PREDICTIVE MODELLING PROJECT REPORT

Prepared By: Barkha Agarwal

PROBLEM STATEMENT-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.

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

Perform the following in given order:

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, Multivariate Analysis.

1.2 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.

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 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.

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.

INTRODUCTION TO PROBLEM STATEMENT-1

The purpose of this business report is perform linear regression model using various attributes. We will build linear equation model to predict 'usr'(Portion of time (%) that cpus run in user mode) and to find out how each attribute affects the system to be in 'usr' mode using a list of system attributes.

1.1 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.

We will import the file **compactiv.xlsx**
Sample of the dataset.

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	ppgout	pgfree	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap	usr
0	1	0	2147	79	68	0.2	0.2	40671.0	53995.0	0.0	0.0	0.0	0.0	0.0	1.6	2.6	16.00	26.40	CPU_Bound	4670	1730946	95
1	0	0	170	18	21	0.2	0.2	448.0	8385.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	15.63	16.83	Not_CPU_Bound	7278	1869002	97
2	15	3	2162	159	119	2.0	2.4	NaN	31950.0	0.0	0.0	0.0	0.0	1.2	6.0	9.4	150.20	220.20	Not_CPU_Bound	702	1021237	87
3	0	0	160	12	16	0.2	0.2	NaN	8670.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	15.60	16.80	Not_CPU_Bound	7248	1863704	98
4	5	1	330	39	38	0.4	0.4	NaN	12185.0	0.0	0.0	0.0	0.0	0.0	1.0	1.2	37.80	47.60	Not_CPU_Bound	633	1760253	90

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

```
 #   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
```

There are a total of 8192 rows and 22 columns in the dataset out of which 13 are float data types, 8 are integer data types and 1 object data type.

We will then perform univariate analysis, bivariate analysis and multivariate analysis to get a clear picture of the variables. There are many variables so the figure is not attached here. We can refer to the code file attached to get the picture.

1.2 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.

Check for missing values in the dataset:

```
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
freemem                    0
freeswap                   0
usr                        0
runqsz_Not_CPU_Bound       0
dtype: int64
```

From the above results we can see that there are missing values present in the dataset in variables rchar(Number of characters transferred per second by system read calls) and wchar(Number of characters transferred per second by system write calls).

