Stat 467 Term Project

PROJECT REPORT SUBMITTED
IN FULFILMENT OF THE REQUIREMENTS FOR COURSE
STAT 467 – MULTIVARIATE ANALYSIS
DEPARTMENT OF STATISTICS OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

Berkay Türetken 2502433

Sidar Yük 2291052

January 2024

ABSTRACT

This study investigates family planning in India and if it has effects on different realities of life and vice versa in India, from children being well fed to the government on being sufficient about family planning. A research about a topic like this is important for understanding how India uses family planning methods to tackle it's issues amongst it's people and if it is affected by other variables. Multiple analyses have been done on the data of this research for the purpose of analysing the relationships between multiple factors. It can be said that while some results can surprise the reader some were fairly logical and thinkable beforehand and it can be said that much or less the efforts of family planning in India are worth a while.

1. Introduction

Family planning has an important role on the socio-economic development and balance, welfare of the people and public health of any country. For a country like India which has the largest population in the world ,which is 1.4 billion, family planning is much more critical then someone can imagine. Knowing that, the Ministry of Health and Family Welfare of India conducted a survey that took 2 years to complete from 2019 to 2021 across all 28 States and 8 Union Territories which took place in 707 districts. This survey provides important information about the general health, nutrition, welfare, usage of family planning, education, of the Indian people. In the survey the Ministry of Health and Family Welfare also asks if the actions taken by the government workers about family planning is enough or not for encouraging the people to start using modern methods of family planning.

This study utilizes a comprehensive approach, combining multiple analyses to divide and explain the relationship between different factors that are either affecting the usage of family planning or be an outcome of family planning itself in India. The methods that were used during this study are as follows, Comparison of Several Multivariate Means for testing

whether the response variables that were selected by us varied with respect to the corresponding state or union territory, Principal Component Analysis for interpreting the data whilst simplifying it for us and finding the trends and patterns amongst the variables, Principal Component Regression for estimating the unknown coefficients of a linear regression model, Factor Analysis and Rotation to summarize the relationships of the variables and minimize the complexity, Discrimination and Classification for finding the linear combinations of variables and grouping them, and finally Clustering for discovering the structure that the data has within itself from the start.

1.1 Data Description

The data set that was used for this study has 707 rows and 66 columns. From these 66 columns 2 of them are categorical variables, 4 of them are discrete variables and the remaining 60 columns are continuous variables. The data is the result of a survey conducted by the Ministry of Health and Family Welfare of India the survey started in 2019 and ended in 2021.

1.2 Research Questions

The main question that was asked in this study is does family planning affect different aspects of life, if it does how and vice versa. Also another question that the study asks is actually if the governments policies and government workers have an influence on the use of family planning and its results. And if the response variable varies with respect to state/united territory.

1.3 The Aim of the Study

The study aims to achieve the following goals; to find and understand the determinants that changes and forms the people of India’s toughts and actions regarding family planning, to see if education plays a role in this equation, to calculate the impact of family planning on the welfare of the Indian people especially children, to identify and explain the patterns between these relationships among the variables that are being examined and finally to conduct a thorough analysis about everything mentioned before. By achieving these goals the study desires to have a better understanding than before.

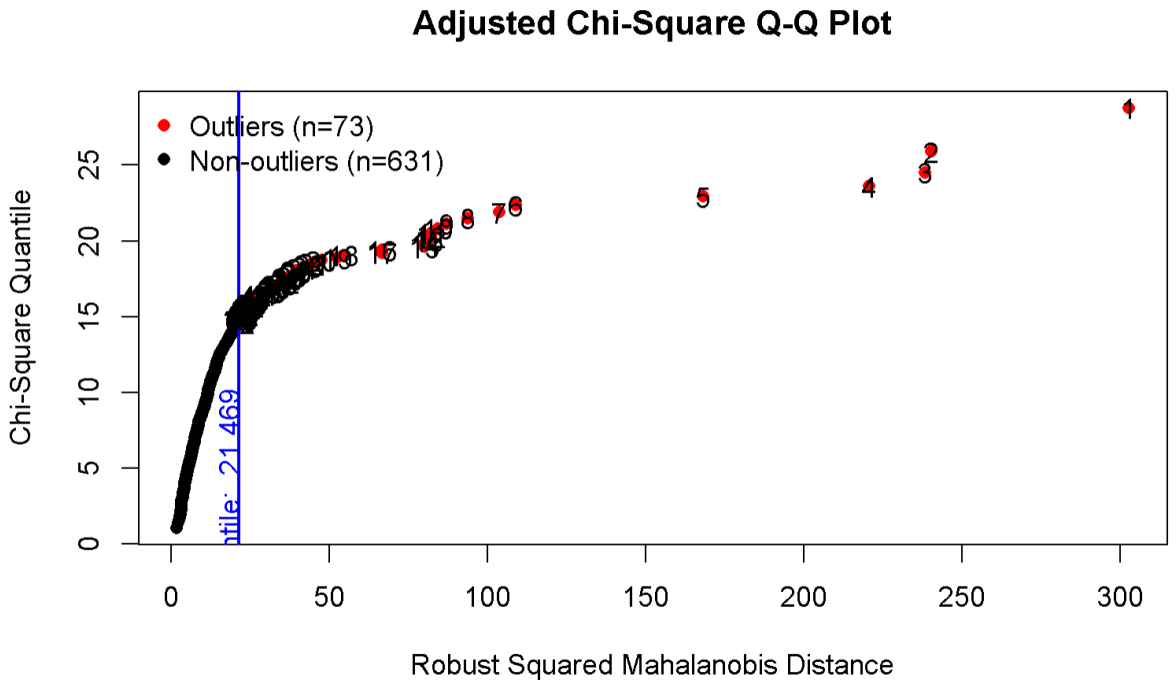
2. Methodology /Analysis

For the purpose of visualizing the data boxplots, histograms and qqplots were created in R. Also Mahalanobis Distance was used to determine how many outliers were in the data and decided not to omit them based on the amount being more than 10%. Carried out Hotelling's T-square test for comparing two mean vectors and computed confidence intervals for displaying the probability that a parameter will fall between a pair of values around the mean and visualized these intervals. The comparison of several multivariate means is used to determine whether the means of response variables varies with respect to the categorical variables. After that normality is checked Box's M test is used and if it fails to reject the null hypothesis from the MANOVA then Bartlett's or Levene's test is used for looking at the Pr(>F) value which shows the p value. Principal Component Analysis(PCA) is an analysis that is mainly used for simplfying, interpreting and describing the data and creating new variables called principal components that are actually mixtures of the original variables. Principal component Regression combines PCA and multiple linear regression so that a response variable can be combined with the principal components. Factor analysis reduces the dimensions of the data and it aims to divide the variables by their correlations and explain the relations between the variables. The discriminant analysis is mainly used for classification the main goal of this analysis is to seperate or characterize different groups in the data.Then finally Clustering is used to group similar data points together depending on their attributes. With doing this it can achieve pattern recognition, segmentation and data understanding.

