`

ADVANCED STATISTICS PROJECT REPORT

2022

Sangeeth A

PGP-DSBA Online

July - 2022

`

CONTENTS

`

LIST OF FIGURES

LIST OF TABLES

`

`

PROBLEM 1 – SUMMARY

The hardness of metal implants in dental cavities depends on multiple factors, such as the method of implant, the temperature at which the metal is treated, and the alloy used as well as on the dentists who may favour one method above another and may work better in his/her favourite method. The response is the variable of interest. The dataset contains these factors and the response variables, we will conduct an ANOVA test on these variables based on the alloys used for the metal implants and we have to find out the interaction between these variables.

INTRODUCTION

The purpose of this exercise is to conduct an ANOVA test on the given dentist implant data and to find out the interaction between the different factors affecting metal implants in dental cavities. The provided dataset contains 90 different observations on dental implant data based on the method of implant, the temperature at which the metal is treated, and the alloy used as well as on the dentists who may favour one method above another and may work better in his/her favourite method. The response is the variable of interest.

DATA DESCRIPTION

1. Dentist: Data collected from five different dentists
2. Method: The method of metal implant totals 3 type
3. Alloy: alloy used in the metal implant (2 types)
4. Temperature: The temperature at which the metal is treated (3 types)
5. Response: The variable of interest

SAMPLE OF THE DATASET

	Dentist	Method	Alloy	Temp	Response
0	1.0	1.0	1.0	1500.0	813.0
1	1.0	1.0	1.0	1600.0	792.0
2	1.0	1.0	1.0	1700.0	792.0
3	1.0	1.0	2.0	1500.0	907.0
4	1.0	1.0	2.0	1600.0	792.0

Table 1 – Sample dataset

`

The dataset has 5 variables which affect the metal implants in dental cavities.

EXPLORATORY DATA ANALYSIS

The dataset contains 107 entries out of which 17 entries are null values. After dropping the null values the dataset is

NO.	Column	Non – Null content	Data Type
1	Dentist	90	Float64
2	Method	90	Float64
3	Alloy	90	Float64
4	Temperature	90	Float64
5	Response	90	Float64

Table 2 – Exploratory Data Analysis

There are 90 entries and 5 columns in the dataset after removing null values

DESCRIPTIVE DATA ANALYSIS

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Dentist	90	5	Dentist_1	18	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Method	90	3	Method_1	30	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Alloy	90	2	Alloy_1	45	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Temp	90.0	NaN	NaN	NaN	1600.0	82.107083	1500.0	1500.0	1600.0	1700.0	1700.0
Response	90.0	NaN	NaN	NaN	741.777778	145.767845	289.0	698.0	767.0	824.0	1115.0

Table 3 – Descriptive Data Analysis

1. Dentist have five unique value, method has three and alloy has two.
2. Total 90 entries are there.
3. Maximum value of temperature is 1700 and minimum value is 1500.

`

Problem 1 – ANOVA

Dental implant data: The hardness of metal implants in dental cavities depends on multiple factors, such as the method of implant, the temperature at which the metal is treated, the alloy used as well as the dentists who may favour one method above another and may work better in his/her favourite method. The response is the variable of interest.

1. **Test whether there is any difference among the dentists on the implant hardness. State the null and alternative hypotheses. Note that both types of alloys cannot be considered together. You must state the null and alternative hypotheses separately for the two types of alloys.**

The problem statement is to consider the alloys separately and to test whether any difference among the dentist on the implant hardness. So we have to group by the parent dataset based on alloy 1 and alloy 2. We have changed the 1,2,3 values in dentist column to Dentist_1, Dentist_2 upto Dentist_5, the values in method column to Method_1 upto Method_3 and alloy changes to Alloy_1 and Alloy_2. The updated dataset for alloy 1 and alloy 2 is,

	Dentist	Method	Alloy	Temp	Response
0	Dentist_1	Method_1	Alloy_1	1500.0	813.0
1	Dentist_1	Method_1	Alloy_1	1600.0	792.0
2	Dentist_1	Method_1	Alloy_1	1700.0	792.0
6	Dentist_1	Method_2	Alloy_1	1500.0	782.0
7	Dentist_1	Method_2	Alloy_1	1600.0	698.0

Table 1.1 – Sample dataset for Alloy 1

	Dentist	Method	Alloy	Temp	Response
3	Dentist_1	Method_1	Alloy_2	1500.0	907.0
4	Dentist_1	Method_1	Alloy_2	1600.0	792.0
5	Dentist_1	Method_1	Alloy_2	1700.0	835.0
9	Dentist_1	Method_2	Alloy_2	1500.0	1115.0
10	Dentist_1	Method_2	Alloy_2	1600.0	835.0

Table 1.2 – Sample dataset for Alloy 2

`

Hypothesis for Alloy 1

Null hypothesis states that the mean of response of implant hardness is equal for all the dentists.

H_0 : μ_{D1}= μ_{D2}=μ_{D3}=μ_{D4}=μ_{D5}

H_0: The means of response of implant hardness between the various dentists are equal.

Alternative hypothesis states that there will be an effect of "Dentist" on at least one of the levels in response of implant hardness. The mean response of implant hardness for at least one category of Dentist are unequal.

H_1: $\mu_{D1} \neq \mu_{D2}$=μ_{D3}=μ_{D4}=μ_{D5}

H_1: $\mu_{D1} = \mu_{D2} \neq \mu_{D3}$=$\mu_{D4}$=$\mu_{D5}$

H_1: $\mu_{D1} = \mu_{D2}$=$\mu_{D3} \neq \mu_{D4}$=μ_{D5}

H_1: $\mu_{D1} = \mu_{D2}$=μ_{D3}=$\mu_{D4} \neq \mu_{D5}$

H_1: $\mu_{D1} \neq \mu_{D2} \neq \mu_{D3}$=$\mu_{D4}$=$\mu_{D5}$

H_1: $\mu_{D1} = \mu_{D2} \neq \mu_{D3} \neq \mu_{D4}$=$\mu_{D5}$

H_1: $\mu_{D1} = \mu_{D2}$=$\mu_{D3} \neq \mu_{D4} \neq \mu_{D5}$

H_1: $\mu_{D1} \neq \mu_{D2}$=μ_{D3}=$\mu_{D4} \neq \mu_{D5}$

H_1: $\mu_{D1} \neq \mu_{D2} \neq \mu_{D3} \neq \mu_{D4}$=$\mu_{D5}$

H_1: $\mu_{D1} = \mu_{D2} \neq \mu_{D3} \neq \mu_{D4} \neq \mu_{D5}$

H_1: $\mu_{D1} \neq \mu_{D2}$=$\mu_{D3} \neq \mu_{D4} \neq \mu_{D5}$

H_1: $\mu_{D1} \neq \mu_{D2} \neq \mu_{D3}$=$\mu_{D4} \neq \mu_{D5}$

H_1: $\mu_{D1} \neq \mu_{D2} \neq \mu_{D3} \neq \mu_{D4} \neq \mu_{D5}$

H_1: At least one of the means between the responses of implant hardness with respect to the various dentists are unequal.

`

Hypothesis for Alloy 2

Similarly for alloy 2, null hypothesis states that the mean of response of implant hardness is equal for all the dentists.

$H_0 : \mu_{D1} = \mu_{D2} = \mu_{D3} = \mu_{D4} = \mu_{D5}$

H_0: The means of response of implant hardness between the various manufacturers are equal.

Alternative hypothesis states that there will be an effect of "Dentist" on at least one of the levels in response of implant hardness. The mean response of implant hardness for at least one category of Dentist are unequal.

$H_1: \mu_{D1} \neq \mu_{D2} = \mu_{D3} = \mu_{D4} = \mu_{D5}$

$H_1: \mu_{D1} = \mu_{D2} \neq \mu_{D3} = \mu_{D4} = \mu_{D5}$

$H_1: \mu_{D1} = \mu_{D2} = \mu_{D3} \neq \mu_{D4} = \mu_{D5}$

$H_1: \mu_{D1} = \mu_{D2} = \mu_{D3} = \mu_{D4} \neq \mu_{D5}$

$H_1: \mu_{D1} \neq \mu_{D2} \neq \mu_{D3} = \mu_{D4} = \mu_{D5}$

$H_1: \mu_{D1} = \mu_{D2} \neq \mu_{D3} \neq \mu_{D4} = \mu_{D5}$

$H_1: \mu_{D1} = \mu_{D2} = \mu_{D3} \neq \mu_{D4} \neq \mu_{D5}$

$H_1: \mu_{D1} \neq \mu_{D2} = \mu_{D3} = \mu_{D4} \neq \mu_{D5}$

$H_1: \mu_{D1} \neq \mu_{D2} \neq \mu_{D3} = \mu_{D4} = \mu_{D5}$

$H_1: \mu_{D1} = \mu_{D2} \neq \mu_{D3} \neq \mu_{D4} \neq \mu_{D5}$

$H_1: \mu_{D1} \neq \mu_{D2} = \mu_{D3} \neq \mu_{D4} \neq \mu_{D5}$

$H_1: \mu_{D1} \neq \mu_{D2} \neq \mu_{D3} = \mu_{D4} \neq \mu_{D5}$

$H_1: \mu_{D1} \neq \mu_{D2} \neq \mu_{D3} \neq \mu_{D4} \neq \mu_{D5}$

H_1: At least one of the means between the responses of implant hardness with respect to the various dentists are unequal.

2. **Before the hypotheses may be tested, state the required assumptions. Are the assumptions fulfilled? Comment separately on both alloy types.**

The assumptions for ANOVA are,

1. All populations under consideration have a normal distribution
2. All populations under consideration have equal variances.
3. The sample is a random sample, i.e. the observations are collected independently of each other.

We will conduct the Shapiro test on the subset dataset for the normal distribution of the data. If the p-value of the Shapiro test is less than 0.05 we will consider the data is normally distributed. For the subset of alloy 1 and alloy 2 dataset p-value is greater than 0.05. Levene's test value of the p-value is also greater than 0.05.

3. **Irrespective of your conclusion in 7.2, we will continue with the testing procedure. What do you conclude regarding whether implant hardness depends on dentists? Clearly state your conclusion. If the null hypothesis is rejected, is it possible to identify which pairs of dentists differ?**

Irrespective of the result from question number 2 we will continue with the testing procedure and the output from the ANOVA test between the response and dentist for alloy 1 is shown below,

	df	sum_sq	mean_sq	F	PR(>F)
Dentist	4.0	106683.688889	26670.922222	1.977112	0.116567
Residual	40.0	539593.555556	13489.838889	NaN	NaN

Table 3.1 – AOV table for Response and Dentist for alloy 1

The ANOVA test statistics for alloy 2 is shown below,

	df	sum_sq	mean_sq	F	PR(>F)
Dentist	4.0	5.679791e+04	14199.477778	0.524835	0.718031
Residual	40.0	1.082205e+06	27055.122222	NaN	NaN

Table 3.2 – AOV table for Response and Dentist for alloy 2

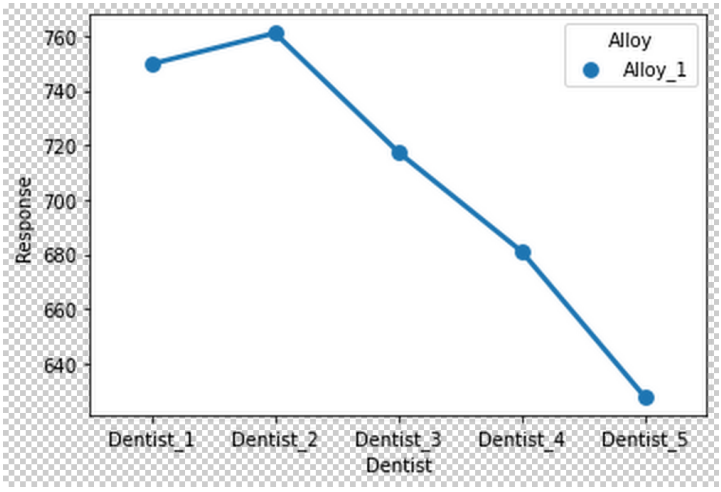

Fig 3.1 – Response to dentist for Alloy 1

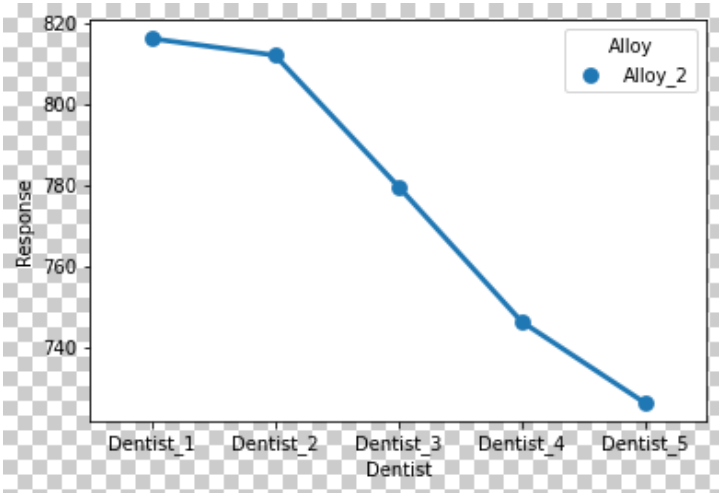

Fig 3.2 – Response to dentist for Alloy 2

From the output it is clear that the p-value for the ANOVA test between response and dentist is greater than 0.05, therefore we cannot reject the null hypothesis. Therefore mean of response of implant hardness is equal for all the dentists.

If the p-value from the ANOVA test is less than 0.05 we can reject the null hypothesis and we can go for multicomparison test to find out which pairs of dentist differ.

`

4. **Now test whether there is any difference among the methods on the hardness of dental implant, separately for the two types of alloys. What are your conclusions? If the null hypothesis is rejected, is it possible to identify which pairs of methods differ?**

Now we consider whether there is any difference among the method on the hardness of dental implant, separately for the two type of alloys.

Let us check the null and alternate hypothesis based on method for alloy 1 and alloy 2 separately.

Hypothesis for Alloy 1

Null hypothesis states that the mean of response of implant hardness is equal for all the methods.

H_0 : $\mu_{M1} = \mu_{M2} = \mu_{M3}$

H_0: The means of response of implant hardness between the various methods are equal.

Alternative hypothesis states that there will be an effect of "Method" on at least one of the levels in response of implant hardness. The mean response of implant hardness for at least one category of method are unequal.

H_1: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$

H_1: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$

H_1: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$

H_1: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H1: At least one of the means between the responses of implant hardness with respect to the various methods are unequal.

Hypothesis for Alloy 2

Null hypothesis states that the mean of response of implant hardness is equal for all the methods.

H_0 : $\mu_{M1} = \mu_{M2} = \mu_{M3}$

H_0: The means of response of implant hardness between the various methods are equal.

Alternative hypothesis states that there will be an effect of "Method" on at least one of the levels in response of implant hardness. The mean response of implant hardness for at least one category of method are unequal.

$H_1: \mu_{M1} \neq \mu_{M2} = \mu_{M3}$

$H_1: \mu_{M1} = \mu_{M2} \neq \mu_{M3}$

$H_1: \mu_{M1} = \mu_{M3} \neq \mu_{M2}$

$H_1: \mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H1: At least one of the means between the responses of implant hardness with respect to the various methods are unequal.

Now let us check the ANOVA test for response and method for alloy 1 and alloy 2 separately. The output from the ANOVA test between the response and method for alloy 1 is shown below,

	df	sum_sq	mean_sq	F	PR(>F)
Method	2.0	148472.177778	74236.088889	6.263327	0.004163
Residual	42.0	497805.066667	11852.501587	NaN	NaN

Table 4.1 – AOV table for Response and Method for alloy 1

The ANOVA test statistics for alloy 2 is shown below,

	df	sum_sq	mean_sq	F	PR(>F)
Method	2.0	499640.4	249820.200000	16.4108	0.000005
Residual	42.0	639362.4	15222.914286	NaN	NaN

Table 4.2 – AOV table for Response and Method for alloy 2

From the output table it is clear that the p value of the ANOVA test is less than 0.05 therefore we can reject null hypothesis in this case, that means at least one of the means between the responses of implant hardness with respect to the various methods are unequal.

To find out which pair of method is different we can go for multicomparison.

The output result of multicomparison for alloy 1 is shown below,

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Method_1	Method_2	-6.1333	0.987	-102.714	90.4473	False
Method_1	Method_3	-124.8	0.0085	-221.3807	-28.2193	True
Method_2	Method_3	-118.6667	0.0128	-215.2473	-22.086	True

Table 4.3 – Multicomparison between method for alloy 1

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Method_1	Method_2	27.0	0.8212	-82.4546	136.4546	False
Method_1	Method_3	-208.8	0.0001	-318.2546	-99.3454	True
Method_2	Method_3	-235.8	0.0	-345.2546	-126.3454	True

Table 4.4 – Multicomparison between method for alloy 2

From the multicomparison table it is clear that we fail to reject the null hypothesis for Method_1 and Method_2 for alloy 1 and alloy 2, for all other pairs we can reject the null hypothesis. That is the pair Method_1 Method_2 pair is statistically not significant and will not affect the implant hardness but other two pairs have p-value less than 0.05 and we can reject the null hypothesis and these pairs are statistically significant, and these 4 pairs will affect the implant hardness.

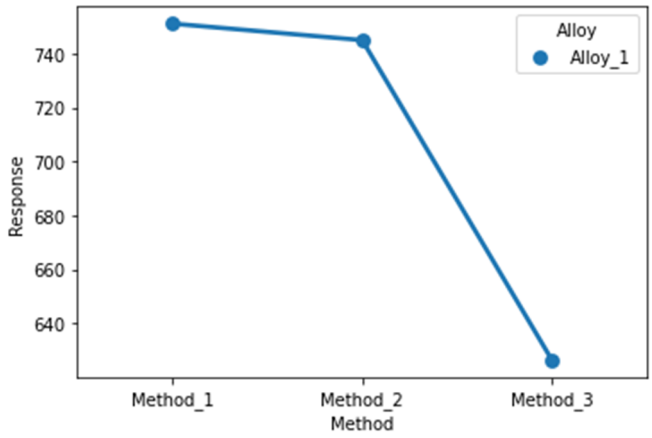

Fig 4.1 – Response to method for Alloy 1

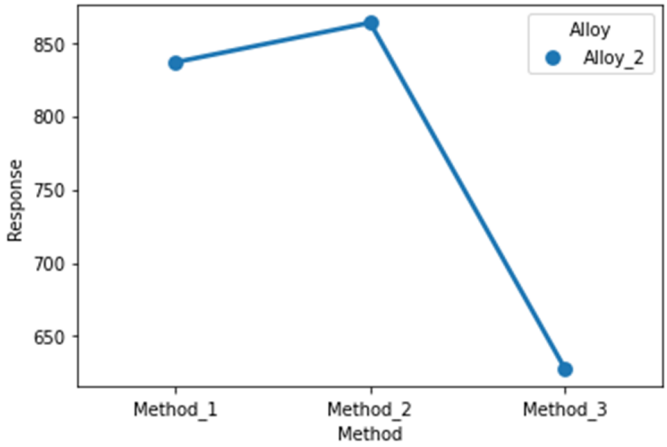

Fig 4.2 – Response to method for Alloy 2

From the point plot it is clear that the distance between the method 1 and method 2 is less compared to other pairs for both alloy 1 and alloy 2. So we can consider method 1 and method 2 and reject other pairs.

5. **Now test whether there is any difference among the temperature levels on the hardness of dental implant, separately for the two types of alloys. What are your conclusions? If the null hypothesis is rejected, is it possible to identify which levels of temperatures differ?**

Now we consider whether there is any difference among the temperature on the hardness of dental implant, separately for the two type of alloys.

Let us check the null and alternate hypothesis based on temperature for alloy 1 and alloy 2 separately.

We have 3 different temperature value here, we can consider

1500 = T_1 , 1600 = T_2 and 1700 = T_3

Hypothesis for Alloy 1

Null hypothesis states that the mean of response of implant hardness is equal for all the temperature.

H_0 : $\mu_{T1} = \mu_{T2} = \mu_{T3}$

H_0: The means of response of implant hardness between the various temperatures are equal.

Alternative hypothesis states that there will be an effect of "Temperature" on at least one of the levels in response of implant hardness. The mean response of implant hardness for at least one value of temperature are unequal.

H_1: $\mu_{T1} \neq \mu_{T2} = \mu_{T3}$

H_1: $\mu_{T1} = \mu_{T2} \neq \mu_{T3}$

H_1: $\mu_{T1} = \mu_{T3} \neq \mu_{T2}$

H_1: $\mu_{T1} \neq \mu_{T2} \neq \mu_{T3}$

H1: At least one of the means between the responses of implant hardness with respect to the various temperature are unequal.

Hypothesis for Alloy 2

Null hypothesis states that the mean of response of implant hardness is equal for all the temperature.

H_0 : $\mu_{T1} = \mu_{T2} = \mu_{T3}$

H_0: The means of response of implant hardness between the various temperatures are equal.

Alternative hypothesis states that there will be an effect of "Temperature" on at least one of the levels in response of implant hardness. The mean response of implant hardness for at least one value of temperature are unequal.

$H_1: \mu_{T1} \neq \mu_{T2} = \mu_{T3}$

$H_1: \mu_{T1} = \mu_{T2} \neq \mu_{T3}$

$H_1: \mu_{T1} = \mu_{T3} \neq \mu_{T2}$

$H_1: \mu_{T1} \neq \mu_{T2} \neq \mu_{T3}$

H1: At least one of the means between the responses of implant hardness with respect to the various temperature are unequal.

Now let us check the ANOVA test for response and temperature for alloy 1 and alloy 2 separately. The output from the ANOVA test between the response and temperature for alloy 1 is shown below,

	df	sum_sq	mean_sq	F	PR(>F)
Temp	1.0	10083.333333	10083.333333	0.681527	0.413618
Residual	43.0	636193.911111	14795.207235	NaN	NaN

Table 5.1 – AOV table for Response and temperature for alloy 1

	df	sum_sq	mean_sq	F	PR(>F)
Temp	1.0	8.629603e+04	86296.033333	3.524941	0.067246
Residual	43.0	1.052707e+06	24481.552713	NaN	NaN

Table 5.2 – AOV table for Response and temperature for alloy 2

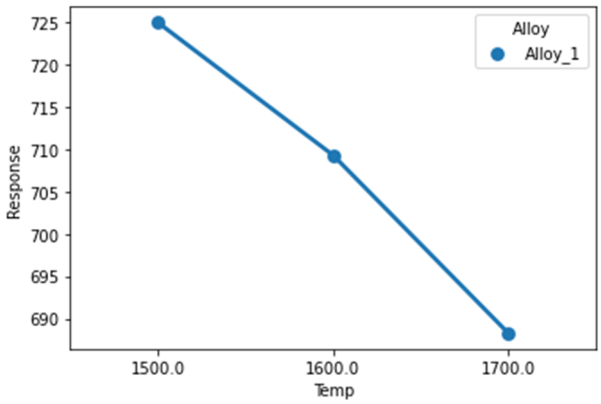

Fig 5.1 – Response to temperature for Alloy 1

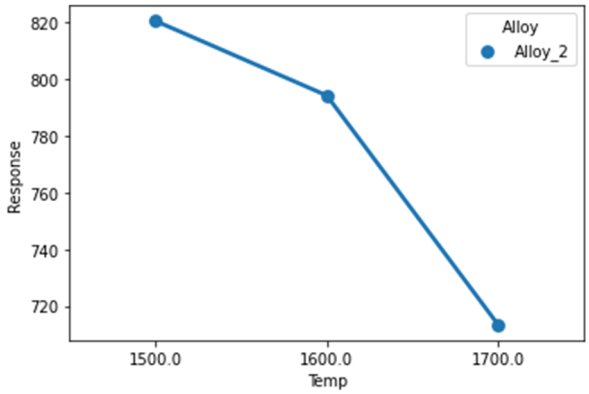

Fig 5.2 – Response to temperature for Alloy 2

From the output it is clear that the p-value for the ANOVA test between response and temperature is greater than 0.05, therefore we cannot reject the null hypothesis. Therefore mean of response of implant hardness is equal for all the temperature.

If the p-value from the ANOVA test is less than 0.05 we can reject the null hypothesis and we can go for multicomparison test to find out which pairs of temperature differ.

6. Consider the interaction effect of dentist and method and comment on the interaction plot, separately for the two types of alloys?

Now let us consider the interaction effect between the dentist and method on the hardness of dental implants for alloy 1,

	df	sum_sq	mean_sq	F	PR(>F)
Dentist	4.0	106683.688889	26670.922222	3.899638	0.011484
Method	2.0	148472.177778	74236.088889	10.854287	0.000284
Dentist:Method	8.0	185941.377778	23242.672222	3.398383	0.006793
Residual	30.0	205180.000000	6839.333333	NaN	NaN

Table 6.1 – AOV table for interaction between Response, dentist and method for alloy 1

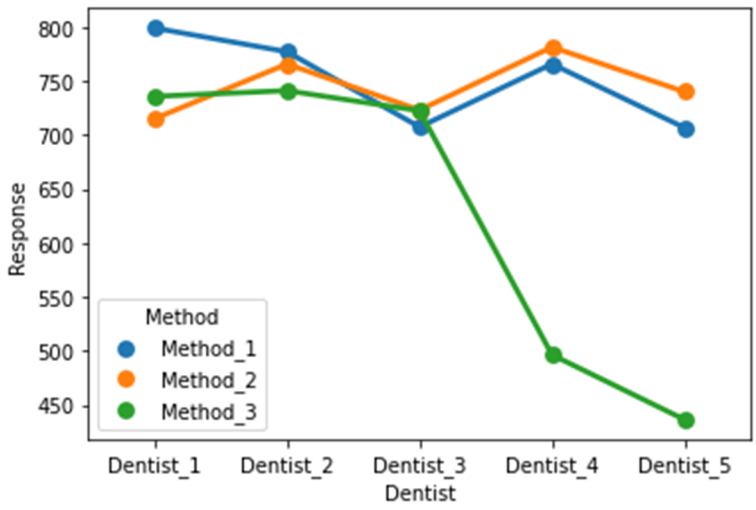

Fig 6.1 –interaction plot between Response, dentist and method for alloy 1

From the output of the ANOVA test between the interaction of dentist and method it is clearly visible that all the variables, dentist and method individually and when we consider the effect of dentist and method together on the response of the hardness of the dental implant, are statistically significant and they all significantly affect the response of the hardness of the dental implant. The p-value of all the variables is less than 0.05, therefore we can reject the null hypothesis in the case of interaction effect for alloy 1.

Now let us consider the point plot for the same, from the point plot it is clearly visible that the interaction between the method_1, method_2 and method_3. At the points dentist_1, dentist_2 and dentist_3 the lines are overlapping to each other which implies the interaction between these.

