**PCA India Data\_Census**

|  |  |
| --- | --- |
| **Part 2 - PCA: Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.** | **4** |
| **Part 2 - PCA: 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** | **6** |
| **Part 2 - PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?** | **1** |
| **Part 2 - PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.** | **3** |
| **Part 2 - PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.** | **4** |
| **Part 2 - PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.** | **3** |
| **Part 2 - PCA: Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables.** | **4** |
| **Part 2 - PCA: Write linear equation for first PC.** | **2** |

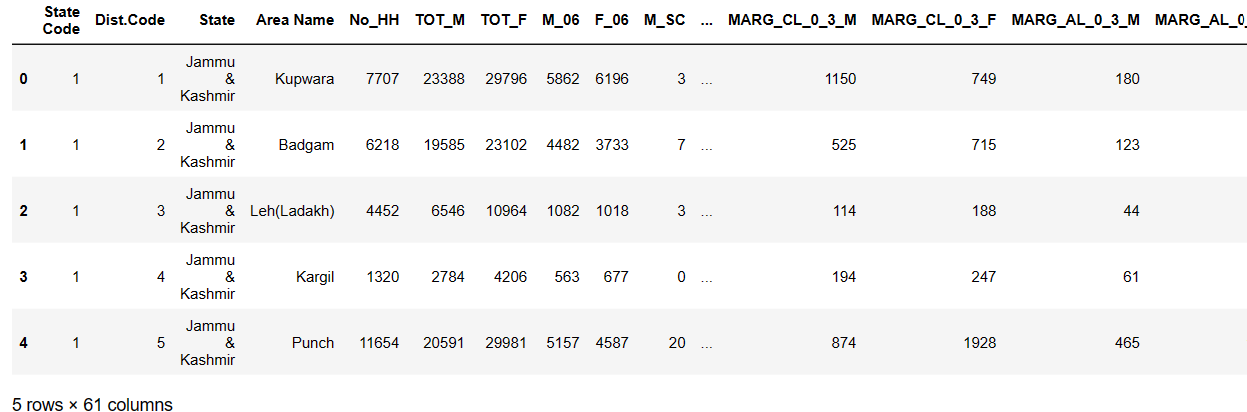
**Part 2: PCA Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc (4 Marks)**

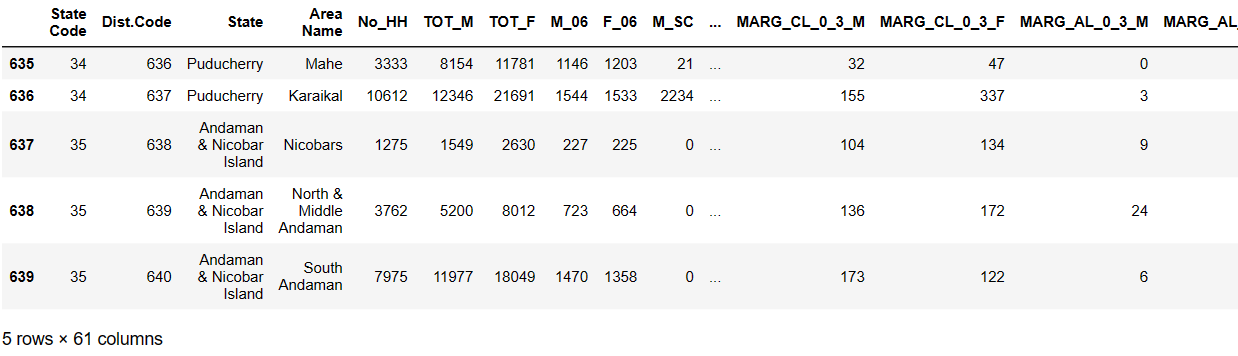
**Ans:-**

We have loaded the all the required packages and loaded PCA India Data\_Census Data file using Pandas.

Dataset has 640 rows and 61 columns.

We have viewed first and last few rows using head() and tail() functions respectively.





We can view dataset information using info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 640 entries, 0 to 639

Data columns (total 61 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 State Code 640 non-null int64

1 Dist.Code 640 non-null int64

2 State 640 non-null object

3 Area Name 640 non-null object

4 No\_HH 640 non-null int64

5 TOT\_M 640 non-null int64

6 TOT\_F 640 non-null int64

7 M\_06 640 non-null int64

8 F\_06 640 non-null int64

9 M\_SC 640 non-null int64

10 F\_SC 640 non-null int64

11 M\_ST 640 non-null int64

12 F\_ST 640 non-null int64

13 M\_LIT 640 non-null int64

14 F\_LIT 640 non-null int64

15 M\_ILL 640 non-null int64

16 F\_ILL 640 non-null int64

17 TOT\_WORK\_M 640 non-null int64

18 TOT\_WORK\_F 640 non-null int64

19 MAINWORK\_M 640 non-null int64

20 MAINWORK\_F 640 non-null int64

21 MAIN\_CL\_M 640 non-null int64

22 MAIN\_CL\_F 640 non-null int64

23 MAIN\_AL\_M 640 non-null int64

24 MAIN\_AL\_F 640 non-null int64

25 MAIN\_HH\_M 640 non-null int64

26 MAIN\_HH\_F 640 non-null int64

27 MAIN\_OT\_M 640 non-null int64

28 MAIN\_OT\_F 640 non-null int64

29 MARGWORK\_M 640 non-null int64

30 MARGWORK\_F 640 non-null int64

31 MARG\_CL\_M 640 non-null int64

32 MARG\_CL\_F 640 non-null int64

33 MARG\_AL\_M 640 non-null int64

34 MARG\_AL\_F 640 non-null int64

35 MARG\_HH\_M 640 non-null int64

36 MARG\_HH\_F 640 non-null int64

37 MARG\_OT\_M 640 non-null int64

38 MARG\_OT\_F 640 non-null int64

39 MARGWORK\_3\_6\_M 640 non-null int64

40 MARGWORK\_3\_6\_F 640 non-null int64

41 MARG\_CL\_3\_6\_M 640 non-null int64

42 MARG\_CL\_3\_6\_F 640 non-null int64

43 MARG\_AL\_3\_6\_M 640 non-null int64

44 MARG\_AL\_3\_6\_F 640 non-null int64

45 MARG\_HH\_3\_6\_M 640 non-null int64

46 MARG\_HH\_3\_6\_F 640 non-null int64

47 MARG\_OT\_3\_6\_M 640 non-null int64

48 MARG\_OT\_3\_6\_F 640 non-null int64

49 MARGWORK\_0\_3\_M 640 non-null int64

50 MARGWORK\_0\_3\_F 640 non-null int64

51 MARG\_CL\_0\_3\_M 640 non-null int64

52 MARG\_CL\_0\_3\_F 640 non-null int64

53 MARG\_AL\_0\_3\_M 640 non-null int64

54 MARG\_AL\_0\_3\_F 640 non-null int64

55 MARG\_HH\_0\_3\_M 640 non-null int64

56 MARG\_HH\_0\_3\_F 640 non-null int64

57 MARG\_OT\_0\_3\_M 640 non-null int64

58 MARG\_OT\_0\_3\_F 640 non-null int64

59 NON\_WORK\_M 640 non-null int64

60 NON\_WORK\_F 640 non-null int64

dtypes: int64(59), object(2)

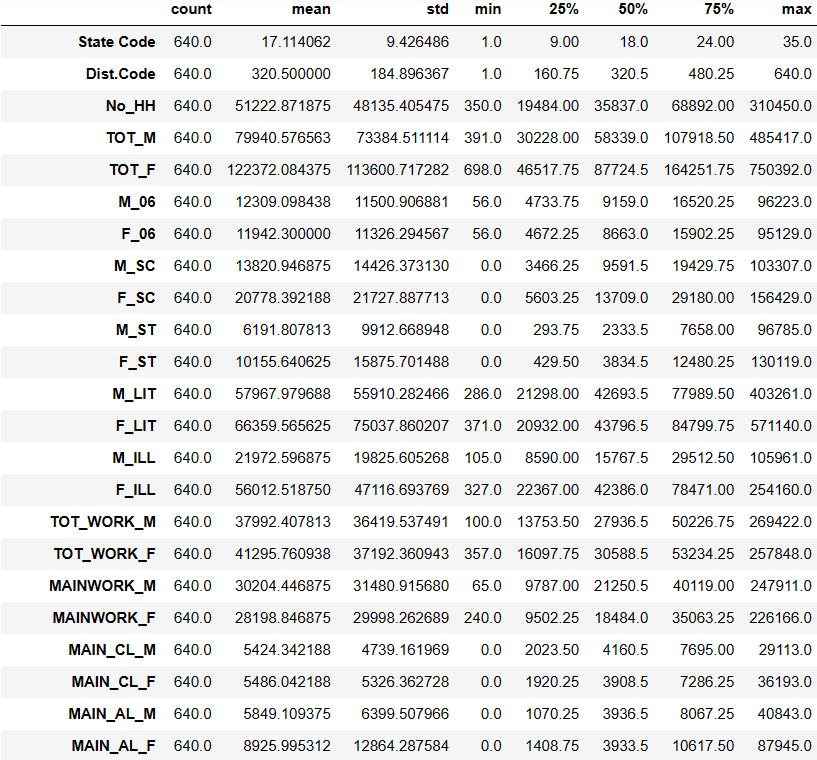
memory usage: 305.1+ KB

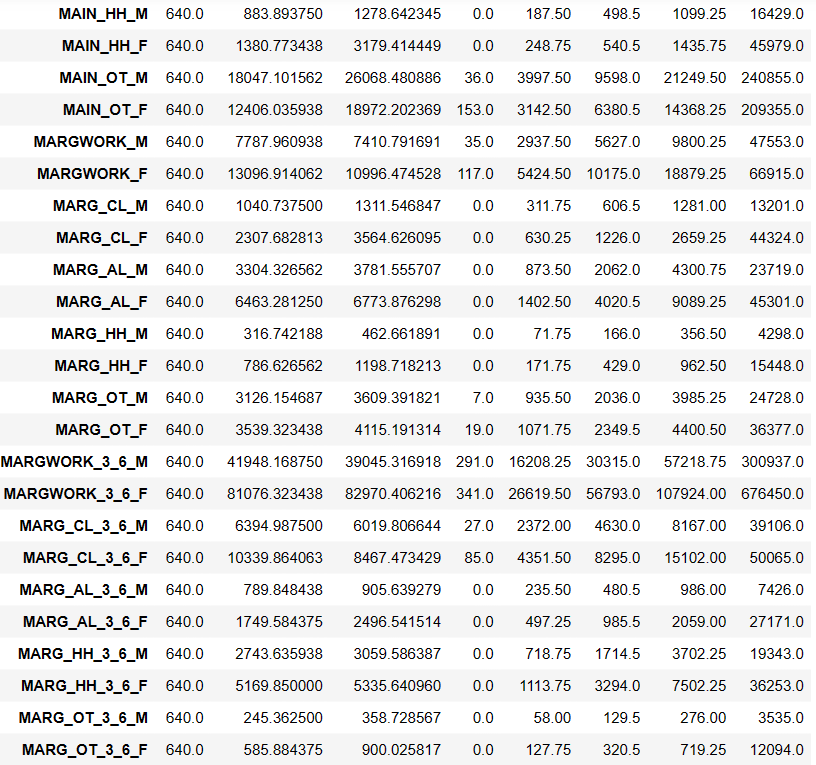
We have 59 integer column and 2 categorical columns. There are no null values in the dataset

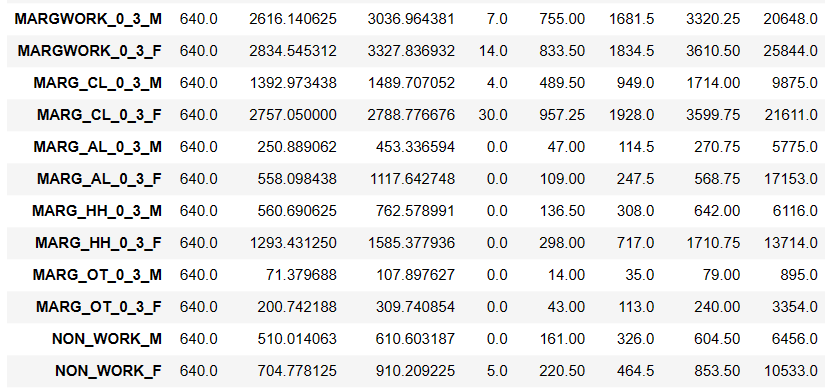
There are no duplicate values in the data.

We can describe data using .describe() function.

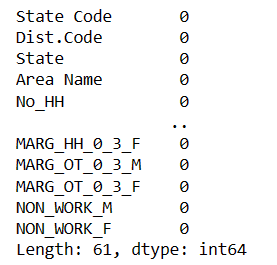
1. There are maximum 35 state codes and minimum of 1 state code.
2. There are maximum 640 dist codes and minimum 1 dist code in a given state.
3. No of Households may vary between 350 and 310450.0
4. Average number of total male population is 79940.576563 and Average population of female is 122372.084375.
5. Total count of each record in the dataset is 640 as there are no null values
6. We can see average value,minimum and maximum values for male and female in every sector of work and even in non working background.
7. Every State code has a State Name and every Dist Name has a Area Name of its own.







There are no null values in the dataset



There are no duplicate values to treat in the dataset.

**Part 2 PCA: 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**

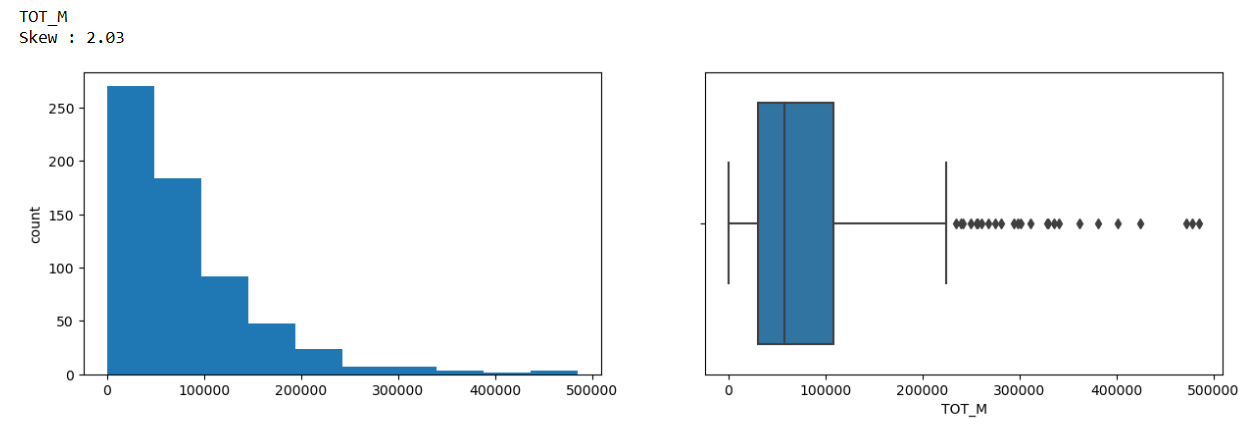
**Ans:-**

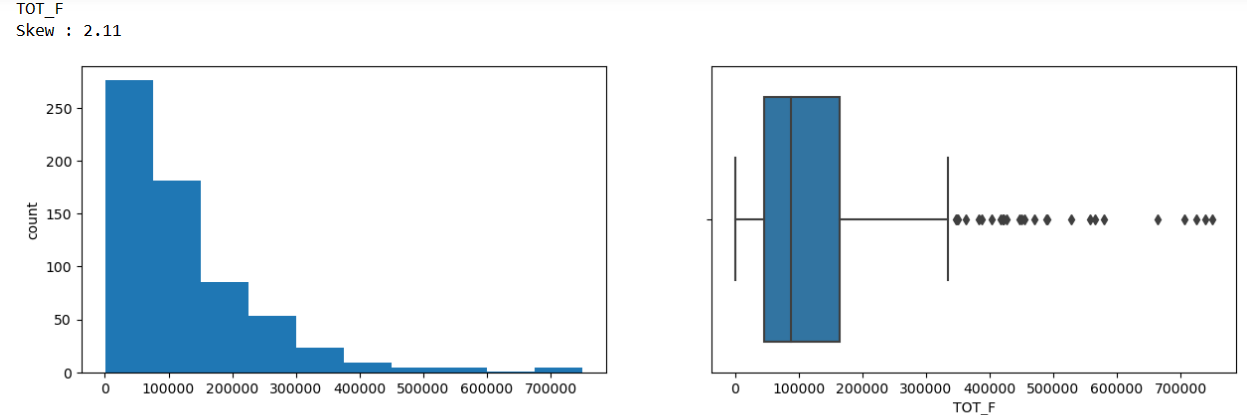
As there are no null or duplicate values in the dataset so we will not be treating any values.