We will impute 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
```

After imputing the missing values with mean, we can see that there are no missing values in the dataset now.

In the dataset we can see that there are many zeroes present in many variables and they hold meaning also, so we don't need to change them or drop them.

Here, we have to create a new feature which is converting the categorical variable to dummy variables. The variable runqsz(Process run queue size (The number of kernel threads in memory that are waiting for a CPU to run) has categorical variables.

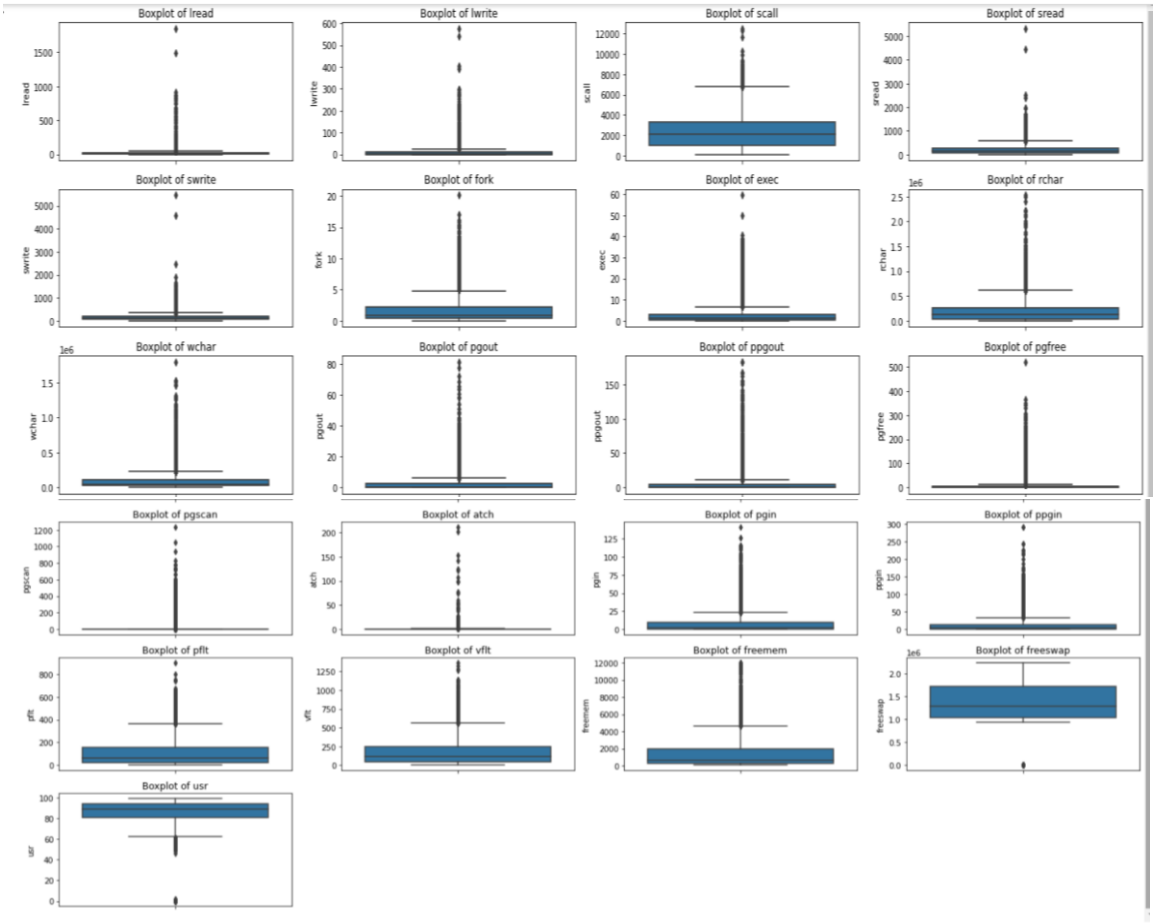
There are two categorical data present in column runqsz which will be replaced by 0 and 1.

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	freemem	freeswap	usr	runqsz_Not_CPU_Bound
0	1	0	2147	79	68	0.2	0.2	40671.000000	53995.0	0.0	...	0.0	0.0	1.6	2.6	16.00	26.40	4670	1730946	95	0
1	0	0	170	18	21	0.2	0.2	448.000000	8385.0	0.0	...	0.0	0.0	0.0	0.0	15.63	16.83	7278	1869002	97	1
2	15	3	2162	159	119	2.0	2.4	197385.728363	31950.0	0.0	...	0.0	1.2	6.0	9.4	150.20	220.20	702	1021237	87	1
3	0	0	160	12	16	0.2	0.2	197385.728363	8670.0	0.0	...	0.0	0.0	0.2	0.2	15.60	16.80	7248	1863704	98	1
