Anemia in Indian Women 15-49: A Multivariate Analysis

Baran Kılınç
Medine Yazıcı
Fitnat Koç

OUTLINE

- Introduction
- Inference about Means
- One and Two-Way Anova
- PCA and Regression
- Factor, LDA and Classification
- Cluster Analysis
- Conclusion

Data Description

“India National Family Health Survey (NFHS) 2019-21"

- 707 observation
- 109 variables

Anemia in women aged 15-49

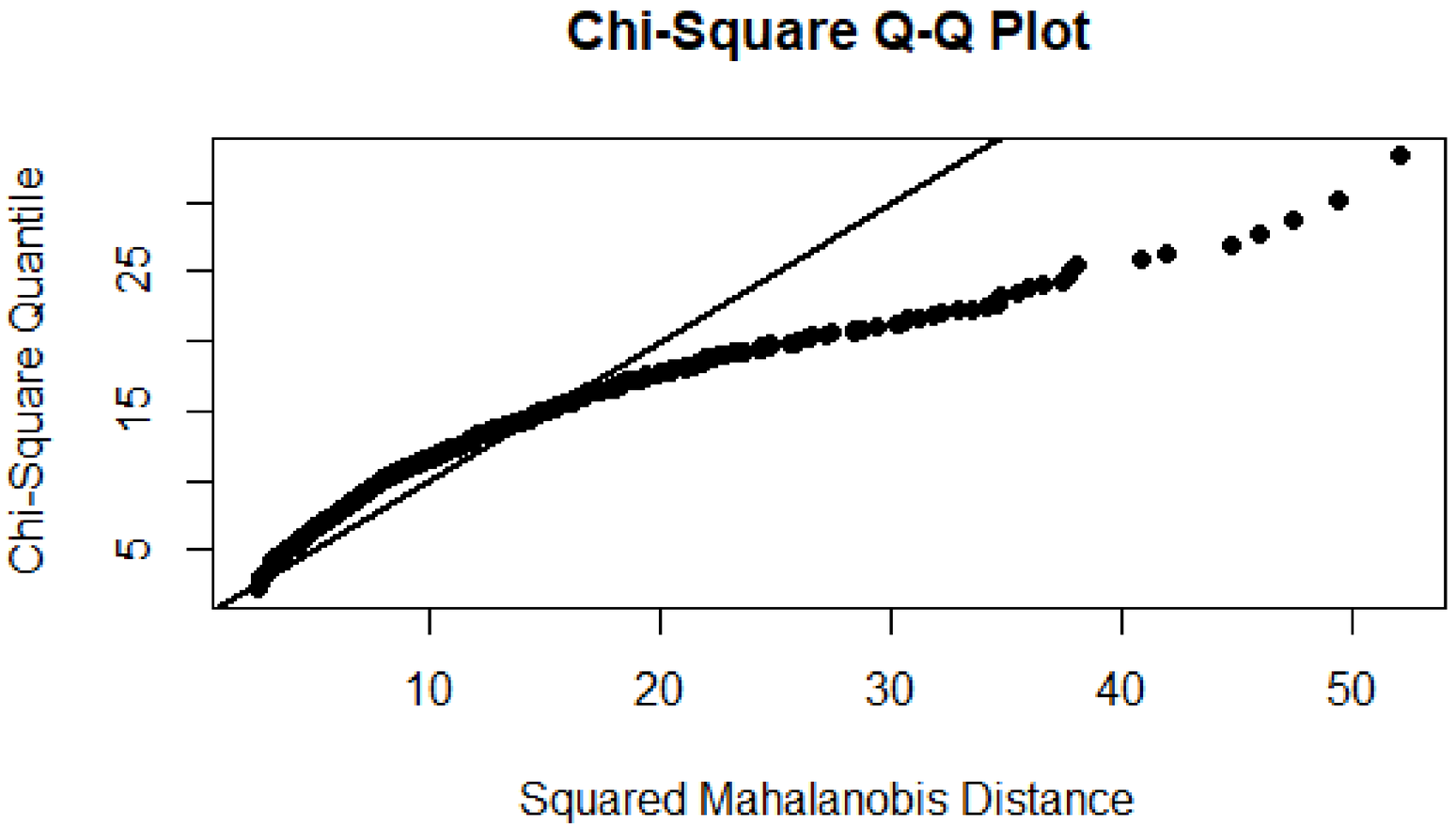
- BMI
- Obese
- WaisToHipRatio
- NonPregnantAnaemic
- PregnantAnameic
- AllAnaemic
- HighBloodSugarLevel
- VeryHighBloodSugarLevel
- MildlyBloodPressure
- ModeratelyBloodPressure
- ElevatedBloodPressure
- Tobacco

MAIN RESEARCH QUESTIONS

- What are the main causes of anemia in women aged 15-49 in India, and how do these differ between rural and urban areas?
- Is anemia more common in pregnant women than in non-pregnant women, and what are the main reasons for this difference?
- What are the best health clusters in the population, and how do they show different risks for anemia and other conditions?
- How do rural or urban residence and obesity affect anemia (in pregnant and non-pregnant women) and blood pressure levels (elevated, mild, moderate) in Indian women aged 15-49?

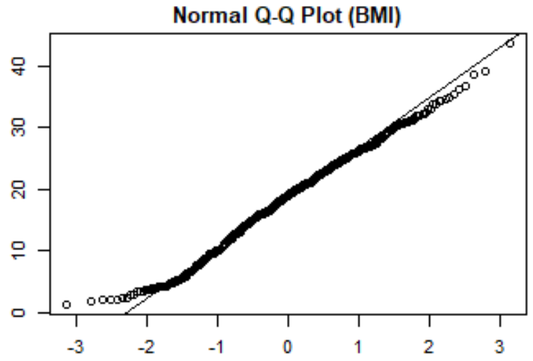
p value is less than alpha

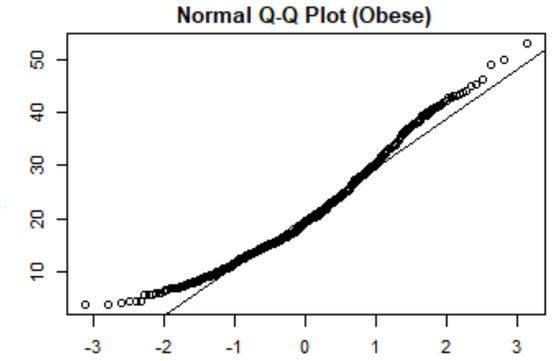
reject H0

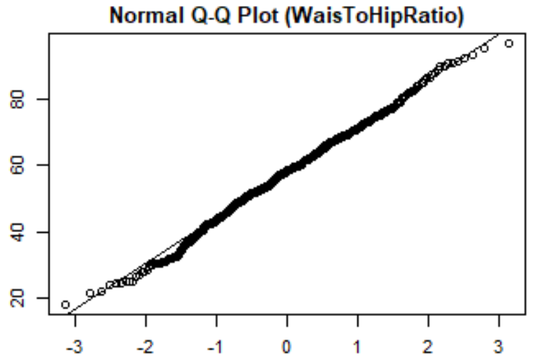


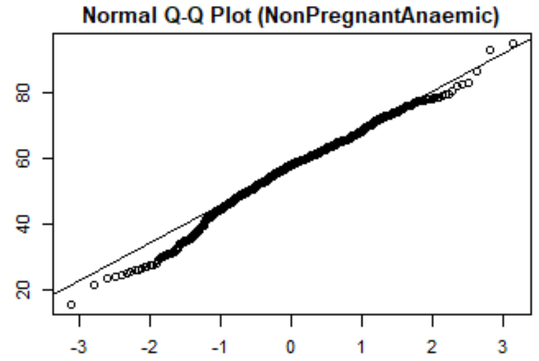
	Test	H	p value	MVN
1	Royston	300.6321	3.379982e-59	NO