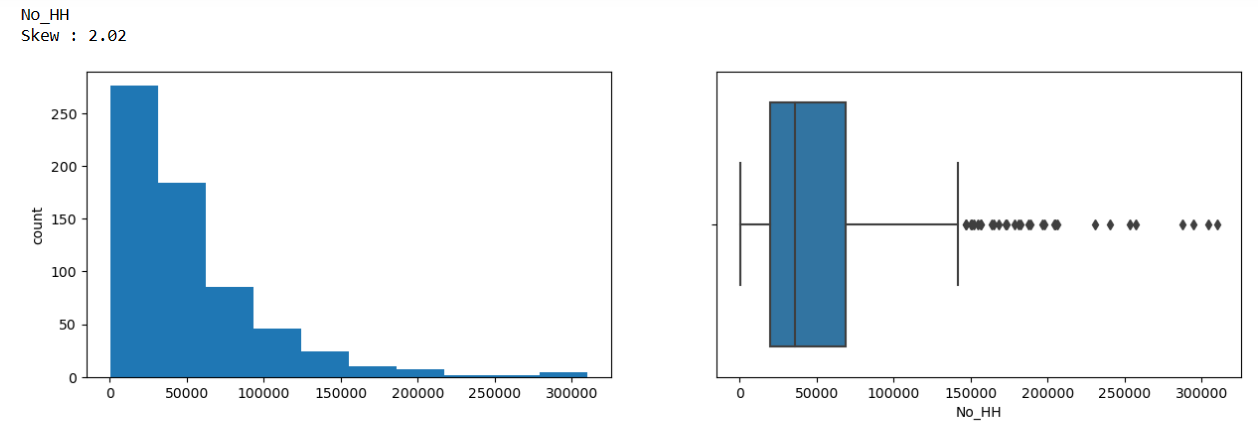
As there are 62 features in the dataset so we are performing EDA on only TOT\_M, TOT\_F,M\_SC,F\_SC,No\_HH.

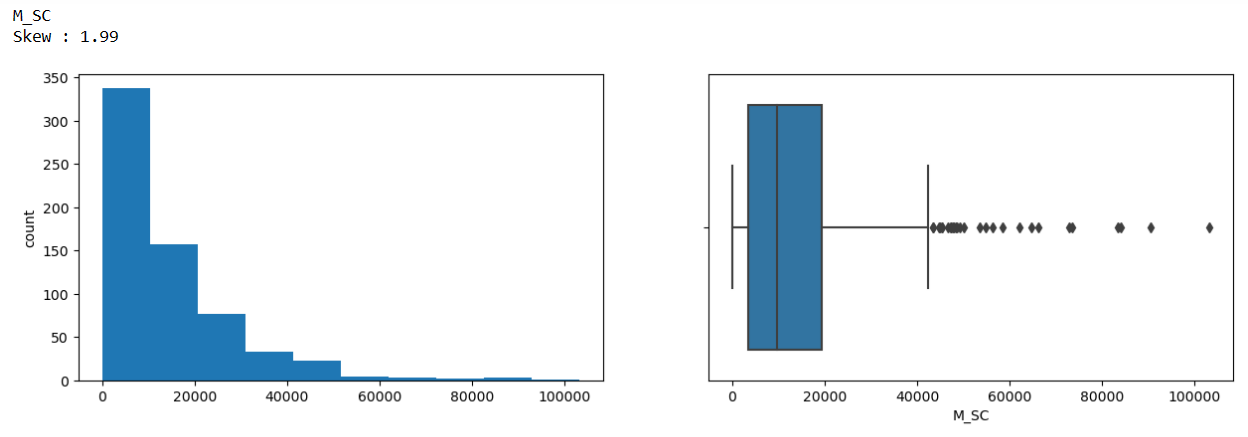
Univariate Analysis.

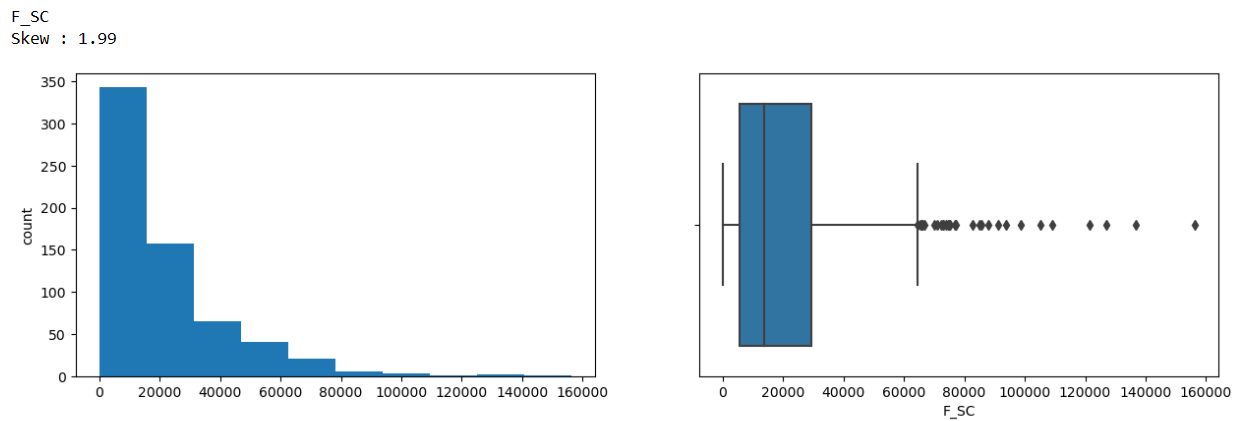
Here we will be doing only analysis for numeric data as in above given 24 varaiables there are only numeric variables.







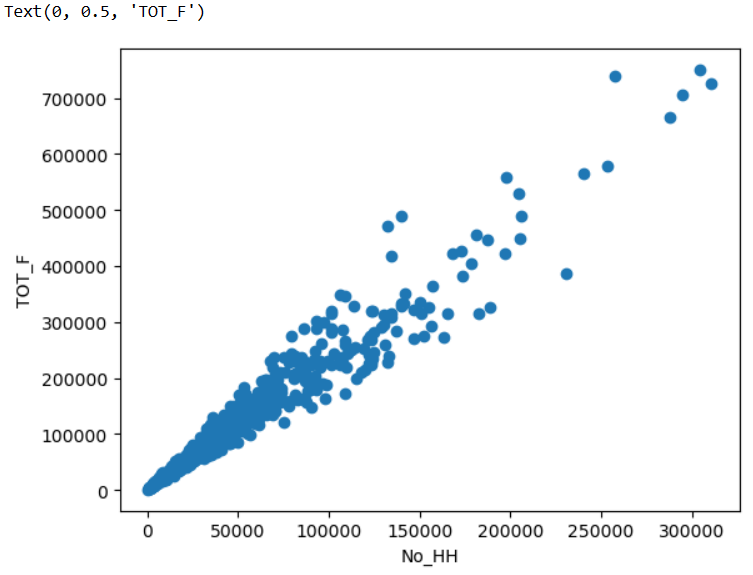




Observation:

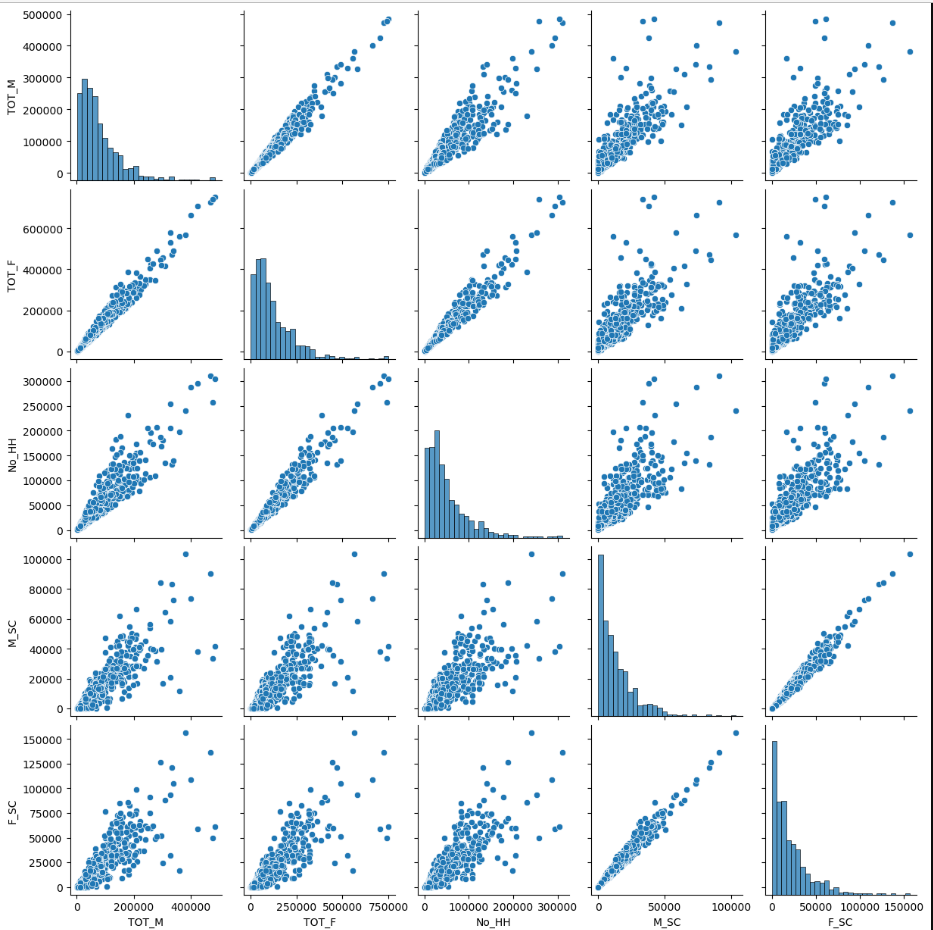
1. As observed, Most of the variables have skewed distributions.
2. All the given variables are highly right skewed and has outliers to right end.

Bivariate Analysis



We can see that Total population of Female and No of households are positively correlated.

As total population of female grows No of household also increases.



Insights:

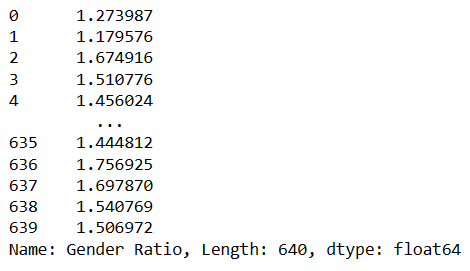
1. All the variables in the data EDA\_data are positively correlated to each other.



All the variables are strongly correlated with each other.

**Which state has highest gender ratio and which has the lowest**

We have calculated Gender ratio from total population of female to total population of Males.



State

Andhra Pradesh 1.895093

Chhattisgarh 1.869501

Tamil Nadu 1.861113

Odisha 1.830305

Arunachal Pradesh 1.757634

Maharashtra 1.743672

Nagaland 1.717555

Kerala 1.700197

Puducherry 1.673442

Uttarakhand 1.650205

Himachal Pradesh 1.616170

Goa 1.613884

Tripura 1.608367

Madhya Pradesh 1.587745

Andaman & Nicobar Island 1.581870

Mizoram 1.581292

Karnataka 1.578222

Manipur 1.563842

Dadara & Nagar Havelli 1.551275

West Bengal 1.541931

Sikkim 1.524218

Gujarat 1.513508

Jharkhand 1.481463

Rajasthan 1.470241

Assam 1.460374

Chandigarh 1.428496

Daman & Diu 1.427386

Jammu & Kashmir 1.380956

Meghalaya 1.366029

Bihar 1.344965

Punjab 1.339593

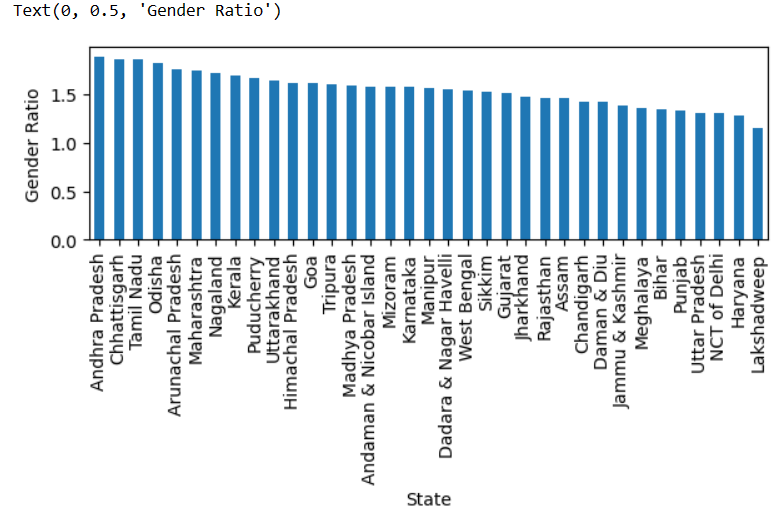
Uttar Pradesh 1.315773

NCT of Delhi 1.307783

Haryana 1.287548

Lakshadweep 1.151993

Name: Gender Ratio, dtype: float64



We can clearly say that Andra Pradesh has highest gender ratio and Lakshdweep has lowest gender ratio.

# Which district has the highest & lowest gender ratio

# 

# 

It is clear from the data that Krishna district has highest gender ratio and lakshdweep district has lowest gender ratio.

