greatlearning
Power Ahead

ADVANCED STATISTICS PROJECT BUSINESS REPORT

VAISHNAV U
PGP-DSBA ONLINE
20/11/2022

Table of Contents

Content

Problem-1

List of tables

List of figures

Dental Implant data

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 favor one method above another and may work better in his/her favorite method. The response is the variable of interest.

Introduction

The data consists of 90 rows and 5 columns. Dentists, method, alloy, and temperature are the different factors that affect implant hardness. The response column in the dataset represents the implant hardness. There are 5 different dentists,3 methods, and 2 types of alloys that are treated at 3 different temperatures in the dataset. Implant hardness may depend on these factors. Here we will be analyzing the implant hardness on these factors and which factor differs.

Data description

1	Dentist	Number of dentists categorised be dentist_1,dentist_2 etc.
2	Method	Number of methods(Method_1,Method_2&Method_3)
3	Alloy	Types of alloys used(Alloy_1&Allot_2)
4	Temp	Temperature at which metal is treated(3 values)
5	Response	Hardness of metal implant

Sample of 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 1.1

The above table shows factors such as dentist,method and alloy as floats.We will convert then to object type such as DENTIST_1,METHOD_1&ALLOY_ etc..The data set after the conversion is shown as below.

	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 1.2

Exploratory Data Analysis

1	Dentist	90 non-null	object
2	Method	90 non-null	object
3	Alloy	90 non-null	object
4	Temp	90 non-null	float 64
5	Response	90 non-null	float 64

Descriptive statistics of data

	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 1.3

There are 90 rows and 5 columns in the data set. The minimum implant hardness response is noted as 289 the maximum is 1115. The average implant hardness is 741.77. A total of 5 different dentists are considered, and 3 methods and 2 types of alloys constitute the factors in the data set.The temperature at which the metal is treated is noted in 3 different values

.

Problem-1

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.?**

For ALLOY_1

H0:Mean implant hardness is same with different dentists

H0: μM1 = μM2 = μM3 = μM4 = μM5 (For five different dentists)

H1:Mean implant hardness is different with at-least one dentist.

H1: μM1 ≠ μM2 = μM3 = μM4 = μM5 Or H1: μM1 = μM2 ≠ μM3 = μM4 = μM5
Or H1: μM1 = μM2 = μM3 ≠ μM4 = μM5 Or H1: μM1 = μM2 = μM3 = μM4 ≠ μM5
Or H1: μM1 ≠ μM3 = μM2 = μM4 = μM5 Or H1: μM1 ≠ μM4 = μM2 = μM3 = μM5
Or H1: μM1 ≠ μM5 = μM2 = μM3 = μM4 Or H1: μM2 ≠ μM4 = μM1 = μM3 = μM5
Or H1: μM2 ≠ μM5 = μM1 = μM3 = μM4 Or H1: μM3 ≠ μM5 = μM1 = μM2 = μM4
Or H1: μM1 ≠ μM2 ≠ μM3 = μM4 = μM5 Or H1: μM1 ≠ μM2 ≠ μM3 = μM4 = μM5
Or H1: μM2 ≠ μM3 ≠ μM4 = μM5 = μM1 Or H1: μM3 ≠ μM4 ≠ μM5 = μM1 = μM2
Or H1: μM1 ≠ μM3 ≠ μM4 = μM5 = μM2 Or H1: μM1 ≠ μM4 ≠ μM5 = μM2 = μM3
Or H1: μM2 ≠ μM4 ≠ μM5 = μM1 = μM3 Or H1: μM2 ≠ μM5 ≠ μM1 = μM3 = μM4
Or H1: μM3 ≠ μM5 ≠ μM1 = μM2 = μM4 Or H1: μM1 ≠ μM2 ≠ μM3 ≠ μM4 = μM5
Or H1: μM5 ≠ μM2 ≠ μM3 ≠ μM4 = μM1 Or H1: μM1 ≠ μM3 ≠ μM5 ≠ μM4 = μM2
Or H1: μM1 ≠ μM2 ≠ μM5 ≠ μM4 = μM3 Or H1: μM1 ≠ μM2 ≠ μM5 ≠ μM4 ≠ μM3

Where μM1,μM1 is the mean implant hardness for dentist_1,dentist_2 etc. For Alloy_1

For ALLOY_2

H0:Mean implant hardness is same with different dentists

H0: μM1 = μM2 = μM3 = μM4 = μM5 (For five different dentists)

H1:Mean implant hardness is different with at-least one dentist.

H1: μM1 ≠ μM2 = μM3 = μM4 = μM5 Or H1: μM1 = μM2 ≠ μM3 = μM4 = μM5

Or H1: μM1 = μM2 = μM3 ≠ μM4 = μM5 Or H1: μM1 = μM2 = μM3 = μM4 ≠ μM5

Or H1: μM1 ≠ μM3 = μM2 = μM4 = μM5 Or H1: μM1 ≠ μM4 = μM2 = μM3 = μM5

Or H1: μM1 ≠ μM5 = μM2 = μM3 = μM4 Or H1: μM2 ≠ μM4 = μM1 = μM3 = μM5

Or H1: μM2 ≠ μM5 = μM1 = μM3 = μM4 Or H1: μM3 ≠ μM5 = μM1 = μM2 = μM4

Or H1: μM1 ≠ μM2 ≠ μM3 = μM4 = μM5 Or H1: μM1 ≠ μM2 ≠ μM3 = μM4 = μM5

Or H1: μM2 ≠ μM3 ≠ μM4 = μM5 = μM1 Or H1: μM3 ≠ μM4 ≠ μM5 = μM1 = μM2

Or H1: μM1 ≠ μM3 ≠ μM4 = μM5 = μM2 Or H1: μM1 ≠ μM4 ≠ μM5 = μM2 = μM3

Or H1: μM2 ≠ μM4 ≠ μM5 = μM1 = μM3 Or H1: μM2 ≠ μM5 ≠ μM1 = μM3 = μM4

Or H1: μM3 ≠ μM5 ≠ μM1 = μM2 = μM4 Or H1: μM1 ≠ μM2 ≠ μM3 ≠ μM4 = μM5

Or H1: μM5 ≠ μM2 ≠ μM3 ≠ μM4 = μM1 Or H1: μM1 ≠ μM3 ≠ μM5 ≠ μM4 = μM2

Or H1: μM1 ≠ μM2 ≠ μM5 ≠ μM4 = μM3 Or H1: μM1 ≠ μM2 ≠ μM5 ≠ μM4 ≠ μM3

Where μM1,μM2 is the mean implant hardness for dentist_1,dentist_2 etc. For Alloy_2

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

Assumptions for ANOVA

1.Dependent variable should be measured at the continuous level.

2.Two independent variables should each consist of two or more categorical, independent groups.

3.There should be no significant outliers.

4.Dependent variable should be approximately normally distributed for each combination of the groups of the two independent variables.

3 **Irrespective of your conclusion in 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?**

Considering all the assumptions are met we will use one-way Anova

For ALLOY_1

H0:Mean implant hardness is same with different dentists

H1:Mean implant hardness is different with at-least one dentist

	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 1.4

For ALLOY_2

H0:Mean implant hardness is same with different dentists

H1:Mean implant hardness is different with at-least one dentist

	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 1.5

Here p value is greater than level of significance for ALLOY_1 & ALLOY _2. So we fail to reject null hypothesis.

So the mean implant hardness is same for different dentists while separately considering alloys.If the null hypothesis was rejected it is possible to identify which pairs differ using Tukey's honest significance test, or Tukey's HSD.

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?**

For ALLOY_1

H0:Mean implant hardness is same with different methods.

H1:Mean implant hardness is different with at-least one method.

	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 1.6

For ALLOY_2

H0:Mean implant hardness is same with different methods.

H1:Mean implant hardness is different with at-least one method.

	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 1.7

Here we reject null hypothesis for alloy_1 and 2.We can conclude that mean implant hardness is different with at-least one method.It is possible to identify which pairs differ using Tukey's honest significance test, or Tukey's HSD.

To identify which pair differs we will conduct Multi-comparison for ALLOY_1 & ALLOY_2 separately.

For ALLOY_1 :

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 1.8

For ALLOY_2 :

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 1.9

Multiple Comparison of Means - Tukey HSD shows that there is a difference between METHOD_1 and METHOD_3 also METHOD_2 and METHOD_3 for both alloys.Also it shows that there is no significance between METHOD_1 and METHOD_2 on implant hardness.

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?**

For ALLOY_1

H0:Mean implant hardness is same with different temperature levels.

H1:Mean implant hardness is different with at-least one temperature level.

	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 1.10

For ALLOY_2

H0:Mean implant hardness is same with different temperature levels.

H1:Mean implant hardness is different with at-least one temperature level.

	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 1.11

Here p value is greater than level of significance for ALLOY_1 & ALLOY _2. So we fail to reject null hypothesis.So the mean implant hardness is same with different temperature levels.If the null hypothesis was rejected it is possible to identify which pairs differ using Tukey's honest significance test, or Tukey's HSD.

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

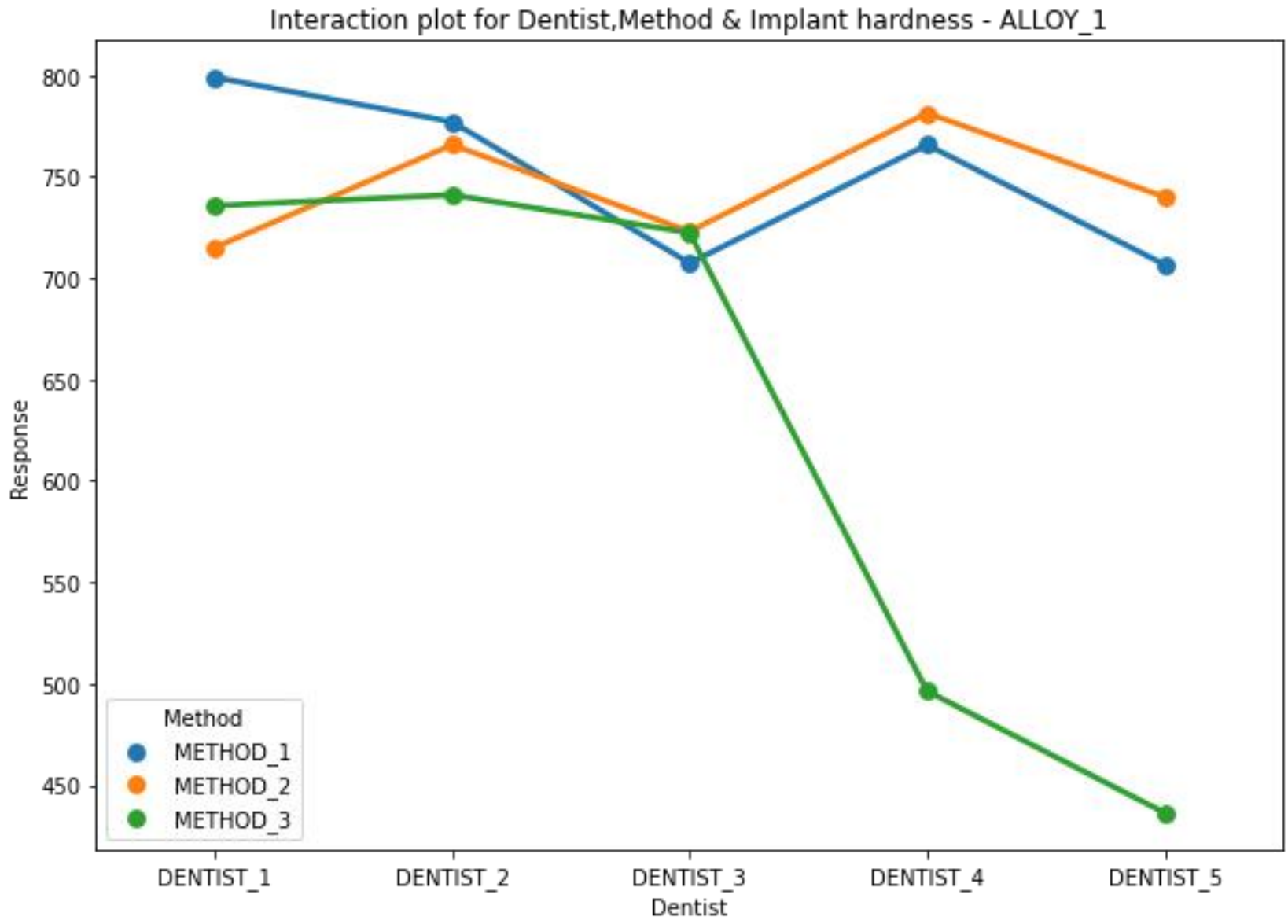


Fig.1.1

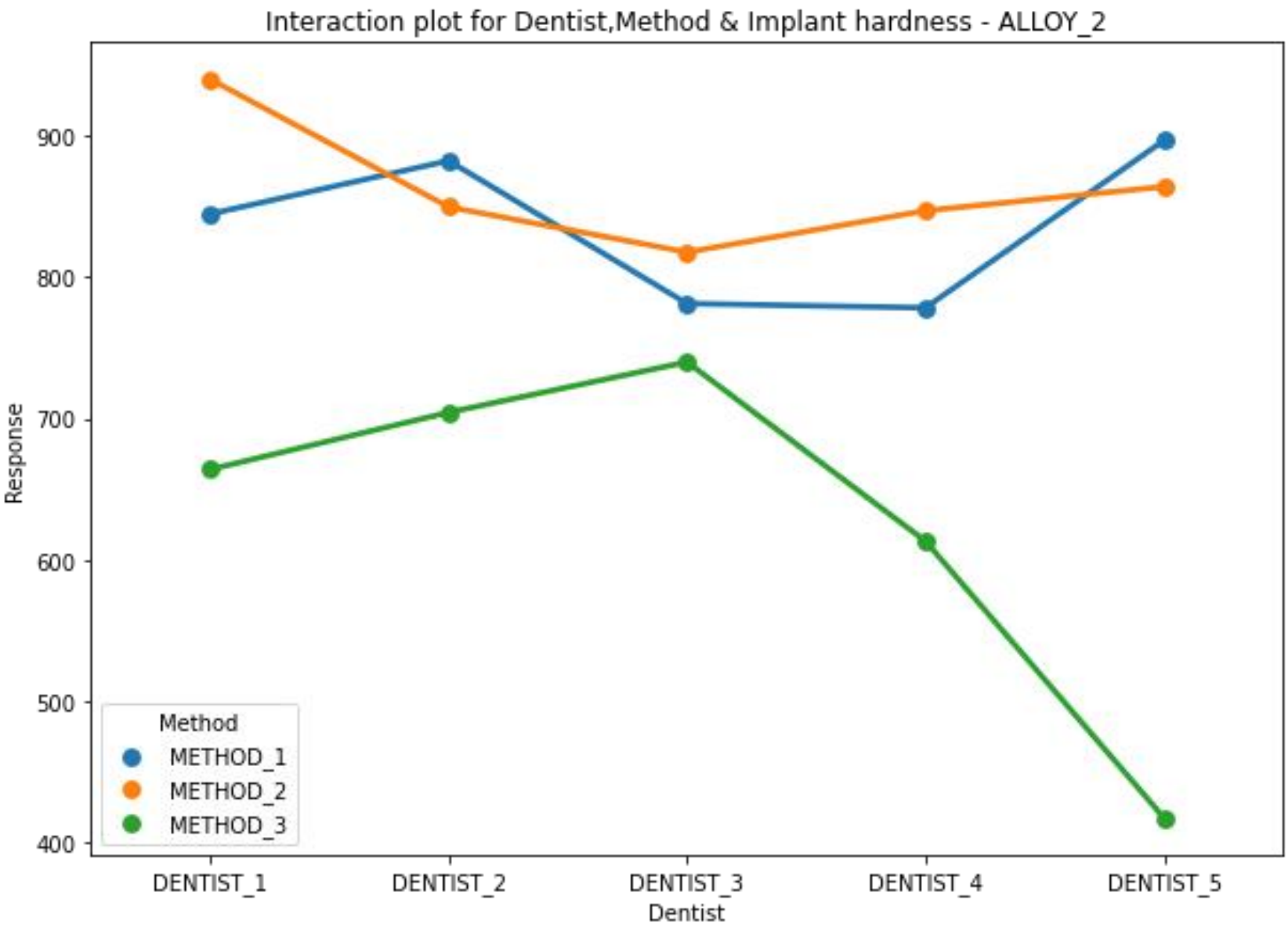


Fig.1.2

Table 1.1 & 1.13 shows the effect of factors such as dentist and method on implant hardness.Also the table consists of interaction of dentist and method on implant hardness.

Interaction plot-inference:ALLOY_1

The interaction plot for the dentist, method on implant hardness for ALLOY_1 shows there is an interaction between these factors. Average Implant hardness is highest for dentist _1 using method_1 followed by method_3 and method_2. Dentist_2 seems to prefer method_1 followed by methods 2 & 3. For dentist_3 average implant, hardness is clustered around a point, so we can conclude that response is approximately equal for all three methods.Dentist_4 & 5 prioritize method_2 where method_3 has the least response.

Interaction plot-inference:ALLOY_2

Here the interaction plot is different from ALLOY_1.Method_3 for all dentists does not show interaction with other methods. Average Implant hardness is highest for dentist _1 using method_2, followed by method_1 and method_3.Dentist_2 seems to prefer method_1 followed by method_2 & 3. For dentist_1,dentist_3, and dentist_4 method_2 seems to have a better response and for dentist _2 and dentist _5 method_1 has better response.Method_3 shows less response value for all dentists.

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?**

Without Considering Interaction effect

For ALLOY_1

H0:Mean implant hardness is same with different dentists & methods

H1:Mean implant hardness is different with at-least one dentist & method

	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 1.12

For ALLOY_2

H0:Mean implant hardness is same with different dentists & methods

H1:Mean implant hardness is different with at-least one dentist & method

	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 1.13

It is clear from both tables that method is a significant factor for implant hardness ,but dentist is not a significant factor.

Considering Interaction effect

For ALLOY_1

H0:Mean implant hardness is same with different dentists & methods

H1:Mean implant hardness is different with at-least one dentist & method

	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 1.14

For ALLOY_2

H0:Mean implant hardness is same with different dentists & methods

H1:Mean implant hardness is different with at-least one dentist & method

	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 1.15

For Alloy_1 all the p_values are less than 0.05. So we reject the null hypothesis. Dentist, method, and interaction between dentist & method have a significant effect on implant hardness. Also, the method and interaction between these factors are more significant than the dentist factor considering implant hardness.

For Alloy_2 method is the most significant factor that affects implant hardness. Here p_value for dentist and interaction between is greater than 0.05. So we fail to reject the null hypothesis. But the p_value for the method is much less than 0.05 proving method is a significant factor for implant hardness.

On basis of these two alloy types, we can conclude that method of the implant is a major factor in determining implant hardness.

To identify which pair differs we will conduct Multi-comparison for ALLOY_1 & ALLOY_2 separately.

For ALLOY_1 :

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 1.16

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 1.17

For ALLOY_2 :

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 1.18

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 1.19

Multiple Comparison of Means - Tukey HSD shows that for both alloys there is a difference between the means of method_1 and method_3 concerning implant hardness. Also, method_2 and method_3 show the difference in means. For all dentists, there is no difference in means for both alloys.

Education Post 12th Standard

Summary

The dataset consists of information about various university and colleges.We will be conducting EDA on various factors such as number of applications received,number of students enrolled etc.

Introduction

The data consists of 777 rows and 18 columns . There are 777 college and university and we will be analysing data based on the colleges.Column labels are renamed for better readability. Univariate ,bivariate analysis of the data will be conducted here based on the various factors.