Now let us consider the interaction effect between the dentist and method on the hardness of dental implants for alloy 2,

	df	sum_sq	mean_sq	F	PR(>F)
Dentist	4.0	56797.911111	14199.477778	1.106152	0.371833
Method	2.0	499640.400000	249820.200000	19.461218	0.000004
Dentist:Method	8.0	197459.822222	24682.477778	1.922787	0.093234
Residual	30.0	385104.666667	12836.822222	NaN	NaN

Table 6.2 – AOV table for interaction between Response, dentist and method for alloy 2

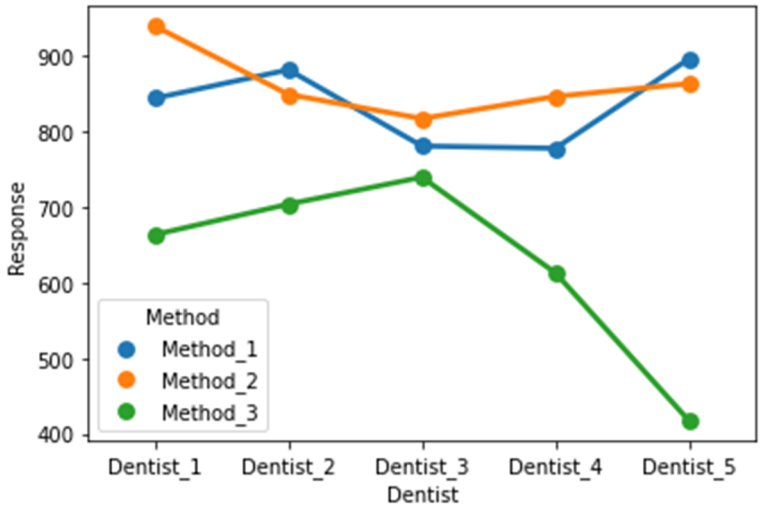

Fig 6.2 –interaction plot between Response, dentist and method for alloy 2

From the output from the ANOVA test the variable dentist and the interaction between the dentist and method are statistically not significant, the p-value of these variables are greater than 0.05 therefore we cannot reject the null hypothesis in this case. But the variable method is statistically significant and the p value is less than 0.05 therefore we can reject null hypothesis in this case, i.e. for the method variable at least one of the means is different with respect to response of the hardness of the dental implant.

From the interaction plot also it is clearly visible the interaction between the method _1 and method_2 since they are overlapping to each other.

7. **Now consider the effect of both factors, dentist, and method, separately on each alloy. What do you conclude? Is it possible to identify which dentists are different, which methods are different, and which interaction levels are different?**

Now we can perform two way ANOVA by considering the factors dentist and method on reaction separately for alloy 1 and alloy 2.

Let us consider alloy 1 first,

Consider the below table which shows output of two way ANOVA by considering dentist and method on implant hardness for alloy 1,

	df	sum_sq	mean_sq	F	PR(>F)
Dentist	4.0	106683.688889	26670.922222	2.591255	0.051875
Method	2.0	148472.177778	74236.088889	7.212522	0.002211
Residual	38.0	391121.377778	10292.667836	NaN	NaN

Table 7.1 – Two way ANOVA table for dentist and method on reaction

The p-value obtained from ANOVA for dentist is not statistically significant (p-value is 0.0518, which is greater than 0.05), therefore it won't affect the hardness of the dental implant. In the case of method the p value is 0.0022 which less than 0.05 and we can reject the null hypothesis in this case, which means the variable method is statistically significant and will affect the hardness of the dental implant.

The below table shows the output of the ANOVA test when we consider dentist and method on reaction with interaction for alloy 1

	df	sum_sq	mean_sq	F	PR(>F)
Dentist	4.0	106683.688889	26670.922222	3.899638	0.011484
Method	2.0	148472.177778	74236.088889	10.854287	0.000284
Dentist:Method	8.0	185941.377778	23242.672222	3.398383	0.006793
Residual	30.0	205180.000000	6839.333333	NaN	NaN

Table 7.2 – Two way ANOVA table for dentist and method on reaction with interaction

The p-value obtained from ANOVA for all the variables considering the interaction between dentist and method on hardness of dental implant is statistically significant because all the p – values are less than 0.05 therefore we can reject the null hypothesis in this case.

Now consider the interaction plot between dentist and method for alloy 1. From the interaction plot it is clearly visible that the lines of Method_1, Method_2 and Method_3 is overlapping each other at the point dentist_1, dentist_2 and dentist_3. But in the case of dentist_4 and dentist_5 the lines are not coinciding each other that means there is no interaction between method_1, method_2 and method_3.

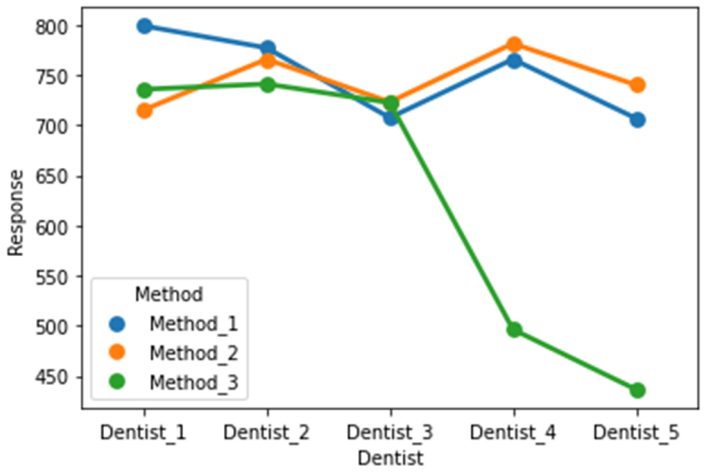

Fig 7.1 –interaction plot between Response, dentist and method for alloy 1

Now Let us consider alloy 2,

	df	sum_sq	mean_sq	F	PR(>F)
Dentist	4.0	56797.911111	14199.477778	0.926215	0.458933
Method	2.0	499640.400000	249820.200000	16.295479	0.000008
Residual	38.0	582564.488889	15330.644444	NaN	NaN

Table 7.3 – Two way ANOVA table for dentist and method on reaction

The p-value obtained from ANOVA for dentist is not statistically significant (p-value is 0.458, which is greater than 0.05), therefore it won't affect the hardness of the dental implant. In the case of method the p value is 0.000008 which less than 0.05 and we can reject the null hypothesis in this case, which means the variable method is statistically significant and will affect the hardness of the dental implant.

	df	sum_sq	mean_sq	F	PR(>F)
Dentist	4.0	56797.911111	14199.477778	1.106152	0.371833
Method	2.0	499640.400000	249820.200000	19.461218	0.000004
Dentist:Method	8.0	197459.822222	24682.477778	1.922787	0.093234
Residual	30.0	385104.666667	12836.822222	NaN	NaN

Table 7.4 – Two way ANOVA table for dentist and method on reaction with interaction

The p-value obtained from ANOVA considering the interaction between the dentist and method is shown in the above table. The p-value of dentist and the interaction of dentist and method are statistically not significant (the p-value is greater than 0.05), so these variables will not affect the hardness of the dental implant. In the case of method the p value is 0.000004 which is less than 0.05, therefore the variable method is statistically significant and will affect the hardness of the dental implant.

Now we can consider the interaction plot between the reaction and dentist based on the method for alloy 2,

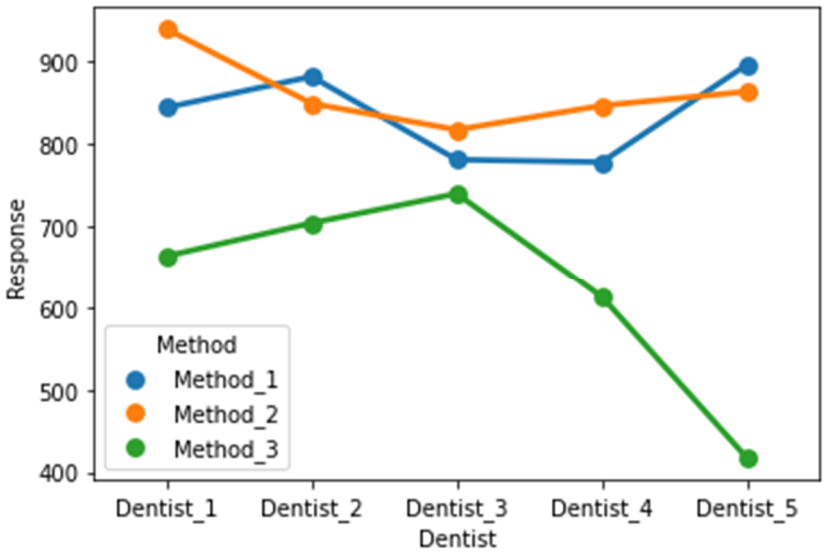

Fig 7.2 –interaction plot between Response, dentist and method for alloy 2

From the plot it is clearly visible that there is interaction between method_1 and method_2 at the points dentist_1 – dentist_2, dentist_2 – dentist_3 and dentist_4 – dentist_5. But in the case of method_3 it does not have any interaction between method_1 and method_2.

Now we can check multicomparison for dentist and method to find out which pair affect the hardness of the dental implant.

Consider the below table, the multicomparison between different pairs of dentist for alloy 1, there are total 10 pairs of dentist are there, from the table it is clear that all the p-adj values of all the pairs are grater than 0.05 which implies that they are not statistically significant, we cannot reject the null hypothesis in this case. So we can conclude that in the case of alloy 1 all pairs of dentist does not have any effect on hardness of the dental implant.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Dentist_1	Dentist_2	11.3333	0.9996	-145.0423	167.709	False
Dentist_1	Dentist_3	-32.3333	0.9757	-188.709	124.0423	False
Dentist_1	Dentist_4	-68.7778	0.7189	-225.1535	87.5979	False
Dentist_1	Dentist_5	-122.2222	0.1889	-278.5979	34.1535	False
Dentist_2	Dentist_3	-43.6667	0.9298	-200.0423	112.709	False
Dentist_2	Dentist_4	-80.1111	0.5916	-236.4868	76.2646	False
Dentist_2	Dentist_5	-133.5556	0.1258	-289.9312	22.8201	False
Dentist_3	Dentist_4	-36.4444	0.9626	-192.8201	119.9312	False
Dentist_3	Dentist_5	-89.8889	0.4805	-246.2646	66.4868	False
Dentist_4	Dentist_5	-53.4444	0.8643	-209.8201	102.9312	False

Table 7.5 – Multicomparison between dentists for alloy 1

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Dentist_1	Dentist_2	-4.1111	1.0	-225.5687	217.3465	False
Dentist_1	Dentist_3	-36.5556	0.9895	-258.0131	184.902	False
Dentist_1	Dentist_4	-70.0	0.8941	-291.4576	151.4576	False
Dentist_1	Dentist_5	-90.1111	0.7724	-311.5687	131.3465	False
Dentist_2	Dentist_3	-32.4444	0.9933	-253.902	189.0131	False
Dentist_2	Dentist_4	-65.8889	0.9132	-287.3465	155.5687	False
Dentist_2	Dentist_5	-86.0	0.8008	-307.4576	135.4576	False
Dentist_3	Dentist_4	-33.4444	0.9925	-254.902	188.0131	False
Dentist_3	Dentist_5	-53.5556	0.9574	-275.0131	167.902	False
Dentist_4	Dentist_5	-20.1111	0.999	-241.5687	201.3465	False

Table 7.6 – Multicomparison between dentists for alloy 2

The above table show the multicomparison between different pairs of dentist for alloy 2, there are total 10 pairs of dentist are there, from the table it is clear that all the p-adj values of all the pairs are greater than 0.05 which implies that they are not statistically significant, we cannot reject the null hypothesis in this case. So we can conclude that in the case of alloy 2 all pairs of dentist does not have any effect on hardness of the dental implant.

Now let us check the case of method,

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Method_1	Method_2	-6.1333	0.987	-102.714	90.4473	False
Method_1	Method_3	-124.8	0.0085	-221.3807	-28.2193	True
Method_2	Method_3	-118.6667	0.0128	-215.2473	-22.086	True

Table 7.7 – Multicomparison between method for alloy 1

The above table shows multicomparison between different pairs of methods for alloy 1, from the table it is clear that the first pair has a p value of 0.987 which is greater than 0.05 therefore this pair is statistically not significant and we cannot reject the null hypothesis, this pair has no effect on hardness of the dental implant. In the case of second and third pair their p value is less than 0.05, therefore they are statistically significant, we can reject the null hypothesis in these cases. These 2 pairs of method have effect on hardness of the dental implant.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Method_1	Method_2	27.0	0.8212	-82.4546	136.4546	False
Method_1	Method_3	-208.8	0.0001	-318.2546	-99.3454	True
Method_2	Method_3	-235.8	0.0	-345.2546	-126.3454	True

Table 7.8 – Multicomparison between method for alloy 2

The above table shows multicomparison between different pairs of methods for alloy 2, from the table it is clear that the first pair has a p value of 0.8212 which is greater than 0.05, therefore this pair is statistically not significant and we cannot reject the null hypothesis, this pair has no effect on hardness of the dental implant. In the case of second and third pair their p value is less than 0.05, therefore they are statistically significant, we can reject the null hypothesis in these cases. These 2 pairs of method have effect on hardness of the dental implant.

`

PROBLEM 2 – SUMMARY

The dataset contains the information about different colleges where we need to take admission post 12th standard. We have to conduct exploratory data analysis on this dataset , univariate, bivariate and multivariate analysis on this dataset and find the relation between different variables in the dataset and draw a conclusion based on the output of the analysis.

INTRODUCTION

The purpose of this exercise is to conduct exploratory data analysis on the provided dataset, univariate, bivariate as well as multivariate analysis. Find the relation between different variables present in the dataset and draw a conclusion based on the observations.

DATA DESCRIPTION

1. Names: Names of various university and colleges
2. Apps: Number of applications received
3. Accept: Number of applications accepted
4. Enroll: Number of new students enrolled
5. Top10perc: Percentage of new students from top 10% of Higher Secondary class
6. Top25perc: Percentage of new students from top 25% of Higher Secondary class
7. F.Undergrad: Number of full-time undergraduate students
8. P.Undergrad: Number of part-time undergraduate students
9. Outstate: Number of students for whom the particular college or university is Out-of-state tuition
10. Room.Board: Cost of Room and board
11. Books: Estimated book costs for a student
12. Personal: Estimated personal spending for a student
13. PhD: Percentage of faculties with Ph.D.’s
14. Terminal: Percentage of faculties with terminal degree
15. S.F.Ratio: Student/faculty ratio
16. perc.alumni: Percentage of alumni who donate
17. Expend: The Instructional expenditure per student

18.Grad.Rate: Graduation rate

SAMPLE OF THE DATASET

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280
2	Adrian College	1428	1097	336	22	50	1036	99	11250
3	Agnes Scott College	417	349	137	60	89	510	63	12960
4	Alaska Pacific University	193	146	55	16	44	249	869	7560

Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
3300	450	2200	70	78	18.1	12	7041	60
6450	750	1500	29	30	12.2	16	10527	56
3750	400	1165	53	66	12.9	30	8735	54
5450	450	875	92	97	7.7	37	19016	59
4120	800	1500	76	72	11.9	2	10922	15

Table 4 – Sample dataset

`

Problem 2 – EDA

The dataset Education - Post 12th Standard.csv contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given.

8. **Perform Exploratory Data Analysis [Univariate, Bivariate, and Multivariate analysis to be performed]. What insight do you draw from the EDA?**

We need to perform the exploratory data analysis on this dataset, first let us check the basic information of the data,

NO.	Column	Null Content	Data type
1	Names	777 non – null	Object
2	Apps	777 non – null	Int 64
3	Accept	777 non – null	Int 64
4	Enroll	777 non – null	Int 64
5	Top10perc	777 non – null	Int 64
6	Top25perc	777 non – null	Int 64
7	F.Undergrad	777 non – null	Int 64
8	P.Undergrad	777 non – null	Int 64
9	Outstate	777 non – null	Int 64
10	Room.Board	777 non – null	Int 64
11	Books	777 non – null	Int 64
12	Personal	777 non – null	Int 64
13	PhD	777 non – null	Int 64
14	Terminal	777 non – null	Int 64
15	S.F.Ratio	777 non – null	Float 64
16	perc.alumni	777 non – null	Int 64
17	Expend	777 non – null	Int 64
18	Grad.Rate	777 non – null	Int 64

Table 8.1 – Data information

There are total 777 entries in the dataset with 18 columns, The name of the college is object type data and student/faculty ration in float data type, all other variables are in int 64 data type.

Now let us check the descriptive data analysis,

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Names	777	777	Abilene Christian University	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Apps	777.0	NaN	NaN	NaN	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	NaN	NaN	NaN	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	NaN	NaN	NaN	779.972973	929.17619	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	NaN	NaN	NaN	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	NaN	NaN	NaN	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	NaN	NaN	NaN	3699.907336	4850.420531	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	NaN	NaN	NaN	855.298584	1522.431887	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	NaN	NaN	NaN	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	NaN	NaN	NaN	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	NaN	NaN	NaN	549.380952	165.10536	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	NaN	NaN	NaN	1340.642214	677.071454	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	NaN	NaN	NaN	72.660232	16.328155	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	NaN	NaN	NaN	79.702703	14.722359	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	NaN	NaN	NaN	14.089704	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	NaN	NaN	NaN	22.743887	12.391801	0.0	13.0	21.0	31.0	64.0
Expend	777.0	NaN	NaN	NaN	9660.171171	5221.76844	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	NaN	NaN	NaN	65.46332	17.17771	10.0	53.0	65.0	78.0	118.0

Table 8.2 – Descriptive data analysis

1. The dataset contains the information about 777 colleges
2. Maximum number of application received is 48094 and minimum number is 81.
3. The mean of book cost is around 549, and maximum cost of books is 2340.
4. The mean of personal cost is 1340 and the maximum personal cost is 6800.
5. The mean of student faculty ratio is 14 and the maximum student faculty ration is 39.8
6. The mean graduation rate is 65.46.

There are no null value as well as duplicate value present in the data set.

Now let us perform univariate, bivariate and multivariate analysis on this dataset.

UNIVARIATE ANALYSIS

Let us check the box plot and hist plot of the variables in the dataset,

Check the below shown box plot and hist plot of application received and application accepted both the plots are right skewed and from the data it is clear that both the variables contains outliers.

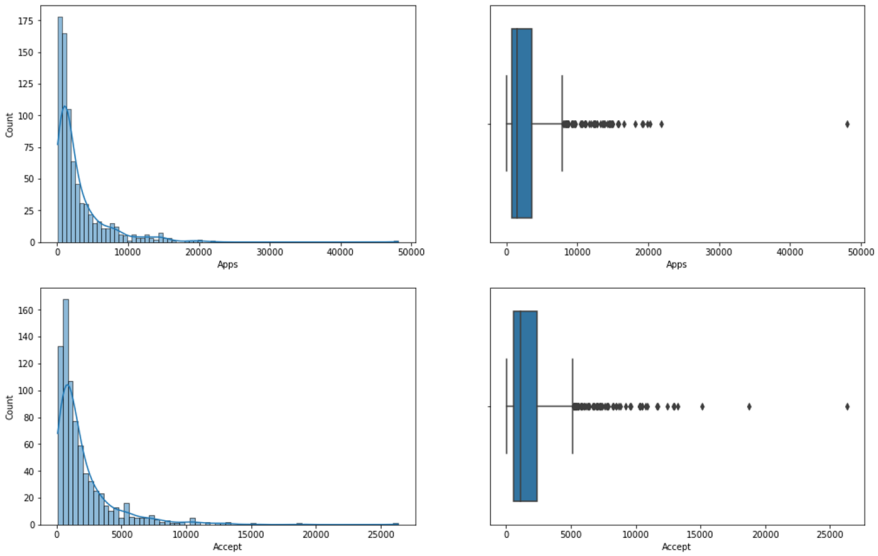

Fig 8.1 – Univariate application received and accepted

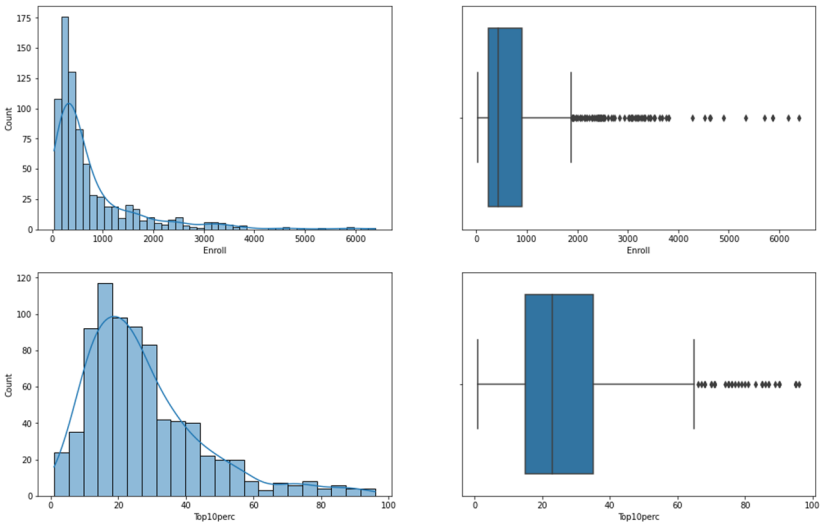

Fig 8.2 – Univariate students enrolled and top 10%

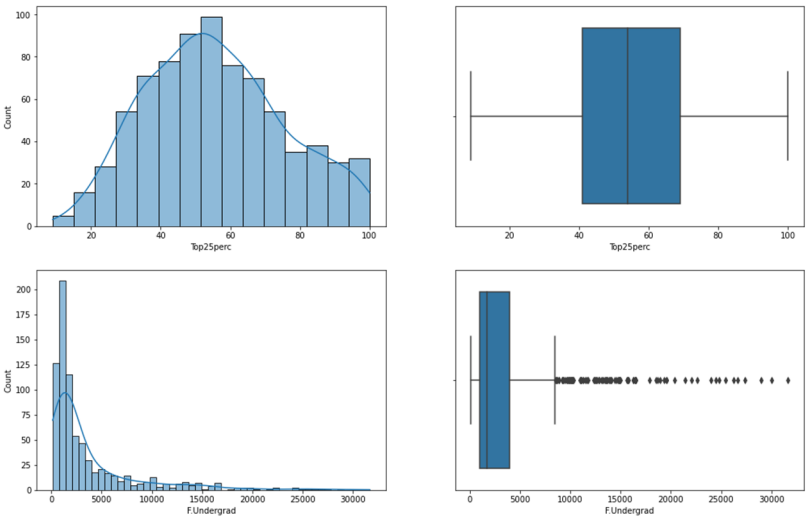

Fig 8.3 – Univariate top 25% and full time graduate

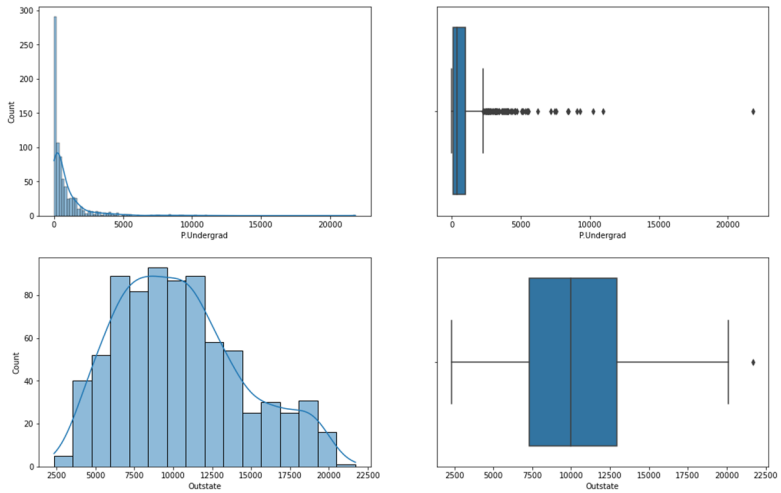

Fig 8.4 – Univariate top Partial graduate and Out of state tuition

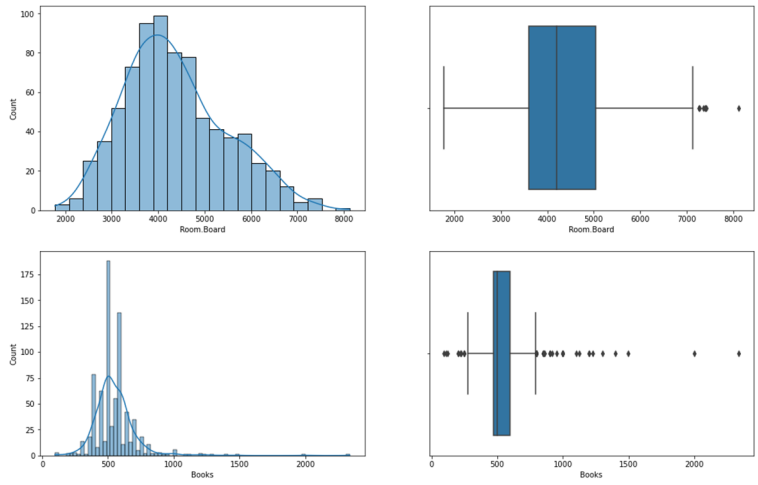

Fig 8.5 – Univariate cost of room and book

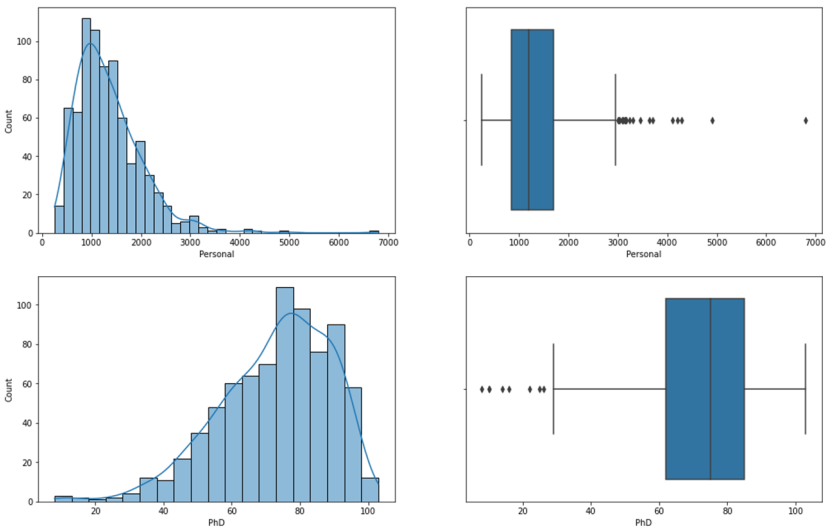

Fig 8.6 – Univariate personal spending and faculty with Phd

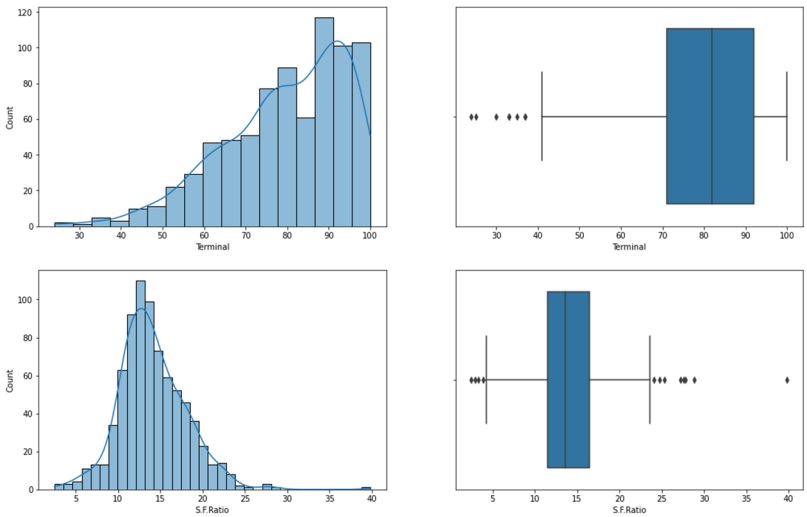

Fig 8.7 – Univariate faculty with terminal degree and student faculty ratio

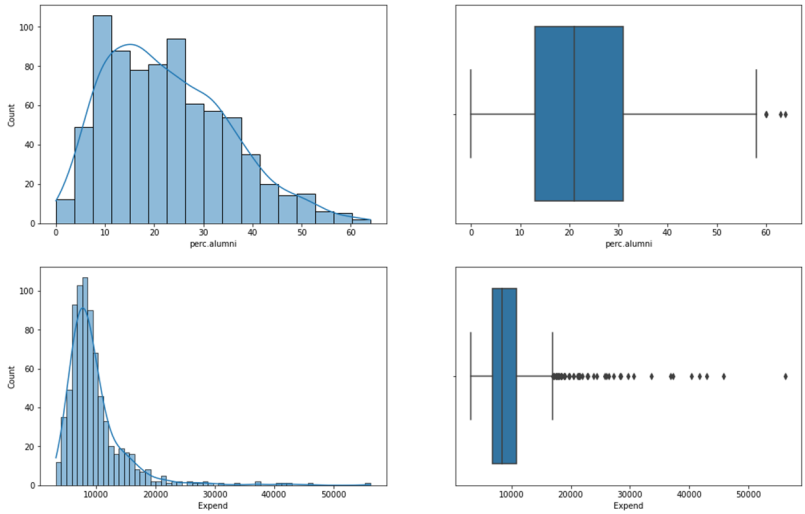

Fig 8.8 – Univariate percentage of alumni donate expenditure per student

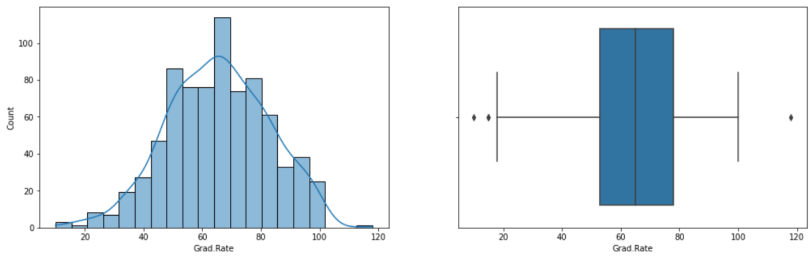

Fig 8.9 – Univariate graduation rate

Consider all the hist plot and box plot shown above in the data set most of the data is left skewed, In the case of out of state tuition, faculty with terminal degree and faculty with PhD and graduation rate the data is right skewed. For student faculty ratio, cost of book and cost of room the data is almost normally distributed. All the

variables contains outliers. These are the conclusions we can reach from the univariate analysis.