# Which state has highest and lowest number of Households

State

Kerala 141618.785714

West Bengal 137646.526316

Andhra Pradesh 135969.000000

Tamil Nadu 92646.875000

Maharashtra 89606.114286

Karnataka 75620.500000

Uttar Pradesh 56434.802817

Bihar 51963.868421

Gujarat 47507.692308

NCT of Delhi 43869.000000

Chhattisgarh 40812.000000

Odisha 40069.100000

Punjab 39801.500000

Goa 39797.500000

Rajasthan 35670.424242

Assam 31606.444444

Uttarakhand 30205.230769

Tripura 30161.750000

Madhya Pradesh 30075.280000

Jharkhand 28412.708333

Haryana 26951.000000

Himachal Pradesh 25443.083333

Chandigarh 24016.000000

Meghalaya 18120.714286

Puducherry 13487.500000

Manipur 9257.777778

Jammu & Kashmir 9256.590909

Nagaland 4856.636364

Mizoram 4633.625000

Lakshadweep 4445.000000

Andaman & Nicobar Island 4337.333333

Dadara & Nagar Havelli 4288.000000

Sikkim 4172.500000

Daman & Diu 3727.500000

Arunachal Pradesh 2194.625000

Name: No\_HH, dtype: float64



It is clear that Kerala has Highest number of households and Arunachal Pradesh has the lowest

# Which District has highest and lowest number of Households

# 

# 

North Twenty Four Parganas district has highest number of households and

Dibang valley has lowest number district has lowest number of households.

# 

# Which State has highest number of scheduled cast Male

State

West Bengal 45231.578947

Andhra Pradesh 26597.608696

Uttar Pradesh 25741.225352

Punjab 23850.700000

Karnataka 22146.166667

Tamil Nadu 20241.250000

NCT of Delhi 20175.333333

Kerala 19478.500000

Maharashtra 17455.057143

Bihar 15476.631579

Rajasthan 12088.727273

Haryana 11374.904762

Chandigarh 9499.000000

Himachal Pradesh 9290.833333

Uttarakhand 8298.846154

Odisha 8257.033333

Tripura 7815.750000

Madhya Pradesh 7040.260000

Chhattisgarh 6284.722222

Jharkhand 6264.875000

Gujarat 6234.846154

Assam 4171.814815

Puducherry 3203.250000

Jammu & Kashmir 2019.227273

Goa 1159.000000

Manipur 616.000000

Sikkim 336.500000

Daman & Diu 261.500000

Meghalaya 167.142857

Dadara & Nagar Havelli 158.000000

Mizoram 1.875000

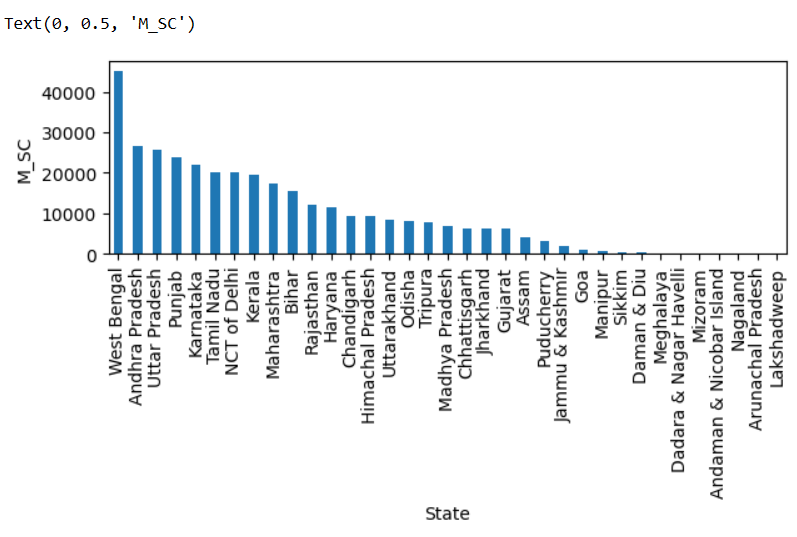
Andaman & Nicobar Island 0.000000

Nagaland 0.000000

Arunachal Pradesh 0.000000

Lakshadweep 0.000000

Name: M\_SC, dtype: float64



West Bengal has highest scheduled cast male and Laskshdweep has the lowest.

# Which State has highest Scheduled cast female

State

West Bengal 69293.210526

Andhra Pradesh 48850.826087

Tamil Nadu 36378.812500

Uttar Pradesh 34615.746479

Karnataka 33799.666667

Punjab 31646.900000

Kerala 31164.714286

Maharashtra 28867.800000

NCT of Delhi 24528.333333

Bihar 20889.921053

Rajasthan 16823.515152

Haryana 14395.380952

Odisha 14381.033333

Himachal Pradesh 13870.583333

Chandigarh 12155.000000

Tripura 12118.000000

Uttarakhand 11965.384615

Chhattisgarh 11780.777778

Madhya Pradesh 10794.640000

Jharkhand 8914.708333

Gujarat 8875.653846

Assam 5976.888889

Puducherry 5454.250000

Jammu & Kashmir 2737.727273

Goa 1769.500000

Manipur 950.333333

Sikkim 506.000000

Daman & Diu 354.000000

Dadara & Nagar Havelli 247.000000

Meghalaya 208.285714

Mizoram 2.500000

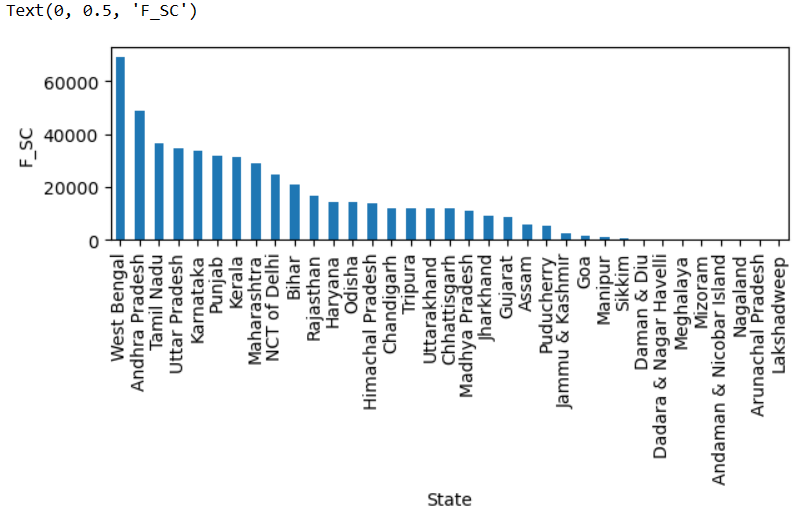
Andaman & Nicobar Island 0.000000

Nagaland 0.000000

Arunachal Pradesh 0.000000

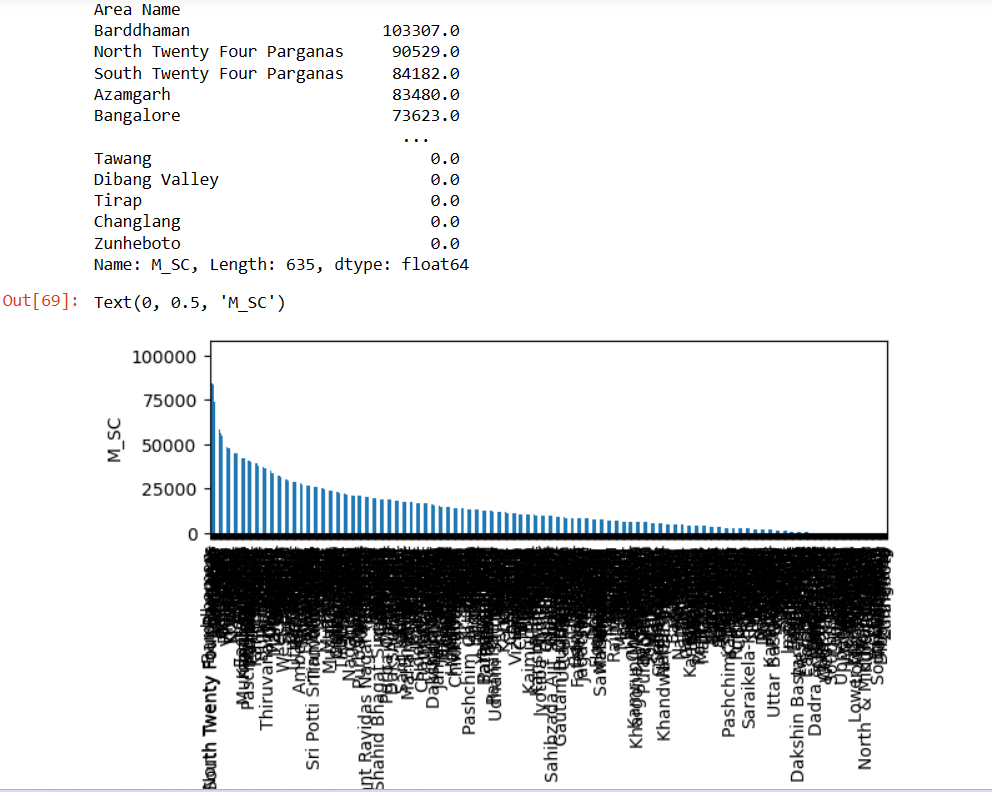
Lakshadweep 0.000000

Name: F\_SC, dtype: float64



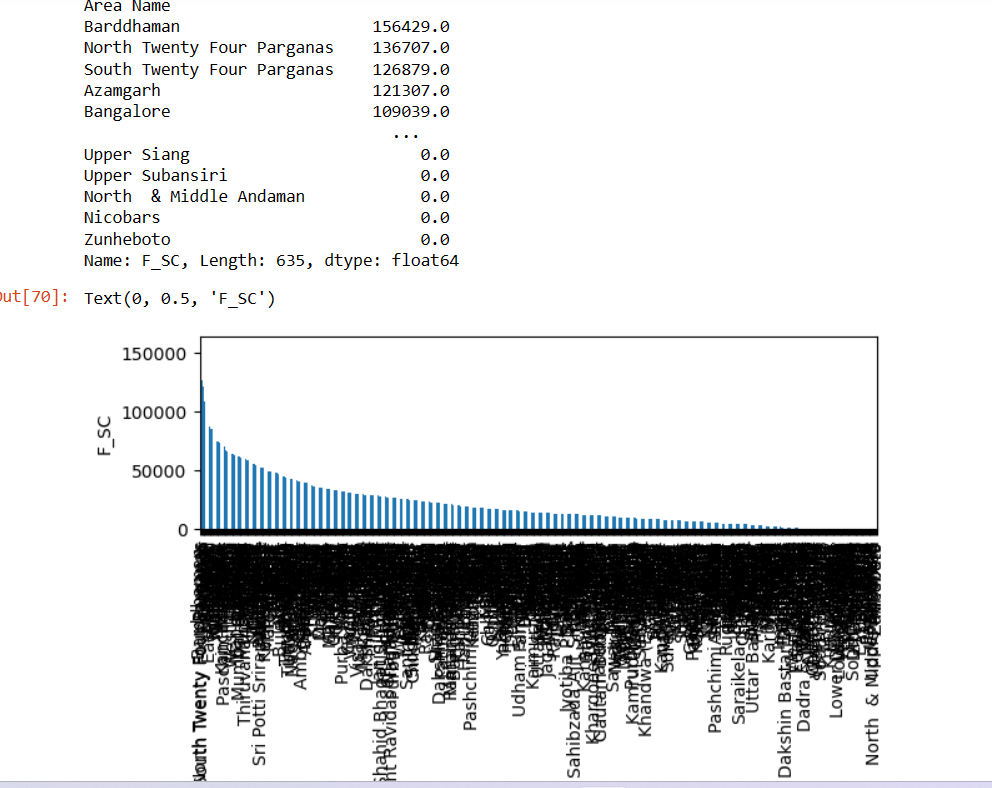
West Bengal has highest SC females and Lakshdweep has lowest SC females.

# Which District has highest scheduled cast males



Burddhaman has highest SC Males and Zunheboto has lowest SC Males.

# Which District has highest scheduled cast females



Burddhaman has highest SC Females and Zunheboto has lowest SC Females

**Part 2 - PCA: We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?**

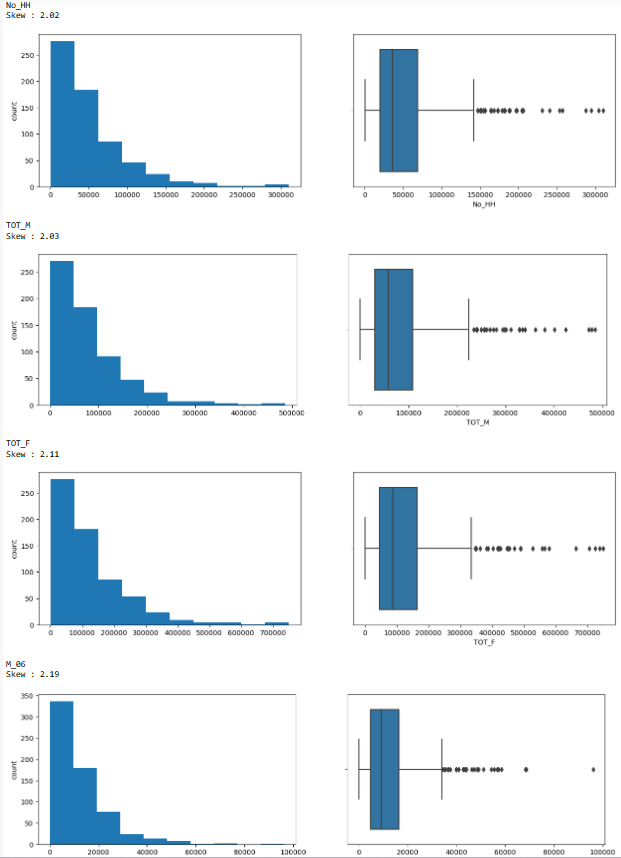
**Ans:-** In this case there we don’t need to Treat Outliers. Outliers in this case provides important information. Outlier can be removed when the values are incorrectly added or the value is wrong. But in our case Outlier provide crucial information.