3. Results and Findings

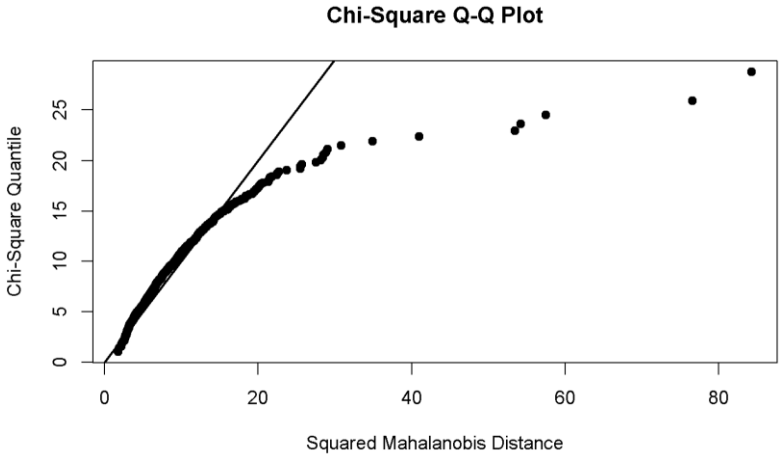
3.1 Exploratory Data Analysis

First the data is checked for outliers with using Mahalanobis distance .



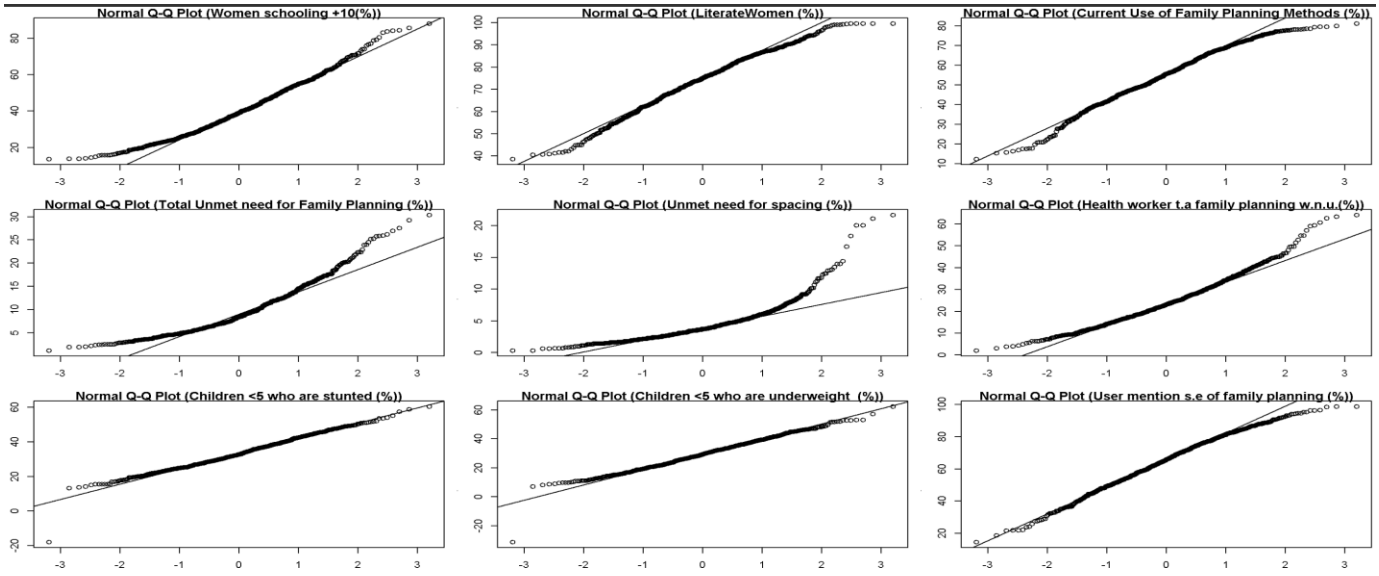
As it can be seen there are 73 outliers which is more that 10% so it is decided not to omit the outliers.

The QQ-Plot below can be seen as left skewed.

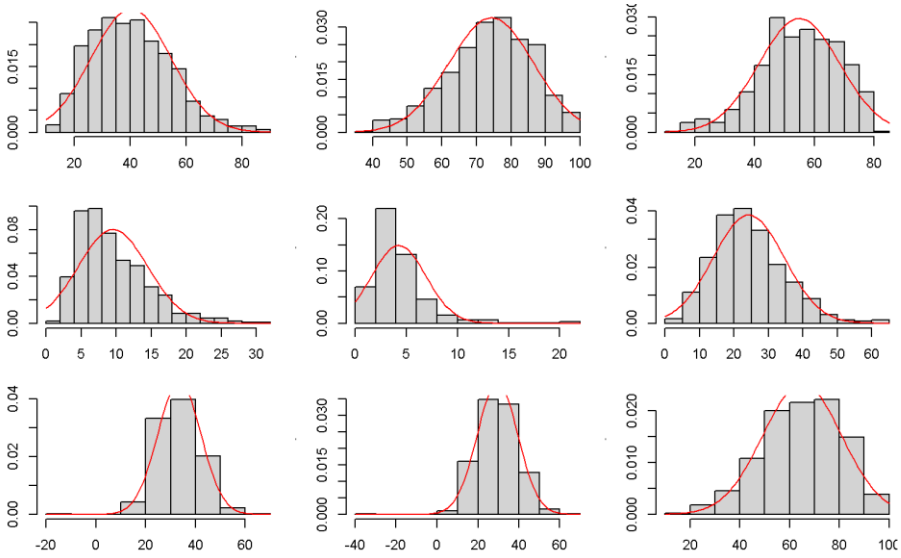


	Test	Variable	Statistic	p value	Normality
	<S3: AsIs>	<S3: AsIs>	<S3: AsIs>	<S3: AsIs>	<S3: AsIs>
1	Shapiro-Wilk	Women schooling +10(%)	0.9769	<0.001	NO
2	Shapiro-Wilk	LiterateWomen (%)	0.9864	<0.001	NO
3	Shapiro-Wilk	Current Use of Family Planning Methods (%)	0.9807	<0.001	NO
4	Shapiro-Wilk	Total Unmet need for Family Planning (%)	0.9238	<0.001	NO
5	Shapiro-Wilk	Unmet need for spacing (%)	0.8043	<0.001	NO
6	Shapiro-Wilk	Health worker t.a family planning w.n.u.(%)	0.9715	<0.001	NO
7	Shapiro-Wilk	Children <5 who are stunted (%)	0.9853	<0.001	NO
8	Shapiro-Wilk	Children <5 who are underweight (%)	0.9850	<0.001	NO
9	Shapiro-Wilk	User mention s.e of family planning (%)	0.9885	<0.001	NO

According to the Shapiro Wilk test no variable can ensure normality. So now Normal QQ-Plots and Histograms should be examined to be sure.



