Predictive Modelling [PM] Project Report

By

Name: Abhishek Pradhan

BATCH: PGPDSBA.O.FEB23.B

Contents

Problem 1: **Linear Regression**

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

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

Dataset for Problem 1: compactiv.xlsx

DATA DICTIONARY:

System measures used:

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

Problem 2: Logistic Regression, LDA and CART

You are a statistician at the Republic of Indonesia Ministry of Health and you are provided with a data of 1473 females collected from a Contraceptive Prevalence Survey. The samples are

married women who were either not pregnant or do not know if they were at the time of the survey.

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

Dataset for Problem 2: Contraceptive_method_dataset.xlsx

Data Dictionary:

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

Problem 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.

Ans.:

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freesv
0	1	0	2147	79	68	0.2	0.20	40671.0	53995.0	0.00	...	0.00	0.0	1.60	2.60	16.00	26.40	CPU_Bound	4670	1730
1	0	0	170	18	21	0.2	0.20	448.0	8385.0	0.00	...	0.00	0.0	0.00	0.00	15.63	16.83	Not_CPU_Bound	7278	1869
2	15	3	2162	159	119	2.0	2.40	NaN	31950.0	0.00	...	0.00	1.2	6.00	9.40	150.20	220.20	Not_CPU_Bound	702	1021
3	0	0	160	12	16	0.2	0.20	NaN	8670.0	0.00	...	0.00	0.0	0.20	0.20	15.60	16.80	Not_CPU_Bound	7248	1863
4	5	1	330	39	38	0.4	0.40	NaN	12185.0	0.00	...	0.00	0.0	1.00	1.20	37.80	47.60	Not_CPU_Bound	633	1760
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
8187	16	12	3009	360	244	1.6	5.81	405250.0	85282.0	8.02	...	55.11	0.6	35.87	47.90	139.28	270.74	CPU_Bound	387	986
8188	4	0	1596	170	146	2.4	1.80	89489.0	41764.0	3.80	...	0.20	0.8	3.80	4.40	122.40	212.60	Not_CPU_Bound	263	1055
8189	16	5	3116	289	190	0.6	0.60	325948.0	52640.0	0.40	...	0.00	0.4	28.40	45.20	60.20	219.80	Not_CPU_Bound	400	969
8190	32	45	5180	254	179	1.2	1.20	62571.0	29505.0	1.40	...	18.04	0.4	23.05	24.25	93.19	202.81	CPU_Bound	141	1022
8191	2	0	985	55	46	1.6	4.80	111111.0	22256.0	0.00	...	0.00	0.2	3.40	6.20	91.80	110.00	CPU_Bound	659	1756