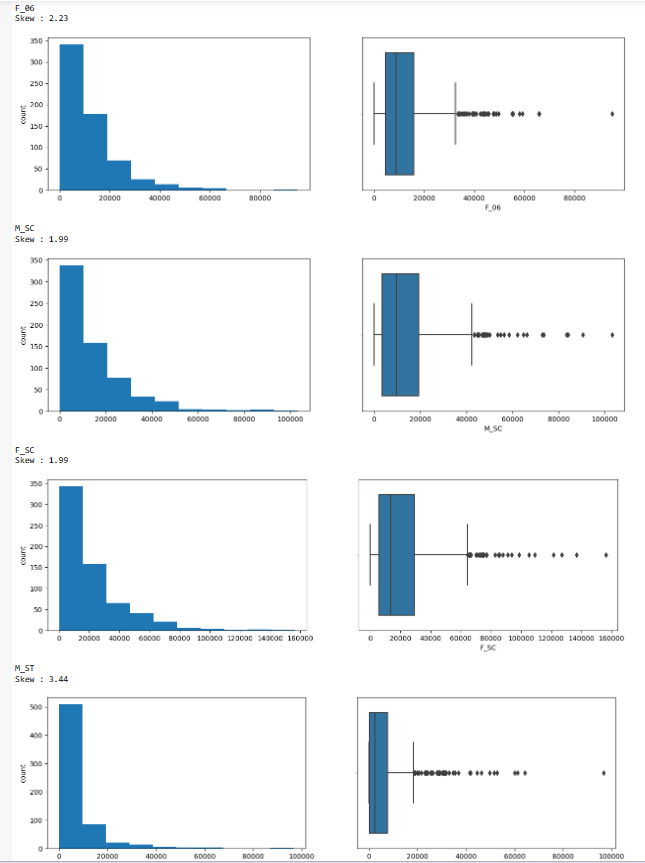
**Part 2 - PCA: Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.**

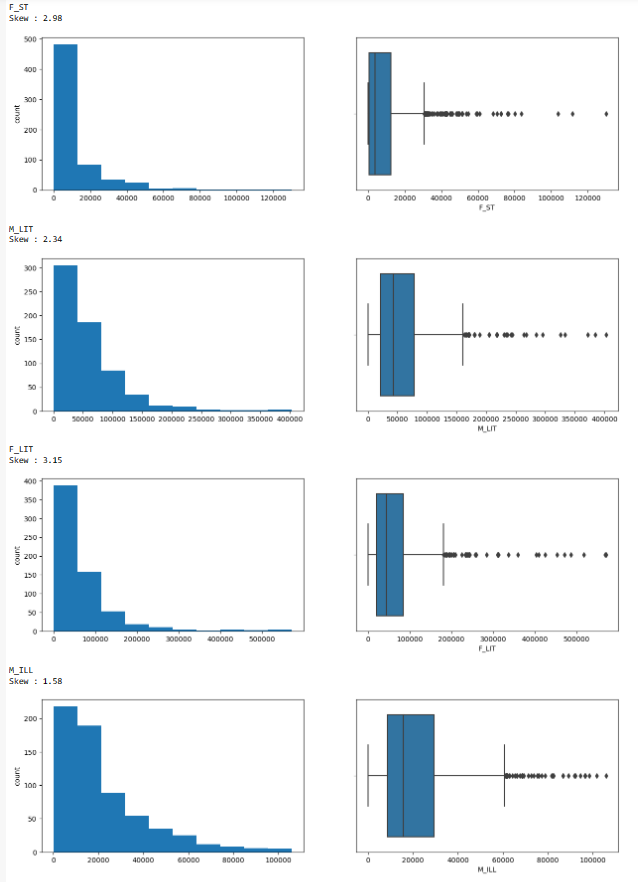
**Ans:-**

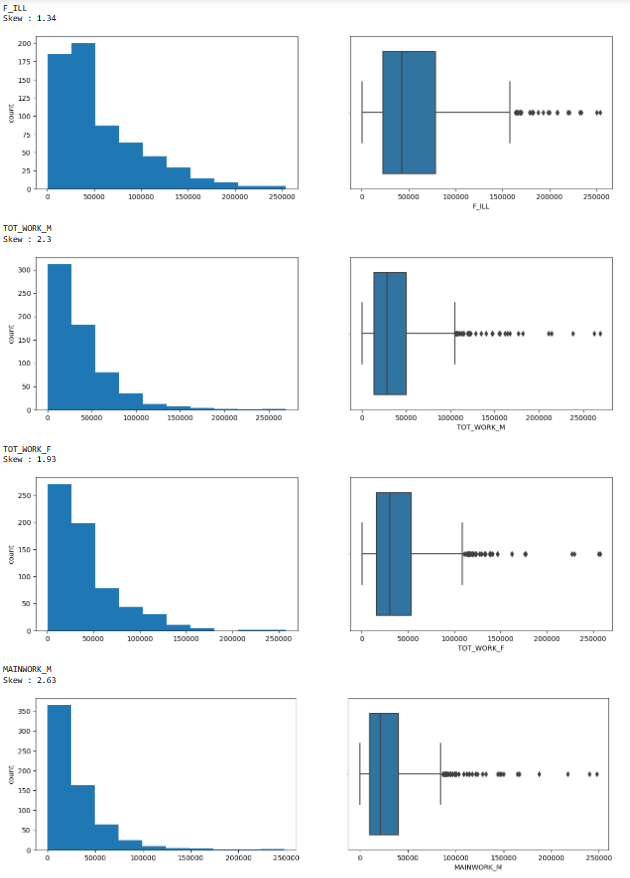
First of all we have created new DataFrame data\_df with only numeric values.

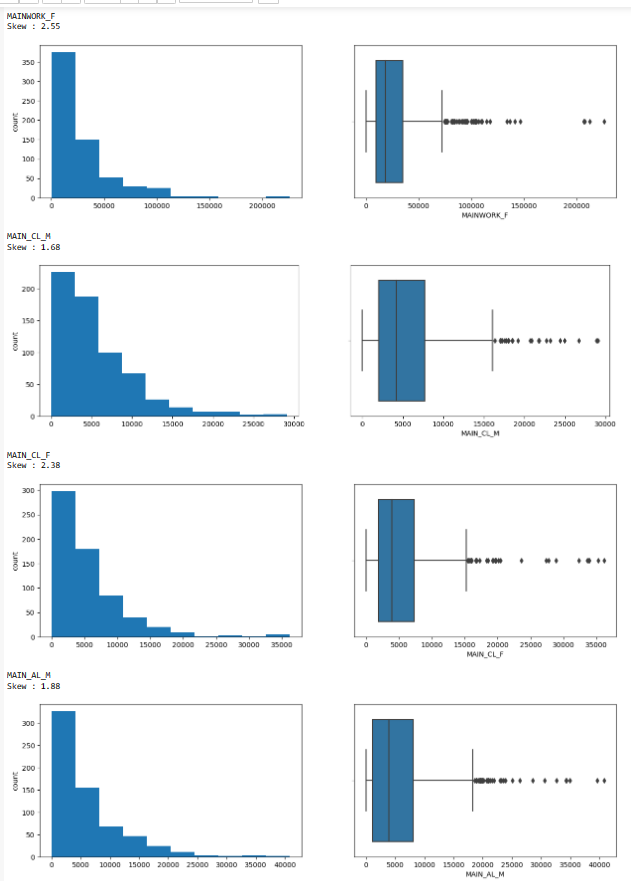
Outliers can be detected by visualization.

