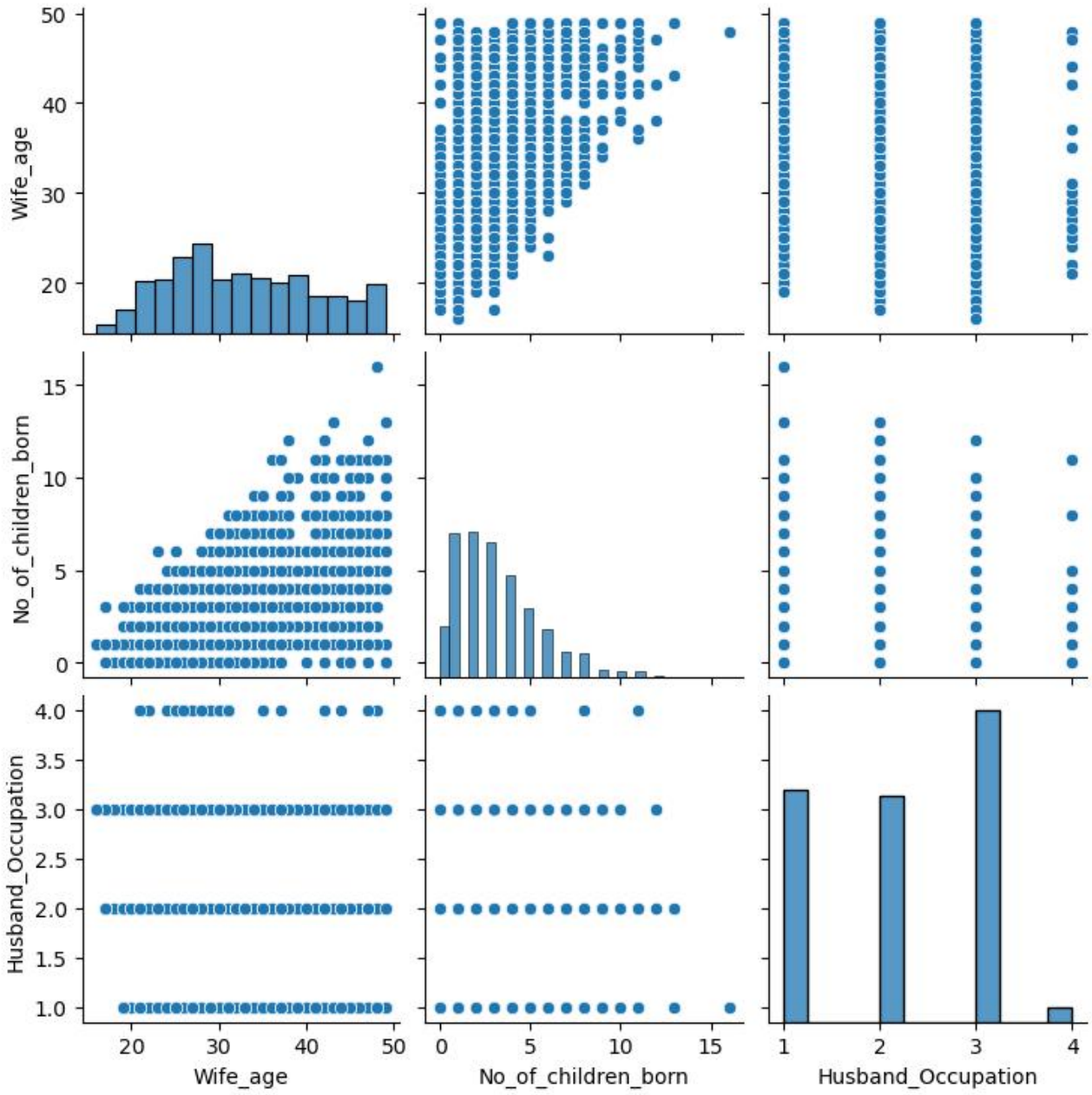

Figure 16 : Pair Plot of Numerical variables

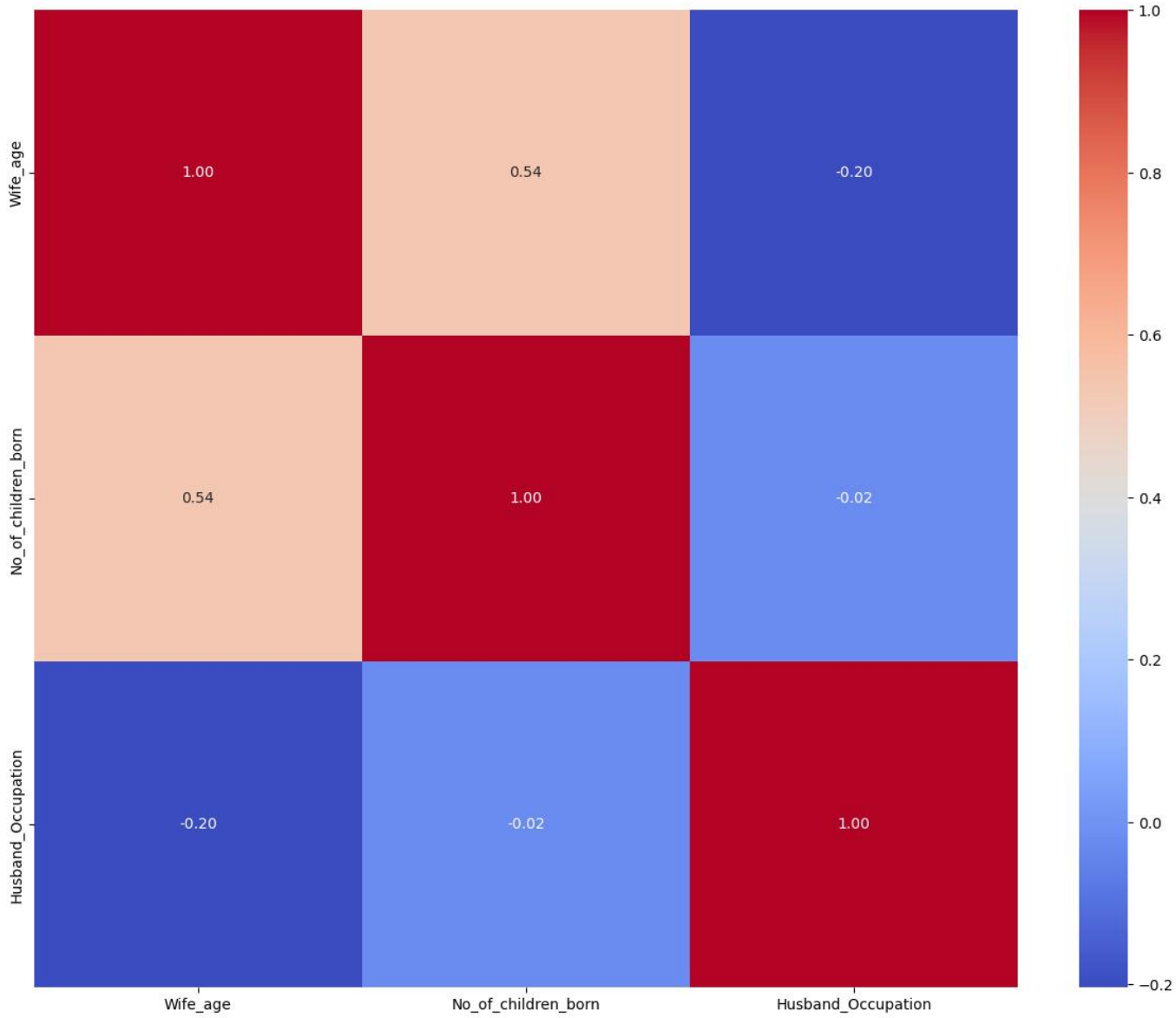

Figure 17 : Heat Map

- None of the correlations are particularly strong, with the highest being 0.538 between " Wife_age" and "No_of_children_born". This implies that while there is some relationshi p between these variables, it's not overwhelmingly strong.

- The positive correlation between "Wife_age" and "No_of_children_born" aligns with ex pectations, whereas the negative correlations involving "Husband_Occupation" are very weak, suggesting minimal linear relationship.

Data Pre-processing:

1. Missing Value Treatment:

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	

Table 17: Missing values in Dataframe

There are 71 missing values are present in "wife_age" and 21 in "no_of_children_born". We replaced the missing values using median values.

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

Table 18: Missing values in Dataframe after imputation

2. Outlier Checks:

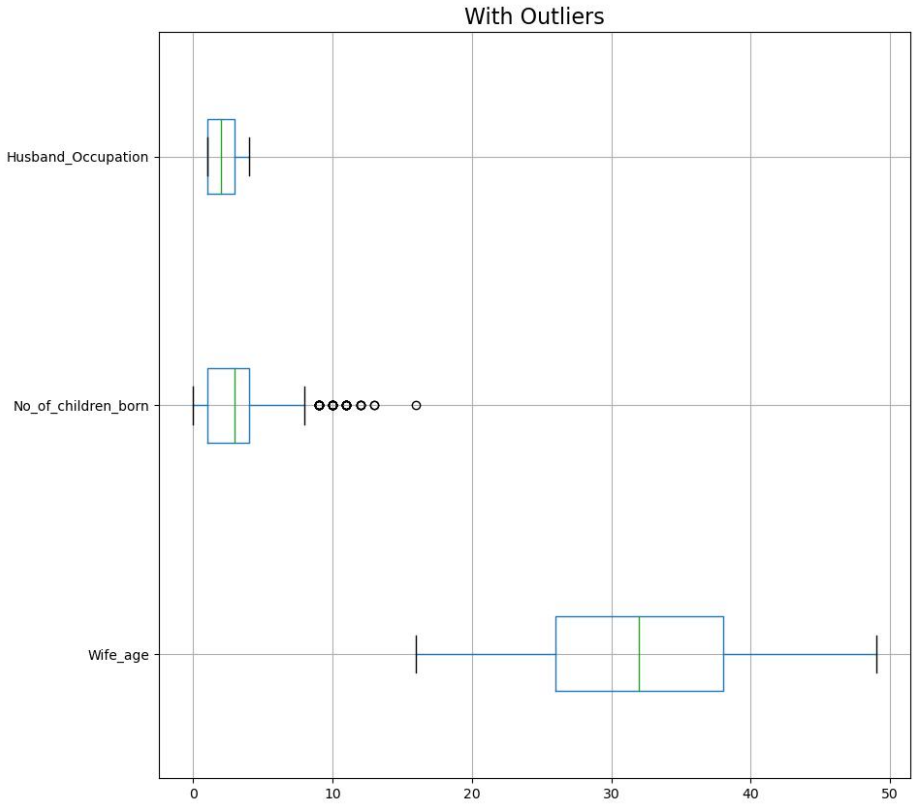

Figure 18 : Outliers in dataframe

No need to treat outliers.

3. Train Test Split:

- The drop method is used to exclude the Contraceptive_method_used column, as this column is the target variable we aim to predict. The remaining columns are considered predictor variables, which will be used to build the model.
- Then splits the predictor variables (X) and the target variable (y) into training and testing sets.

Logistic Regression model:

Proportion of correctly predicted class labels out of the total predictions made on the training data is calculated. This accuracy metric provides an indication of how well the model performs on data it has been trained on.

Accuracy of 64.5% suggests the model has moderate performance on the training data.