When looked at the Normal QQ-Plots it can be said that Children under 5 years old who are stunted and Children under 5 years old who are underweight can ensure normality.



And looking at the histograms of those 2 variables it can be said that they can ensure normality

3.2 Inference About Mean

The aim is to model the responses, Total Unmet need for Family Planning and Unmet need for spacing.

Means of the responses:

```
Total Unmet need for Family Planning (%)
                                9.526420
             Unmet need for spacing (%)
                                4.260653
```

μT0=[10,5]

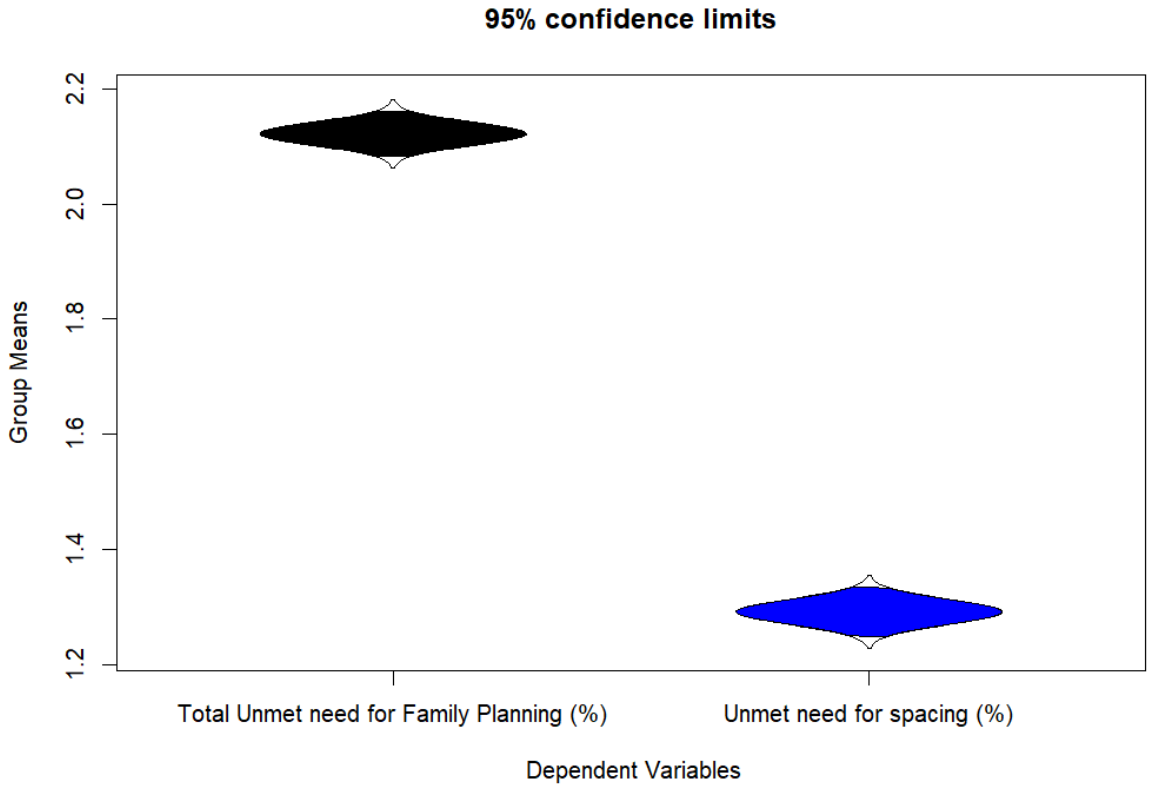
n>20 royston is used

```
> test$multivariateNormality
     Test        H      p value MVN
1 Royston 164.7273 7.047211e-37  NO
```

The response matrix does not follow normal distribution so it can be considered as the log of the data.

```
> test2$univariateNormality
             Test                                  Variable Statistic   p value Normality
1 Anderson-Darling Total Unmet need for Family Planning (%)    0.8206    0.0339    NO
2 Anderson-Darling        Unmet need for spacing (%)           1.5692    0.0005    NO
>
```

The response matrix does not follow normal distribution but the assumption of normality is satisfied ın this case, note that mu0 will not be tested, instead log(mu0) will be tested.