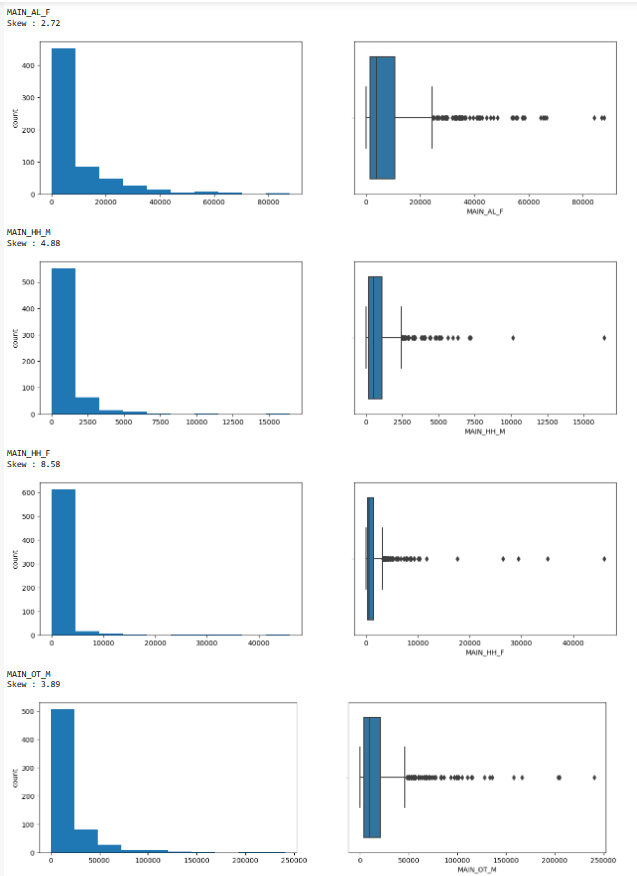


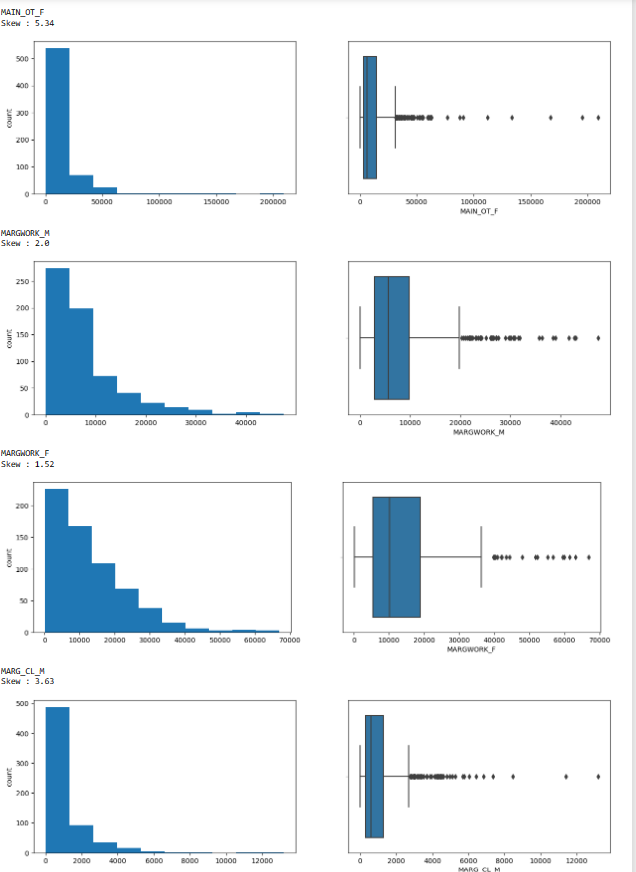


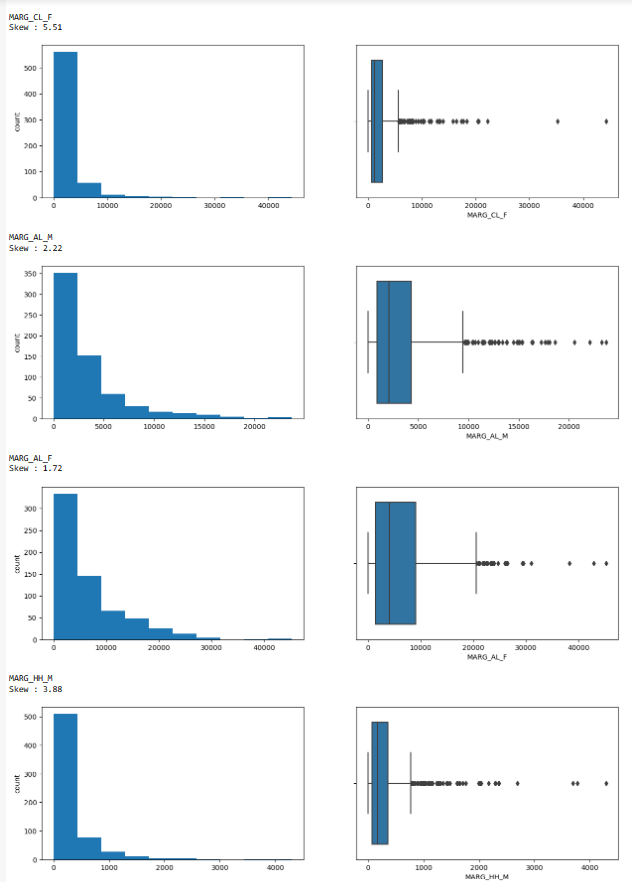


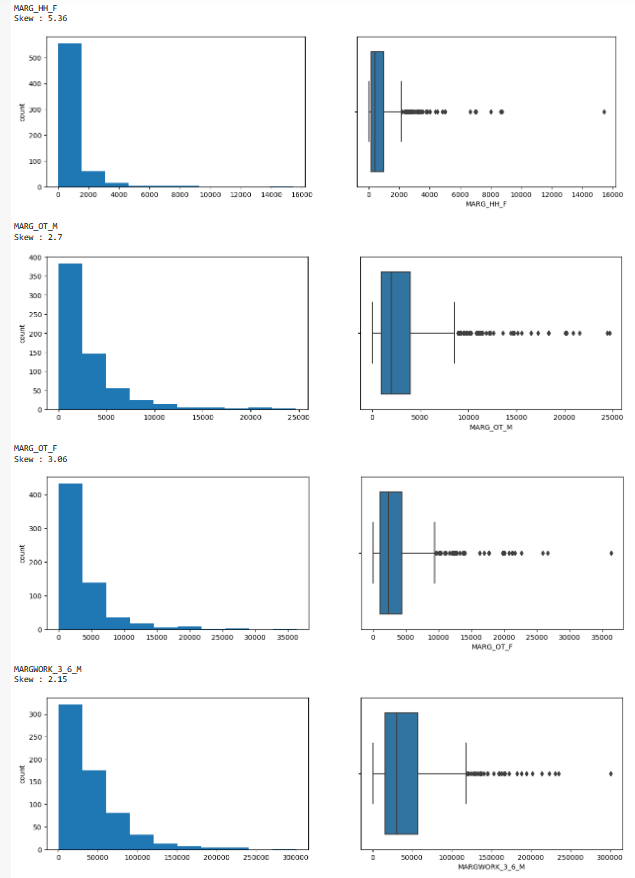


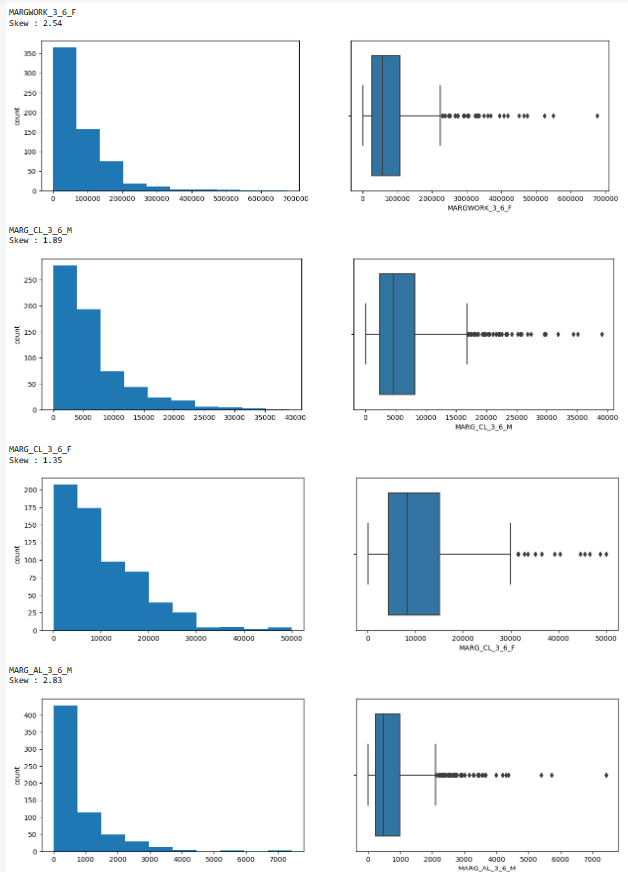


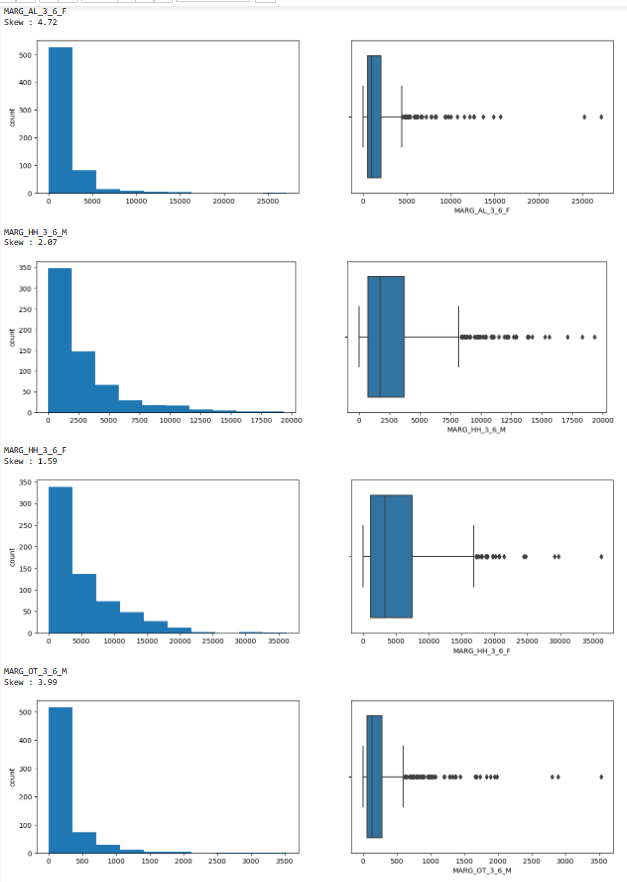


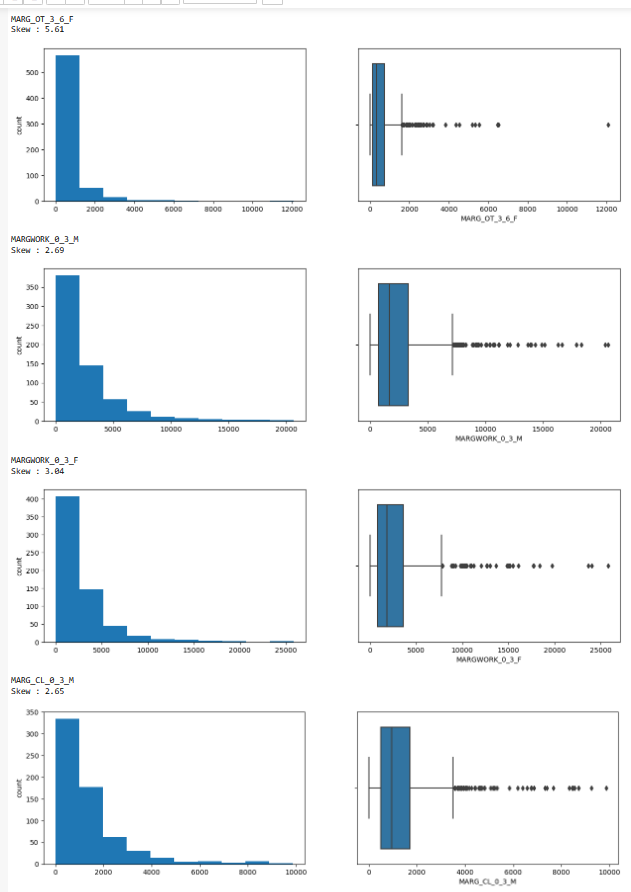


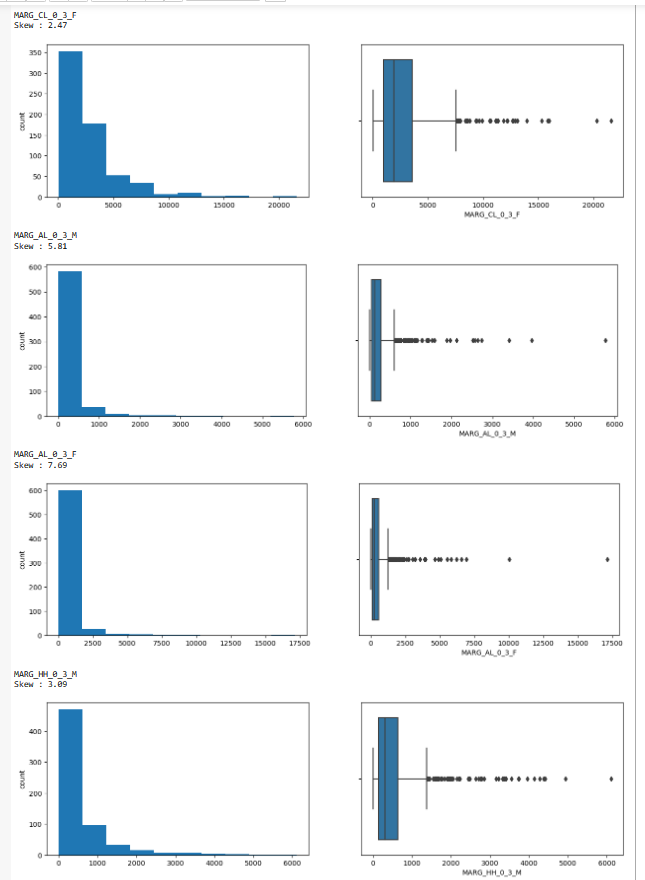


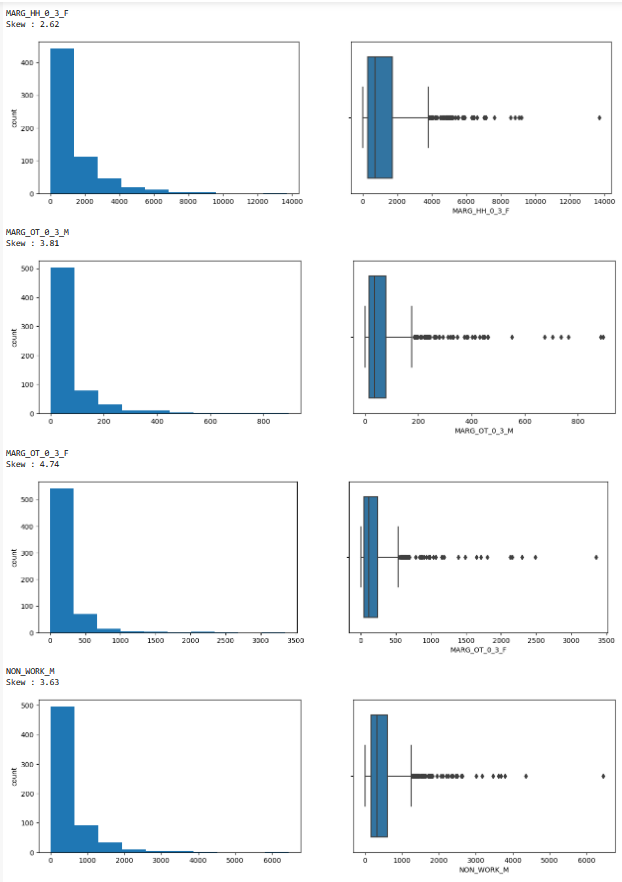


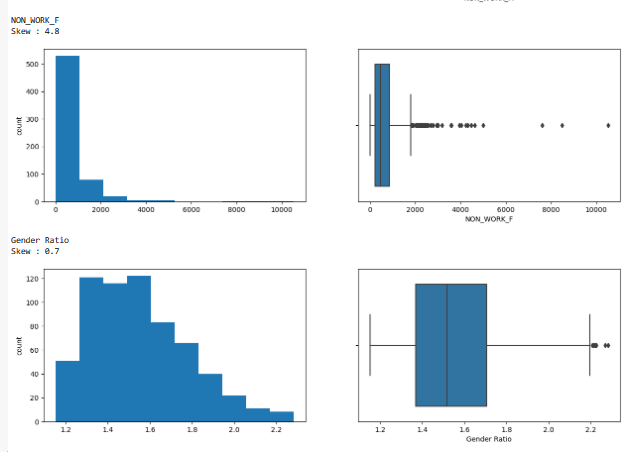




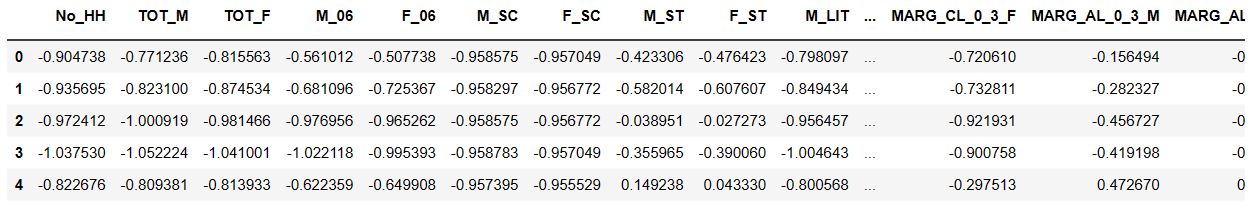




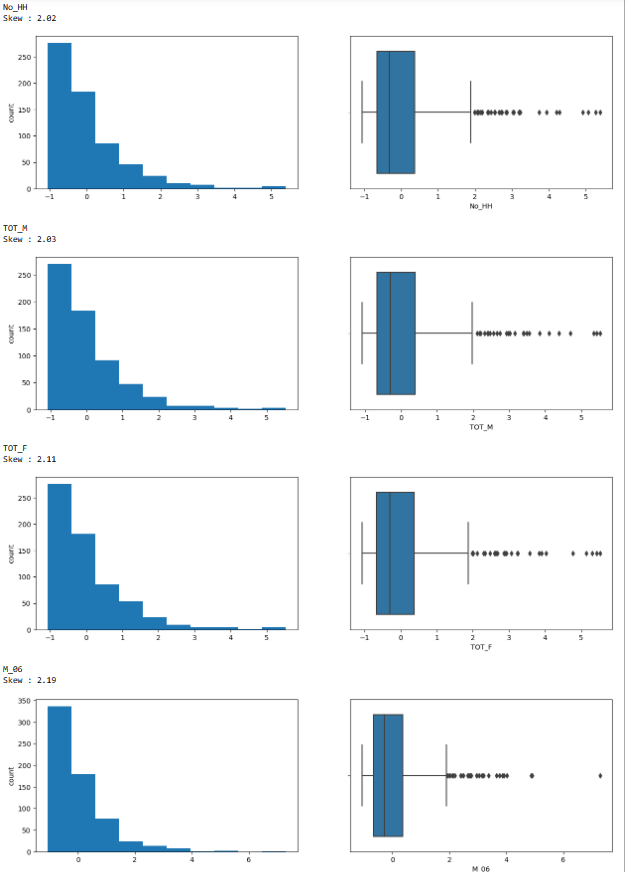


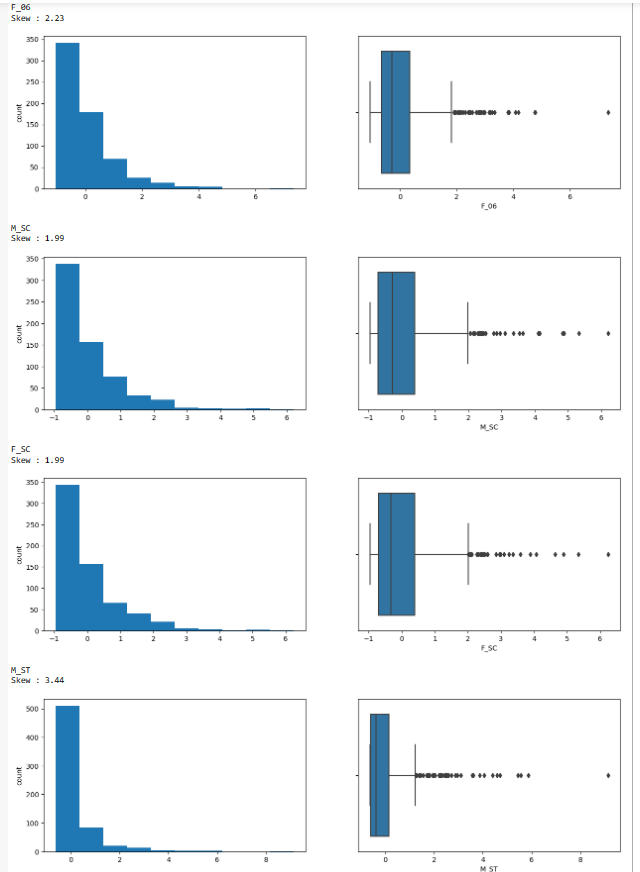


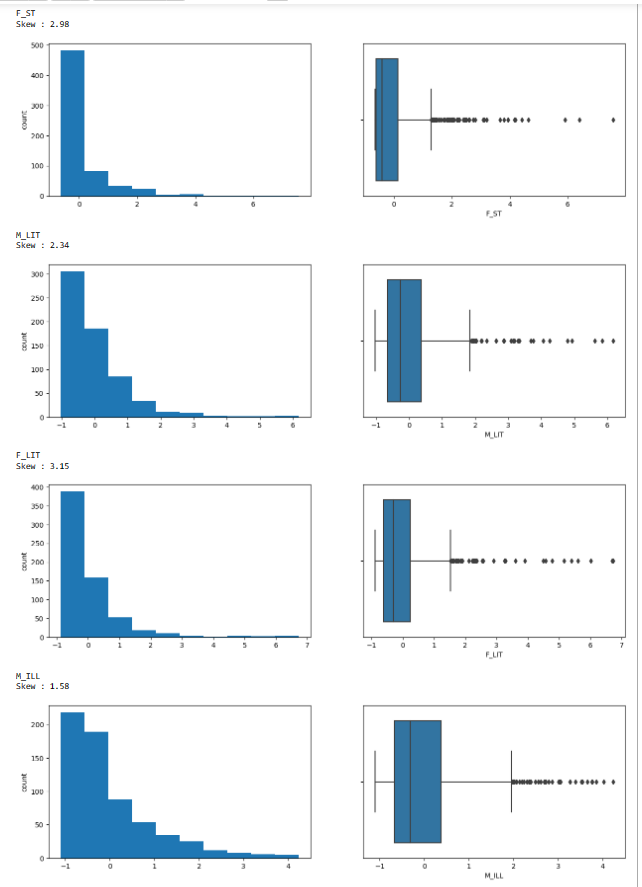
We have scaled the data using Zscore technique.