8192 rows × 22 columns

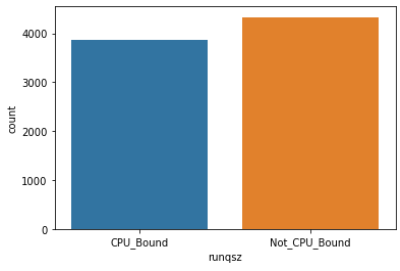
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8192 entries, 0 to 8191
Data columns (total 22 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   lread     8192 non-null   int64
 1   lwrite    8192 non-null   int64
 2   scall     8192 non-null   int64
 3   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
```

There are a total of 8192 rows and 22 columns in the dataset.

Out of 22, 13 are float 8 are integer type and 1 object type variable

	count	mean	std	min	25%	50%	75%	max
lread	8192.0	1.955969e+01	53.353799	0.0	2.0	7.0	20.000	1845.00
lwrite	8192.0	1.310620e+01	29.891726	0.0	0.0	1.0	10.000	575.00
scall	8192.0	2.306318e+03	1633.617322	109.0	1012.0	2051.5	3317.250	12493.00
sread	8192.0	2.104800e+02	198.980146	6.0	86.0	166.0	279.000	5318.00
swrite	8192.0	1.500582e+02	160.478980	7.0	63.0	117.0	185.000	5456.00
fork	8192.0	1.884554e+00	2.479493	0.0	0.4	0.8	2.200	20.12
exec	8192.0	2.791998e+00	5.212456	0.0	0.2	1.2	2.800	59.56
rchar	8088.0	1.973857e+05	239837.493526	278.0	34091.5	125473.5	267828.750	2526649.00
wchar	8177.0	9.590299e+04	140841.707911	1498.0	22916.0	46619.0	106101.000	1801623.00
pgout	8192.0	2.285317e+00	5.307038	0.0	0.0	0.0	2.400	81.44
ppgout	8192.0	5.977229e+00	15.214590	0.0	0.0	0.0	4.200	184.20
pgfree	8192.0	1.191971e+01	32.363520	0.0	0.0	0.0	5.000	523.00
pgscan	8192.0	2.152685e+01	71.141340	0.0	0.0	0.0	0.000	1237.00
atch	8192.0	1.127505e+00	5.708347	0.0	0.0	0.0	0.600	211.58
pgin	8192.0	8.277960e+00	13.874978	0.0	0.6	2.8	9.765	141.20
ppgin	8192.0	1.238859e+01	22.281318	0.0	0.6	3.8	13.800	292.61
pflt	8192.0	1.097938e+02	114.419221	0.0	25.0	63.8	159.600	899.80
vflt	8192.0	1.853158e+02	191.000603	0.2	45.4	120.4	251.800	1365.00
freemem	8192.0	1.763456e+03	2482.104511	55.0	231.0	579.0	2002.250	12027.00
freeswap	8192.0	1.328126e+06	422019.426957	2.0	1042623.5	1289289.5	1730379.500	2243187.00
usr	8192.0	8.396887e+01	18.401905	0.0	81.0	89.0	94.000	99.00

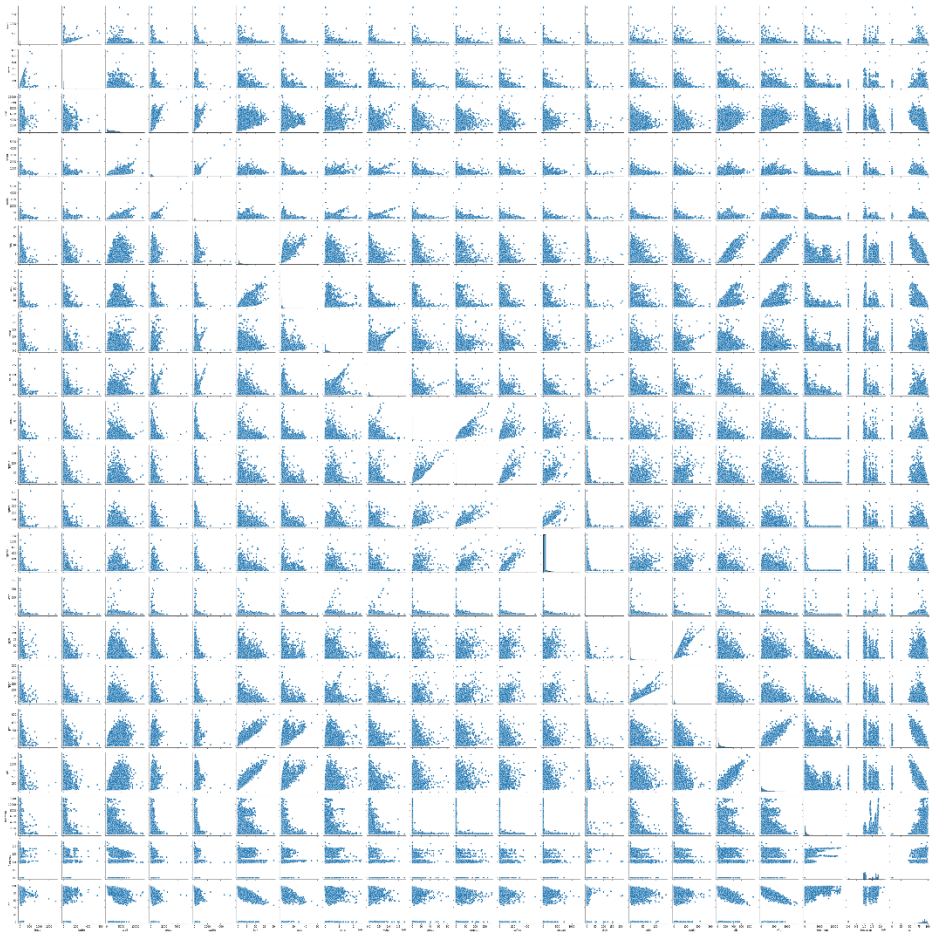
Univariate analysis for categorical data:



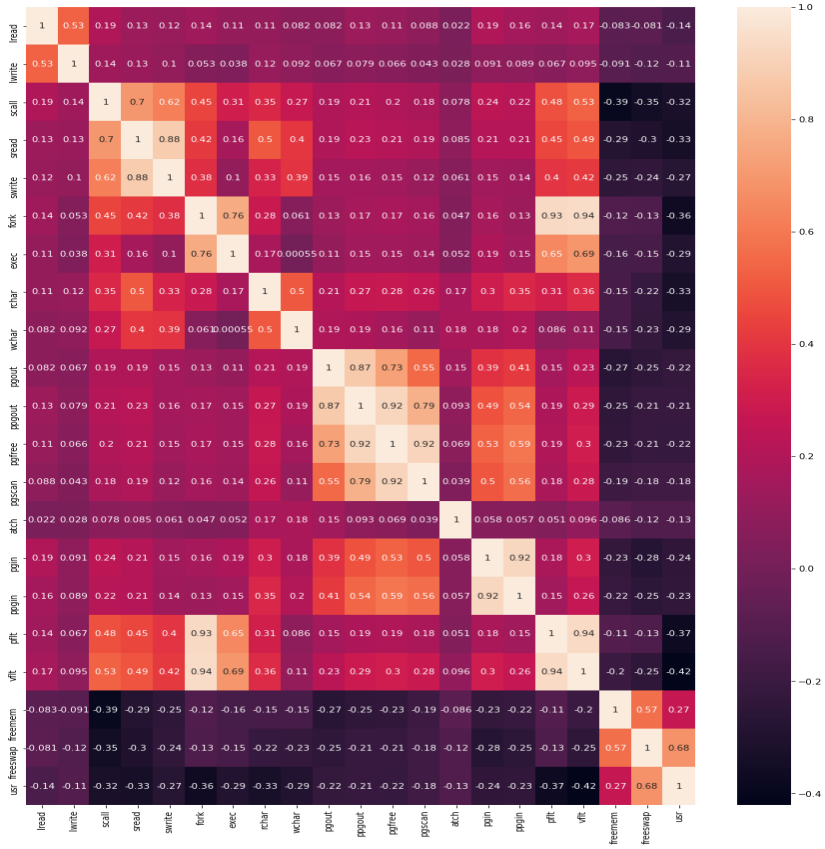
From the analysis we can say we have total ‘Process run queue size’

Not_CPU_Bound -4331 and CPU_Bound - 38

Bivariate analysis: Pairplot (pairwise relationships between variables):



Bivariate analysis: Heatmap (Check for presence of correlations):



we can see the presence of correlations.

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

Ans.:

```
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
```

There are null values at ‘rchar’ – 104 and ‘wchar’-15 whereas no duplicates.
So we can impute the null values with the mean.

lread	675	lread	8.239746
lwrite	2684	lwrite	32.763672
scall	0	scall	0.000000
sread	0	sread	0.000000
swrite	0	swrite	0.000000
fork	21	fork	0.256348
exec	21	exec	0.256348
rchar	0	rchar	0.000000
wchar	0	wchar	0.000000
pgout	4878	pgout	59.545898
ppgout	4878	ppgout	59.545898
pgfree	4869	pgfree	59.436035
pgscan	6448	pgscan	78.710938
atch	4575	atch	55.847168
pgin	1220	pgin	14.892578
ppgin	1220	ppgin	14.892578
pflt	3	pflt	0.036621
vflt	0	vflt	0.000000
runqsz	0	runqsz	0.000000
freemem	0	freemem	0.000000
freeswap	0	freeswap	0.000000
usr	283	usr	3.454590
dtype: int64		dtype: float64	

These many Zeros are present in the dataset and their percentages.

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 appropriatemethod from stats model. Create multiple models and check the performance of Predictions on Trainand Test sets using R-square, RMSE & Adj R-square. Compare these models and select the best onewith appropriate reasoning.

Ans:

	Variable	VIF
0	lread	inf
1	lwrite	6.423744
2	scall	9.017225
3	sread	18.594655
4	swrite	16.966453
5	fork	25.333287
6	exec	5.955285
7	rchar	4.253690
8	wchar	3.352619
9	pgout	16.205025
10	ppgout	43.013793
11	pgfree	24.106513
12	pgscan	NaN
13	atch	2.750902
14	pgin	23.215304
15	ppgin	23.342616
16	pflt	24.272827
17	vflt	32.809519
18	runqsz	inf
19	freemem	3.428430
20	freeswap	24.692144
21	usr	24.342889

From the above I can conclude that variables have moderate correlation.

Training Data - R-squared: 0.7821160469525934

Training Data - Mean Absolute Error: 3.2702481960745065

Training Data - Mean Squared Error: 20.416016073903695

Training Data - Root Mean Squared Error: 4.51840857757504

Testing Data - R-squared: 0.7777106377151379

Testing Data - Mean Absolute Error: 3.3977003943872917

Testing Data - Mean Squared Error: 21.792039222216523

Testing Data - Root Mean Squared Error: 4.668194428493368

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.781
Model:	OLS	Adj. R-squared:	0.781
Method:	Least Squares	F-statistic:	1535.
Date:	Sun, 23 Jul 2023	Prob (F-statistic):	0.00
Time:	17:13:22	Log-Likelihood:	-24055.
No. Observations:	8192	AIC:	4.815e+04
Df Residuals:	8172	BIC:	4.829e+04
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.7534	0.252	336.772	0.000	84.260	85.247
lread	-0.0348	0.004	-9.179	0.000	-0.042	-0.027
lwrite	0.0635	0.011	5.683	0.000	0.042	0.085
scall	-0.0008	5.45e-05	-14.290	0.000	-0.001	-0.001
sread	0.0016	0.001	1.876	0.061	-7.43e-05	0.003
swrite	-0.0056	0.001	-4.502	0.000	-0.008	-0.003
fork	-0.0288	0.114	-0.252	0.801	-0.252	0.195
exec	-0.2891	0.044	-6.564	0.000	-0.376	-0.203
rchar	-5.791e-06	4.15e-07	-13.945	0.000	-6.61e-06	-4.98e-06
wchar	-7.459e-06	8.89e-07	-8.389	0.000	-9.2e-06	-5.72e-06
pgout	-0.4151	0.078	-5.347	0.000	-0.567	-0.263
ppgout	-0.0077	0.069	-0.112	0.911	-0.143	0.128
pgfree	0.0706	0.042	1.684	0.092	-0.012	0.153
pgscan	2.935e-14	1.03e-16	283.989	0.000	2.91e-14	2.96e-14
atch	0.6623	0.122	5.419	0.000	0.423	0.902
pgin	0.0181	0.024	0.744	0.457	-0.030	0.066
ppgin	-0.0628	0.017	-3.720	0.000	-0.096	-0.030
pflt	-0.0325	0.002	-19.234	0.000	-0.036	-0.029

Based on the provided evaluation metrics for the linear regression model predicting the 'usr' mode using system activity measures, we can draw the following conclusions:

1. Model Performance:
    - The R-squared values for both the training data (0.7821) and testing data (0.7777) are relatively high, indicating that the model explains a significant portion of the variance in the target variable 'usr'. This suggests that the selected system activity measures are reasonably good predictors of the 'usr' mode.
2. Accuracy:
    - The mean absolute error (MAE) for the training data (3.2702) and testing data (3.3977) are relatively close, which means that, on average, the model's predictions are off by about 3.27% to 3.40% in the training and testing data, respectively. Lower MAE values indicate better accuracy, and the values obtained are reasonably low.
3. Error:

- The mean squared error (MSE) and root mean squared error (RMSE) for both the training and testing data are also reasonably low. These metrics quantify the average squared and squared root differences between the predicted and actual values, respectively. Lower MSE and RMSE values indicate better model performance, and the values obtained are relatively good.

Overall, the model performs well in predicting the 'usr' mode based on the system activity measures. It demonstrates a good fit to the training data and generalizes well to unseen testing data. The model's R-squared value indicates that approximately 78% of the variance in the 'usr' mode can be explained by the system activity measures.

However, as with any modeling task, it's essential to consider the specific context and domain knowledge. Further analysis and domain expertise may be necessary to validate the model's results and understand the practical implications of the findings. Additionally, depending on the application, you may explore different regression techniques, feature engineering, or further data preprocessing to improve the model's accuracy and interpretability.

Problem 2: Logistic Regression, LDA and CART

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

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

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.

Ans:

```
<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
```

The dataset of 10 variables in which there are 7 objects, 2 float type and 1 integer type variable

Contraceptive_method_used is the dependent variable.We have 1473 rows and 10 columns in our Data-set.

	count	mean	std	min	25%	50%	75%	max
Wife_age	1402.0	32.606277	8.274927	16.0	26.0	32.0	39.0	49.0
No_of_children_born	1452.0	3.254132	2.365212	0.0	1.0	3.0	4.0	16.0
Husband_Occupation	1473.0	2.137814	0.864857	1.0	1.0	2.0	3.0	4.0

```
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
```

There are blanks in wife age and no of children born.

```
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
```

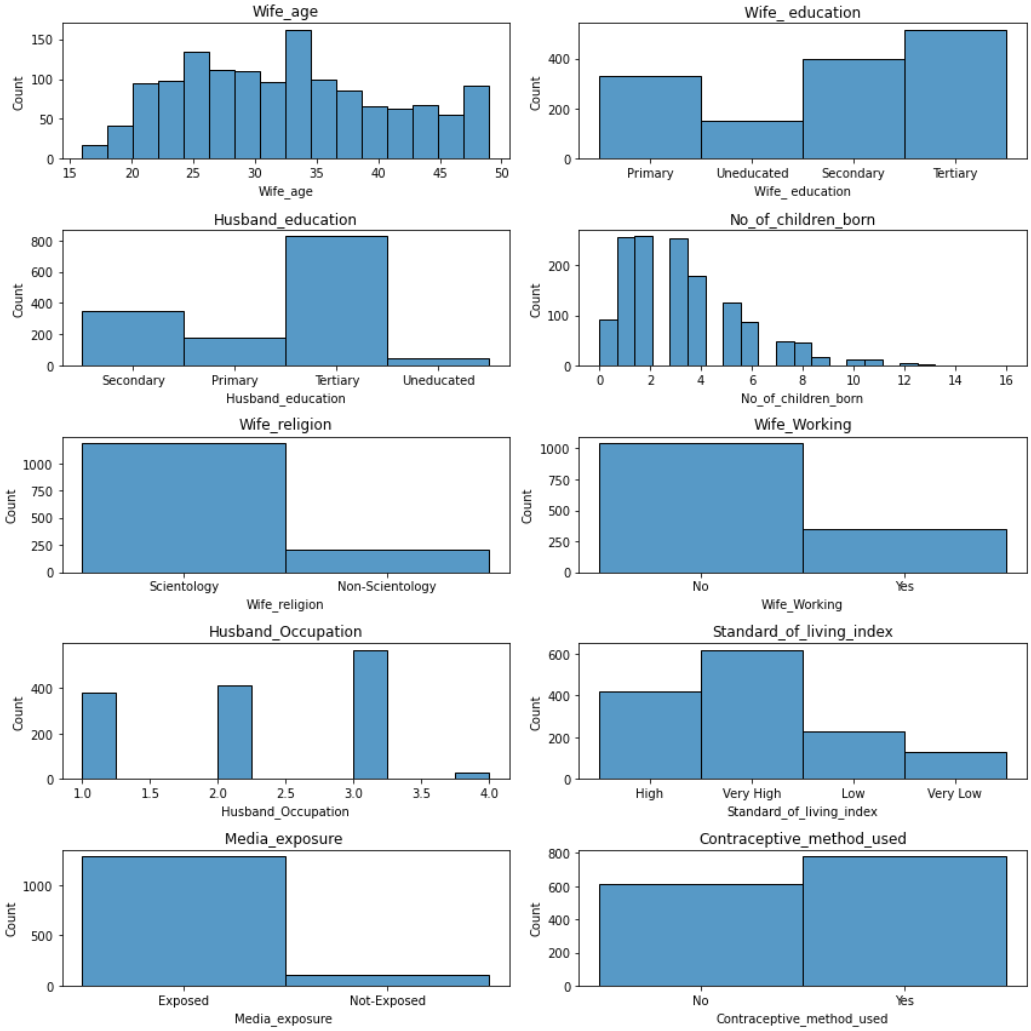
After treating the null values.

