PREDICTIVE MODELLING PROJECT

Submitted by,
VIDYA V

PGPDSBA.O.2023.B
06.08.2023

CONTENTS

List of Figures

List of Figures

Case 1: Compactiv Data- 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.

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

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

Fig.1.1 Dataset info

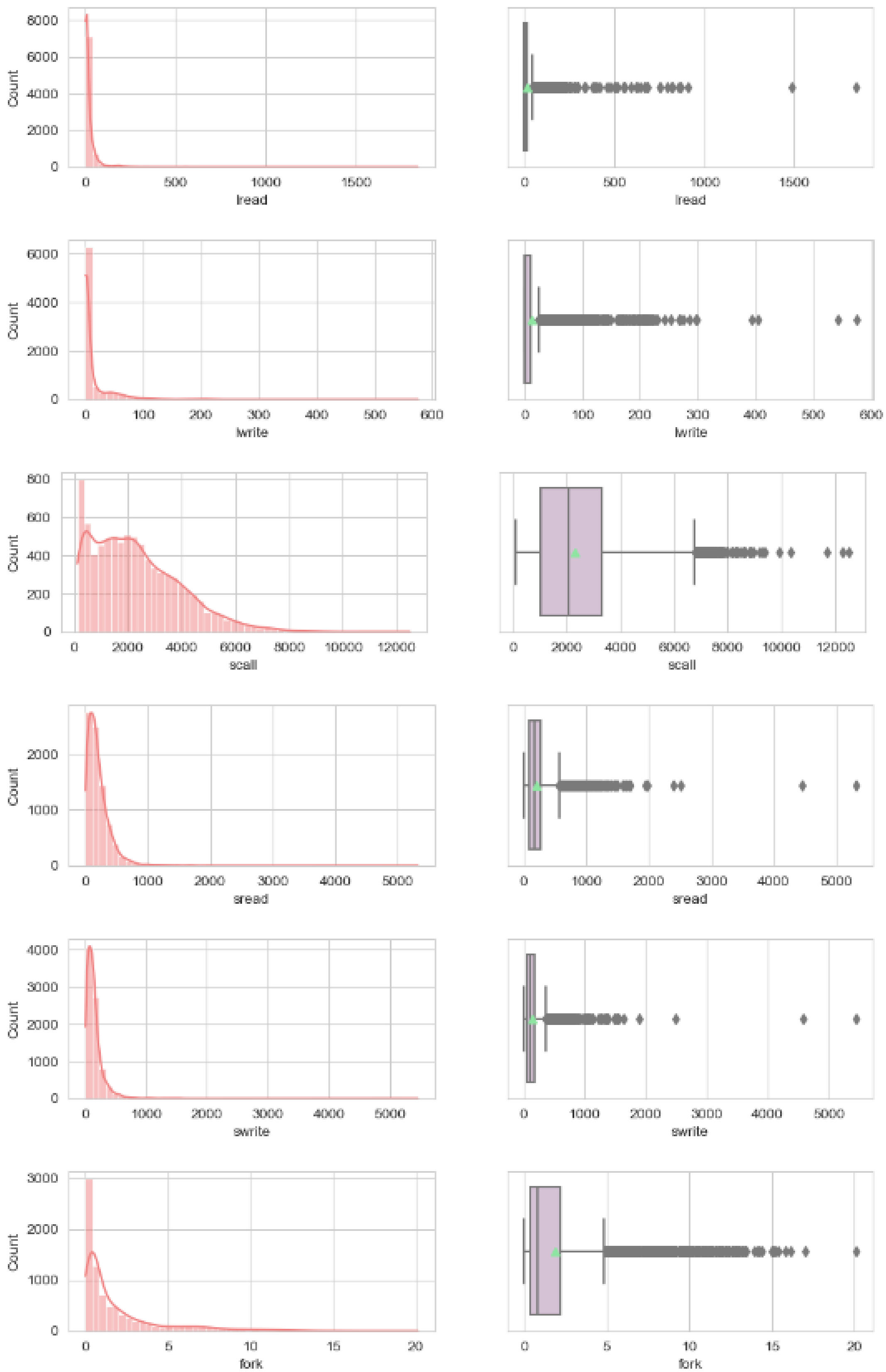
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
lread	8192.0	NaN	NaN	NaN	19.56	53.35	0.0	2.0	7.0	20.0	1845.0
lwrite	8192.0	NaN	NaN	NaN	13.11	29.89	0.0	0.0	1.0	10.0	575.0
scall	8192.0	NaN	NaN	NaN	2306.32	1633.62	109.0	1012.0	2051.5	3317.25	12493.0
sread	8192.0	NaN	NaN	NaN	210.48	198.98	6.0	86.0	166.0	279.0	5318.0
swrite	8192.0	NaN	NaN	NaN	150.06	160.48	7.0	63.0	117.0	185.0	5456.0
fork	8192.0	NaN	NaN	NaN	1.88	2.48	0.0	0.4	0.8	2.2	20.12
exec	8192.0	NaN	NaN	NaN	2.79	5.21	0.0	0.2	1.2	2.8	59.56
rchar	8192.0	NaN	NaN	NaN	194879.85	239332.57	0.0	31606.5	122035.0	265394.75	2526649.0
wchar	8192.0	NaN	NaN	NaN	95727.39	140772.42	0.0	22846.75	46434.5	106037.0	1801623.0
pgout	8192.0	NaN	NaN	NaN	2.29	5.31	0.0	0.0	0.0	2.4	81.44
ppgout	8192.0	NaN	NaN	NaN	5.98	15.21	0.0	0.0	0.0	4.2	184.2
pgfree	8192.0	NaN	NaN	NaN	11.92	32.36	0.0	0.0	0.0	5.0	523.0
pgscan	8192.0	NaN	NaN	NaN	21.53	71.14	0.0	0.0	0.0	0.0	1237.0
atch	8192.0	NaN	NaN	NaN	1.13	5.71	0.0	0.0	0.0	0.6	211.58
pgin	8192.0	NaN	NaN	NaN	8.28	13.87	0.0	0.6	2.8	9.76	141.2
ppgin	8192.0	NaN	NaN	NaN	12.39	22.28	0.0	0.6	3.8	13.8	292.61
pflt	8192.0	NaN	NaN	NaN	109.79	114.42	0.0	25.0	63.8	159.6	899.8
vflt	8192.0	NaN	NaN	NaN	185.32	191.0	0.2	45.4	120.4	251.8	1365.0
runqsz	8192	2	Not_CPU_Bound	4331	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freemem	8192.0	NaN	NaN	NaN	1763.46	2482.1	55.0	231.0	579.0	2002.25	12027.0
freeswap	8192.0	NaN	NaN	NaN	1328125.96	422019.43	2.0	1042623.5	1289289.5	1730379.5	2243187.0
usr	8192.0	NaN	NaN	NaN	83.97	18.4	0.0	81.0	89.0	94.0	99.0