Data description

1	University	Names of various university and colleges
2	Application received	Number of applications received
3	Application accepted	Number of applications accepted
4	Students enrolled	Number of new students enrolled
5	Top_10_perc_HSC	Percentage of new students from top 10% of Higher Secondary class
6	Top_25_perc_HSC	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	Out of state tuition	Number of students for whom the particular college or university is Out-of-state tuition
10	Cost of room and board	Cost of Room and board
11	Estimated book cost	Estimated book costs for a student
12	Personal spending	Estimated personal spending for a student
13	Fac_PhD	Percentage of faculties with Ph.D.’s
14	Fac_Terminal	Percentage of faculties with terminal degree
15	Stu_Fac Ratio	Student/faculty ratio
16	Alumni_donation_perc	Percentage of alumni who donate
17	Instructional expenditure	The Instructional expenditure per student
18	Graduation rate	Graduation rate

Sample of dataset

University	Application received	Application accepted	Students enrolled	Top_10_perc_HSC	Top_25_perc_HSC	F_Undergrad	P_Undergrad	Out of state tuition	Cost of room and board	Estimated book cost	Personal spending	Fac_PhD
Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70
Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29
Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53
Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92
Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76

Table 2.1

Problem-2

Exploratory Data Analysis

	Column	Non-Null Count	Data type
1	University	777 non-null	object
2	Application received	777 non-null	int64
3	Application accepted	777 non-null	int64
4	Students enrolled	777 non-null	int64
5	Top_10_perc_HSC	777 non-null	int64
6	Top_25_perc_HSC	777 non-null	int64
7	F_Undergrad	777 non-null	int64
8	P_Undergrad	777 non-null	int64
9	Out of state tuition	777 non-null	int64
10	Cost of room and board	777 non-null	int64
11	Estimated book cost	777 non-null	int64
12	Personal spending	777 non-null	int64
13	Fac_PhD	777 non-null	int64
14	Fac_Terminal	777 non-null	float64
15	Stu_Fac Ratio	777 non-null	int64

16	Alumni_donation_perc	777 non-null	int64
17	Instructional expenditure	777 non-null	int64
18	Graduation rate	777 non-null	int64

Descriptive statistics of data:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
University	777	777	Abilene Christian University	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Application received	777.0	NaN	NaN	NaN	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Application accepted	777.0	NaN	NaN	NaN	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Students enrolled	777.0	NaN	NaN	NaN	779.972973	929.17619	35.0	242.0	434.0	902.0	6392.0
Top_10_perc_HSC	777.0	NaN	NaN	NaN	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top_25_perc_HSC	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
Out of state tuition	777.0	NaN	NaN	NaN	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Cost of room and board	777.0	NaN	NaN	NaN	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Estimated book cost	777.0	NaN	NaN	NaN	549.380952	165.10536	96.0	470.0	500.0	600.0	2340.0
Personal spending	777.0	NaN	NaN	NaN	1340.642214	677.071454	250.0	850.0	1200.0	1700.0	6800.0
Fac_PhD	777.0	NaN	NaN	NaN	72.660232	16.328155	8.0	62.0	75.0	85.0	103.0
Fac_Terminal	777.0	NaN	NaN	NaN	79.702703	14.722359	24.0	71.0	82.0	92.0	100.0
Stu_Fac Ratio	777.0	NaN	NaN	NaN	14.089704	3.958349	2.5	11.5	13.6	16.5	39.8
Alumni_donation_perc	777.0	NaN	NaN	NaN	22.743887	12.391801	0.0	13.0	21.0	31.0	64.0
Instructional expenditure	777.0	NaN	NaN	NaN	9660.171171	5221.76844	3186.0	6751.0	8377.0	10830.0	56233.0
Graduation rate	777.0	NaN	NaN	NaN	65.46332	17.17771	10.0	53.0	65.0	78.0	118.0

Table 2.2

Descriptive statistics of data shows that there are 777 rows in data set.There are 777 colleges in the provided data.Null values are not present in dataset.All the columns except names are numerical in nature.

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

Univariate analysis

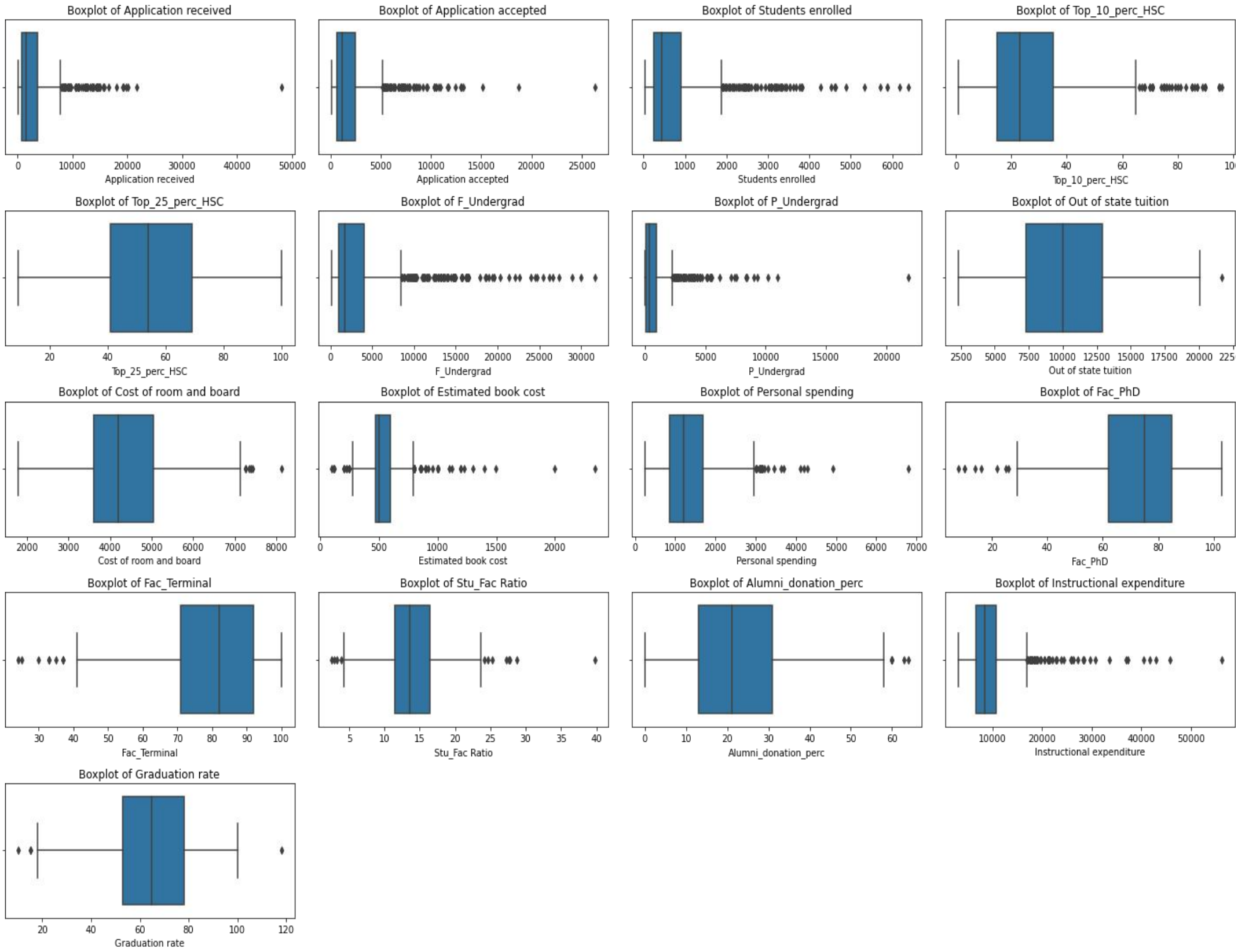


Fig 2.1

Boxplots for all the numerical data are shown above. From the boxplot, it is evident that all the columns except the column showing percentage of new students from the top 25% of Higher Secondary classes have outliers. Also, most of the columns are right-skewed.Top_25_perc_HSC, Out of state tuition, Stu_Fac Ratio, and Graduation rate seem to follow a normal distribution.Fac_PhD and Fac_Terminal columns are left skewed. Columns Estimated book cost, Fac_PhD, Fac_Terminal, Stu_Fac Ratio, and Graduation rate have data points below the calculated minimum value, ie these data are below Q1-(1.5*IQR). All other outliers are above the upper limit, ie Q3+(1.5*IQR).

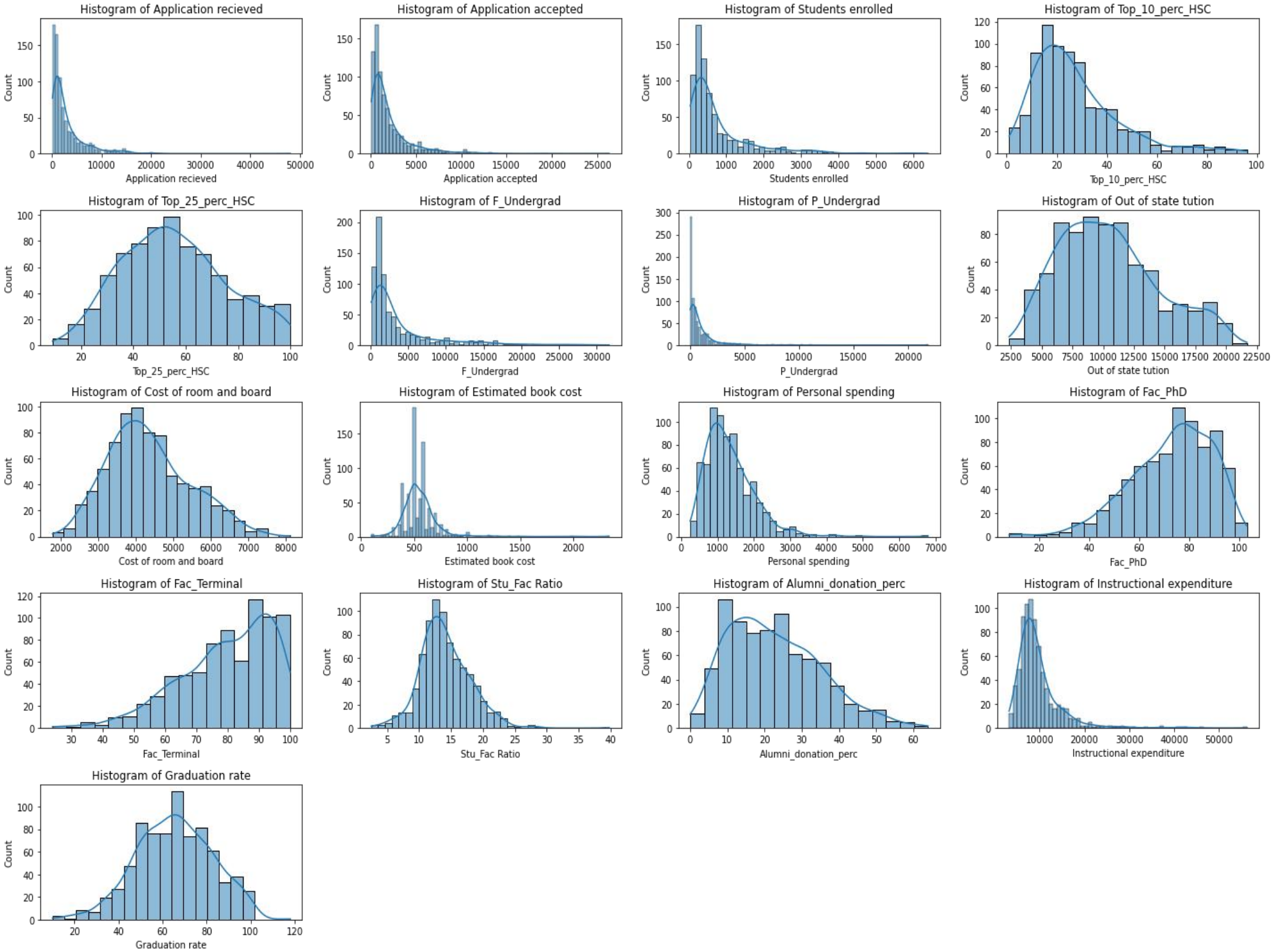


Fig 2.2

Histograms are plotted for all the numerical columns as shown above. It is evident from the histogram that Top_25_perc_HSC, Out of state tuition, Stu_Fac Ratio, and Graduation rate graphically follow a normal distribution. Here most columns asymptote to the x-axis. This indicates the presence of outliers in the data. Also, it represents the data is skewed to right.

Bivariate & Multi-variate analysis

Listing top 10 college based on columns all numeric columns.Below is an sample data with most number of application accepted.

University	Application received
Rutgers at New Brunswick	48094
Purdue University at West Lafayette	21804
Boston University	20192
University of California at Berkeley	19873
Pennsylvania State Univ. Main Campus	19315
University of Michigan at Ann Arbor	19152
Michigan State University	18114
Indiana University at Bloomington	16587
University of Virginia	15849
Virginia Tech	15712

Table 2.3

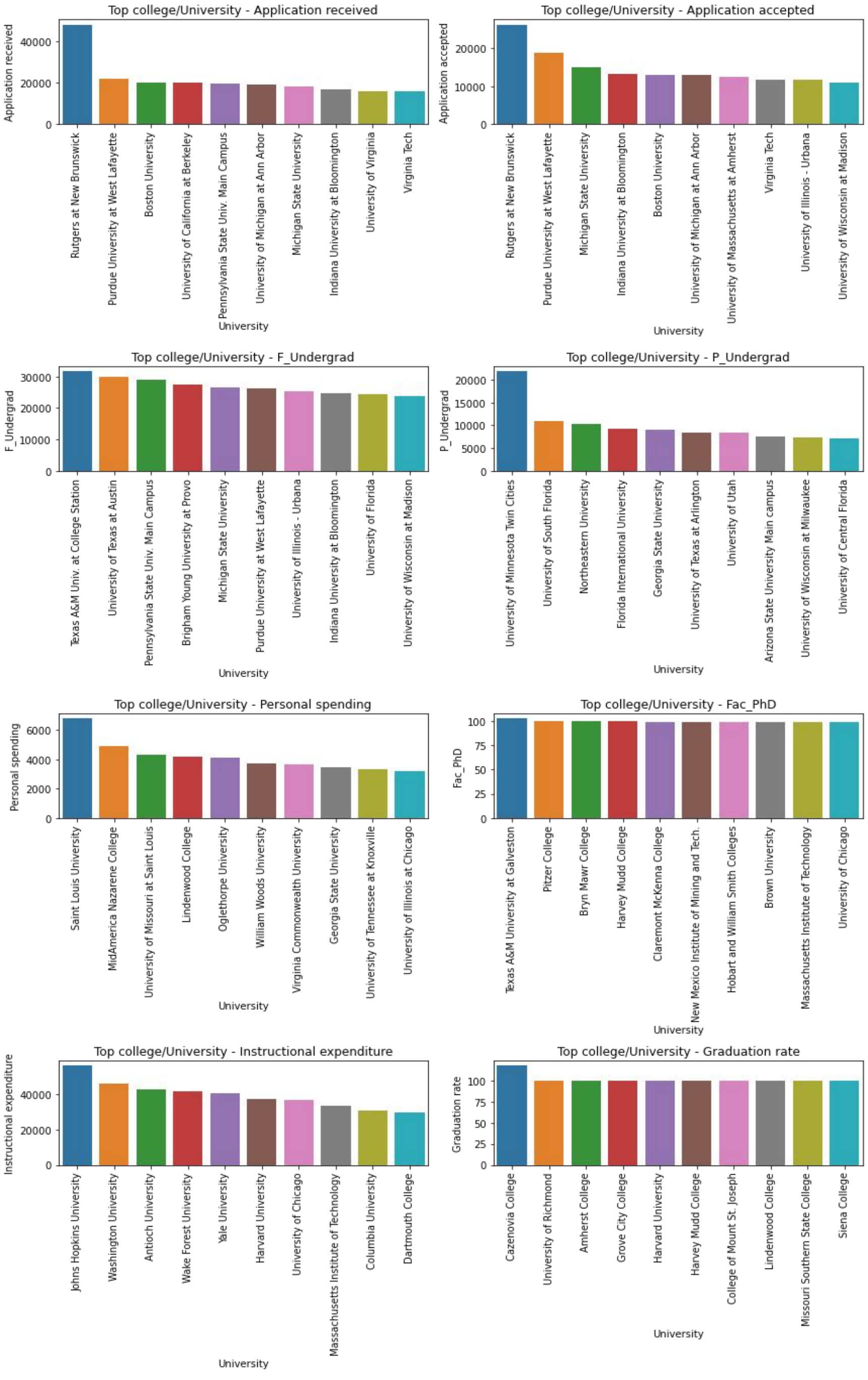
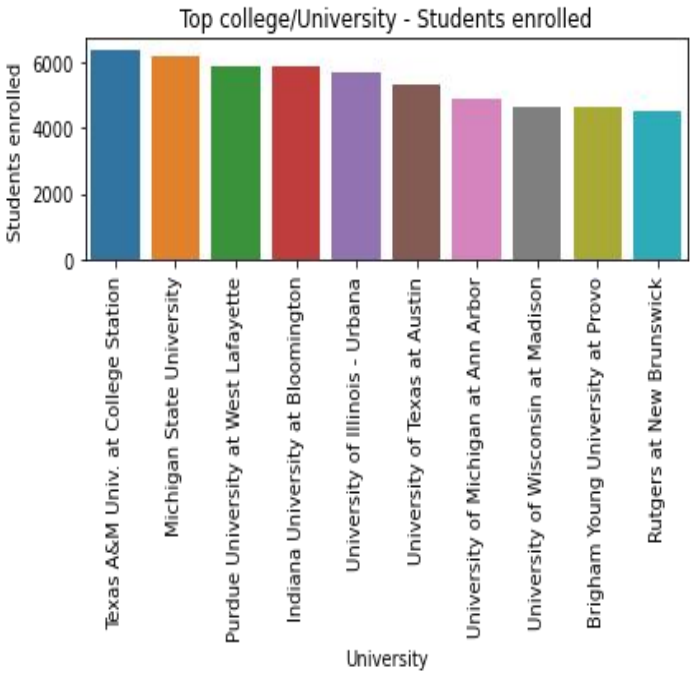
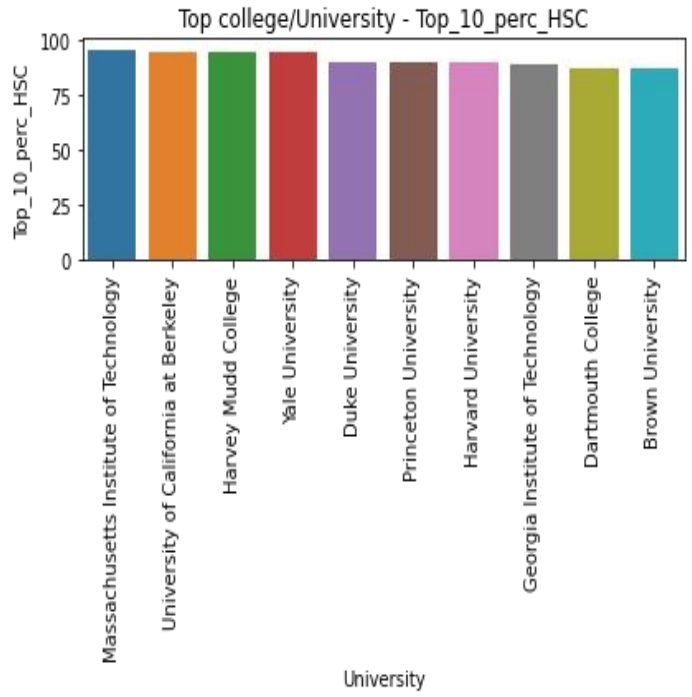
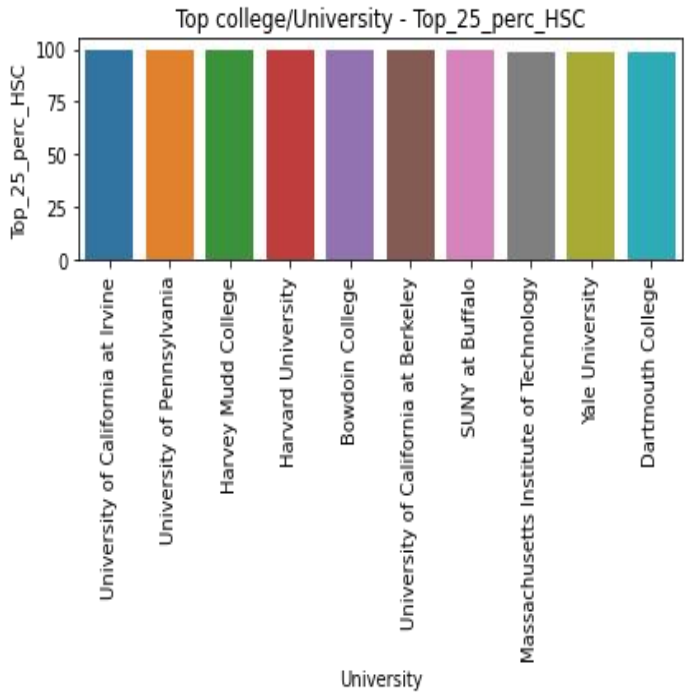
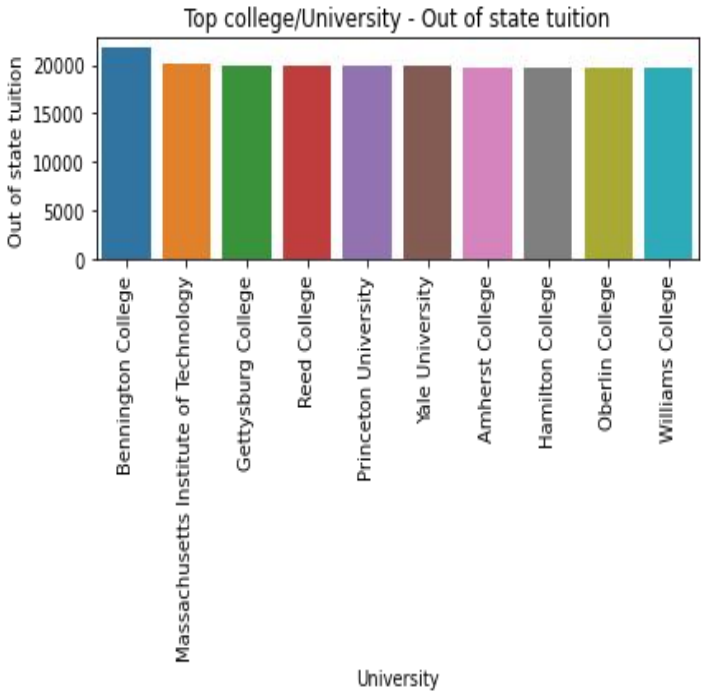