```
contra.duplicated().sum()
```

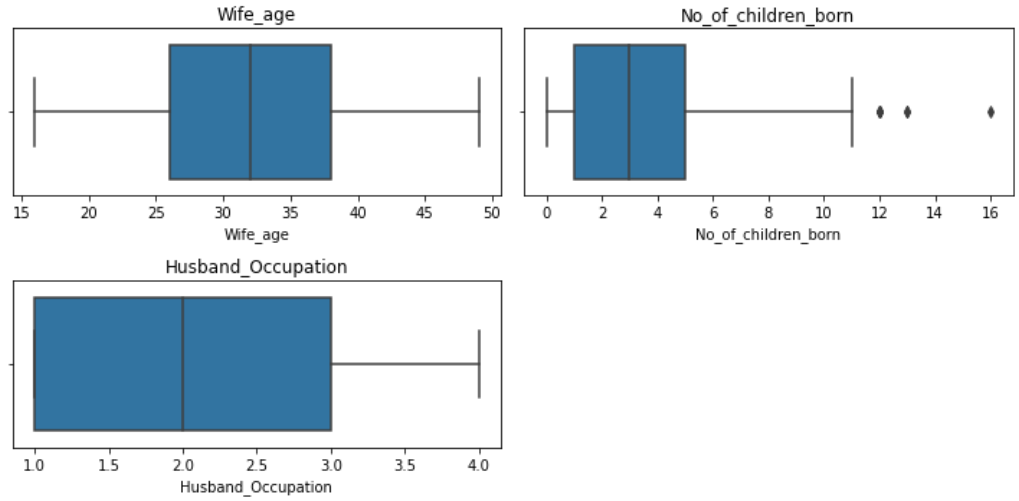
```
80
```

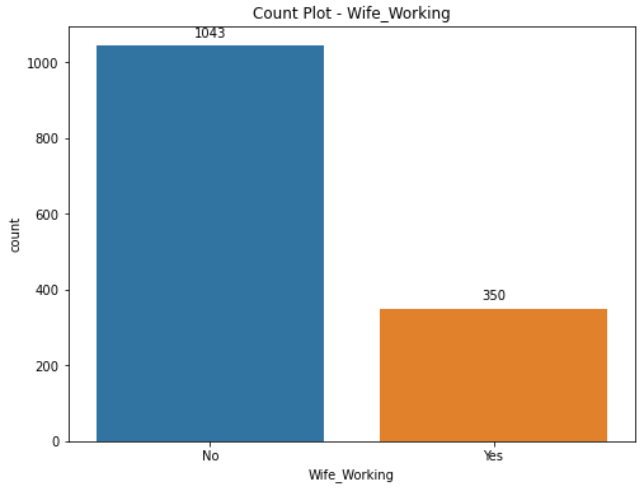
There are 80 duplicate rows. So removing them.

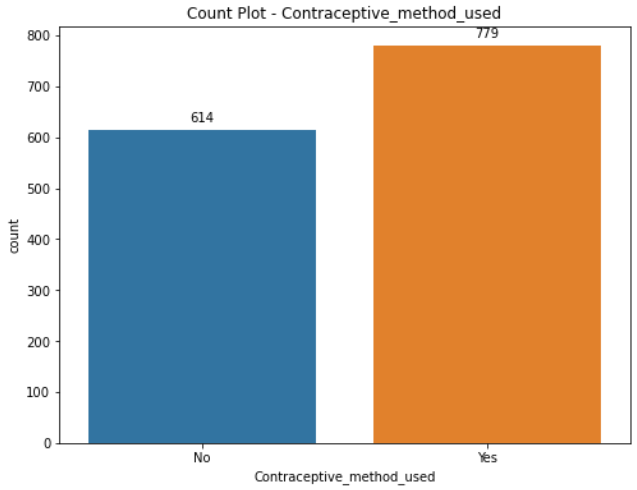
Univariant Analysis:

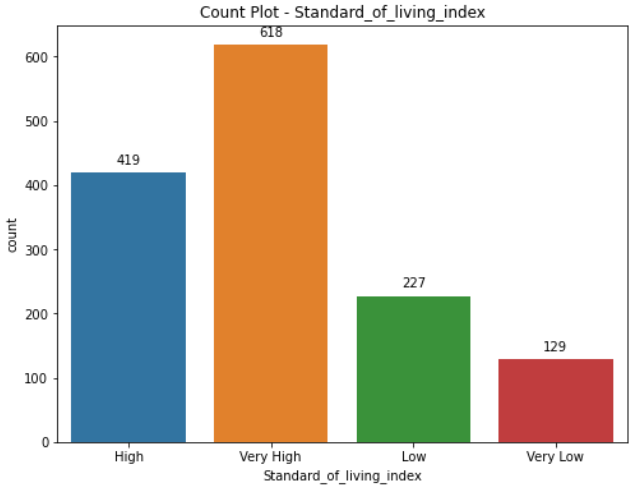

Wife_age
Count
Wife_age
Wife_ education
Count
Primary
Uneducated
Secondary
Tertiary
Wife_ education
Husband_education
Count
Secondary
Primary
Tertiary
Uneducated
Husband_education
No_of_children_born
Count
No_of_children_born
Wife_religion
Count
Scientology
Non-Scientology
Wife_religion
Wife_Working
Count
No
Yes
Wife_Working
Husband_Occupation
Count
Husband_Occupation
Standard_of_living_index
Count
High
Very High
Low
Very Low
Standard_of_living_index
Media_exposure
Count
Exposed
Not-Exposed
Media_exposure
Contraceptive_method_used
Count
No
Yes
Contraceptive_method_used