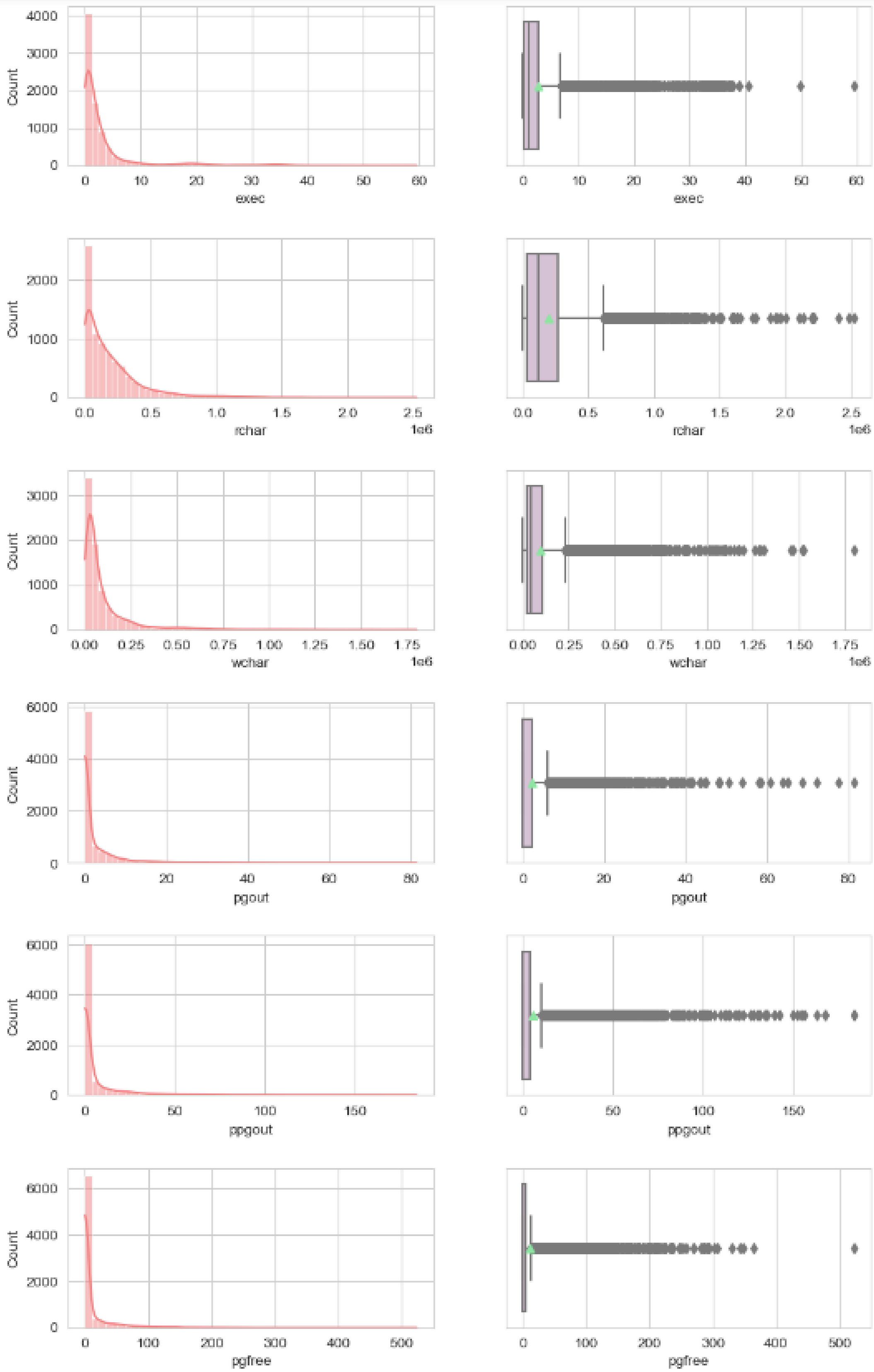
Fig.1.2. Data Description

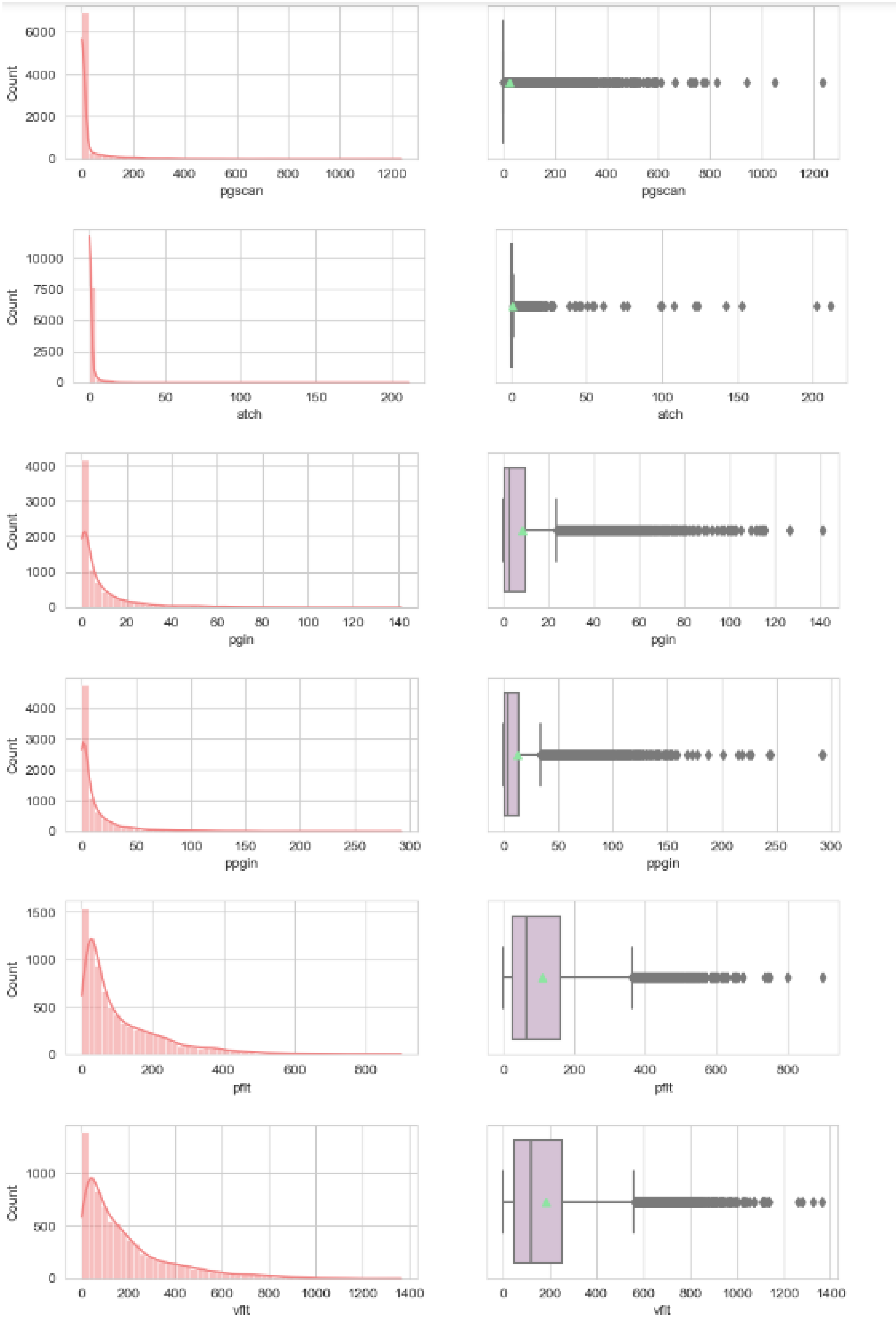
Observations:

- There are 8192 datapoints and 22 fields in the given dataset
- All the fields are numerical, except for 'runqsz' which has two categories- CPU bound and Non-CPU-bound
- The field 'usr', which is the percentage of time the CPU runs in user mode, is the target variable for the linear regression
- The fields 'rchar' and 'wchar' have NaN values. These represent the number of characters transferred during system read and write calls.
- Based on domain knowledge, it is plausible that either of these can be 0, but not both 0 for a given datapoint. So, for further analysis, we choose to impute with 0. As the number is very limited when compared to the depth of the dataset (119 in 8192), the choice is justified.
- There are no bad values in the dataset
- Some fields have outliers
- There are no duplicate values