BIVARIATE ANALYSIS

Let us consider the top ten colleges based on the application received and accepted and perform bivariate analysis for the dataset.

Consider the below shown bar plots between the colleges and other variables, let us check how they differ according to colleges.

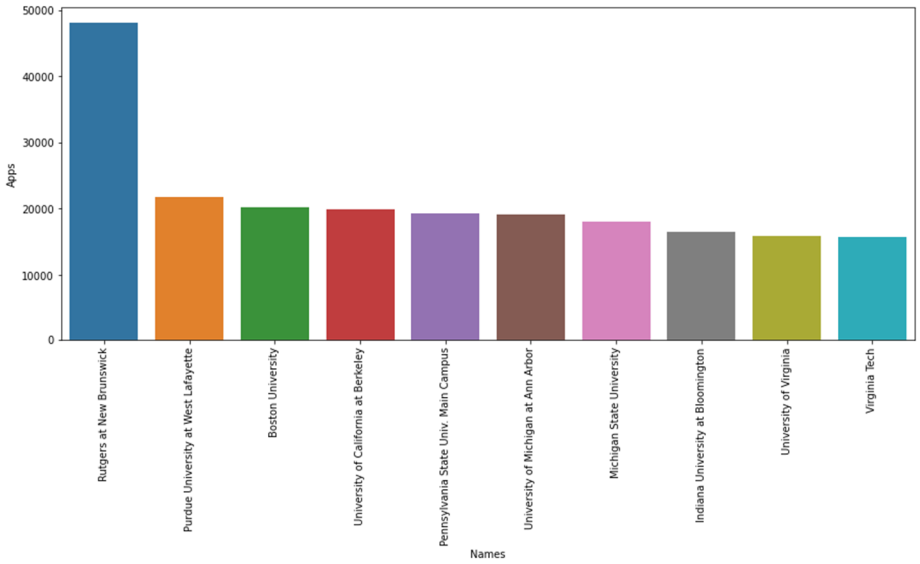

Fig 8.10 – Bivariate application received

When we consider top ten colleges based on the application received, Rutgers at New Brunswick received most number of applications.

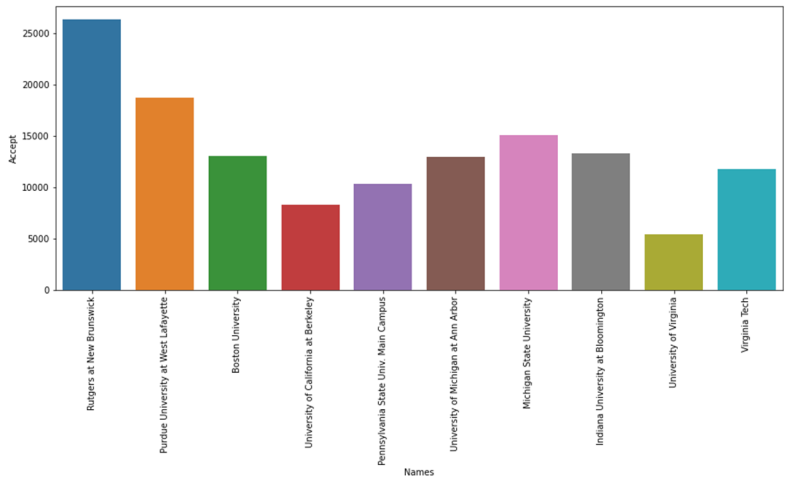

Fig 8.11 – Bivariate application accepted

Rutgers at New Brunswick accepted most number of applications, and university of Virginia accepted least number of applications among the top ten colleges.

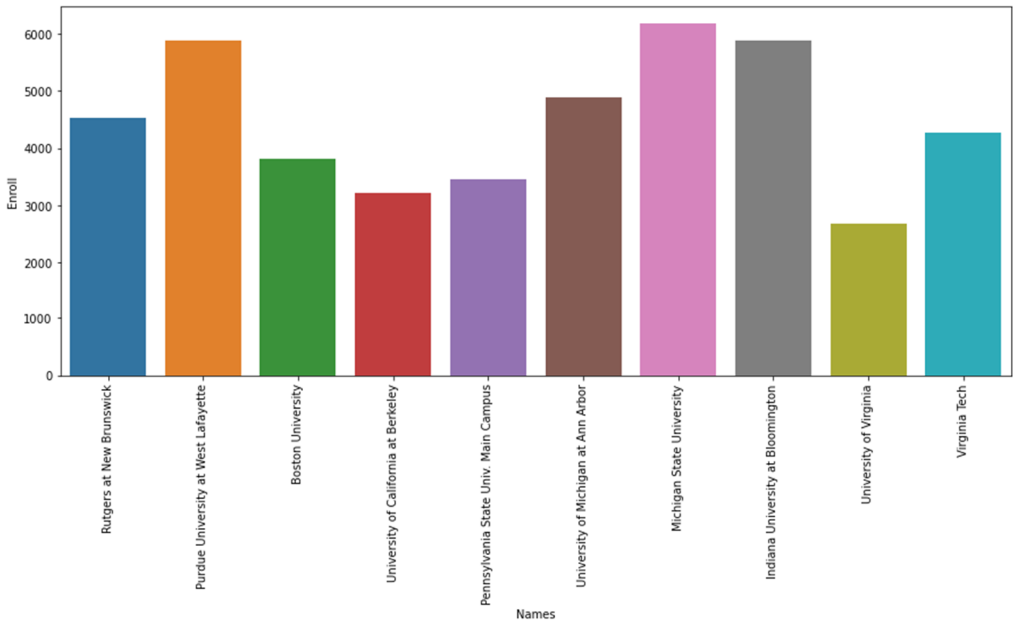

Fig 8.12 – Bivariate Students enrolled

Among the top ten university most students enrolled in Michigan state university

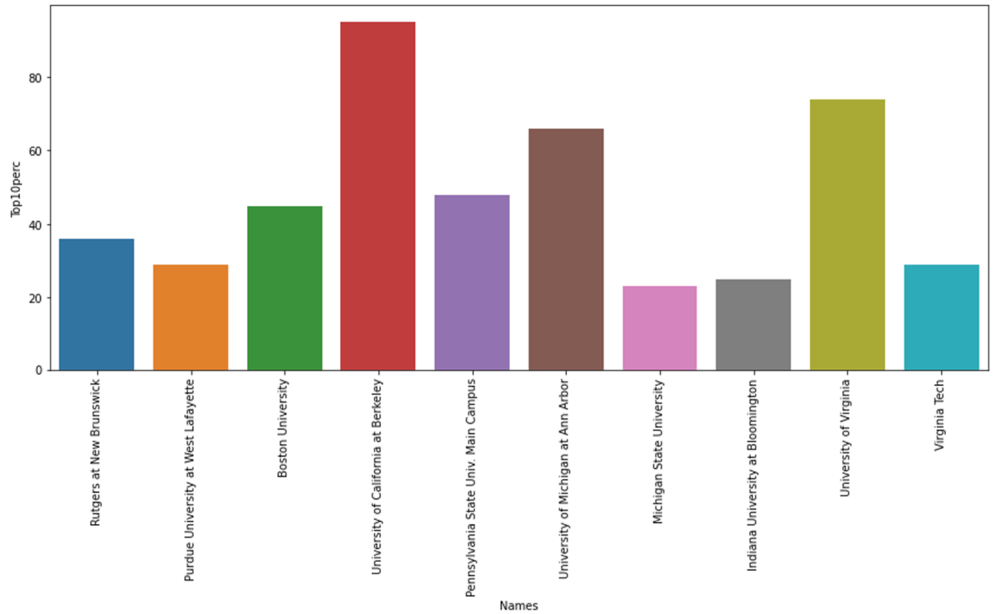

Fig 8.13 – Bivariate Percentage of new students from top 10% of Higher Secondary class

New students from top 10% of Higher Secondary class joined at University of California at Berkley

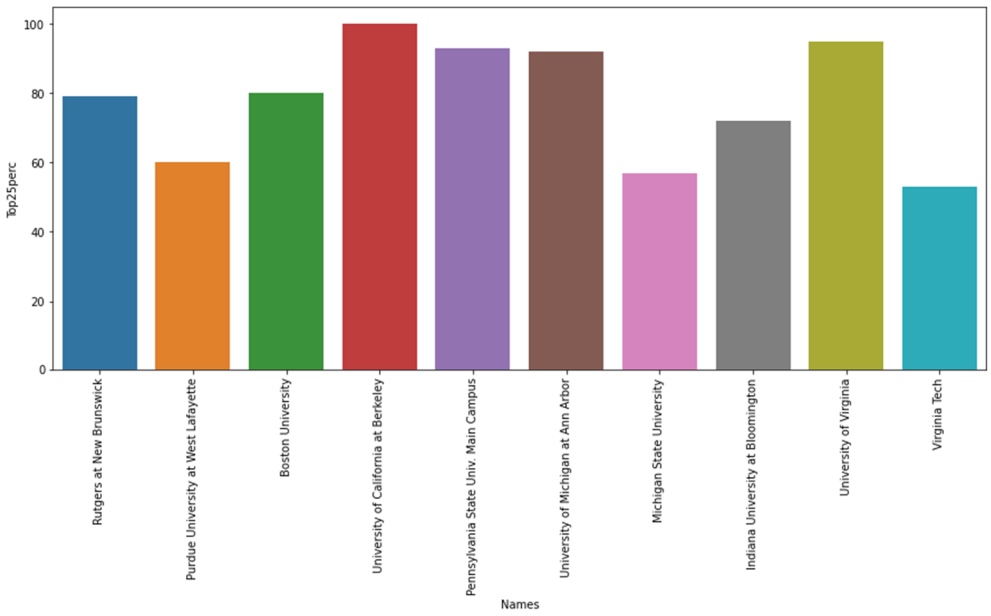

Fig 8.14 – Bivariate Percentage of new students from top 25% of Higher Secondary class

New students from top 25% of Higher Secondary class joined at University of California at Berkley

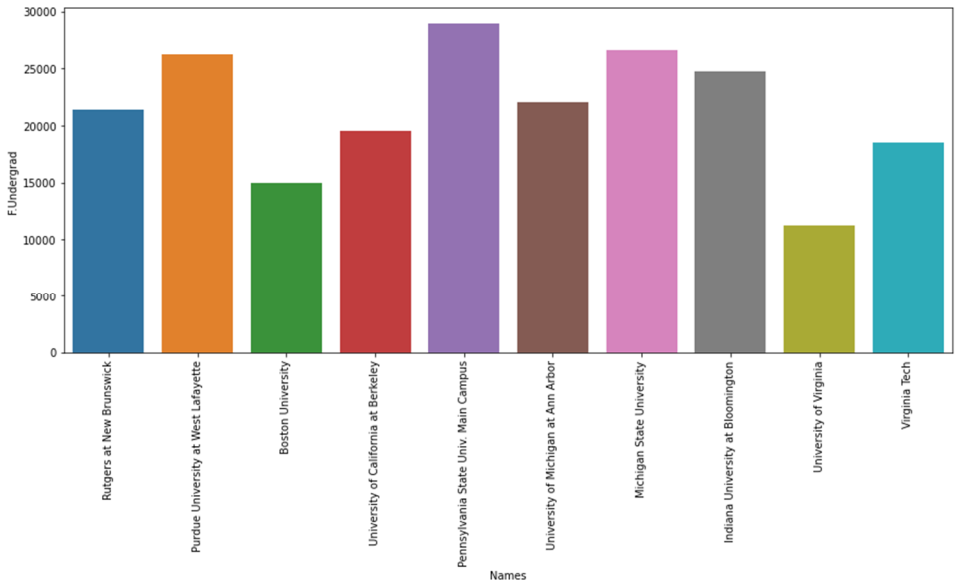

Fig 8.15 – Bivariate Number of full-time undergraduate students

Most of the full time graduate students joined at Pennsylvania state university main campus.

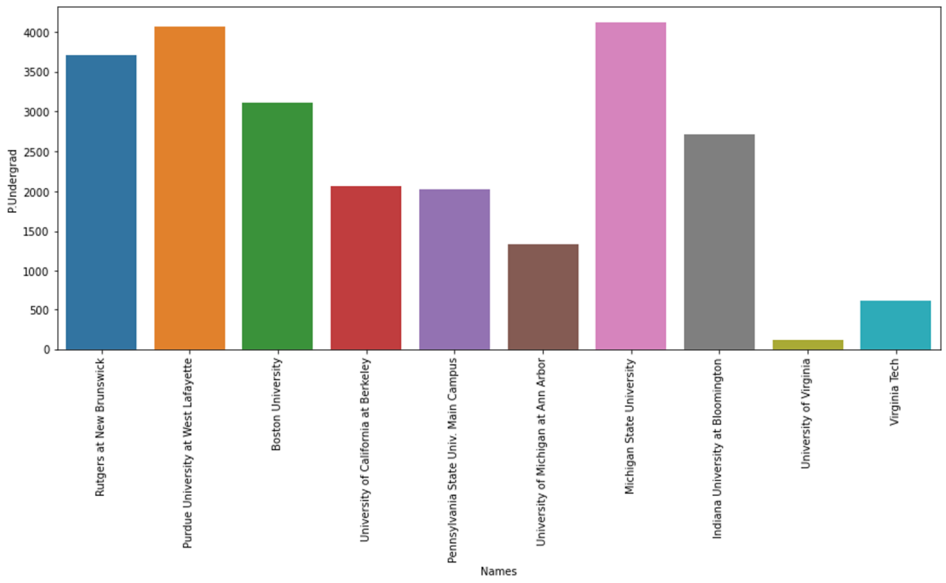

Fig 8.16 – Bivariate Number of part-time undergraduate students

Most of the full time graduate students joined at Michigan state university.

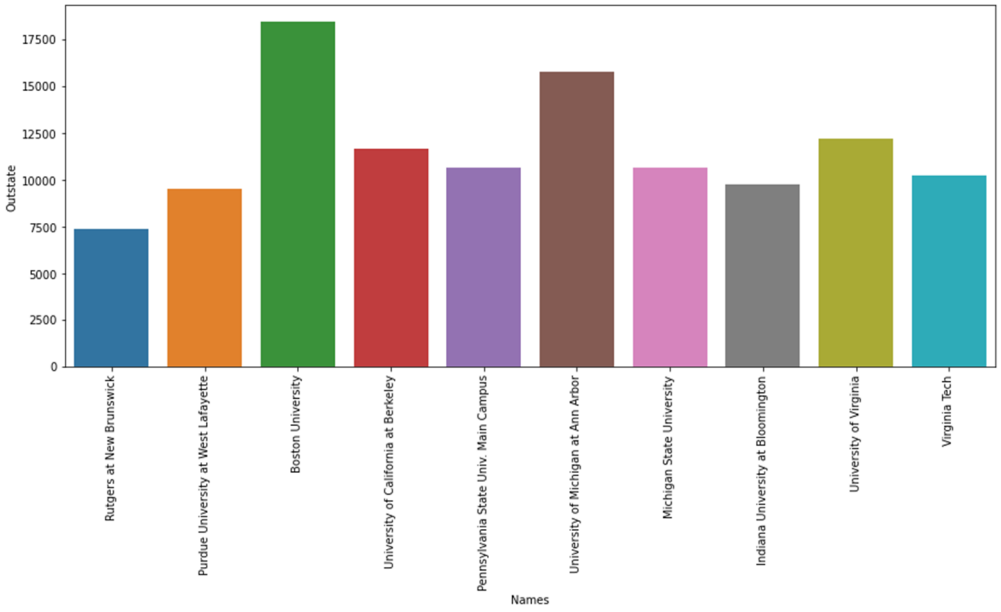

Fig 8.17 – Bivariate Out-of-state tuition

Most of the students whose particular college or university is Out-of-state tuition is joined at Boston University.

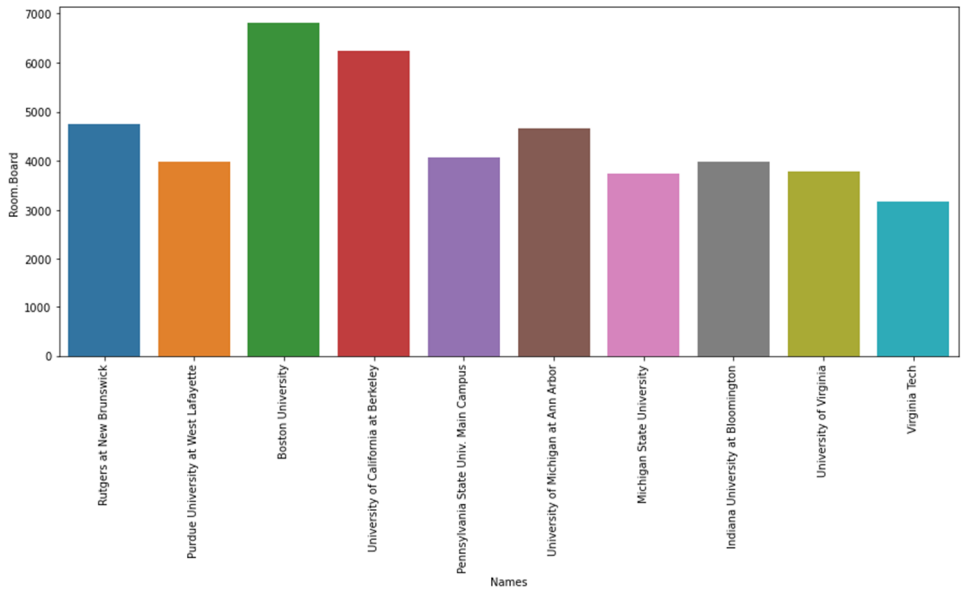

Fig 8.18 – Bivariate Room or board cost

The room cost is more in Boston University compared to other universities

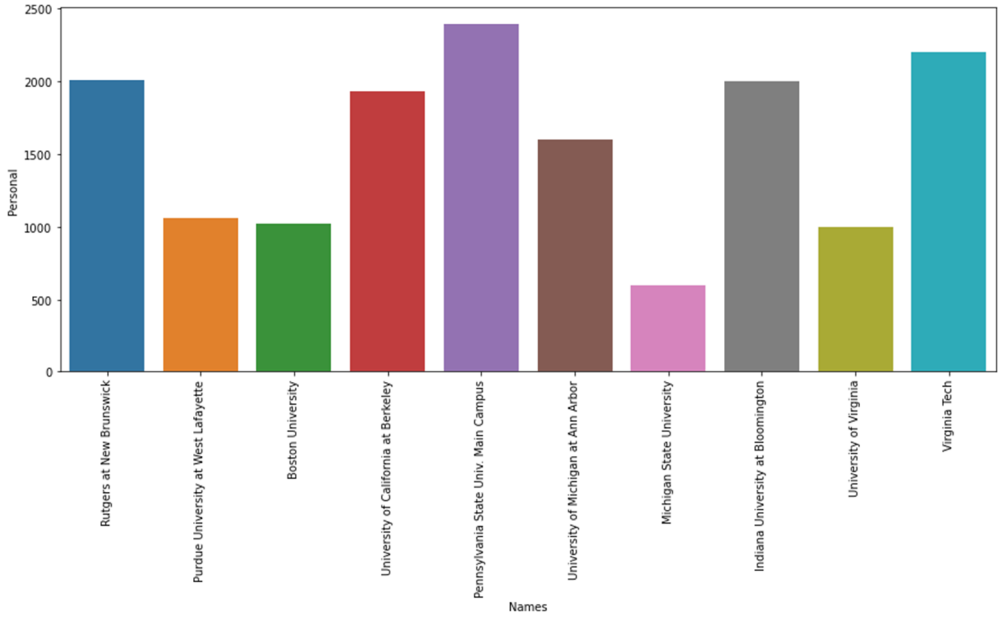

Fig 8.19 – Bivariate Personal cost

The personal cost is more in Pennsylvania state university main campus compared to other universities.

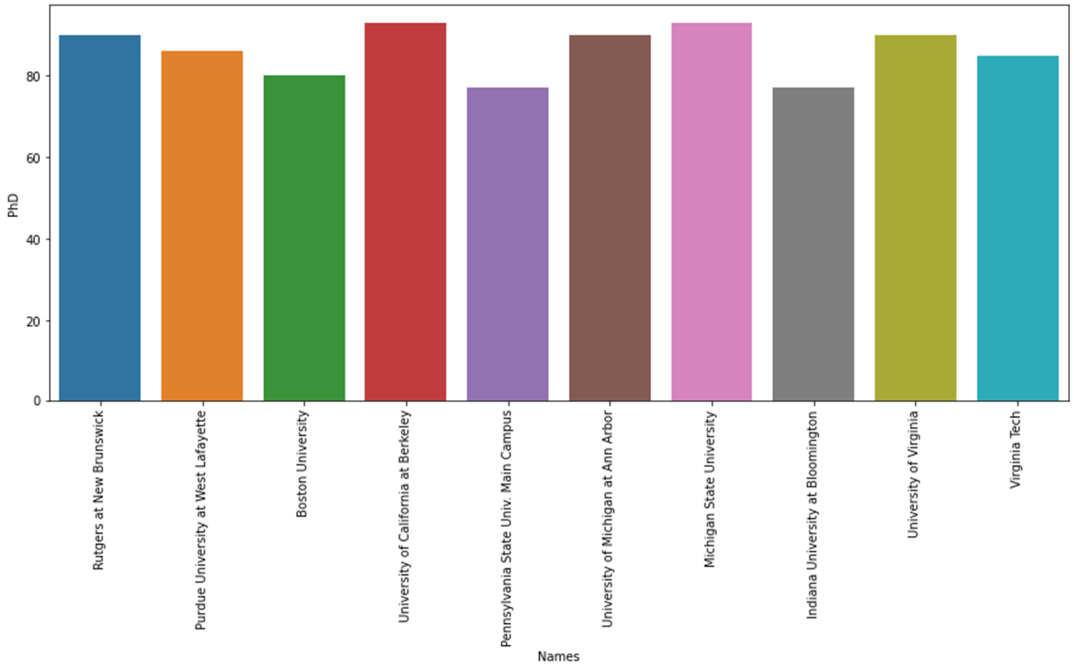

Fig 8.20 – Bivariate Percentage of faculties with Ph.D.'s

All most all the universities contains equal number of faculties with PhD's, the data seems to be normally distributed.

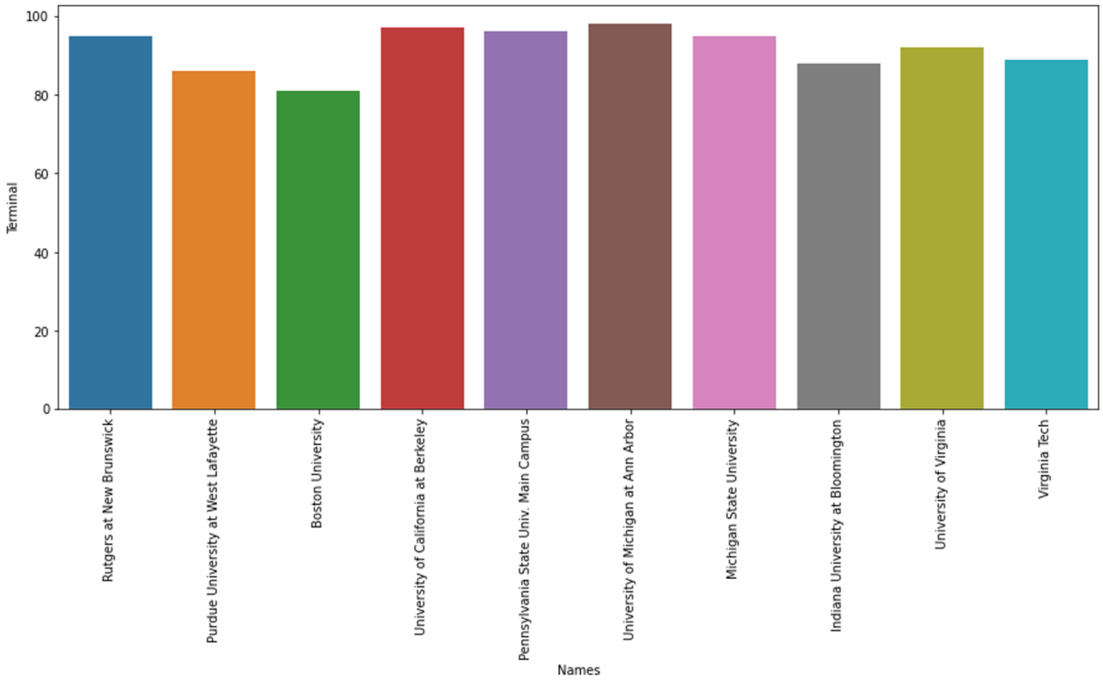

Fig 8.21 – Bivariate Percentage of faculties with Terminal degree.

All most all the universities contains equal number of faculties with Terminal degree, the data seems to be normally distributed.