Top college/University - Application received
Application received
University
Rutgers at New Brunswick
Purdue University at West Lafayette
Boston University
University of California at Berkeley
Pennsylvania State Univ. Main Campus
University of Michigan at Ann Arbor
Michigan State University
Indiana University at Bloomington
University of Virginia
Virginia Tech
Top college/University - Application accepted
Application accepted
University
Rutgers at New Brunswick
Purdue University at West Lafayette
Michigan State University
Indiana University at Bloomington
Boston University
University of Michigan at Ann Arbor
University of Massachusetts at Amherst
Virginia Tech
University of Illinois - Urbana
University of Wisconsin at Madison
Top college/University - F_Undergrad
F_Undergrad
University
Texas A&M Univ. at College Station
University of Texas at Austin
Pennsylvania State Univ. Main Campus
Brigham Young University at Provo
Michigan State University
Purdue University at West Lafayette
University of Illinois - Urbana
Indiana University at Bloomington
University of Florida
University of Wisconsin at Madison
Top college/University - P_Undergrad
P_Undergrad
University
University of Minnesota Twin Cities
University of South Florida
Northeastern University
Florida International University
Georgia State University
University of Texas at Arlington
University of Utah
Arizona State University Main campus
University of Wisconsin at Milwaukee
University of Central Florida
Top college/University - Personal spending
Personal spending
University
Saint Louis University
MidAmerica Nazarene College
University of Missouri at Saint Louis
Lindenwood College
Oglethorpe University
William Woods University
Virginia Commonwealth University
Georgia State University
University of Tennessee at Knoxville
University of Illinois at Chicago
Top college/University - Fac_PhD
Fac_PhD
University
Texas A&M University at Galveston
Pitzer College
Bryn Mawr College
Harvey Mudd College
Claremont McKenna College
Mexico Institute of Mining and Tech.
Hobart and William Smith Colleges
Brown University
assachusetts Institute of Technology
University of Chicago
Top college/University - Instructional expenditure
Instructional expenditure
University
Johns Hopkins University
Washington University
Antioch University
Wake Forest University
Yale University
Harvard University
University of Chicago
Massachusetts Institute of Technology
Columbia University
Dartmouth College
Top college/University - Graduation rate
Graduation rate
University
Cazenovia College
University of Richmond
Amherst College
Grove City College
Harvard University
Harvey Mudd College
College of Mount St. Joseph
Lindenwood College
Missouri Southern State College
Siena College

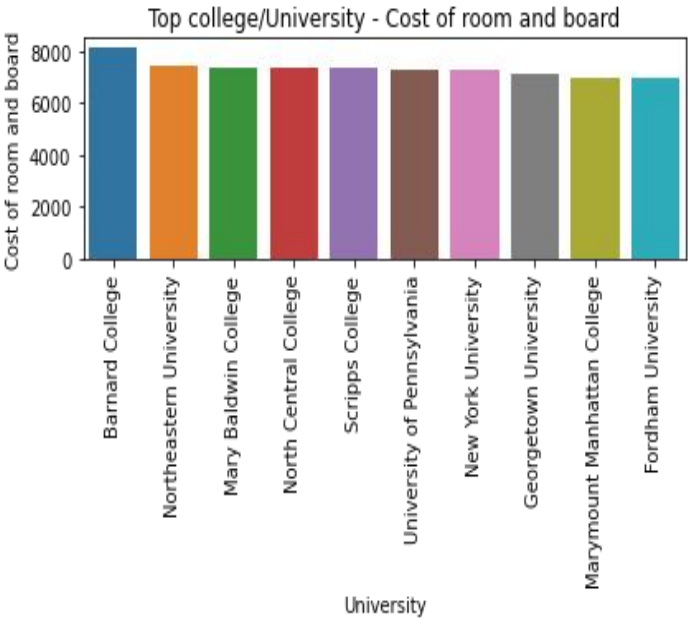
Fig 2.3

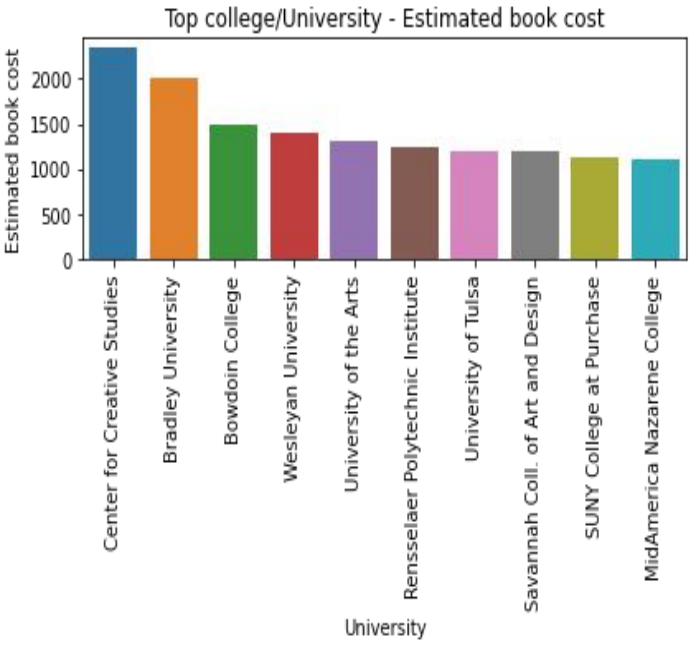

Top college/University - Students enrolled
Students enrolled
0
2000
4000
6000
Texas A&M Univ. at College Station
Michigan State University
Purdue University at West Lafayette
Indiana University at Bloomington
University of Illinois - Urbana
University of Texas at Austin
University of Michigan at Ann Arbor
University of Wisconsin at Madison
Brigham Young University at Provo
Rutgers at New Brunswick
University

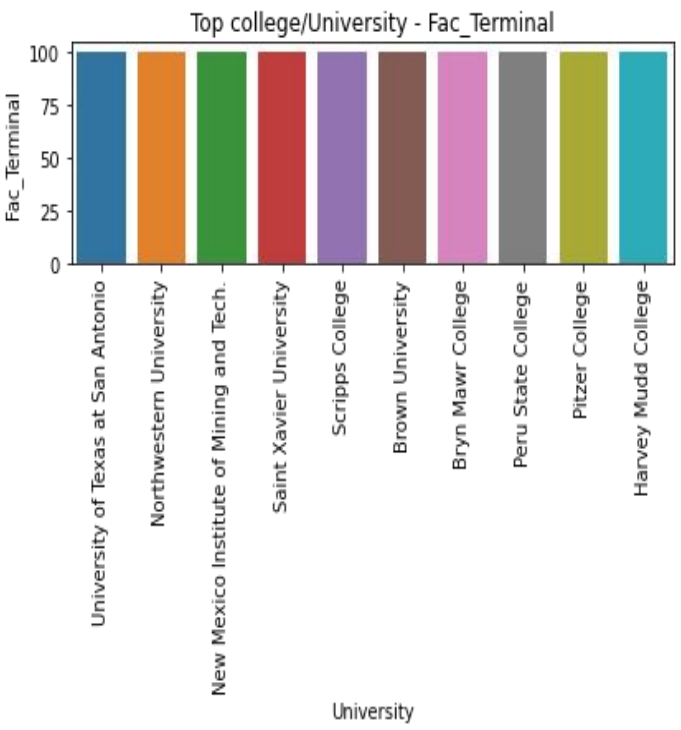

Top college/University - Top_10_perc_HSC
Top_10_perc_HSC
0
25
50
75
100
Massachusetts Institute of Technology
University of California at Berkeley
Harvey Mudd College
Yale University
Duke University
Princeton University
Harvard University
Georgia Institute of Technology
Dartmouth College
Brown University
University

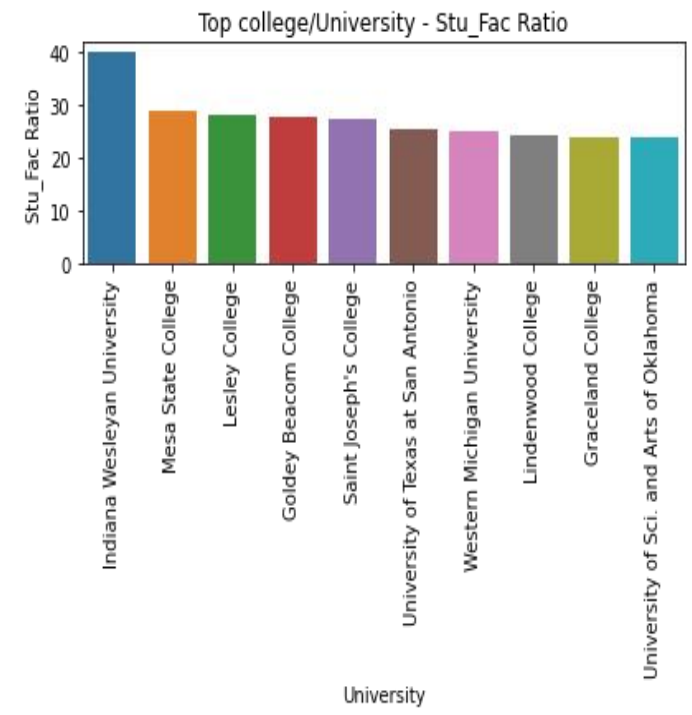

Top college/University - Top_25_perc_HSC
Top_25_perc_HSC
0
25
50
75
100
University of California at Irvine
University of Pennsylvania
Harvey Mudd College
Harvard University
Bowdoin College
University of California at Berkeley
SUNY at Buffalo
Massachusetts Institute of Technology
Yale University
Dartmouth College
University


Top college/University - Out of state tuition
Out of state tuition
0
5000
10000
15000
20000
Bennington College
Massachusetts Institute of Technology
Gettysburg College
Reed College
Princeton University
Yale University
Amherst College
Hamilton College
Oberlin College
Williams College
University


Top college/University - Cost of room and board
Cost of room and board
0
2000
4000
6000
8000
Barnard College
Northeastern University
Mary Baldwin College
North Central College
Scripps College
University of Pennsylvania
New York University
Georgetown University
Marymount Manhattan College
Fordham University
University


Top college/University - Estimated book cost
Estimated book cost
0
500
1000
1500
2000
Center for Creative Studies
Bradley University
Bowdoin College
Wesleyan University
University of the Arts
Rensselaer Polytechnic Institute
University of Tulsa
Savannah Coll. of Art and Design
SUNY College at Purchase
MidAmerica Nazarene College
University


Top college/University - Fac_Terminal
Fac_Terminal
0
25
50
75
100
University of Texas at San Antonio
Northwestern University
Colorado School of Mining and Tech.
Saint Xavier University
Scripps College
Brown University
Bryn Mawr College
Peru State College
Pitzer College
Harvey Mudd College
University


Top college/University - Stu_Fac Ratio
Stu_Fac Ratio
0
10
20
30
40
Indiana Wesleyan University
Mesa State College
Lesley College
Goldey Beacom College
Saint Joseph's College
University of Texas at San Antonio
Western Michigan University
Lindenwood College
Graceland College
University of Sci. and Arts of Oklahoma
University

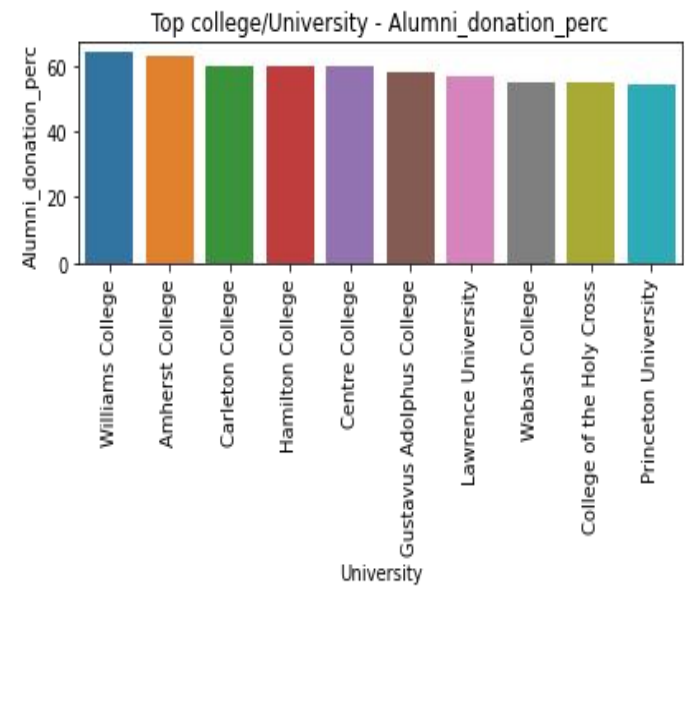

Top college/University - Alumni_donation_perc
Alumni_donation_perc
0
20
40
60
Williams College
Amherst College
Carleton College
Hamilton College
Centre College
Gustavus Adolphus College
Lawrence University
Wabash College
College of the Holy Cross
Princeton University
University

Fig 2.4

We have selected 10 colleges based on the data provided on y-axis. All of the numeric columns are analyzed based on these colleges. Rutgers at New Brunswick university received the most number of applications and they accepted the most number of students. Virginia Tech receives the least number of applications of the 10 colleges we have selected.Texas A&M Univ. at College Station has the most number of students enrolled followed by Michigan State University. Rutgers at New is the least of the 10 colleges where new students are enrolled. The top 10% of the Higher Secondary class are from MIT followed by University of California at Berkeley. The most number of full-time undergraduate students are from Texas A&M Univ. at College Station .University of Minnesota Twin Cities has the most number of part-time undergraduate students, and University of Central Florida has the least number. The cost of room and board is highest in Barnard college, but estimated book cost for students is more at center for creative studies. Students at Saint Louis University seem to spend more on personal spending comparing other students in the selected colleges. The percentage of faculties with Ph.D. and terminal degree is approximately equal in top 10 Colleges. Student/faculty ratio is better at Indiana Wesleyan University followed by Mesa State College .The main aspect of colleges is their graduation rate.Cazenovia College has the highest graduation rate all other colleges in the top 10 list have same graduation rate.

Pairplot of all the columns in dataset:

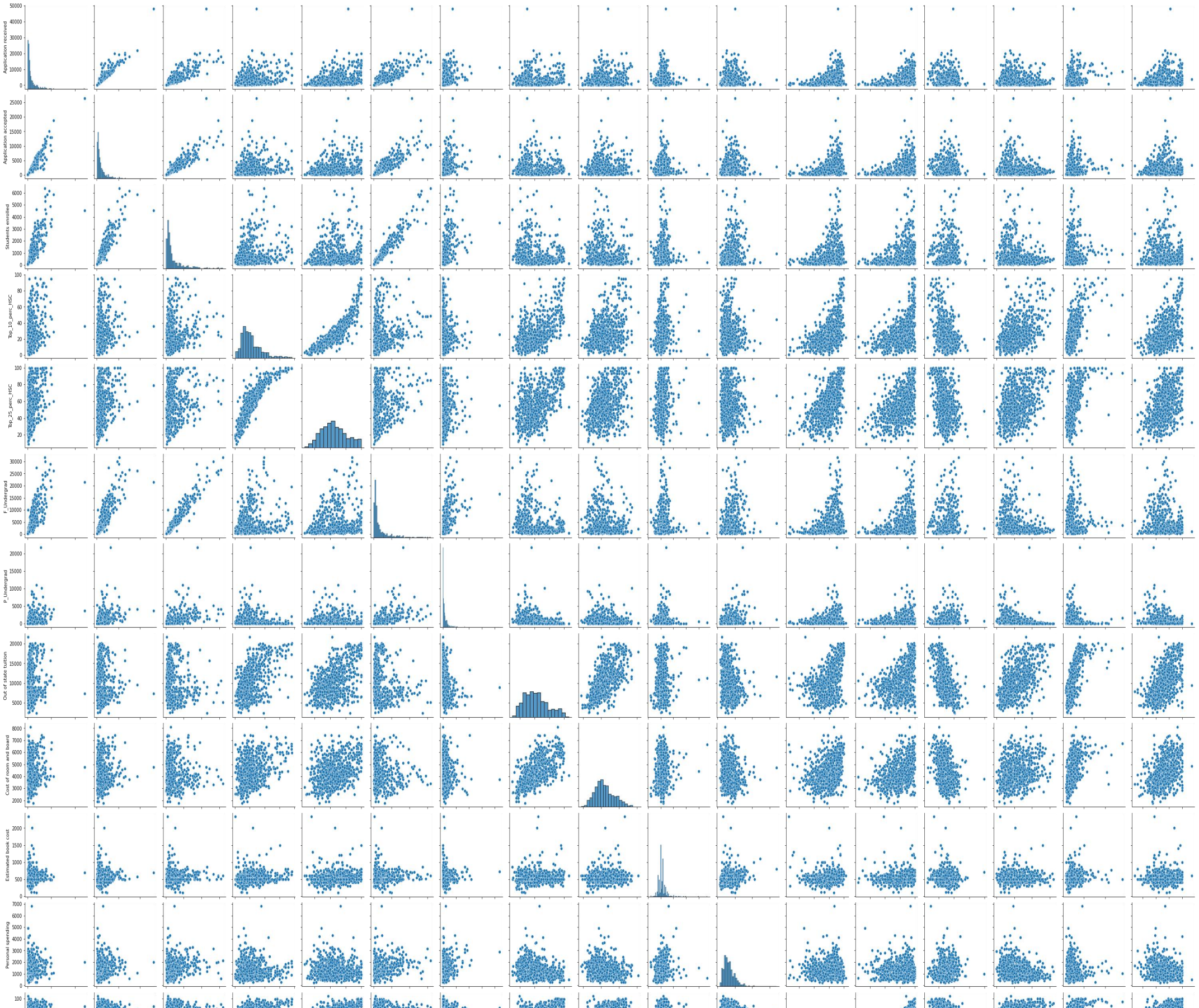

Application received
Application accepted
Students enrolled
Top_10_perc_HSC
Top_25_perc_HSC
F_Undergrad
P_Undergrad
Out of state tuition
Cost of room and board
Estimated book cost
Personal spending
Fac_PhD
Fac_Terminal
Stu_Fac Ratio
Alumni_donation_perc
Instructional expenditure
Graduation rate

Fig 2.5

Heat map to show correlation of all the columns in dataset:

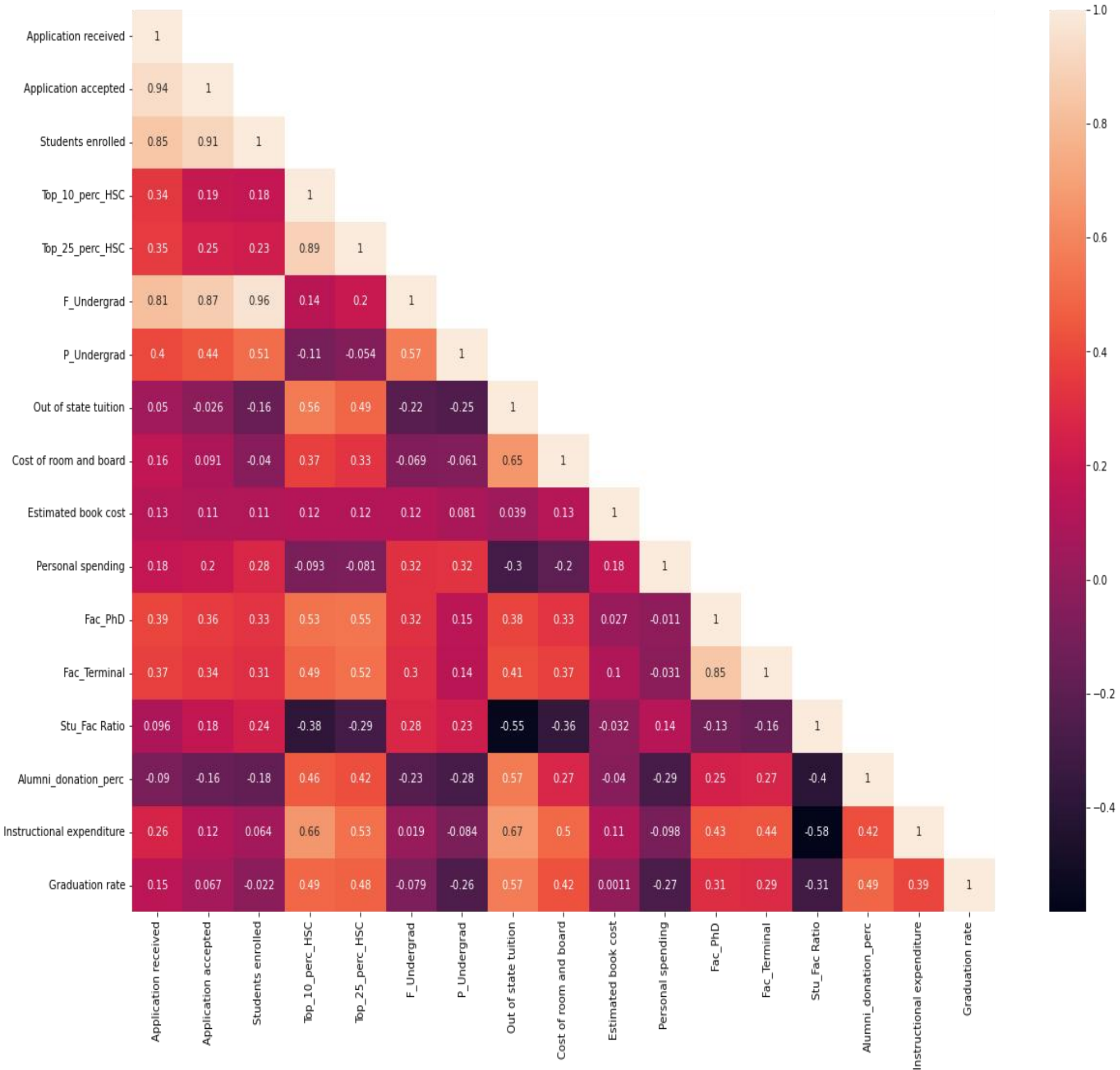


Fig 2.6

The heat map of the dataset shows that there is a correlation between various columns. Application received, accepted and students enrolled show high collinearity. ie When the number of applications received increases the number of students who get accepted increases and the number of students enrolled increases. The number of full-time undergraduate students is also highly correlated with the Application received, accepted and students enrolled factors. But the number of part-time undergraduate students has medium collinearity with these factors. 0.7 to 0.9 is considered a high correlation, 0.5 to 0.7 is moderately correlated and below 0.3 to 0.5 indicate variables that have a low correlation.