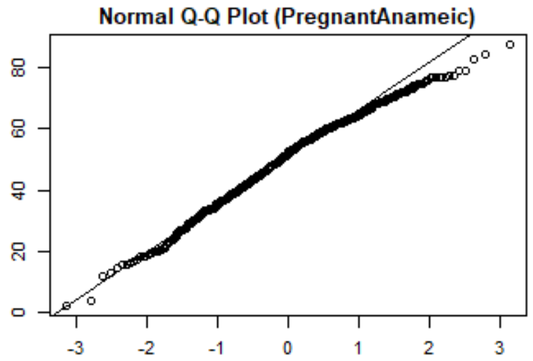
UNIVARIATE Q-Q PLOTS

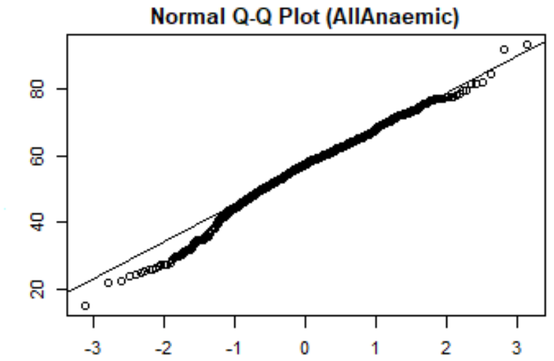


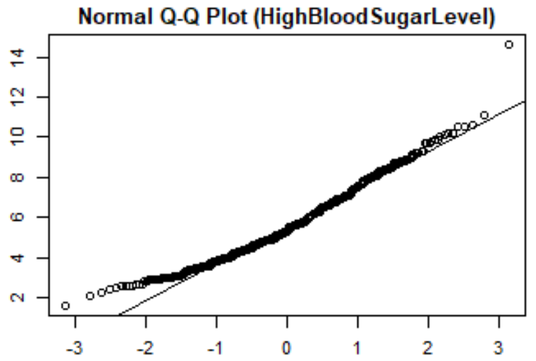


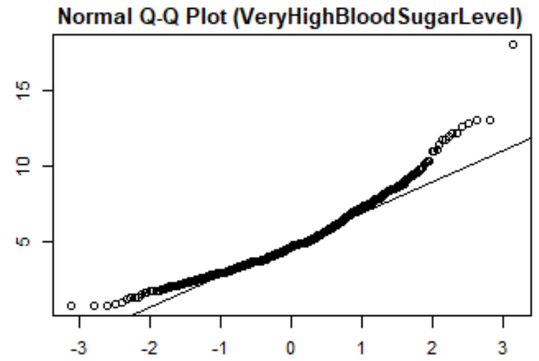


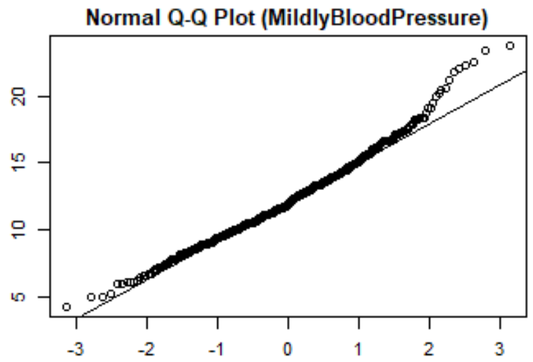


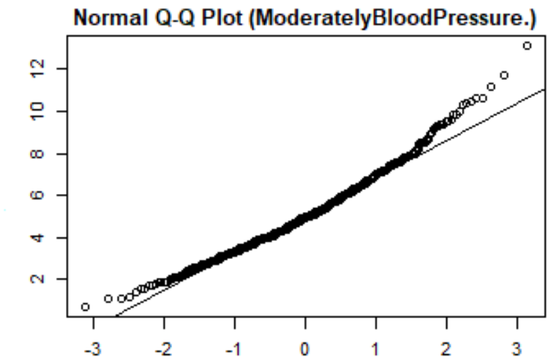


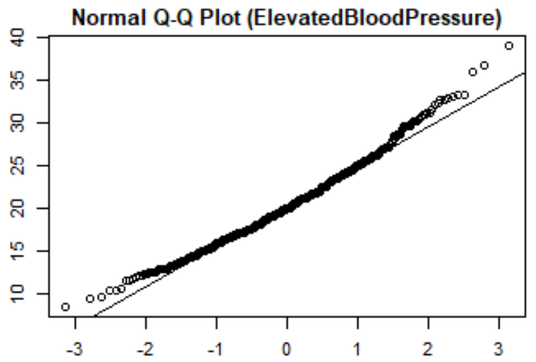


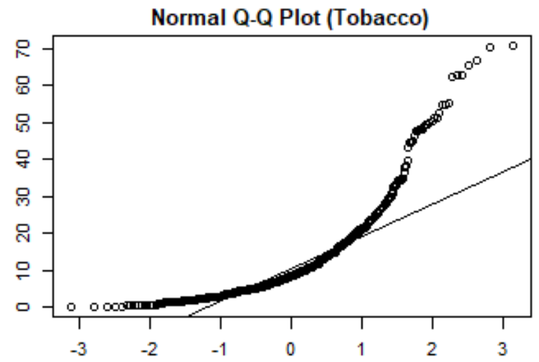








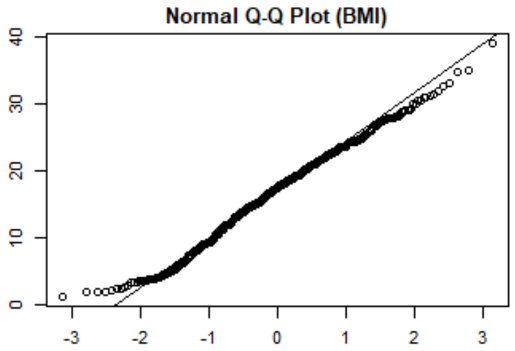


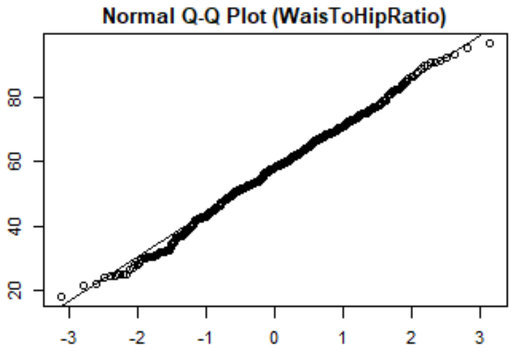


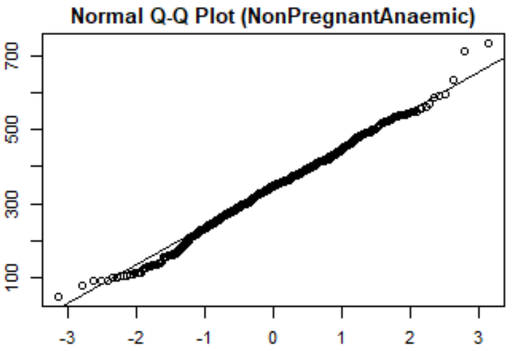
	Test	Variable	Statistic	p value	Normality
1	Anderson-Darling	BMI	1.0401	0.0097	NO
2	Anderson-Darling	Obese	5.7368	<0.001	NO
3	Anderson-Darling	WaisToHipRatio	0.5605	0.1469	YES
4	Anderson-Darling	NonPregnantAnaemic	2.5476	<0.001	NO
5	Anderson-Darling	PregnantAnameic	2.0867	<0.001	NO
6	Anderson-Darling	AllAnaemic	2.7117	<0.001	NO
7	Anderson-Darling	HighBloodSugarLevel	3.8262	<0.001	NO
8	Anderson-Darling	VeryHighBloodSugarLevel	7.9952	<0.001	NO
9	Anderson-Darling	MildlyBloodPressure	1.8098	1e-04	NO
10	Anderson-Darling	ModeratelyBloodPressure.	2.6417	<0.001	NO
11	Anderson-Darling	ElevatedBloodPressure	1.8564	1e-04	NO
12	Anderson-Darling	Tobacco	36.1033	<0.001	NO