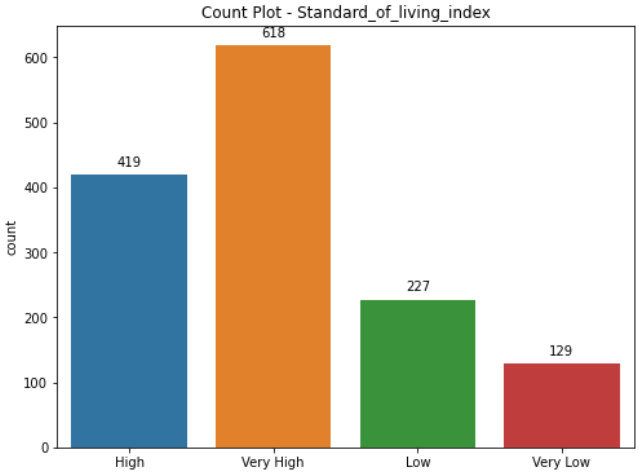
Box Plots :

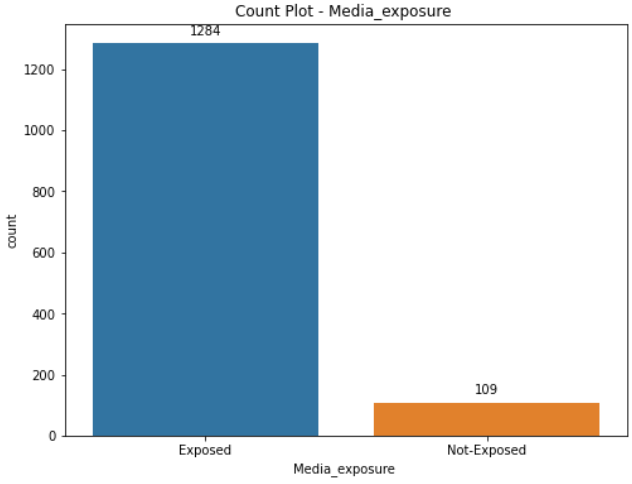

Wife_age
Wife_age
No_of_children_born
No_of_children_born
Husband_Occupation
Husband_Occupation

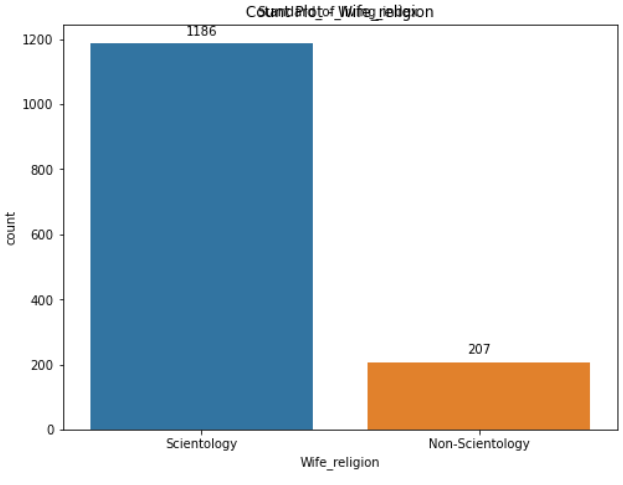

Count Plot - Wife_Working
1043
350
count
Wife_Working
No
Yes

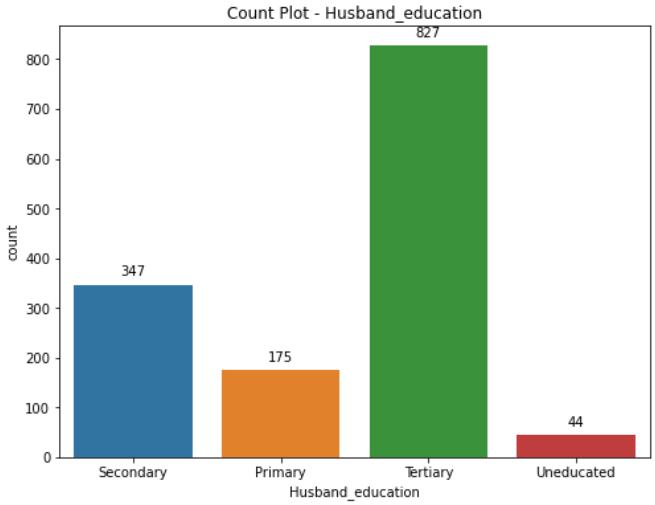

Count Plot - Contraceptive_method_used
614
779
count
Contraceptive_method_used
No
Yes


Count Plot - Standard_of_living_index
419
618
227
129
count
Standard_of_living_index
High
Very High
Low
Very Low

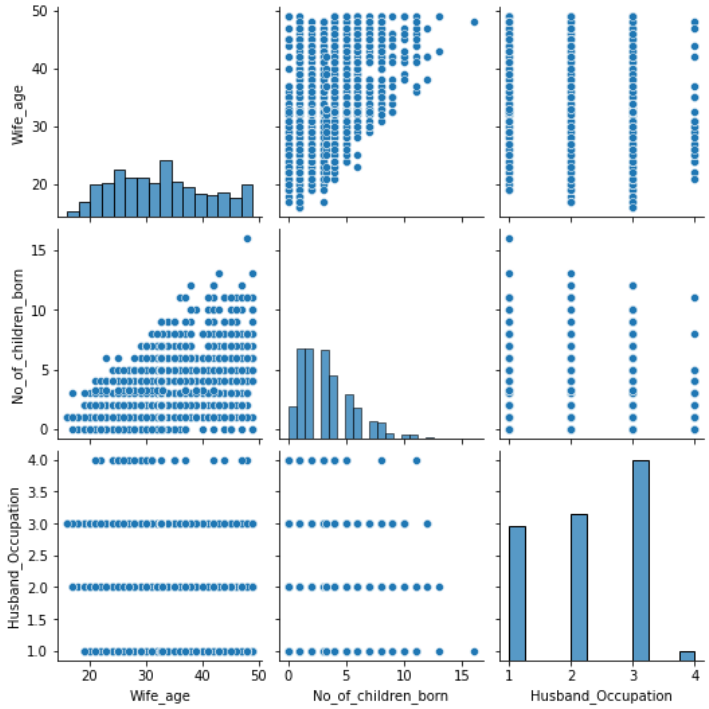

Count Plot - Standard_of_living_index
419
618
227
129
count
High
Very High
Low
Very Low

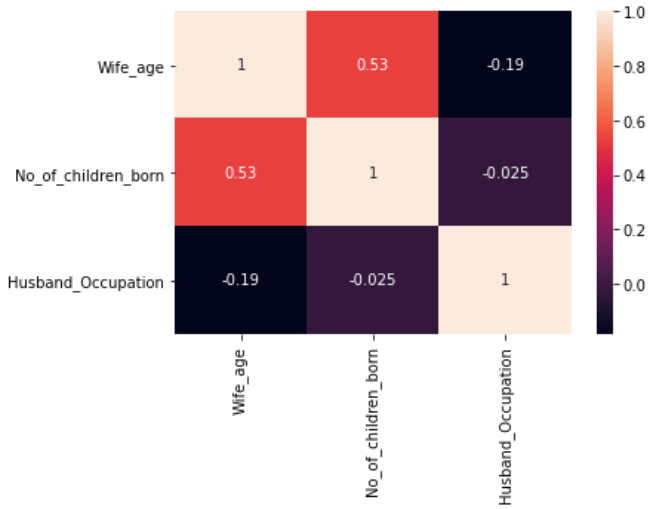

Count Plot - Media_exposure
1284
109
count
Media_exposure
Exposed
Not-Exposed


1186
207
count
Wife_religion
Scientology
Non-Scientology



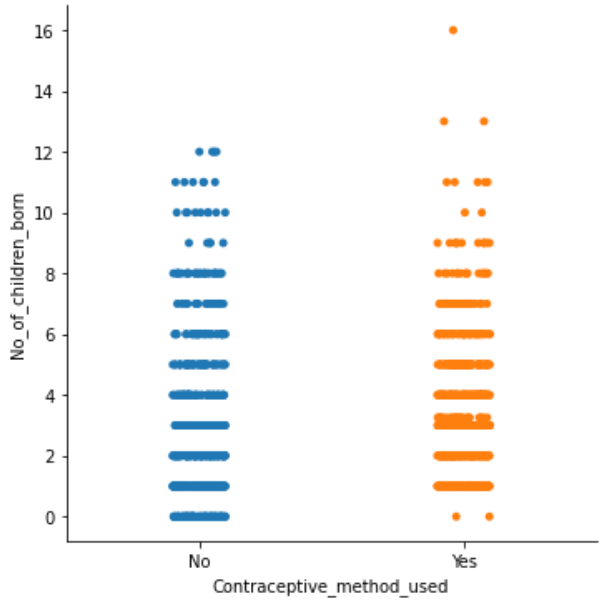