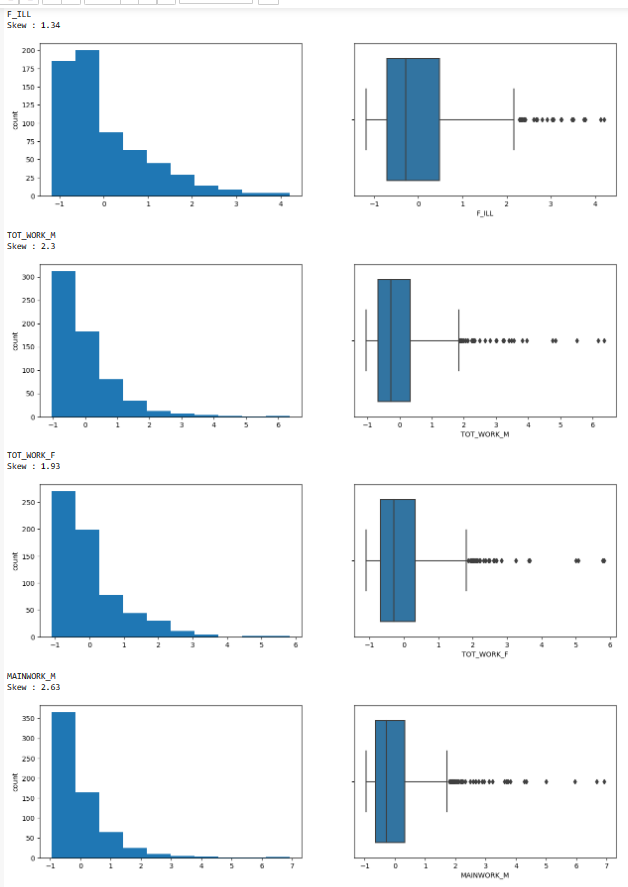


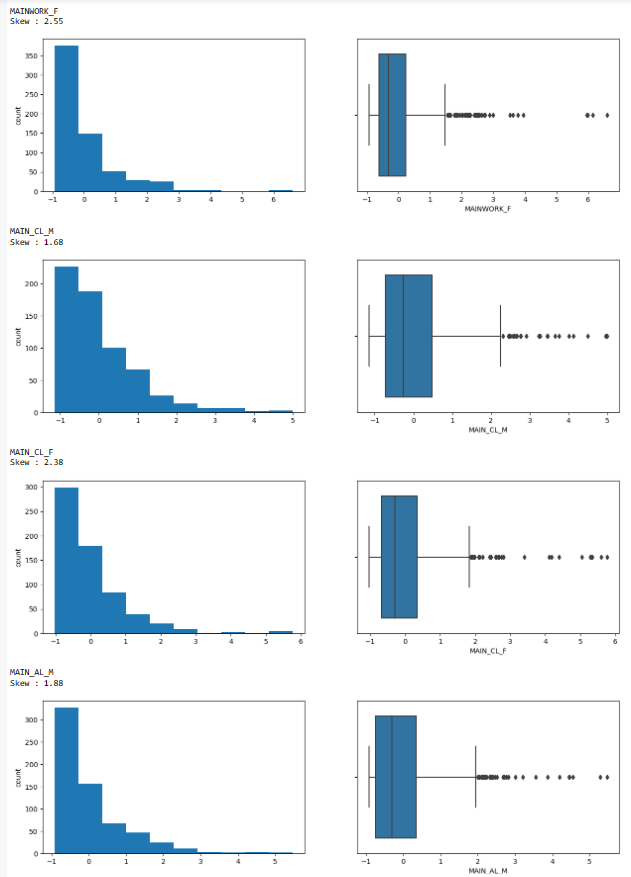
Checking for outliers on scaled data

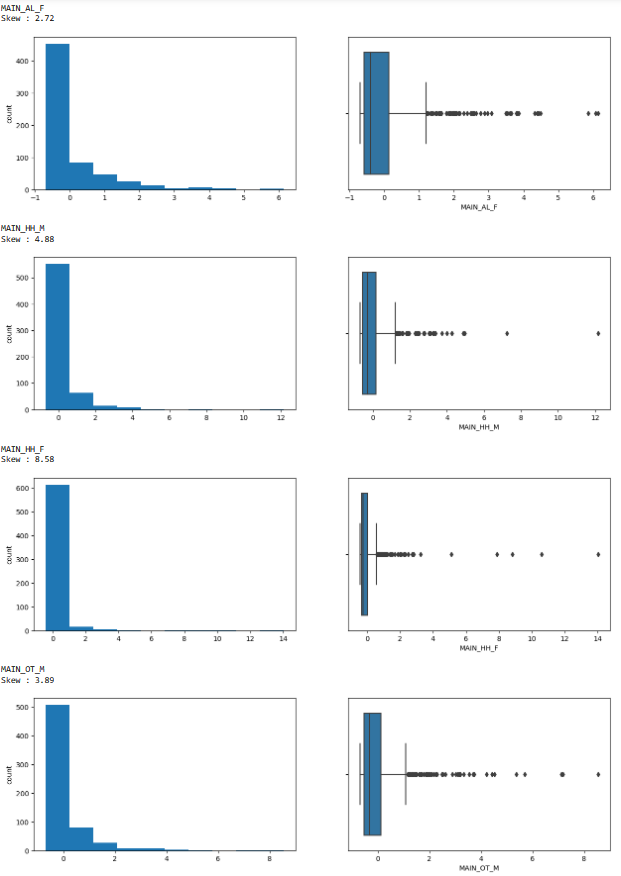




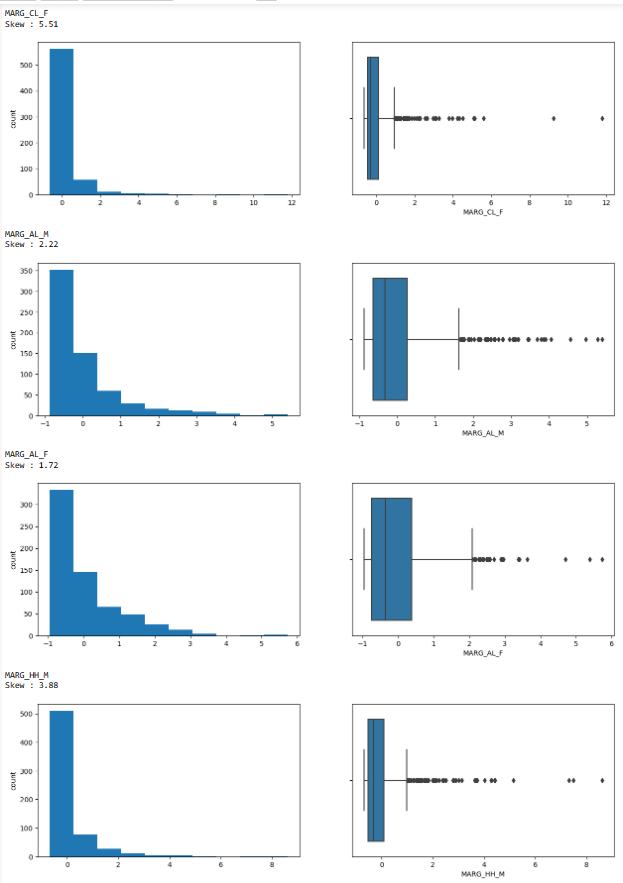


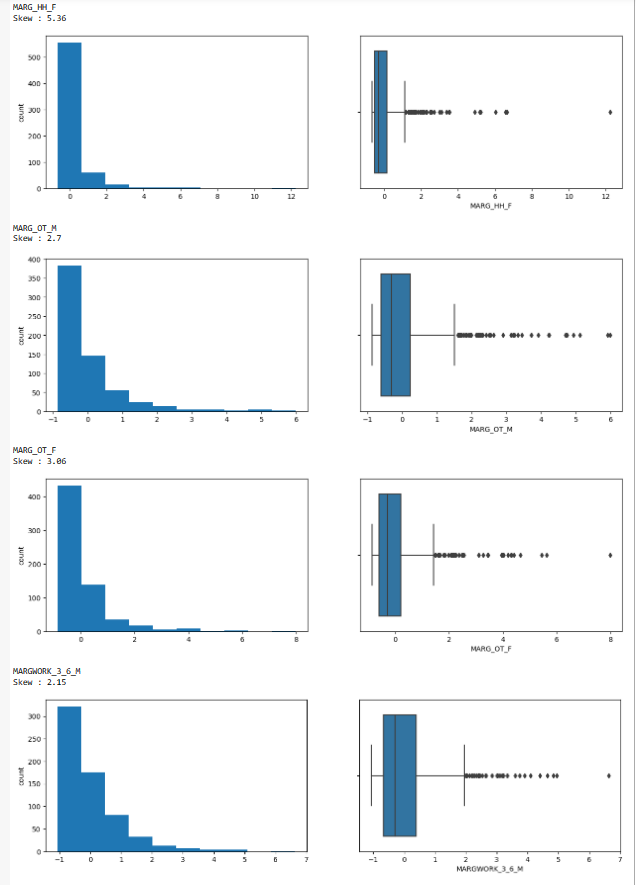


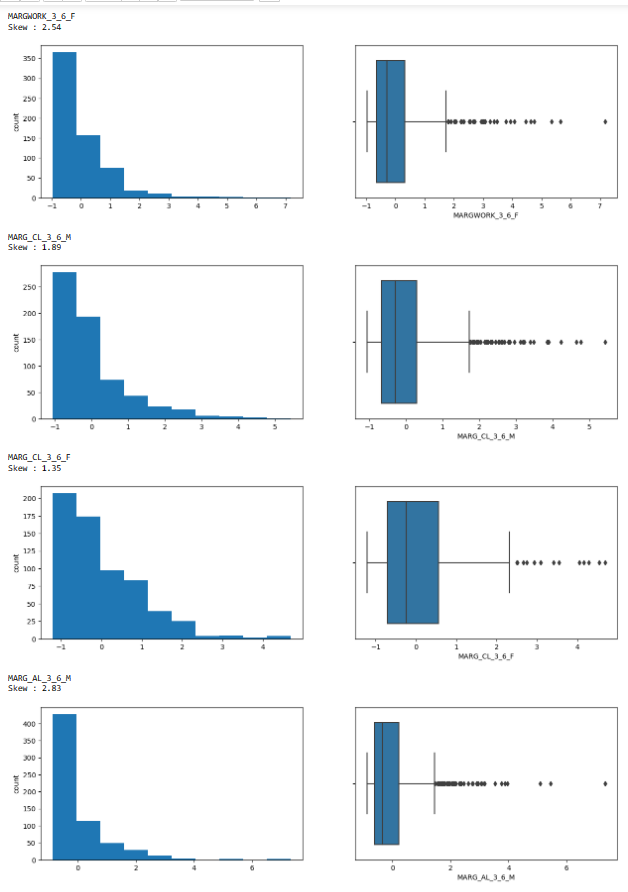


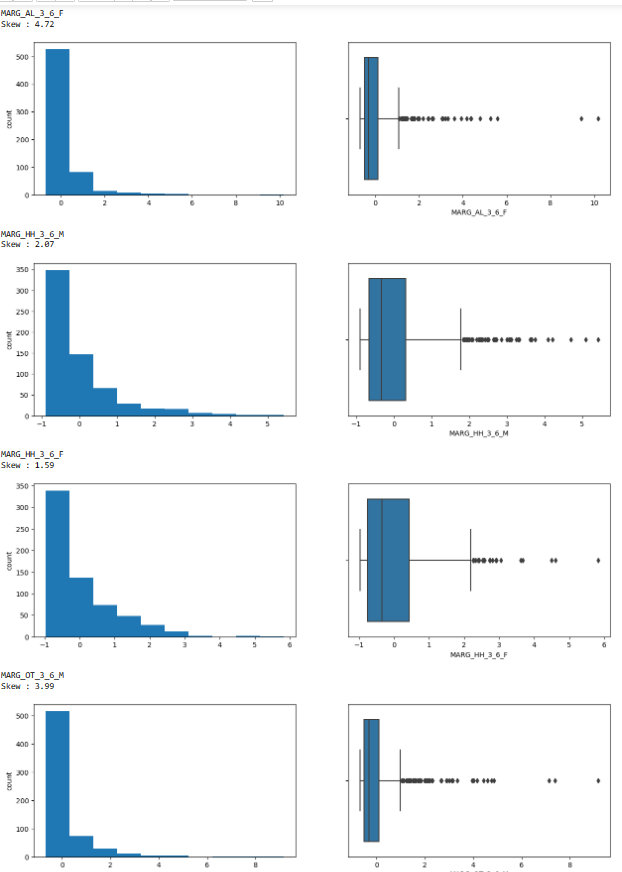


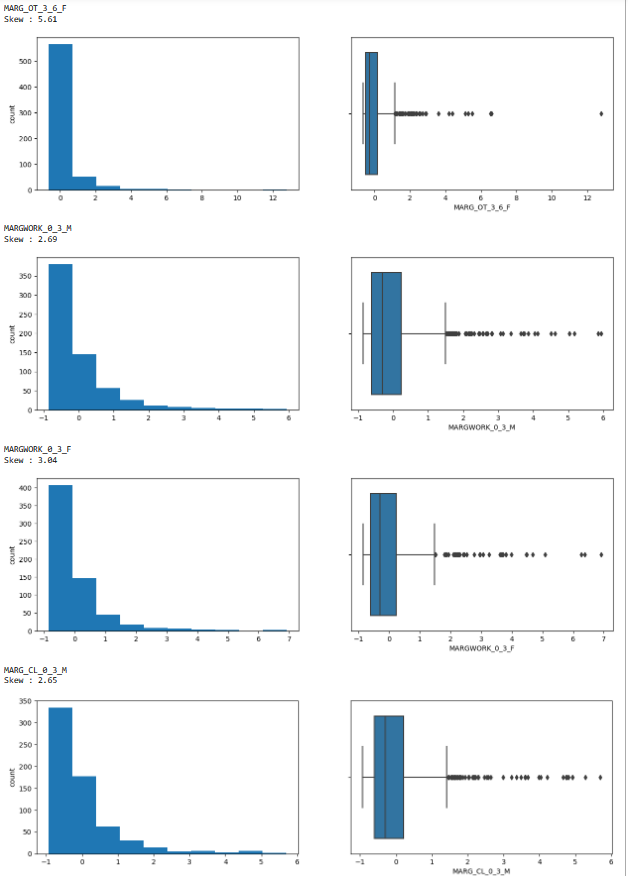


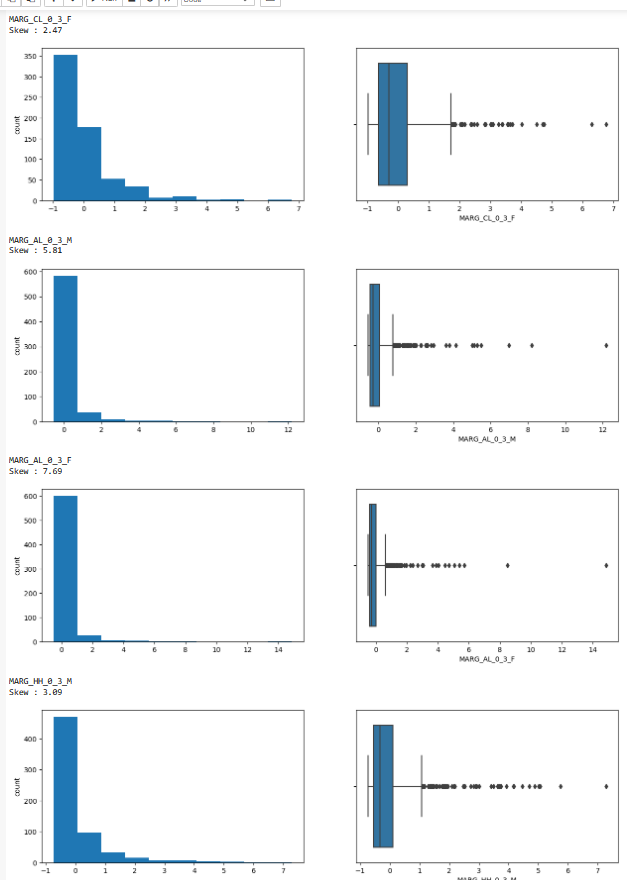


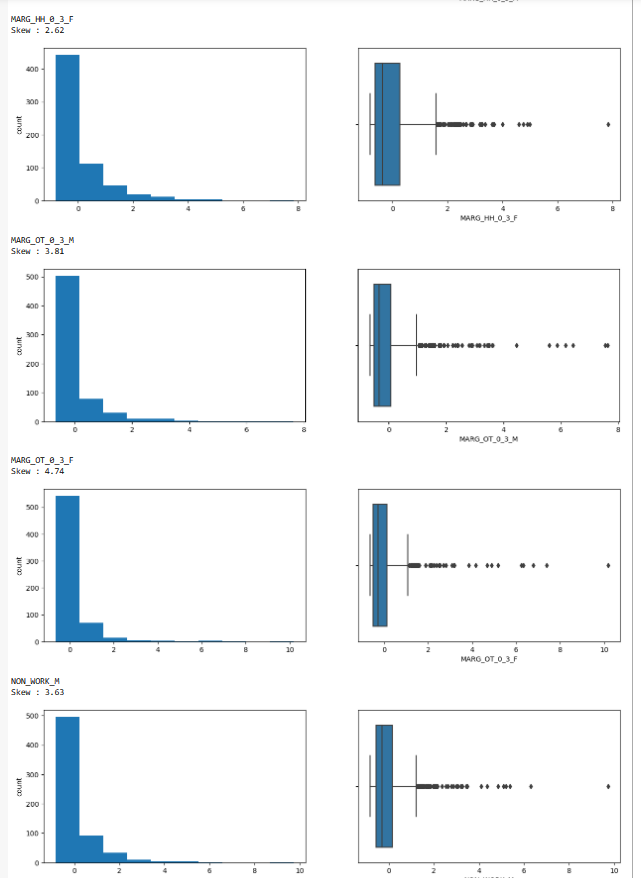


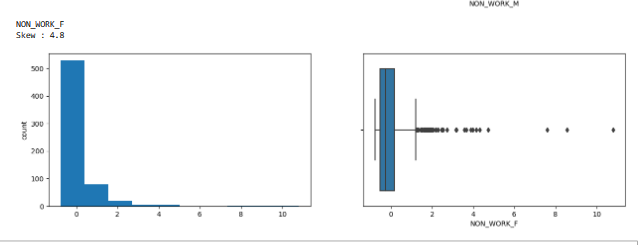


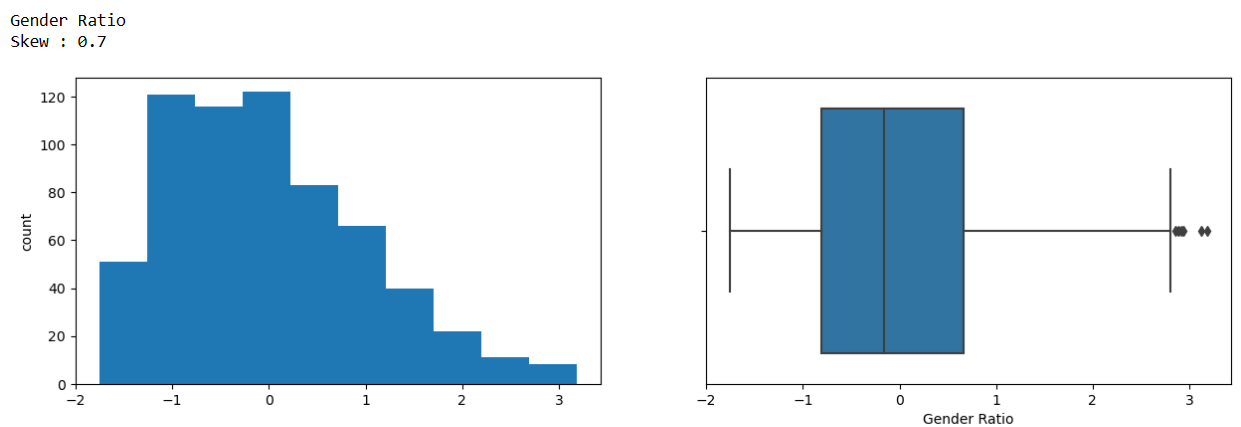












There is no impact on Outliers after scaling.Scaling makes all the variables with extreme values on level inshort it standardizes the data.All the variable which had outliers before scaling has outliers after scaling. There is no change as such.

**PCA: Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector(4 Marks)**

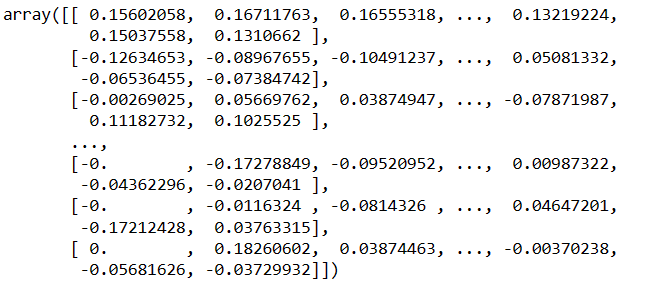
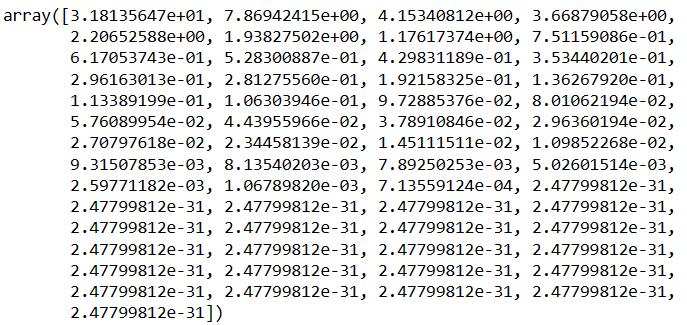
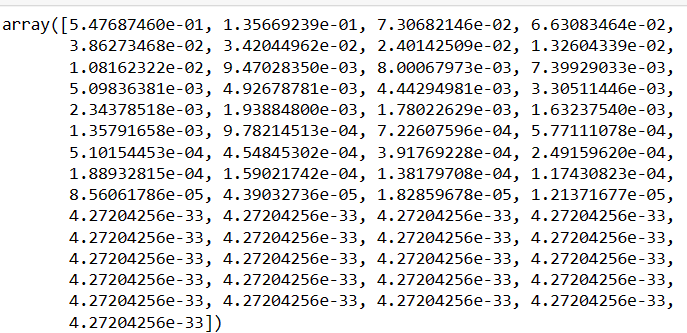
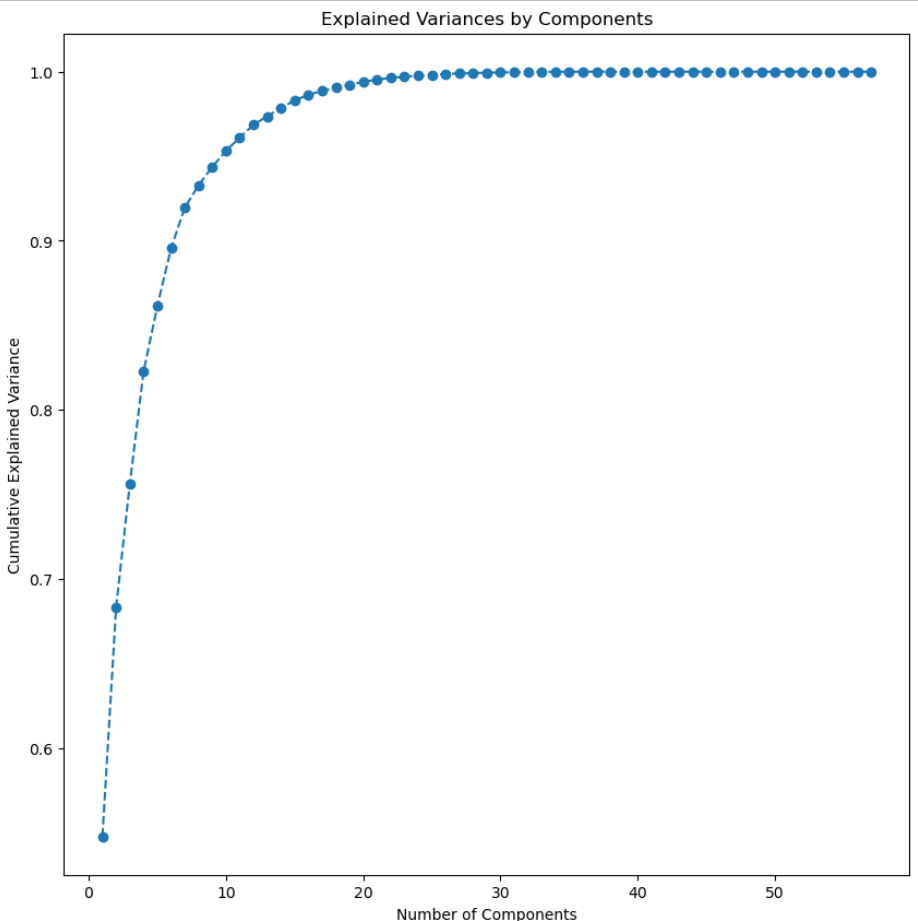
**Ans:-**

We will first Confirm the statistical significance of correlations.

H0: Correlations are not significant, H1: There are significant correlations

Reject H0 if p-value < 0.05