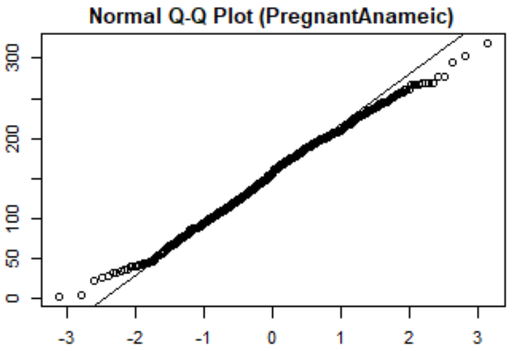
TRANSFORMATION

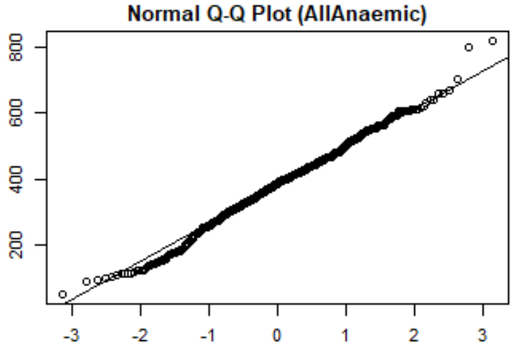
UNIVARIATE Q-Q PLOTS

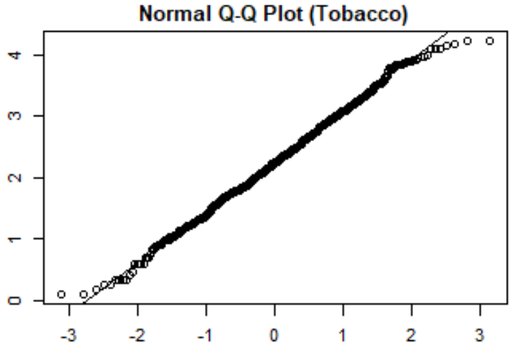


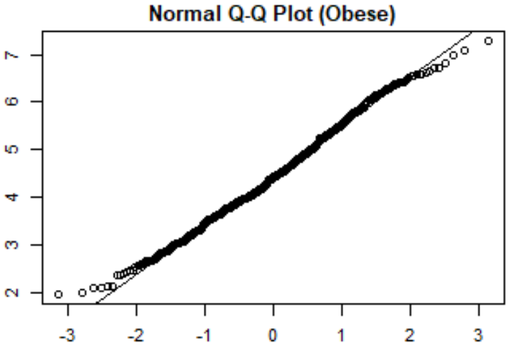


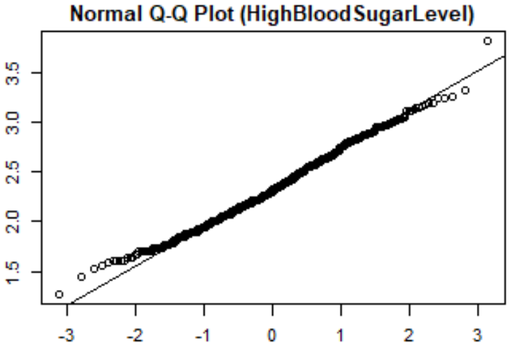


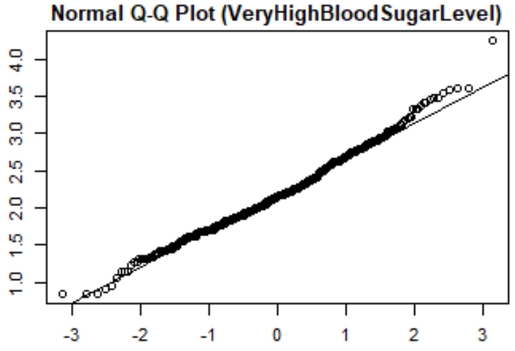


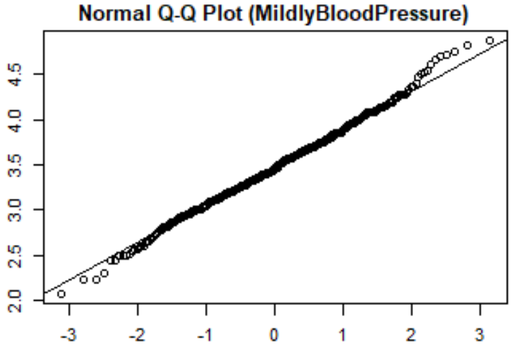


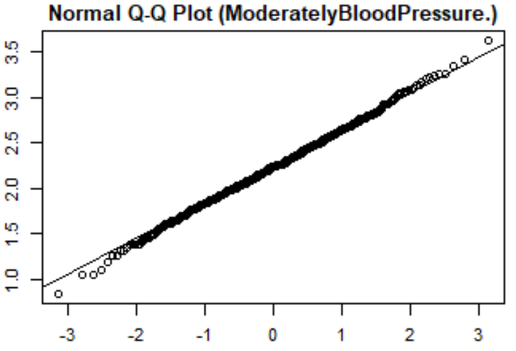


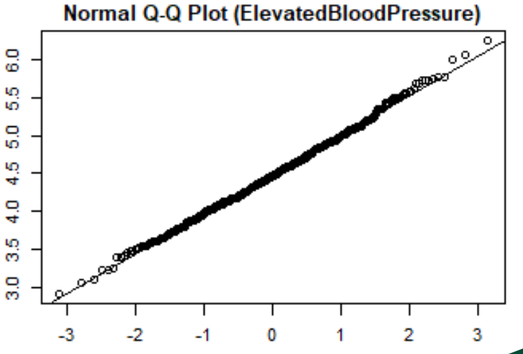




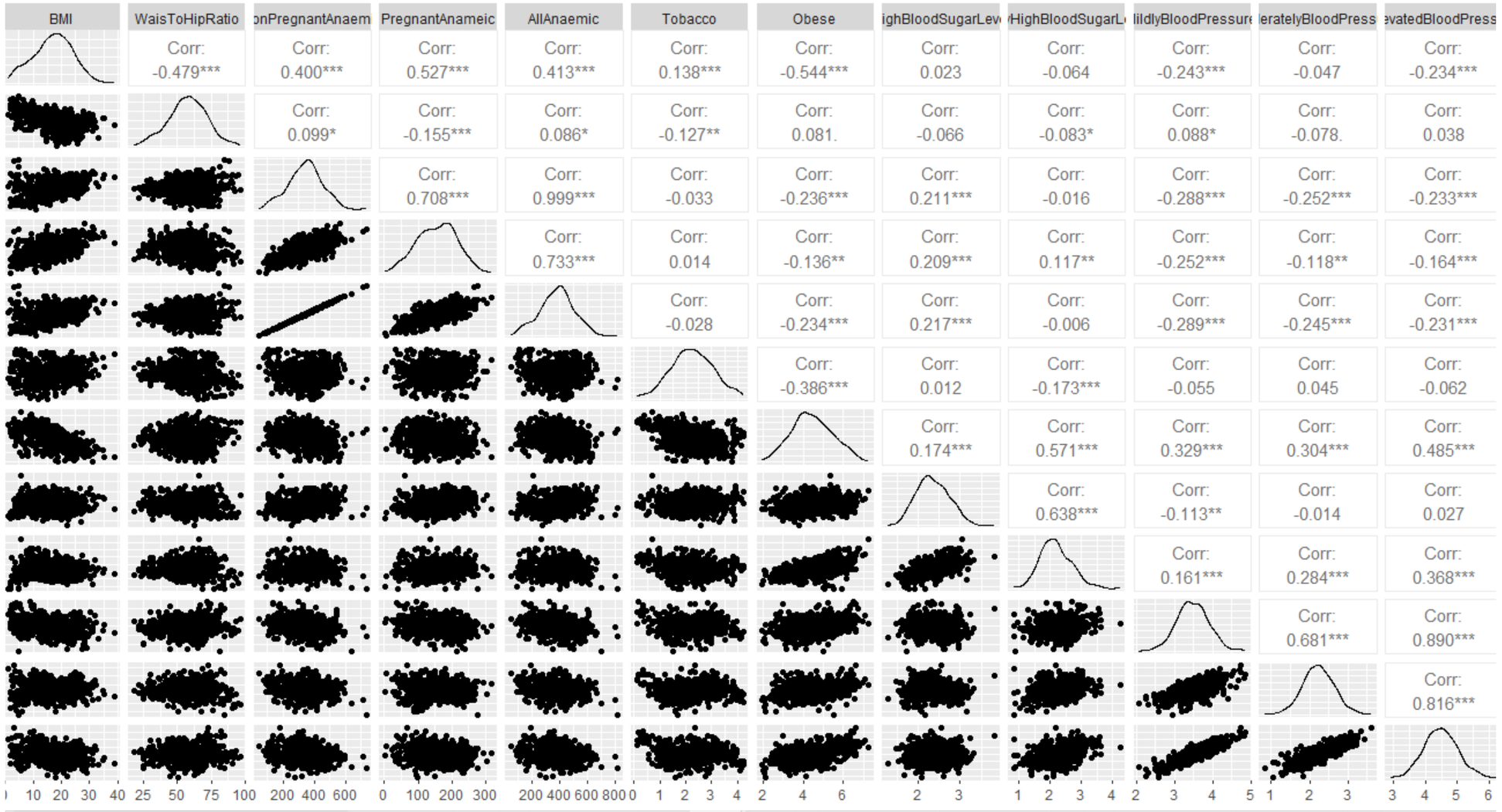








	Test	Variable	Statistic	p value	Normality
1	Anderson-Darling	BMI	1.0401	0.0097	NO
2	Anderson-Darling	WaisToHipRatio	0.5605	0.1469	YES
3	Anderson-Darling	NonPregnantAnaemic	0.6777	0.0764	YES
4	Anderson-Darling	PregnantAnameic	1.0368	0.0099	NO
5	Anderson-Darling	AllAnaemic	0.6717	0.0791	YES
6	Anderson-Darling	Obese	0.8400	0.0303	NO
7	Anderson-Darling	HighBloodSugarLevel	1.0829	0.0076	NO
8	Anderson-Darling	VeryHighBloodSugarLevel	1.5347	6e-04	NO
9	Anderson-Darling	MildlyBloodPressure	0.4174	0.329	YES
10	Anderson-Darling	ModeratelyBloodPressure.	0.2818	0.6376	YES
11	Anderson-Darling	ElevatedBloodPressure	0.3484	0.4756	YES
12	Anderson-Darling	Tobacco	8.1723	<0.001	NO



LINEARITY

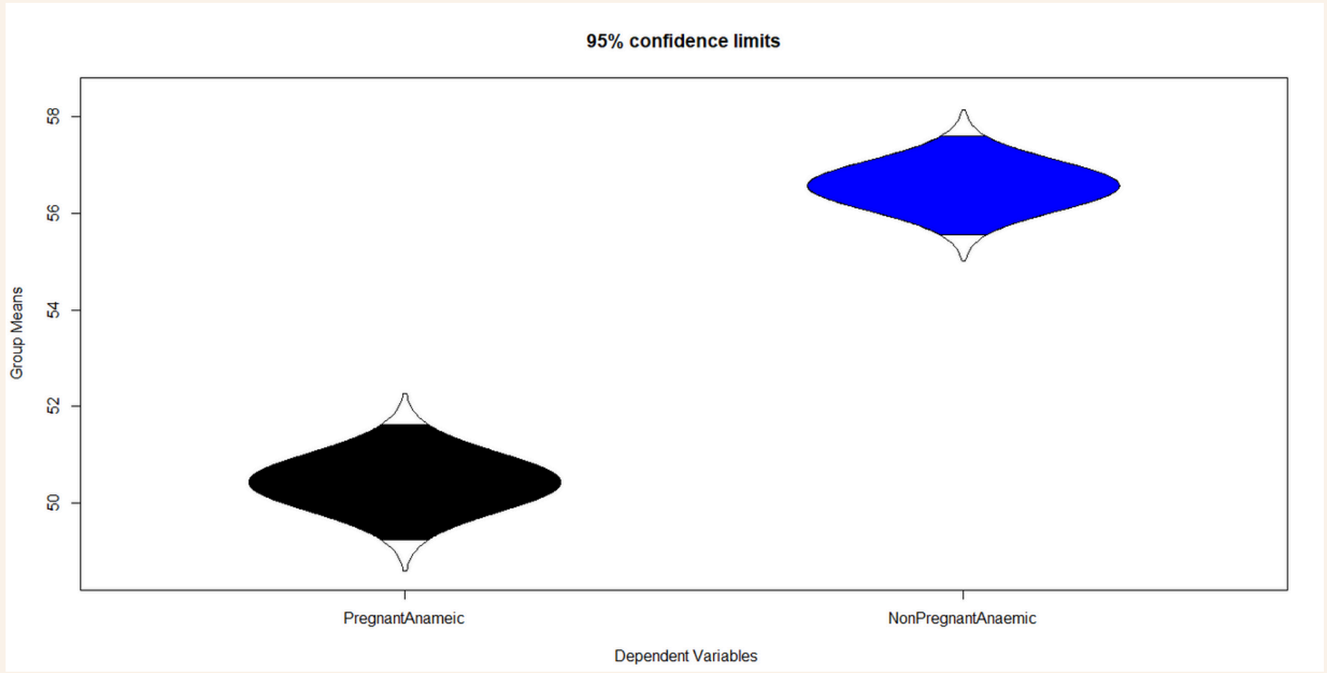
Linearity Assumption is satisfied

Hypothesis Testing

We will test whether there is a difference in the response variable with respect to Rural or Urban area.