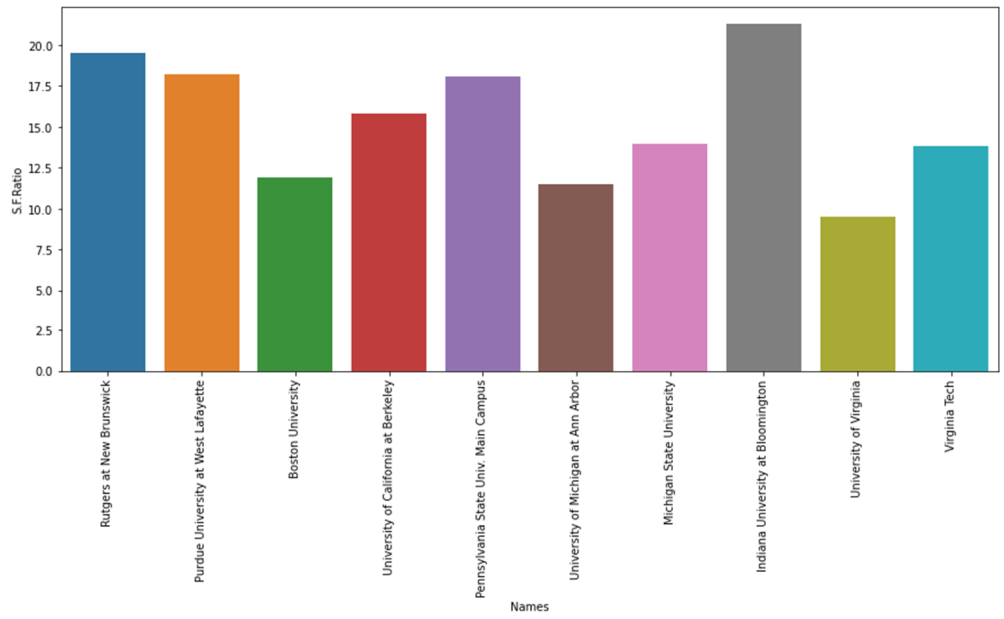

Fig 8.22 – Bivariate Student/faculty ratio

The student faculty ratio is more in Indiana University Bloomington.

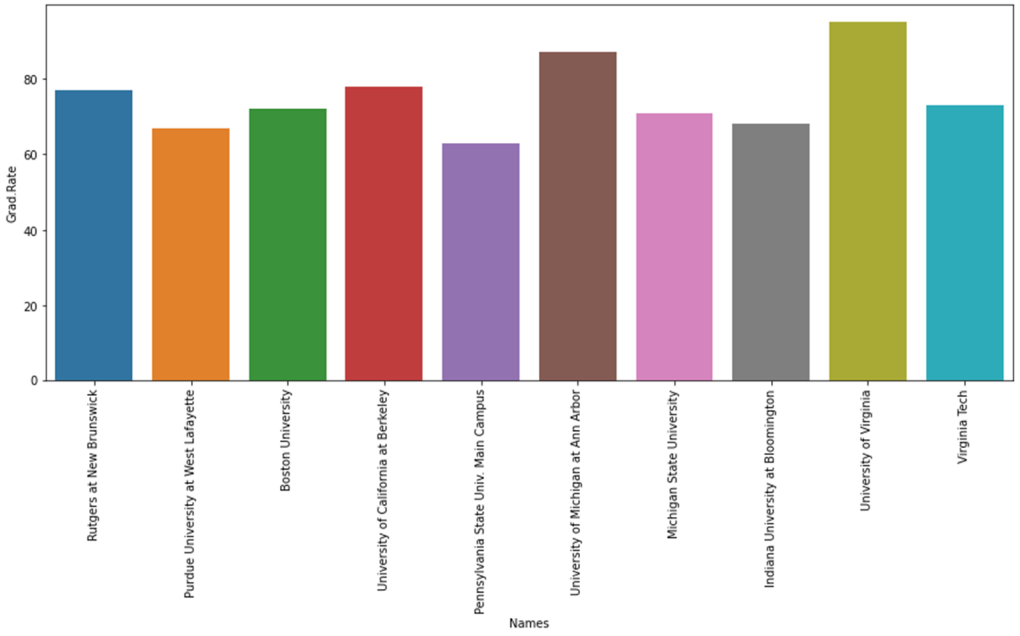

Fig 8.23 – Bivariate Graduation ratio

The graduation ratio is more in University of Virginia compared to other universities.

MULTIVARIATE ANALYSIS

Now let us consider the multivariate analysis among the variables in the dataset. For multivariate analysis we will consider pair plot and heat map in this case.

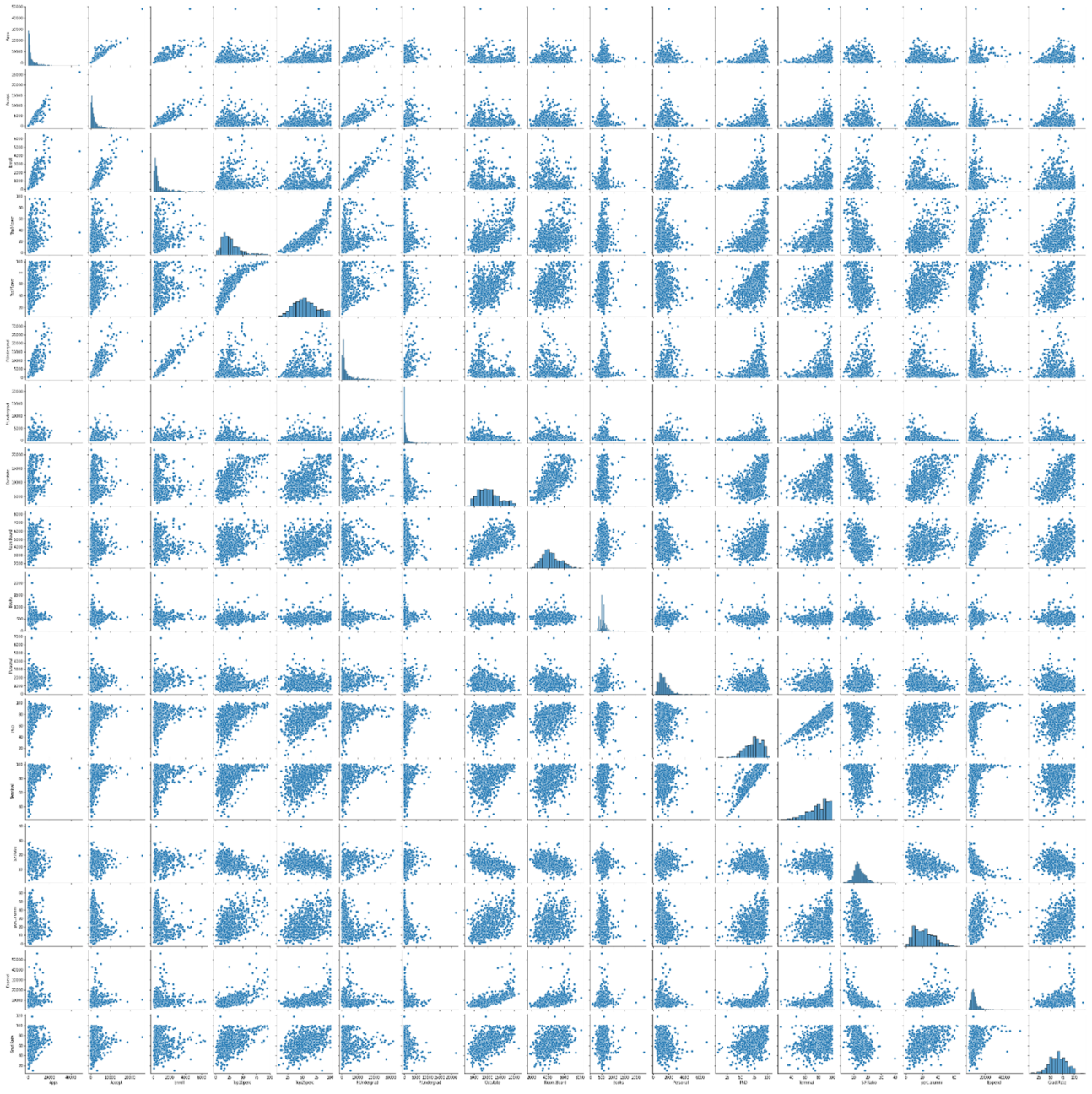

Fig 8.24 – Multivariate pair plot

Pair plot shows the correlation between the numeric variables in the data set. While checking the pair plot, scatter plot show thicker for most of the variables this shows the relation between these variables. Let us check the Heat map also for more information.

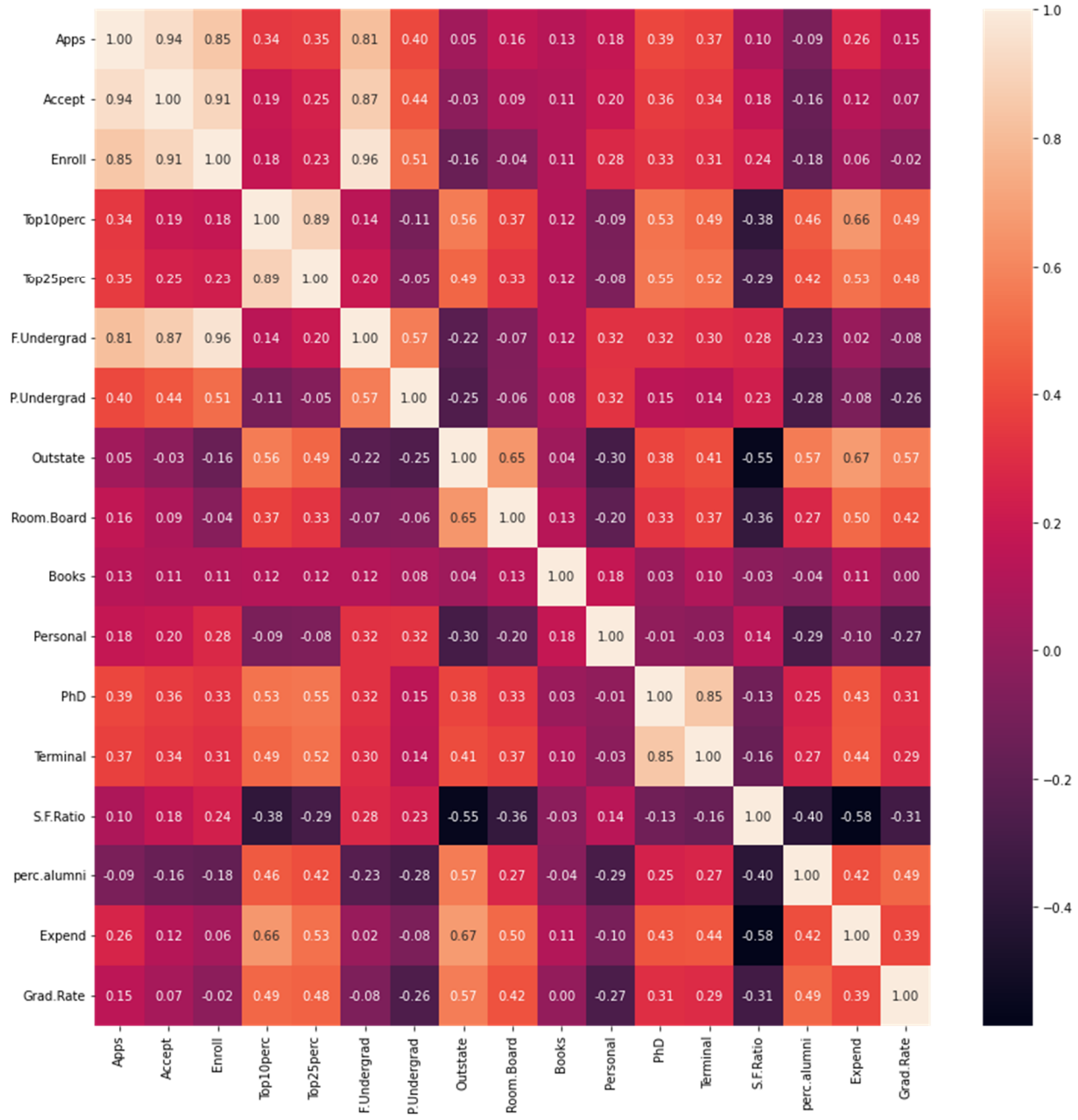

Fig 8.25 – Multivariate Heat map

From the heat map we can conclude that there is good correlation between the full time graduate and the application received, application accepted and the total students enrolled.

INSIGHTS FROM EDA

From the given dataset after exploratory data analysis we can conclude that the correlation between the variables in the data set is very low. Most of the variables in

`

the dataset is left skewed and some variable show almost uniform distribution. All the variables in the dataset contains outliers.

`

PROBLEM 3 – SUMMARY

The Indian Census has the reputation of being one of the best in the world. The first Census in India was conducted in the year 1872. This was conducted at different points of time in different parts of the country. In 1881 a Census was taken for the entire country simultaneously. Since then, Census has been conducted every ten years, without a break. Thus, the Census of India 2011 was the fifteenth in this unbroken series since 1872, the seventh after independence and the second census of the third millennium and twenty first century. The census has been uninterruptedly continued despite of several adversities like wars, epidemics, natural calamities, political unrest, etc. The Census of India is conducted under the provisions of the Census Act 1948 and the Census Rules, 1990. The Primary Census Abstract which is important publication of 2011 Census gives basic information on Area, Total Number of Households, Total Population, Scheduled Castes, Scheduled Tribes Population, Population in the age group 0-6, Literates, Main Workers and Marginal Workers classified by the four broad industrial categories, namely, (i) Cultivators, (ii) Agricultural Laborers, (iii) Household Industry Workers, and (iv) Other Workers and also Non-Workers. The characteristics of the Total Population include Scheduled Castes, Scheduled Tribes, Institutional and Houseless Population and are presented by sex and rural-urban residence. Census 2011 covered 35 States/Union Territories, 640 districts, 5,924 sub-districts, 7,935 Towns and 6, 40,867 Villages.

The data collected has so many variables thus making it difficult to find useful details without using Data Science Techniques. You are tasked to perform detailed EDA and identify Optimum Principal Components that explains the most variance in data.

INTRODUCTION

The data collected is a primary census abstract for female headed households excluding institutional households (India & States/UTs - District Level), Scheduled tribes – 2011. The purpose of this exercise is to perform detailed EDA and identify Optimum Principal Components that explains the most variance in data.

DATA DESCRIPTION

1. State: State Code
2. District: District Code
3. Name: Name

`

4. TRU1: Area Name
5. No_HH: No of Household
6. TOT_M: Total population Male
7. TOT_F: Total population Female
8. M_06: Population in the age group 0-6 Male
9. F_06: Population in the age group 0-6 Female
10. M_SC: Scheduled Castes population Male
11. F_SC: Scheduled Castes population Female
12. M_ST: Scheduled Tribes population Male
13. F_ST: Scheduled Tribes population Female
14. M_LIT: Literates population Male
15. F_LIT: Literates population Female
16. M_ILL: Illiterate Male
17. F_ILL: Illiterate Female
18. TOT_WORK_M: Total Worker Population Male
19. TOT_WORK_F: Total Worker Population Female
20. MAINWORK_M: Main Working Population Male
21. MAINWORK_F: Main Working Population Female
22. MAIN_CL_M: Main Cultivator Population Male
23. MAIN_CL_F: Main Cultivator Population Female
24. MAIN_AL_M: Main Agricultural Laborers Population Male
25. MAIN_AL_F: Main Agricultural Labourers Population Female
26. MAIN_HH_M: Main Household Industries Population Male
27. MAIN_HH_F: Main Household Industries Population Female
28. MAIN_OT_M: Main Other Workers Population Male
29. MAIN_OT_F: Main Other Workers Population Female
30. MARGWORK_M: Marginal Worker Population Male
31. MARGWORK_F: Marginal Worker Population Female
32. MARG_CL_M: Marginal Cultivator Population Male
33. MARG_CL_F: Marginal Cultivator Population Female
34. MARG_AL_M: Marginal Agriculture Labourers Population Male
35. MARG_AL_F: Marginal Agriculture Labourers Population Female
36. MARG_HH_M: Marginal Household Industries Population Male
37. MARG_HH_F: Marginal Household Industries Population Female
38. MARG_OT_M: Marginal Other Workers Population Male

39. MARG_OT_F: Marginal Other Workers Population Female
40. MARGWORK_3_6_M: Marginal Worker Population 3-6 Male
41. MARGWORK_3_6_F: Marginal Worker Population 3-6 Female
42. MARG_CL_3_6_M: Marginal Cultivator Population 3-6 Male
43. MARG_CL_3_6_F: Marginal Cultivator Population 3-6 Female
44. MARG_AL_3_6_M: Marginal Agriculture Labourers Population 3-6 Male
45. MARG_AL_3_6_F: Marginal Agriculture Labourers Population 3-6 Female
46. MARG_HH_3_6_M: Marginal Household Industries Population 3-6 Male
47. MARG_HH_3_6_F: Marginal Household Industries Population 3-6 Female
48. MARG_OT_3_6_M: Marginal Other Workers Population Person 3-6 Male
49. MARG_OT_3_6_F: Marginal Other Workers Population Person 3-6 Female
50. MARGWORK_0_3_M: Marginal Worker Population 0-3 Male
51. MARGWORK_0_3_F: Marginal Worker Population 0-3 Female
52. MARG_CL_0_3_M: Marginal Cultivator Population 0-3 Male
53. MARG_CL_0_3_F: Marginal Cultivator Population 0-3 Female
54. MARG_AL_0_3_M: Marginal Agriculture Labourers Population 0-3 Male
55. MARG_AL_0_3_F: Marginal Agriculture Labourers Population 0-3 Female
56. MARG_HH_0_3_M: Marginal Household Industries Population 0-3 Male
57. MARG_HH_0_3_F: Marginal Household Industries Population 0-3 Female
58. MARG_OT_0_3_M: Marginal Other Workers Population 0-3 Male
59. MARG_OT_0_3_F: Marginal Other Workers Population 0-3 Female
60. NON_WORK_M: Non Working Population Male
61. NON_WORK_F: Non Working Population Female

SAMPLE OF THE DATASET

	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	F_LIT	M_ILL	F_ILL	TOT_WORK_M
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	0	1999	2598	13381	11364	10007	18432	6723
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	6	427	517	10513	7891	9072	15211	6982
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	6	5806	9723	4534	5840	2012	5124	2775
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	0	2666	3968	1842	1962	942	2244	1002
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	33	7670	10843	13243	13477	7348	16504	5717

`

TOT_WORK_F	MAINWORK_M	MAINWORK_F	MAIN_CL_M	MAIN_CL_F	MAIN_AL_M	MAIN_AL_F	MAIN_HH_M	MAIN_HH_F	MAIN_OT_M	MAIN_OT_F
3752	2763	1275	486	235	407	143	78	86	1792	811
4200	4628	1733	1098	357	442	108	538	343	2550	925
4800	1940	2923	519	1205	36	71	19	55	1366	1592
1118	491	408	35	102	8	24	9	6	439	276
7692	2523	2267	743	766	254	237	35	64	1491	1200

MARGWORK_M	MARGWORK_F	MARG_CL_M	MARG_CL_F	MARG_AL_M	MARG_AL_F	MARG_HH_M	MARG_HH_F	MARG_OT_M	MARG_OT_F
3960	2477	619	580	2052	641	142	244	1147	1012
2354	2467	384	661	915	547	369	627	686	632
835	1877	360	1250	44	157	15	32	416	438
511	710	135	286	63	176	10	43	303	205
3194	5425	1327	2462	1037	1069	62	319	768	1575

MARGWORK_3_6_M	MARGWORK_3_6_F	MARG_CL_3_6_M	MARG_CL_3_6_F	MARG_AL_3_6_M	MARG_AL_3_6_F	MARG_HH_3_6_M	MARG_HH_3_6_F
16665	26044	2810	1728	439	343	1372	389
12603	18902	1829	1752	261	432	729	399
3771	6164	721	1689	316	1161	41	123
1782	3088	317	463	74	158	50	126
14874	22289	2320	3497	862	1419	832	767

MARG_OT_3_6_M	MARG_OT_3_6_F	MARGWORK_0_3_M	MARGWORK_0_3_F	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F
110	198	889	798	1150	749	180	237
293	449	546	472	525	715	123	229
15	28	349	377	114	188	44	89
6	33	187	146	194	247	61	128
38	214	588	1097	874	1928	465	1043

MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
680	252	32	46	258	214
186	148	76	178	140	160
3	34	0	4	67	61
13	50	4	10	116	59
205	302	24	105	180	478

Table 5 – Sample dataset

`

Problem 3 – PCA

PCA FH (FT): Primary census abstract for female headed households excluding institutional households (India & States/UTs - District Level), Scheduled tribes - 2011

PCA for Female Headed Household Excluding Institutional Household

The Indian Census has the reputation of being one of the best in the world. The first Census in India was conducted in the year 1872. This was conducted at different points of time in different parts of the country. In 1881 a Census was taken for the entire country simultaneously. Since then, Census has been conducted every ten years, without a break. Thus, the Census of India 2011 was the fifteenth in this unbroken series since 1872, the seventh after independence and the second census of the third millennium and twenty first century. The census has been uninterruptedly continued despite of several adversities like wars, epidemics, natural calamities, political unrest, etc. The Census of India is conducted under the provisions of the Census Act 1948 and the Census Rules, 1990. The Primary Census Abstract which is important publication of 2011 Census gives basic information on Area, Total Number of Households, Total Population, Scheduled Castes, Scheduled Tribes Population, Population in the age group 0-6, Literates, Main Workers and Marginal Workers classified by the four broad industrial categories, namely, (i) Cultivators, (ii) Agricultural Laborers, (iii) Household Industry Workers, and (iv) Other Workers and also Non-Workers. The characteristics of the Total Population include Scheduled Castes, Scheduled Tribes, Institutional and Houseless Population and are presented by sex and rural-urban residence. Census 2011 covered 35 States/Union Territories, 640 districts, 5,924 sub-districts, 7,935 Towns and 6,40,867 Villages.

The data collected has so many variables thus making it difficult to find useful details without using Data Science Techniques. You are tasked to perform detailed EDA and identify Optimum Principal Components that explains the most variance in data. Use Sklearn only.

9. **Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc**

The data contains Primary census abstract for female headed households excluding institutional households, the sample of the data set shown below,

`

	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	F_LIT	M_ILL	F_ILL	TOT_WOR
0	1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	0	1999	2598	13381	11364	10007	18432	
1	1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	6	427	517	10513	7891	9072	15211	
2	1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	6	5806	9723	4534	5840	2012	5124	
3	1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	0	2666	3968	1842	1962	942	2244	
4	1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	33	7670	10843	13243	13477	7348	16504	

TOT_WORK_F	MAINWORK_M	MAINWORK_F	MAIN_CL_M	MAIN_CL_F	MAIN_AL_M	MAIN_AL_F	MAIN_HH_M	MAIN_HH_F	MAIN_OT_M	MAIN_OT_F
3752	2763	1275	486	235	407	143	78	86	1792	811
4200	4628	1733	1098	357	442	108	538	343	2550	925
4800	1940	2923	519	1205	36	71	19	55	1366	1592
1118	491	408	35	102	8	24	9	6	439	276
7692	2523	2267	743	766	254	237	35	64	1491	1200

MARGWORK_M	MARGWORK_F	MARG_CL_M	MARG_CL_F	MARG_AL_M	MARG_AL_F	MARG_HH_M	MARG_HH_F	MARG_OT_M	MARG_OT_F
3960	2477	619	580	2052	641	142	244	1147	1012
2354	2467	384	661	915	547	369	627	686	632
835	1877	360	1250	44	157	15	32	416	438
511	710	135	286	63	176	10	43	303	205
3194	5425	1327	2462	1037	1069	62	319	768	1575

MARGWORK_3_6_M	MARGWORK_3_6_F	MARG_CL_3_6_M	MARG_CL_3_6_F	MARG_AL_3_6_M	MARG_AL_3_6_F	MARG_HH_3_6_M	MARG_HH_3_6_F
16665	26044	2810	1728	439	343	1372	389
12603	18902	1829	1752	261	432	729	399
3771	6164	721	1689	316	1161	41	123
1782	3088	317	463	74	158	50	126
14874	22289	2320	3497	862	1419	832	767

MARG_OT_3_6_M	MARG_OT_3_6_F	MARGWORK_0_3_M	MARGWORK_0_3_F	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F
110	198	889	798	1150	749	180	237
293	449	546	472	525	715	123	229
15	28	349	377	114	188	44	89
6	33	187	146	194	247	61	128
38	214	588	1097	874	1928	465	1043

MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
680	252	32	46	258	214
186	148	76	178	140	160
3	34	0	4	67	61
13	50	4	10	116	59
205	302	24	105	180	478

Table 9.1 – Head of the Dataset

EXPLORATORY DATA ANALYSIS

The basic information of the data is shown in below table

No.	Column	Non null Count	Data type
1	State Code	640 non – null	Int 64
2	Dist.Code	640 non – null	Int 64
3	State	640 non – null	Object
4	Area Name	640 non – null	Object
5	No_HH	640 non – null	Int 64
6	TOT_M	640 non – null	Int 64
7	TOT_F	640 non – null	Int 64
8	M_06	640 non – null	Int 64
9	F_06	640 non – null	Int 64
10	M_SC	640 non – null	Int 64

`

11	F_SC	640 non – null	Int 64
12	M_ST	640 non – null	Int 64
13	F_ST	640 non – null	Int 64
14	M_LIT	640 non – null	Int 64
15	F_LIT	640 non – null	Int 64
16	M_ILL	640 non – null	Int 64
17	F_ILL	640 non – null	Int 64
18	TOT_WORK_M	640 non – null	Int 64
19	TOT_WORK_F	640 non – null	Int 64
20	MAINWORK_M	640 non – null	Int 64
21	MAINWORK_F	640 non – null	Int 64
22	MAIN_CL_M	640 non – null	Int 64
23	MAIN_CL_F	640 non – null	Int 64
24	MAIN_AL_M	640 non – null	Int 64
25	MAIN_AL_F	640 non – null	Int 64
26	MAIN_HH_M	640 non – null	Int 64
27	MAIN_HH_F	640 non – null	Int 64
28	MAIN_OT_M	640 non – null	Int 64
29	MAIN_OT_F	640 non – null	Int 64
30	MARGWORK_M	640 non – null	Int 64
31	MARGWORK_F	640 non – null	Int 64
32	MARG_CL_M	640 non – null	Int 64
33	MARG_CL_F	640 non – null	Int 64
34	MARG_AL_M	640 non – null	Int 64
35	MARG_AL_F	640 non – null	Int 64
36	MARG_HH_M	640 non – null	Int 64
37	MARG_HH_F	640 non – null	Int 64
38	MARG_OT_M	640 non – null	Int 64
39	MARG_OT_F	640 non – null	Int 64
40	MARGWORK_3_6_M	640 non – null	Int 64
41	MARGWORK_3_6_F	640 non – null	Int 64
42	MARG_CL_3_6_M	640 non – null	Int 64
43	MARG_CL_3_6_F	640 non – null	Int 64
44	MARG_AL_3_6_M	640 non – null	Int 64
45	MARG_AL_3_6_F	640 non – null	Int 64
46	MARG_HH_3_6_M	640 non – null	Int 64
47	MARG_HH_3_6_F	640 non – null	Int 64
48	MARG_OT_3_6_M	640 non – null	Int 64
49	MARG_OT_3_6_F	640 non – null	Int 64