1.1. EDA


Count
lread
lwrite
scall
sread
swrite
fork


Count
exec
rchar
1e6
wchar
1e6
pgout
ppgout
pgfree


Count
pgscan
pgscan
Count
atch
atch
Count
pgin
pgin
Count
ppgin
ppgin
Count
pflt
pflt
Count
vflt
vflt

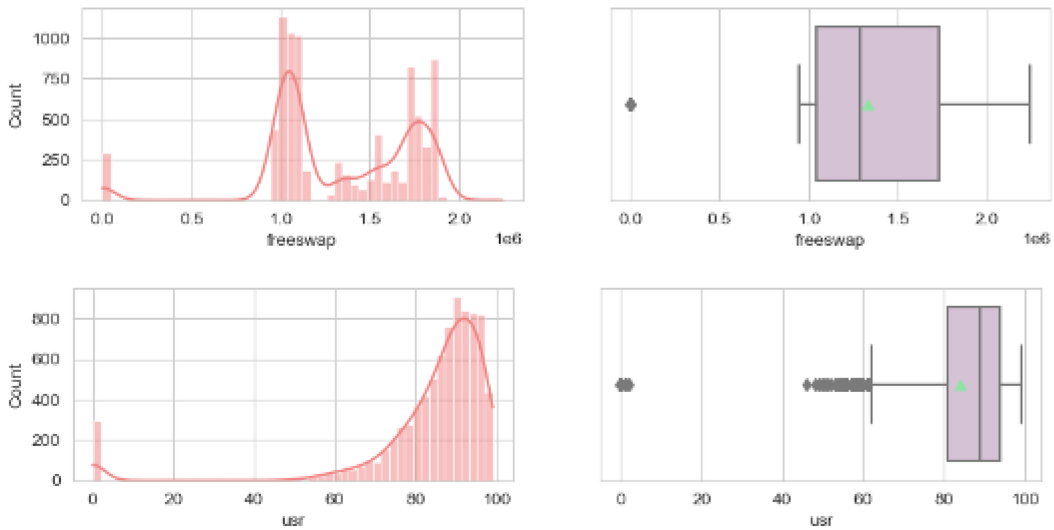


Fig.1.3 Univariate Analysis- Numerical Columns

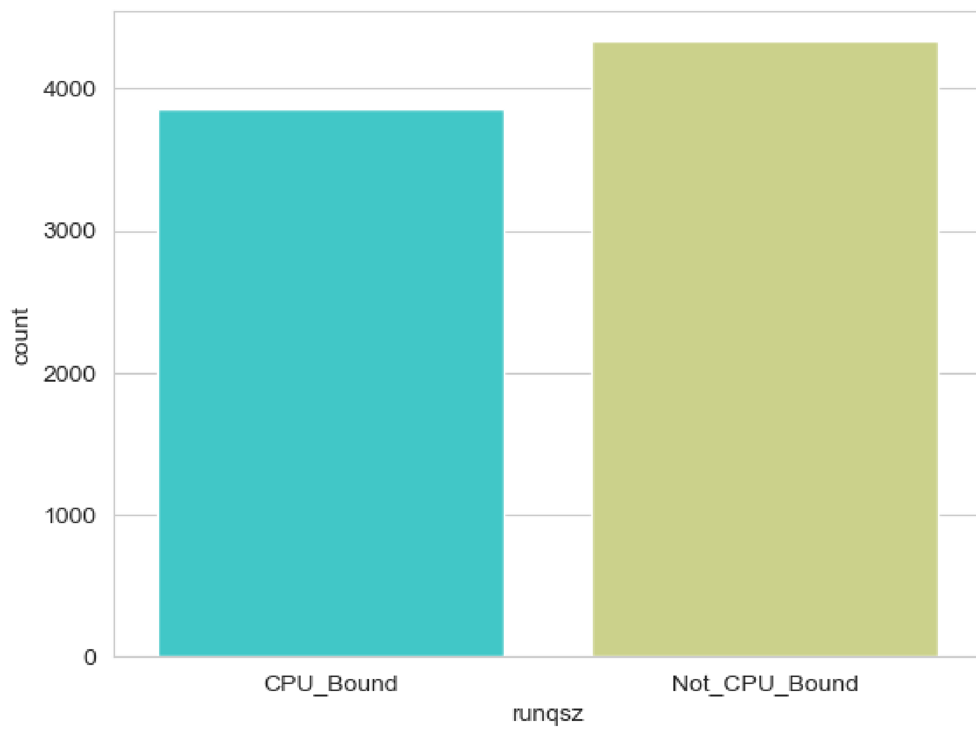


Fig.1.4. Univariate Analysis- Categorical Column

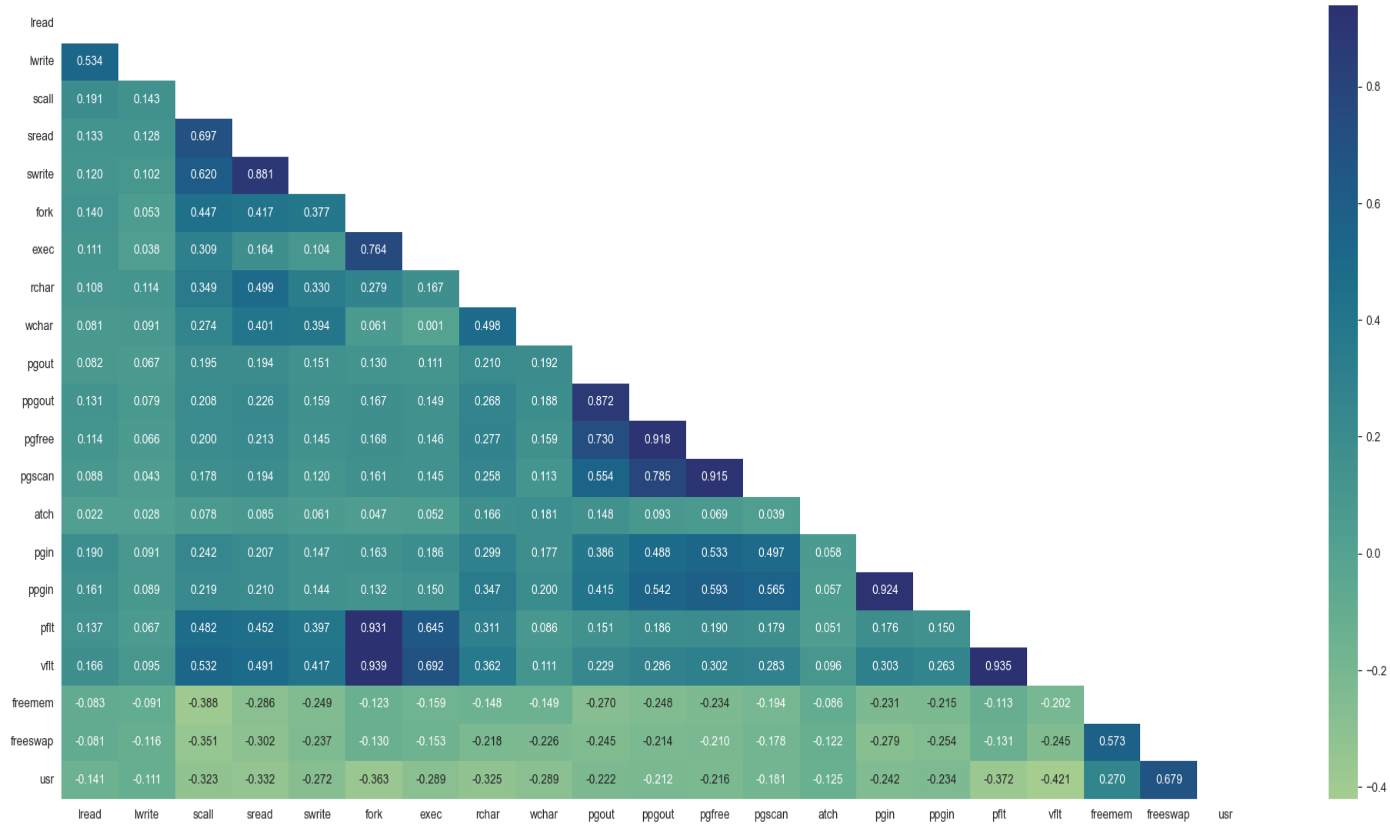


Fig.1.5 Heatmap of fields

Observations:

- Univariate analysis of numerical fields:
    - All fields have outliers and hence are skewed
    - The 'freeswap' field has a bimodal distribution
- Univariate analysis of categorical field:
    - There are more instances of 'Not_CPU_Bound' than 'CPU_Bound'
- Bivariate Analysis:
    - Few fields like pflt, vflt and fork, pgout and ppgout, pgscan and pgfree exhibit a very strong correlation
    - However, none of the fields have a very strong correlation to the target variable- usr

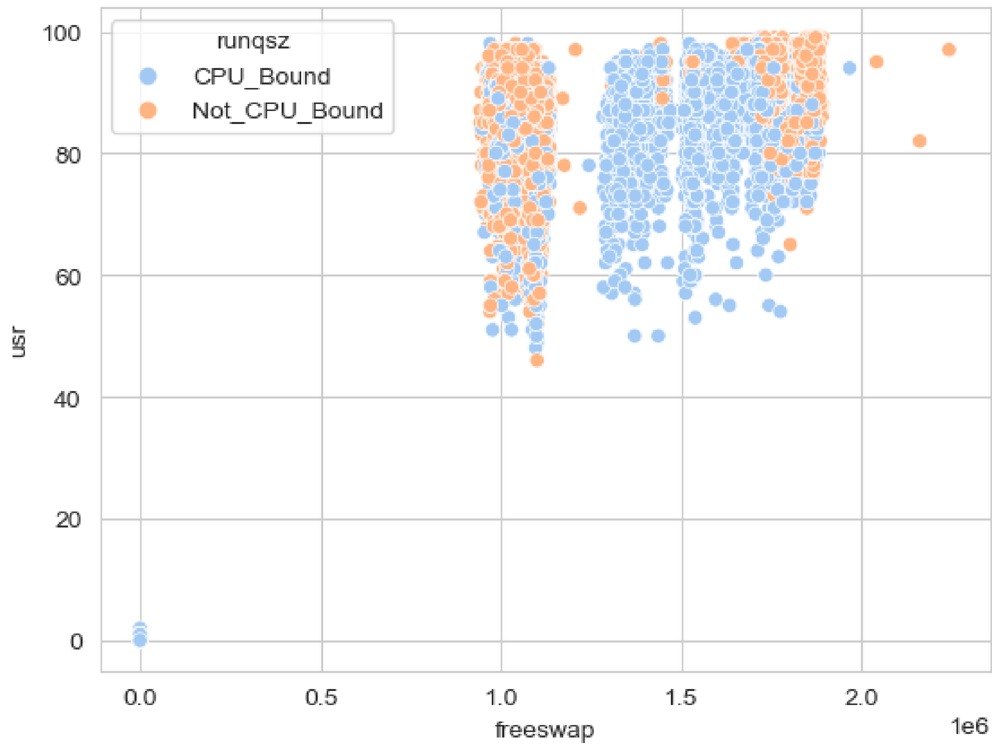


Fig.1.6. Multivariate Analysis

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

Observations:

- The fields 'rchar' and 'wchar' have NaN values. These represent the number of characters transferred during system read and write calls.
- Based on domain knowledge, it is plausible that either of these can be 0, but not both 0 for a given datapoint. So, on further exploration we find that atleast one of these is non zero for all datapoints. So, as such, there is no missing data and hence are imputed with 0
- There are no bad values in the dataset
- Some fields have outliers and have been treated with IQR approach
- There are no duplicate values
- Feature engineering:
    - The features of the given dataset were combined as per the formulae below to derive new features:
        - Total_io: Total_IO_Activity = lread + lwrite
        - Total_disk_io : Total_Disk_IO = sread + swrite
        - Total_pg: Total_Page_Activities = pgout + ppgout + pgfree + pgscan + atch + pgin + ppgin + pflt + vflt
        - Total_proc: Total_Process_Activities = fork + exec
        - Disk_mem_usage: Total_Disk_Memory_Usage = freemem + freeswap
        - Total_scalls: Total_System_Calls = scall + fork + exec
        - Char_total: Total_Characters_Transferred = rchar + wchar
        - Io_scall_ratio: Total_io/Total_scalls

- Total_pg_faults= pflt+vflt

3. **Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.**

For the sake of analysis- 2 different approaches were adopted, based on which 2 different models were built

Model 1: Without feature Engineering

Model 2: With Feature Engineering

Performance Metric	MODEL 1- No Feature Engineering						MODEL 2- With Feature engineering					
	Scikit learn Linear Regression		Stats models OLS		Post-VIF drop		Scikit learn Linear Regression		Stats models OLS		Post-VIF drop	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
R square/Score	0.786	0.793	0.786	0.797	0.963	0.967	0.779	0.785	0.779	0.788	0.961	0.963
Adj R square	-	-	0.785	0.795	0.963	0.967	-	-	0.778	0.787	0.961	0.963
RMSE	4.546	4.357	219	19			4.620	4.436	21	19		

Table.1.1 Linear Regression- Performance Comparison

Observations:

- Comparison between Models:
  - Model 1, i.e., the model without feature engineering performs better in all aspects
- Comparison Between Methods:
  - Sklearn linear Regressions have better RMSE for both the models, but statsmodels OLS have better R square values
- Variance Influence factor analysis:
  - For both the models, after analyzing the most factor that most influences variance and dropping, the R square and adjusted R square values made a significant jump.
  - The difference of the R square and adjusted R square values before and after dropping the factor of most influence was 0.18.
  - However, beyond this, the r square and adjusted r square values showed no improvement for subsequent dropping of the most influential factors.

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

Linear Equation- sklearn

usr = (-0.05420 * lread) + (0.04506 * lwrite) - (0.00076 * scall) + (0.00228 * sread) - (0.00495 * swrite) - (0.14707 * fork) - (0.23054 * exec) - (0.00000 * rchar) - (0.00000 * wchar) - (0.45885 * pgout) + (0.03462 * ppgout) + (0.04065 * pgfree) + (0.00000 * pgscan) + (0.51061 * atch) + (0.00700 * pgin) - (0.05910 * ppgin) - (0.03145 * pflt) - (0.00639 * vflt) - (0.00053 * runqsz) + (0.00001 * freemem) + (1.84976 * freeswap)

Observations and Inferences:

- Based on the above equation, it can be inferred that the variables- freeswap, atch, and pgout contribute most to the percentage of time spent in usr mode, and hence are the most significant
- Variables with positive coefficients (e.g., `lwrite`, `sread`, `ppgout`, `pgfree`, `atch`, `pgin`, `freemem`, `freeswap`) have a positive impact on the target variable `usr`. An increase in these variables is associated with an increase in the portion of time that CPUs run in user mode.
- Variables with negative coefficients (e.g., `lread`, `scall`, `swrite`, `fork`, `exec`, `pgout`, `ppgin`, `pflt`, `vflt`, `runqsz`) have a negative impact on the target variable `usr`. An increase in these variables is associated with a decrease in the portion of time that CPUs run in user mode.
- Some input variables have zero coefficients (`rchar`, `wchar`, `pgscan`). This indicates that changes in these variables do not significantly impact the target variable.