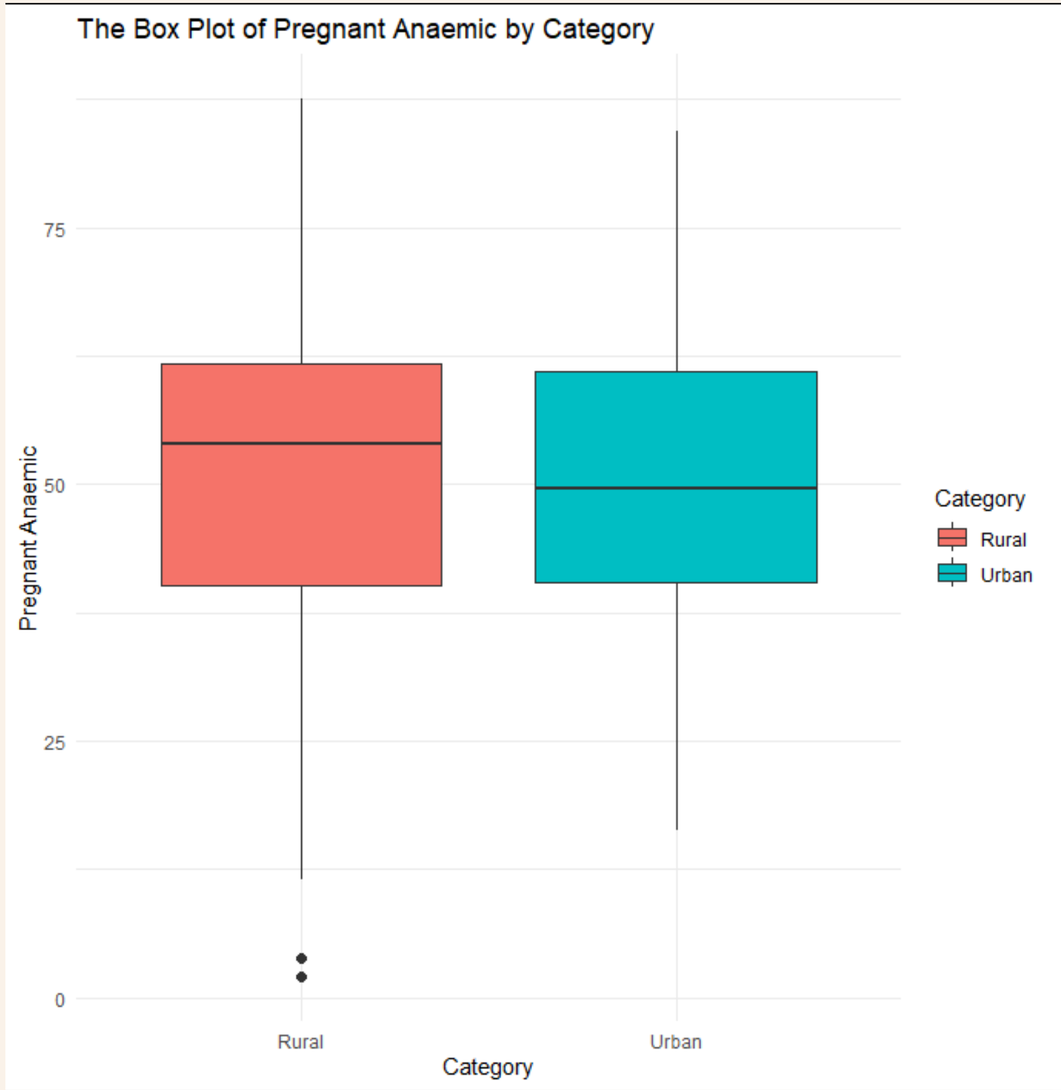
Response: Pregnant
Anameic ,Non-Pregnant
Anaemic

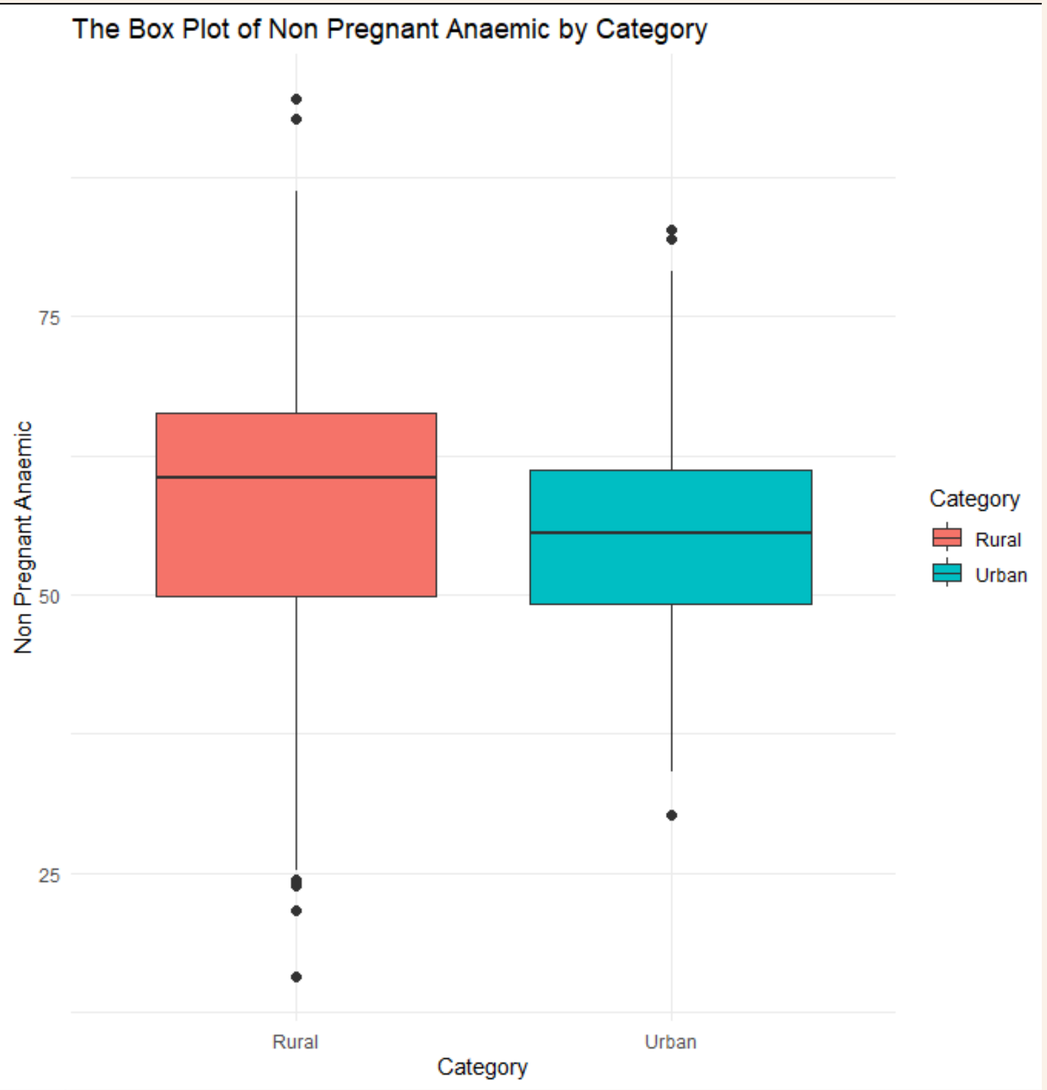
Inference About Means




```
> HotellingsT2(y,mu=mu0)

        Hotelling's one sample T2-test

data:  y
T.2 = 1303.8, df1 = 2, df2 = 571, p-value < 2.2e-16
alternative hypothesis: true location is not equal to
```