`

50	MARGWORK_0_3_M	640 non – null	Int 64
51	MARGWORK_0_3_F	640 non – null	Int 64
52	MARG_CL_0_3_M	640 non – null	Int 64
53	MARG_CL_0_3_F	640 non – null	Int 64
54	MARG_AL_0_3_M	640 non – null	Int 64
55	MARG_AL_0_3_F	640 non – null	Int 64
56	MARG_HH_0_3_M	640 non – null	Int 64
57	MARG_HH_0_3_F	640 non – null	Int 64
58	MARG_OT_0_3_M	640 non – null	Int 64
59	MARG_OT_0_3_F	640 non – null	Int 64
60	NON_WORK_M	640 non – null	Int 64
61	NON_WORK_F	640 non – null	Int 64

Table 9.2 – Exploratory Data Analysis

64 columns and total 640 entries

dtypes: int64(59), object(2)

DESCRIPTIVE DATA ANALYSIS

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
State Code	640.0	NaN	NaN	NaN	17.114062	9.426486	1.0	9.0	18.0	24.0	35.0
Dist.Code	640.0	NaN	NaN	NaN	320.5	184.896367	1.0	160.75	320.5	480.25	640.0
State	640	35	Uttar Pradesh	71	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Area Name	640	635	Raigarh	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_HH	640.0	NaN	NaN	NaN	51222.871875	48135.405475	350.0	19484.0	35837.0	68892.0	310450.0
TOT_M	640.0	NaN	NaN	NaN	79940.576563	73384.511114	391.0	30228.0	58339.0	107918.5	485417.0
TOT_F	640.0	NaN	NaN	NaN	122372.084375	113600.717282	698.0	46517.75	87724.5	164251.75	750392.0
M_06	640.0	NaN	NaN	NaN	12309.098438	11500.906881	56.0	4733.75	9159.0	16520.25	96223.0
F_06	640.0	NaN	NaN	NaN	11942.3	11326.294567	56.0	4672.25	8663.0	15902.25	95129.0
M_SC	640.0	NaN	NaN	NaN	13820.946875	14426.37313	0.0	3466.25	9591.5	19429.75	103307.0
F_SC	640.0	NaN	NaN	NaN	20778.392188	21727.887713	0.0	5603.25	13709.0	29180.0	156429.0
M_ST	640.0	NaN	NaN	NaN	6191.807813	9912.668948	0.0	293.75	2333.5	7658.0	96785.0
F_ST	640.0	NaN	NaN	NaN	10155.640625	15875.701488	0.0	429.5	3834.5	12480.25	130119.0
M_LIT	640.0	NaN	NaN	NaN	57967.979688	55910.282466	286.0	21298.0	42693.5	77989.5	403261.0
F_LIT	640.0	NaN	NaN	NaN	66359.565625	75037.860207	371.0	20932.0	43796.5	84799.75	571140.0
M_ILL	640.0	NaN	NaN	NaN	21972.596875	19825.605268	105.0	8590.0	15767.5	29512.5	105961.0
F_ILL	640.0	NaN	NaN	NaN	56012.51875	47116.693769	327.0	22367.0	42386.0	78471.0	254160.0
TOT_WORK_M	640.0	NaN	NaN	NaN	37992.407813	36419.537491	100.0	13753.5	27936.5	50226.75	269422.0
TOT_WORK_F	640.0	NaN	NaN	NaN	41295.760938	37192.360943	357.0	16097.75	30588.5	53234.25	257848.0
MAINWORK_M	640.0	NaN	NaN	NaN	30204.446875	31480.91568	65.0	9787.0	21250.5	40119.0	247911.0

MAINWORK_F	640.0	NaN	NaN	NaN	28198.846875	29998.262689	240.0	9502.25	18484.0	35063.25	226166.0
MAIN_CL_M	640.0	NaN	NaN	NaN	5424.342188	4739.161969	0.0	2023.5	4160.5	7695.0	29113.0
MAIN_CL_F	640.0	NaN	NaN	NaN	5486.042188	5326.362728	0.0	1920.25	3908.5	7286.25	36193.0
MAIN_AL_M	640.0	NaN	NaN	NaN	5849.109375	6399.507966	0.0	1070.25	3936.5	8067.25	40843.0
MAIN_AL_F	640.0	NaN	NaN	NaN	8925.995312	12864.287584	0.0	1408.75	3933.5	10617.5	87945.0
MAIN_HH_M	640.0	NaN	NaN	NaN	883.89375	1278.642345	0.0	187.5	498.5	1099.25	16429.0
MAIN_HH_F	640.0	NaN	NaN	NaN	1380.773438	3179.414449	0.0	248.75	540.5	1435.75	45979.0
MAIN_OT_M	640.0	NaN	NaN	NaN	18047.101562	26068.480886	36.0	3997.5	9598.0	21249.5	240855.0
MAIN_OT_F	640.0	NaN	NaN	NaN	12406.035938	18972.202369	153.0	3142.5	6380.5	14368.25	209355.0
MARGWORK_M	640.0	NaN	NaN	NaN	7787.960938	7410.791691	35.0	2937.5	5627.0	9800.25	47553.0
MARGWORK_F	640.0	NaN	NaN	NaN	13096.914062	10996.474528	117.0	5424.5	10175.0	18879.25	66915.0
MARG_CL_M	640.0	NaN	NaN	NaN	1040.7375	1311.546847	0.0	311.75	606.5	1281.0	13201.0
MARG_CL_F	640.0	NaN	NaN	NaN	2307.682813	3564.626095	0.0	630.25	1226.0	2659.25	44324.0
MARG_AL_M	640.0	NaN	NaN	NaN	3304.326562	3781.555707	0.0	873.5	2062.0	4300.75	23719.0
MARG_AL_F	640.0	NaN	NaN	NaN	6463.28125	6773.876298	0.0	1402.5	4020.5	9089.25	45301.0
MARG_HH_M	640.0	NaN	NaN	NaN	316.742188	462.661891	0.0	71.75	166.0	356.5	4298.0
MARG_HH_F	640.0	NaN	NaN	NaN	786.626562	1198.718213	0.0	171.75	429.0	962.5	15448.0
MARG_OT_M	640.0	NaN	NaN	NaN	3126.154687	3609.391821	7.0	935.5	2036.0	3985.25	24728.0
MARG_OT_F	640.0	NaN	NaN	NaN	3539.323438	4115.191314	19.0	1071.75	2349.5	4400.5	36377.0
MARGWORK_3_6_M	640.0	NaN	NaN	NaN	41948.16875	39045.316918	291.0	16208.25	30315.0	57218.75	300937.0
MARGWORK_3_6_F	640.0	NaN	NaN	NaN	81076.323438	82970.406216	341.0	26619.5	56793.0	107924.0	676450.0
MARG_CL_3_6_M	640.0	NaN	NaN	NaN	6394.9875	6019.806644	27.0	2372.0	4630.0	8167.0	39106.0
MARG_CL_3_6_F	640.0	NaN	NaN	NaN	10339.864063	8467.473429	85.0	4351.5	8295.0	15102.0	50065.0
MARG_AL_3_6_M	640.0	NaN	NaN	NaN	789.848438	905.639279	0.0	235.5	480.5	986.0	7426.0
MARG_AL_3_6_F	640.0	NaN	NaN	NaN	1749.584375	2496.541514	0.0	497.25	985.5	2059.0	27171.0
MARG_HH_3_6_M	640.0	NaN	NaN	NaN	2743.635938	3059.586387	0.0	718.75	1714.5	3702.25	19343.0
MARG_HH_3_6_F	640.0	NaN	NaN	NaN	5169.85	5335.64096	0.0	1113.75	3294.0	7502.25	36253.0
MARG_OT_3_6_M	640.0	NaN	NaN	NaN	245.3625	358.728567	0.0	58.0	129.5	276.0	3535.0
MARG_OT_3_6_F	640.0	NaN	NaN	NaN	585.884375	900.025817	0.0	127.75	320.5	719.25	12094.0
MARGWORK_0_3_M	640.0	NaN	NaN	NaN	2616.140625	3036.964381	7.0	755.0	1681.5	3320.25	20648.0
MARGWORK_0_3_F	640.0	NaN	NaN	NaN	2834.545312	3327.836932	14.0	833.5	1834.5	3610.5	25844.0
MARG_CL_0_3_M	640.0	NaN	NaN	NaN	1392.973438	1489.707052	4.0	489.5	949.0	1714.0	9875.0
MARG_CL_0_3_F	640.0	NaN	NaN	NaN	2757.05	2788.776676	30.0	957.25	1928.0	3599.75	21611.0
MARG_AL_0_3_M	640.0	NaN	NaN	NaN	250.889062	453.336594	0.0	47.0	114.5	270.75	5775.0
MARG_AL_0_3_F	640.0	NaN	NaN	NaN	558.098438	1117.642748	0.0	109.0	247.5	568.75	17153.0
MARG_HH_0_3_M	640.0	NaN	NaN	NaN	560.690625	762.578991	0.0	136.5	308.0	642.0	6116.0
MARG_HH_0_3_F	640.0	NaN	NaN	NaN	1293.43125	1585.377936	0.0	298.0	717.0	1710.75	13714.0
MARG_OT_0_3_M	640.0	NaN	NaN	NaN	71.379688	107.897627	0.0	14.0	35.0	79.0	895.0
MARG_OT_0_3_F	640.0	NaN	NaN	NaN	200.742188	309.740854	0.0	43.0	113.0	240.0	3354.0
NON_WORK_M	640.0	NaN	NaN	NaN	510.014063	610.603187	0.0	161.0	326.0	604.5	6456.0
NON_WORK_F	640.0	NaN	NaN	NaN	704.778125	910.209225	5.0	220.5	464.5	853.5	10533.0

Table 9.3 – Descriptive Data Analysis

1. There are total 640 entries in the dataset from 35 different states,
2. The mean of non-working female is more than the mean of non-working male

`

3. The mean of literate female is greater than the mean of literate male.
4. The mean of illiterate female is less than the mean of illiterate male.
5. The mean of total population of male is less than the mean of total population of female.
6. Total worker population of male is less than the total worker population of female

From exploratory analysis data table it is clear that there are 640 entries in the data set and there are 640 non-null values are present in each column therefore there is no null value present in the column.

Using python checked for duplicate values also. There is no duplicate value present in the dataset.

10. **Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio? (Example Questions). Pick 5 variables out of the given 24 variables below for EDA: No_HH, TOT_M, TOT_F, M_06, F_06, M_SC, F_SC, M_ST, F_ST, M_LIT, F_LIT, M_ILL, F_ILL, TOT_WORK_M, TOT_WORK_F, MAINWORK_M, MAINWORK_F, MAIN_CL_M, MAIN_CL_F, MAIN_AL_M, MAIN_AL_F, MAIN_HH_M, MAIN_HH_F, MAIN_OT_M, MAIN_OT_F**

Let us consider No_HH, TOT_M, TOT_F, M_06, F_06 variable with state and perform exploratory data analysis on the data set,

Let us check answers for the following questions,

1. Which state has highest number of males and which one has lowest number of males.

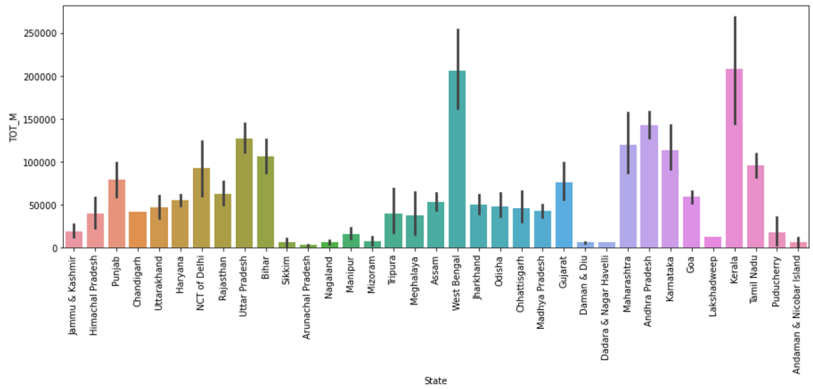

Fig 10.1 – State v/s total male

From the above plot west Bengal and Kerala have more number of male, if we look closer we can conclude that more number of male present in Kerala. In the case of lease number of males Arunachal Pradesh has least number of male.

2. Which state contains more number of households which contains least?

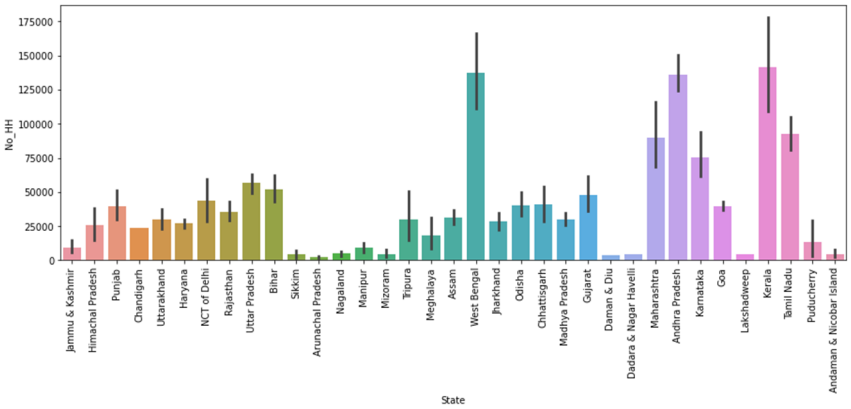

Fig 10.2 – State v/s house hold

From the above plot in the case of house hold also Kerala have more number of households compared to other states and Arunachal Pradesh has least number of household.

3. Which state has more number of females and which one has less?

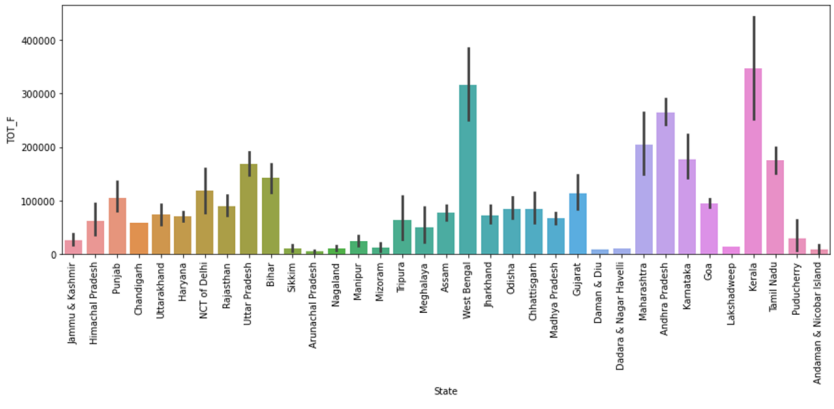

Fig 10.3 – State v/s total female

In the case of female also Kerala has more number of female and Arunachal Pradesh has least number of females.

4. Comment on the correlation between the variables

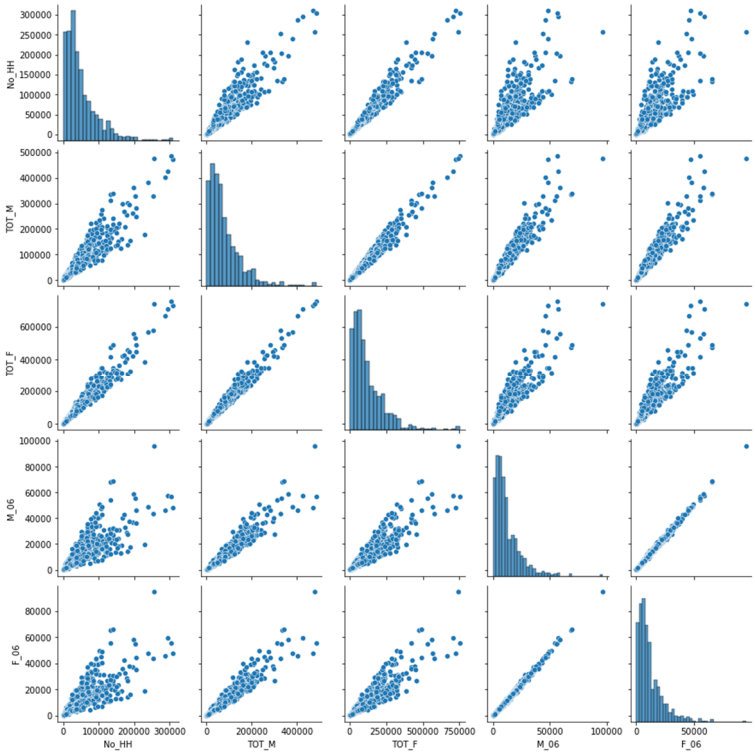

Fig 10.4 – Pair plot

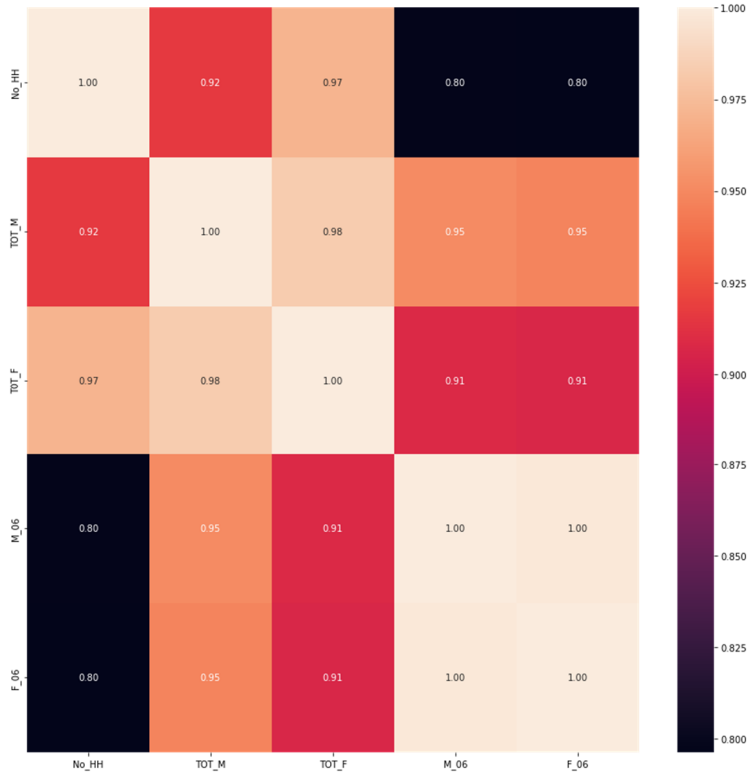

Fig 10.5 – Heat map

From the above heat map and pair plot we can say that all the variables are highly correlated to each other. All the correlation values in heat map is above 0.8, pair plot also shows the same.

5. How the number of households and other variables are related?

Consider the below reg plot between the households and other variables. All the variables shows high correlation to the household. All most all the points are very much closer to each other which shows a good correlation between these variables.

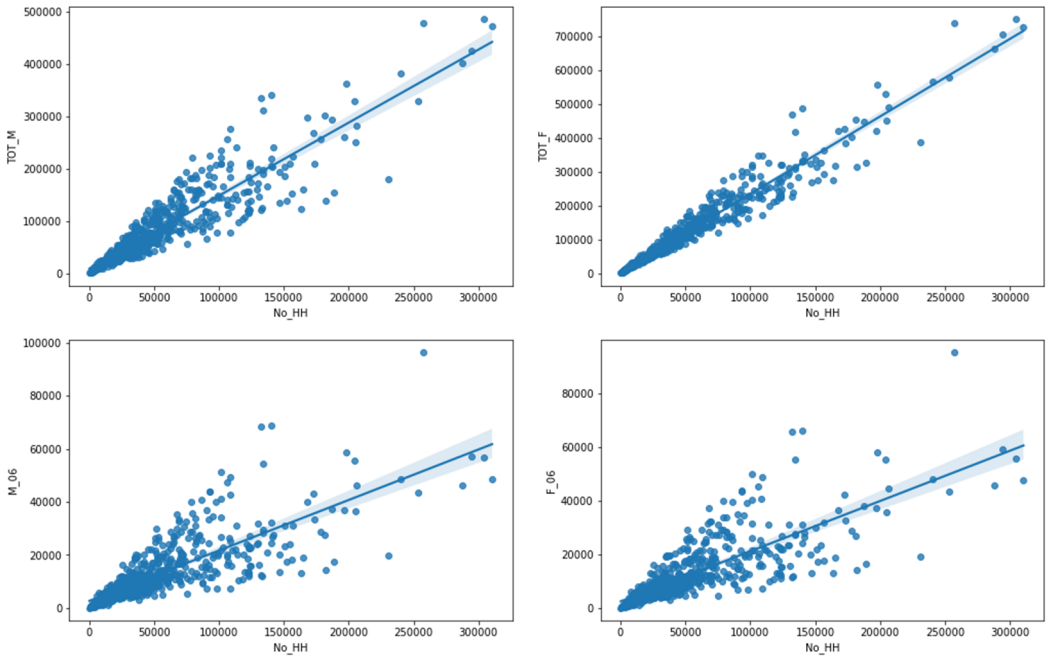

Fig 10.6 – Reg plot

11. We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

It is not acceptable to drop an observation just because it is an outlier. They can be legitimate observations and are sometimes the most interesting ones. It's important to investigate the nature of the outlier before deciding.

1. If it is obvious that the outlier is due to incorrectly entered or measured data, you should drop the outlier:
2. If the outlier does not change the results but does affect assumptions, you may drop the outlier.

3. More commonly, the outlier affects both results and assumptions. In this situation, it is *not* legitimate to simply drop the outlier. You may run the analysis both with and without it, but you should state in at least a footnote the dropping of any such data points and how the results changed.

Here the given data is a Primary census abstract for female headed households, so the probability of occurring a outlier because of an incorrectly entered or measured data is very less and they can be legitimate observations. So it is better not to treat outlier in this case, unless it affect our assumptions.

12. Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.

We can drop the columns state code, district code, state and area before performing scaling on the dataset because it will not affect the output and these are normally object values.

Let us first check the boxplot of all the columns before performing the scaling. Now we have 57 columns, and their box plot shown below,

