PM Project Report

Contents

List of Figures

List of Tables

PM PROJECT

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.

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.

Data Types:

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap
0	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
1	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
2	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
3	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
4	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

```
<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
```

```
float64    13
int64       8
object      1
dtype: int64
```

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.

Shape:
We have 8192 rows and 22 columns in our Data set.

Summary:

	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:

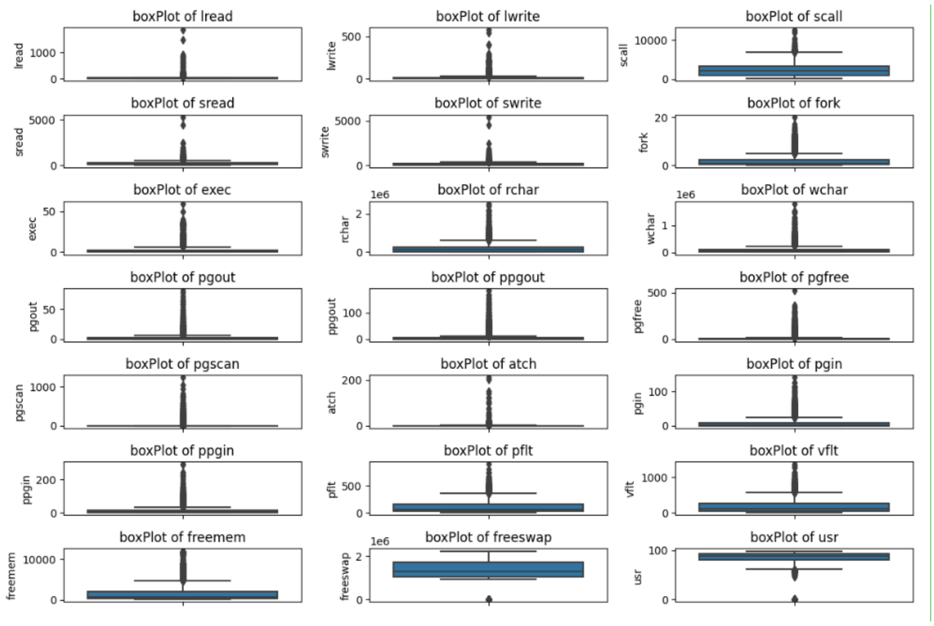


Figure 1: Boxplots of the numerical variables

There are outliers present in our current data set.

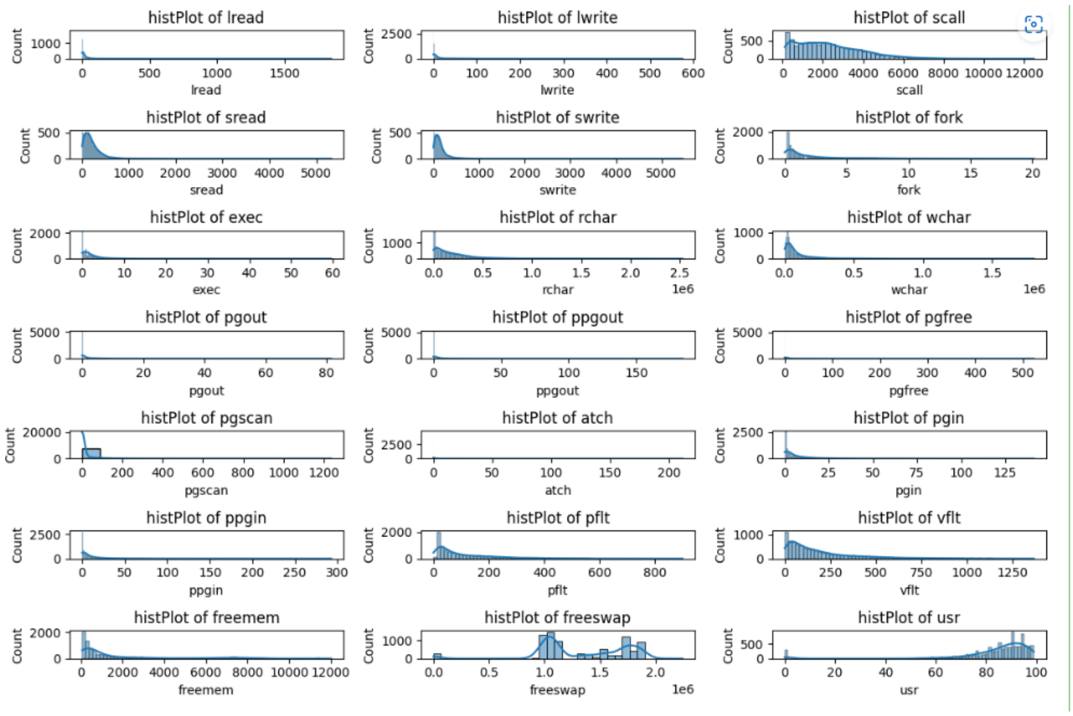


Figure 2: Histogram of the numerical variables

There are skewness in our current data sets.
User data is left skewed and right skweness are identifed in freemem, freeswap, vflt, scall and etc.

Process run queue size

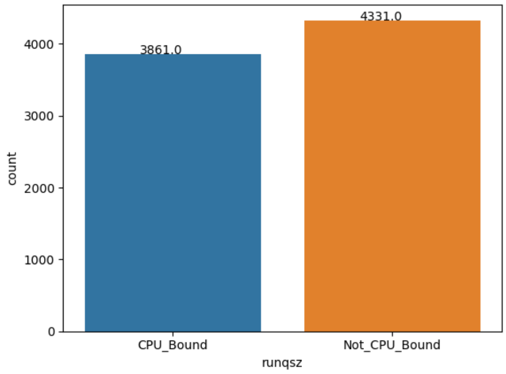


Figure 3: Countplot of the Process run queue size

From the countplot of Process run queue size, CPU_Bound are 3861 and Not_CPU_Bound are 4331.