One-Way Anova



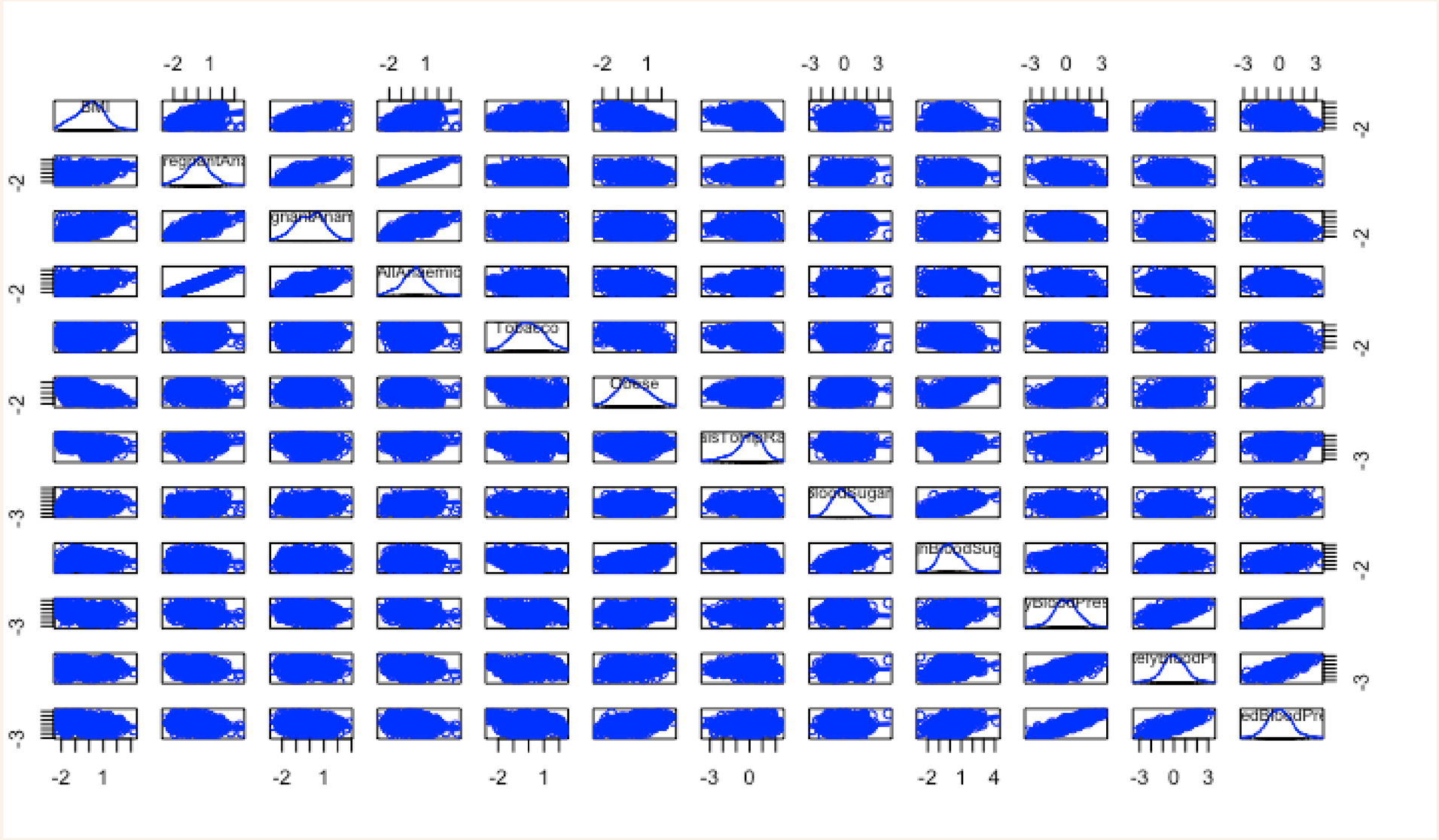


Two-Way Anova

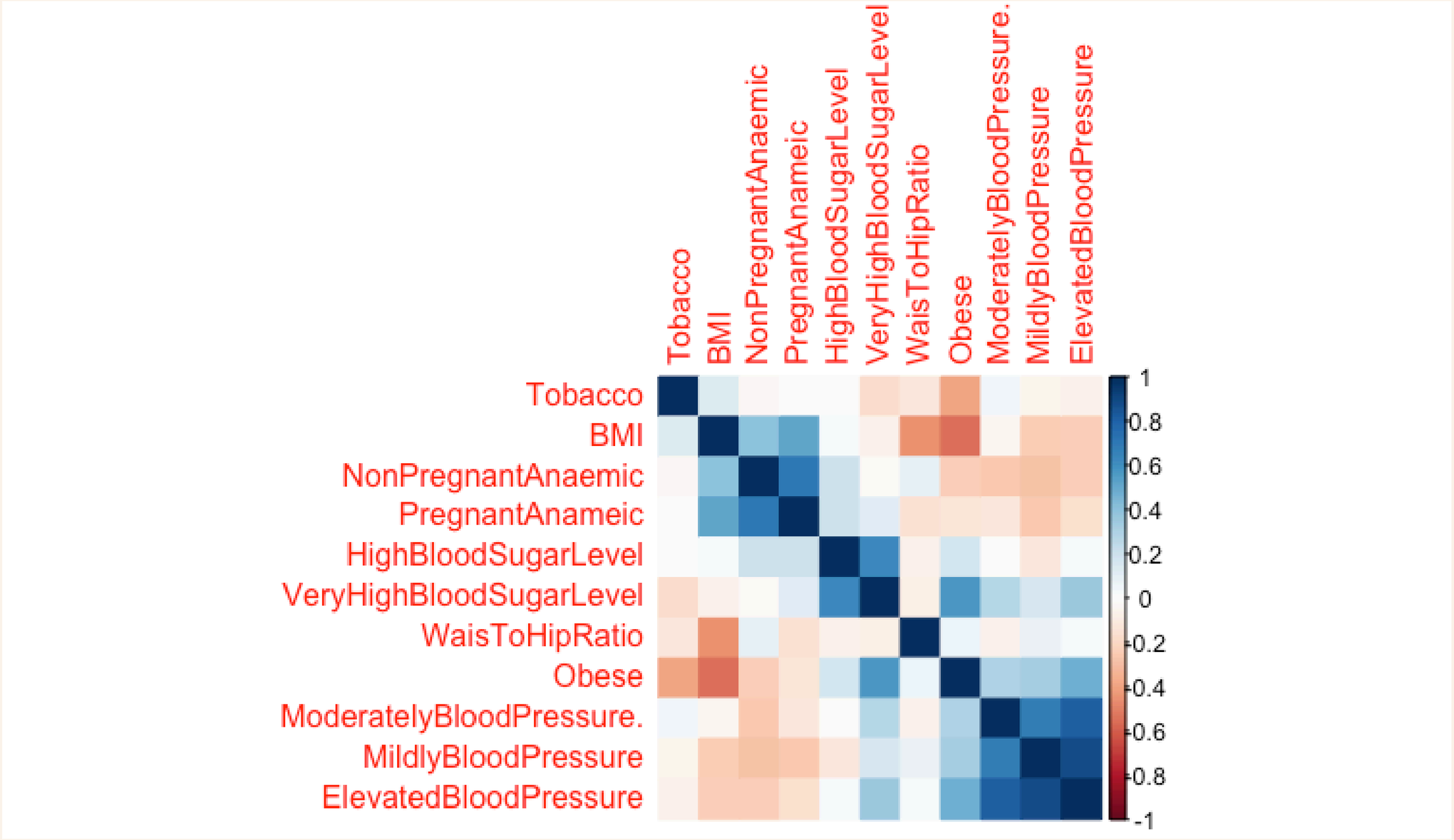
```
> m2 <- manova(cbind(PregnantAnameic,NonPregnantAnaemic) ~ Category*ObeseSituation, data = subset_data1)
> summary(m2)
                        Df   Pillai approx F num Df den Df Pr(>F)
Category                 1 0.009749    2.796      2    568 0.0619 .
ObeseSituation           1 0.155257   52.197      2    568 <2e-16 ***
Category:ObeseSituation  1 0.000530    0.150      2    568 0.8603
Residuals              569
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