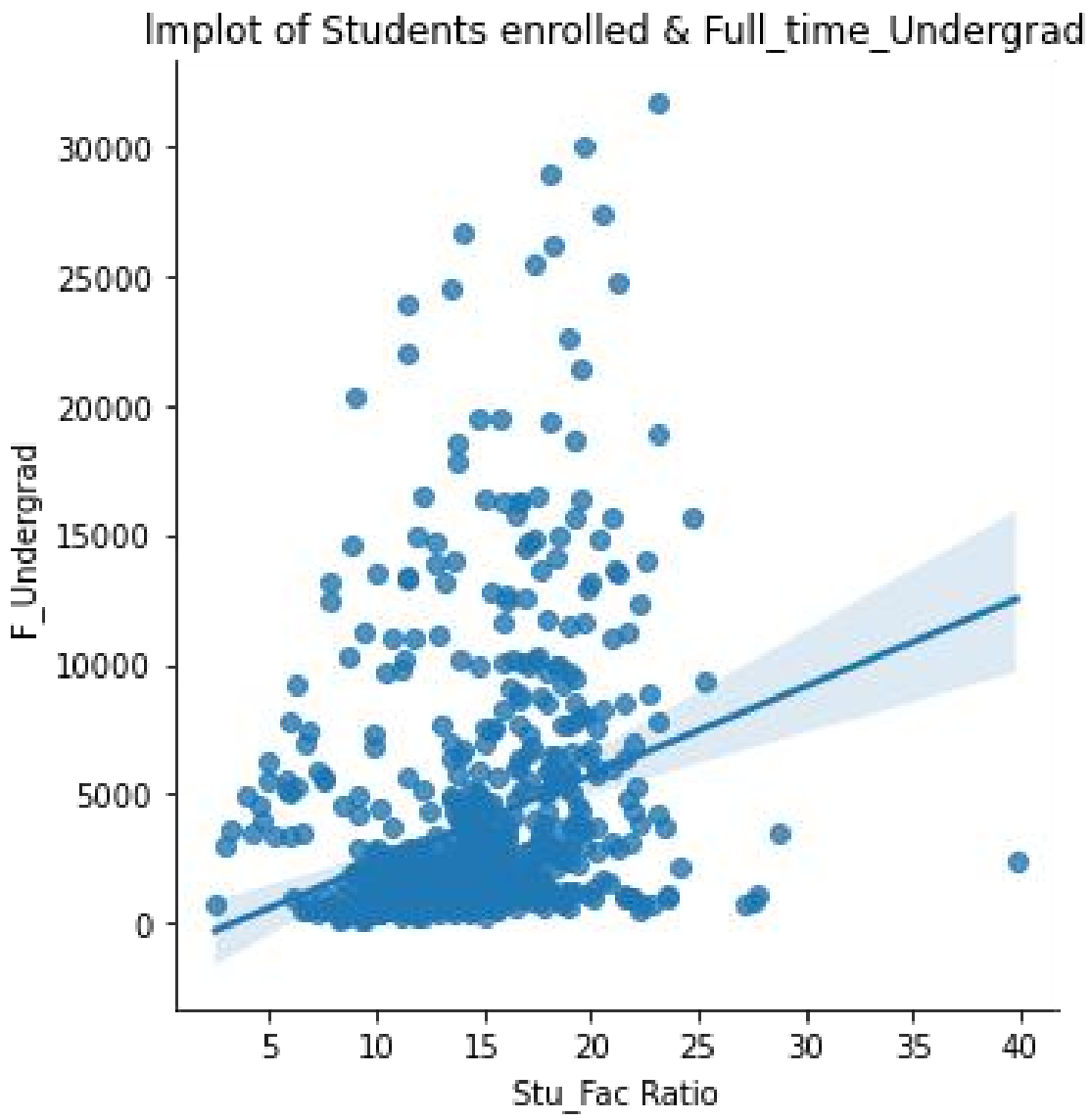


Fig 2.7

This is a clear representation of collinearity between Students enrolled & full-time Under gradate students. Similarly, there will be a negative correlation between the columns, ie one factor increases, and the other factor decreases.Below shown is a negative correlation plot.

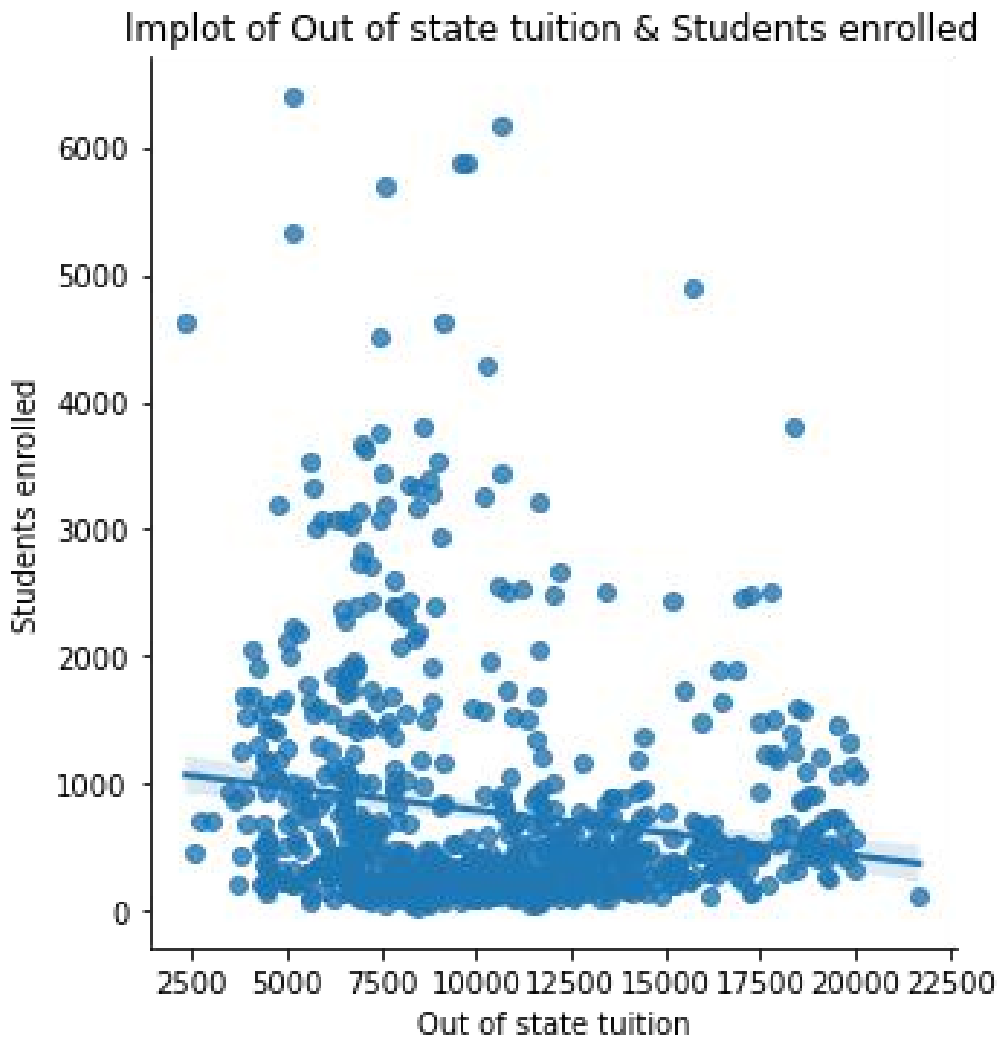


Fig 2.8

Multivariate analysis of Application received,Application accepted & Students enrolled:

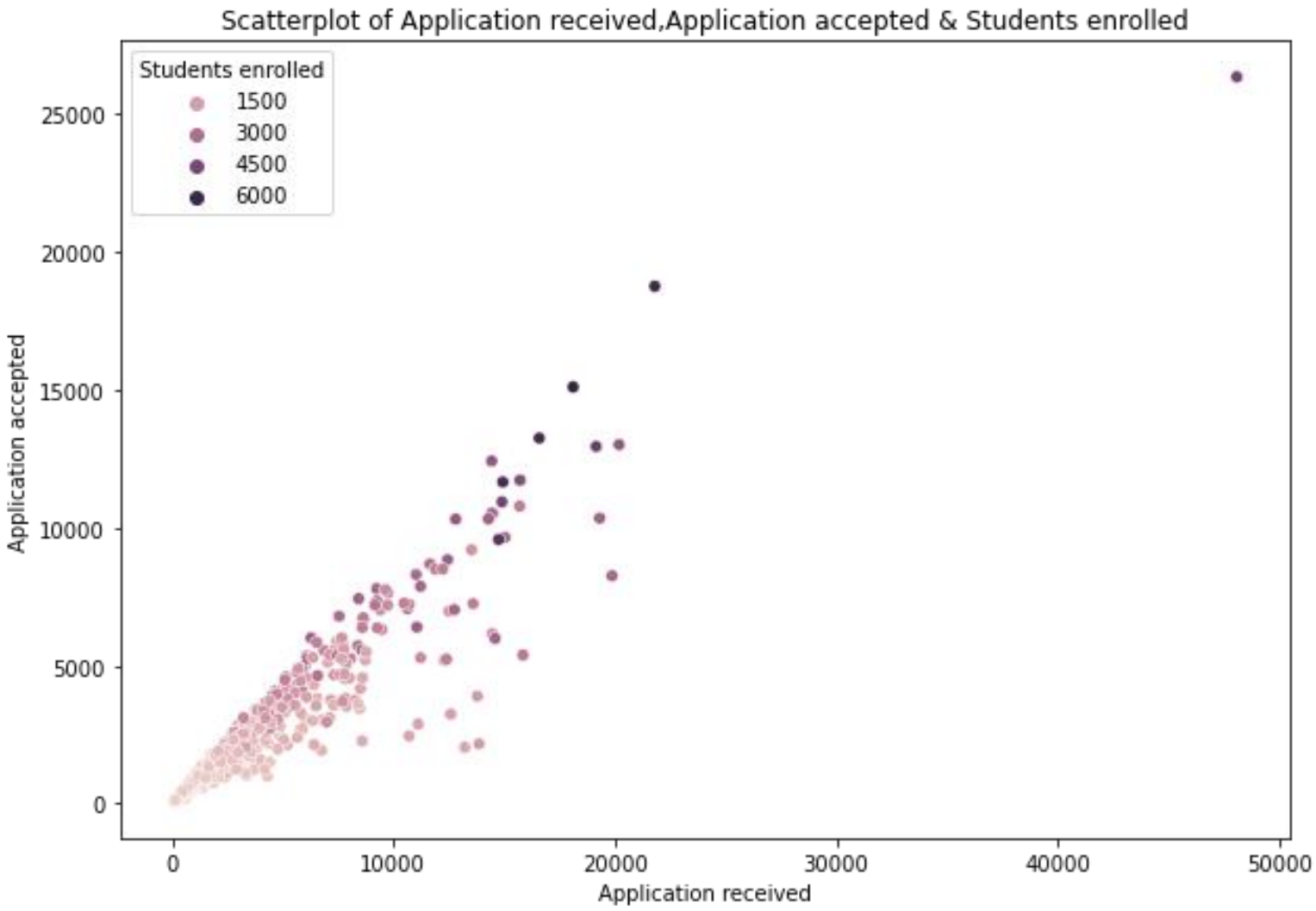


Fig 2.9

When three factors such as Application received,Application accepted & Students enrolled are considered and based on this scatter plot is plotted we will obtain the above graph.It is evident from the heat map that these columns follow high correlation thus the data points in plot are not scattered along the x or y axis of the plot.

India Data Census

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 Labourers, (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 House less 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.

Introduction

There are 640 rows and 61 columns in dataset.The data collected has so many variables.We will perform detailed EDA and identify Optimum Principal Components that explains the most variance in data.

Data description

1	State Code	State Code
2	Dist.Code	District Code
3	State	State name
4	Area Name	Area name in each state
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 Labourers 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 dataset

State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F
Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	0	1999	...	1150	749	180	237
Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	6	427	...	525	715	123	229
Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	6	5806	...	114	188	44	89
Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	0	2666	...	194	247	61	128
Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	33	7670	...	874	1928	465	1043
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Puducherry	Mahe	3333	8154	11781	1146	1203	21	30	0	...	32	47	0	0
Puducherry	Karaikal	10612	12346	21691	1544	1533	2234	4155	0	...	155	337	3	14
Andaman & Nicobar Island	Nicobars	1275	1549	2630	227	225	0	0	1012	...	104	134	9	4
Andaman & Nicobar Island	North & Middle Andaman	3762	5200	8012	723	664	0	0	28	...	136	172	24	44
Andaman & Nicobar Island	South Andaman	7975	11977	18049	1470	1358	0	0	161	...	173	122	6	2

Table 3.1

Columns named State Code and Dist.Code are removed.These columns are not useful in analysis of the data.Now the data set consists of 640 rows and 59 columns.

Problem-3

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

Checking for null values:

	Column	Non-Null Count	Data type
1	State	640 non-null	object
2	Area Name	640 non-null	object
3	No_HH	640 non-null	int64
4	TOT_M	640 non-null	int64
5	TOT_F	640 non-null	int64
6	M_06	640 non-null	int64
7	F_06	640 non-null	int64
8	M_SC	640 non-null	int64
9	F_SC	640 non-null	int64
10	M_ST	640 non-null	int64
11	F_ST	640 non-null	int64
12	M_LIT	640 non-null	int64
13	F_LIT	640 non-null	int64
14	M_ILL	640 non-null	int64
15	F_ILL	640 non-null	int64
16	TOT_WORK_M	640 non-null	int64
17	TOT_WORK_F	640 non-null	int64
18	MAINWORK_M	640 non-null	int64
19	MAINWORK_F	640 non-null	int64
20	MAIN_CL_M	640 non-null	int64
21	MAIN_CL_F	640 non-null	int64
22	MAIN_AL_M	640 non-null	int64

23	MAIN_AL_F	640 non-null	int64
24	MAIN_HH_M	640 non-null	int64
25	MAIN_HH_F	640 non-null	int64
26	MAIN_OT_M	640 non-null	int64
27	MAIN_OT_F	640 non-null	int64
28	MARGWORK_M	640 non-null	int64
29	MARGWORK_F	640 non-null	int64
30	MARG_CL_M	640 non-null	int64
31	MARG_CL_F	640 non-null	int64
32	MARG_AL_M	640 non-null	int64
33	MARG_AL_F	640 non-null	int64
34	MARG_HH_M	640 non-null	int64
35	MARG_HH_F	640 non-null	int64
36	MARG_OT_M	640 non-null	int64
37	MARG_OT_F	640 non-null	int64
38	MARGWORK_3_6_M	640 non-null	int64
39	MARGWORK_3_6_F	640 non-null	int64
40	MARG_CL_3_6_M	640 non-null	int64
41	MARG_CL_3_6_F	640 non-null	int64
42	MARG_AL_3_6_M	640 non-null	int64
43	MARG_AL_3_6_F	640 non-null	int64
44	MARG_HH_3_6_M	640 non-null	int64
45	MARG_HH_3_6_F	640 non-null	int64
46	MARG_OT_3_6_M	640 non-null	int64
47	MARG_OT_3_6_F	640 non-null	int64
48	MARGWORK_0_3_M	640 non-null	int64
49	MARGWORK_0_3_F	640 non-null	int64
50	MARG_CL_0_3_M	640 non-null	int64

51	MARG_CL_0_3_F	640 non-null	int64
52	MARG_AL_0_3_M	640 non-null	int64
53	MARG_AL_0_3_F	640 non-null	int64
54	MARG_HH_0_3_M	640 non-null	int64
55	MARG_HH_0_3_F	640 non-null	int64
56	MARG_OT_0_3_M	640 non-null	int64
57	MARG_OT_0_3_F	640 non-null	int64
58	NON_WORK_M	640 non-null	int64
59	NON_WORK_F	640 non-null	int64

Dataset head :

	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F
0	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	0	1999	...	1150	749	180	237
1	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	6	427	...	525	715	123	229
2	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	6	5806	...	114	188	44	89
3	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	0	2666	...	194	247	61	128
4	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	33	7670	...	874	1928	465	1043

5 rows × 59 columns

Table 3.2

There are no null values or duplicate values in the dataset.

Descriptive statistics of data:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
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 3.3

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**

Selected columns are :

1.Number of Household - No_HH

2.Total population Male - TOT_M

3.Total population Female - TOT_F

4.Literates population Male - M_LIT

5.Literates population Female - F_LIT

10.1 Compare Number of Household,Total population Male,Total population Female, Literates population Male and Literates population Female.

Fig 3.1

It is evident from the box plot that there are outliers in the selected columns of data.Also, all the columns selected have right-skewed data.While the total population of males and females is compared it is observed that there are more females than males.The average population of females is more than males in India.The number of literate males and females seems to be approximately equal,but closely comparing these two it is evident that there are slightly more female literates than males.