Bivariate Analysis:

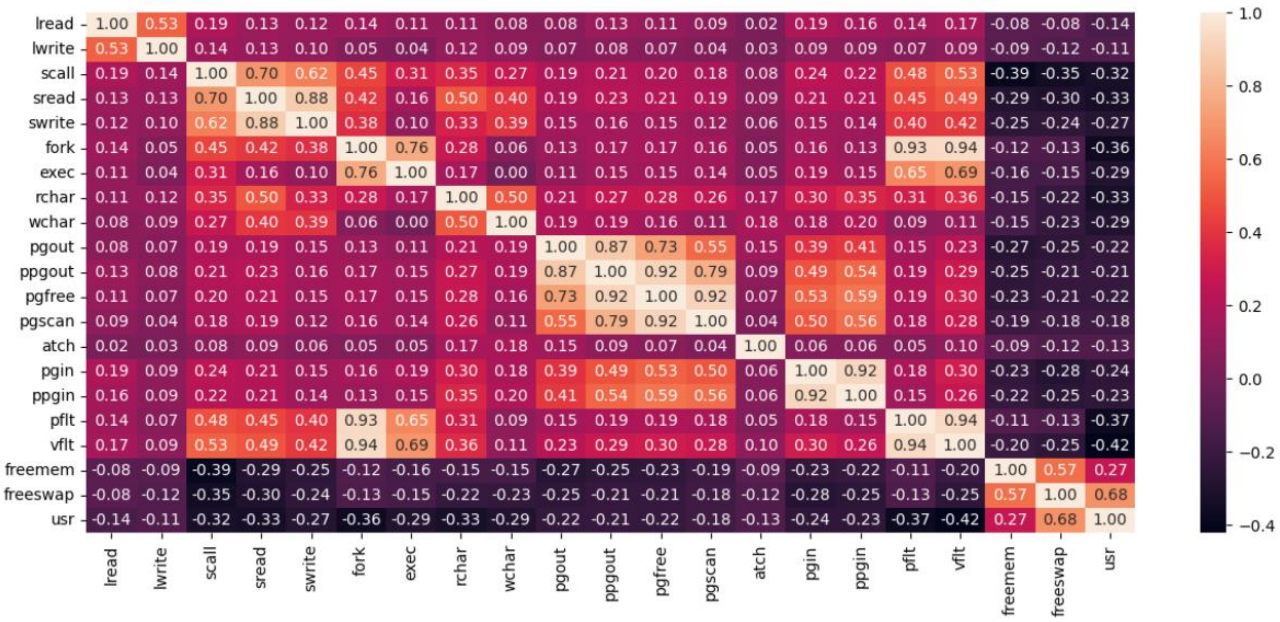


Figure 4: Heat Map of the numerical variables

Pair Plot

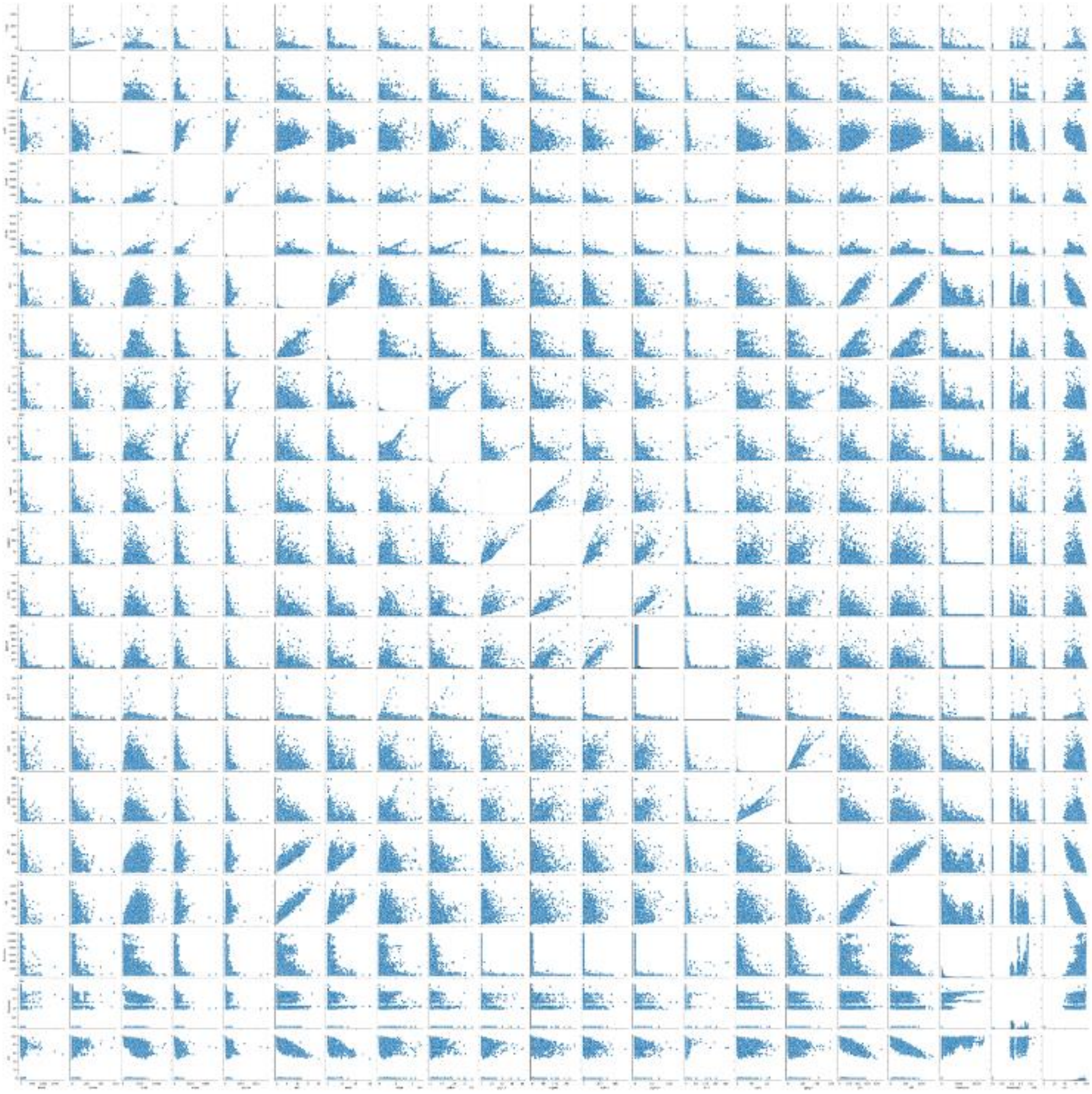


Figure 5: Pair plot of the numerical variables

From the above HeatMap and PairPlot , we can say there is a correlations.

- Highest positive corelation exist between Number of page faults caused by address translation (vflt) and Number of page faults caused by protection error (pflt) 94%.
- Highest positive corelation exist between Number of page faults caused by address translation (vflt) and Number of system fork calls per second (fork) 94%.
- Highest positive corelation exist between Number of page faults caused by protection error (pflt) and Number of system fork calls per second (fork) 93%.
- Highest positive corelation exist between Number of pages paged in per second (ppgin) and Number of page-in requests per second (pgin) 92%.

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.

Null Value Check:

```
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 in rchar and wchar. Null values replaced by 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
```

Zero Values

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

We can keep Zeros in our data frame since there might be chance when system is idle.