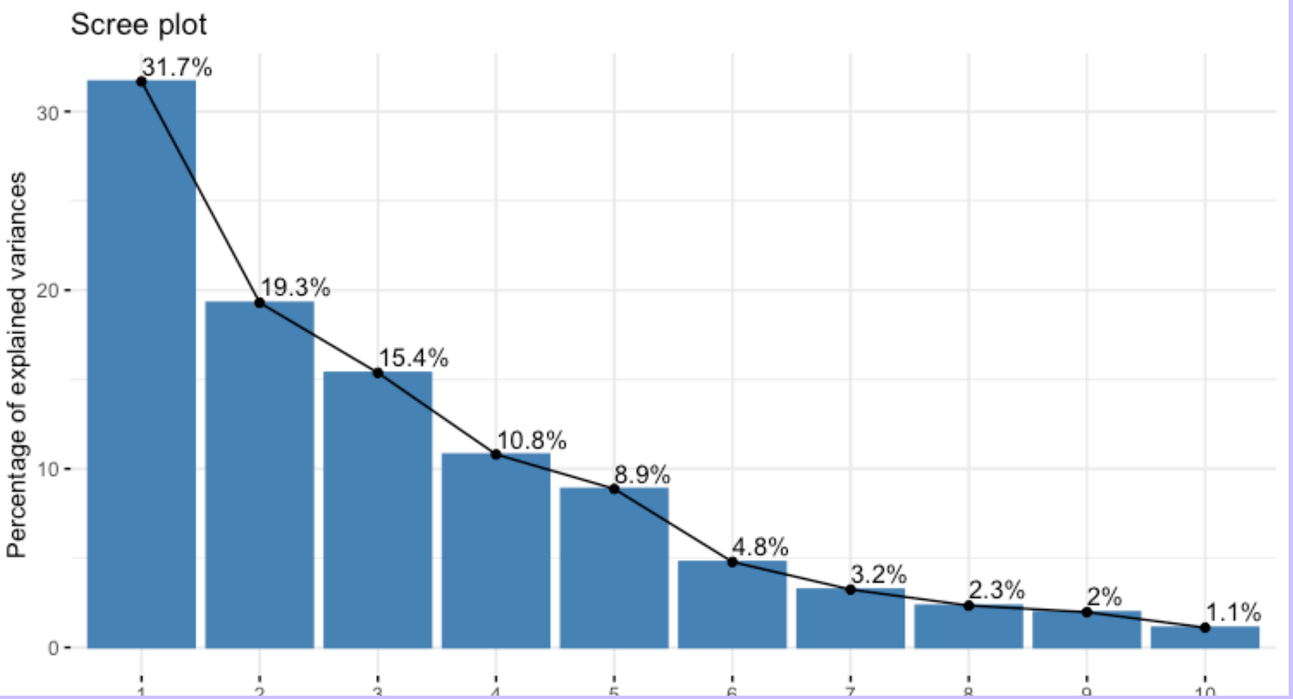
According to this output, the p-value is 0.0619, which is above 0.05. This shows that there is no significant difference between the rural and urban categories.

PRINCIPLE COMPONENT ANALYSIS



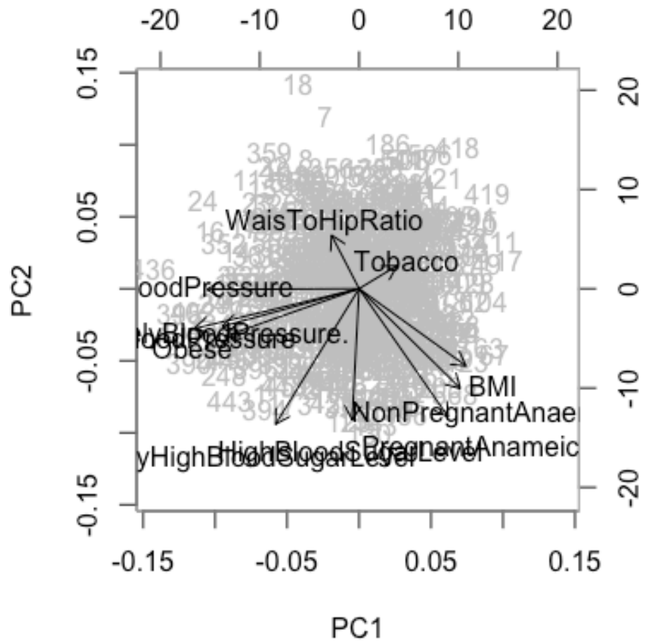
Correlation Heatmap for PCA Analysis

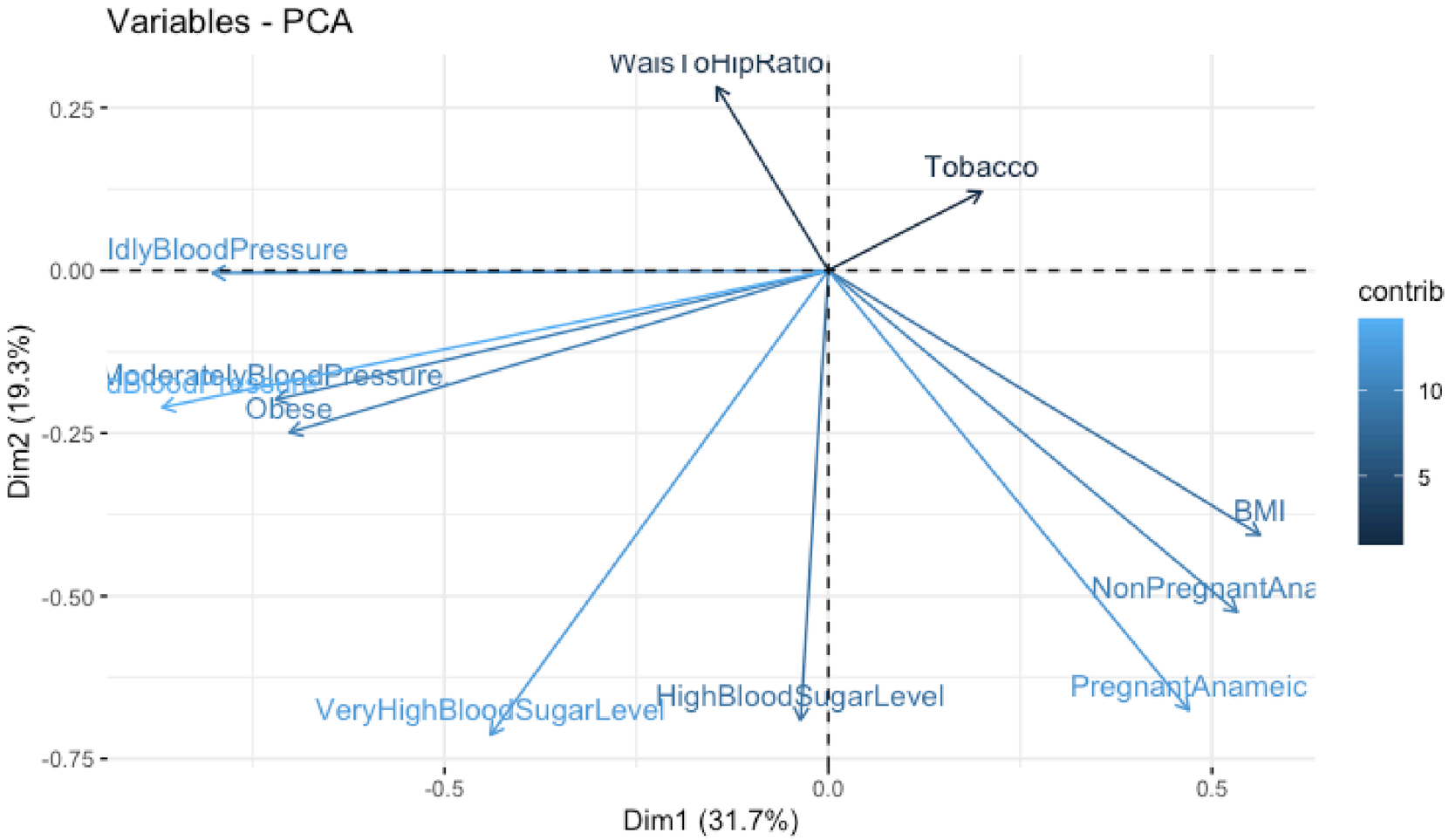




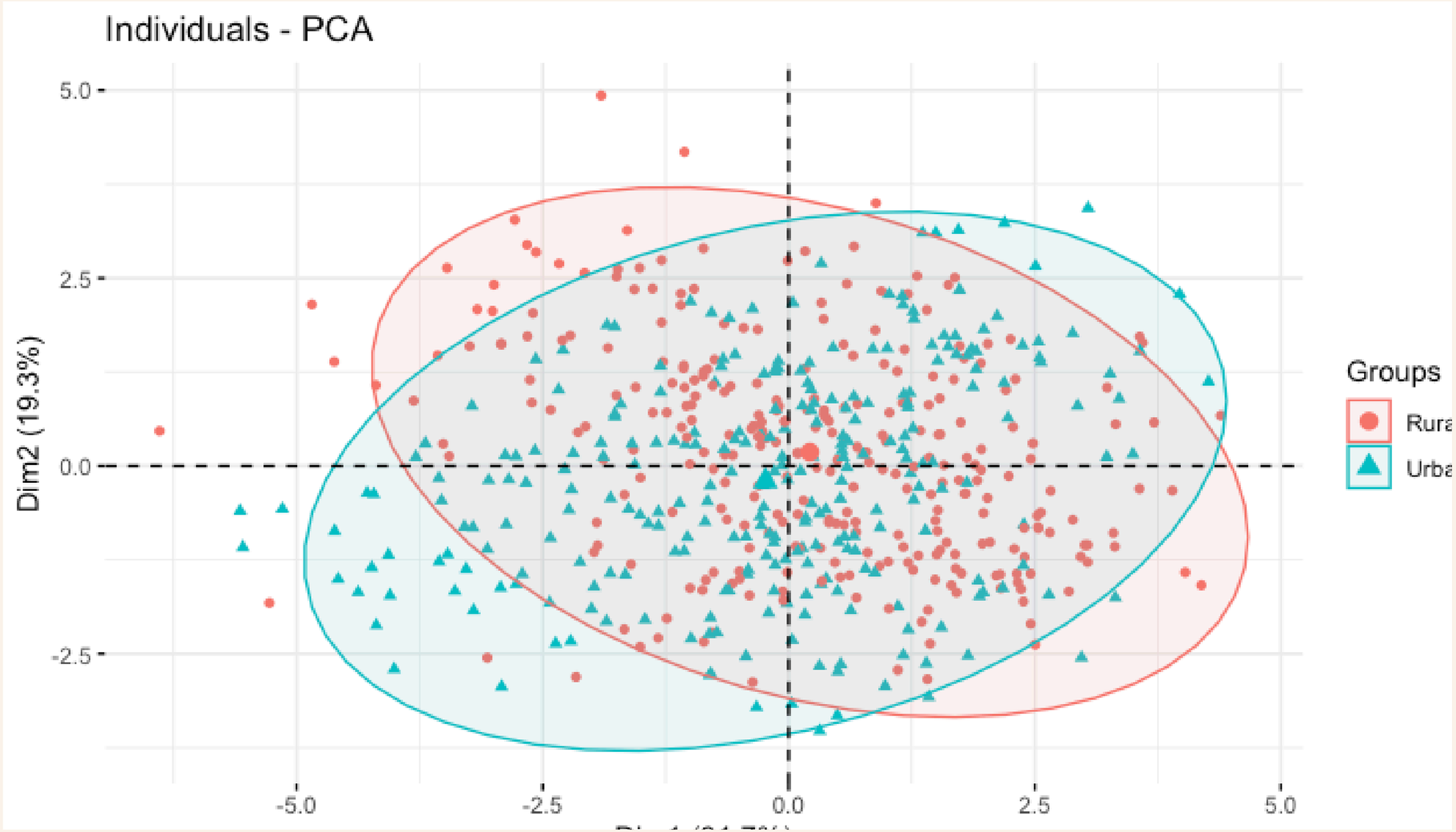
The first four principal components (PC1 to PC4) together explain about 77.5% of the total variance.
PC4 is the elbow point.

BMI, PregnantAnaemic, and NonPregnantAnaemic have strong contributions to PC1 . Tobacco shows a unique