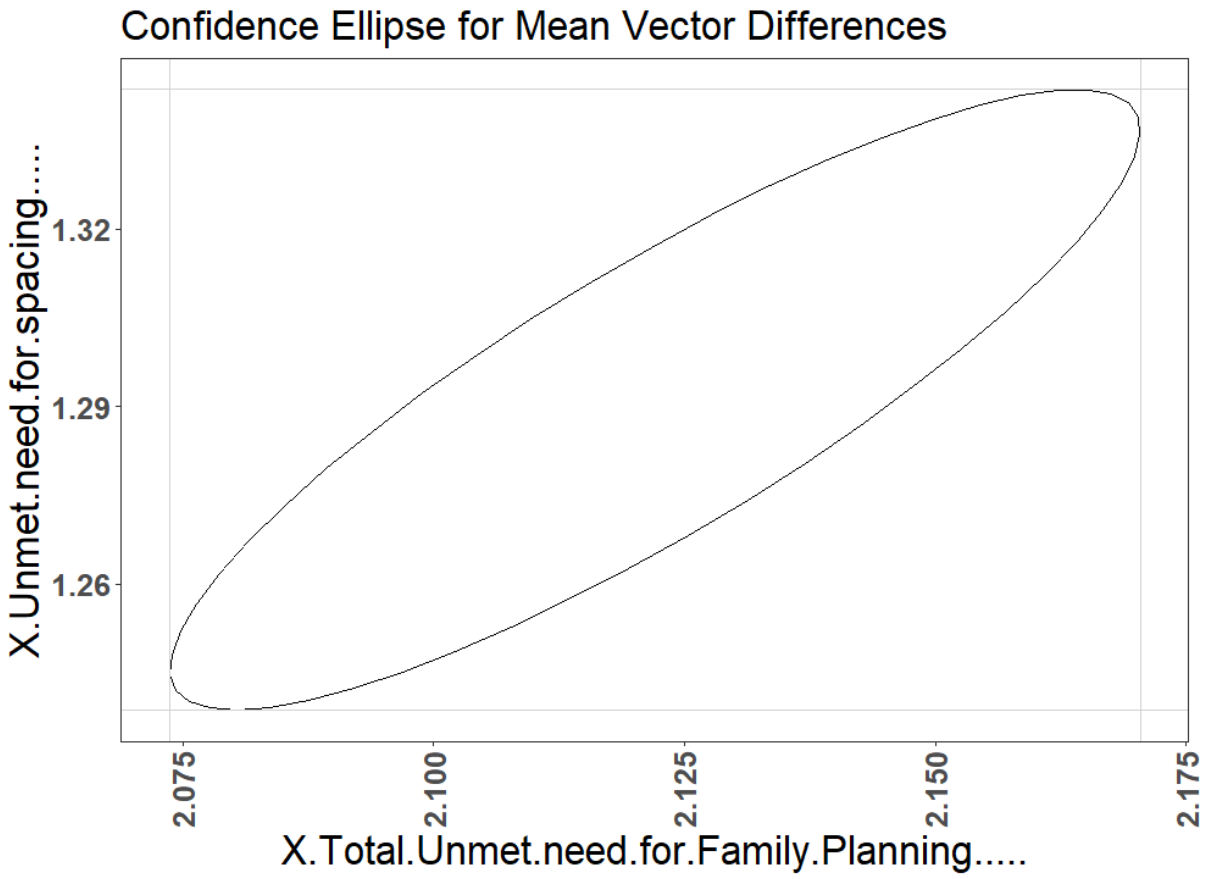


There is an outstanding difference between the means of the groups.

```
> HotellingsT2Test(log_y, mu = log(mu0))

        Hotelling's one sample T2-test

data:  log_y
T.2 = 139.01, df1 = 2, df2 = 702,
p-value < 2.2e-16
alternative hypothesis: true location is not equal to c(2.30258509299405,1.6094379124341)
```

Since p < alpha H0 is rejected. Therefore, there is not enough evidence to conclude that the log of the mean vector equals to log(10,5).

```
> MVcis(log_y)
                                                  [,1]     [,2]
X.Total.Unmet.need.for.Family.Planning..... 2.073709 2.170430
X.Unmet.need.for.spacing.....               1.238694 1.343663
>
```



While mu0 values are in the simultaneous confidence intervals for each variable since it is not in the confidence region (i.e., the point is not in the ellipse in the plot), the null hypothesis is rejected.

3.3 Comparısons of Several Multıvarıate Means

Here whether the response variables (Total Unmet need for Family Planning ,Unmet need for spacing) varie with respect to State/UT will be tested.

Shapiro Wilk test is used to check normality.

```
# A tibble: 64 × 4
   `State/UT`                variable statistic       p
   <fct>                     <chr>        <dbl>   <dbl>
 1 Andaman & Nicobar Islands log_s        1.00  0.985
 2 Andaman & Nicobar Islands log_t        0.996 0.878
 3 Andhra Pradesh            log_s        0.943 0.500
 4 Andhra Pradesh            log_t        0.930 0.341
 5 Arunachal Pradesh         log_s        0.956 0.471
 6 Arunachal Pradesh         log_t        0.911 0.0663
 7 Assam                     log_s        0.972 0.531
 8 Assam                     log_t        0.988 0.970
 9 Bihar                     log_s        0.903 0.00301
10 Bihar                     log_t        0.916 0.00743
```

It is assumed that the normality is satisfied to find homogeneity of the variance Box's M-test is used.

```
        Box's M-test for Homogeneity of Covariance Matrices

data:  cbind(filtered_mysubset$log_t, filtered_mysubset$log_s)
Chi-Sq (approx.) = 247.46, df = 93, p-value =
5.756e-16
```

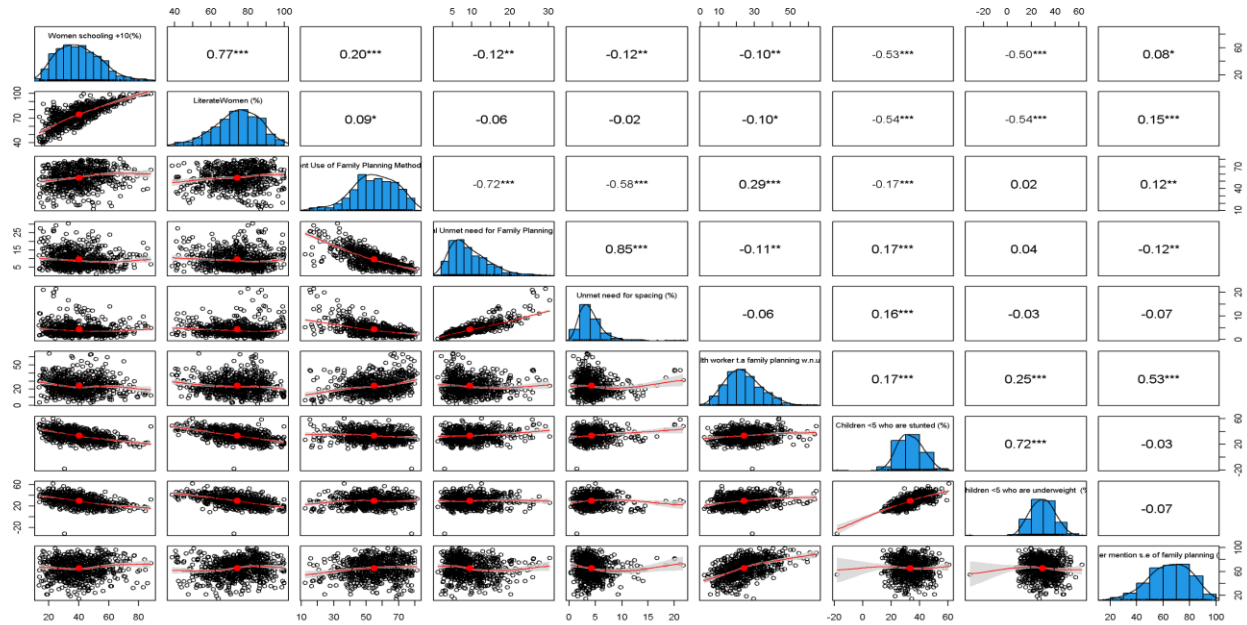
It is assumed that null hypothesis can not be rejected and conclude that variance-covariance matrices are equal for each combination of the dependent variable formed by each group in the independent variable.