4	5	1	330	39	38	0.4	0.4	197385.728363	12185.0	0.0	...	0.0	0.0	1.0	1.2	37.80	47.60	633	1760253	90	1

5 rows × 22 columns

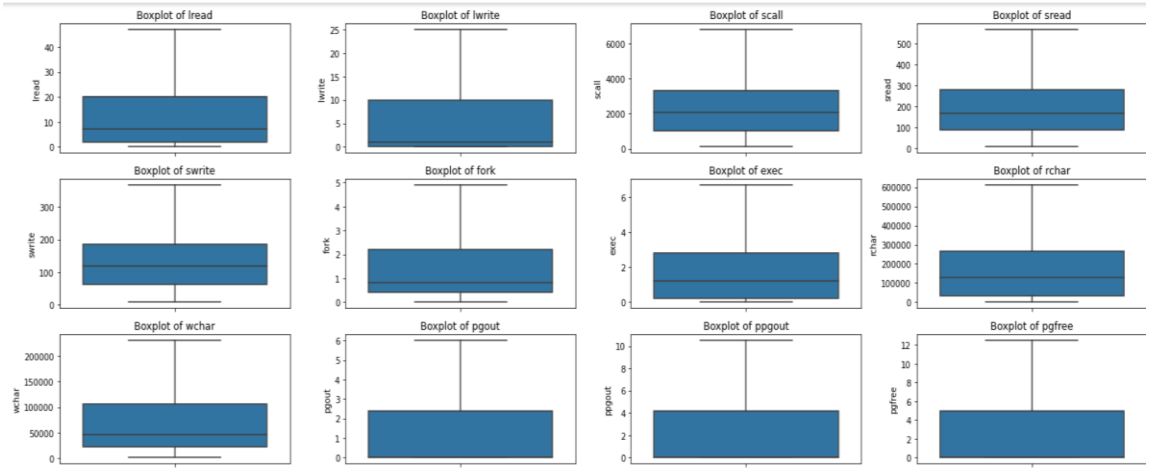
Check for outliers:

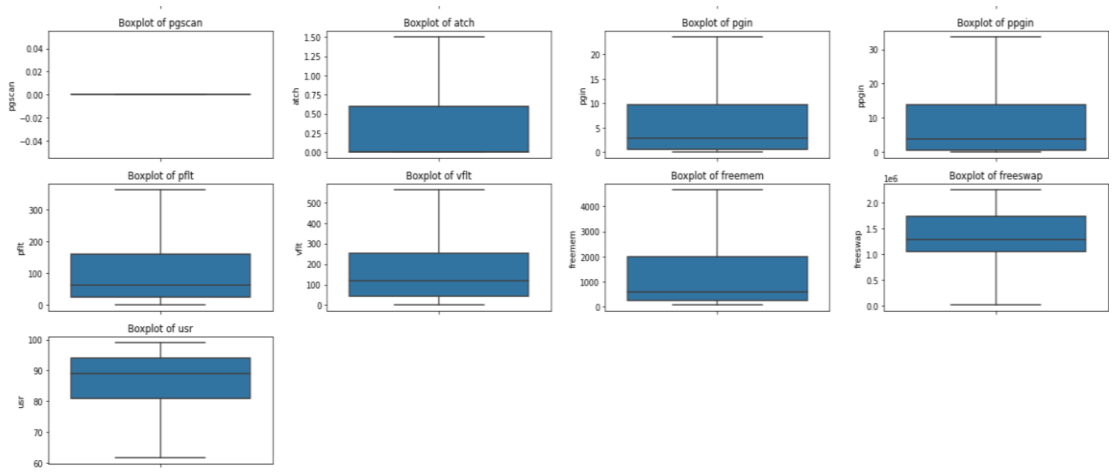
By plotting boxplots we can check whether there are outliers present in the dataset. Outliers can be checked with numerical data with continuous data.



By the above diagram we can see that there are outliers present in the dataset which needs to be treated.

Below is the diagram after treating the outliers.





Checking for duplicates:
There are no duplicates present in the dataset.

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 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.

Train-Test split:
We will split the data into X and Y. X will contain all the independent features and Y will consist of target label. We will split the data here using Linear regression scikit learn in the ratio of 70:30.
Linear Regression Model:
Firstly, we will train the model and get all the features after training the model.
After training the model, we will find out the co-efficients which is the beta co-efficient.