direction, indicating less correlation with other variables and contributing independently.





This biplot shows the contributions of the variables to the principal components. The blue arrows are the variables with the highest contributions.



There is significant overlap between the Rural and Urban groups, suggesting that the variance explained by the first two dimensions (Dim1 and Dim2) does not fully differentiate these groups.

Principle Component Regression

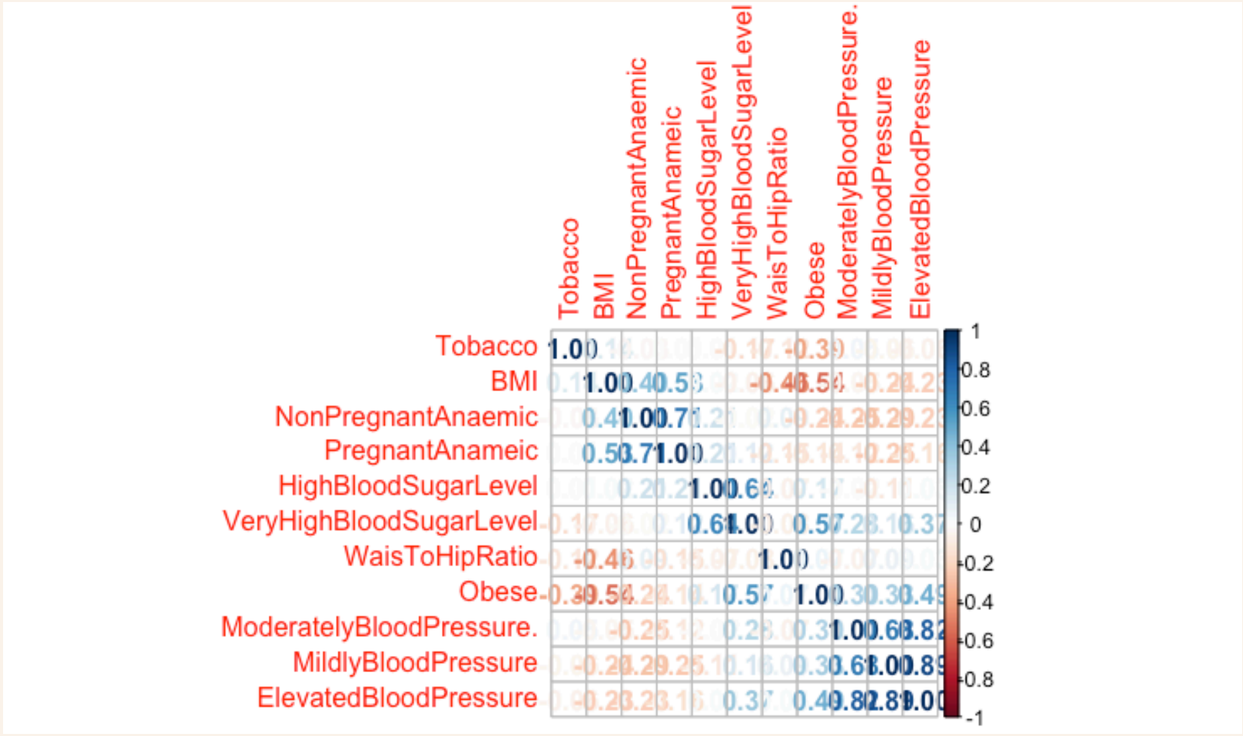
```
Residuals:
    Min      1Q  Median      3Q     Max
-210.27  -61.22    2.46   60.42  405.88