To show the distribution of data we can plot histogram as shown below:

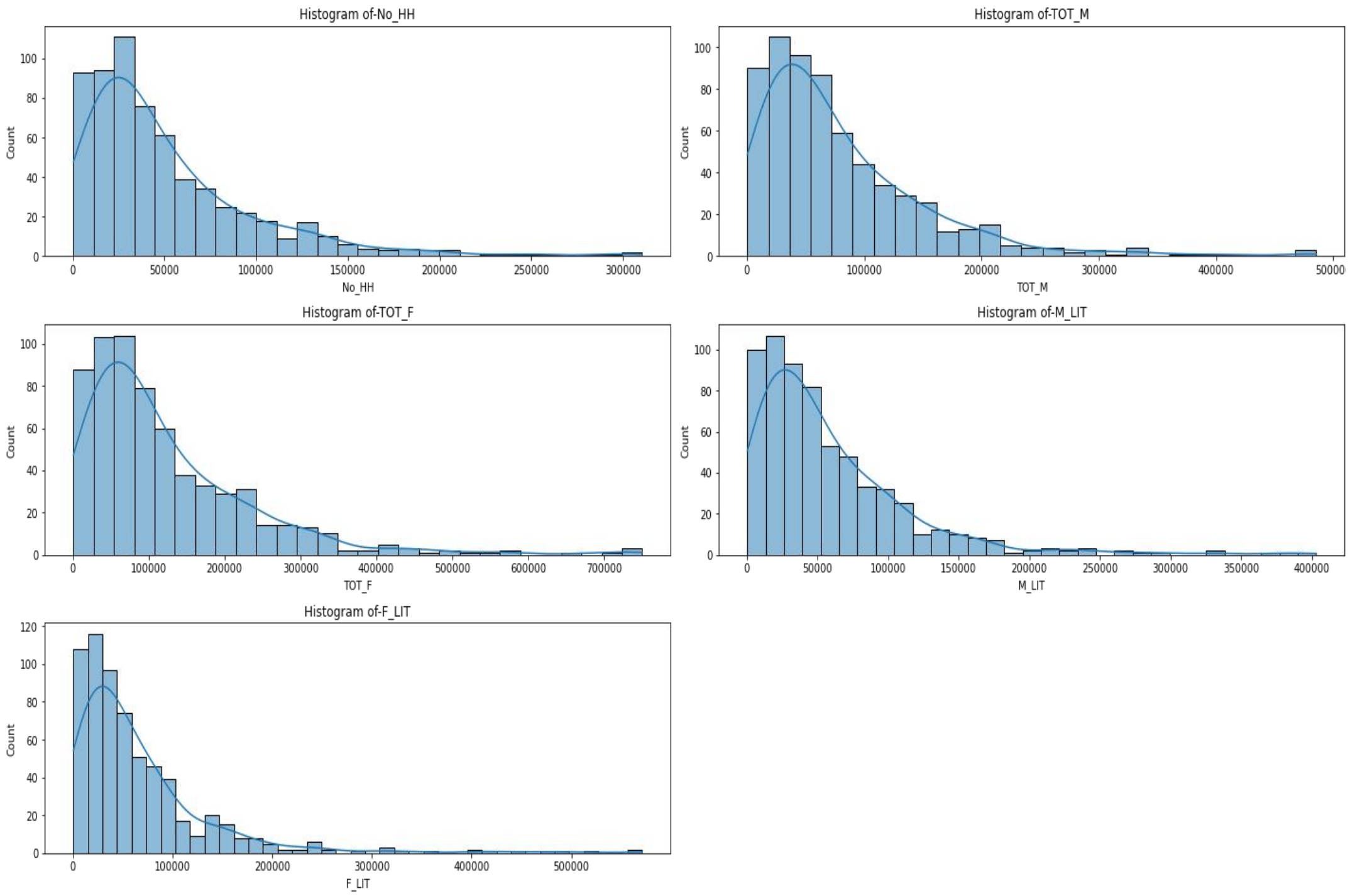


Fig 3.2

We have plotted and compared the data within each other .To analyse the data further we will conduct bi-variate analysis on these columns based on the states.

10.2 Which state has the highest number of house holds.?Find the total and average number of house holds.

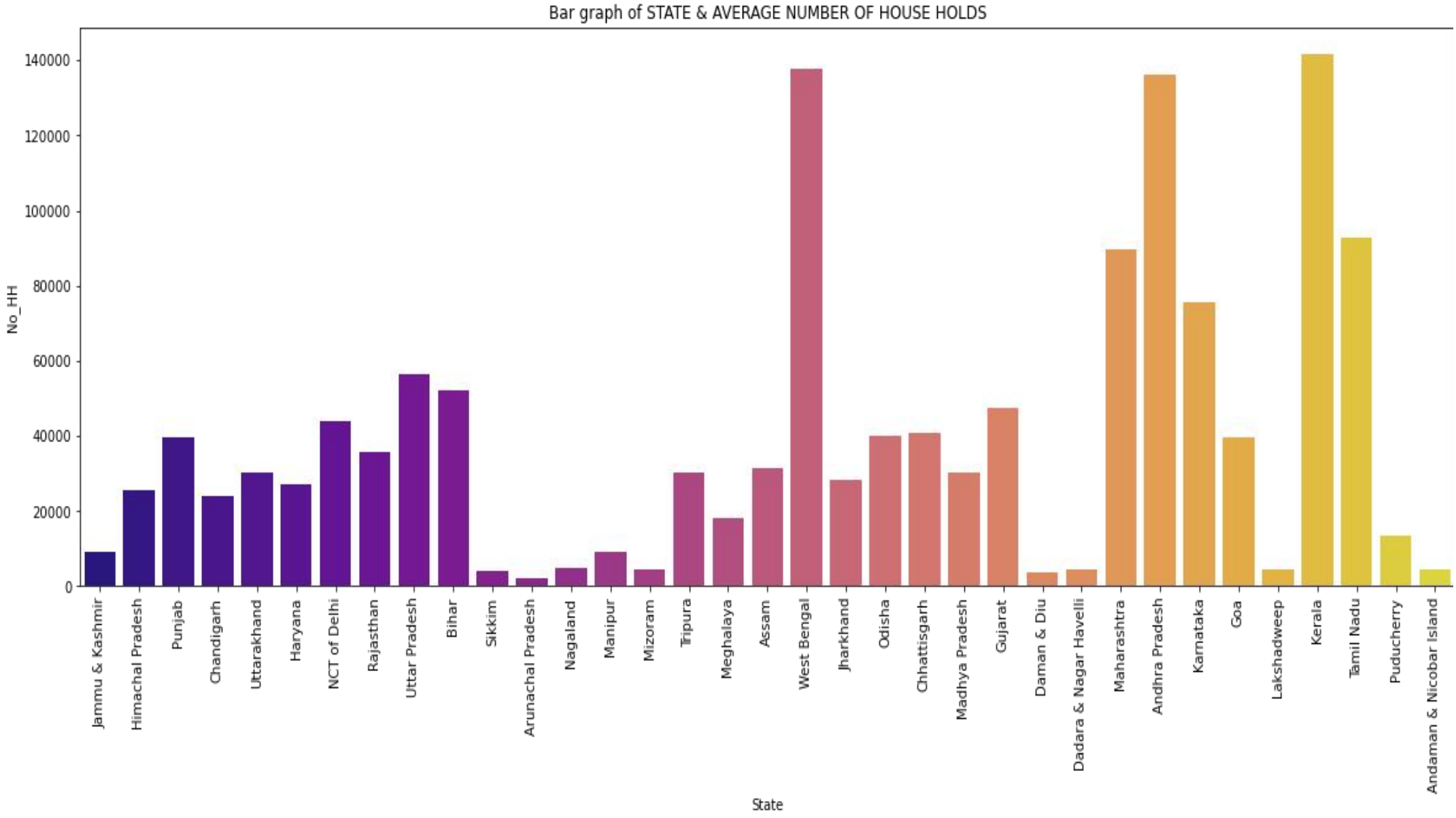

Bar graph of STATE & AVERAGE NUMBER OF HOUSE HOLDS
No_HH
140000
120000
100000
80000
60000
40000
20000
0
Jammu & Kashmir
Himachal Pradesh
Punjab
Chandigarh
Uttarakhand
Haryana
NCT of Delhi
Rajasthan
Uttar Pradesh
Bihar
Sikkim
Arunachal Pradesh
Nagaland
Manipur
Mizoram
Tripura
Meghalaya
Assam
West Bengal
Jharkhand
Odisha
Chhattisgarh
Madhya Pradesh
Gujarat
Daman & Diu
Dadara & Nagar Havelli
Maharashtra
Andhra Pradesh
Karnataka
Goa
Lakshadweep
Kerala
Tamil Nadu
Puducherry
Andaman & Nicobar Island
State

Fig 3.3

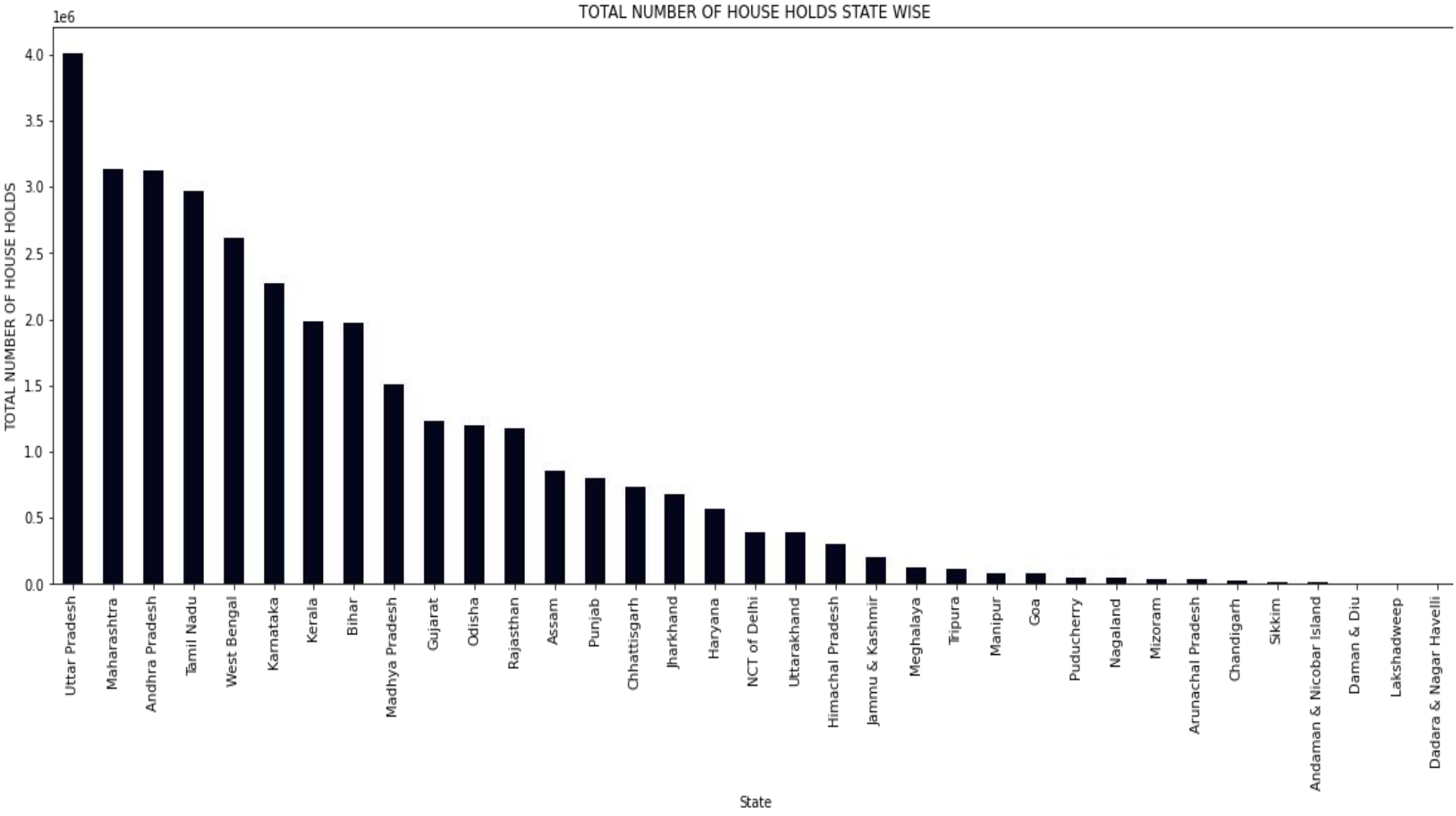

TOTAL NUMBER OF HOUSE HOLDS STATE WISE
1e6
TOTAL NUMBER OF HOUSE HOLDS
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Uttar Pradesh
Maharashtra
Andhra Pradesh
Tamil Nadu
West Bengal
Karnataka
Kerala
Bihar
Madhya Pradesh
Gujarat
Odisha
Rajasthan
Assam
Punjab
Chhattisgarh
Jharkhand
Haryana
NCT of Delhi
Uttarakhand
Himachal Pradesh
Jammu & Kashmir
Meghalaya
Tripura
Manipur
Goa
Puducherry
Nagaland
Mizoram
Arunachal Pradesh
Chandigarh
Sikkim
Andaman & Nicobar Island
Daman & Diu
Lakshadweep
Dadara & Nagar Havelli
State

Fig 3.4

On average most households are present in Kerala followed by Assam and Andhra Pradesh. The least average of households is in Arunachal Pradesh. Based on the total number of households in a state Uttar Pradesh has the highest and Dadara & Nagar Haveli has the least.

10.3 Which state has the highest male and female population ?Find the total and average number of population in each state.

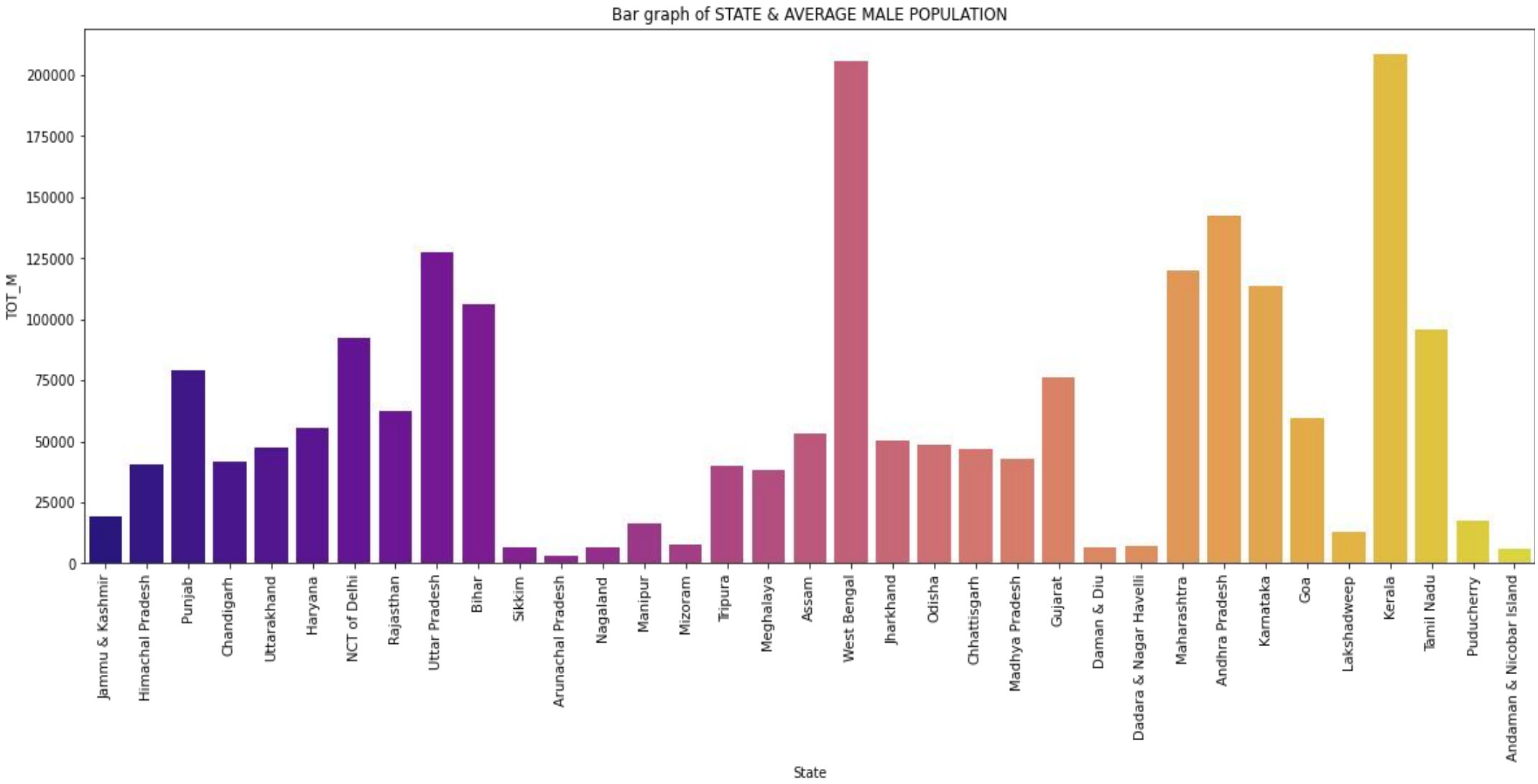


Fig 3.5

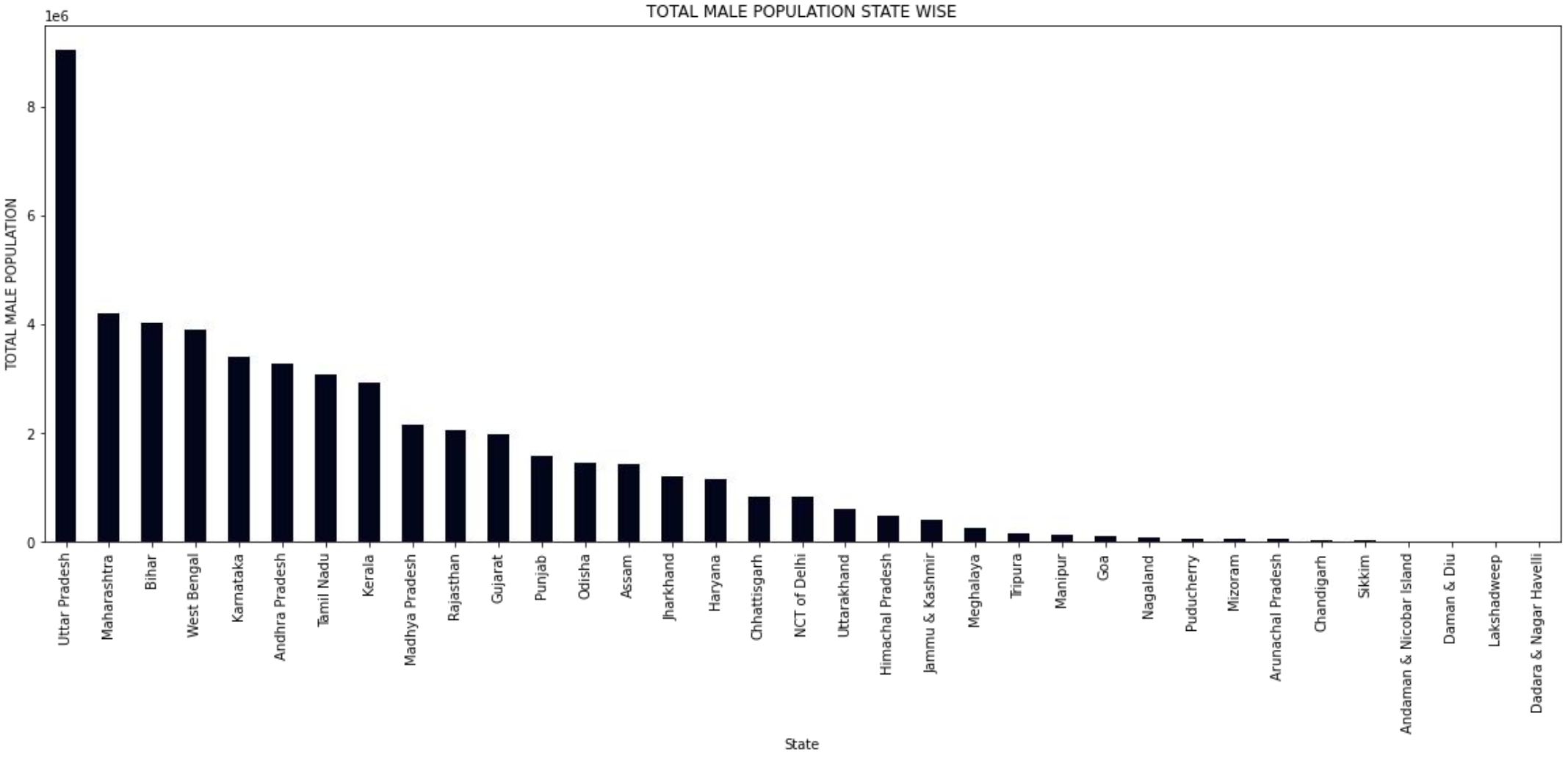


Fig 3.6

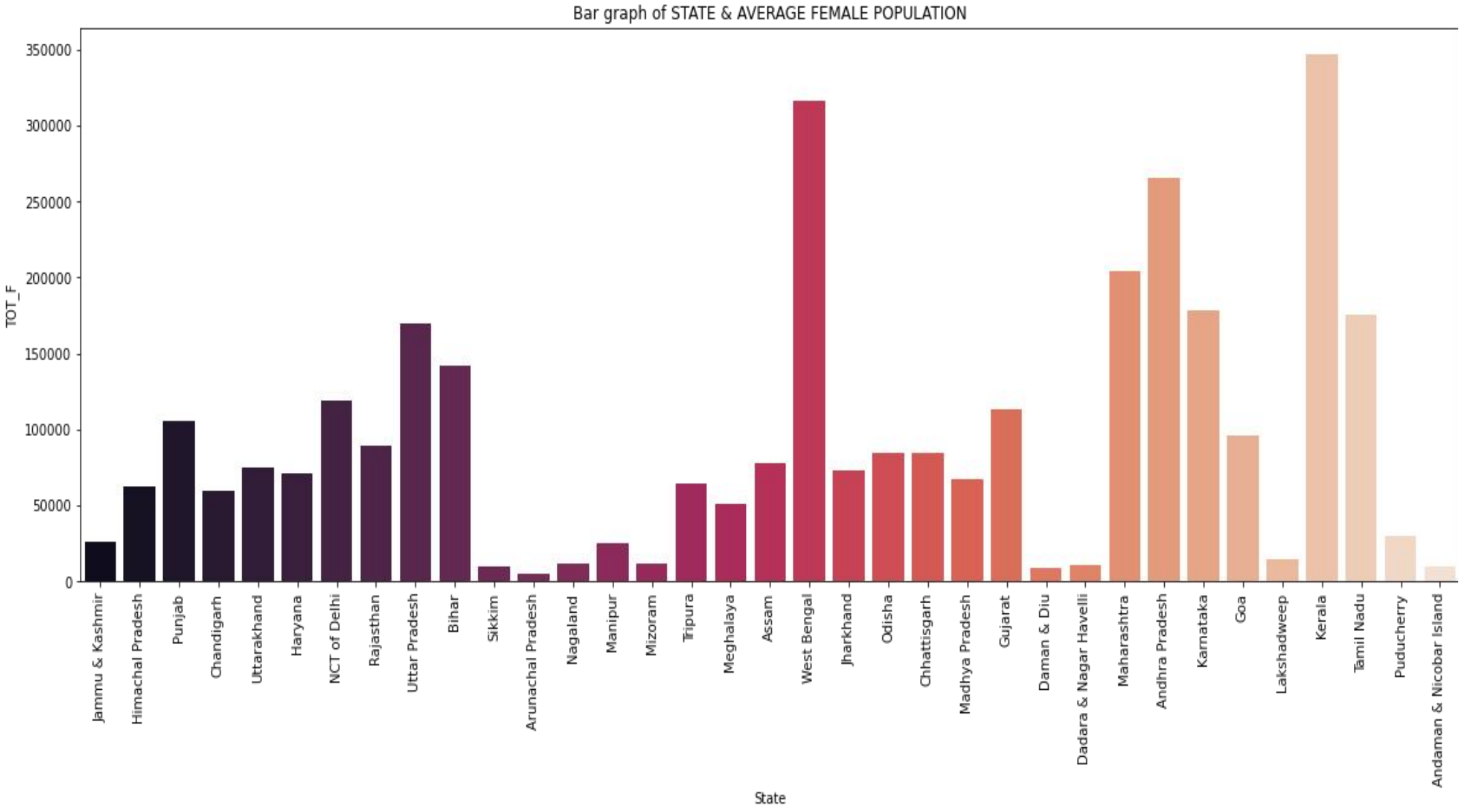

Bar graph of STATE & AVERAGE FEMALE POPULATION
TOT_F
350000
300000
250000
200000
150000
100000
50000
0
Jammu & Kashmir
Himachal Pradesh
Punjab
Chandigarh
Uttarakhand
Haryana
NCT of Delhi
Rajasthan
Uttar Pradesh
Bihar
Sikkim
Arunachal Pradesh
Nagaland
Manipur
Mizoram
Tripura
Meghalaya
Assam
West Bengal
Jharkhand
Odisha
Chhattisgarh
Madhya Pradesh
Gujarat
Daman & Diu
Dadara & Nagar Havelli
Maharashtra
Andhra Pradesh
Karnataka
Goa
Lakshadweep
Kerala
Tamil Nadu
Puducherry
Andaman & Nicobar Island
State