```
The coefficient for lread is -0.0634099731929712
The coefficient for lwrite is 0.04801838613982387
The coefficient for scall is -0.000664352337468163
The coefficient for sread is 0.0003385875789163073
The coefficient for swrite is -0.0054598818142991765
The coefficient for fork is 0.0296329957065836
The coefficient for exec is -0.3210632504852232
The coefficient for rchar is -5.2118791560622775e-06
The coefficient for wchar is -5.346335455397222e-06
The coefficient for pgout is -0.3668522981358798
The coefficient for ppgout is -0.07860920074522432
The coefficient for pgfree is 0.08525820125741179
The coefficient for pgscan is 4.440892098500626e-16
The coefficient for atch is 0.6304380351240503
The coefficient for pgin is 0.01975385591267015
The coefficient for ppgin is -0.06715372112646986
The coefficient for pflt is -0.03359198966258064
The coefficient for vflt is -0.005464921751511641
The coefficient for freemem is -0.0004576614008076046
The coefficient for freeswap is 8.829320420421227e-06
The coefficient for runqsz_Not_CPU_Bound is 1.6137372384501187
```

Here, we can see that there are some negative values as co-efficients.
After this we will find the intercept of the model. Here, the intercept of our model is 84.13.
The R square of the training model is 79.6%.
The R square of the testing model is 76.8%.

79.6% of the variation in the usr(Portion of time (%) that cpus run in user mode) is explained by the predictors in the model for train set. We can say that the model is performing good.
We will calculate RMSE value as well for train and test data.
RMSE value for training data is 4.42
RMSE value for testing data is 4.65

Linear Regression using Statsmodel:
Here we will take the data in original form as we have all the independent and dependent features in the training data. So, here we have all the features in the training data as well as testing data.
The concept here is to create a formula or expression.
We will get the same beta value which we got earlier.
After getting the summary of the model, we will focus on the probability value, the co-efficients and the model which has been taken here.

[ ]

Dep. Variable:	usr	R-squared:	0.796
Model:	OLS	Adj. R-squared:	0.795
Method:	Least Squares	F-statistic:	1116.
Date:	Wed, 07 Dec 2022	Prob (F-statistic):	0.00
Time:	04:34:09	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]