Bivariate analysis(Pair-plot) :





Heatmap: Wife age And No_of_children_born are slightly correlated.

Catplot for categorical vs numerical analysis:



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

Ans.:

Wife_ education: Uneducated = 1, Primary = 2, Secondary = 3, Tertiary = 4.

Husband_education: Uneducated = 1, Primary = 2, Secondary = 3, Tertiary = 4.

Wife_religion: Scientology = 1 and non-Scientology = 2.

Wife_Working: Yes = 1 and No = 2.

Standard_of_living_index: Very Low = 1, Low = 2, High = 3, Very High = 4.

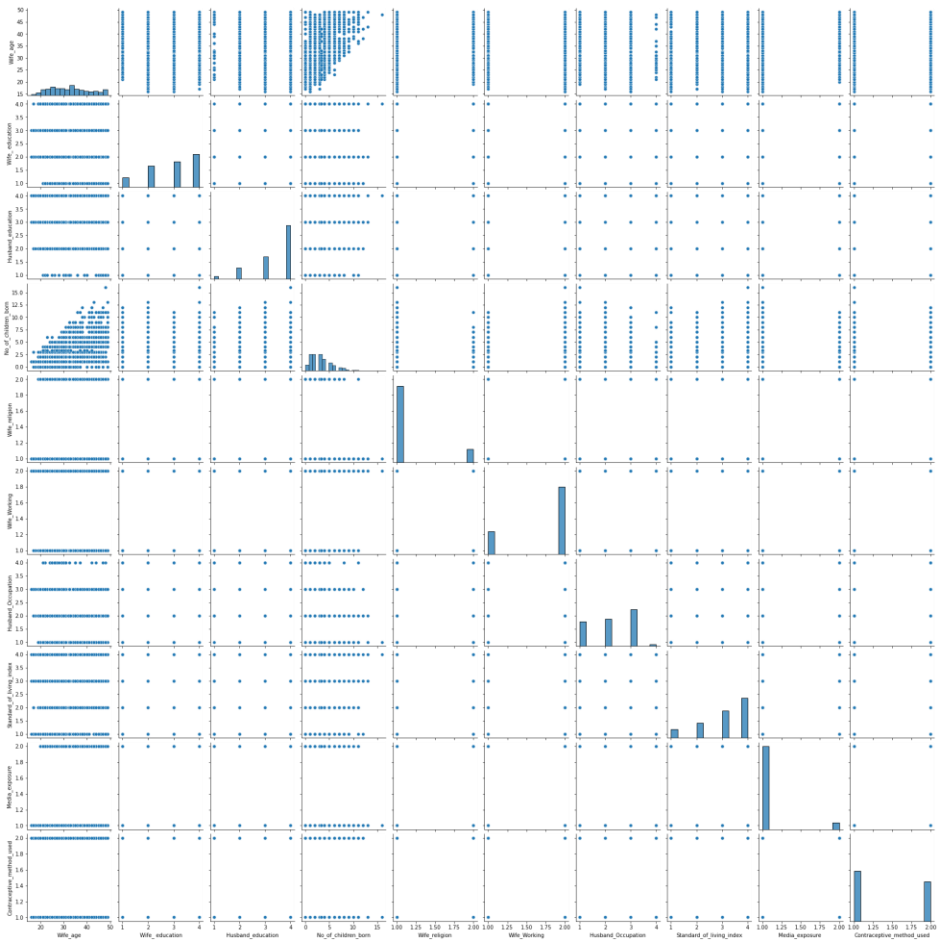
Media_exposure: Exposed = 1 and Not-Exposed = 2.

Contraceptive_method_used: Yes = 1 and No = 0

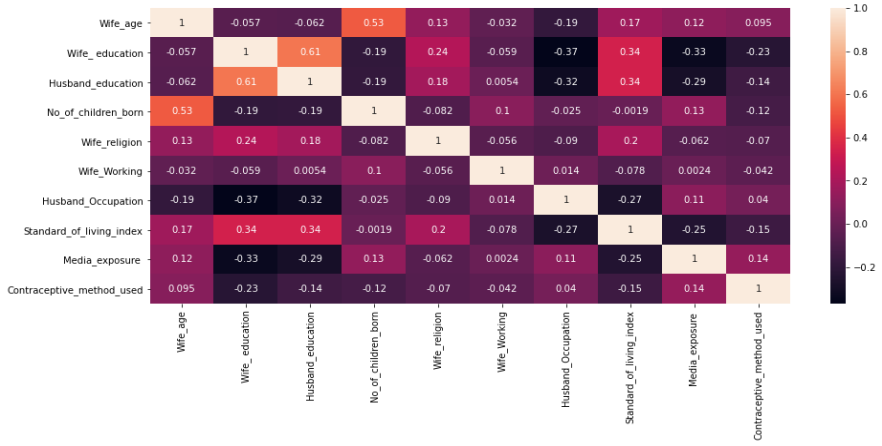
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1393 entries, 0 to 1472
Data columns (total 10 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   Wife_age                   1393 non-null   float64
 1   Wife_ education            1393 non-null   int64  
 2   Husband_education          1393 non-null   int64  
 3   No_of_children_born        1393 non-null   float64
 4   Wife_religion              1393 non-null   int64  
 5   Wife_Working               1393 non-null   int64  
 6   Husband_Occupation         1393 non-null   int64  
 7   Standard_of_living_index   1393 non-null   int64  
 8   Media_exposure             1393 non-null   int64  
 9   Contraceptive_method_used  1393 non-null   int64  
dtypes: float64(2), int64(8)
memory usage: 152.0 KB
```

Now we don’t have any object datatypes left.

Pairplot:



Heat Map :



	precision	recall	f1-score	support
1	0.67	0.81	0.73	779
2	0.67	0.49	0.56	614
accuracy			0.67	1393
macro avg	0.67	0.65	0.65	1393
weighted avg	0.67	0.67	0.66	1393