Manova

```
            Df  Pillai approx F num Df den Df    Pr(>F)
`State/UT`  31 0.66126   10.612     62   1332 < 2.2e-16 ***
Residuals  666
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

With a 95% confidence level, it is concluded that at least one state exhibits a statistically significant difference from the others, as indicated by p < alpha.

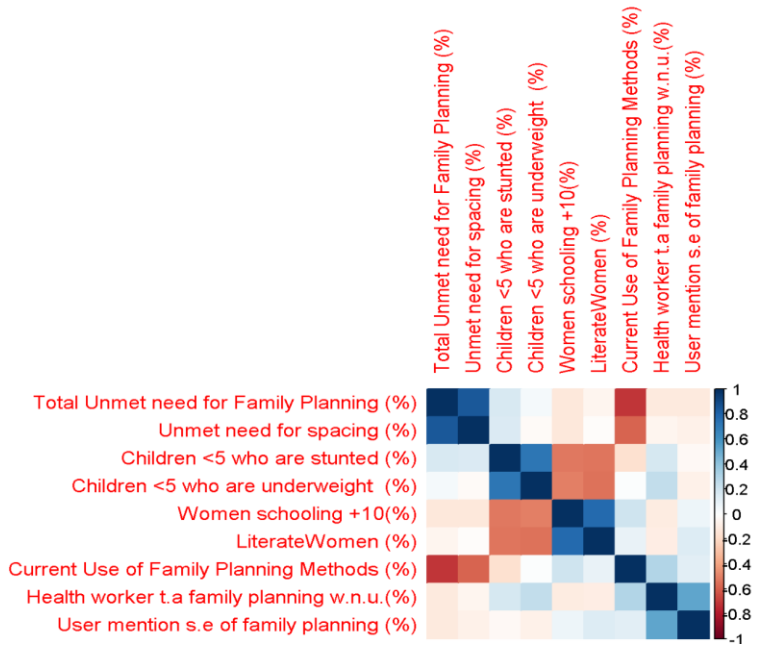
3.4 Principal Component Analysis



When looked at the table above it can be seen that amongst the variables that were picked for looking at their correlations amongst themselves some of the correlations catch the eye with strong and and moderate levels of correlations for the readers benefit here is a list of the correlated variables:

- Women with 10 or more years of education - Literate Women (0.77)
- Women with 10 or more years of education – Children under 5 years old who are stunted (-0.53)
- Women with 10 or more years of education - Children under 5 years old who are underweight (-0.50)
- Literate Women - Children under 5 years old who are stunted (-0.54)
- Literate Women - Children under 5 years old who are underweight (-0.54)
- Current Usage of Family Planning Methods - Total Unmet Need for Family Planning (-0.72)
- Current Usage of Family Planning Methods – Unmet Need for Spacing (-0.58)
- Total Unmet Need for Family Planning - Unmet Need for Spacing (0.85)
- Health Worker Ever Talked to Non-users about Family Planning – Current User Mentioned about the Side Effects of Family Planning (0.53)
- Children under 5 years old who are stunted - Children under 5 years old who are underweight (0.72)

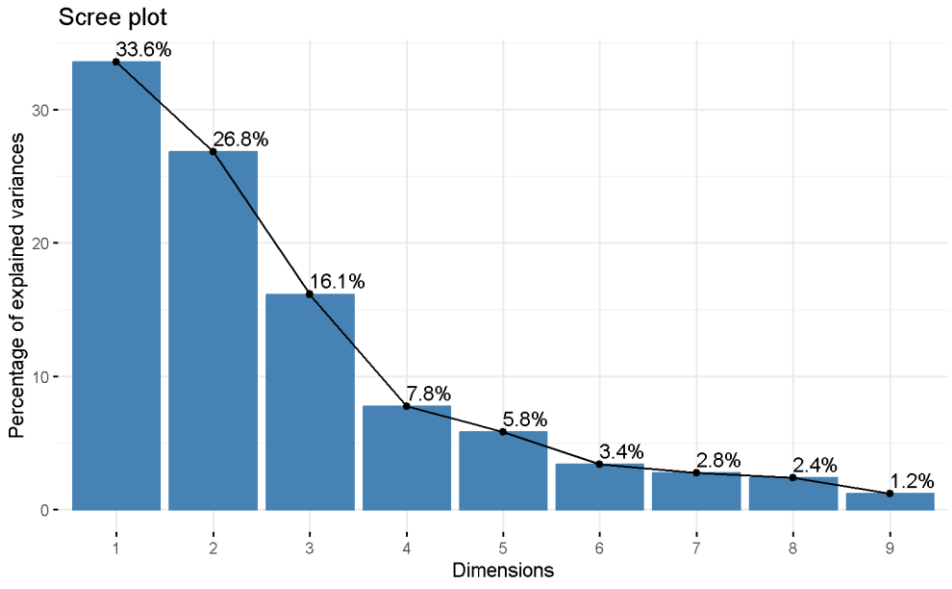
Below is a correlation map for as an alternative visualization for the correlations:



After the variables were scaled for the PCA the summary of PCA was calculated.

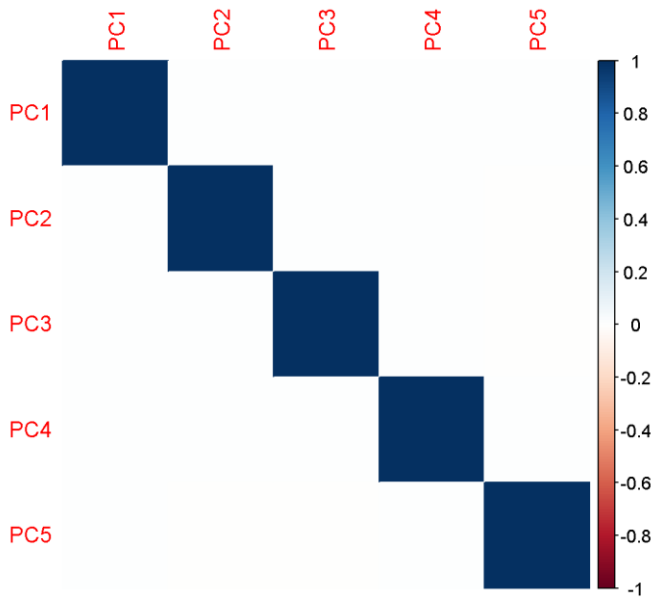
```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     1.7383 1.5536 1.2053 0.83656 0.72383 0.55555 0.50035 0.46721 0.33299
Proportion of Variance 0.3357 0.2682 0.1614 0.07776 0.05821 0.03429 0.02782 0.02425 0.01232
Cumulative Proportion  0.3357 0.6039 0.7653 0.84310 0.90132 0.93561 0.96343 0.98768 1.00000
```

For the purpose of visualization a scree plot was created.



As it can be seen on the plot above first 5 components can explain nearly 90% of the variability in the data. So the first 5 is picked to continue the research.

For the purpose of checking the orthogonality a correlation plot of the PC's were created.



As it can be seen the components are linearly independent so there is no collinearity.

The correlation between the PC's and the numerical variables can be seen below.

	PC1	PC2	PC3	PC4	PC5
Women schooling +10(%)	-0.77833110	-0.2952574	-0.11391267	0.43536608	0.05852881
LiterateWomen (%)	-0.74715796	-0.3845246	-0.20908377	0.34800212	0.13383142
Current Use of Family Planning Methods (%)	-0.49788112	0.6939213	0.05946638	0.15369829	-0.34977286
Total Unmet need for Family Planning (%)	0.52728978	-0.7457821	-0.26023221	0.08181528	-0.12005311
Unmet need for spacing (%)	0.47293006	-0.7094759	-0.32393377	0.07138377	-0.25509609
Health worker t.a family planning w.n.u.(%)	0.06362427	0.4933147	-0.74340325	0.04964010	-0.32798254
Children <5 who are stunted (%)	0.77536040	0.3098296	-0.05109300	0.34507233	0.24447087
Children <5 who are underweight (%)	0.68508196	0.4841905	0.01914394	0.41159375	0.07584513
User mention s.e of family planning (%)	-0.18274750	0.2526387	-0.81501206	-0.25065136	0.35730951

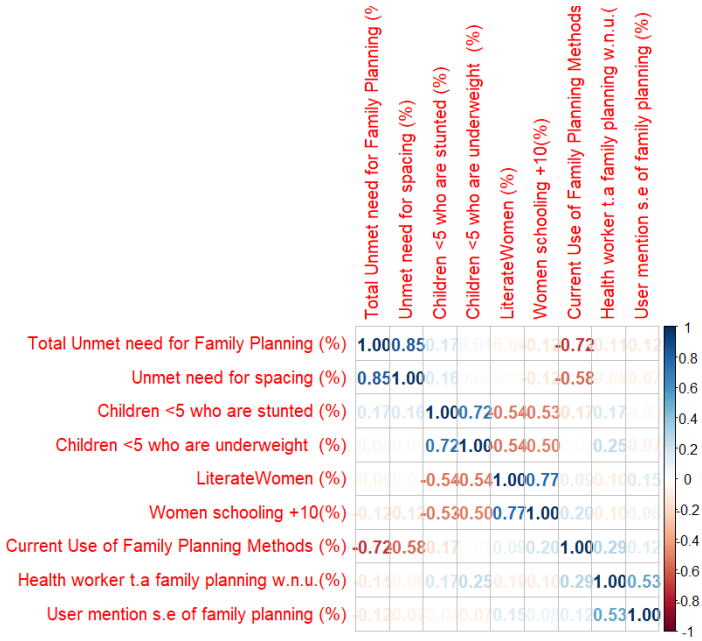
Judging by the table it can be said that Component 1 is strongly negatively related with Women schooling +10 and Literate Women, and strongly positively correlated with Children <5 who are stunted and positively correlated with children <5 who are underweight.

Component 2 is strongly negatively correlated with Total unmet need for family planning and unmet need for spacing and positively correlated with current use of family planning methods.

Component 3 is strongly negatively correlated with Health Worker Ever Talked to Non-users about Family Planning and User mention side effects of family planning.

3.5 Factor Analysis and Factor Rotation

Numeric variables of our data set is considered for here.



It can be observed that there are some correlated variables.

Kaiser-Meyer-Olkin (KMO) and Bartlett Tests:

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cm)
Overall MSA =  0.64
MSA for each item =
                          LiterateWomen (%)                       Women schooling +10(%)  Current Use of Family Planning Methods (%)
                                       0.70                                        0.69                                        0.67
   Total Unmet need for Family Planning (%)                 Unmet need for spacing (%) Health worker t.a family planning w.n.u.(%)
                                       0.58                                        0.61                                        0.51
            Children <5 who are stunted (%)       Children <5 who are underweight  (%)      User mention s.e of family planning (%)
                                       0.72                                        0.67                                        0.47
```