Fig 3.7

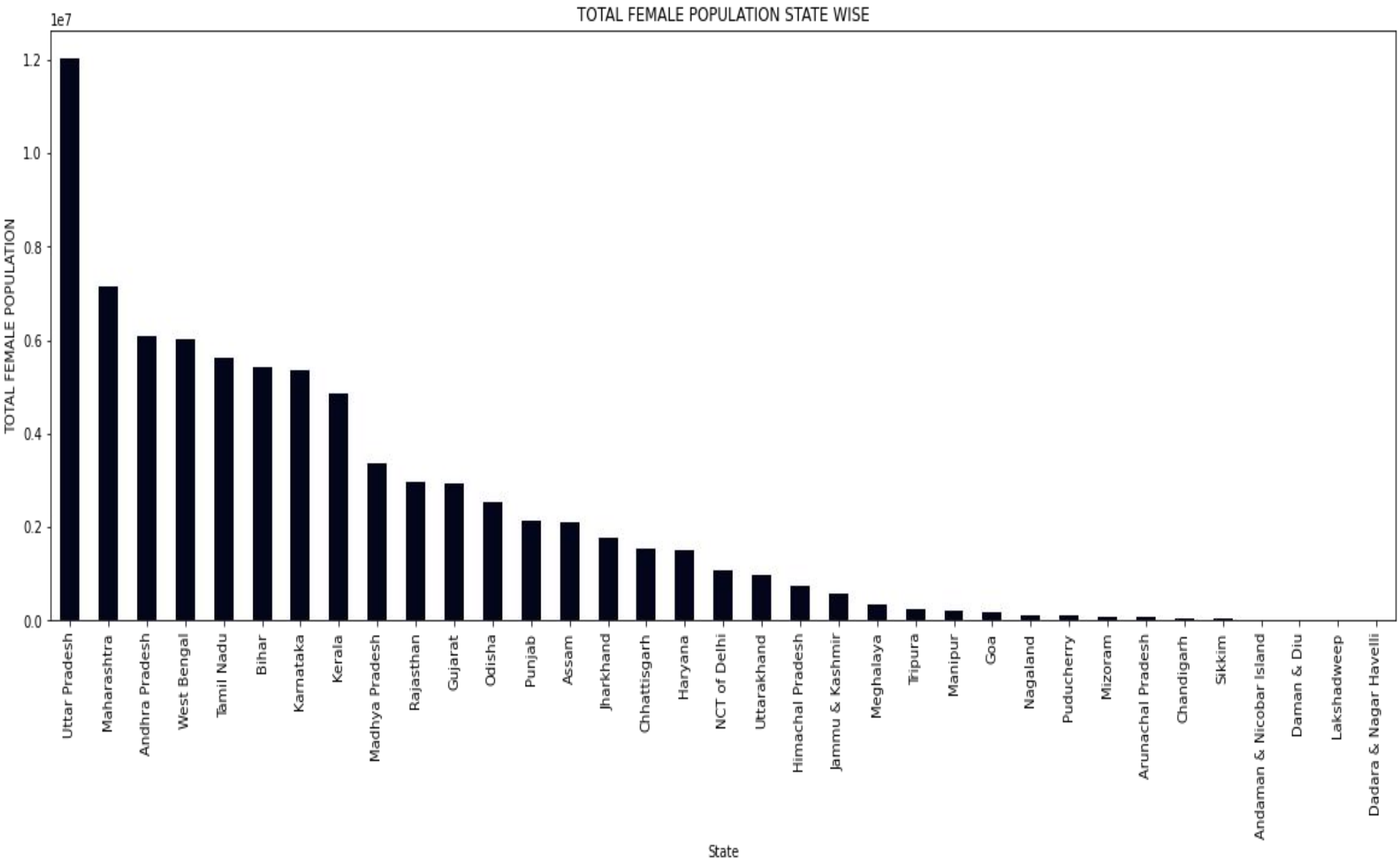

TOTAL FEMALE POPULATION STATE WISE
1e7
TOTAL FEMALE POPULATION
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Uttar Pradesh
Maharashtra
Andhra Pradesh
West Bengal
Tamil Nadu
Bihar
Karnataka
Kerala
Madhya Pradesh
Rajasthan
Gujarat
Odisha
Punjab
Assam
Jharkhand
Chhattisgarh
Haryana
NCT of Delhi
Uttarakhand
Himachal Pradesh
Jammu & Kashmir
Meghalaya
Tripura
Manipur
Goa
Nagaland
Puducherry
Mizoram
Arunachal Pradesh
Chandigarh
Sikkim
Andaman & Nicobar Island
Daman & Diu
Lakshadweep
Dadara & Nagar Havelli
State

Fig 3.8

Kerala and West Bengal has the highest average male and female population in India.When considering the total number of male and female Uttar Pradesh and Maharashtra shows the highest values.Arunachal Pradesh has the least average male and female population.

10.4 Which state has the highest literacy rate of male and female population ?

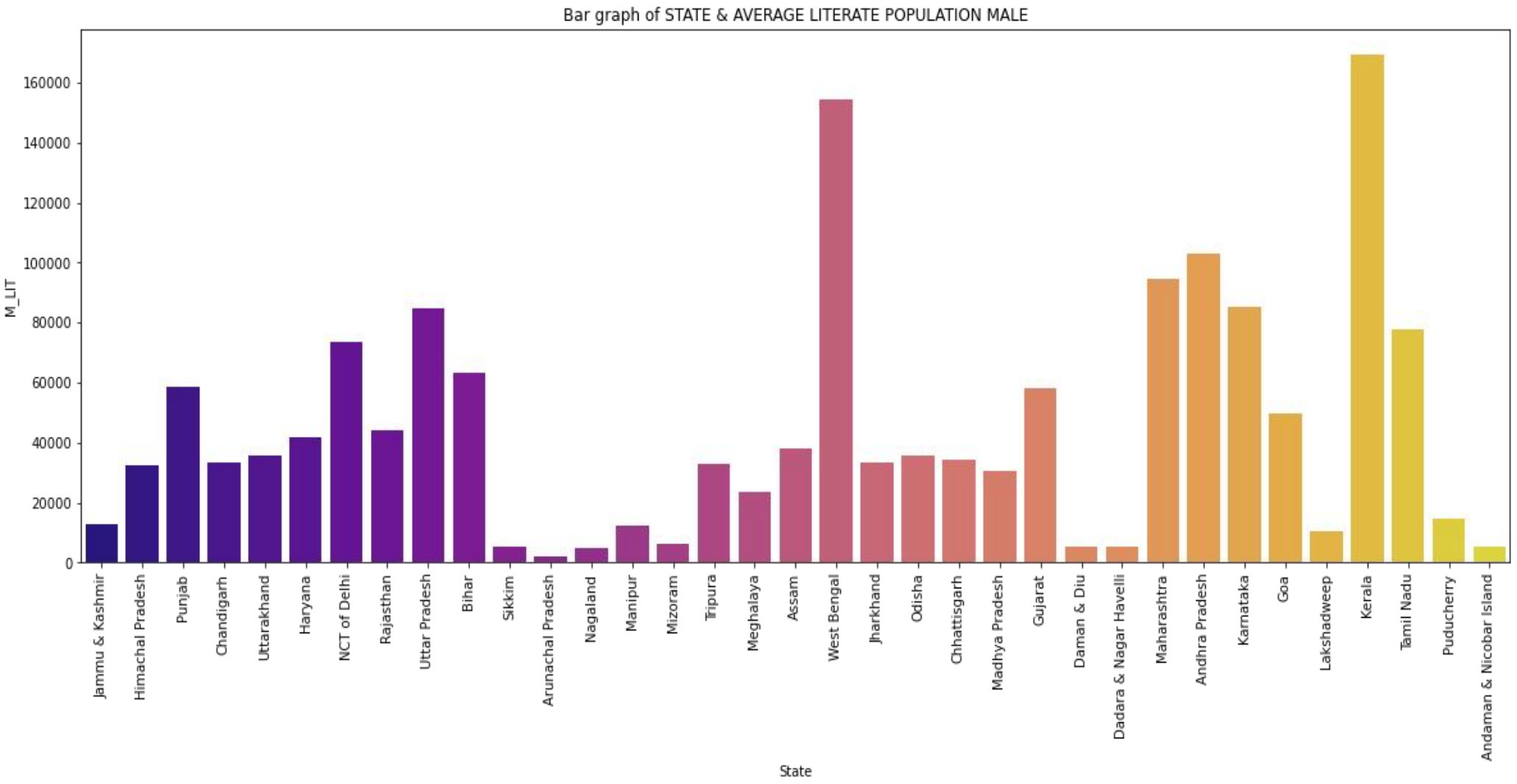


Fig 3.9

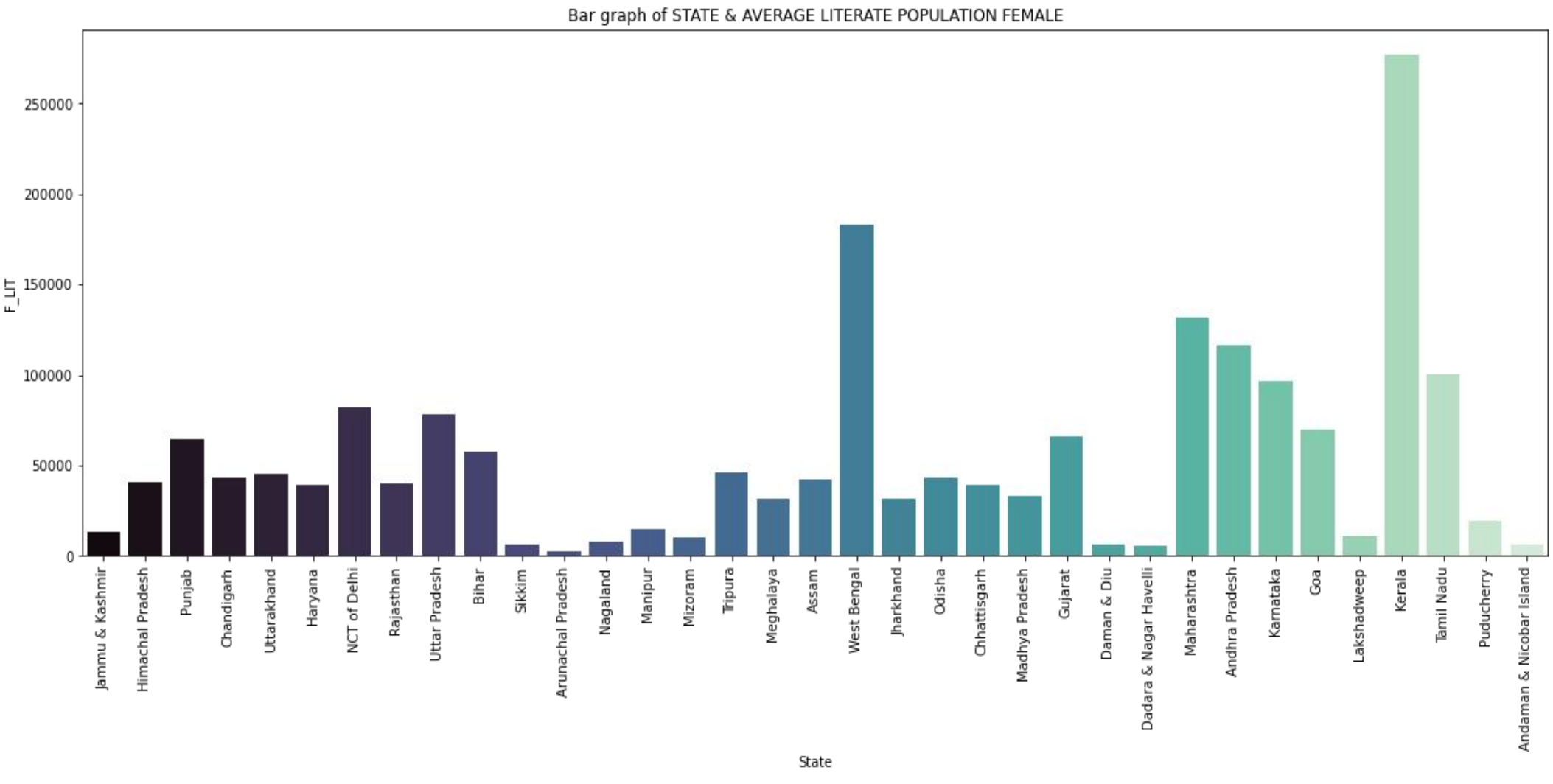


Fig 3.10

Kerala and West Bengal has the highest literacy rate for both male and female population..Arunachal Pradesh has the least average among all states in India.Literacy rate for females in Kerala is much higher than other states.

Heat map to show the correlation between the selected columns:

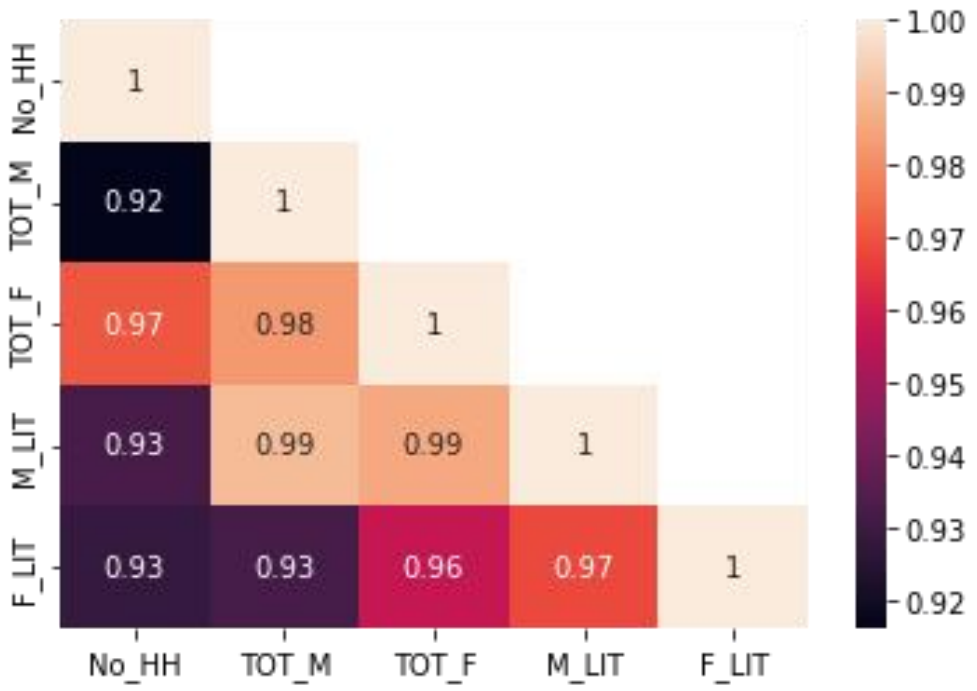


Fig 3.11

Number of Household - No_HH , Total population Male - TOT_M , Total population Female - TOT_F , Literates population Male - M_LIT , Literates population Female - F_LIT all these factors show high correlation .

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

We don't have to treat outliers here.The data provided is the census.Here there are no abnormal observations as these are population data.In some scenario treating outlier is not good because it negatively affect the statistical analysis and the training process of a machine learning algorithm resulting in lower accuracy.

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

Scaling does not treat outliers.Scaling is a concern between the columns whereas outliers are found within the column.