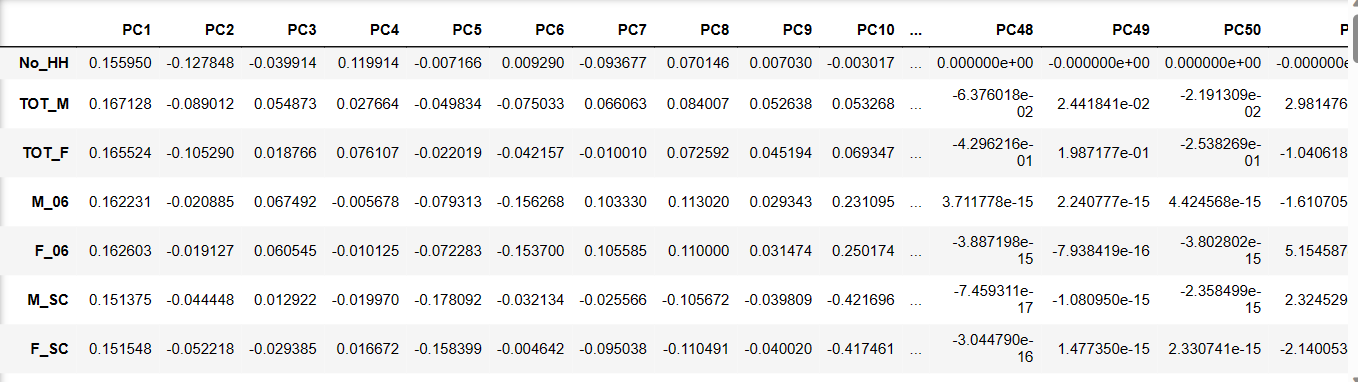
Our p-value is 0 so we can go ahead.

1. We will first perform fit\_transform function on scaled which will give us Eigen vectors.
2. We have perform PCA on all the features of dataset (57 variable)
3. 
4. We have eigen values which are always returned in array format in descending order.
5. 
6. We will Check the explained variance for each PC
7. Explained variance = (eigen value of each PC)/(sum of eigen values of all PCs)
8. 
9. 
10. We can clearly see that cumulative Explained Variance increases as the Number of Components increases.

**Part 2 - PCA: Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.(3 Marks)**

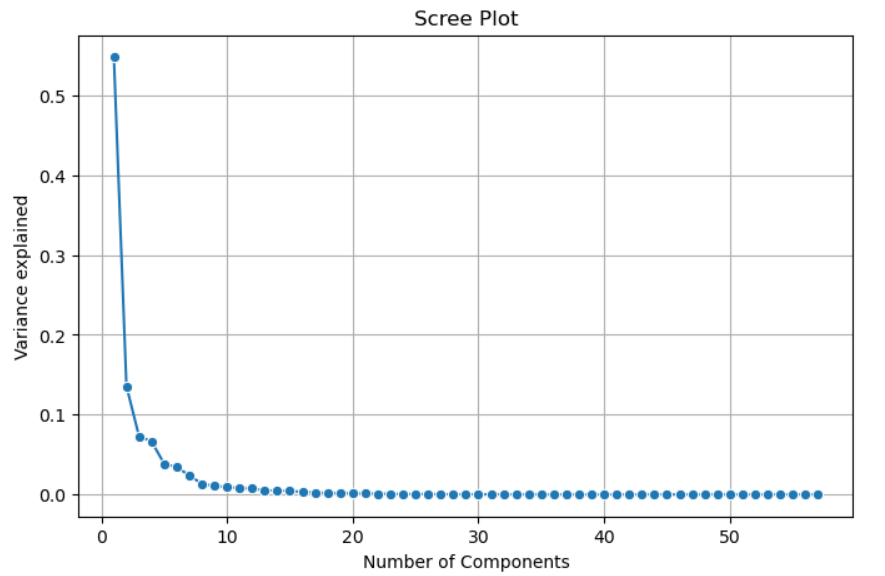
**Ans:-**

We have now created a dataframe containing coefficients of all the Principal components.



57 PCs were created initially.

Lets create Scree Plot on explained\_variance\_ratio



From the graph we can Conclude that Variance explained’ s value decreases as the number of componenst increases.They have negative correlation between them.

Lets find the least number of components that can explain more than 90% variance

For that first we will calculate number of components with more than 90% variance using a function.