Case 2: Contraceptive dataset- Logistic Regression, LDA and CART

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.**

```
<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
```

Fig.2.1. Info of dataset

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Wife_age	1326.0	NaN	NaN	NaN	32.56	8.29	16.0	26.0	32.0	39.0	49.0
Wife_ education	1393	4	Tertiary	515	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_education	1393	4	Tertiary	827	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_of_children_born	1372.0	NaN	NaN	NaN	3.29	2.4	0.0	1.0	3.0	5.0	16.0
Wife_religion	1393	2	Scientology	1186	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Wife_Working	1393	2	No	1043	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_Occupation	1393.0	NaN	NaN	NaN	2.17	0.85	1.0	1.0	2.0	3.0	4.0
Standard_of_living_index	1393	4	Very High	618	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Media_exposure	1393	2	Exposed	1284	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Contraceptive_method_used	1393	2	Yes	779	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Fig.2.2 Data description

Observations:

- Wife_age, No_of_children_born contains null values
- Of these, it is possible for the Number of children to be 0, hence need not be imputed
- There are blank values present in No_of_children_born. These are imputed with mode of the field
- There were 80 duplicate records, which were removed.
- No_of_children field has outliers, but not very significant

2.1. EDA

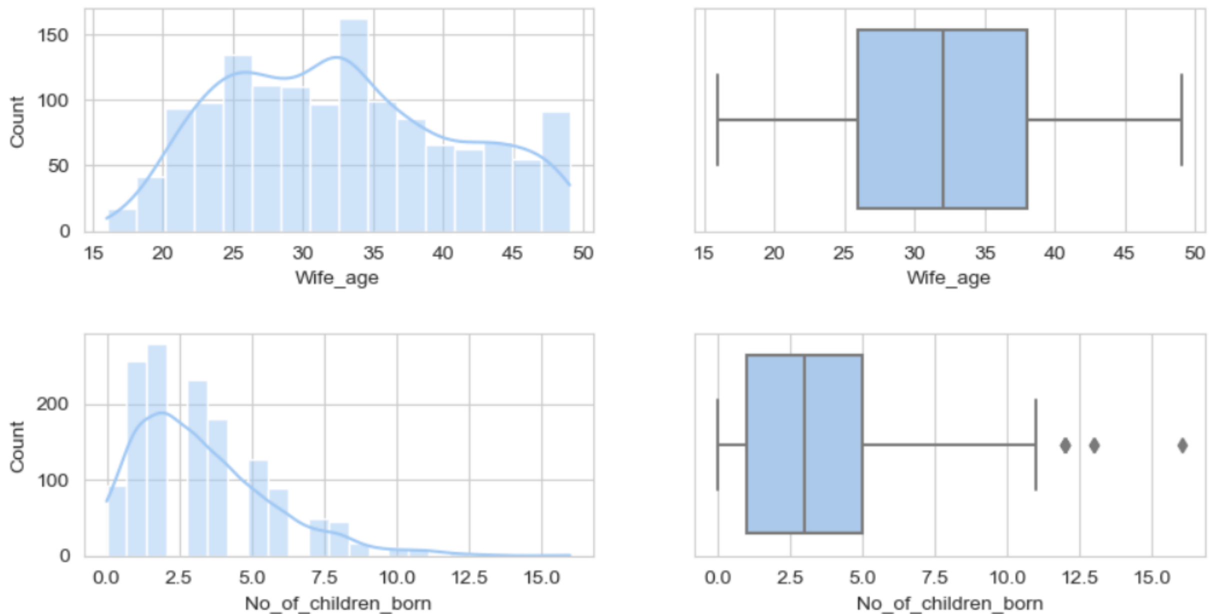


Fig.2.3. Univariate analysis- Numerical columns

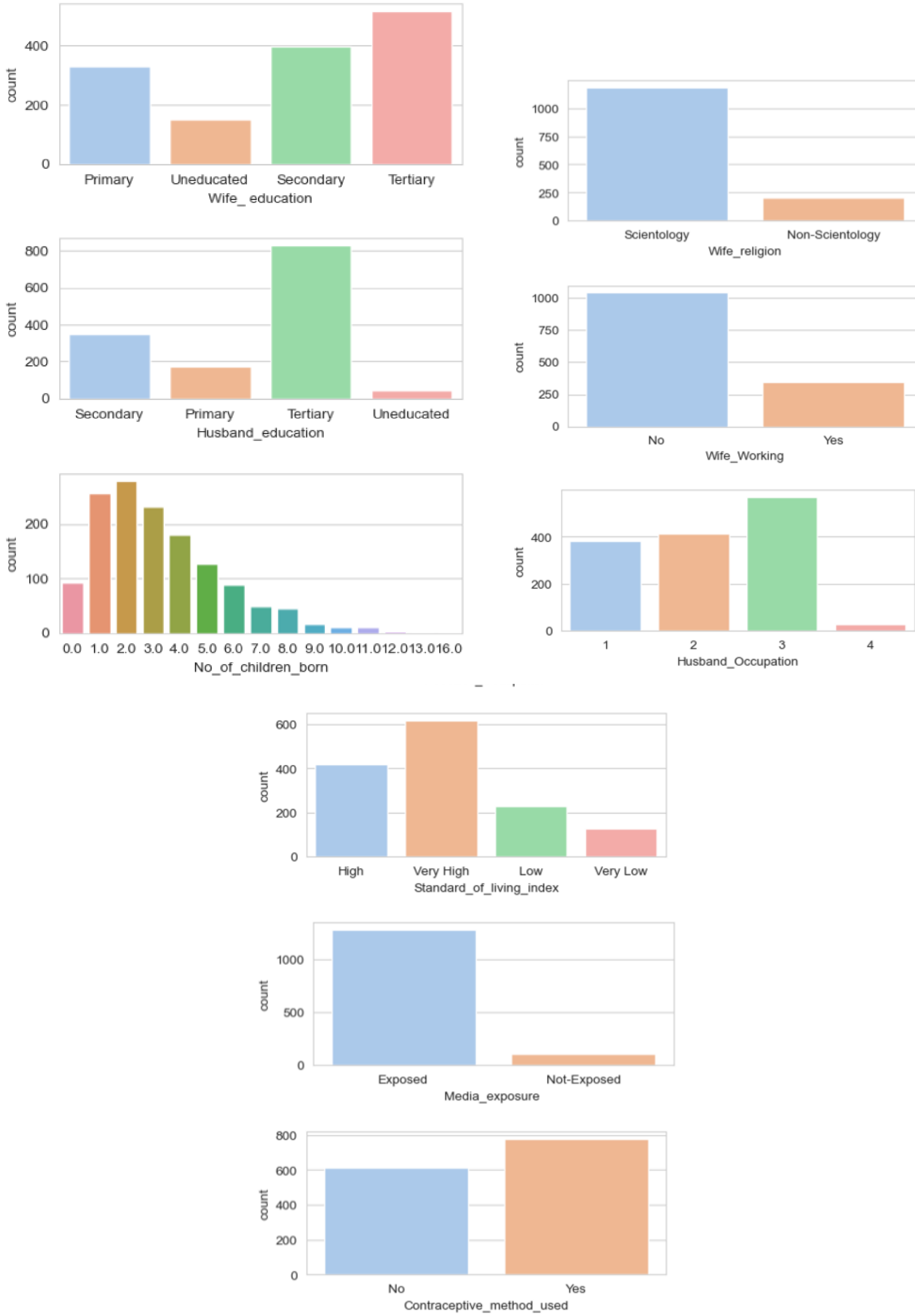
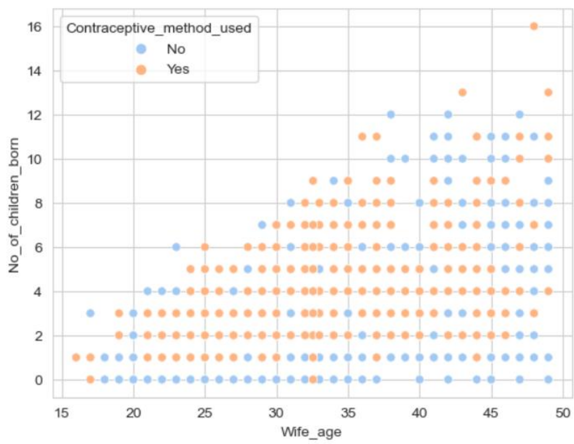
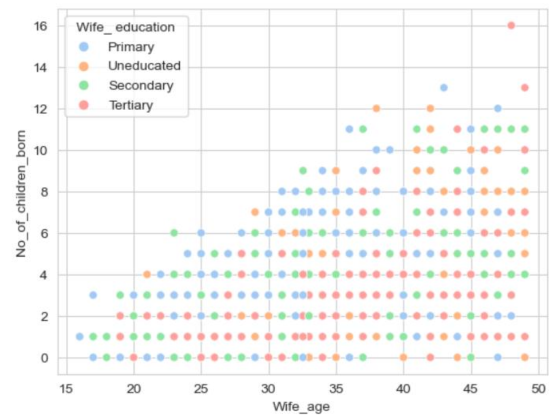