1. AUC and ROC for the training data:

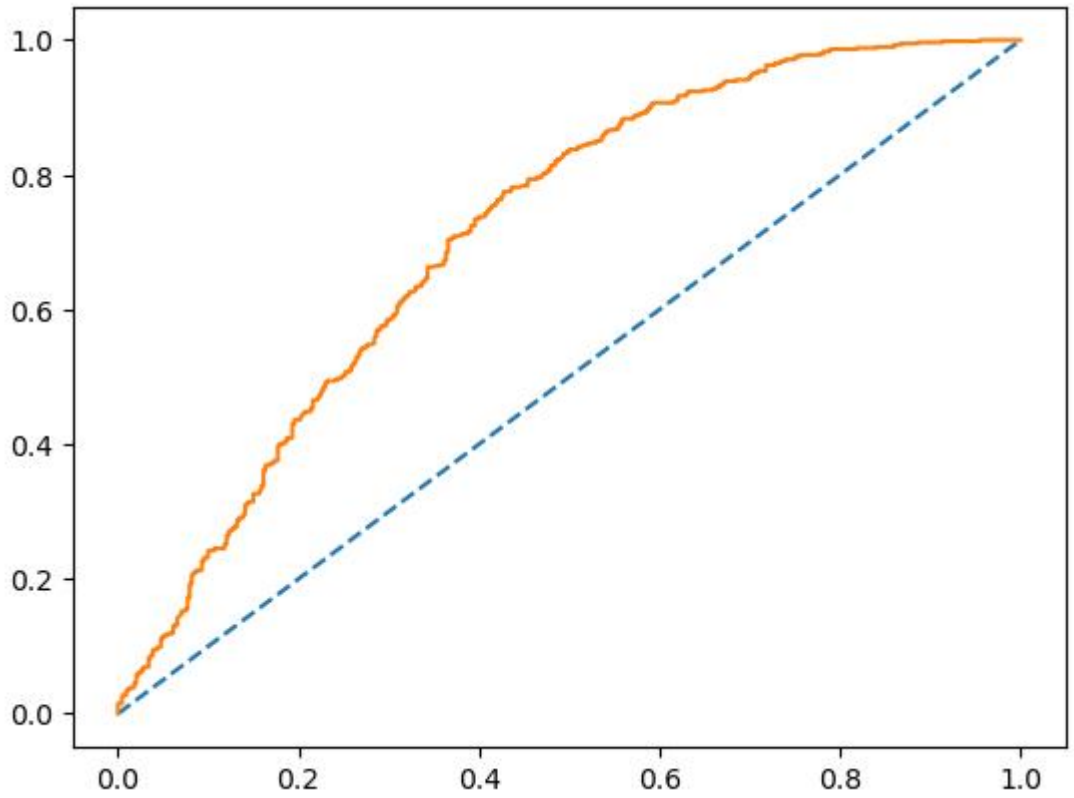

Figure 19 : Receiver Operating Characteristic (ROC) Curve (Train Data)

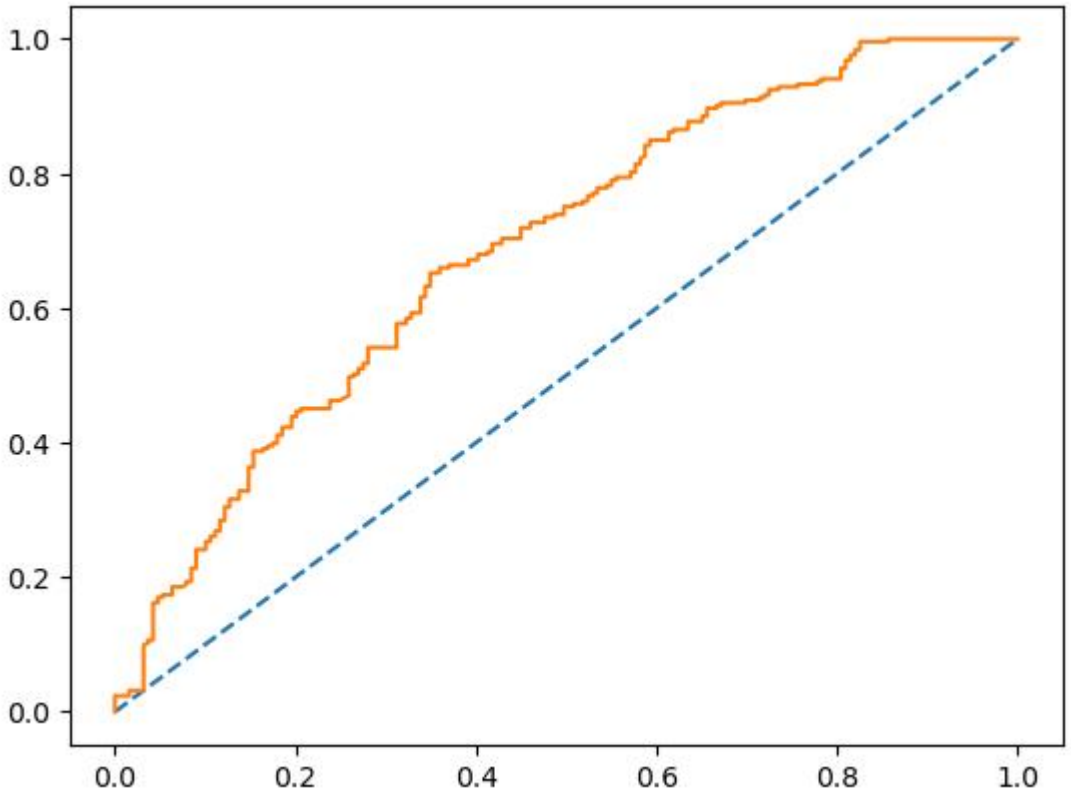

Figure 20 : Receiver Operating Characteristic (ROC) Curve (Train Data)