Intercept	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
pgscan	5.543e-15	5.16e-17	107.427	0.000	5.44e-15	5.64e-15
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

From the above image we can see that sread(Number of system read calls per second), fork(Number of system fork calls per second), ppgout(Number of pages, paged out per second), pgfree(Number of pages per second placed on the free list), pgin(Number of pages paged in per second) are not significant to the model.
If p value is less than 0.05 then it means that it is significant to the model.
We can also check the R square and the adjusted R square values here.

After the we will calculate the MSE value, which is (y^ - y)2 =19.53

And, Root Mean Squared Error – RMSE = 4.42

We can see from above two models that the output we are getting is same for both the models. Hence, we can use either Linear regression using scikit learn or Linear Regression using statsmodel.

According to my analysis statsmodel is a better option to get the values of Rsquare, RMSE & Adj Rsquare in one single table.

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.

By performing the calculation using two models for calculating Linear Regression using scikit learn and statsmodel we get the below linear regression equation.

usr = b0 + b1 * lread + b2 * lwrite + b3 * scall + b4 * sread + b5 * swrite + b6 * fork + b7 * exec + b8 * rchar + b9 * wchar + b10 * pgout + b11 * ppgout + b12 * pgfree + b13 * pgscan + b14 * atch + b15 * pgin + b16 * ppgin + b17* pflt + b18* vflt + b19* freemem + b20* freeswap +b21* runqsz_Not_CPU_Bound

usr=(84.13) * Intercept + (-0.06) * lread + (0.05) * lwrite + (-0.0) * scall + (0.0) * sread + (-0.01) * swrite + (0.03) * fork + (-0.32) * exec + (-0.0) * rchar + (-0.0) * wchar + (-0.37) * pgout + (-0.08) * ppgout + (0.09) * pgfree + (0.0) * pgscan + (0.63) * atch + (0.02) * pgin + (-0.07) * ppgin + (-0.03) * pflt + (-0.01) * vflt + (-0.0) * freemem + (0.0) * freeswap + (1.61) * runqsz_Not_CPU_Bound

By looking into the equation we can assume that When lwrite increases by 1 unit, usr increases by 0.05 units keeping all other predictors constant. similarly, when fork increases by 1 unit, usr increases by 0.03 units keeping all other predictors constant.

There are also some negative co-efficient values, for instance, lread has negative coefficient which means that when lread(Reads (transfers per second) between system memory and user memory) the usr descreases by -0.06 units.