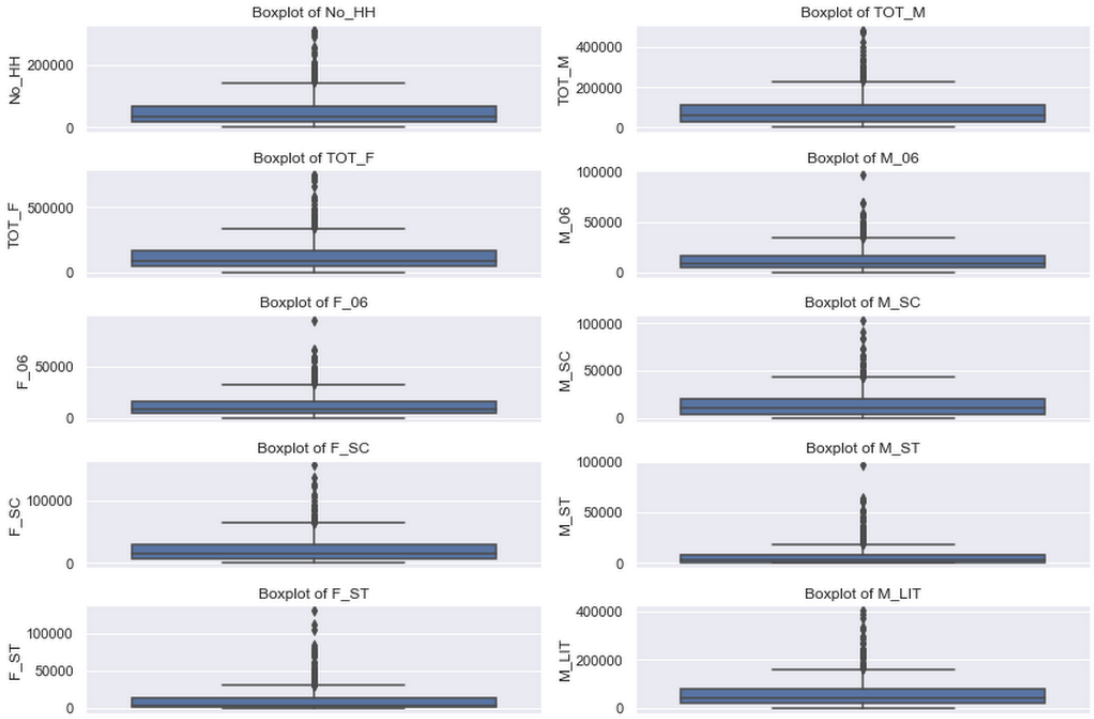

Fig 12.1 – Boxplot before scaling

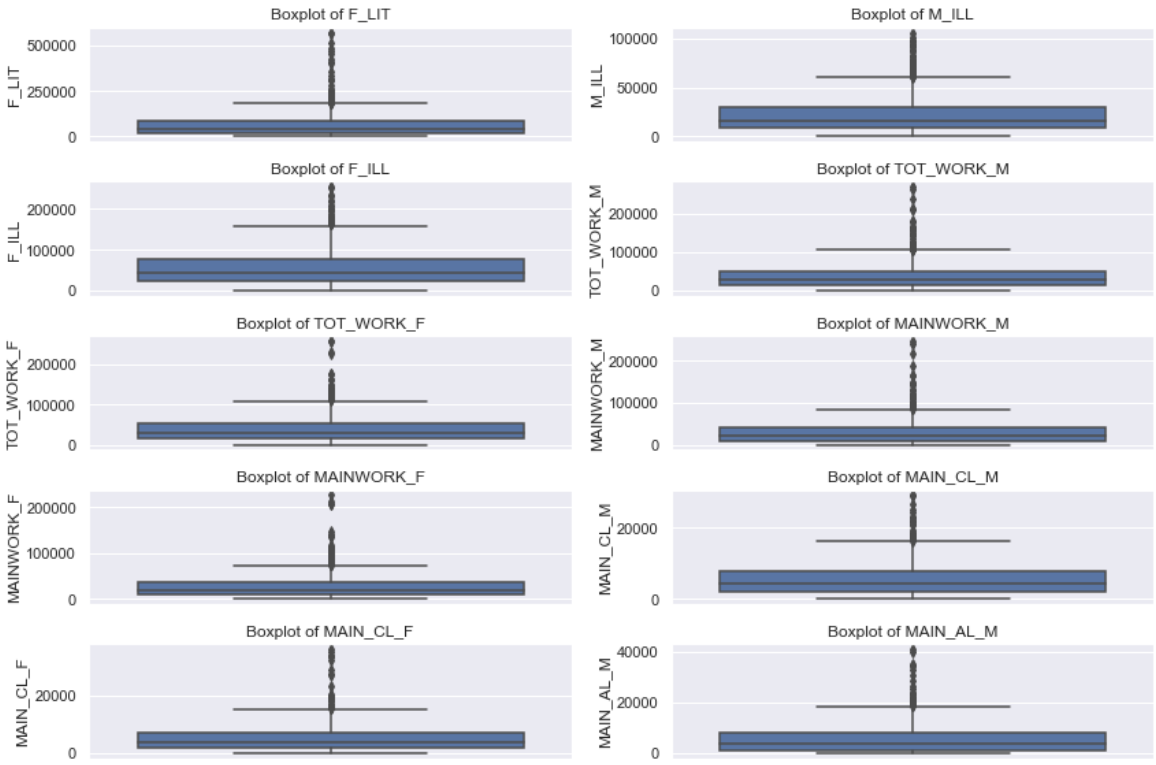

Fig 12.2 – Boxplot before scaling

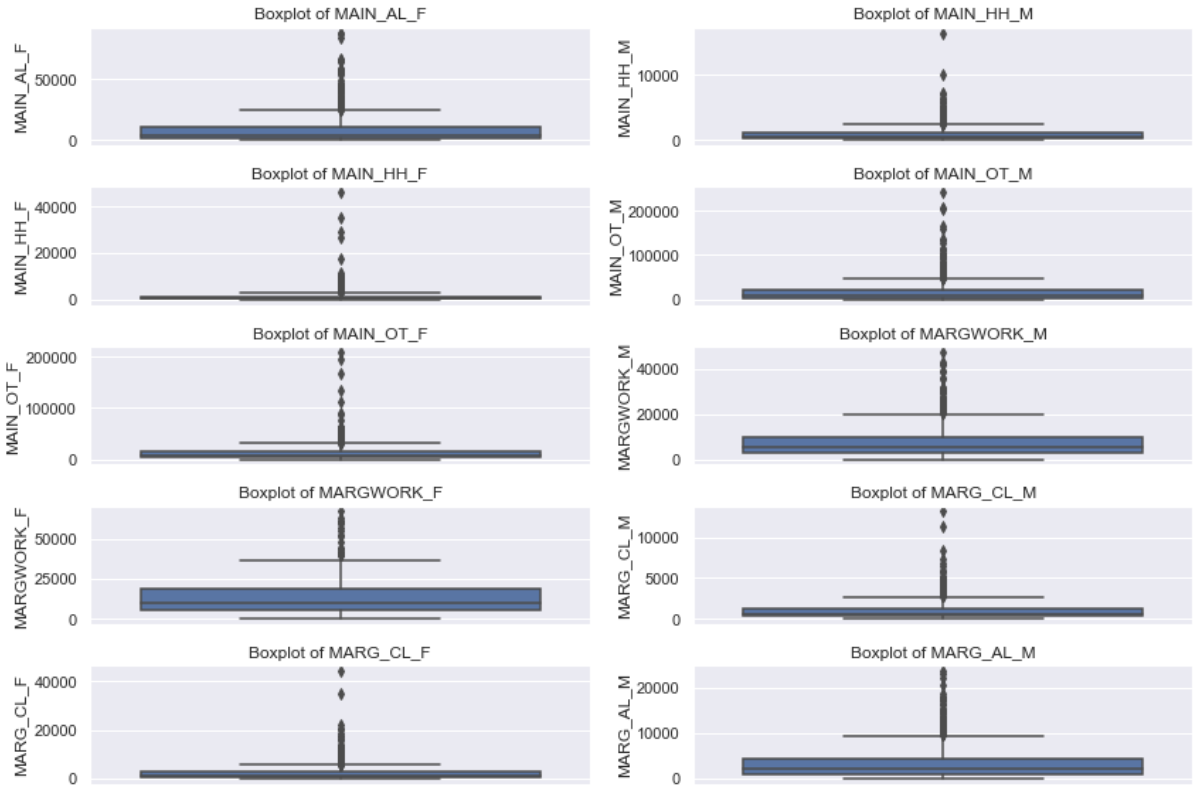

Fig 12.3 – Boxplot before scaling

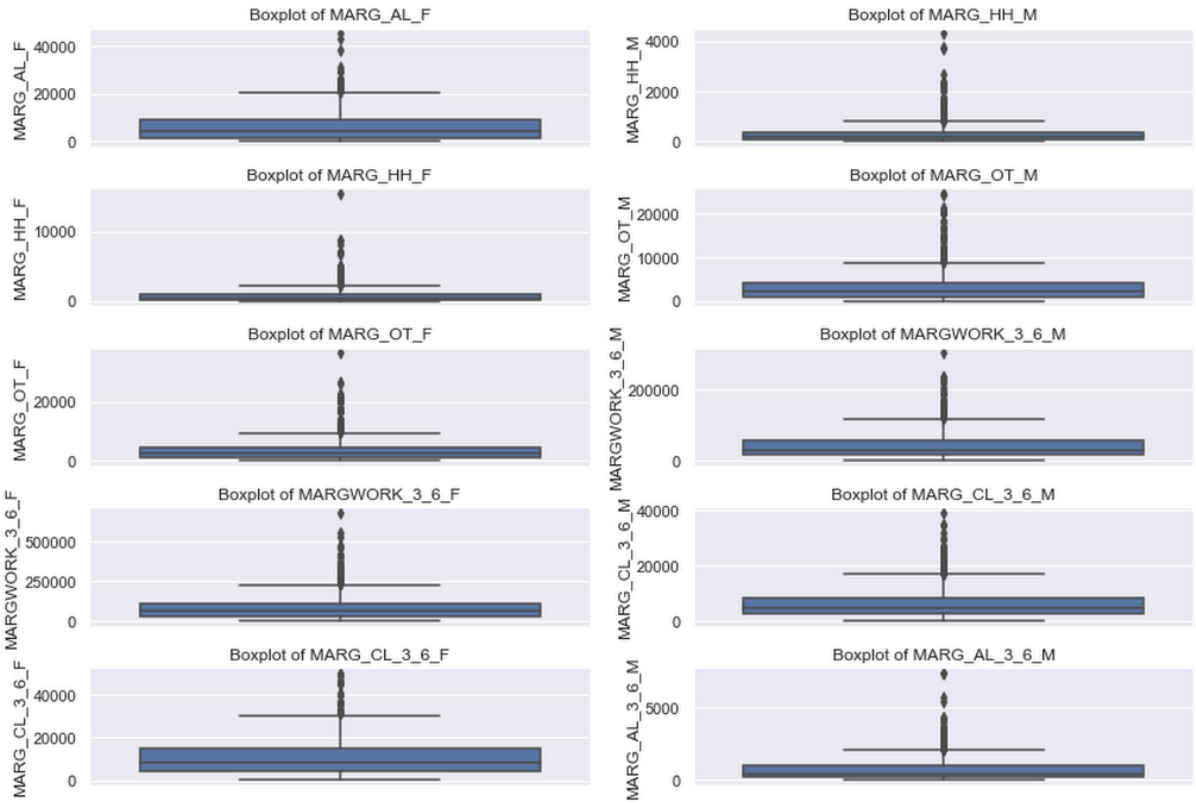

Fig 12.4 – Boxplot before scaling

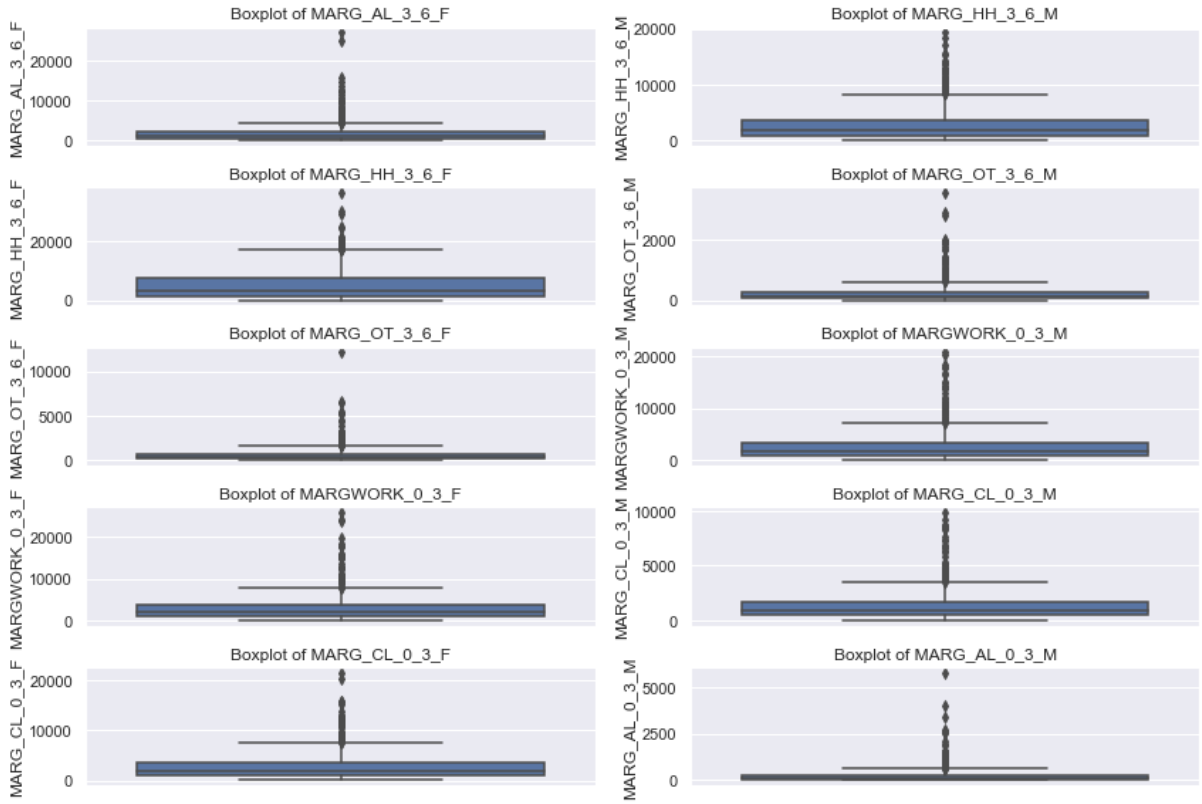

Fig 12.5 – Boxplot before scaling

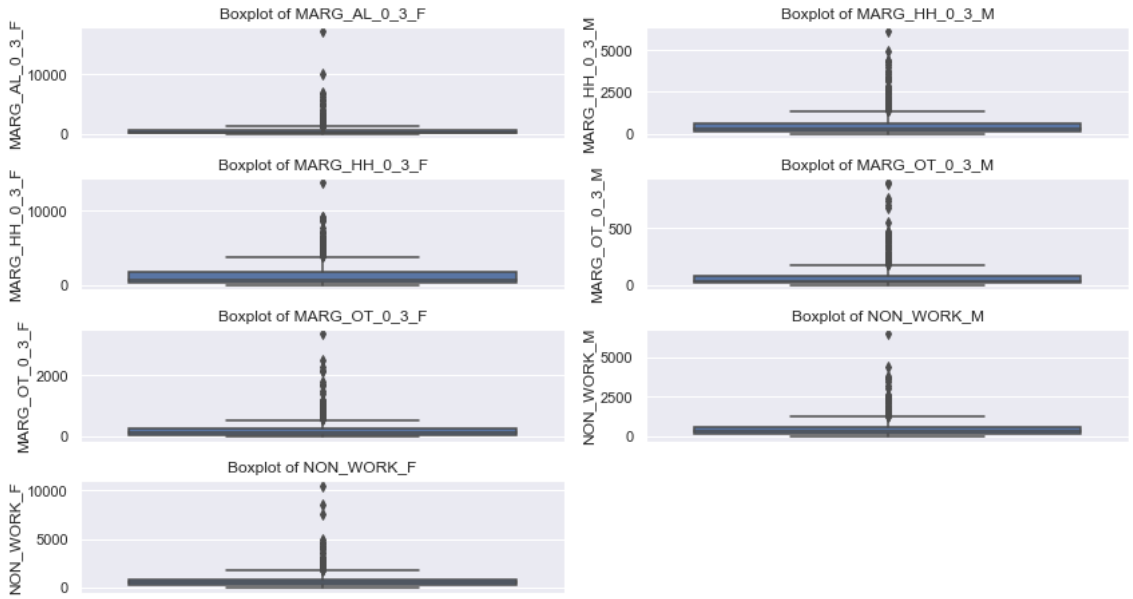

Fig 12.6 – Boxplot before scaling

From the above shown boxplots it is very clear that all the columns have outliers present in them, since we already decided not to treat the outliers no actions need to be taken here. Now let us scale the data set using z score. The scaled data set is shown below,

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	F_LIT	M_ILL	F_ILL	TOT_WORK_M
0	-0.904738	-0.771236	-0.815563	-0.561012	-0.507738	-0.958575	-0.957049	-0.423306	-0.476423	-0.798097	-0.733477	-0.604015	-0.798229	-0.859260
1	-0.935695	-0.823100	-0.874534	-0.681096	-0.725367	-0.958297	-0.956772	-0.582014	-0.607607	-0.849434	-0.779797	-0.651213	-0.866645	-0.852143
2	-0.972412	-1.000919	-0.981466	-0.976956	-0.965262	-0.958575	-0.956772	-0.038951	-0.027273	-0.956457	-0.807151	-1.007596	-1.080898	-0.967749
3	-1.037530	-1.052224	-1.041001	-1.022118	-0.995393	-0.958783	-0.957049	-0.355965	-0.390060	-1.004643	-0.858872	-1.061609	-1.142070	-1.016469
4	-0.822676	-0.809381	-0.813933	-0.622359	-0.649908	-0.957395	-0.955529	0.149238	0.043330	-0.800568	-0.705296	-0.738239	-0.839181	-0.886905

TOT_WORK_F	MAINWORK_M	MAINWORK_F	MAIN_CL_M	MAIN_CL_F	MAIN_AL_M	MAIN_AL_F	MAIN_HH_M	MAIN_HH_F	MAIN_OT_M	MAIN_OT_F
-1.010238	-0.872367	-0.898216	-1.042844	-0.986630	-0.851060	-0.683276	-0.630766	-0.407555	-0.624042	-0.611637
-0.998183	-0.813078	-0.882936	-0.913606	-0.963707	-0.845587	-0.685999	-0.270728	-0.326659	-0.594942	-0.605624
-0.982038	-0.898530	-0.843236	-1.035875	-0.804375	-0.909079	-0.688878	-0.676945	-0.417313	-0.640396	-0.570440
-1.081114	-0.944594	-0.927140	-1.138083	-1.011620	-0.913457	-0.692534	-0.684772	-0.432737	-0.675984	-0.639858
-0.904219	-0.879997	-0.865121	-0.988572	-0.886859	-0.874987	-0.675964	-0.664422	-0.414480	-0.635597	-0.591118

MARGWORK_M	MARGWORK_F	MARG_CL_M	MARG_CL_F	MARG_AL_M	MARG_AL_F	MARG_HH_M	MARG_HH_F	MARG_OT_M	MARG_OT_F
-0.516943	-0.966512	-0.321809	-0.485053	-0.331426	-0.860192	-0.377984	-0.453026	-0.548764	-0.614625
-0.733823	-0.967422	-0.501127	-0.462312	-0.632331	-0.874080	0.113039	-0.133269	-0.676586	-0.707038
-0.938955	-1.021117	-0.519440	-0.296948	-0.862840	-0.931699	-0.652697	-0.630020	-0.751449	-0.754217
-0.982709	-1.127325	-0.691127	-0.567595	-0.857811	-0.928892	-0.663513	-0.620837	-0.782781	-0.810881
-0.620386	-0.698216	0.218434	0.043325	-0.600044	-0.796959	-0.551032	-0.390411	-0.653850	-0.477708

MARGWORK_3_6_M	MARGWORK_3_6_F	MARG_CL_3_6_M	MARG_CL_3_6_F	MARG_AL_3_6_M	MARG_AL_3_6_F	MARG_HH_3_6_M	MARG_HH_3_6_F
-0.648040	-0.663795	-0.595998	-1.017848	-0.387707	-0.563854	-0.448658	-0.896723
-0.752155	-0.749942	-0.759087	-1.015011	-0.584407	-0.528177	-0.658982	-0.894847
-0.978530	-0.903586	-0.943290	-1.022457	-0.523629	-0.235944	-0.884025	-0.946615
-1.029511	-0.940689	-1.010455	-1.167360	-0.791053	-0.638014	-0.881081	-0.946052
-0.693946	-0.709088	-0.677459	-0.808767	0.079732	-0.132521	-0.625291	-0.825823

MARG_OT_3_6_M	MARG_OT_3_6_F	MARGWORK_0_3_M	MARGWORK_0_3_F	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F
-0.377635	-0.431307	-0.569151	-0.612451	-0.163229	-0.720610	-0.156494	-0.287524
0.132899	-0.152208	-0.682181	-0.710490	-0.583103	-0.732811	-0.282327	-0.294688
-0.642666	-0.620339	-0.747099	-0.739059	-0.859212	-0.921931	-0.456727	-0.420050
-0.667774	-0.614779	-0.800484	-0.808528	-0.805468	-0.900758	-0.419198	-0.385127
-0.578501	-0.413516	-0.668341	-0.522533	-0.348645	-0.297513	0.472670	0.434200

MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
0.156577	-0.657412	-0.365258	-0.499977	-0.413053	-0.539614
-0.491731	-0.723062	0.042855	-0.073481	-0.606455	-0.598988
-0.731894	-0.795026	-0.662068	-0.635680	-0.726103	-0.707839
-0.718770	-0.784926	-0.624966	-0.616294	-0.645791	-0.710038
-0.466796	-0.625849	-0.439461	-0.309346	-0.540895	-0.249344

Table 12.1 – Scaled dataset

Now we have the scaled dataset with us, let us check the boxplot of the columns of the scaled dataset,