Remove Outliers

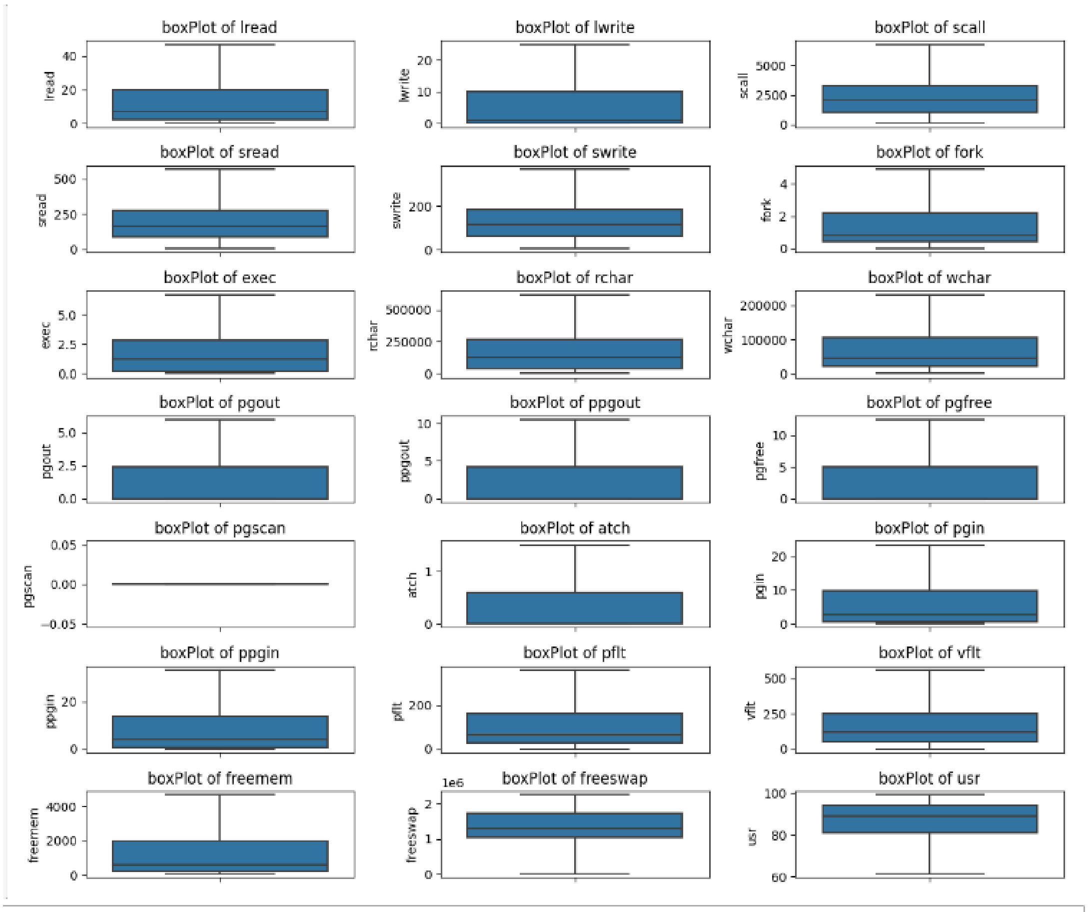


Figure 6: Box Plot of the numerical variables After removing outliers

Duplicate Check

There is no duplicate records in our data set.

New Feature

Runsqz is converted to continuous data like below

cData["runqsz"] = cData["runqsz"].replace({'CPU_Bound': 0, 'Not_CPU_Bound':1})

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.

Checking Multicollinearity using VIF:

```
VIF values:

lread        1.659493
lwrite       1.680327
scall        6.607273
sread       14.056800
swrite       9.678751
fork        27.468314
exec         3.900465
rchar        3.289783
wchar        2.223182
pgout        6.679123
ppgout      18.149664
pgfree      23.218580
pgscan      10.111498
atch         1.129015
pgin        10.877871
ppgin       11.311240
pflt        22.564882
vflt        37.080281
runqsz       2.078007
freemem      2.468755
freeswap     5.538409
dtype: float64
```

So, variables have moderate correlations. (VIF Values exceeding 5)

We fit the dataset to model to LinearRegression()

Coefficients:

```
The coefficient for lread is -0.019898242591582342
The coefficient for lwrite is 0.004822549499005826
The coefficient for scall is 0.0010078328708177846
The coefficient for sread is -0.00042925110899032817
The coefficient for swrite is -0.0020785052844854283
The coefficient for fork is -1.721635260301748
The coefficient for exec is -0.08962572330407521
The coefficient for rchar is -4.114249883094468e-06
The coefficient for wchar is -1.1603100029289946e-05
The coefficient for pgout is -0.17414405160284666
The coefficient for ppgout is 0.09896424632675815
The coefficient for pgfree is -0.0702837828644868
The coefficient for pgscan is 0.008611010098028173
The coefficient for atch is -0.07829685978947061
The coefficient for pgin is 0.09136880232552246
The coefficient for ppgin is -0.0593593716268139
The coefficient for pflt is -0.04150261126432213
The coefficient for vflt is 0.022282136803892492
The coefficient for runqsz is 7.788368806940666
The coefficient for freemem is -0.0016166383185141119
The coefficient for freeswap is 3.219084535188488e-05
```

```
The intercept for our model is 44.64681588526172
R square on Training data 0.6428635339285307
R square on Testing data 0.6311655542667606
```

```
RMSE on Training data 10.812852066268919
RMSE on Testing data 11.59482423619469
```

Linear Regression using OS Models:

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.643
Model:	OLS	Adj. R-squared:	0.642
Method:	Least Squares	F-statistic:	489.6
Date:	Sun, 28 May 2023	Prob (F-statistic):	0.00
Time:	14:26:36	Log-Likelihood:	-21787.
No. Observations:	5734	AIC:	4.362e+04
Df Residuals:	5712	BIC:	4.377e+04
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	44.6468	0.746	59.838	0.000	43.184	46.110
lread	-0.0199	0.003	-6.217	0.000	-0.026	-0.014
lwrite	0.0048	0.006	0.799	0.424	-0.007	0.017
scall	0.0010	0.000	7.449	0.000	0.001	0.001
sread	-0.0004	0.002	-0.234	0.815	-0.004	0.003
swrite	-0.0021	0.002	-1.037	0.300	-0.006	0.002
fork	-1.7216	0.244	-7.050	0.000	-2.200	-1.243
exec	-0.0896	0.048	-1.879	0.060	-0.183	0.004
rchar	-4.114e-06	8.29e-07	-4.961	0.000	-5.74e-06	-2.49e-06
wchar	-1.16e-05	1.28e-06	-9.091	0.000	-1.41e-05	-9.1e-06
pgout	-0.1741	0.064	-2.721	0.007	-0.300	-0.049
ppgout	0.0990	0.037	2.702	0.007	0.027	0.171
pgfree	-0.0703	0.020	-3.505	0.000	-0.110	-0.031
pgscan	0.0086	0.006	1.361	0.174	-0.004	0.021
atch	-0.0783	0.027	-2.939	0.003	-0.131	-0.026
pgin	0.0914	0.029	3.107	0.002	0.034	0.149
ppgin	-0.0594	0.019	-3.127	0.002	-0.097	-0.022
pflt	-0.0415	0.004	-9.696	0.000	-0.050	-0.033
vflt	0.0223	0.003	6.665	0.000	0.016	0.029
runqsz	7.7884	0.303	25.684	0.000	7.194	8.383
freemem	-0.0016	7.53e-05	-21.482	0.000	-0.002	-0.001
freeswap	3.219e-05	4.53e-07	70.984	0.000	3.13e-05	3.31e-05