2. Confusion Matrix for the training data:

```
array([[222, 218],
       [ 99, 492]], dtype=int64)
```

Table 19: Confusion Matrix for the training data

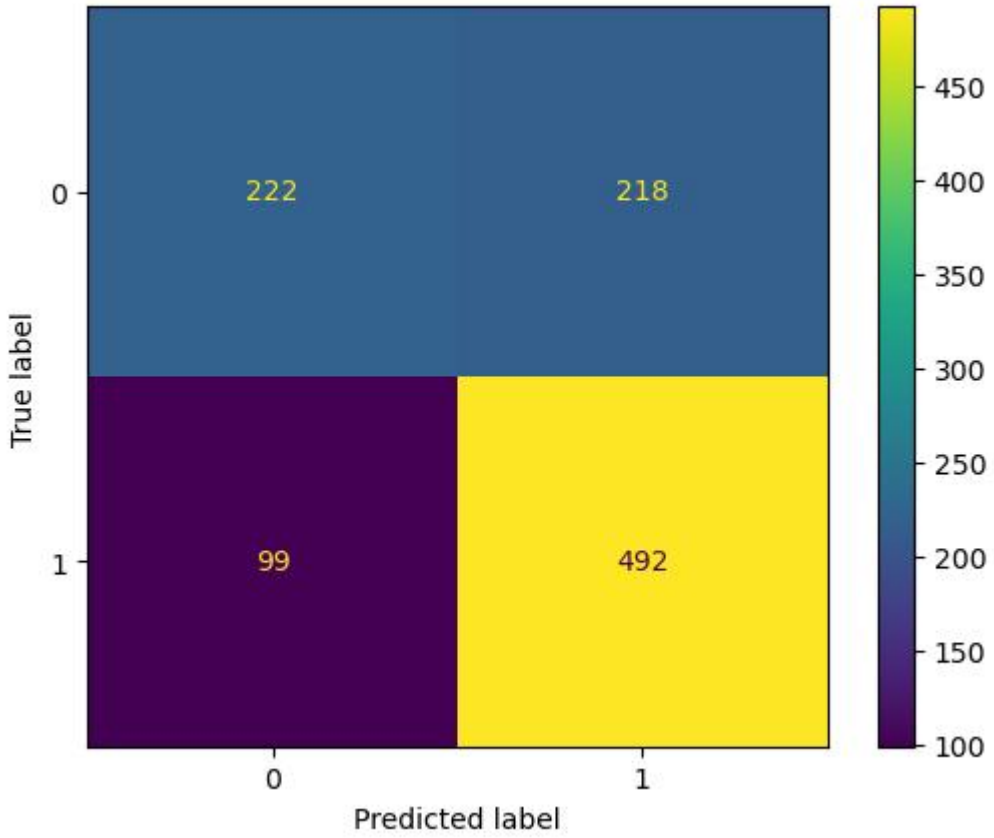

Figure 21 : Confusion Matrix Plot for training data

	precision	recall	f1-score	support
0	0.69	0.50	0.58	440
1	0.69	0.83	0.76	591
accuracy			0.69	1031
macro avg	0.69	0.67	0.67	1031
weighted avg	0.69	0.69	0.68	1031

Table 20: classification Report for Training Data

- The model performs better in identifying instances of class 1 (contraceptive method used) than class 0. This is evident from the higher recall and F1-score for class 1.

- Given that there are more instances of class 1 than class 0, the model's performance metrics are skewed towards the majority class (class 1). This is why the recall for class 1 is higher compared to class 0.

3. Confusion Matrix for test data:

```
array([[ 82, 107],
       [ 52, 201]], dtype=int64)
```

Table 21: Confusion Matrix for the test data

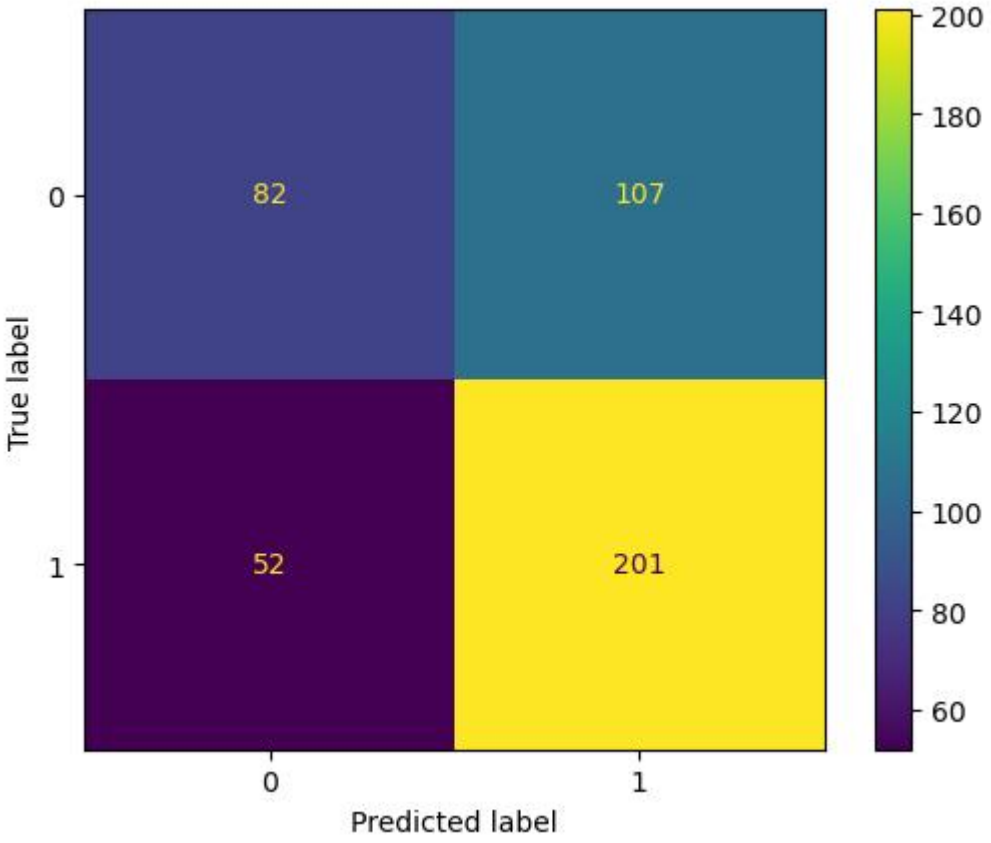

Figure 22 : Confusion Matrix Plot for test data

	precision	recall	f1-score	support
0	0.61	0.43	0.51	189
1	0.65	0.79	0.72	253
accuracy			0.64	442
macro avg	0.63	0.61	0.61	442
weighted avg	0.64	0.64	0.63	442

Table 22: classification Report for Test Data

- 79% of actual class 1 instances are correctly predicted, showing good performance in identifying class 1 in the test set.

- 43% of actual class 0 instances are correctly predicted, indicating that the model struggles to identify class 0 instances in the test set.

- The accuracy of 64% and the F1-scores suggest that the model performs reasonably well.