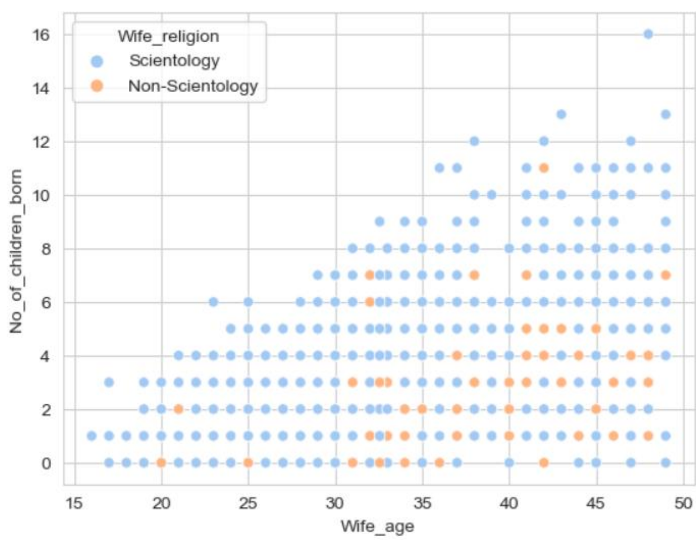
Fig.2.4. Univariate analysis- categorical columns

Observations:

- The age of wife is almost normally distributed
- There are a few outliers in the Number_of_children_born field
- Most women are educated, with the highest number of women having completed teritiary education
- The husband_education field also has maximum values in teritiary education field, with only a very low count of uneducated males.
- Most have upto 4 children. However, the data also shows people having more than 10 children. These might be genuine, or bad values.
- Most women adopt scientology as religion
- Most of the women are ot working
- A majority of the husbands have occupation -3
- This data has high counts of people having high or very high standard of living index
- Most of these people are media exposed
- most of these people have used contraceptive methods







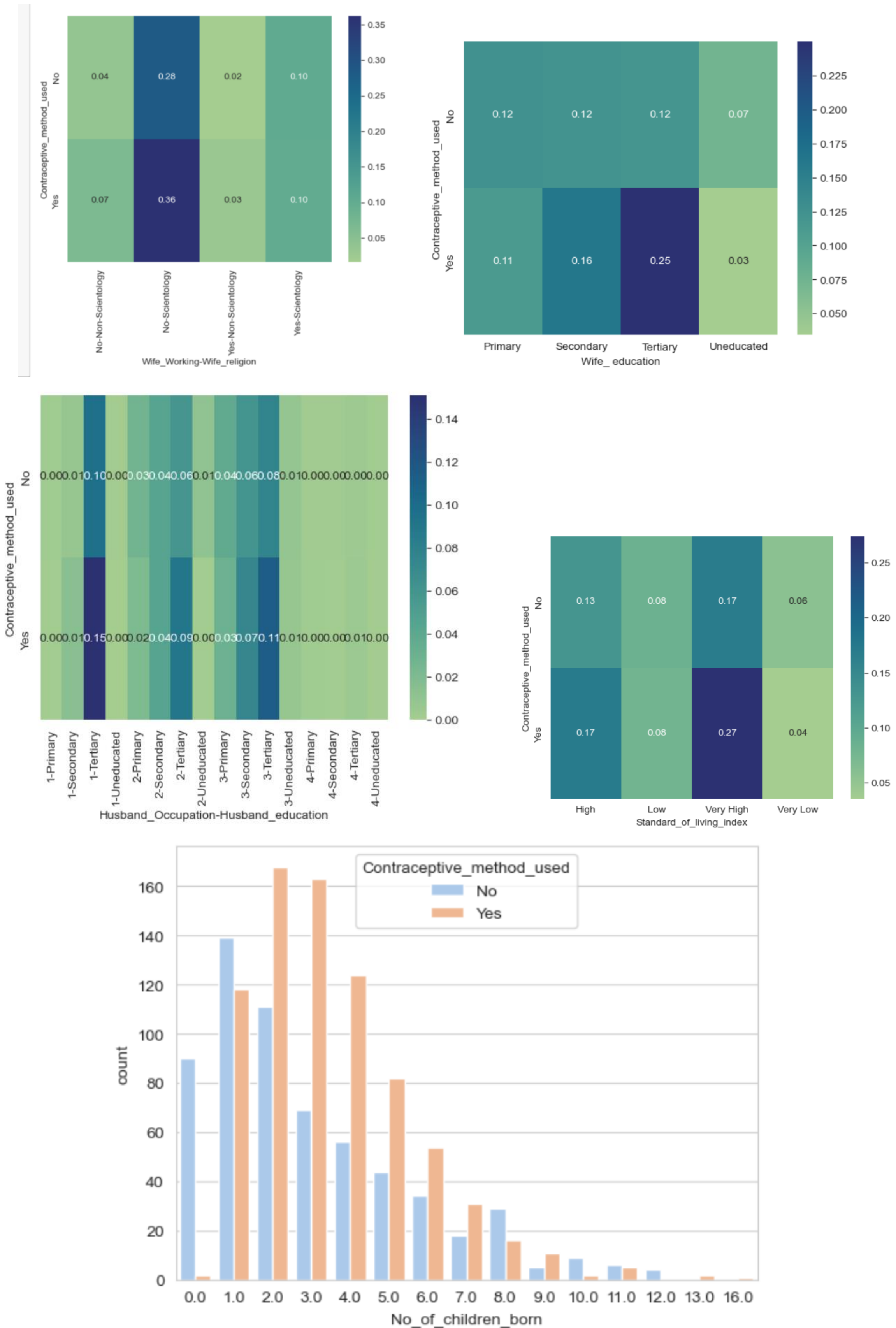


Fig.2.5. Multivariate Analysis

Observations:

- No discernable patterns emerge from the scatter plots
- Some minor correlation detected between the status and usage of contraceptives

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.**

   The following equations are obtained after the application of logistic regression and LDA respectively:

   log(contraceptive usage)=-0.425 - 0.077 * Wife_age + 0.567 * Wife_education - 0.048 * Husband_education + 0.308 * No_of_children_born - 0.265 * Wife_religion - 0.065 * Wife_working + 0.069 * Husband_occupation + 0.158 * Standard_of_living_index + 0.314 * Media_exposure

   contraceptive_usage= 0.286- 0.075 * Wife_age + 0.572 * Wife_education - 0.057 * Husband_education + 0.297 * No_of_children_born - 0.287 * Wife_religion - 0.071 * Wife_working + 0.069 * Husband_occupation + 0.158 * Standard_of_living_index + 0.299* Media_exposure