We have performed various steps to get this linear regression equation. The steps performed are:

- Loading the data.
- Performing EDA(Exploratory Data Analysis)
- Performing Univariate, Bivariate and Multivariate analysis.
- Checking of missing values and replacing them with mean.
- Converting categorical variables to dummy variables using 0 and 1.
- Performing outlier check and treating them.
- Checking for duplicates.
- Splitting the data into train and test.
- Performing linear regression using scikit learn.
- Performing linear regression using statsmodel.
- Drawing conclusion.

PROBLEM STATEMENT-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 condition check, check for duplicates and outliers and write an inference on it. Perform Univariate and Bivariate Analysis and Multivariate Analysis.

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.

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

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.

INTRODUCTION TO PROBLEM STATEMENT-2

Here we will perform logistic regression, linear discriminant analysis (LDA) and CART to study the samples of 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 if the do/don't use a contraceptive method of choice based on their demographic and socio-economic characteristics.

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.

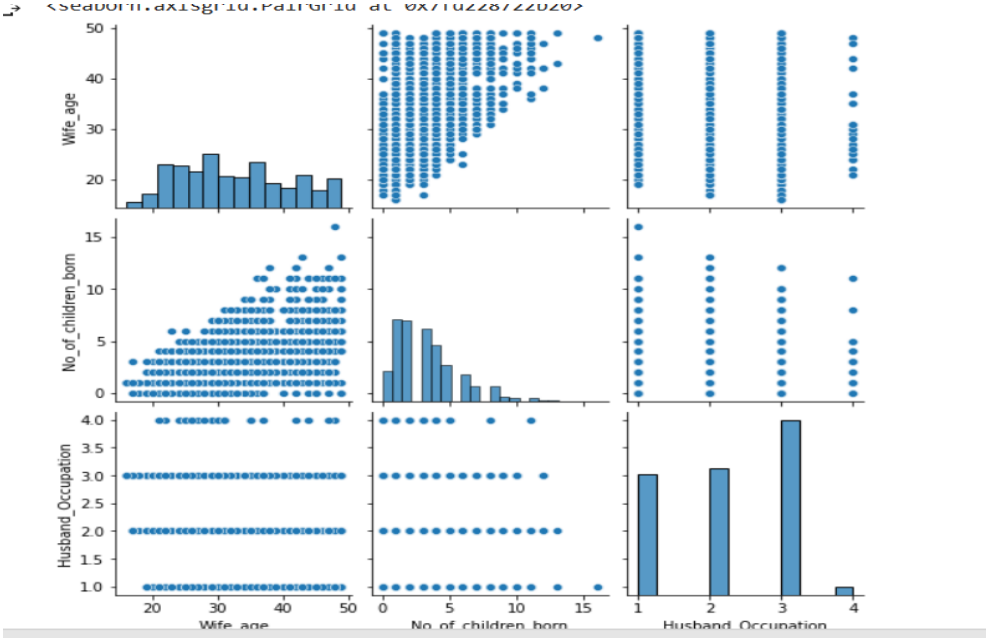
We will import file Contraceptive_method_dataset.xlsx

Then get a summary of the dataset provided to us.i.e; checking the data type.

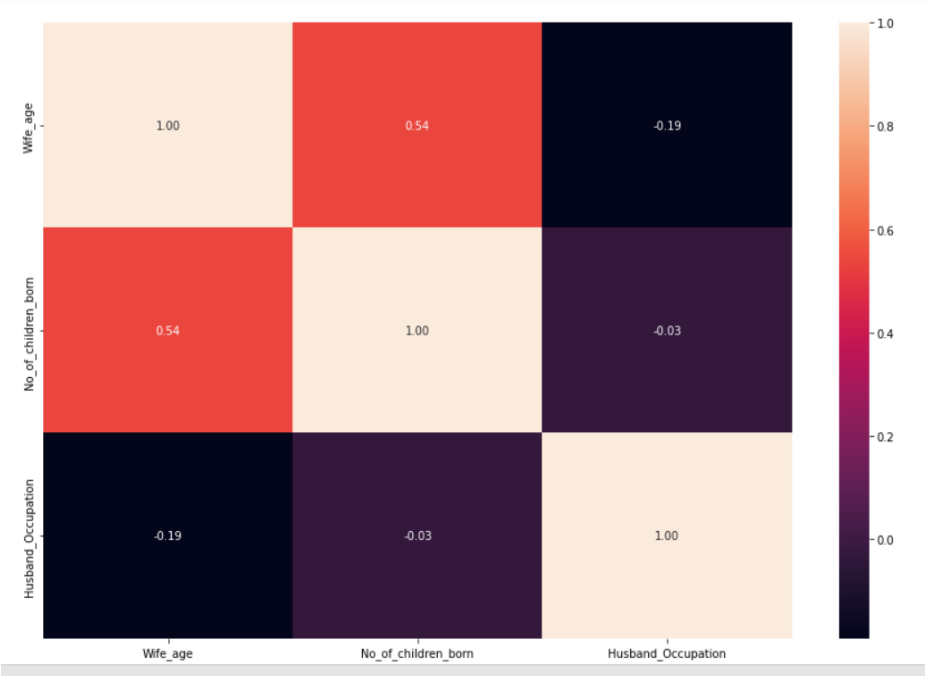
```
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Wife_age                    1402 non-null   float64
 1   Wife_ education              1473 non-null   object
 2   Husband_education           1473 non-null   object
 3   No_of_children_born         1452 non-null   float64
 4   Wife_religion               1473 non-null   object
 5   Wife_Working                1473 non-null   object
 6   Husband_Occupation          1473 non-null   float64
 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(3), object(7)
memory usage: 115.2+ KB
```

We will check the detailed summary of the data, the shape of the dataset, also we will check if any duplicates are there in the data set.

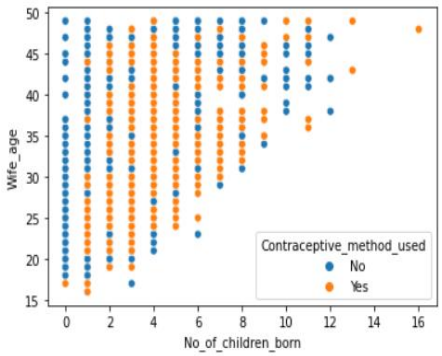