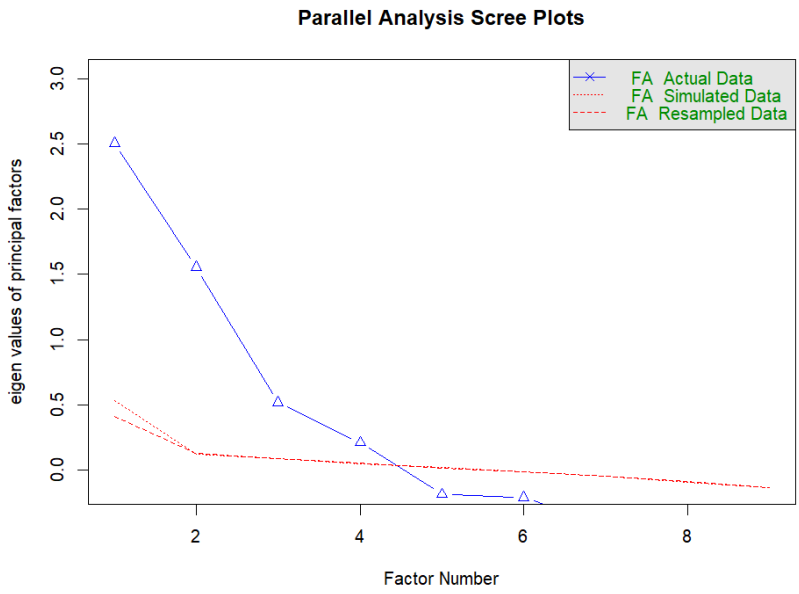
MSA value is greater than 0.5

```
$chisq
[1] 3443.101

$p.value
[1] 0

$df
[1] 36
```

P- value is less than 0.05 significance level

So variables are suitable for a factor analysis.

Scree Plot and Factanal Test Result:

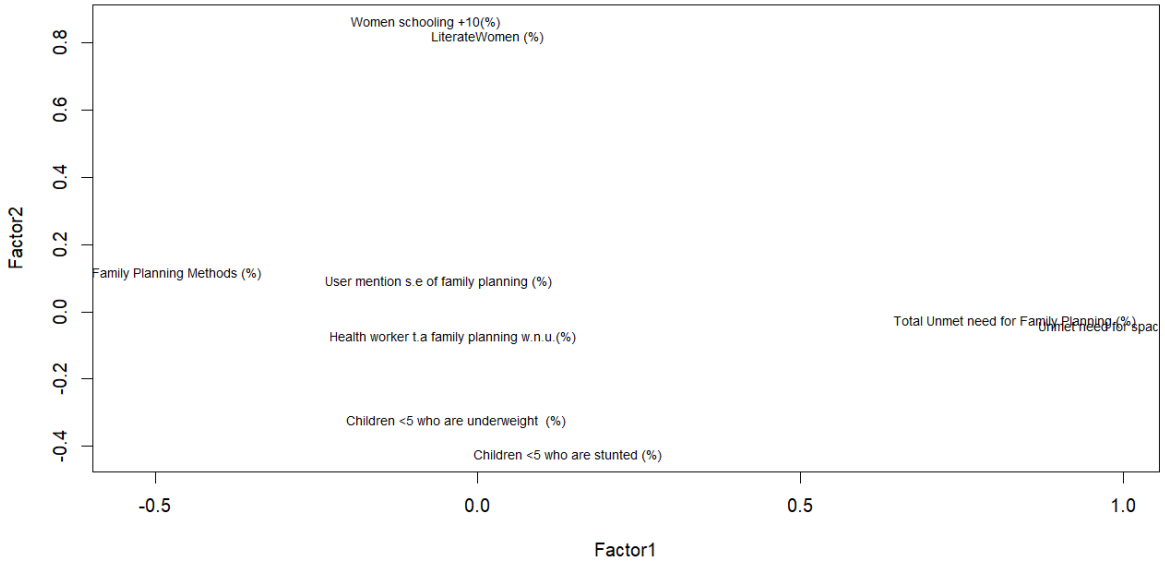



```
Parallel analysis suggests that the number of factors =  4  and the number of components
=  NA
> factanal(myfactordata, factors = 4)$PVAL
   objective
1.168781e-16
> factanal(myfactordata, factors = 5)$PVAL
   objective
1.890441e-06
```