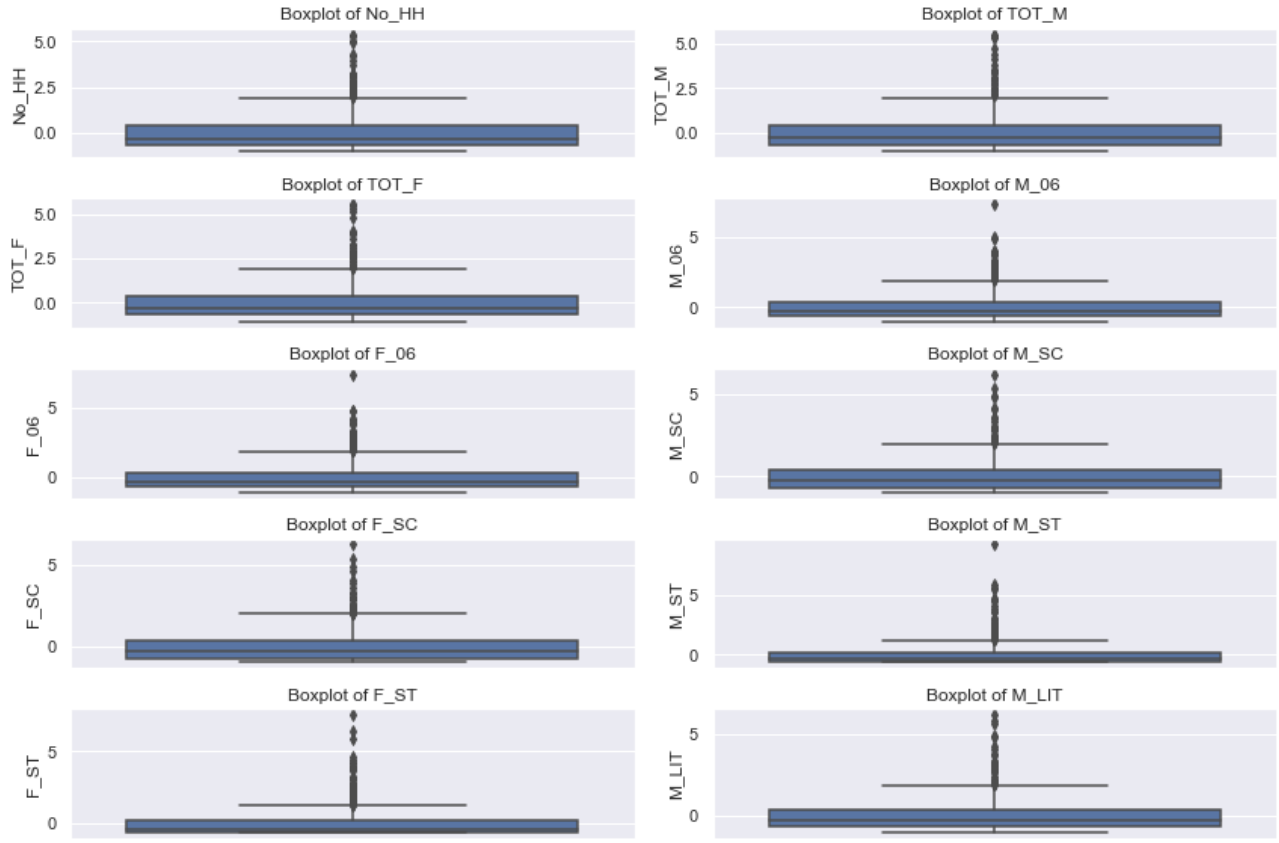

Fig 12.7 – Boxplot after scaling

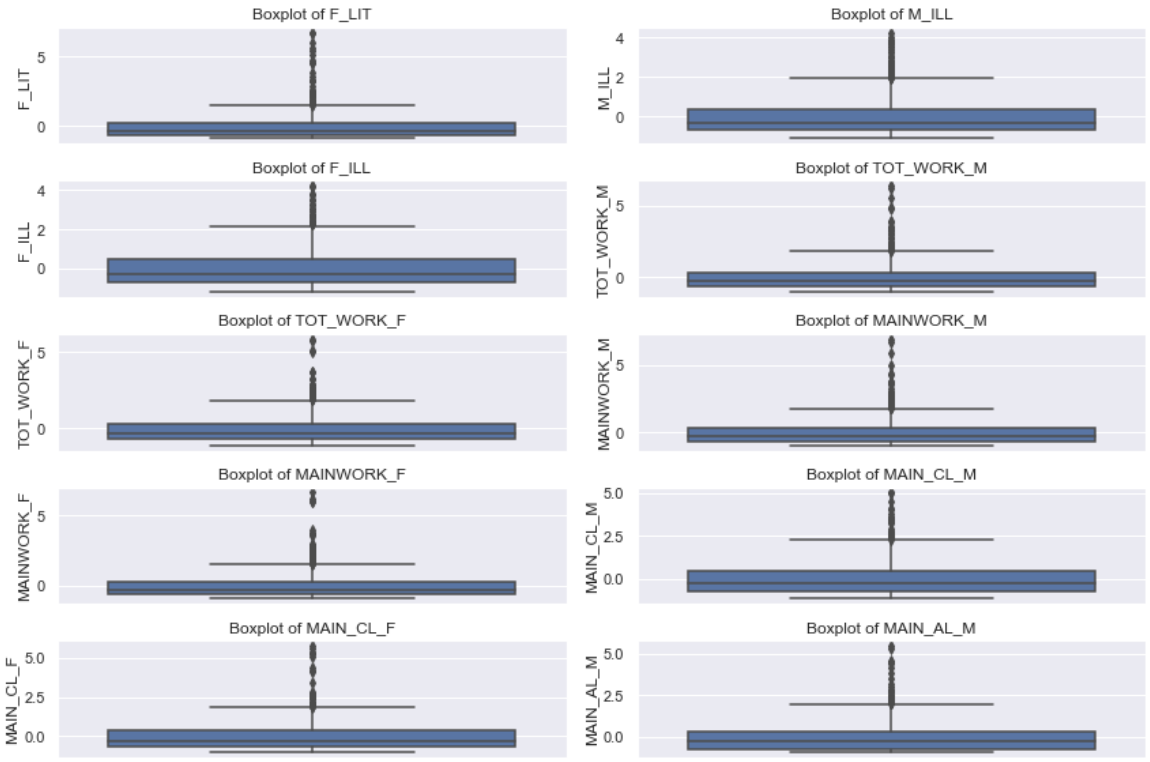

Fig 12.8 – Boxplot after scaling

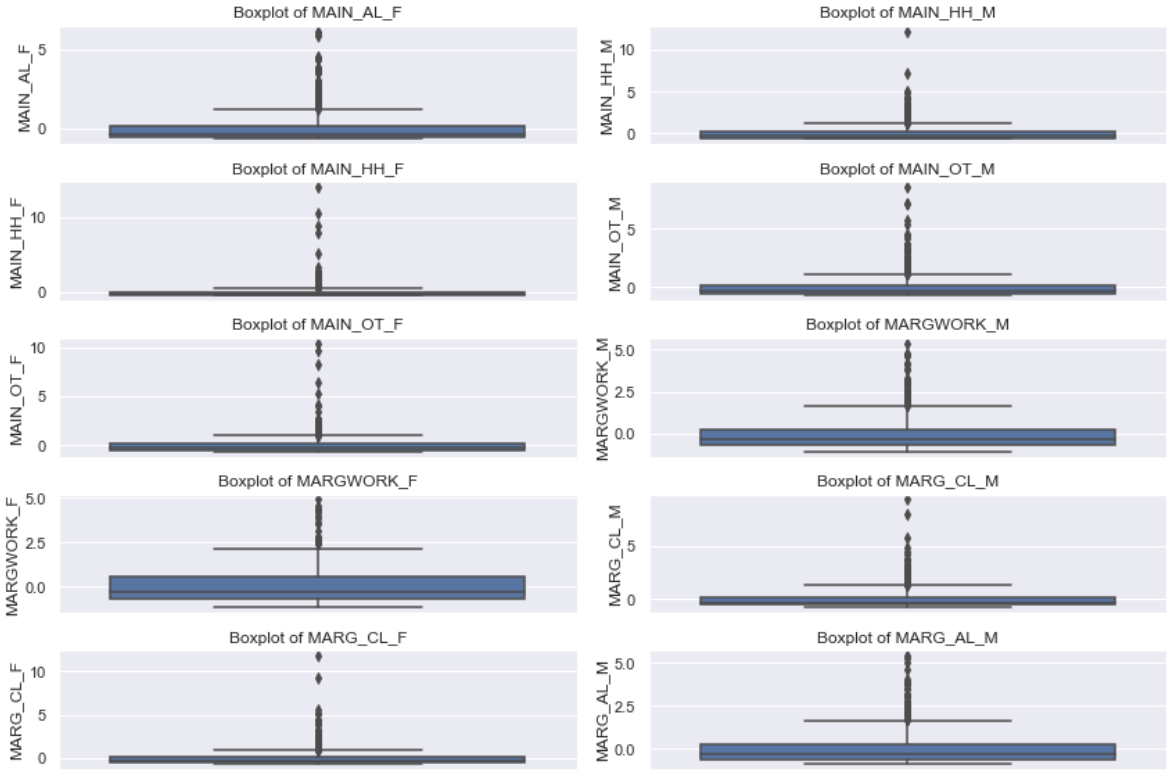

Fig 12.9 – Boxplot after scaling

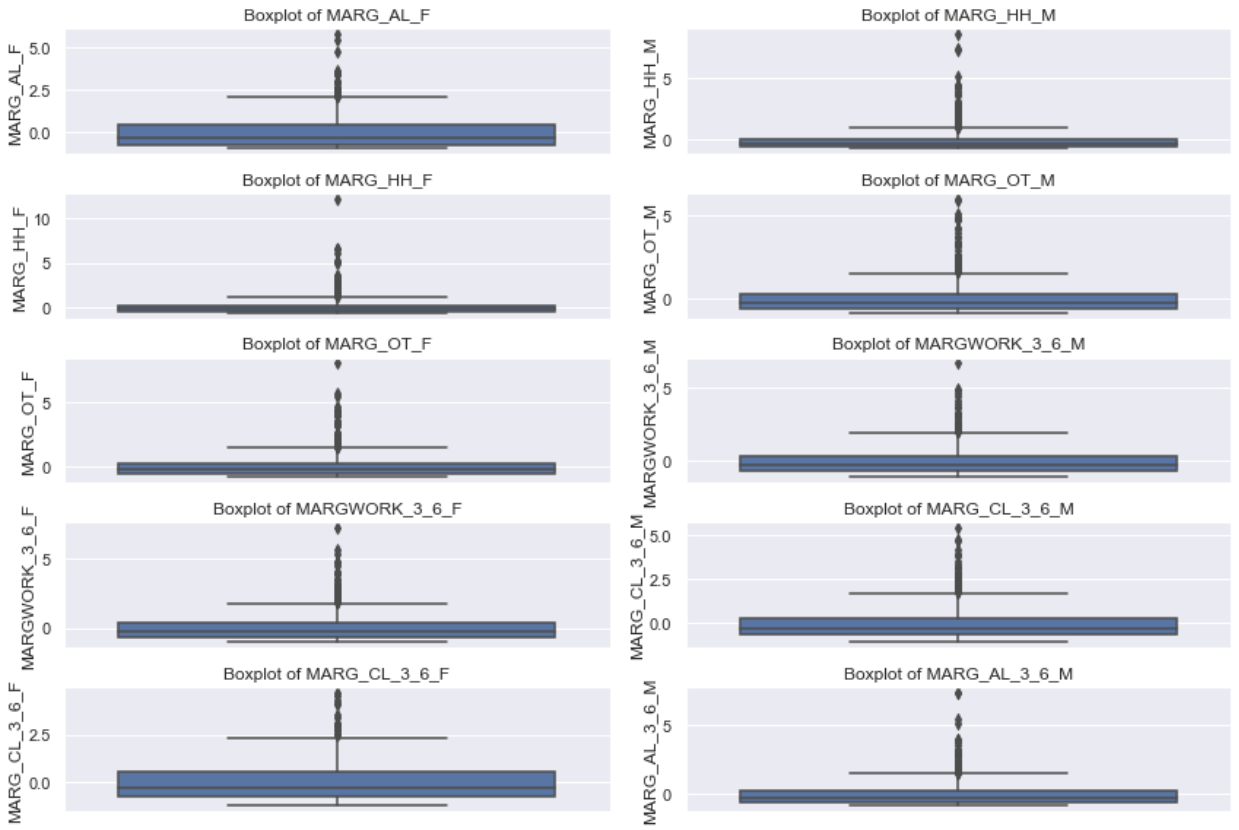

Fig 12.10 – Boxplot after scaling

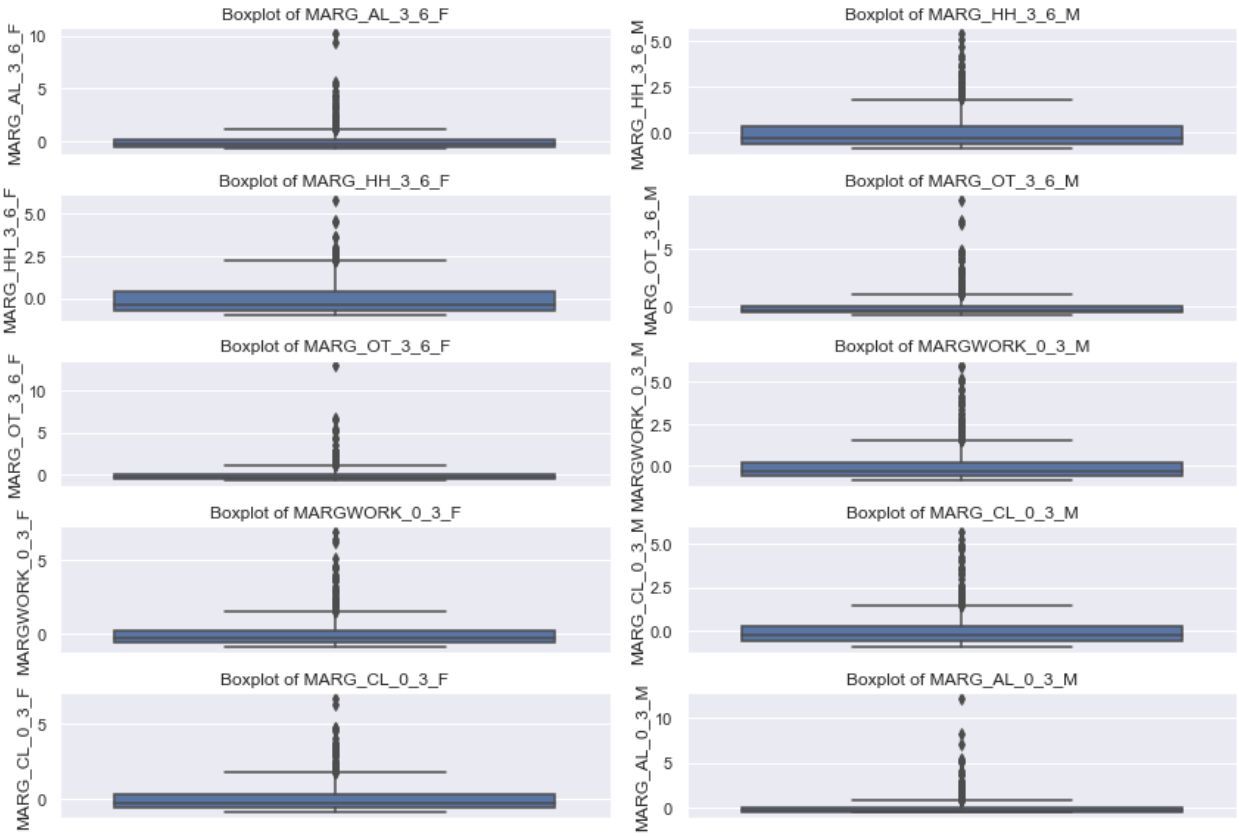

Fig 12.11 – Boxplot after scaling

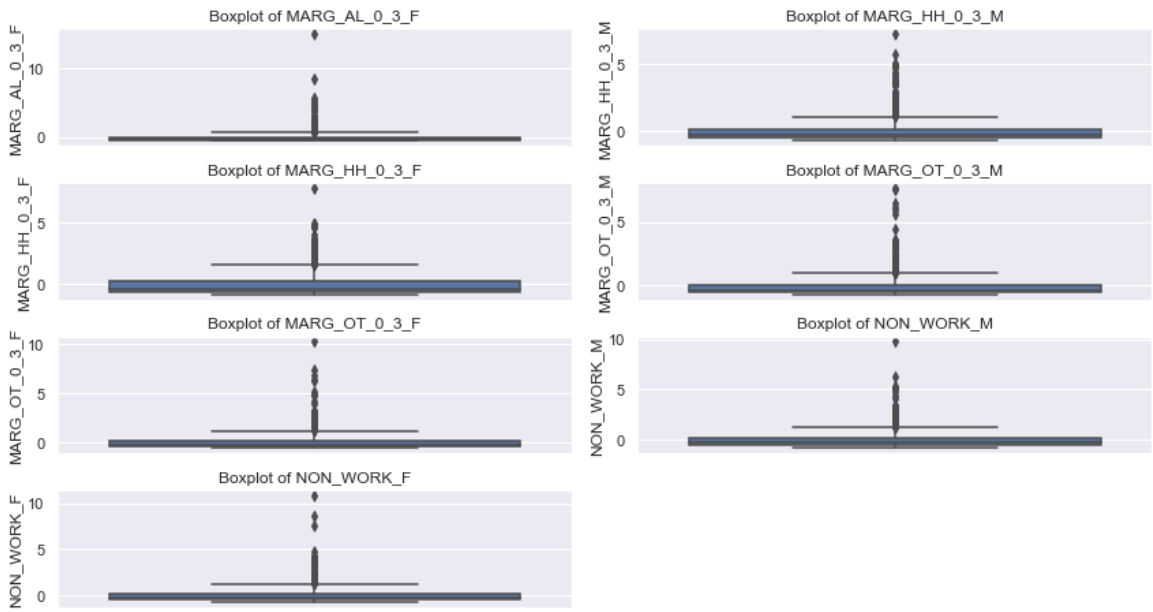

Fig 12.12 – Boxplot after scaling

While comparing the box plots before and after the scaling of the data it is very clear that the outliers are present even after the scaling of the data, Therefore we can conclude that the scaling does not affect the outliers present in the data.

13. Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.

First let us check the correlation between the variables after scaling of the data.

Below shown heat map show the correlation between the variables in the data set. By checking the color and the correlation values in the heat map we can understand that there is high correlation between the variables in the dataset. The color in the top portion as well as the bottom portion of the heat map shows the high correlation value. But in the case of male scheduled tribe and female scheduled tribe the correlation between other variables is low compared to the other columns in the dataset.

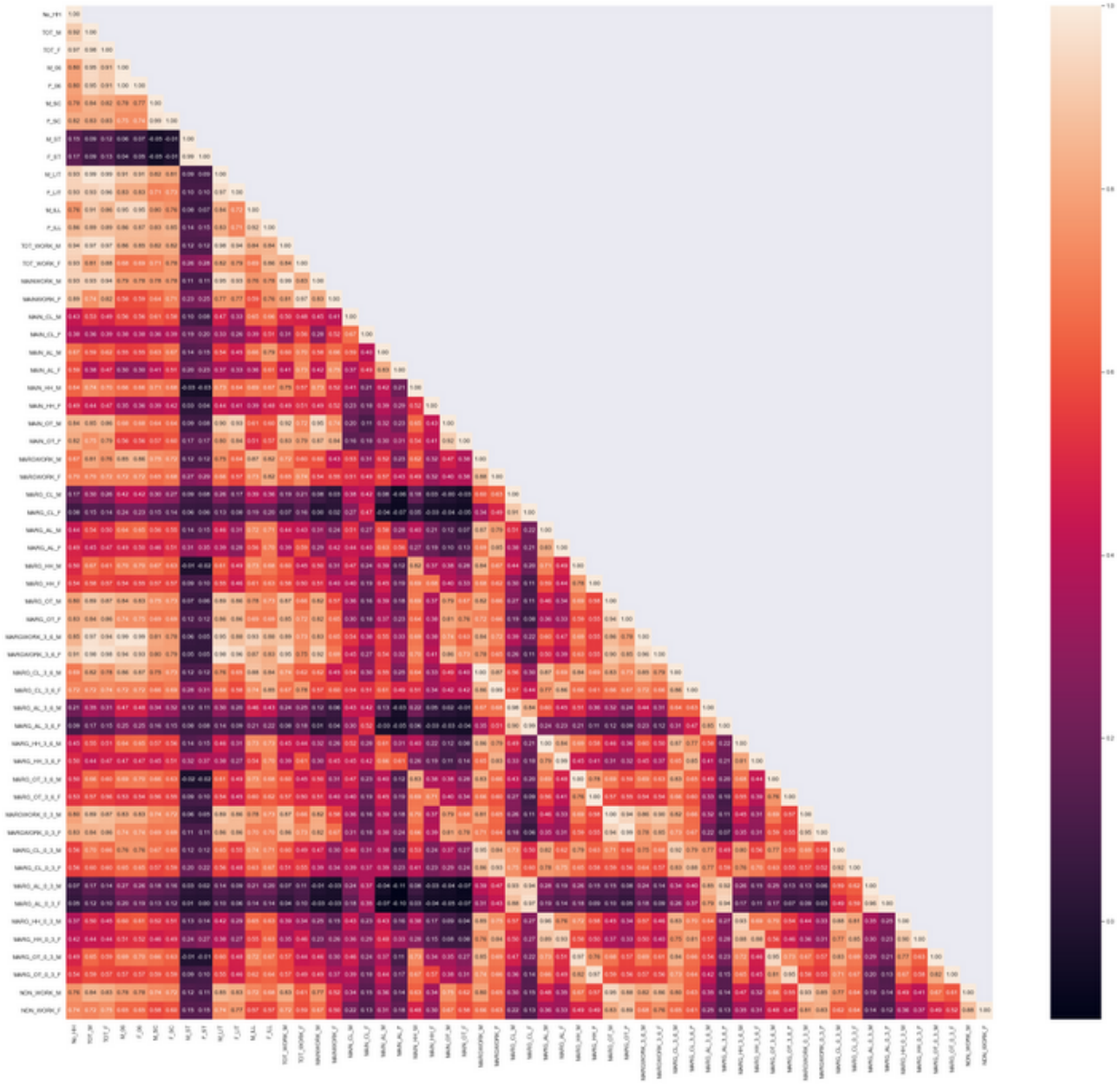

Fig 13.1 – Heat map after scaling

Before performing the necessary steps in the PCA we can confirm the statistical significance of correlations using the bartlett_sphericity from factor analyser, if the p value is less than 0.05 we can reject the null hypothesis. In this case we got the p value as zero, therefore we can reject the null hypothesis and accept the alternate hypothesis which states that there are significant correlation between the variables.

Now let us confirm the adequacy of the sample size, kmo model from factor analyser. If the value is above 0.7 it is good and if it is below 0.5 it is not acceptable. We have got the kmo model value from python is 0.8039 which is good.

Let us create the covariance matrix using the scaled data,

```
Covariance Matrix
%s [[ 1.00156495e+00  9.17603640e-01  9.72108714e-01  7.98806907e-01
   7.97618919e-01  7.76522369e-01  8.25136683e-01  1.49861475e-01
   1.65360681e-01  9.33396787e-01  9.29538954e-01  7.64234752e-01
   8.63422916e-01  9.39667104e-01  9.26707323e-01  9.28078721e-01
   8.92700480e-01  4.32077194e-01  3.83279052e-01  6.74691780e-01
   5.86772401e-01  6.42378897e-01  4.91676318e-01  8.45082456e-01
   8.23643362e-01  6.75430316e-01  6.99039472e-01  1.69019949e-01
   8.07298427e-02  4.37790432e-01  4.89923160e-01  5.03009772e-01
   5.39103798e-01  8.02224745e-01  8.34538508e-01  8.48133805e-01
   9.15577221e-01  6.93460151e-01  7.24567926e-01  2.10970704e-01
   9.45852793e-02  4.49083401e-01  4.97753194e-01  5.02113508e-01
   5.32970978e-01  7.99910847e-01  8.30450430e-01  5.57812590e-01
   5.56412458e-01  6.75308932e-02  4.62004765e-02  3.69167588e-01
   4.18100022e-01  4.87508348e-01  5.37694329e-01  7.63577220e-01
   7.36843781e-01]
 [ 9.17603640e-01  1.00156495e+00  9.84178228e-01  9.52312991e-01
   9.49274978e-01  8.41239667e-01  8.27592378e-01  9.15645222e-02
   8.63149540e-02  9.90860057e-01  9.33166178e-01  9.12965224e-01
   8.86746418e-01  9.71935898e-01  8.09159268e-01  9.34292216e-01
   7.45533191e-01  5.32566581e-01  3.56444388e-01  5.94348459e-01
   3.80342435e-01  7.41512399e-01  4.44206228e-01  8.49178302e-01
   7.46407716e-01  8.07617007e-01  7.02938295e-01  3.02005404e-01
   1.52087484e-01  5.43098289e-01  4.47527349e-01  6.68012295e-01
   5.80604565e-01  8.93825626e-01  8.40843501e-01  9.75837847e-01
   9.84794613e-01  8.21151968e-01  7.16386139e-01  3.53435289e-01
   1.65330938e-01  5.47453302e-01  4.37110738e-01  6.65258684e-01
   5.70553316e-01  8.92161675e-01  8.43624982e-01  6.99402835e-01
   5.96628075e-01  1.67666490e-01  1.15761019e-01  4.96703648e-01
   4.41048742e-01  6.52623598e-01  5.89100695e-01  8.46218438e-01
   7.17181809e-01]
[ 9.72108714e-01  9.84178228e-01  1.00156495e+00  9.09396230e-01
  9.07975479e-01  8.18237750e-01  8.34058837e-01  1.23818997e-01
  1.28847537e-01  9.86983011e-01  9.58509732e-01  8.59541939e-01
  8.88304623e-01  9.70470995e-01  8.77604172e-01  9.42488524e-01
  8.24109808e-01  4.88420440e-01  3.85976046e-01  6.24699606e-01
  4.73487627e-01  7.02053474e-01  4.67029194e-01  8.61586560e-01
  7.95377017e-01  7.65600124e-01  7.20077021e-01  2.60172620e-01
  1.35980950e-01  5.02354444e-01  4.75445042e-01  6.07208659e-01
  5.66112546e-01  8.73238759e-01  8.58859719e-01  9.44526418e-01
  9.77918860e-01  7.80074826e-01  7.37570114e-01  3.07212718e-01
  1.49571587e-01  5.08307799e-01  4.71728794e-01  6.04250359e-01
  5.56879542e-01  8.71166697e-01  8.57542688e-01  6.56372952e-01
  5.99887862e-01  1.38979900e-01  9.95931546e-02  4.51716738e-01
  4.43825041e-01  5.94734505e-01  5.72748001e-01  8.28948515e-01
  7.47750967e-01]
[ 7.98806907e-01  9.52312991e-01  9.09396230e-01  1.00156495e+00
  9.99713045e-01  7.82342397e-01  7.48699648e-01  5.53606688e-02
  4.40169441e-02  9.14185662e-01  8.33811774e-01  9.46888857e-01
  8.64674691e-01  8.57112380e-01  6.84563382e-01  7.90929409e-01
  5.85894847e-01  5.62042387e-01  3.82591721e-01  5.50717765e-01
  2.96713228e-01  6.60794280e-01  3.55281633e-01  6.85361373e-01
  5.58259562e-01  8.52331936e-01  7.17020791e-01  4.19430745e-01
  2.40181288e-01  6.40897777e-01  4.88375980e-01  6.97883268e-01
  5.43243856e-01  8.36671901e-01  7.45809701e-01  9.90371955e-01
  9.38256653e-01  8.60757852e-01  7.17751777e-01  4.73721541e-01
  2.54006605e-01  6.42078525e-01  4.67038742e-01  6.92036325e-01
  5.28773243e-01  8.36304537e-01  7.43890251e-01  7.61799838e-01
  6.48010572e-01  2.67091246e-01  1.98648856e-01  6.02030662e-01
  5.14859840e-01  6.91681514e-01  5.65914158e-01  7.86189192e-01
  6.52162312e-01]
```

Table 13.1 – Sample of covariance matrix

`

Now using sklearn let us create the eigen vectors,

```
array([[ 1.56020579e-01,  1.67117635e-01,  1.65553179e-01,
         1.62192948e-01,  1.62566396e-01,  1.51357849e-01,
         1.51566500e-01,  2.72341946e-02,  2.81833150e-02,
         1.61992837e-01,  1.46872680e-01,  1.61749445e-01,
         1.65248187e-01,  1.59871988e-01,  1.45935804e-01,
         1.46200730e-01,  1.23970284e-01,  1.03127159e-01,
         7.45397856e-02,  1.13355712e-01,  7.38821590e-02,
         1.31572584e-01,  8.33826397e-02,  1.23526242e-01,
         1.11021264e-01,  1.64615479e-01,  1.55395618e-01,
         8.23885414e-02,  4.91953957e-02,  1.28598563e-01,
         1.14305073e-01,  1.40853227e-01,  1.27669598e-01,
         1.55262872e-01,  1.47286584e-01,  1.64971950e-01,
         1.61253433e-01,  1.65501611e-01,  1.55647049e-01,
         9.30142064e-02,  5.15358640e-02,  1.28576116e-01,
         1.10645843e-01,  1.39592763e-01,  1.24545909e-01,
         1.54293786e-01,  1.46285654e-01,  1.50125706e-01,
         1.40157047e-01,  5.25417829e-02,  4.17859530e-02,
         1.21840354e-01,  1.16011410e-01,  1.39868774e-01,
         1.32192245e-01,  1.50375578e-01,  1.31066203e-01],
       [-1.26346525e-01, -8.96765481e-02, -1.04912371e-01,
        -2.20945086e-02, -2.02705495e-02, -4.51109032e-02,
        -5.19237543e-02,  2.76790387e-02,  3.02225550e-02,
        -1.15354767e-01, -1.53109487e-01, -6.62537318e-03,
        -9.10743681e-03, -1.33529221e-01, -8.50869689e-02,
        -1.76368057e-01, -1.51412544e-01,  6.24149874e-02,
         8.64767269e-02, -3.10403498e-02, -5.86880214e-02,
        -7.60210677e-02, -8.24766375e-02, -2.12984254e-01,
        -2.10071166e-01,  9.29935012e-02,  1.25269967e-01,
         2.69449716e-01,  2.46546811e-01,  1.65830750e-01,
         1.40957749e-01,  6.80679428e-02,  2.42164125e-02,
        -8.94419720e-02, -1.17899307e-01, -4.39949601e-02,
        -1.05501898e-01,  7.71926975e-02,  1.03173976e-01,
         2.64409408e-01,  2.44261317e-01,  1.58782773e-01,
         1.25286970e-01,  6.22623250e-02,  1.47659019e-02,
        -9.31585894e-02, -1.25595577e-01,  1.50680869e-01,
         1.80690375e-01,  2.51328442e-01,  2.40719745e-01,
         1.85277342e-01,  1.80615650e-01,  8.48690452e-02,
         5.08133220e-02, -6.53645529e-02, -7.38474208e-02],
```

Table 13.2 – Sample of Eigen vectors

Now let us create the Eigen values of for the corresponding using sklearn,

```
array([3.18135647e+01, 7.86942415e+00, 4.15340812e+00, 3.66879058e+00,
       2.20652588e+00, 1.93827502e+00, 1.17617374e+00, 7.51159086e-01,
       6.17053743e-01, 5.28300887e-01, 4.29831189e-01, 3.53440201e-01,
       2.96163013e-01, 2.81275560e-01, 1.92158325e-01, 1.36267920e-01,
       1.13389199e-01, 1.06303946e-01, 9.72885376e-02, 8.01062194e-02,
       5.76089954e-02, 4.43955966e-02, 3.78910846e-02, 2.96360194e-02,
       2.70797618e-02, 2.34458139e-02, 1.45111511e-02, 1.09852268e-02,
       9.31507853e-03, 8.13540203e-03, 7.89250253e-03, 5.02601514e-03,
       2.59771182e-03, 1.06789820e-03, 7.13559124e-04, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31]).
```

Table 13.3 –Eigen values

14. Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.

Let us calculate the explained variance of the each principle components using sklearn,

```
array([3.18135647e+01, 7.86942415e+00, 4.15340812e+00, 3.66879058e+00
       2.20652588e+00, 1.93827502e+00, 1.17617374e+00, 7.51159086e-01
       6.17053743e-01, 5.28300887e-01, 4.29831189e-01, 3.53440201e-01
       2.96163013e-01, 2.81275560e-01, 1.92158325e-01, 1.36267920e-01
       1.13389199e-01, 1.06303946e-01, 9.72885376e-02, 8.01062194e-02
       5.76089954e-02, 4.43955966e-02, 3.78910846e-02, 2.96360194e-02
       2.70797618e-02, 2.34458139e-02, 1.45111511e-02, 1.09852268e-02
       9.31507853e-03, 8.13540203e-03, 7.89250253e-03, 5.02601514e-03
       2.59771182e-03, 1.06789820e-03, 7.13559124e-04, 2.47799812e-31
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31
       2.47799812e-31])
```

Table 14.1 –PCA explained variance

Now we can check explained variance ratio of each principle component,

```
array([5.57260632e-01, 1.37844354e-01, 7.27529548e-02, 6.42641771e-02,
       3.86504944e-02, 3.39516923e-02, 2.06023855e-02, 1.31576386e-02,
       1.08085894e-02, 9.25395468e-03, 7.52911540e-03, 6.19101667e-03,
       5.18772384e-03, 4.92694855e-03, 3.36593119e-03, 2.38692984e-03,
       1.98617593e-03, 1.86206747e-03, 1.70414955e-03, 1.40317638e-03,
       1.00910494e-03, 7.77653131e-04, 6.63717190e-04, 5.19117774e-04,
       4.74341222e-04, 4.10687364e-04, 2.54183814e-04, 1.92422147e-04,
       1.63167083e-04, 1.42503342e-04, 1.38248605e-04, 8.80379297e-05,
       4.55026824e-05, 1.87057826e-05, 1.24990208e-05, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33, 4.34057237e-33, 4.34057237e-33, 4.34057237e-33,
       4.34057237e-33])
```

Table 14.2 –PCA explained variance ratio

After calculating explained variance ratio let create a data frame containing the loadings or coefficients of all PCs, sample of the dataframe shown below,

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
No_HH	0.156021	-0.126347	-0.002690	-0.125293	-0.007022	0.004083	-0.118110	0.057238	0.004265	0.019985	0.010592	0.086193	-0.104175
TOT_M	0.167118	-0.089677	0.056698	-0.019942	-0.033026	-0.073389	0.089554	0.111431	0.018872	-0.024501	0.011145	0.018850	-0.035665
TOT_F	0.165553	-0.104912	0.038749	-0.070873	-0.012847	-0.043647	-0.002124	0.088355	0.014911	-0.038041	0.007734	0.093551	-0.056325
M_06	0.162193	-0.022095	0.057788	0.011917	-0.050248	-0.157957	0.165067	0.169595	-0.056773	-0.153574	0.081252	0.104353	0.018338
F_06	0.162566	-0.020271	0.050126	0.014844	-0.043848	-0.154436	0.169082	0.169459	-0.059323	-0.169567	0.081964	0.105280	0.014711
M_SC	0.151358	-0.045111	0.002569	0.012485	-0.173007	-0.064295	-0.001566	-0.129301	0.037480	0.448517	-0.228821	-0.076365	0.179625
F_SC	0.151567	-0.051924	-0.025101	-0.029893	-0.159803	-0.040518	-0.084658	-0.144352	0.041232	0.446968	-0.213023	-0.010991	0.177701
M_ST	0.027234	0.027679	-0.123504	-0.222247	0.433163	0.222591	0.405505	0.021982	0.018632	0.160418	0.067589	0.014766	0.088277
F_ST	0.028183	0.030223	-0.139769	-0.229754	0.438792	0.225531	0.357800	0.014873	0.043866	0.134862	0.053348	0.022340	0.054592
M_LIT	0.161993	-0.115355	0.082168	-0.035163	-0.009101	-0.055465	0.045934	0.099423	0.045193	-0.005752	-0.030219	0.075910	-0.045563
F_LIT	0.146873	-0.153109	0.117098	-0.059559	0.055844	-0.048021	-0.021064	0.110360	0.021997	-0.040669	-0.033355	0.192889	-0.113669
M_ILL	0.161749	-0.006625	-0.021855	0.025348	-0.096580	-0.115234	0.201947	0.132080	-0.057596	-0.074471	0.126474	-0.144302	-0.003521
F_ILL	0.165248	-0.009107	-0.093062	-0.076023	-0.119911	-0.028757	0.028425	0.037269	0.000920	-0.026949	0.071768	-0.081637	0.045225
TOT_WORK_M	0.159872	-0.133529	0.045176	-0.040154	-0.019553	-0.001801	0.045053	0.076869	0.045256	0.080155	-0.031361	-0.104659	-0.119446