Omnibus:	1507.116	Durbin-Watson:	2.057
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4767.078
Skew:	-1.333	Prob(JB):	0.00
Kurtosis:	6.584	Cond. No.	7.48e+06

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.643
Model:                            OLS   Adj. R-squared:                  0.642
Method:                 Least Squares   F-statistic:                     489.6
Date:                Sun, 28 May 2023   Prob (F-statistic):               0.00
Time:                        13:27:15   Log-Likelihood:                -21787.
No. Observations:                5734   AIC:                         4.362e+04
Df Residuals:                    5712   BIC:                         4.377e+04
Df Model:                          21
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         44.6468      0.746     59.838      0.000      43.184      46.110
lread         -0.0199      0.003     -6.217      0.000      -0.026      -0.014
lwrite         0.0048      0.006      0.799      0.424      -0.007       0.017
scall          0.0010      0.000      7.449      0.000       0.001       0.001
sread         -0.0004      0.002     -0.234      0.815      -0.004       0.003
swrite        -0.0021      0.002     -1.037      0.300      -0.006       0.002
fork          -1.7216      0.244     -7.050      0.000      -2.200      -1.243
exec          -0.0896      0.048     -1.879      0.060      -0.183       0.004
rchar      -4.114e-06   8.29e-07     -4.961      0.000   -5.74e-06   -2.49e-06
wchar       -1.16e-05   1.28e-06     -9.091      0.000   -1.41e-05    -9.1e-06
pgout         -0.1741      0.064     -2.721      0.007      -0.300      -0.049
ppgout         0.0990      0.037      2.702      0.007       0.027       0.171
pgfree        -0.0703      0.020     -3.505      0.000      -0.110      -0.031
pgscan         0.0086      0.006      1.361      0.174      -0.004       0.021
atch          -0.0783      0.027     -2.939      0.003      -0.131      -0.026
pgin           0.0914      0.029      3.107      0.002       0.034       0.149
ppgin         -0.0594      0.019     -3.127      0.002      -0.097      -0.022
pflt          -0.0415      0.004     -9.696      0.000      -0.050      -0.033
vflt           0.0223      0.003      6.665      0.000       0.016       0.029
runqsz         7.7884      0.303     25.684      0.000       7.194       8.383
freemem       -0.0016   7.53e-05    -21.482      0.000      -0.002      -0.001
freeswap    3.219e-05   4.53e-07     70.984      0.000    3.13e-05    3.31e-05
==============================================================================
Omnibus:                     1507.116   Durbin-Watson:                   2.057
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4767.078
Skew:                          -1.333   Prob(JB):                         0.00
Kurtosis:                       6.584   Cond. No.                     7.48e+06
==============================================================================
```

Figure 7: OLS Regression Results

R Squared:64.3%

RMSE on Training data 10.812852066268919
RMSE on Testing data 11.59482423619469

Equation

(44.65) * const + (-0.02) * lread + (0.0) * lwrite + (0.0) * scall + (-0.0) * sread + (-0.0) * swrite + (-1.72) * fork + (-0.09) * exec + (-0.0) * rchar + (-0.0) * wchar + (-0.17) * pgout + (0.1) * ppgout + (-0.07) * pgfree + (0.01) * pgscan + (-0.08) * atch + (0.09) * pgin + (-0.06) * ppgin + (-0.04) * pflt + (0.02) * vflt + (7.79) * runqsz + (-0.0) * freemem + (0.0) * freeswap

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

- There is an increment in user by a large factor if Number of page faults caused by address translation (vflt).
- There is a decrement in user by a large factor if Number of system fork calls per second is increased (fork).
- There is an increment in user by a large factor if Number of pages, paged out per second (ppgout).

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.

Null Value Check

```
data_df.isnull().sum()
```

```
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 null values in Wife age and No_of_children_born

Duplicates:
There are eighty duplicates records found in given data set.
Hence duplicates records removed from the data set.

After removing duplicates:

	Wife_age	Wife_education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_e
0	24.0	1	2	3.0	1	2	2	3	
1	45.0	0	2	10.0	1	2	3	4	
2	43.0	1	2	7.0	1	2	3	4	
3	42.0	2	1	9.0	1	2	3	3	
4	36.0	2	2	8.0	1	2	3	2	
...	...	...	...	...	...	...	...	...	
1468	33.0	3	3	NaN	1	1	2	4	
1469	33.0	3	3	NaN	1	2	1	4	
1470	39.0	2	2	NaN	1	1	1	4	
1471	33.0	2	2	NaN	1	1	2	2	
1472	17.0	2	2	1.0	1	2	2	4	

Summary:

	Wife_age	No_of_children_born	Husband_Occupation
count	1402.000000	1452.000000	1473.000000
mean	32.606277	3.254132	2.137814
std	8.274927	2.365212	0.864857
min	16.000000	0.000000	1.000000
25%	26.000000	1.000000	1.000000
50%	32.000000	3.000000	2.000000
75%	39.000000	4.000000	3.000000
max	49.000000	16.000000	4.000000

Average value of wife age is 32 and 75% of women falls under age 39.

Data Types

```
Wife_age                     float64
Wife_ education               object
Husband_education             object
No_of_children_born          float64
Wife_religion                 object
Wife_Working                  object
Husband_Occupation             int64
Standard_of_living_index      object
Media_exposure                object
Contraceptive_method_used     object
dtype: object
```

7 Parameters are object, 2 are float type and 1 Integer type variable.
Contraceptive_method_used is dependent variable

Shape:
Data set having 1473 rows and 10 columns.

Univariate Analysis:

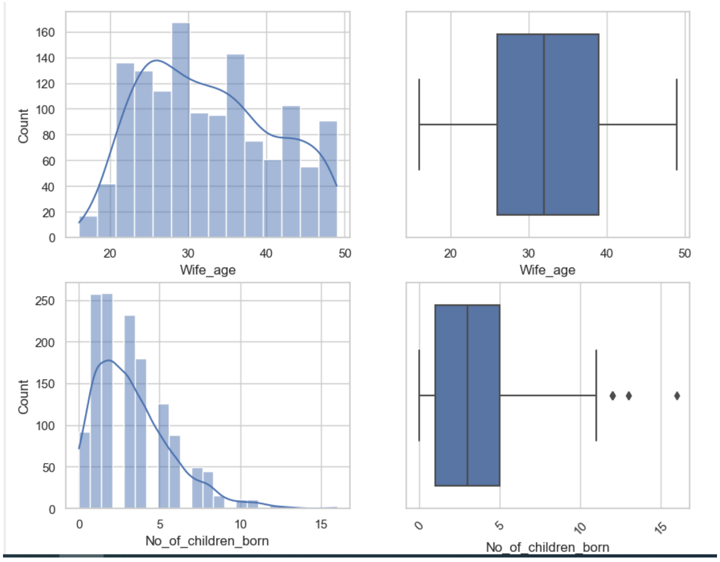


Figure 8: Box Plot and Hist Plot of Wife Age and No_children_born

There are some outliers are present in No of children born.
Wife Age is right skewed which means more number of women identified between 28-40.