It is not enough when the number of factors is 5, but it has the largest p value among all possibilities, so the number of factors is chosen as 5.

Loading Part of Factanal Test and Plot of First Two Factors:

```
Call:
factanal(x = myfactordata, factors = 5)

Uniquenesses:
                        LiterateWomen (%)                                 Women schooling +10(%)  Current Use of Family Planning Methods (%)
                                    0.241                                                  0.185                                       0.005
   Total Unmet need for Family Planning (%)                        Unmet need for spacing (%) Health worker t.a family planning w.n.u.(%)
                                    0.196                                                  0.005                                       0.565
            Children <5 who are stunted (%)          Children <5 who are underweight  (%)      User mention s.e of family planning (%)
                                    0.400                                                  0.005                                       0.005

Loadings:
                                        Factor1 Factor2 Factor3 Factor4 Factor5
LiterateWomen (%)                                0.818  -0.295
Women schooling +10(%)                           0.862  -0.243
Current Use of Family Planning Methods (%)  -0.536  0.114           0.149   0.820
Total Unmet need for Family Planning (%)     0.833                          -0.309
Unmet need for spacing (%)                   0.995
Health worker t.a family planning w.n.u.(%)                  0.194   0.578   0.238
Children <5 who are stunted (%)              0.139  -0.425   0.626
Children <5 who are underweight  (%)                -0.325   0.940
User mention s.e of family planning (%)                              0.983

               Factor1 Factor2 Factor3 Factor4 Factor5
SS loadings      2.003   1.727   1.476   1.341   0.846
Proportion Var   0.223   0.192   0.164   0.149   0.094
Cumulative Var   0.223   0.414   0.578   0.727   0.821

Test of the hypothesis that 5 factors are sufficient.
The chi square statistic is 22.7 on 1 degree of freedom.
The p-value is 1.89e-06
```



Factor 1 is dominated by the Total Unmet need for Family Planning and Unmet need for spacing

Factor 2 is dominated by the Women schooling +10(%) and LiterateWomen (%)

5 factors explains nearly 82% of the variability in the data set.

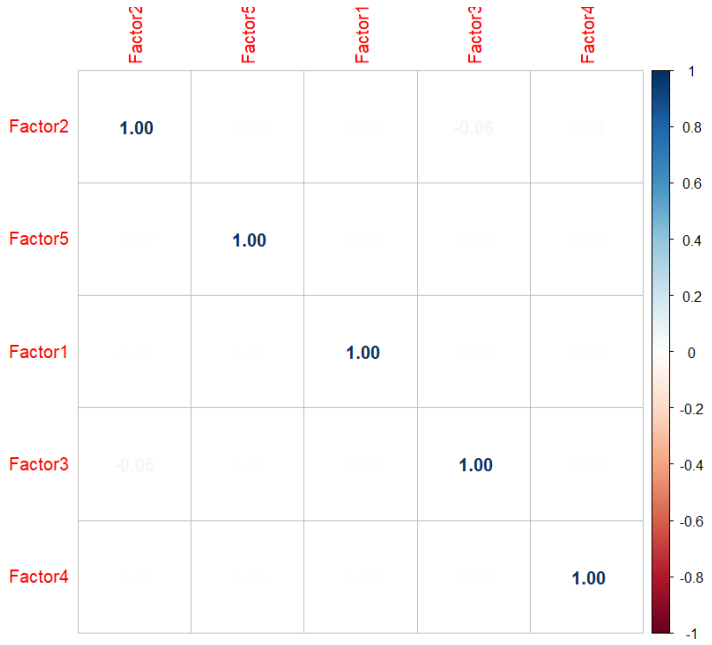
Cronbach's Alpha Test

```
Reliability analysis
 raw_alpha std.alpha G6(smc) average_r S/N   ase mean  sd median_r
      0.64      0.88    0.87      0.72 7.5 0.013   13 6.4     0.72
```

While a Cronbach's alpha exceeding 0.7 is generally considered satisfactory for a factor, the highest alpha value that was observed after evaluating all the factors and their dominating variables was 0.6.

Estiamated Factor Scores For 5 Factors and Their Correlation Plot

```
        Factor1    Factor2    Factor3    Factor4    Factor5
[1,] -0.3189249  0.9909258 -0.1358053 -1.110191 0.06901788
[2,] -1.0169715  0.7712556  1.5313542  1.231457 0.66000926
[3,]  1.7019653  1.0696794 -0.9450179  1.426938 0.28400801
[4,] -0.2091638 -0.4765037 -1.0580617 -1.161133 1.68017045
[5,]  0.2062547 -0.7439483  0.0253829 -1.579691 1.99066195
[6,] -0.6624517  0.1303761  0.4867826 -1.823779 1.04271411
```



At the end of the factor analysis estimated factor scores are calculated and they are almost uncorrelated which guarantees that there is no multicollinearity problem in linear regression.