Scaled data using zscore method is shown below:

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_0_3_M	MARG_CL_0_3_F	M
0	-0.904738	-0.771236	-0.815563	-0.561012	-0.507738	-0.958575	-0.957049	-0.423306	-0.476423	-0.798097	...	-0.163229	-0.720610	
1	-0.935695	-0.823100	-0.874534	-0.681096	-0.725367	-0.958297	-0.956772	-0.582014	-0.607607	-0.849434	...	-0.583103	-0.732811	
2	-0.972412	-1.000919	-0.981466	-0.976956	-0.965262	-0.958575	-0.956772	-0.038951	-0.027273	-0.956457	...	-0.859212	-0.921931	
3	-1.037530	-1.052224	-1.041001	-1.022118	-0.995393	-0.958783	-0.957049	-0.355965	-0.390060	-1.004643	...	-0.805468	-0.900758	
4	-0.822676	-0.809381	-0.813933	-0.622359	-0.649908	-0.957395	-0.955529	0.149238	0.043330	-0.800568	...	-0.348645	-0.297513	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
635	-0.995677	-0.978990	-0.974268	-0.971387	-0.948916	-0.957326	-0.955667	-0.625124	-0.640197	-0.913820	...	-0.914299	-0.972530	
636	-0.844340	-0.921822	-0.886965	-0.936754	-0.919757	-0.803806	-0.765670	-0.625124	-0.640197	-0.853390	...	-0.831668	-0.868461	
637	-1.038465	-1.069066	-1.054885	-1.051356	-1.035331	-0.958783	-0.957049	-0.522953	-0.529880	-1.016367	...	-0.865930	-0.941309	
638	-0.986758	-1.019276	-1.007472	-1.008195	-0.996541	-0.958783	-0.957049	-0.622297	-0.637046	-0.962328	...	-0.844432	-0.927673	
639	-0.899166	-0.926854	-0.919050	-0.943193	-0.935220	-0.958783	-0.957049	-0.608870	-0.623555	-0.856916	...	-0.819576	-0.945616	

640 rows × 57 columns

Table 3.4

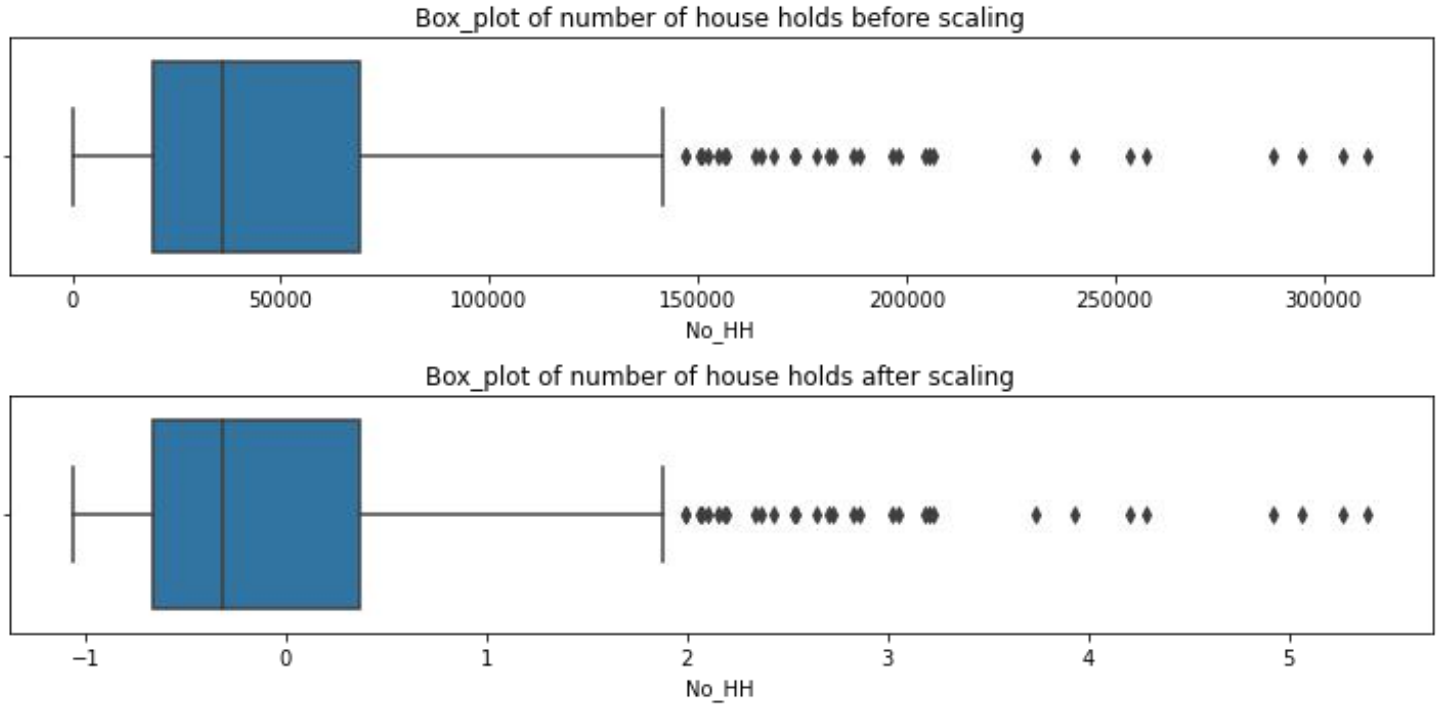


Fig 3.12

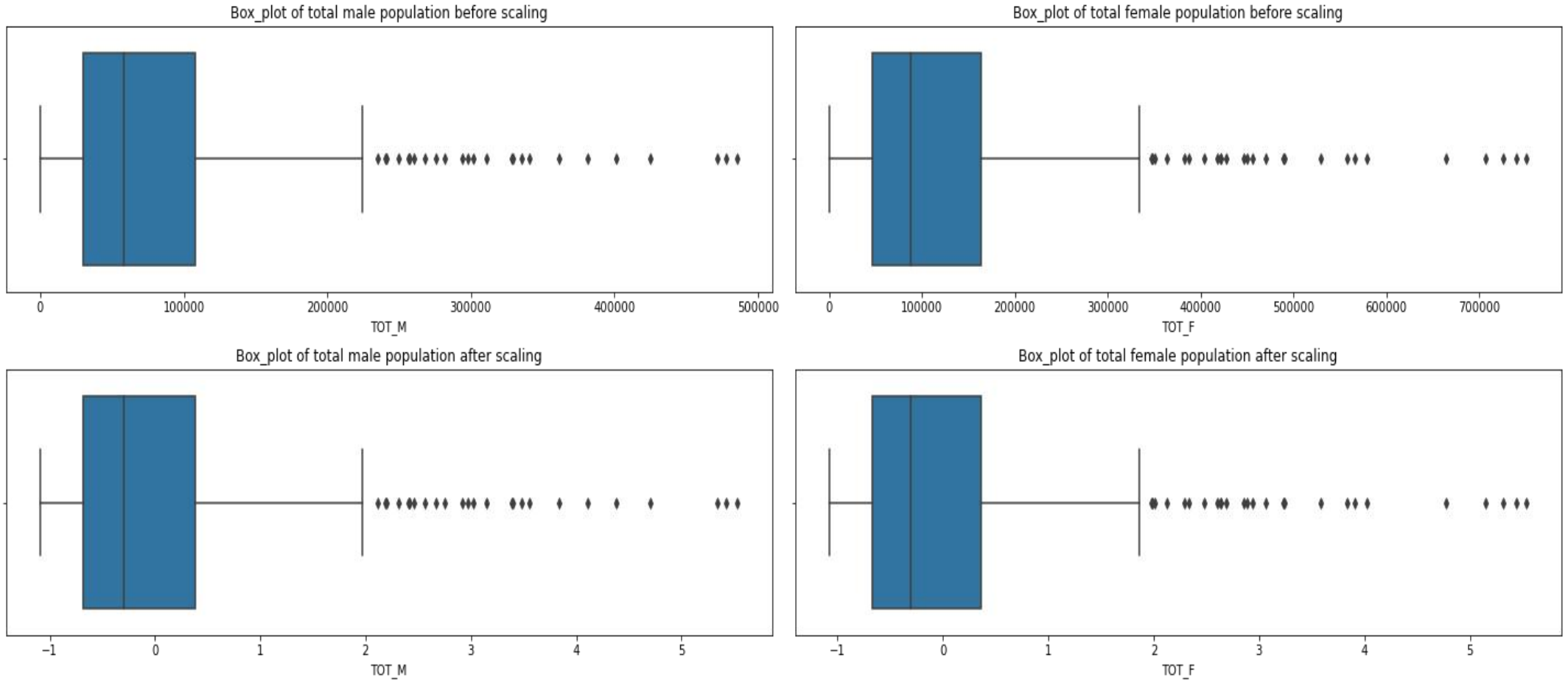


Fig 3.13

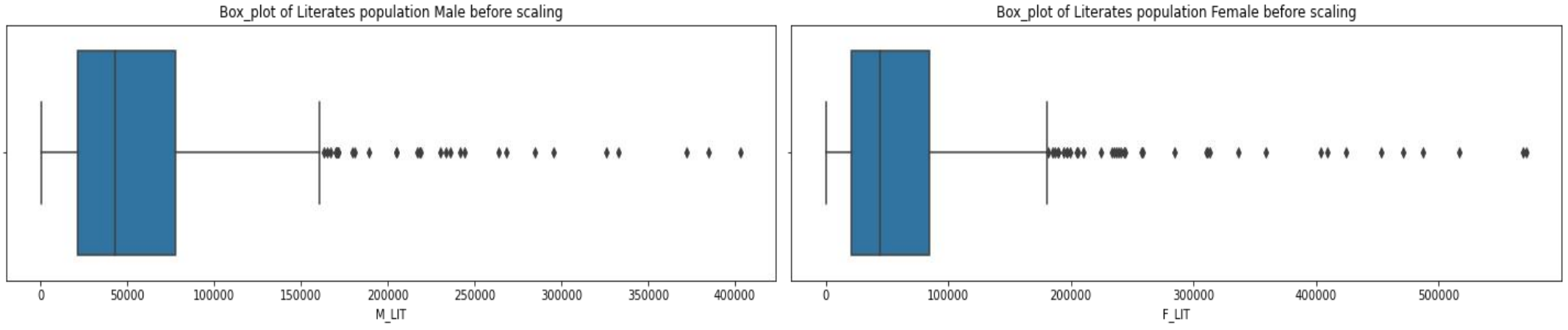


Fig 3.14

We have plotted only the selected columns but it represents the effect of outliers on scaled data.It is evident from the box plots that scaling does not effect the outliers in the data,but scaling reduces the larger scaled dimensions.

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

Checking correlation of scaled data:

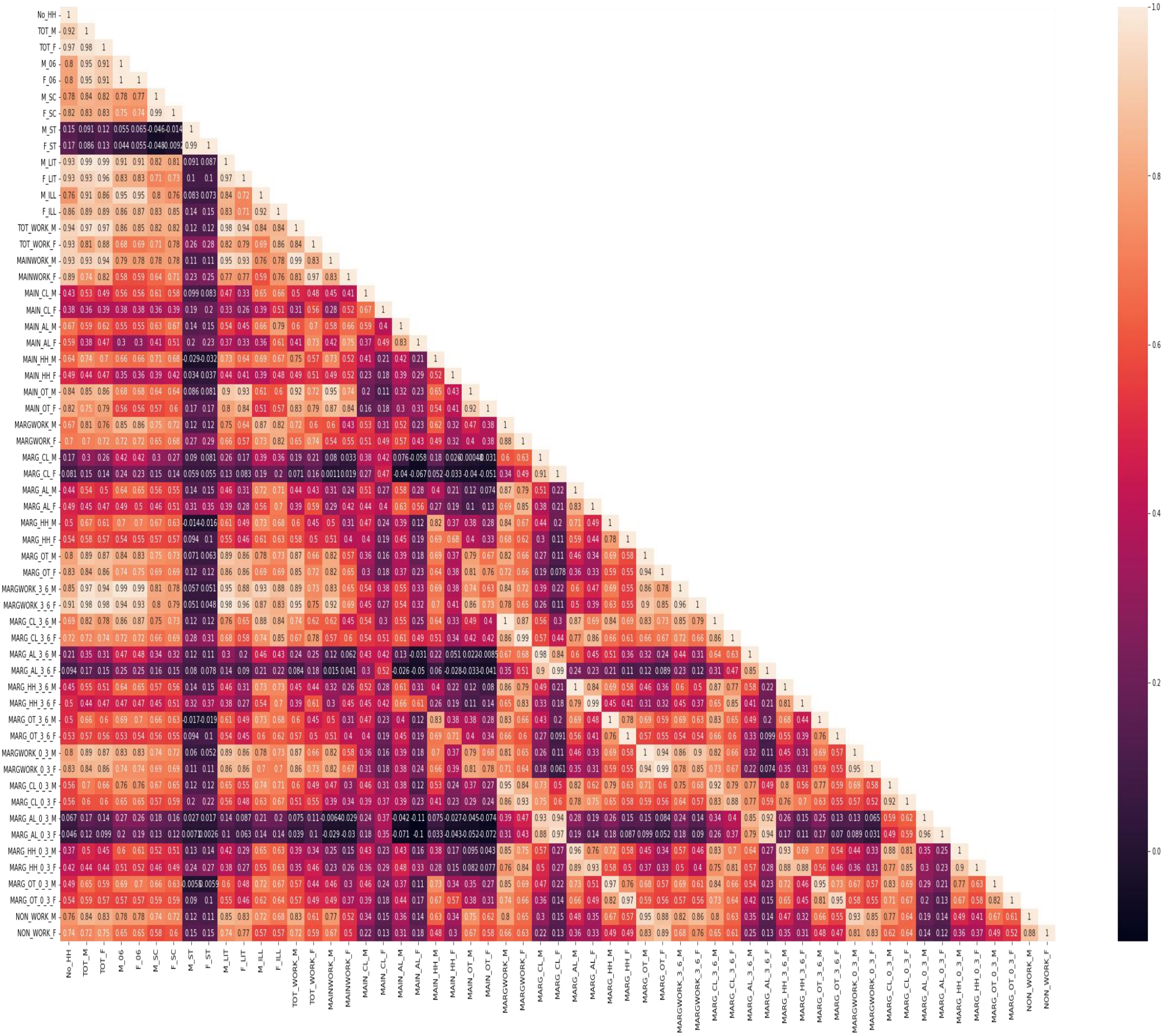


Fig 3.15

COVARIANCE MATRIX :

```
Covariance Matrix
%s [[1.00156495 0.91760364 0.97210871 ... 0.53769433 0.76357722 0.73684378]
 [0.91760364 1.00156495 0.98417823 ... 0.5891007  0.84621844 0.71718181]
 [0.97210871 0.98417823 1.00156495 ... 0.572748   0.82894851 0.74775097]
 ...
 [0.53769433 0.5891007  0.572748   ... 1.00156495 0.61052325 0.52191235]
 [0.76357722 0.84621844 0.82894851 ... 0.61052325 1.00156495 0.88228018]
 [0.73684378 0.71718181 0.74775097 ... 0.52191235 0.88228018 1.00156495]]
```

Table 3.5

PCA_TRANSFORMED:

```
[[-4.61726348e+00  1.38115857e-01  3.28544953e-01 ... -6.06029097e-16
   6.08748032e-16  2.96207163e-16]
 [-4.77166187e+00 -1.05865351e-01  2.44448976e-01 ...  2.27457842e-16
   1.68484483e-16  4.64305393e-16]
 [-5.96483558e+00 -2.94346892e-01  3.67393453e-01 ...  4.15150177e-16
  -4.48318483e-17  8.23349729e-16]
 ...
 [-6.29462500e+00 -6.38126644e-01  1.07482817e-01 ...  3.45399264e-16
   1.72137876e-16 -4.11830008e-16]
 [-6.22319199e+00 -6.72319673e-01  2.71325467e-01 ... -1.01201345e-15
  -4.66369483e-16  1.55035951e-16]
 [-5.89623627e+00 -9.37169526e-01  3.49218364e-01 ... -6.19779140e-17
   3.72269309e-16 -5.80404880e-16]]
```

Table 3.6

Finding eigon_values and eigon_vectors:

Eigon_vector

```
array([[ 0.15602058,  0.16711763,  0.16555318, ...,  0.13219224,
         0.15037558,  0.1310662 ],
       [-0.12634653, -0.08967655, -0.10491237, ...,  0.05081332,
        -0.06536455, -0.07384742],
       [-0.00269025,  0.05669762,  0.03874947, ..., -0.07871987,
         0.11182732,  0.1025525 ],
       ...,
       [ 0.        ,  0.2077636 ,  0.24647657, ..., -0.07217993,
         0.00399206, -0.06929081],
       [ 0.        ,  0.2887035 , -0.20596721, ...,  0.04019745,
        -0.03192722,  0.00778048],
       [-0.        ,  0.18790022,  0.02642675, ..., -0.02597314,
        -0.13972835, -0.02147533]])
```

Table 3.7

Eigon_values

```
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 3.8

Explained variance for each PC

```
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 3.9

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

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	...	PC48	PC49	PC50	PC5
No_HH	0.156021	-0.126347	-0.002690	-0.125293	-0.007022	0.004083	-0.118110	0.057238	0.004265	0.019985	...	-0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+0
TOT_M	0.167118	-0.089677	0.056698	-0.019942	-0.033026	-0.073389	0.089554	0.111431	0.018872	-0.024501	...	1.617181e-01	-3.303320e-01	-2.059894e-01	3.603136e-0
TOT_F	0.165553	-0.104912	0.038749	-0.070873	-0.012847	-0.043647	-0.002124	0.088355	0.014911	-0.038041	...	1.907122e-01	-8.601042e-02	-1.608897e-01	-1.625048e0
M_06	0.162193	-0.022095	0.057788	0.011917	-0.050248	-0.157957	0.165067	0.169595	-0.056773	-0.153574	...	-2.870332e-15	-3.410187e-16	-5.577075e-16	4.228508e-1
F_06	0.162566	-0.020271	0.050126	0.014844	-0.043848	-0.154436	0.169082	0.169459	-0.059323	-0.169567	...	2.475438e-15	2.685312e-16	-4.896323e-16	-2.956643e1
M_SC	0.151358	-0.045111	0.002569	0.012485	-0.173007	-0.064295	-0.001566	-0.129301	0.037480	0.448517	...	2.230854e-15	1.734723e-16	-1.203898e-15	-6.591949e1
F_SC	0.151567	-0.051924	-0.025101	-0.029893	-0.159803	-0.040518	-0.084658	-0.144352	0.041232	0.446968	...	-2.319125e-15	-3.999868e-16	1.219330e-15	5.288230e-1

Table 3.10

All the 57 components are shown in the table above.To identify the optimum number of components we can use scree plot.