Wife Education:

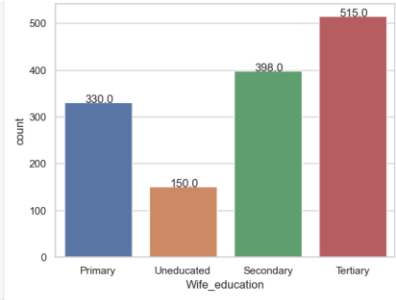


Figure 9: Count Plot of Wife education

Tertiary having more numbers of women.

Husband Education:

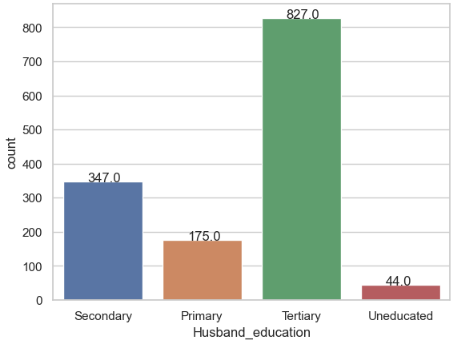


Figure 10: Count Plot of Husband education

Tertiary having more numbers of men.

Wife Religion

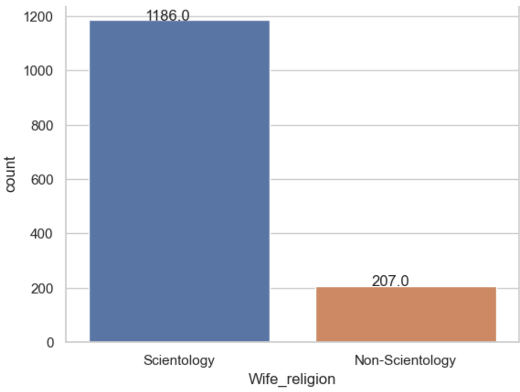


Figure 11: Count Plot of Wife Religion

Wife Working

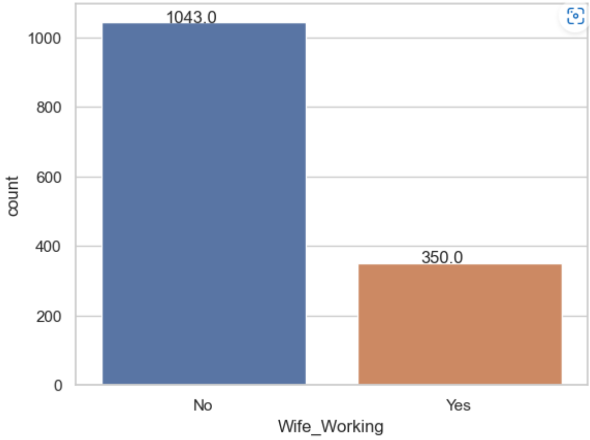


Figure 12: Count Plot of Wife Working

Standard_of_living_index

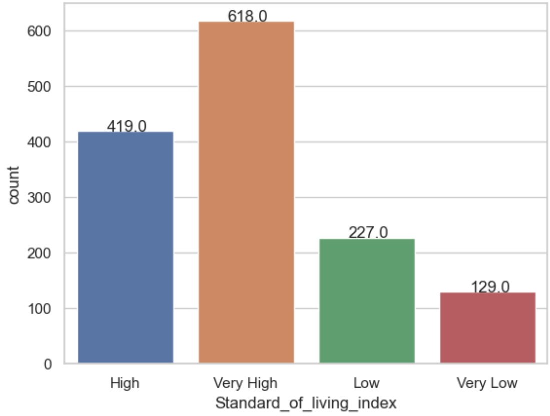


Figure 13: Count Plot Standard_of_living_index

Media exposure

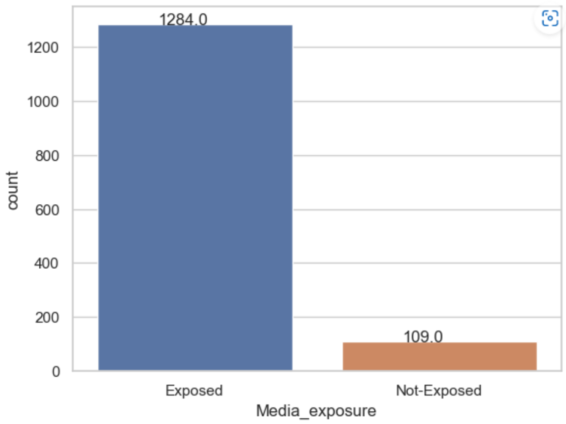


Figure 14: Count Plot Media Exposure

Contraceptive_method_used

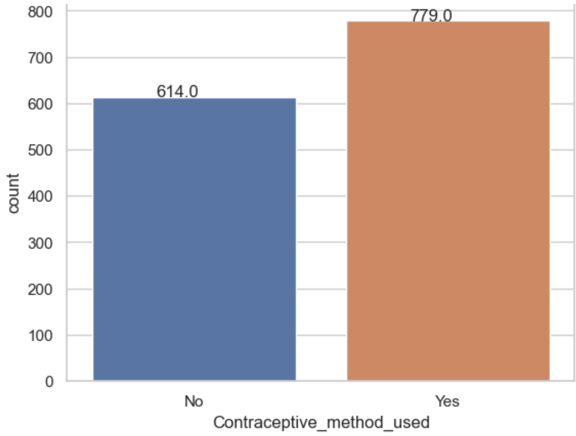


Figure 15: Count Plot Contraceptive_method_used

Bivariate Analysis:

Wife_Age vs Contraceptive_method_used

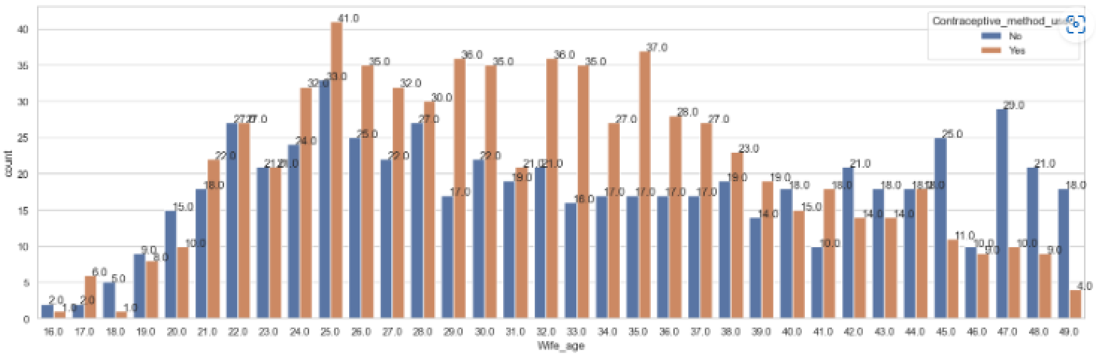


Figure 16: Count Plot between Wife Age and Contraceptive_method_used

Wife_Education Vs Contraceptive_method_used

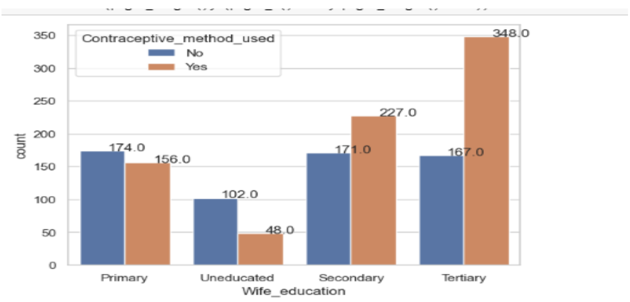


Figure 17: Count Plot between Wife Education and Contraceptive_method_used

Husband_Education Vs Contraceptive_method_used

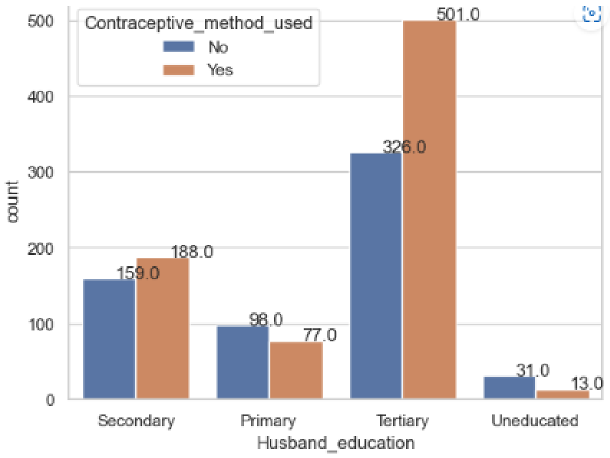


Figure 18: Count Plot between Husband Education and Contraceptive_method_used

Wife_religion Vs Contraceptive_method_used

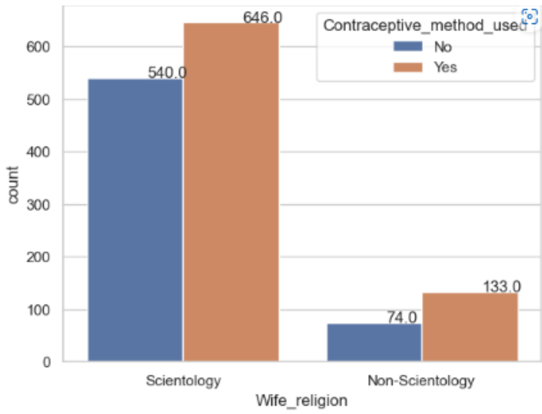


Figure 19: Count Plot between Wife Religion and Contraceptive_method_used

Wife_Working Vs Contraceptive_method_used

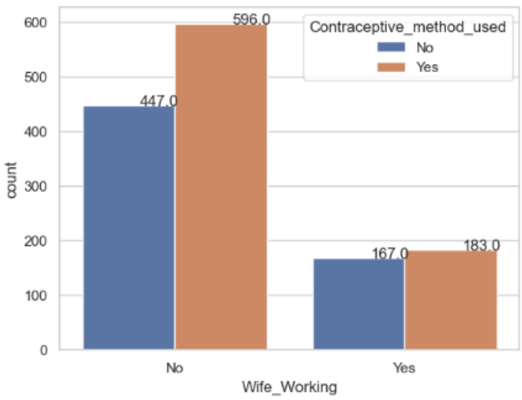


Figure 20: Count Plot between Wife working and Contraceptive_method_used

Standard_of_living_index Vs Contraceptive_method_used

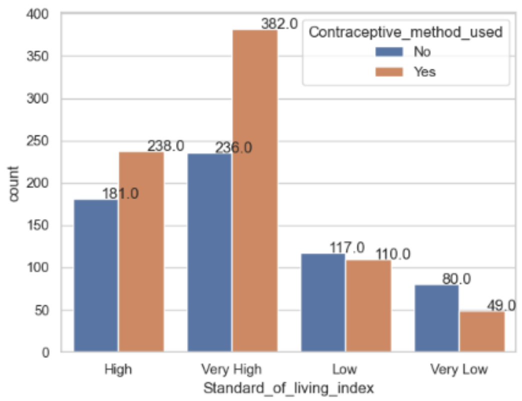


Figure 21: Count Plot between Standard_of_living_index and Contraceptive_method_used

Media_exposure Vs Contraceptive_method_used

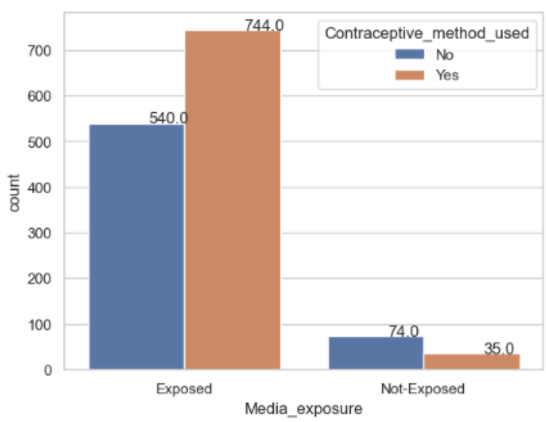


Figure 22: Count Plot between Media_exposure and Contraceptive_method_used

HeatMap

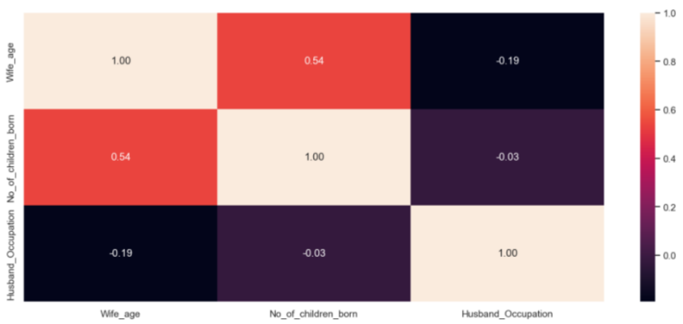


Figure 23: Heat Map of numerical variables

There is no higher corelation between wife_age, no_of_children_born and Husband_Occupation with respect to Contraceptive_method_used