After performing the basic check we will perform univariate analysis, bivariate analysis, multivariate analysis. We will refer to the code file for the graphs we have plotted.
The bivariate analysis chart has been attached here.

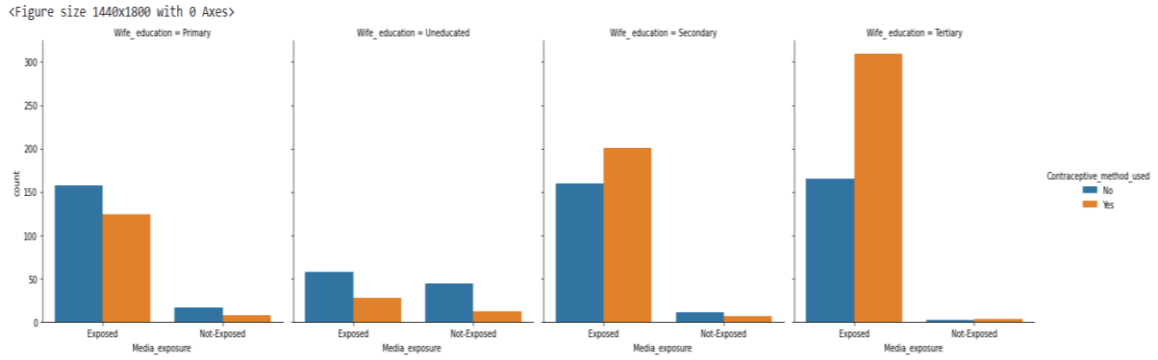


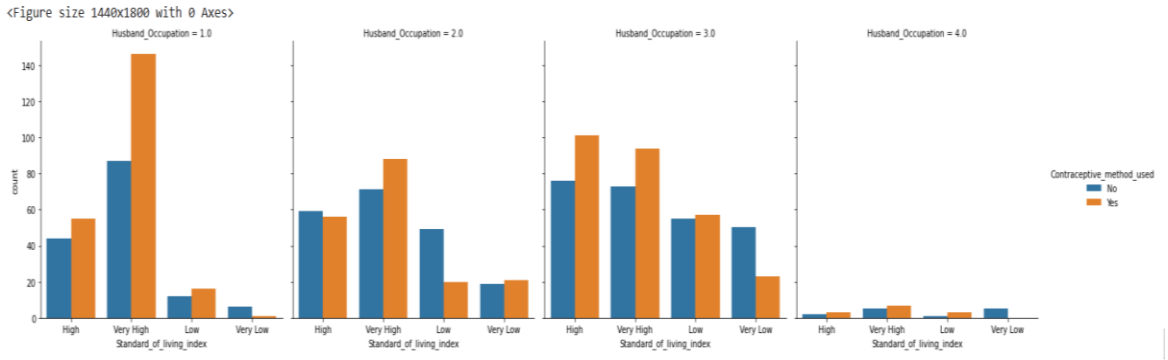
Below is the heatmap attached.

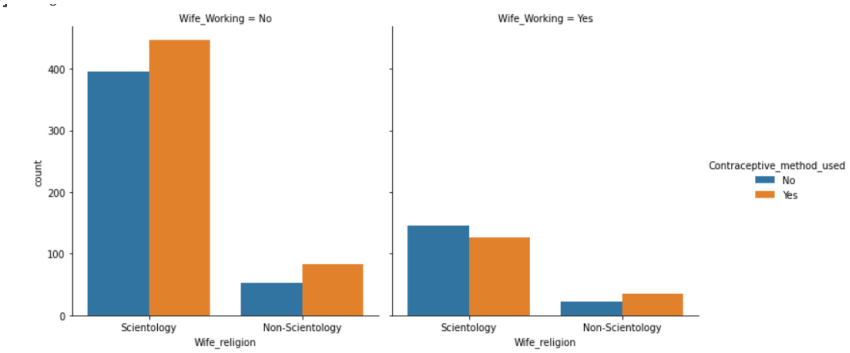


Charts for multivariate analysis are attached below.

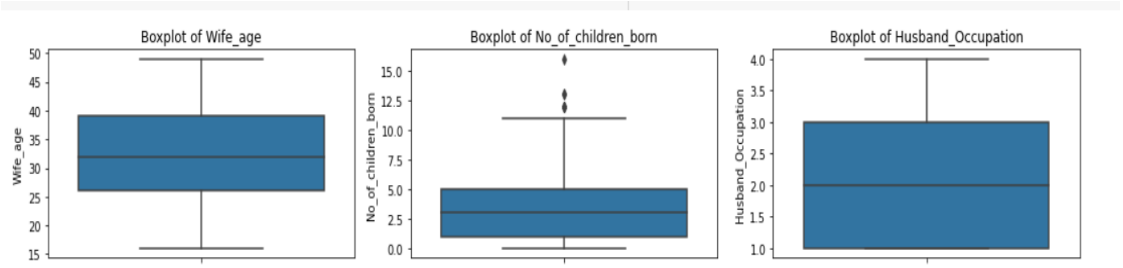




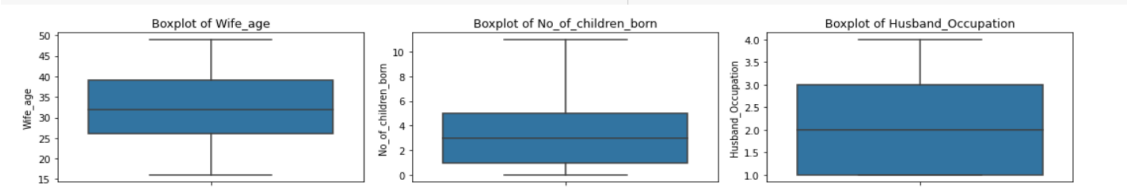




We will check for outliers and treat the outliers.



We can see that there are outliers present in the dataset.
Below is the figure attached after treating the outliers.



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.

After encoding the data structure looks like this:

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
24.0	1	2	3.0	0	0	2.0	0	1	0
45.0	0	2	10.0	0	0	3.0	1	1	0
43.0	1	2	7.0	0	0	3.0	1	1	0
42.0	2	1	9.0	0	0	3.0	0	1	0
36.0	2	2	8.0	0	0	3.0	2	1	0

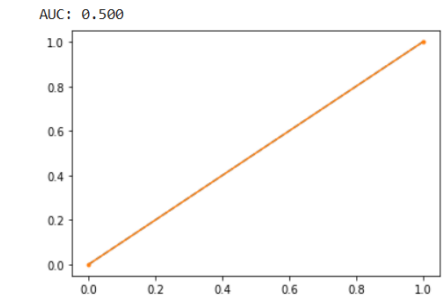
After encoding the data and we will split the data into 70/30 ratio.
We will build a LDA model.

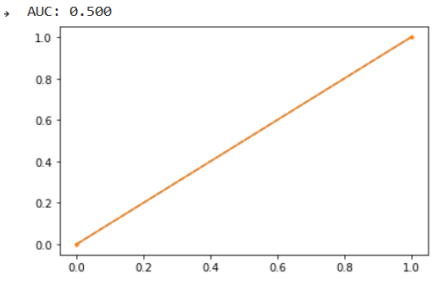
2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

We will build a classification report and then we will change the cut off values. The threshold value taken here is 0.5 and we will run a loop here which will run from 0.1 to 0.9.
We then build a decision tree classifier.
We will see the variables important for us in this dataset.

	Imp
Wife_age	0.274976
Standard_of_living_index	0.144742
No_of_children_born	0.141732
Husband_Occupation	0.126919
Wife_ education	0.126215
Husband_education	0.060689
Wife_Working	0.054017
Wife_religion	0.053823
Media_exposure	0.016887

Now we will go for model evaluation.
Measuring AUC-ROC curve for training and testing data.





Training Testing

Now we will build a confusion matrix foe training and testing data.

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.

Accuracy of training data is 55% Accuracy of testing data is 48%

AUC of training and testing data is the same as 50% Accuracy, AUC, Precision and Recall of test data is almost inline with training data.

This proves that no over fitting and under fitting has happened, and overall the model is good.

The important variables are Wife's age, Standard-of-living index, Number of children ever born, Husband's occupation, Wife's education.