4. **Applying GridSearchCV for Logistic Regression:**

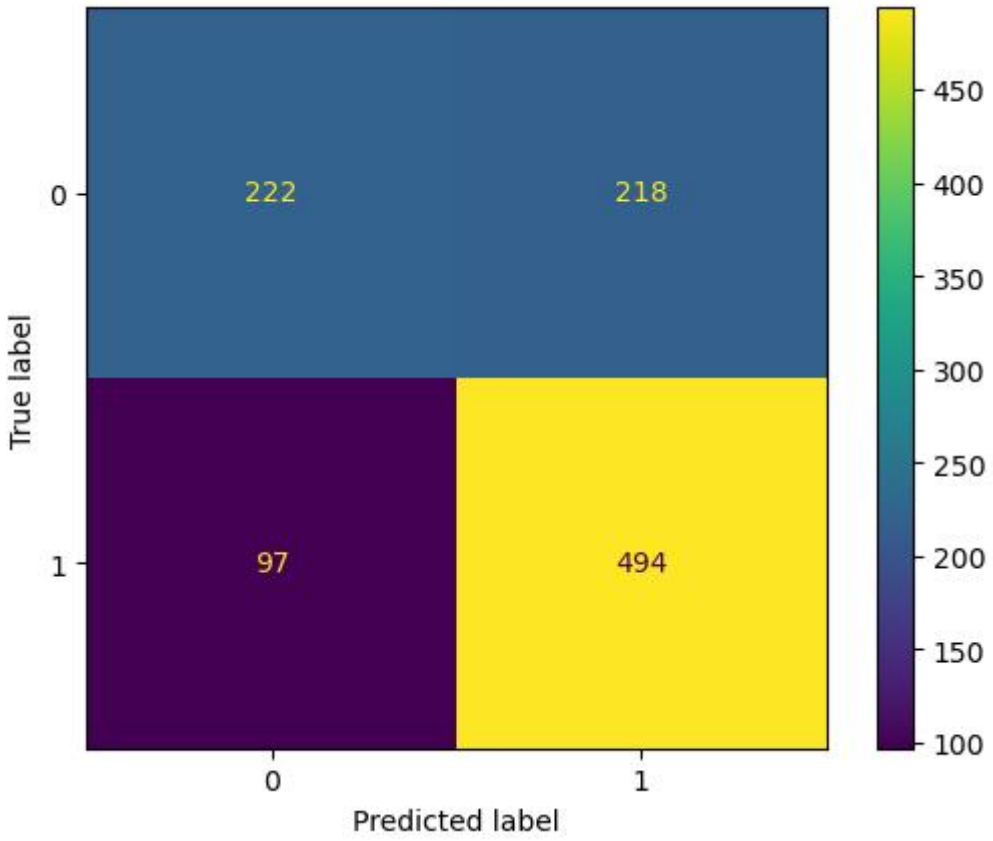

Figure 23 : Confusion Matrix Plot for Best Model (Train data)

	precision	recall	f1-score	support
0	0.70	0.50	0.58	440
1	0.69	0.84	0.76	591
accuracy			0.69	1031
macro avg	0.69	0.67	0.67	1031
weighted avg	0.69	0.69	0.68	1031

Table 23: Classification Report for Train Data(Best Model)

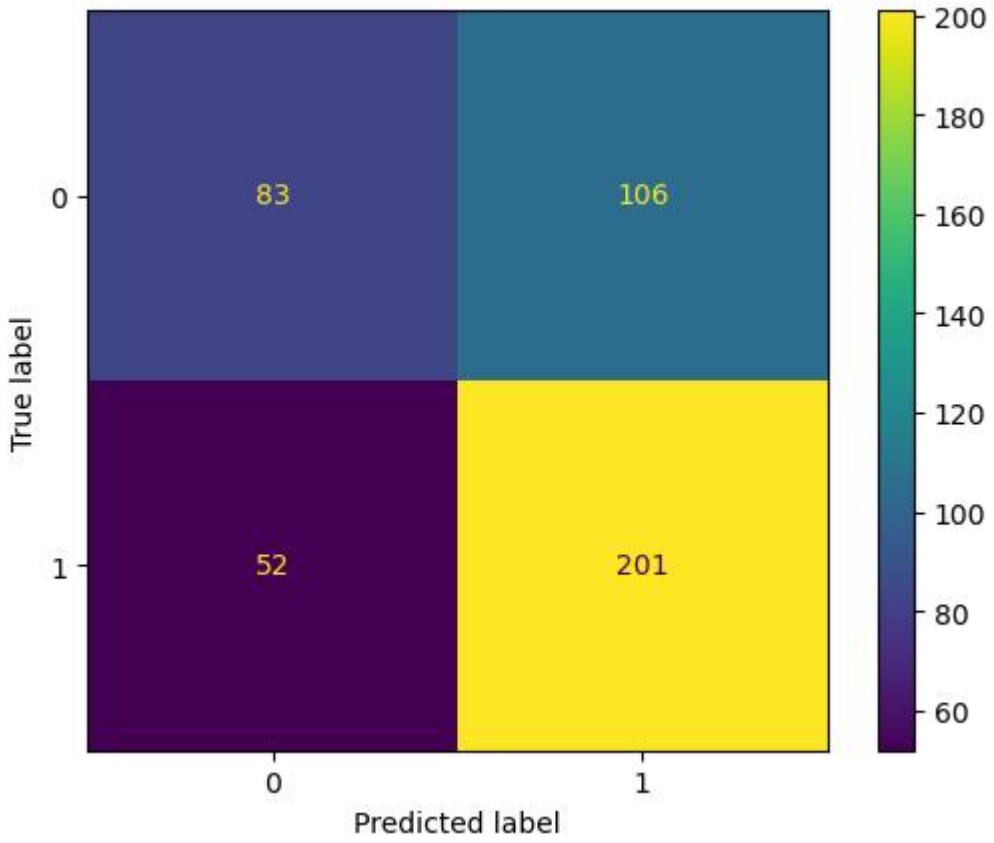

Figure 24 : Confusion Matrix Plot for Best Model (Test data)

	precision	recall	f1-score	support
0	0.61	0.44	0.51	189
1	0.65	0.79	0.72	253
accuracy			0.64	442
macro avg	0.63	0.62	0.62	442
weighted avg	0.64	0.64	0.63	442

Table 24: Classification Report for Test Data(Best Model)

- At 44%, the model misses a substantial portion of actual class 0 instances, indicating poor performance in identifying this class.
- At 83%, the model is quite effective at identifying class 1 instances, showing strong performance for this class.

Linear Discriminant Analysis model:

1. Splitting the Data:

- The dataset is divided into predictor variables (features) and the target variable (which indicates whether the contraceptive method is used).
- The data is then split into training and testing sets, with 30% of the data reserved for testing. This split is done while maintaining the original distribution of the target variable to ensure balanced representation in both sets.
- Information about the dimensions of the training and testing sets is printed to confirm the correct split.

2. Standardization:

- To ensure that all features contribute equally to the model, the data is standardized.
- This process transforms the features to have a mean of zero and a standard deviation of one. This step helps improve the performance and stability of the LDA model.

3. Model Initialization and Training:

- A Linear Discriminant Analysis (LDA) model is created and trained using the standardized training data. LDA is a technique used to find the linear combinations of features that best separate the classes in the dataset.

4. Class Predictions:

- The trained LDA model is used to predict class labels for both the training and testing datasets. Predictions are made with a default cut-off value of 0.5, which means that any probability greater than or equal to 0.5 is classified into one class, and anything below 0.5 is classified into the other.