Pair Plot

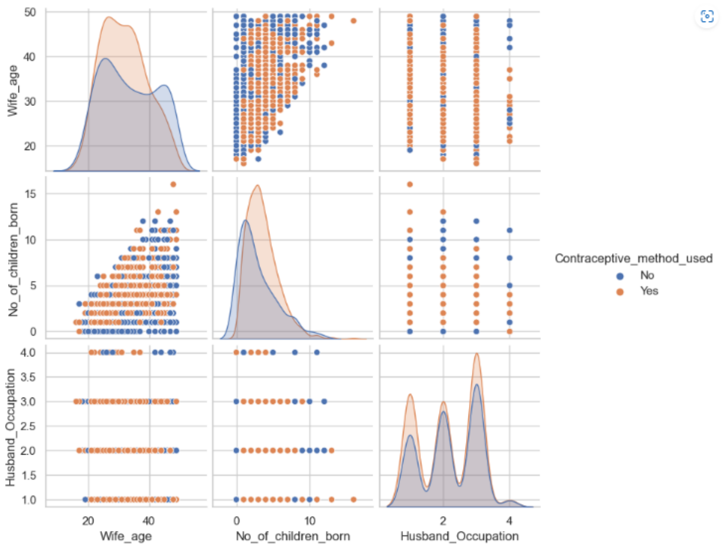


Figure 24: Pair Plot of numerical variables

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.

- Wife_ education is converted numerical values like below
  Uneducated = 1, Primary = 2, Secondary = 3, Tertiary = 4.
- Husband_education is converted numerical values like below
  Uneducated = 1, Primary = 2, Secondary = 3, Tertiary = 4.
- Wife_religion is converted numerical values like below
  Scientology = 1 and non-Scientology = 2.
- Wife_Working is converted numerical values like below
  Yes = 1 and No = 2.
- Standard_of_living_index is converted numerical values like below
  Very Low = 1, Low = 2, High = 3, Very High = 4.
- Media_exposure is converted numerical values like below
  Exposed = 1 and Not-Exposed = 2.
- Contraceptive_method_used is converted numerical values like below
  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                   1326 non-null   float64
 1   Wife_education             1393 non-null   int64  
 2   Husband_education          1393 non-null   int64  
 3   No_of_children_born        1372 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: 119.7 KB
```

	Wife_age	Wife_education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_expo
0	24.0	1	2	3.0	1	2	2	3	
1	45.0	0	2	10.0	1	2	3	4	
2	43.0	1	2	7.0	1	2	3	4	
3	42.0	2	1	9.0	1	2	3	3	
4	36.0	2	2	8.0	1	2	3	2	

Pair Plot:

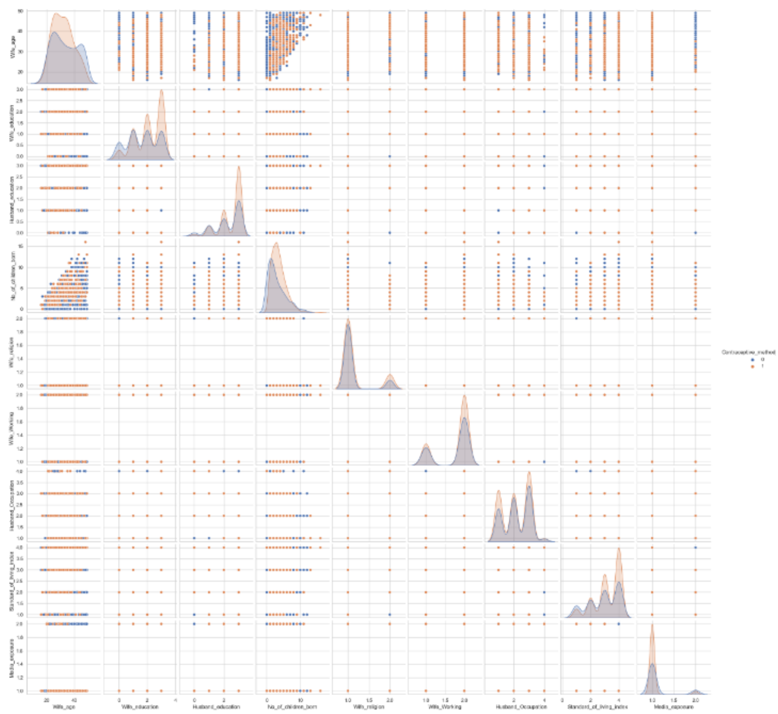


Figure 25: Pair Plot of numerical variables

Heat Map:

<AxesSubplot: >

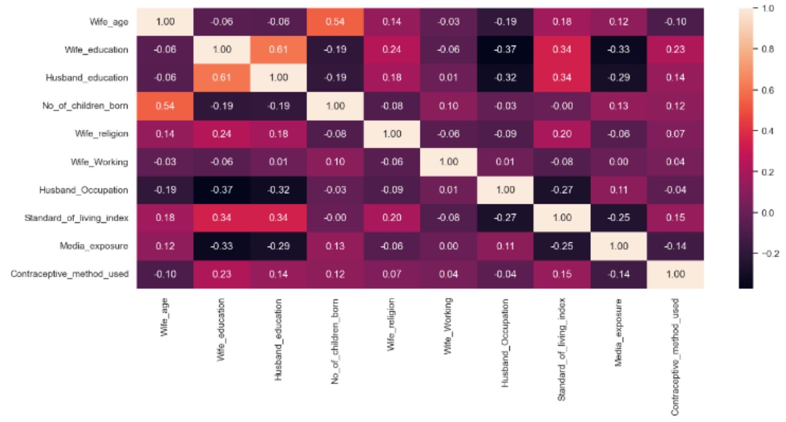


Figure 26: Heat Map of numerical variables

Logistic Regression:

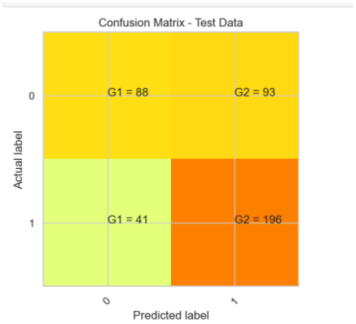
Test Data

```
0.6794258373205742
[[ 88  93]
 [ 41 196]]
              precision    recall  f1-score   support

         0.0       0.68      0.49      0.57       181
         1.0       0.68      0.83      0.75       237

    accuracy                           0.68       418
   macro avg       0.68      0.66      0.66       418
weighted avg       0.68      0.68      0.67       418
```

Confusion Matrix on Test Data:



AUC on Test Data

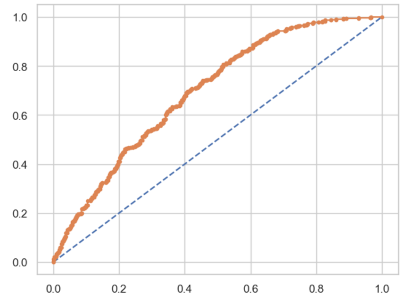


Figure 27: AOC Curve on Test Data

AUC: 0.707

Train Data:

```
0.6512820512820513
[[216 217]
 [123 419]]
              precision    recall  f1-score   support

         0.0       0.64      0.50      0.56       433
         1.0       0.66      0.77      0.71       542

    accuracy                           0.65       975
   macro avg       0.65      0.64      0.64       975
weighted avg       0.65      0.65      0.64       975
```

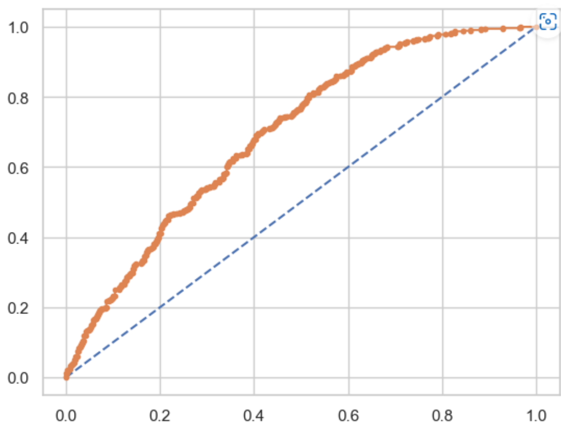


Figure 28: AOC Curve on training Data

AUC: 0.696

LDA

```
LinearDiscriminantAnalysis()
[[ 87  94]
 [ 39 198]]
              precision    recall  f1-score   support

         0.0       0.69      0.48      0.57       181
         1.0       0.68      0.84      0.75       237

    accuracy                           0.68       418
   macro avg       0.68      0.66      0.66       418
weighted avg       0.68      0.68      0.67       418
```

AUC Training Score

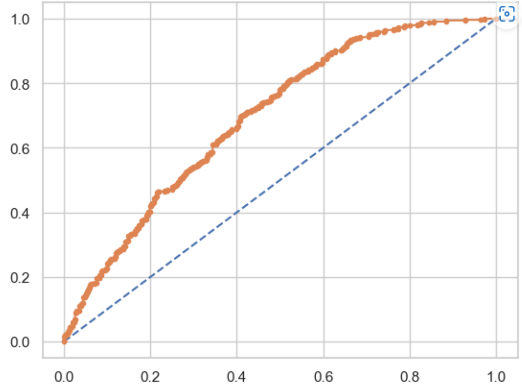


Figure 29: AUC Curve on training Data

AUC: 0.696

AUC Testing Score

AUC: 0.706

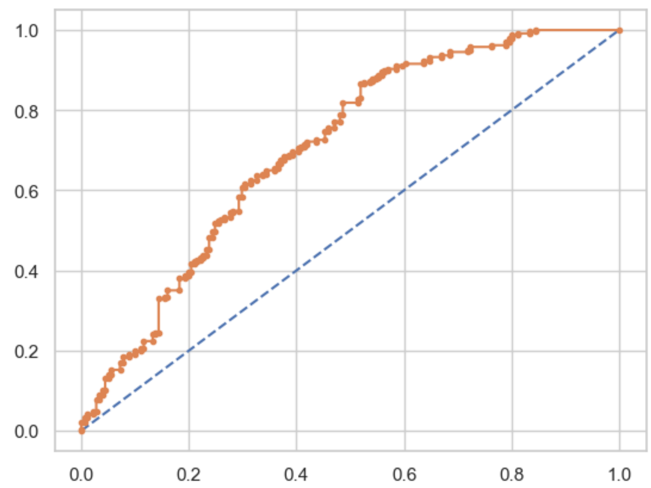


Figure 30: AOC Curve on testing Data

CART

Test Data

```
DecisionTreeClassifier(random_state=1)
[[104  77]
 [ 87 150]]
              precision    recall  f1-score   support

         0.0       0.54      0.57      0.56       181
         1.0       0.66      0.63      0.65       237

    accuracy                           0.61       418
   macro avg       0.60      0.60      0.60       418
weighted avg       0.61      0.61      0.61       418
```

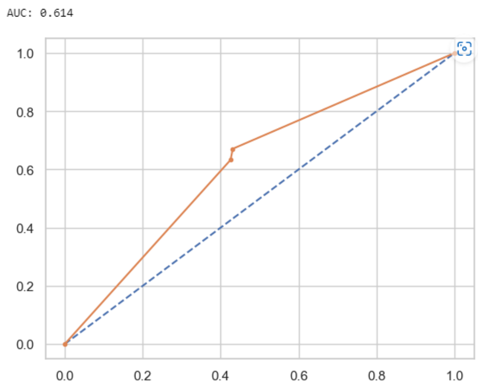


Figure 31: AOC Curve on test Data

AUC: 0.614

Training AUC Score

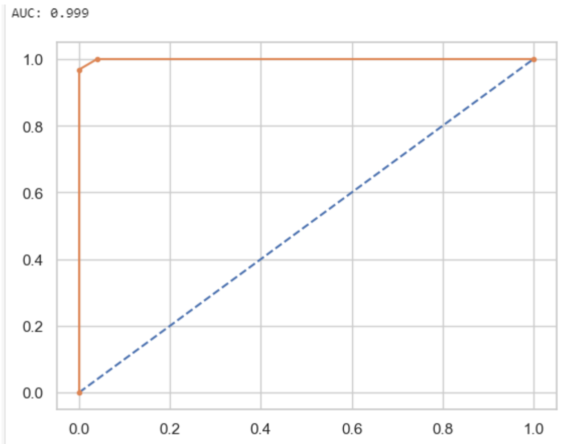


Figure 32: AOC Curve on training Data

AUC: 0.999

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.

	LR Train	LR Test	LDA Train	LDA Test	CART Train	CART Test
Accuracy	0.65	0.67	0.66	0.65	0.74	0.61
AUC	0.70	0.69	0.70	0.69	0.99	0.61
Recall	0.79	0.76	0.80	0.79	0.85	0.77
Precision	0.67	0.66	0.66	0.66	0.74	0.69
F1 score	0.72	0.71	0.73	0.72	0.79	0.73

Table 1: Predictions on Train and Test sets

Comparing all three Linear Regression, Linear Discriminant Analysis and CART we found that all are giving similar results but CART is giving better results.

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

- The EDA analysis clearly indicates that women with a tertiary education and extremely high standard of living used contraceptive methods. Women ranging from 21 to 38 use contraceptive methods more.
- The usage of contraceptive methods need not depend on their demographic or socioeconomic backgrounds since the use of contraceptive methods were the same for both working and non-working women.
- The use of contraceptive method was high for both Scientology and Non-scientology women.