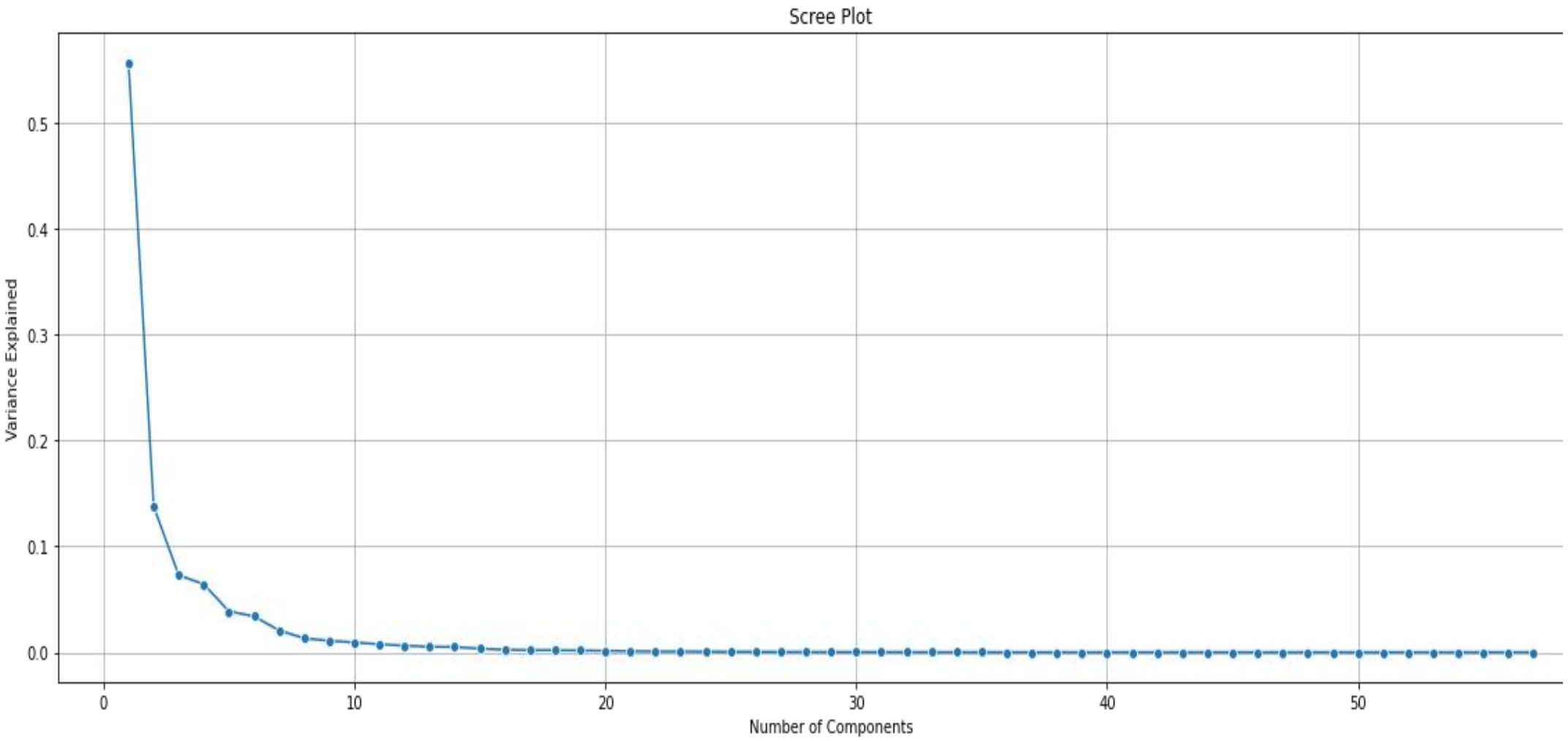


Fig 3.16

Check the cumulative explained variance ratio to find a cut off for selecting the number of PCs:

```
[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 3.11

If we take 90% of the explained variance we will consider 6 components.

Table below shows the selected 6 components.

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_0_3_M	MA
PC1	0.156021	0.167118	0.165553	0.162193	0.162566	0.151358	0.151567	0.027234	0.028183	0.161993	...	0.150126	
PC2	-0.126347	-0.089677	-0.104912	-0.022095	-0.020271	-0.045111	-0.051924	0.027679	0.030223	-0.115355	...	0.150681	
PC3	-0.002690	0.056698	0.038749	0.057788	0.050126	0.002569	-0.025101	-0.123504	-0.139769	0.082168	...	0.054892	
PC4	-0.125293	-0.019942	-0.070873	0.011917	0.014844	0.012485	-0.029893	-0.222247	-0.229754	-0.035163	...	0.087433	
PC5	-0.007022	-0.033026	-0.012847	-0.050248	-0.043848	-0.173007	-0.159803	0.433163	0.438792	-0.009101	...	0.081185	
PC6	0.004083	-0.073389	-0.043647	-0.157957	-0.154436	-0.064295	-0.040518	0.222591	0.225531	-0.055465	...	-0.060715	

Table 3.12

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.

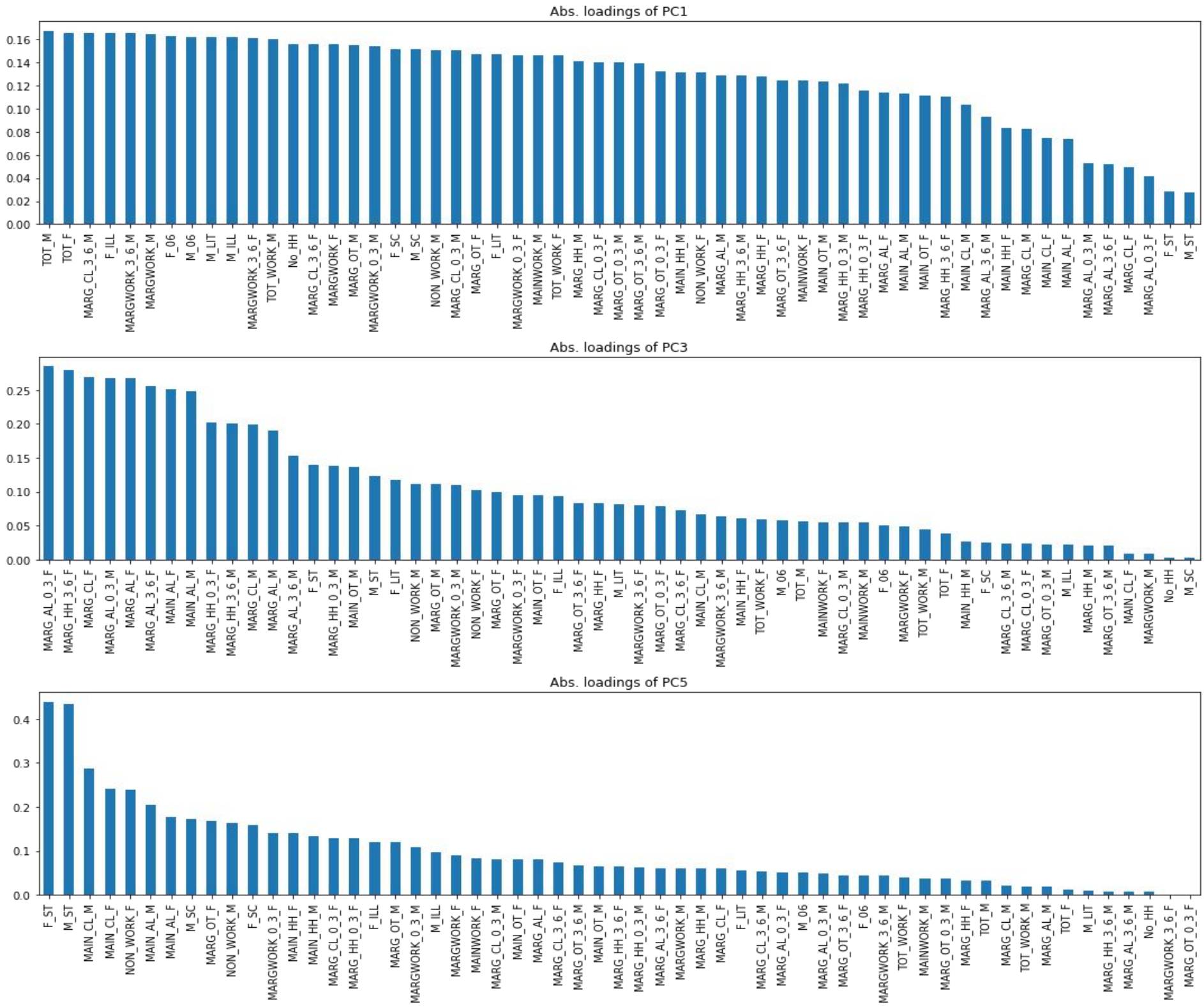


Fig 3.17

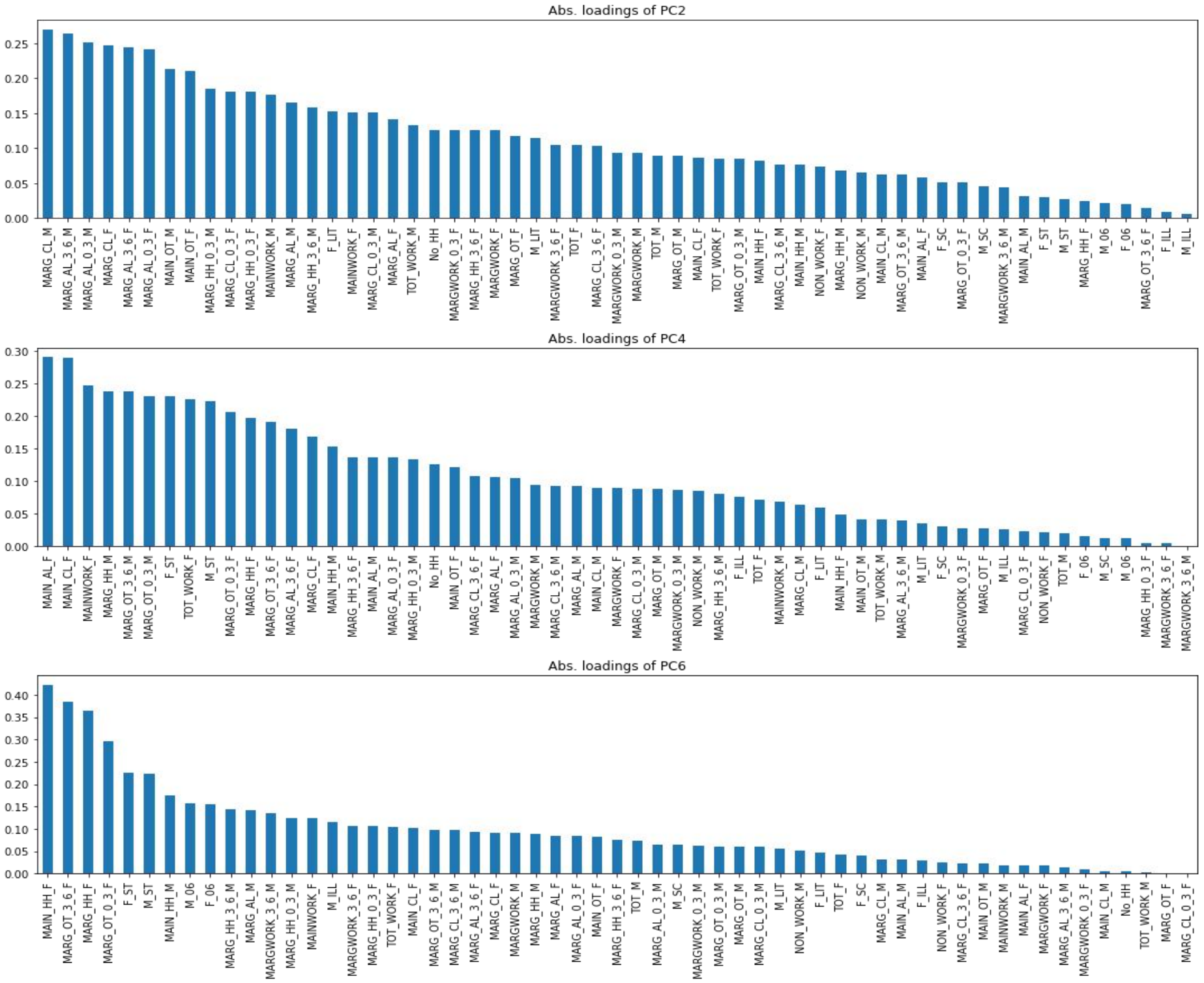


Fig 3.18

Compare how the original features influence various PCs

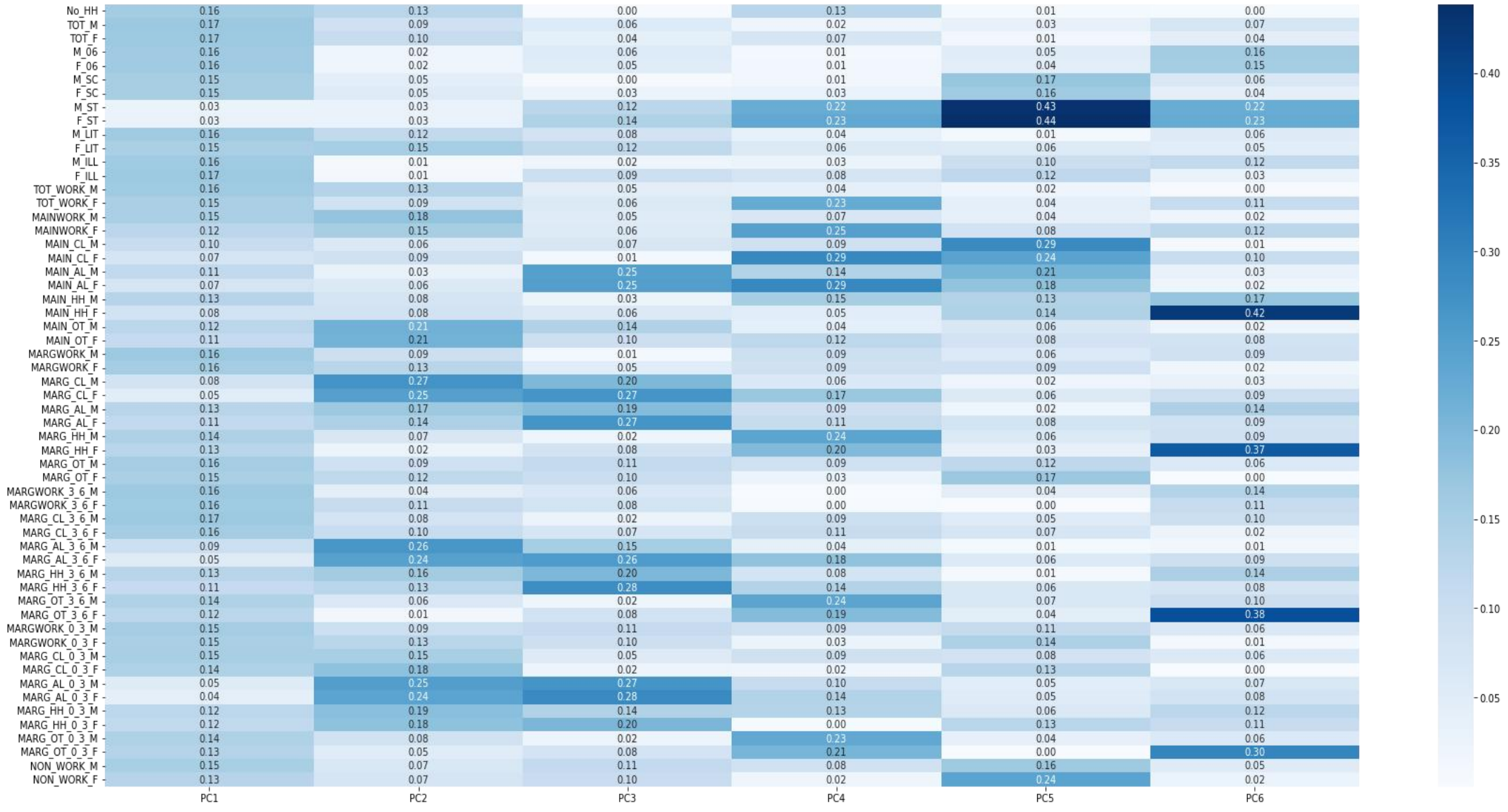


Fig 3.19

It is observed from the cumulative explained variance ratio that PC1 contributes to 55% of census data.PC1 contributes more to the data than others. So we can conclude that 90% of the data is based on these 6 selected components.

Observing PC1 total male population and female population contributes more followed by a column representing marginal Cultivator Population 3-6 Male. Thus all these factors contribute to PC1 in descending order of magnitude(From left).

The same observation can be seen from the heat map. The dark-colored value shows a high magnitude.PC1 is made up of several important characteristics, most factors having approximately equal magnitude. This makes PC1 unique from other PCs.

Final transformed data of the selected 6 components:

	0	1	2	3	4	5	6	7	8	9	...	630	631	632	633
PC1	-4.617263	-4.771662	-5.964836	-6.280796	-4.478566	-3.319963	-5.021393	-4.608709	-5.186703	-4.226190	...	0.231290	2.233075	-0.425308	-6.462810
PC2	0.138116	-0.105865	-0.294347	-0.500384	0.894154	2.823865	-0.346359	0.022370	-0.059097	-1.335080	...	-0.986332	-4.733497	-3.210692	-0.731394
PC3	0.328545	0.244449	0.367394	0.212701	1.078277	3.058460	0.650378	0.398755	0.184397	0.697838	...	-0.788338	0.651248	-0.142504	0.120313
PC4	1.543697	1.963215	0.619543	1.074515	0.535557	-0.447904	0.981072	1.576995	1.735440	1.470509	...	-0.856567	-1.613116	-1.539146	1.168501
PC5	0.353736	-0.153884	0.478199	0.300799	0.804065	0.742445	-0.059778	0.171316	0.169174	0.269146	...	-1.141717	-0.407586	-0.869835	-0.002325
PC6	-0.420948	0.417308	0.276581	0.051157	0.341678	0.634676	-0.246957	-0.139444	0.455039	-0.002576	...	0.071518	0.576427	0.630854	-0.086795

Table 3.13

Correlation of the transformed final data:

	PC1	PC2	PC3	PC4	PC5	PC6
PC1	1.000000e+00	1.535547e-16	-1.076170e-16	3.653845e-17	5.163546e-17	9.759612e-17
PC2	1.535547e-16	1.000000e+00	-1.788472e-17	-1.761625e-16	1.778289e-17	-5.573071e-17
PC3	-1.076170e-16	-1.788472e-17	1.000000e+00	4.192504e-16	-1.005367e-16	-1.106783e-16
PC4	3.653845e-17	-1.761625e-16	4.192504e-16	1.000000e+00	1.429077e-16	5.354019e-17
PC5	5.163546e-17	1.778289e-17	-1.005367e-16	1.429077e-16	1.000000e+00	-1.478158e-16
PC6	9.759612e-17	-5.573071e-17	-1.106783e-16	5.354019e-17	-1.478158e-16	1.000000e+00

Table 3.14

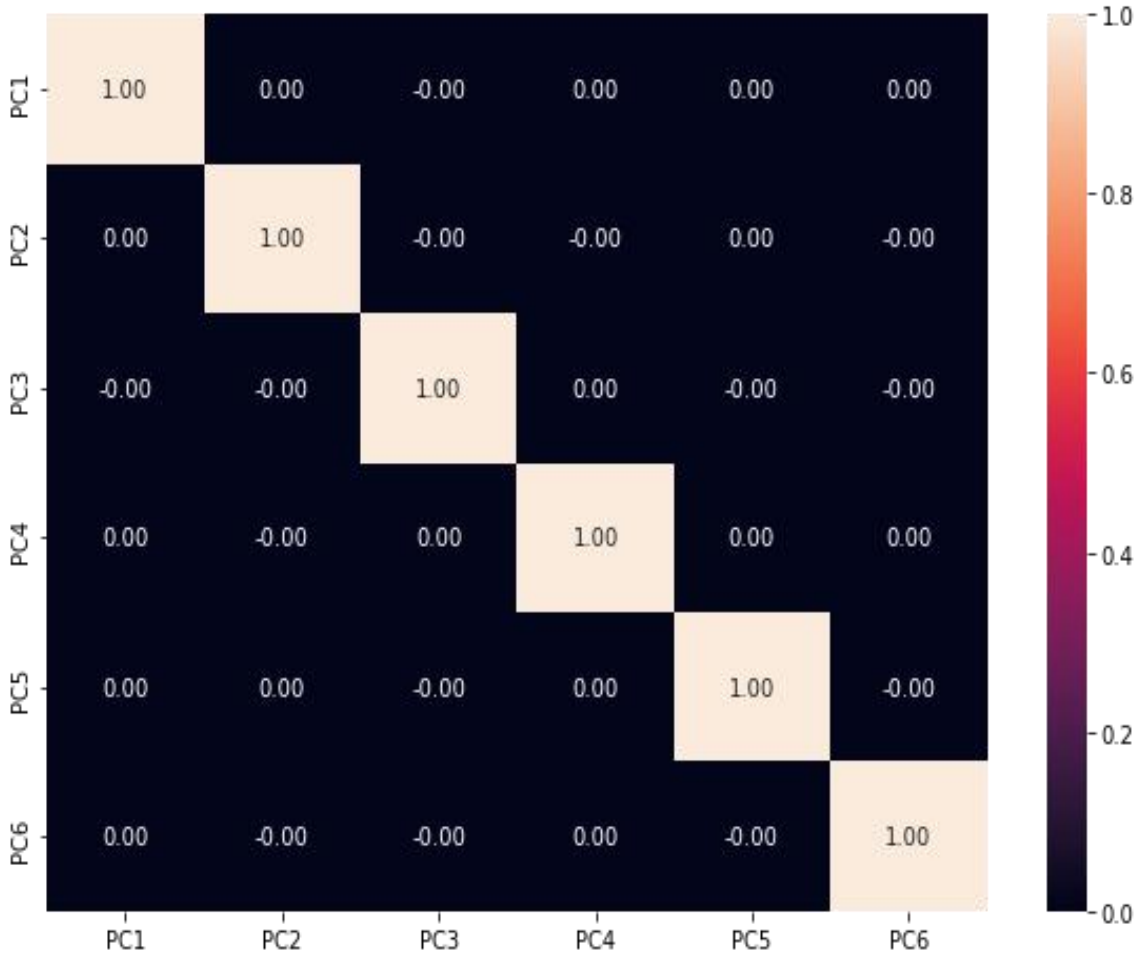


Fig 3.20

The heatmap of the the data shows that there is no correlation between the transformed data.Instead of studying the whole data we can use the selected 6 components for future predictions and analyses .

16 Write linear equation for first PC.

0.1560205785856793 6 * No_HH +
0.1671176348853345 * TOT_M +
0.1655531790906489 3 * TOT_F +
0.1621929482046555 * M_06 +
0.1625663956573483 2 * F_06 +
0.1513578490906058 2 * M_SC +
0.1515665001920887 5 * F_SC +
0.0272341945710042 3 * M_ST +
0.0281833150158726 92 * F_ST +
0.1619928373362915 5 * M_LIT +
0.1468726803014028 5 * F_LIT +
0.1617494446347163 3 * M_ILL +
0.1652481873683337 2 * F_ILL +
0.1598719881620128 4 * TOT_WORK_M +
0.1459358037724762 * TOT_WORK_F +
0.1462007297630599 * MAINWORK_M +
0.1239702835727364 8 * MAINWORK_F +
0.1031271588301986 7 * MAIN_CL_M +
0.0745397855548367 7 * MAIN_CL_F +
0.1133557121815672 3 * MAIN_AL_M +
0.0738821590315588 1 * MAIN_AL_F +
0.131572584022756 * MAIN_HH_M +
0.0833826396743576 6 * MAIN_HH_F +

0.1235262419225308 4 * MAIN_OT_M +
0.111021263913201 3 * MAIN_OT_F +
0.164615478560110 1 * MARGWORK_M +
0.155395618108341 3 * MARGWORK_F +
0.082388541407045 46 * MARG_CL_M +
0.049195395678738256 * MARG_CL_F +
0.128598562946685 65 * MARG_AL_M +
0.114305072789198 95 * MARG_AL_F +
0.140853226961851 4 * MARG_HH_M +
0.127669598014753 64 * MARG_HH_F +
0.155262871623116 03 * MARG_OT_M +
0.14728658356523 39 * MARG_OT_F +
0.164971949937144 54 * MARGWORK_3_6_M +
0.16125343257531 36 * MARGWORK_3_6_F +
0.165501611025806 3 * MARG_CL_3_6_M +
0.155647049144833 85 * MARG_CL_3_6_F +
0.093014206401928 48 * MARG_AL_3_6_M +
0.051535863970152224 * MARG_AL_3_6_F +
0.128576116428678 22 * MARG_HH_3_6_M +
0.110645843236969 22 * MARG_HH_3_6_F +
0.139592762521588 36 * MARG_OT_3_6_M +
0.12454590917258751 * MARG_OT_3_6_F +
0.154293785789160 42 * MARGWORK_0_3_M +
0.14628565406214422 * MARGWORK_0_3_F +
0.150125706102620 66 * MARG_CL_0_3_M +
0.14015704689010391 * MARG_CL_0_3_F +
0.052541782853963 45 * MARG_AL_0_3_M +
0.041785953012010 33 * MARG_AL_0_3_F +
0.121840353879250 24 * MARG_HH_0_3_M +
0.11601141016824 11 * MARG_HH_0_3_F +
0.139868774110428 08 * MARG_OT_0_3_M +
0.13219224458196535 * MARG_OT_0_3_F +
0.150375578044112 97 * NON_WORK_M +
0.13106620313207334 * NON_WORK_F