5. Training Data and Test Data Confusion Matrix Comparison:

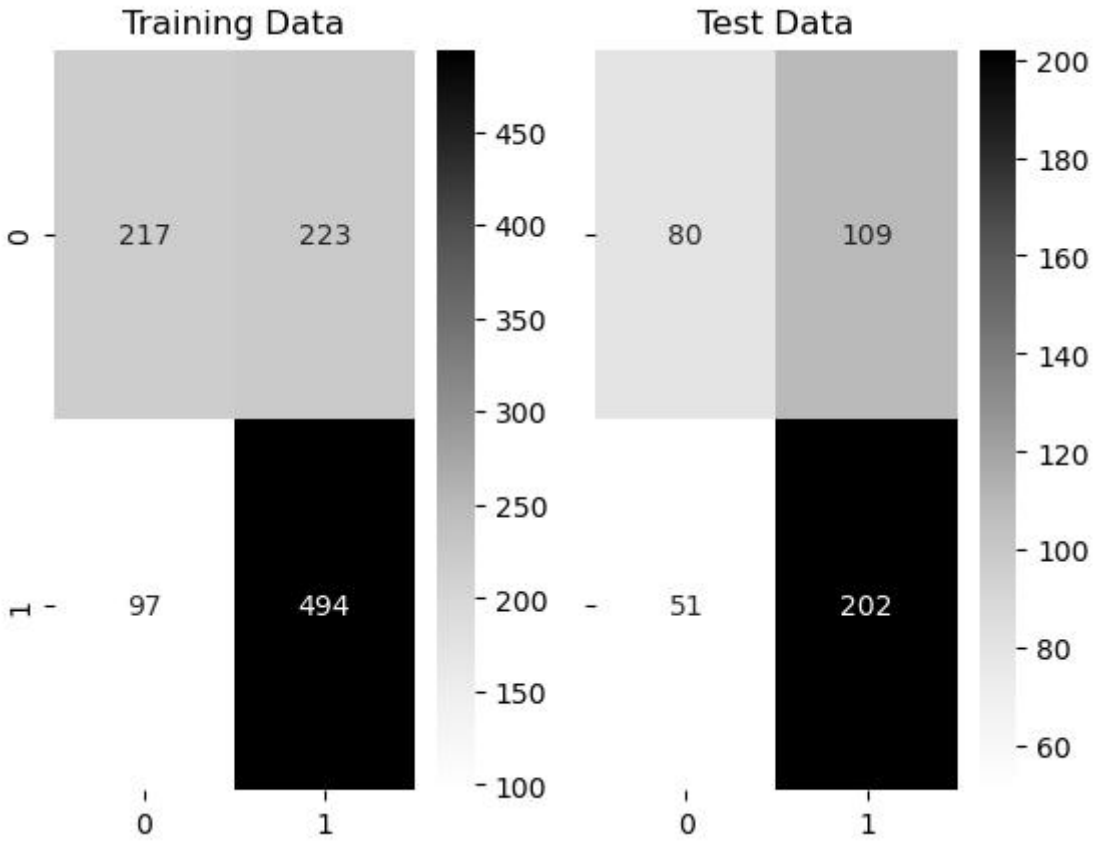

Figure 25 : Confusion Matrix

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.69	0.49	0.58	440
1	0.69	0.84	0.76	591
accuracy			0.69	1031
macro avg	0.69	0.66	0.67	1031
weighted avg	0.69	0.69	0.68	1031

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.61	0.42	0.50	189
1	0.65	0.80	0.72	253
accuracy			0.64	442
macro avg	0.63	0.61	0.61	442
weighted avg	0.63	0.64	0.62	442

Table 25: classification Report

6.INFERENCE:

- Overall accuracy of the model – 64 % of total predictions are correct.
- The training accuracy was slightly higher at 69%, suggesting some overfitting
- For Customer who did not use Contraceptive method (Label 0):

  Precision (61%) – 61% of Customers who did not use Contraceptive method are correctly predicted ,out of all Customers who did not use Contraceptive method that are predicted .

  Recall (42%) – Out of all the Customers who actually did not use Contraceptive method , 42% of Customers who did not Churn have been predicted correctly .

- For Customer who did use Contraceptive method (Label 1):

  Precision (65%) – 65% of Customers who did use Contraceptive method are correctly predicted ,out of all Customers who did use Contraceptive method that are predicted .

  Recall (80%) – Out of all the Customers who actually did use Contraceptive method , 80% of Customers who did use Contraceptive method have been predicted correctly .