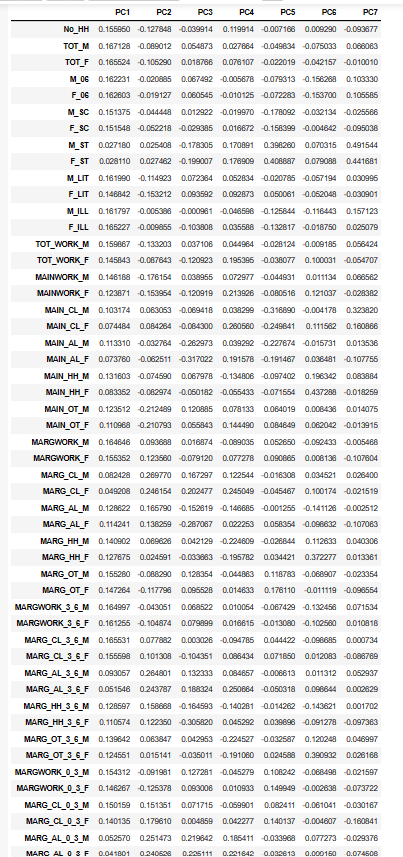


We got 7 as an output.

We will also check cumlative explained variance ratio to find a cut off for selecting the number of PCs



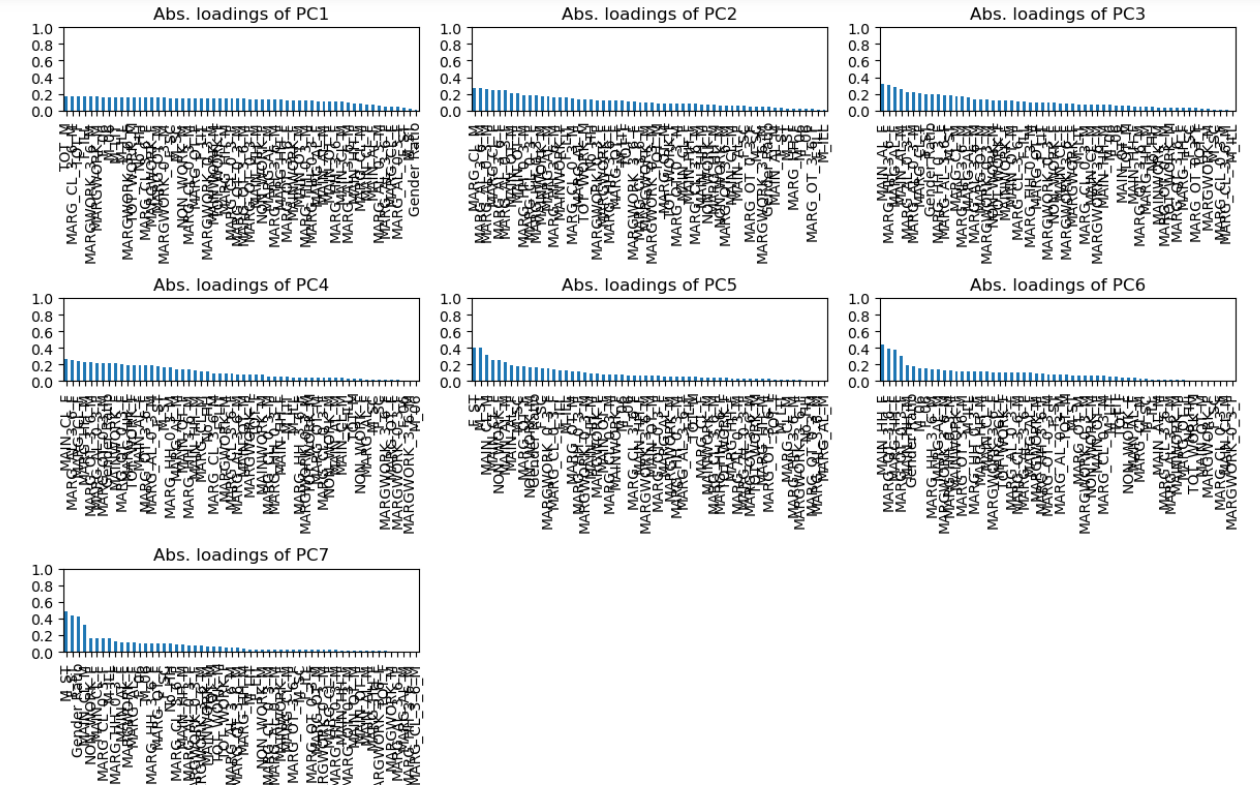
We will choose the Pcs basis cumulative explained variance



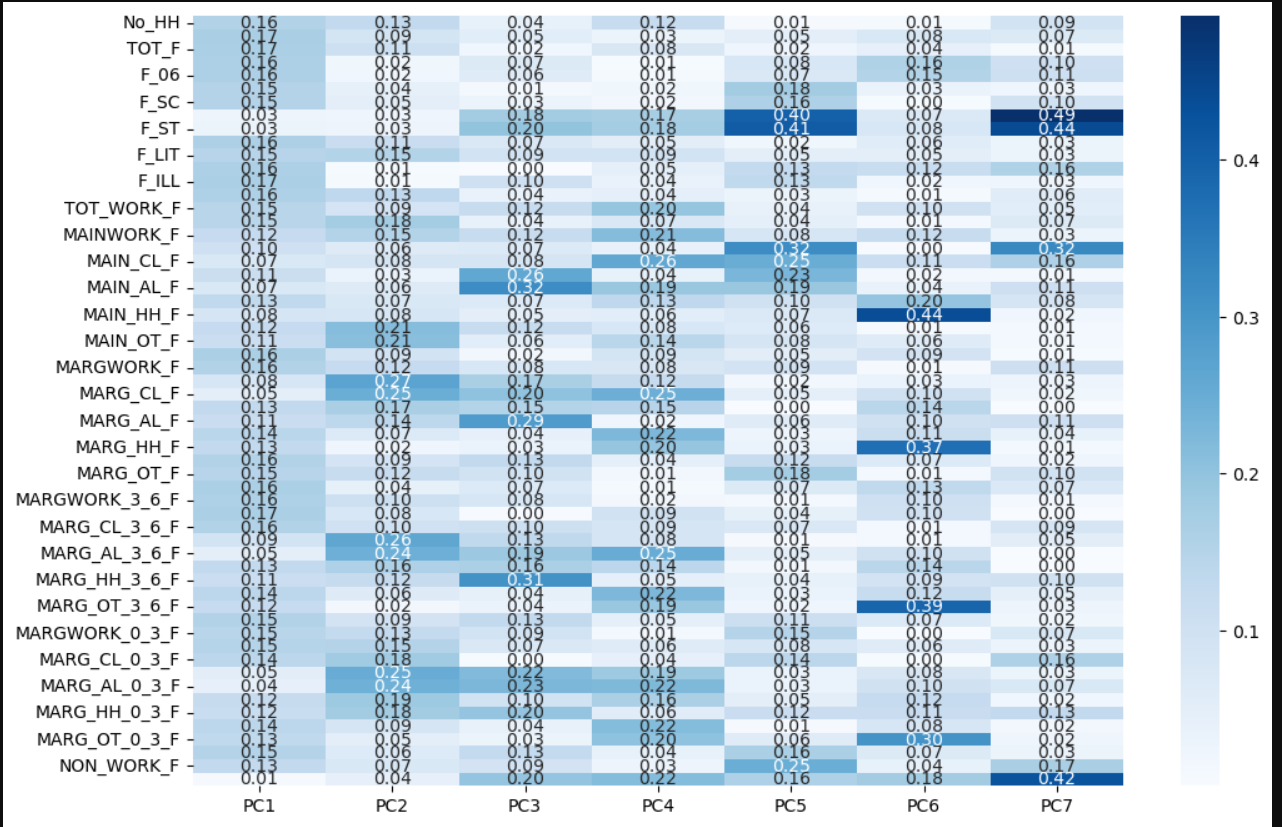
# Part 2 - PCA: Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the Principal components in terms of actual variables(4 Marks)

# Ans:-

We will check how original features matter to the PC



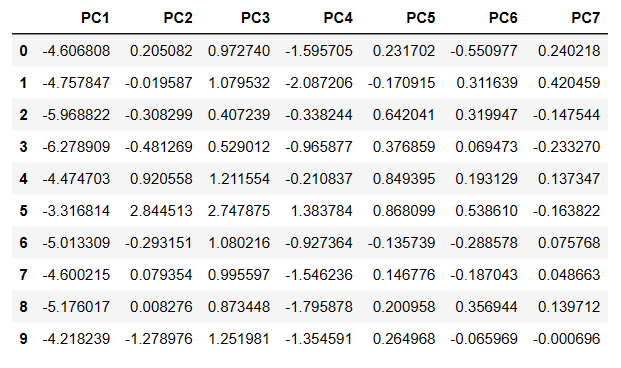
Lets Compare how the original features influence various PCs



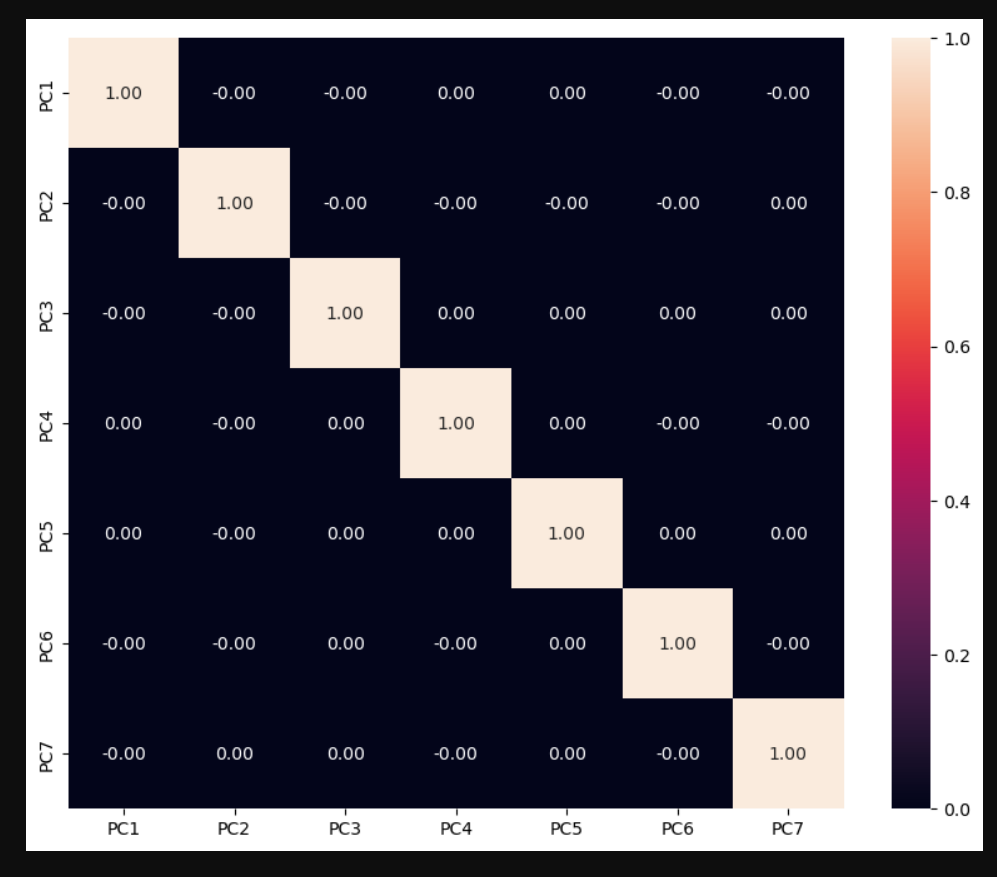
Above data did not decrease the collinearity.

We will Extract the required(as per the cumulative explained variance) number of PCs.

Just create a dataframe out of fit\_transformed for the scaled data.



Check for presence of correlations among teh PCs



There is no correlation between among the PCs. So we can say that 7 has 90% of the data.

We have reduced the data from 57 to 7 dimensions.

**Part 2 - PCA: Write linear equation for first PC.(2Marks)**

(0.16 ) \* No\_HH + ( 0.17 ) \* TOT\_M + ( 0.17 ) \* TOT\_F + ( 0.16 ) \* M\_06 + ( 0.16 ) \* F\_06 + ( 0.15 ) \* M\_SC + ( 0.15 ) \* F\_SC + ( 0.03 ) \* M\_ST + ( 0.03 ) \* F\_ST + ( 0.16 ) \* M\_LIT + ( 0.15 ) \* F\_LIT + ( 0.16 ) \* M\_ILL + ( 0.17 ) \* F\_ILL + ( 0.16 ) \* TOT\_WORK\_M + ( 0.15 ) \* TOT\_WORK\_F + ( 0.15 ) \* MAINWORK\_M + ( 0.12 ) \* MAINWORK\_F + ( 0.1 ) \* MAIN\_CL\_M + ( 0.07 ) \* MAIN\_CL\_F + ( 0.11 ) \* MAIN\_AL\_M + ( 0.07 ) \* MAIN\_AL\_F + ( 0.13 ) \* MAIN\_HH\_M + ( 0.08 ) \* MAIN\_HH\_F + ( 0.12 ) \* MAIN\_OT\_M + ( 0.11 ) \* MAIN\_OT\_F + ( 0.16 ) \* MARGWORK\_M + ( 0.16 ) \* MARGWORK\_F + ( 0.08 ) \* MARG\_CL\_M + ( 0.05 ) \* MARG\_CL\_F + ( 0.13 ) \* MARG\_AL\_M + ( 0.11 ) \* MARG\_AL\_F + ( 0.14 ) \* MARG\_HH\_M + ( 0.13 ) \* MARG\_HH\_F + ( 0.16 ) \* MARG\_OT\_M + ( 0.15 ) \* MARG\_OT\_F + ( 0.16 ) \* MARGWORK\_3\_6\_M + ( 0.16 ) \* MARGWORK\_3\_6\_F + ( 0.17 ) \* MARG\_CL\_3\_6\_M + ( 0.16 ) \* MARG\_CL\_3\_6\_F + ( 0.09 ) \* MARG\_AL\_3\_6\_M + ( 0.05 ) \* MARG\_AL\_3\_6\_F + ( 0.13 ) \* MARG\_HH\_3\_6\_M + ( 0.11 ) \* MARG\_HH\_3\_6\_F + ( 0.14 ) \* MARG\_OT\_3\_6\_M + ( 0.12 ) \* MARG\_OT\_3\_6\_F + ( 0.15 ) \* MARGWORK\_0\_3\_M + ( 0.15 ) \* MARGWORK\_0\_3\_F + ( 0.15 ) \* MARG\_CL\_0\_3\_M + ( 0.14 ) \* MARG\_CL\_0\_3\_F + ( 0.05 ) \* MARG\_AL\_0\_3\_M + ( 0.04 ) \* MARG\_AL\_0\_3\_F + ( 0.12 ) \* MARG\_HH\_0\_3\_M + ( 0.12 ) \* MARG\_HH\_0\_3\_F + ( 0.14 ) \* MARG\_OT\_0\_3\_M + ( 0.13 ) \* MARG\_OT\_0\_3\_F + ( 0.15 ) \* NON\_WORK\_M + ( 0.13 ) \* NON\_WORK\_F +