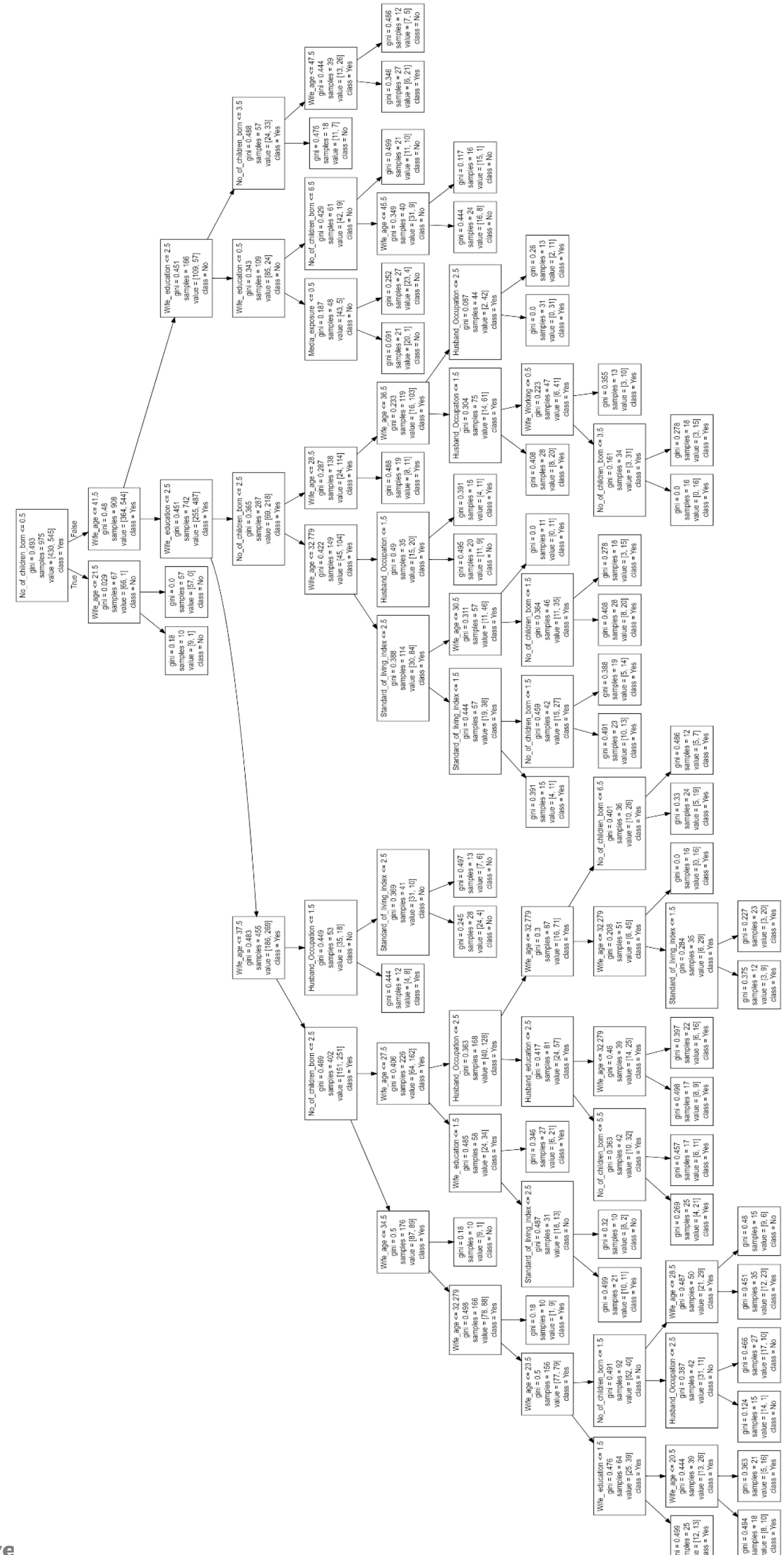


Fig 2.6. Regularized Decision Tree(Pruned)

```
Logistic Regression Score 1.0

Confusion Matrix:
 [[ 91  93]
 [ 55 179]]

 Classification Report:
               precision    recall  f1-score   support

           0       0.62      0.49      0.55       184
           1       0.66      0.76      0.71       234

    accuracy                           0.65       418
   macro avg       0.64      0.63      0.63       418
weighted avg       0.64      0.65      0.64       418
```

Fig.2.7. Logistic regression scores

```
LDA Score 1.0

Confusion Matrix:
 [[ 89  95]
 [ 52 182]]

 Classification Report:
               precision    recall  f1-score   support

           0       0.63      0.48      0.55       184
           1       0.66      0.78      0.71       234

    accuracy                           0.65       418
   macro avg       0.64      0.63      0.63       418
weighted avg       0.65      0.65      0.64       418
```

Fig. 2.8. LDA scores

```
Regularized DTC Model Score 1.0

Confusion Matrix:
 [[100  84]
 [ 47 187]]

 Classification Report:
               precision    recall  f1-score   support

           0       0.68      0.54      0.60       184
           1       0.69      0.80      0.74       234

    accuracy                           0.69       418
   macro avg       0.69      0.67      0.67       418
weighted avg       0.69      0.69      0.68       418
```

Fig.2.9. Regularized Decision Tree scores

Observations- Regularized Decision Tree:

- The output of the model is a pruned tree
- There are 10 levels
- 1 feature has a coefficient of 0- Wife_religion, and 2 have coefficients close to 0- Wife_working, Media_exposure.

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 each model Final Model: Compare Both the models and write inference which model is best/optimized.**

Train data:

Logistic Regression:

	precision	recall	f1-score	support
0	0.65	0.50	0.57	430
1	0.67	0.79	0.72	545
accuracy			0.66	975
macro avg	0.66	0.64	0.64	975
weighted avg	0.66	0.66	0.65	975

LDA:

	precision	recall	f1-score	support
0	0.66	0.48	0.56	430
1	0.66	0.81	0.73	545
accuracy			0.66	975
macro avg	0.66	0.64	0.64	975
weighted avg	0.66	0.66	0.65	975

CART-Decision Tree- regularized:

	precision	recall	f1-score	support
0	0.78	0.62	0.69	430
1	0.74	0.86	0.80	545
accuracy			0.76	975
macro avg	0.76	0.74	0.75	975
weighted avg	0.76	0.76	0.75	975

Test data:

Logistic Regression:

	precision	recall	f1-score	support
0	0.62	0.49	0.55	184
1	0.66	0.76	0.71	234
accuracy			0.65	418
macro avg	0.64	0.63	0.63	418
weighted avg	0.64	0.65	0.64	418

LDA:

	precision	recall	f1-score	support
0	0.63	0.48	0.55	184
1	0.66	0.78	0.71	234
accuracy			0.65	418
macro avg	0.64	0.63	0.63	418
weighted avg	0.65	0.65	0.64	418

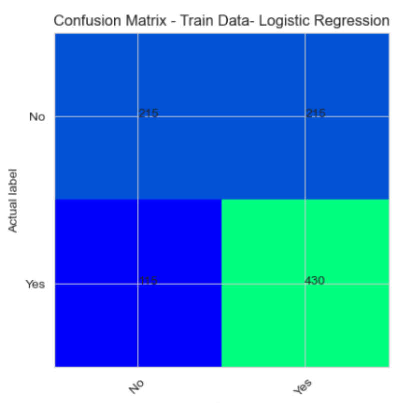
CART-Decision Tree- regularized:

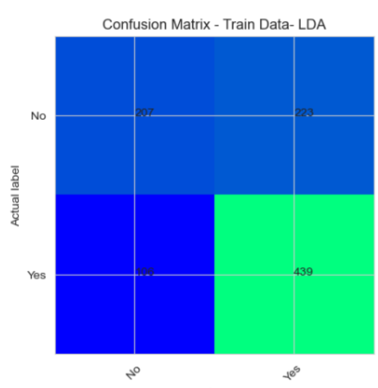
	precision	recall	f1-score	support
0	0.68	0.54	0.60	184
1	0.69	0.80	0.74	234
accuracy			0.69	418
macro avg	0.69	0.67	0.67	418
weighted avg	0.69	0.69	0.68	418

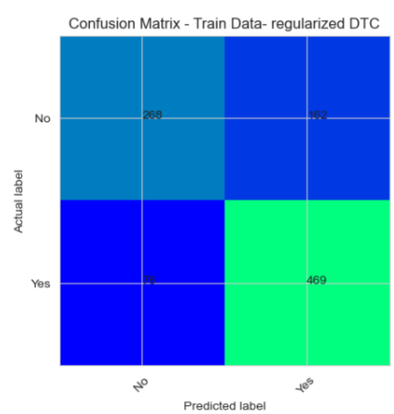
Fig.2.10. Accuracy and scores comparison across models for Train and test Data

Observations:

- Train Data:
  - CART is better in terms of accuracy- 0.76 against 0.66 for LDA and Logistic Regression
  - CART also gives better F1 scores for both the classes (0.69 and 0.80) as against LDA (0.56 and 0.73) and Logistic regression (0.57 and 0.72)
- Test Data:
  - CART is better in terms of accuracy- 0.69 against 0.65 for LDA and Logistic Regression
  - CART also gives better F1 scores for both the classes (0.60 and 0.74) as against LDA and Logistic regression (0.55 and 0.71)







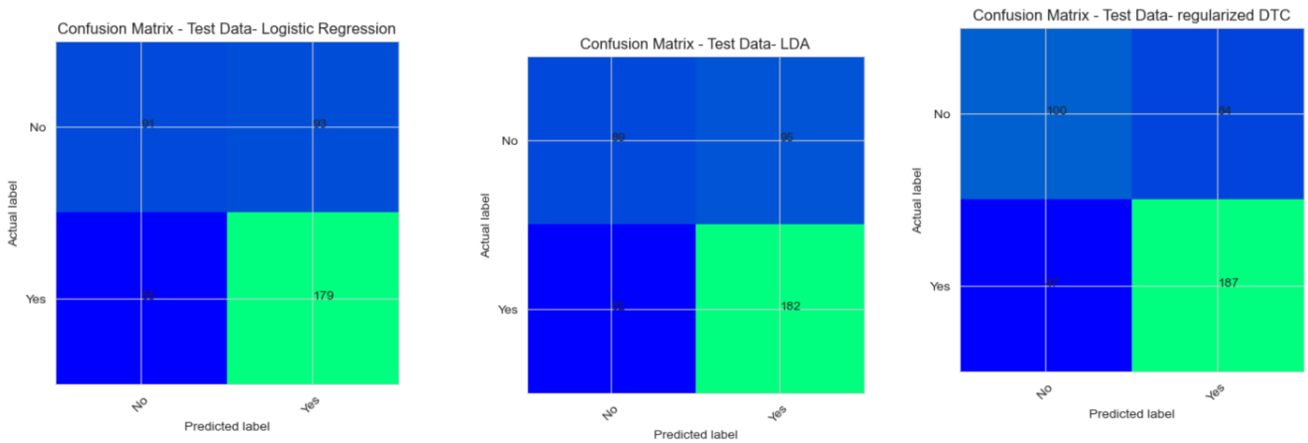


Fig.2.11. Confusion Matrix comparisons

	Logistic Regression		LDA		CART- Regularized	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Type I Error	**215**	**93**	**223**	**96**	**162**	**84**
Type II Error	**115**	**55**	**106**	**52**	**76**	**47**

Table 2.1. Type I and II Error comparison across models

Observations:

- From the above, it can be inferred that CART model outperforms the other two

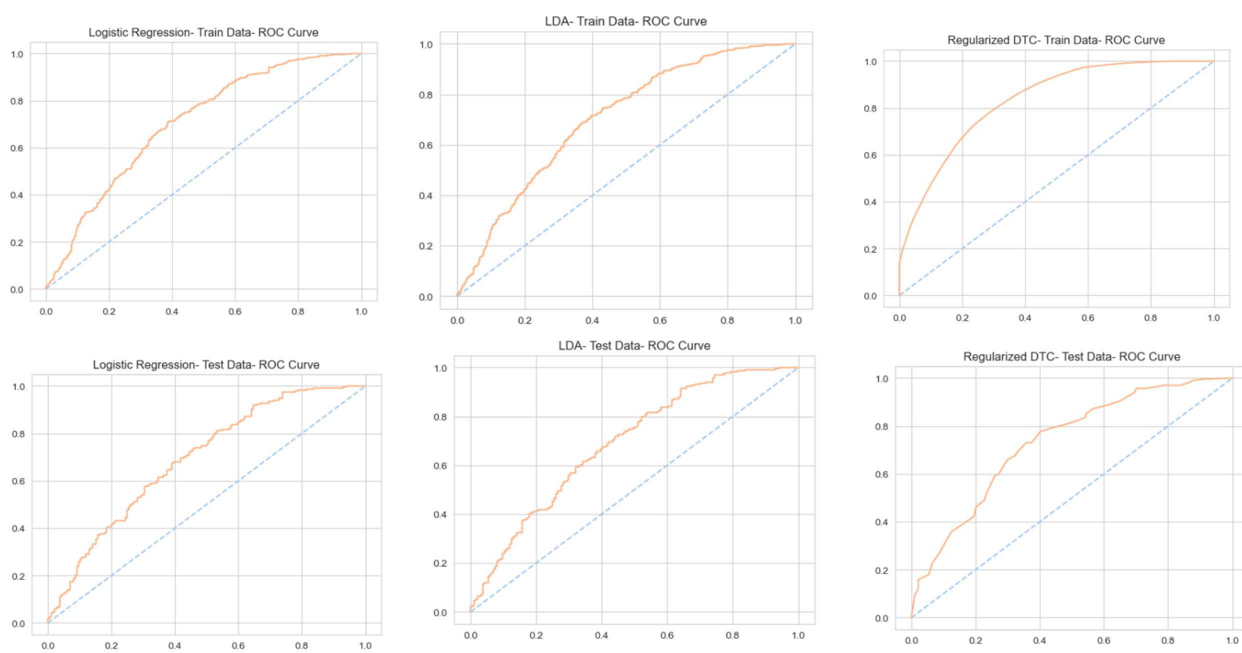


Fig.2.13. ROC Curves across models for train and test data

Train Data: AUC Scores:	Test Data: AUC Scores:
Logistic Regression: 0.704	Logistic Regression: 0.691
LDA: 0.703	LDA: 0.69
Regularized DTC: 0.833	Regularized DTC: 0.733

Fig.2.14. AUC scores comparison

Observations:

- From the graphs and scores above, it can be inferred that of the three models, CART gives better fit in both train and test sets.

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

Summary:

- For the given dataset, the performance of LDA and logistic regression were similar. However, CART gave better results
- This can be an indication of a non-linear relation between the predictors and target
- Based on the equations derived from LDA and logistic regression, the features that create maximum separability are Wife_education, Media_exposure and Number_of_children_born
- Hence, in order to improve contraceptive usage these actors could be targeted