7. Probability prediction for the training and test data:

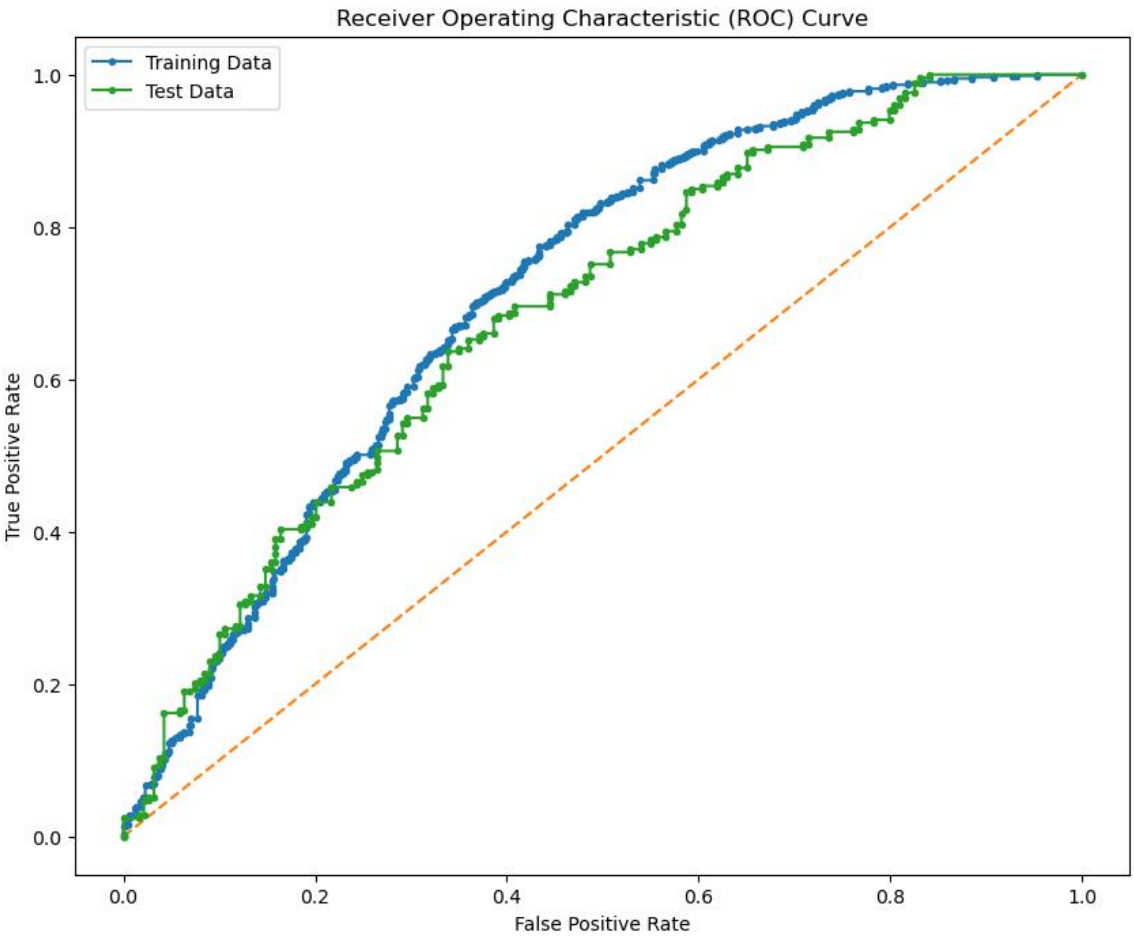

Figure 26 : Confusion Matrix

8. Conclusion:

- Higher standard of living and media exposure are associated with Contraceptive method usage, indicating a correlation between socioeconomic factors and Contraceptive method use.
- Older age and lower education levels tend to push the prediction towards No Contraceptive method use, while more children and higher education levels for the wife push towards use of Contraceptive method.

CART Model:

Model Evaluation:

AUC and ROC for the training data

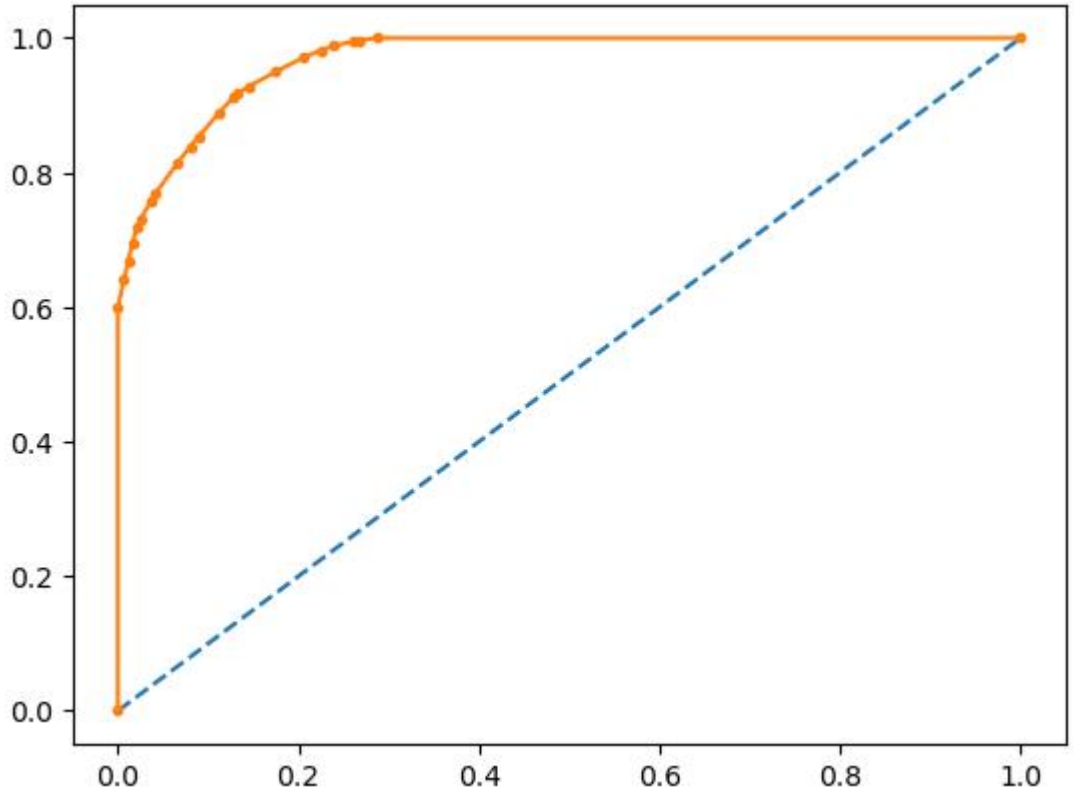

Figure 27 : AUC and ROC for the training data

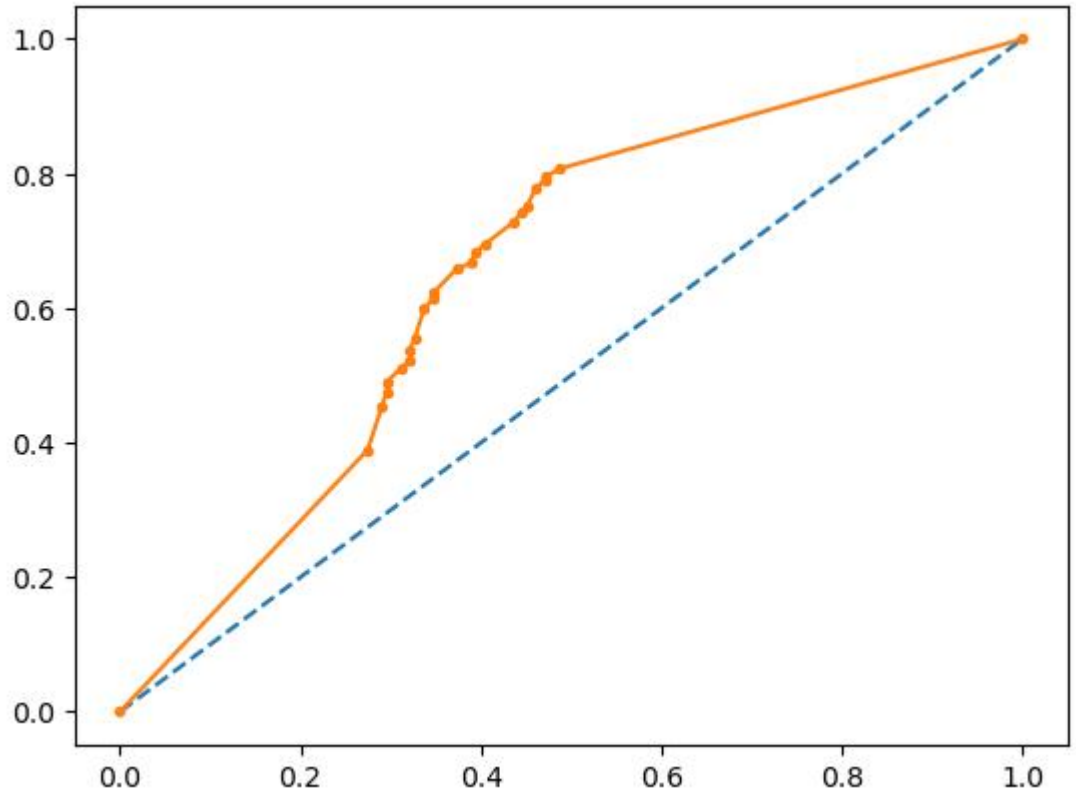

Figure 28 : AUC and ROC for the test data

	precision	recall	f1-score	support
0	0.90	0.86	0.88	436
1	0.90	0.93	0.91	595
accuracy			0.90	1031
macro avg	0.90	0.89	0.89	1031
weighted avg	0.90	0.90	0.90	1031

Table 26: classification Report(Train data)

	precision	recall	f1-score	support
0	0.60	0.60	0.60	193
1	0.69	0.69	0.69	249
accuracy			0.65	442
macro avg	0.65	0.65	0.65	442
weighted avg	0.65	0.65	0.65	442

Table 27: classification Report(Test data)

- The model shows very high performance on the training data with an accuracy of 90%. Both classes have strong precision, recall, and F1-scores.
- There is a noticeable drop in accuracy from the training data (90%) to the test data (65%). This suggests that the model may be overfitting the training data and not generalizing well to unseen data.

Performance Of All The Models Built:

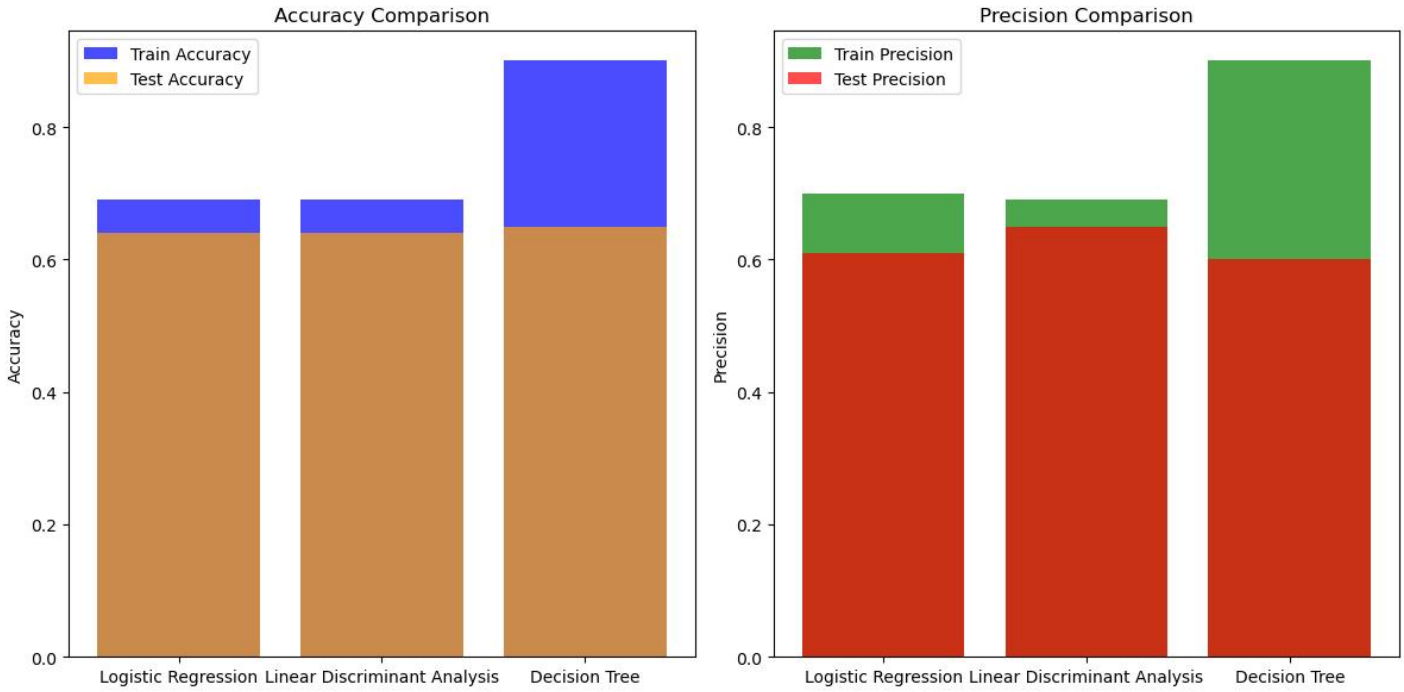

Figure 29 : Accuracy And Precision Comparison of Models

- The Decision Tree has the highest training accuracy but struggles with generalization as shown by its lower test accuracy. Logistic Regression and LDA show consistent performance on both training and test data.

- On the test set, Logistic Regression has slightly higher precision compared to the Decision Tree. Precision for the LDA is slightly higher on the test set compared to the Decision Tree.

- The Decision Tree has higher recall on the training set but performs similarly to Logistic Regression and LDA on the test set.

Conclusions and Recommendations:

- The most significant predictors are the wife's age and the number of children born. These factors are crucial for understanding contraceptive choices. This suggests that age and family size are primary determinants in whether women are likely to use contraceptive methods.

- Economic and living conditions, as reflected by the standard of living index, also play an important role. This indicates that financial stability and living standards significantly influence contraceptive decisions.

- Cultural factors, including religion and education level, have a notable impact. This shows that beliefs and educational background are important considerations in contraceptive use.

- The husband's education level has a lower impact on the prediction. While it contributes to the model, its influence is less compared to other attributes.

- Design targeted outreach programs focusing on women in specific age groups and those with varying numbers of children. Provide tailored information on contraceptive options that align with their life stage and family planning needs.

- Implement programs that cater to different economic groups. For lower-income groups, consider providing affordable or subsidized contraceptive options and education on family planning. For higher-income groups, offer premium services and advanced options.

- Develop culturally sensitive educational materials and counseling services. Ensure that these resources respect religious beliefs and address educational backgrounds to increase their relevance and acceptance.

- Offer specific programs for working women, recognizing their unique schedules and challenges. Provide flexible and convenient options that fit their lifestyle, such as online consultations or workplace-based education sessions.