`

TOT_WORK_F	0.145936	-0.085087	-0.059450	-0.225160	-0.040437	0.105162	-0.119424	-0.041254	0.114254	-0.071732	-0.052991	0.002982	0.040241
MAINWORK_M	0.146201	-0.176368	0.054295	-0.068351	-0.036802	0.019283	0.047367	0.087962	0.062067	0.092100	-0.042866	-0.100842	-0.157420
MAINWORK_F	0.123970	-0.151413	-0.055609	-0.246640	-0.082834	0.123832	-0.090431	-0.018037	0.143679	-0.050134	-0.065198	0.047211	0.068761
MAIN_CL_M	0.103127	0.062415	-0.067399	-0.089769	-0.286039	-0.006170	0.385792	-0.231344	-0.364574	0.065457	-0.257296	-0.236974	-0.159564
MAIN_CL_F	0.074540	0.086477	-0.009238	-0.288965	-0.241936	0.102951	0.207882	-0.299574	-0.113374	-0.358866	-0.299760	0.373462	0.081172
MAIN_AL_M	0.113356	-0.031040	-0.247917	-0.136082	-0.205724	-0.031068	-0.013077	0.051757	-0.179034	0.263965	0.423837	-0.021895	-0.135075
MAIN_AL_F	0.073882	-0.058688	-0.251932	-0.290042	-0.177605	0.019240	-0.158334	-0.113997	0.024901	0.089929	0.403912	0.208630	0.059417
MAIN_HH_M	0.131573	-0.076021	0.026569	0.152366	-0.134089	0.174465	0.119825	-0.135093	0.383825	0.012559	0.173124	-0.063196	0.022776
MAIN_HH_F	0.083383	-0.082477	-0.060523	0.048950	-0.139441	0.422309	-0.139294	0.377112	-0.214205	-0.145089	-0.047917	-0.085043	0.649319
MAIN_OT_M	0.123526	-0.212984	0.137378	-0.040289	0.064638	0.023477	-0.015601	0.142203	0.166356	0.033905	-0.117529	-0.070224	-0.129053
MAIN_OT_F	0.111021	-0.210071	0.095634	-0.120391	0.080743	0.083079	-0.070645	0.069684	0.278023	-0.015183	-0.284779	-0.157410	-0.063169
MARGWORK_M	0.164615	0.092994	-0.008628	0.093018	0.060244	-0.090761	0.020195	0.004105	-0.041253	0.002678	0.027973	-0.085960	0.081709
MARGWORK_F	0.155396	0.125270	-0.049370	-0.088707	0.089202	0.017868	-0.157222	-0.090327	-0.005524	-0.105846	-0.001367	-0.118708	-0.051478
MARG_CL_M	0.082389	0.269450	0.198754	-0.062761	-0.022263	0.031915	0.029072	0.073819	0.015655	0.043908	0.015222	-0.078518	0.015234
MARG_CL_F	0.049195	0.246547	0.268787	-0.168402	-0.059205	0.092086	-0.045884	0.004893	0.023722	-0.001678	0.067293	-0.056365	-0.045496
MARG_AL_M	0.128599	0.165831	-0.189868	0.091787	0.019422	-0.141605	0.020298	0.141823	-0.011349	0.023736	-0.066130	-0.064786	0.008587
MARG_AL_F	0.114305	0.140958	-0.267768	-0.106365	0.080527	-0.085120	-0.150712	-0.025186	0.091925	-0.111567	-0.054118	-0.077522	-0.019512
MARG_HH_M	0.140853	0.068068	-0.021257	0.237985	-0.059971	0.089533	0.108604	-0.206463	0.238752	-0.065611	0.126318	0.078527	0.047710
MARG_HH_F	0.127670	0.024216	-0.082504	0.196321	-0.033602	0.365112	-0.049472	0.074957	-0.147745	0.007443	-0.027110	0.015512	-0.248364
MARG_OT_M	0.155263	-0.089442	0.111713	0.087119	0.119121	-0.061066	-0.004288	-0.140519	-0.109102	-0.026914	0.104995	-0.090152	0.147117
MARG_OT_F	0.147287	-0.117899	0.100046	0.026729	0.166882	0.001739	-0.117886	-0.225984	-0.143587	-0.099906	0.035036	-0.145295	0.006315
MARGWORK_3_6_M	0.164972	-0.043995	0.064423	-0.000026	-0.043834	-0.136253	0.126291	0.137732	-0.006744	-0.120815	0.050199	0.133048	0.044382
MARGWORK_3_6_F	0.161253	-0.105502	0.079704	0.003894	0.000537	-0.106900	0.050625	0.139466	-0.030800	-0.019930	0.034343	0.126751	-0.095157
MARG_CL_3_6_M	0.165502	0.077193	-0.024205	0.092875	0.054073	-0.096708	0.026671	-0.002699	-0.047876	-0.032610	0.051042	-0.150503	0.072177
MARG_CL_3_6_F	0.155647	0.103174	-0.072013	-0.107860	0.073050	0.023773	-0.138021	-0.117628	0.004349	-0.155206	0.014697	-0.210384	-0.048242
MARG_AL_3_6_M	0.093014	0.264409	0.153518	-0.038488	-0.007789	0.013477	0.063274	0.068003	0.006041	-0.009779	-0.030630	-0.142157	0.001680
MARG_AL_3_6_F	0.051536	0.244261	0.256213	-0.179691	-0.061303	0.093993	-0.019221	-0.019446	0.012344	-0.060382	0.025352	-0.088684	-0.057761
MARG_HH_3_6_M	0.128576	0.158783	-0.200119	0.080411	0.008457	-0.144061	0.021451	0.140603	-0.027715	0.012104	-0.034034	-0.117072	-0.003505
MARG_HH_3_6_F	0.110646	0.125287	-0.279866	-0.136240	0.064109	-0.076708	-0.146121	-0.052007	0.079959	-0.129671	-0.009183	-0.130286	-0.023739
MARG_OT_3_6_M	0.139593	0.062262	-0.020618	0.237745	-0.066400	0.097057	0.115068	-0.213660	0.242023	-0.073376	0.161815	0.046763	0.039682

MARG_OT_3_6_F	0.124546	0.014766	-0.082794	0.190511	-0.044810	0.384552	-0.042239	0.084448	-0.169232	-0.015811	-0.007154	-0.035351	-0.219654
MARGWORK_0_3_M	0.154294	-0.093159	0.110285	0.086479	0.108829	-0.062043	-0.001204	-0.142040	-0.097366	-0.065249	0.125481	-0.143512	0.141411
MARGWORK_0_3_F	0.146286	-0.125596	0.095667	0.027275	0.141190	0.008962	-0.091060	-0.224163	-0.080628	-0.137432	0.035034	-0.250325	0.018052
MARG_CL_0_3_M	0.150126	0.150681	0.054892	0.087433	0.081185	-0.060715	-0.007316	0.031324	-0.011756	0.145097	-0.067101	0.180551	0.114811
MARG_CL_0_3_F	0.140157	0.180690	0.023982	-0.022290	0.129936	-0.001727	-0.200877	0.000979	-0.034985	0.053883	-0.050013	0.170702	-0.056509
MARG_AL_0_3_M	0.052542	0.251328	0.268330	-0.104686	-0.048849	0.065409	-0.042295	0.077716	0.033223	0.146568	0.105228	0.056829	0.040718
MARG_AL_0_3_F	0.041786	0.240720	0.284956	-0.135716	-0.051895	0.083743	-0.103407	0.059042	0.048087	0.129527	0.157994	0.018327	-0.016082
MARG_HH_0_3_M	0.121840	0.185277	-0.138628	0.132544	0.062380	-0.124209	0.014587	0.139166	0.054918	0.069137	-0.191385	0.148443	0.056644
MARG_HH_0_3_F	0.116011	0.180616	-0.202198	0.004051	0.128308	-0.105530	-0.152175	0.067420	0.123663	-0.040286	-0.200324	0.107251	-0.003473
MARG_OT_0_3_M	0.139869	0.084869	-0.022599	0.230038	-0.036390	0.061228	0.083122	-0.174950	0.219106	-0.037385	0.003662	0.181251	0.072645
MARG_OT_0_3_F	0.132192	0.050813	-0.078720	0.206201	0.000165	0.295600	-0.068722	0.044704	-0.080036	0.074747	-0.084129	0.162755	-0.322929
NON_WORK_M	0.150376	-0.065365	0.111827	0.084854	0.162862	-0.052386	-0.019354	-0.124166	-0.160651	0.165440	-0.003461	0.180885	0.166299
NON_WORK_F	0.131066	-0.073847	0.102553	0.021124	0.238292	-0.024901	-0.200053	-0.202142	-0.354393	0.050779	0.030313	0.258317	-0.037447

Table 14.3 –Sample of dataframe containing the loadings

Now using PCA explained variance ratio we can create scree plot of the same,

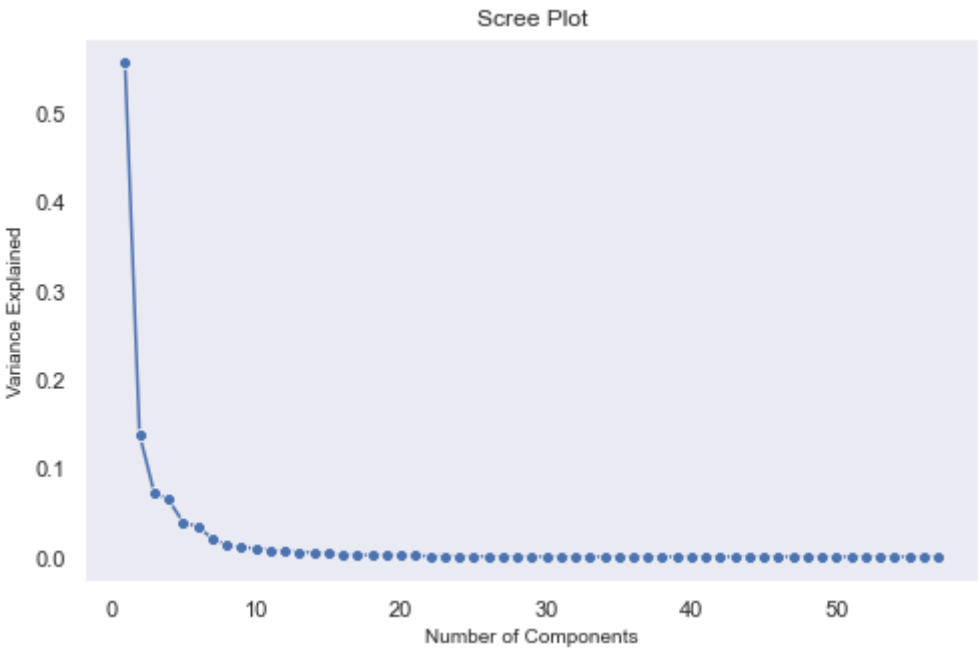

Fig 14.1 –Scree plot

Let us check the cumulative explained variance ratio to find a cut off for selecting the number of PCs

```
array([0.55726063, 0.69510499, 0.76785794, 0.83212212, 0.87077261,
       0.9047243 , 0.92532669, 0.93848433, 0.94929292, 0.95854687,
       0.96607599, 0.97226701, 0.97745473, 0.98238168, 0.98574761,
       0.98813454, 0.99012071, 0.99198278, 0.99368693, 0.99509011,
       0.99609921, 0.99687687, 0.99754058, 0.9980597 , 0.99853404,
       0.99894473, 0.99919891, 0.99939134, 0.9995545 , 0.99969701,
       0.99983525, 0.99992329, 0.9999688 , 0.9999875 , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        ])
```

Table 14.4 – cumulative explained variance ratio

In the question it is clearly states that we have to take at least 90% explained variance for this project, while checking cumulative explained variance ratio table it is clear that we have to consider six principle components to reach the 90% explained variance. So the final data frame after considering 6 PC's is,

	PC1	PC2	PC3	PC4	PC5	PC6
No_HH	0.156021	-0.126347	-0.002690	-0.125293	-0.007022	0.004083
TOT_M	0.167118	-0.089677	0.056698	-0.019942	-0.033026	-0.073389
TOT_F	0.165553	-0.104912	0.038749	-0.070873	-0.012847	-0.043647
M_06	0.162193	-0.022095	0.057788	0.011917	-0.050248	-0.157957
F_06	0.162566	-0.020271	0.050126	0.014844	-0.043848	-0.154436
M_SC	0.151358	-0.045111	0.002569	0.012485	-0.173007	-0.064295
F_SC	0.151567	-0.051924	-0.025101	-0.029893	-0.159803	-0.040518
M_ST	0.027234	0.027679	-0.123504	-0.222247	0.433163	0.222591
F_ST	0.028183	0.030223	-0.139769	-0.229754	0.438792	0.225531
M_LIT	0.161993	-0.115355	0.082168	-0.035163	-0.009101	-0.055465
F_LIT	0.146873	-0.153109	0.117098	-0.059559	0.055844	-0.048021
M_ILL	0.161749	-0.006625	-0.021855	0.025348	-0.096580	-0.115234
F_ILL	0.165248	-0.009107	-0.093062	-0.076023	-0.119911	-0.028757
TOT_WORK_M	0.159872	-0.133529	0.045176	-0.040154	-0.019553	-0.001801
TOT_WORK_F	0.145936	-0.085087	-0.059450	-0.225160	-0.040437	0.105162
MAINWORK_M	0.146201	-0.176368	0.054295	-0.068351	-0.036802	0.019283
MAINWORK_F	0.123970	-0.151413	-0.055609	-0.246640	-0.082834	0.123832
MAIN_CL_M	0.103127	0.062415	-0.067399	-0.089769	-0.286039	-0.006170

Table 14.5 – Sample of the selected PC's

15. Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.

For comparing the selected principal components with the actual columns consider the below shown bar plot and heat map,

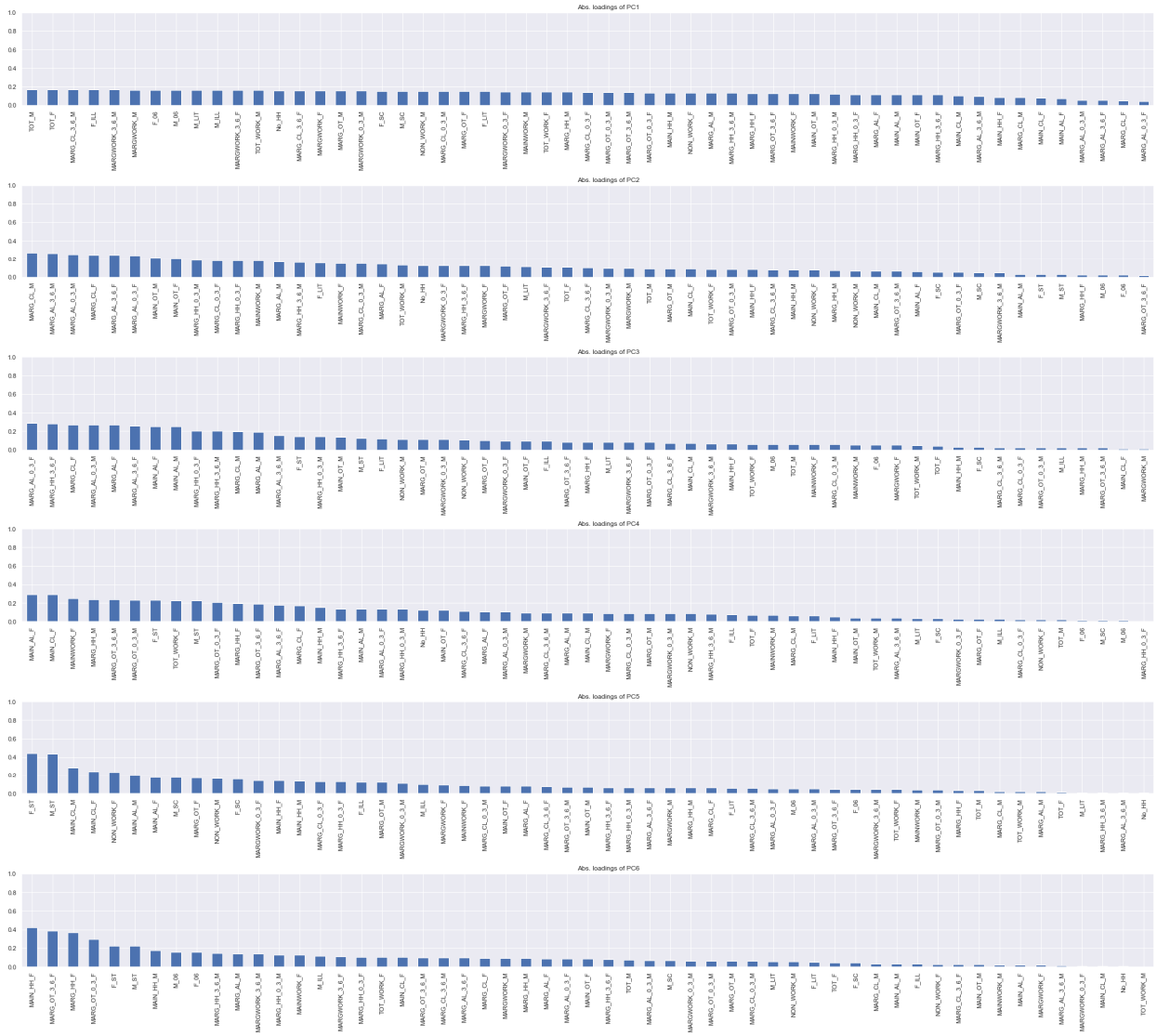

Fig 15.1 – Bar plot of PC's

From the bar plot it is clear that the PC1 shows more variability among the columns compared to other principal components in the data set. Let us consider the heat map for the same for getting more insight about the data.

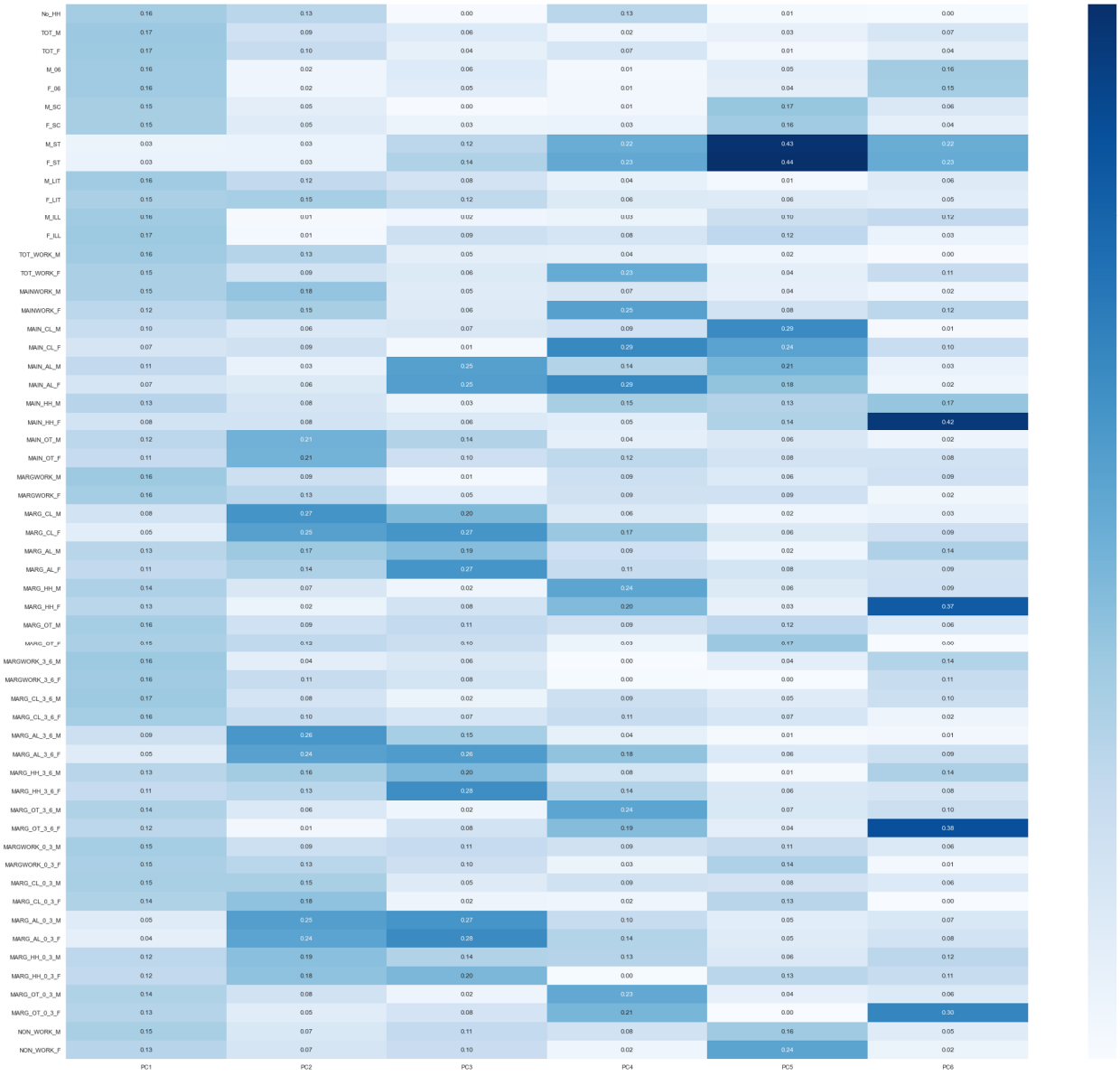

Fig 15.2 – Heat map of PC's

Consider the heat map shown above of the selected 6 PC's, the correlation values of PC1 is almost equal among all the variables. In the case of PC5 and PC6 the correlation value is higher compared to other PC's .

From the bar plot and heat map it is clear that PC1 is the optimum component it has good covariance and It shows almost equal distribution among all the variables in the dataset.

Now let us calculate the PCA final dataframe using this 6 components. The final dataframe obtained after performing the PCA on the selected 6 component dataframe is,

`

	PC1	PC2	PC3	PC4	PC5	PC6
0	-4.617263	0.138116	0.328545	1.543697	0.353736	-0.420948
1	-4.771662	-0.105865	0.244449	1.963215	-0.153884	0.417308
2	-5.964836	-0.294347	0.367394	0.619543	0.478199	0.276581
3	-6.280796	-0.500384	0.212701	1.074515	0.300799	0.051157
4	-4.478566	0.894154	1.078277	0.535557	0.804065	0.341678
5	-3.319963	2.823865	3.058460	-0.447904	0.742445	0.634676
6	-5.021393	-0.346359	0.650378	0.981072	-0.059778	-0.246957
7	-4.608709	0.022370	0.398755	1.576995	0.171316	-0.139444
8	-5.186703	-0.059097	0.184397	1.735440	0.169174	0.455039
9	-4.226190	-1.335080	0.697838	1.470509	0.269146	-0.002576

Table 15.1 – Final dataframe of PC's

Let us check the correlation between these principal components using the heat map,

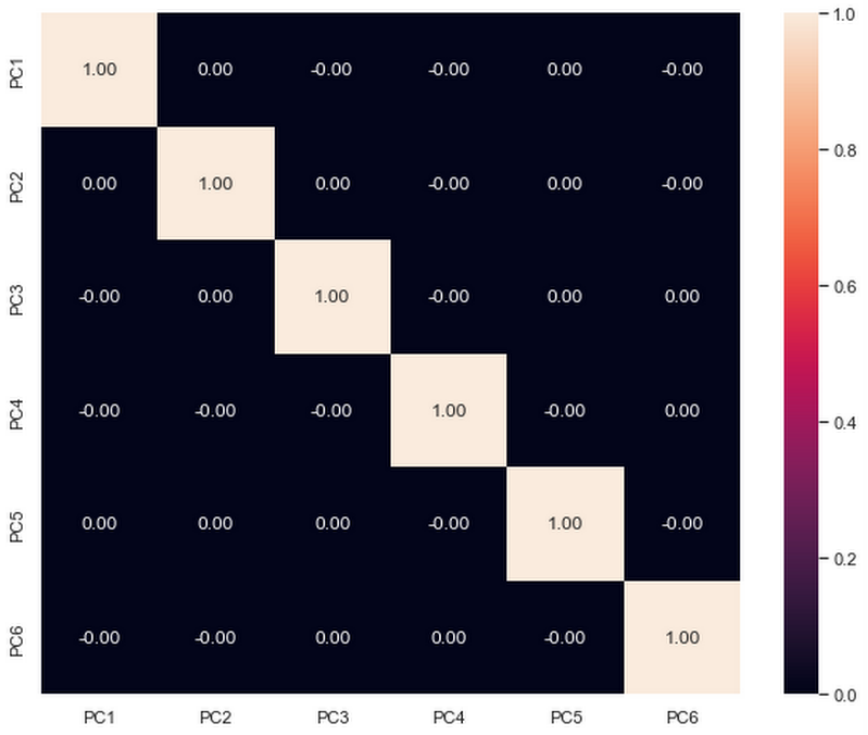

Fig 15.3 – Heat map Final

Now all the principal components are independent, no correlation between the principal components.

`

16.Write linear equation for first PC.

The final equation of PC1 is written below,

0.15602057858567936 * No_HH +

0.1671176348853345 * TOT_M +

0.16555317909064893 * TOT_F +

0.1621929482046555 * M_06 +

0.16256639565734832 * F_06 +

0.15135784909060582 * M_SC +

0.15156650019208875 * F_SC +

0.02723419457100423 * M_ST +

0.028183315015872692 * F_ST +

0.16199283733629155 * M_LIT +

0.14687268030140285 * F_LIT +

0.16174944463471633 * M_ILL +

0.16524818736833372 * F_ILL +

0.15987198816201284 * TOT_WORK_M +

0.1459358037724762 * TOT_WORK_F +

0.1462007297630599 * MAINWORK_M +

0.12397028357273648 * MAINWORK_F +

0.10312715883019867 * MAIN_CL_M +

0.07453978555483677 * MAIN_CL_F +

0.11335571218156723 * MAIN_AL_M +

0.07388215903155881 * MAIN_AL_F +

0.131572584022756 * MAIN_HH_M +

0.08338263967435766 * MAIN_HH_F +

0.12352624192253084 * MAIN_OT_M +

0.1110212639132013 * MAIN_OT_F +

`

0.1646154785601101 * MARGWORK_M +

0.1553956181083413 * MARGWORK_F +

0.08238854140704546 * MARG_CL_M +

0.049195395678738256 * MARG_CL_F +

0.12859856294668565 * MARG_AL_M +

0.11430507278919895 * MARG_AL_F +

0.1408532269618514 * MARG_HH_M +

0.12766959801475364 * MARG_HH_F +

0.15526287162311603 * MARG_OT_M +

0.1472865835652339 * MARG_OT_F +

0.16497194993714454 * MARGWORK_3_6_M +

0.1612534325753136 * MARGWORK_3_6_F +

0.1655016110258063 * MARG_CL_3_6_M +

0.15564704914483385 * MARG_CL_3_6_F +

0.09301420640192848 * MARG_AL_3_6_M +

0.051535863970152224 * MARG_AL_3_6_F +

0.12857611642867822 * MARG_HH_3_6_M +

0.11064584323696922 * MARG_HH_3_6_F +

0.13959276252158836 * MARG_OT_3_6_M +

0.12454590917258751 * MARG_OT_3_6_F +

0.15429378578916042 * MARGWORK_0_3_M +

0.14628565406214422 * MARGWORK_0_3_F +

0.15012570610262066 * MARG_CL_0_3_M +

0.14015704689010391 * MARG_CL_0_3_F +

0.05254178285396345 * MARG_AL_0_3_M +

0.04178595301201033 * MARG_AL_0_3_F +

0.12184035387925024 * MARG_HH_0_3_M +

`

0.1160114101682411 * MARG_HH_0_3_F +

0.13986877411042808 * MARG_OT_0_3_M +

0.13219224458196535 * MARG_OT_0_3_F +

0.15037557804411297 * NON_WORK_M +

0.13106620313207334 * NON_WORK_F +

THE END

CLICK HERE TO GO TO CONTENTS