3.6 Discrimination and Classification

```
         State/UT   LiterateWomen (%) Women schooling +10(%)
Uttar Pradesh : 75   Min.   :38.60     Min.   :13.60
Madhya Pradesh: 51   1st Qu.:66.80     1st Qu.:29.23
Bihar         : 38   Median :75.10     Median :39.15
Maharastra    : 36   Mean   :74.28     Mean   :40.28
Assam         : 33   3rd Qu.:83.70     3rd Qu.:49.83
Gujarat       : 33   Max.   :99.70     Max.   :88.20
(Other)       :438
Current Use of Family Planning Methods (%) Total Unmet need for Family Planning (%)
Min.   :12.30                              Min.   : 1.200
1st Qu.:46.67                              1st Qu.: 5.800
Median :55.60                              Median : 8.450
Mean   :54.90                              Mean   : 9.526
3rd Qu.:65.62                              3rd Qu.:12.300
Max.   :81.20                              Max.   :30.400
```

```
Children <5 who are stunted (%) Children <5 who are underweight  (%)
Min.   :-18.00                  Min.   :-31.30
1st Qu.: 27.30                  1st Qu.: 22.07
Median : 32.85                  Median : 29.35
Mean   : 33.46                  Mean   : 29.47
3rd Qu.: 39.20                  3rd Qu.: 36.30
Max.   : 60.60                  Max.   : 62.40

User mention s.e of family planning (%) Cluster
Min.   :14.60                           1:147
1st Qu.:54.38                           2:295
Median :66.05                           3:262
Mean   :65.04
3rd Qu.:76.92
Max.   :98.90
```

State/Ut is the only factor variable. It has 36 levels 75 people live in uttar pradesh, 51 people live in madhya pradesh 38 people live in bihar, 36 people live maharastra, 33 people live in assam, 33 people live in gujarat and 438 people live in other states.

```
Call:
lda(`State/UT` ~ ., data = train)

Prior probabilities of groups:
             Andaman & Nicobar Islands                          Andhra Pradesh
                           0.005328597                           0.015985790
                     Arunachal Pradesh                                 Assam
                           0.028419183                           0.046181172
                                Bihar                             Chandigarh
                           0.060390764                           0.001776199
                          Chhattisgarh Dadra and Nagar Haveli & Daman and Diu
                           0.035523979                           0.005328597
                                  Goa                                 Gujarat
                           0.003552398                           0.047957371
                              Haryana                        Himachal Pradesh
                           0.030195382                           0.021314387
                        Jammu & Kashmir                             Jharkhand
                           0.023090586                           0.031971581
                            Karnataka                                Kerala
                           0.044404973                           0.021314387
                               Ladakh                            Lakshadweep
                           0.003552398                           0.001776199
                         Madhya Pradesh                           Maharastra
                           0.087033748                           0.049733570
                              Manipur                             Meghalaya
                           0.012433393                           0.012433393
                              Mizoram                              Nagaland
                           0.010657194                           0.015985790
                           NCT of Delhi                                Odisha
                           0.019538188                           0.039076377
                             Puducherry                                Punjab
                           0.007104796                           0.030195382
                              Rajasthan                                Sikkim
                           0.046181172                           0.003552398
                             Tamil Nadu                             Telangana
                           0.047957371                           0.035523979
                                 Tripura                         Uttar Pradesh
                           0.014209591                           0.097690941
                             Uttarakhand                           West Bengal
                           0.014209591                           0.028419183
```

There are 36 levels in the data. The LDA output indicates that for example Uttar Prades = 0.1 that means 10% of the training observation corresponds to patients that live in uttar Prades, 90% people live in other state.

```
> sum(diag(table_train))/sum(table_train)
[1] 0.6891652
```

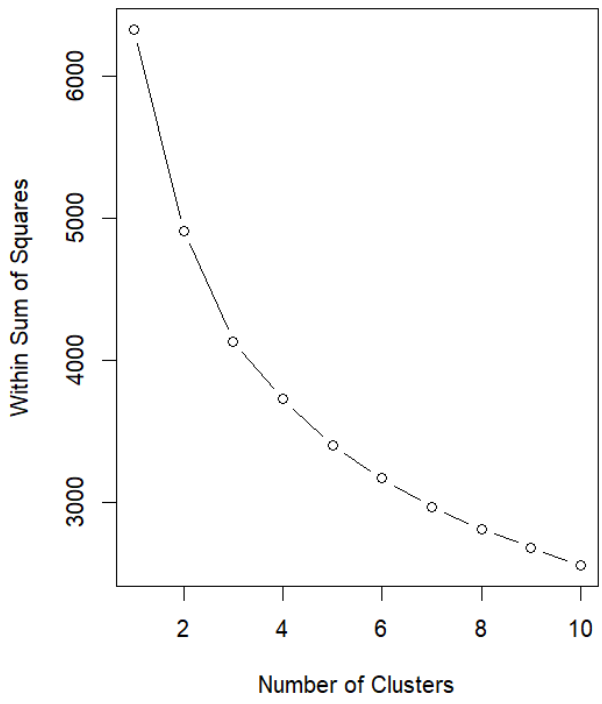
The model correctly classifies the states where people live. 0.69 is the probability for the training data so the classification error rate(misclassification) for training data is 1-0.69=0.31

3.7 Clustering

Firstly K-means Clustering is used, missing values are checked and we the numeric data is standardized.

After that the optimal number of clusters is determined using the elbow method.

Elbow plot



Based on the plot, visually identify the "elbow" where the rate of decrease slows down. The only “elbow” in the plot occurs for three groups, and so now the three-group solution will be examined.

```
> head(data_all$Cluster)
[1] 1 1 3 1 2 1
Levels: 1 2 3
```

There are 3 cluster numbers.

```
> summary(data_all$Cluster)
  1   2   3
295 262 147
```

The cluster number of 295 observations of the data is 1, the cluster number of 262 observations is 2 and the cluster number of 147 observations is 3.

4. Discussion and Conclusion

The usage of family planning in India is a crucial practice as it affects multiple areas of life like general health, nutrition, welfare and education but it is also affected by some of these areas and many others for example governments policies about family planning. So when there is something as crucial as this it should be investigated in a thorough manner because it may be the related to unpleasents truths of life like starving or undernourished children, people not being able to get the education they deserve and women not being able to take the break they need to take and get pregnant at frequent intervals (which is highly risky for them). During this research the variables that were beforehand tought to affect family planning and get affected by it, were put through multiple statistical analyses and examined. As a result of these analyses some variables were found to be strongly correlated with each other and some were surprisingly not. We can see from the results that the unmet need of family planning is highly correlated with unmet need for spacing for women and this tells us that if family planning were used and the states were able to meet the unmet needs women would have gotten the spacing they needed between pregnanices. Even though the people would think that education would increase the usage of family planning it can be seen from the study that it does not but it has a moderate level of correlation with children being stunted and underweight but nothing official can be said about it due to lack of evidince. If there were

more time for this research of course the current results wouldn't have changed but a more delicate and more comprehensive search would be possible and things that were not looked at because they were not looking too promising would have been studied and who knows what kind of results would have uncovered.

References

ODTUCLASS 2023-2024 FALL: Multivariate Analysis (n.d.). https://odtuclass2023f.metu.edu.tr/pluginfile.php/521231/mod_resource/content/5/RC8.html

Brownlee, J. (2019, August 8). How to Transform Data to Better Fit The Normal Distribution. MachineLearningMastery.com. https://machinelearningmastery.com/how-to-transform-data-to-fit-the-normal-distribution/

Appendices

https://www.kaggle.com/datasets/bhanupratapbiswas/national-family-health-survey-nfhs-2019-21