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                -272.0181    66.1254  -4.114 4.48e-05 ***
CategoryUrban               -10.7272     8.0922  -1.326  0.18550
BMI                          10.7688     0.8062  13.357  < 2e-16 ***
WaisToHipRatio               41.9785     4.4501   9.433  < 2e-16 ***
Obese                        20.2184     6.4332   3.143  0.00176 **
HighBloodSugarLevel          85.0483    13.8386   6.146 1.51e-09 ***
VeryHighBloodSugarLevel     -59.4329    13.2945  -4.470 9.44e-06 ***
MildlyBloodPressure         -95.8512    20.9098  -4.584 5.62e-06 ***
ModeratelyBloodPressure.    -94.7363    16.1588  -5.863 7.76e-09 ***
ElevatedBloodPressure       113.5126    23.3442   4.863 1.51e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 86.83 on 562 degrees of freedom
Multiple R-squared:  0.3983,    Adjusted R-squared:  0.3886
F-statistic: 41.33 on 9 and 562 DF,  p-value: < 2.2e-16
```

According to the output in the table, the model is significant, the Adjusted R-squared value is 0.3886.

FACTOR ANALYSIS

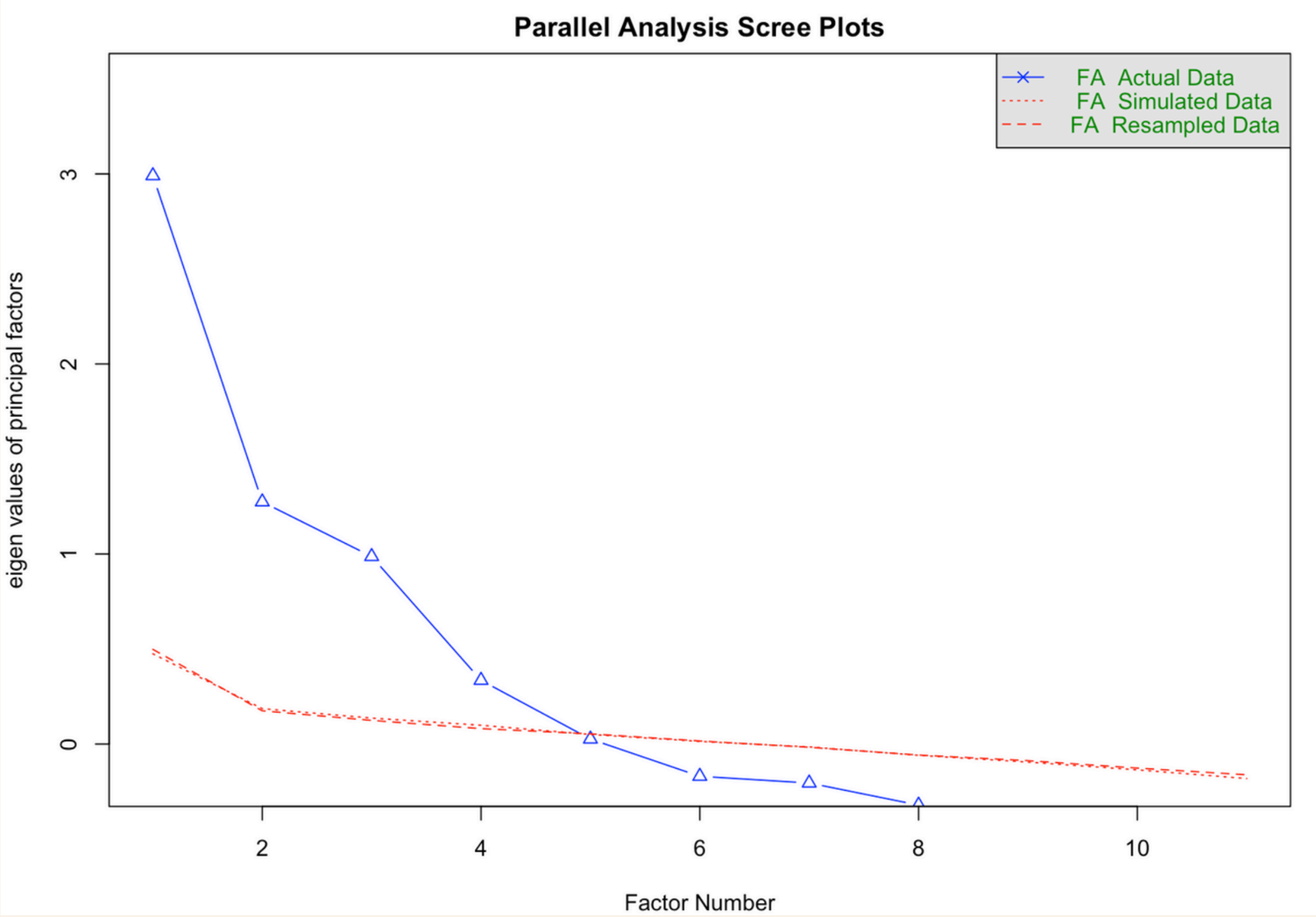
```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cm)
Overall MSA =  0.58
MSA for each item =
                      BMI       NonPregnantAnaemic           PregnantAnameic
                     0.51                     0.59                      0.65
                  Tobacco                    Obese           WaisToHipRatio
                     0.46                     0.55                      0.36
     HighBloodSugarLevel  VeryHighBloodSugarLevel       MildlyBloodPressure
                     0.46                     0.53                      0.66
ModeratelyBloodPressure.    ElevatedBloodPressure
                     0.73                     0.62
```

The Overall MSA value was calculated as 0.58, which indicates that the data set is poorly suitable for factor analysis. Variables with a KMO value below 0.50 were not included in the factor analysis.



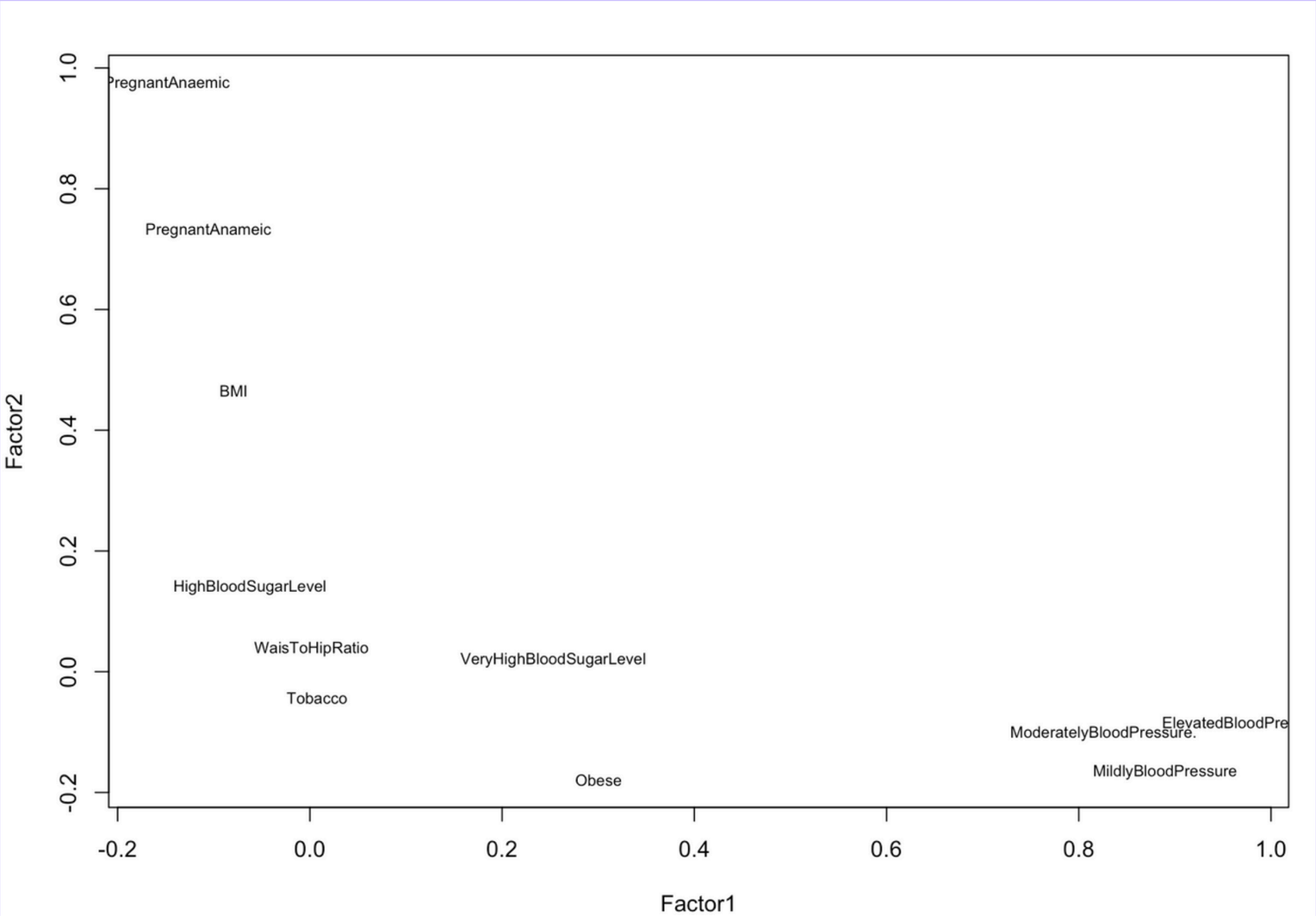
VISUALIZATION

Scree Plot



- The Scree Plot is a crucial visualization used in factor analysis to help decide the optimal number of factors to retain.
- A noticeable "elbow" at the 4th factor suggests retaining 4 factors, as they explain the majority of the variance.
- Confirms that eigenvalues beyond the 4th factor are lower than simulated data, indicating they are not significant.

Factor Loadings Plot

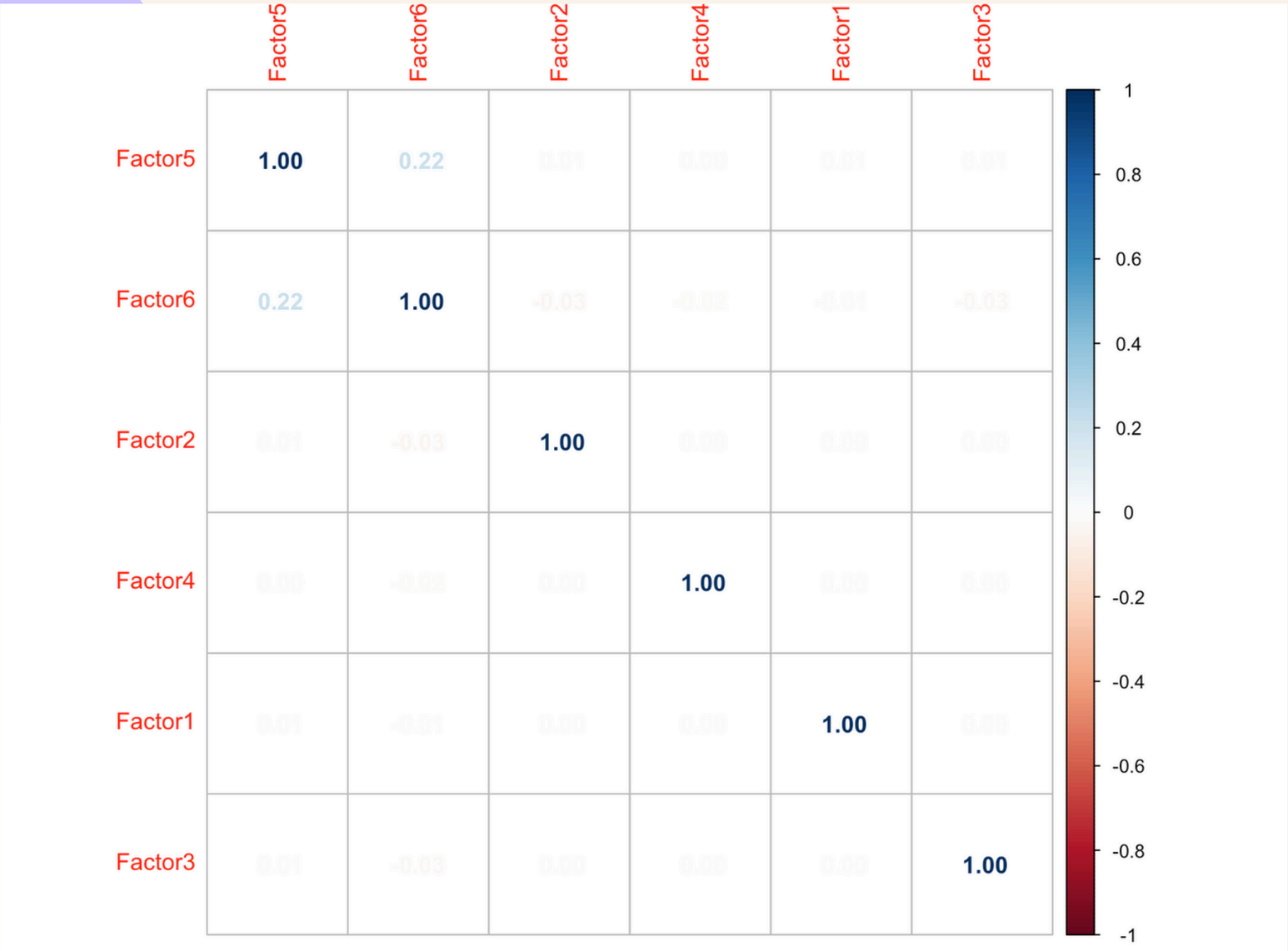


- Displays how variables contribute to the first two factors (Factor1 and Factor2).
- Variables like PregnantAnaemic and BMI strongly contribute to Factor2.
- Variables like ElevatedBloodPressure and MildlyBloodPressure are key for Factor1.
- Some variables, such as Tobacco and WaistToHipRatio, have minimal contributions to either factor.

Factor Scores Correlation Heatmap

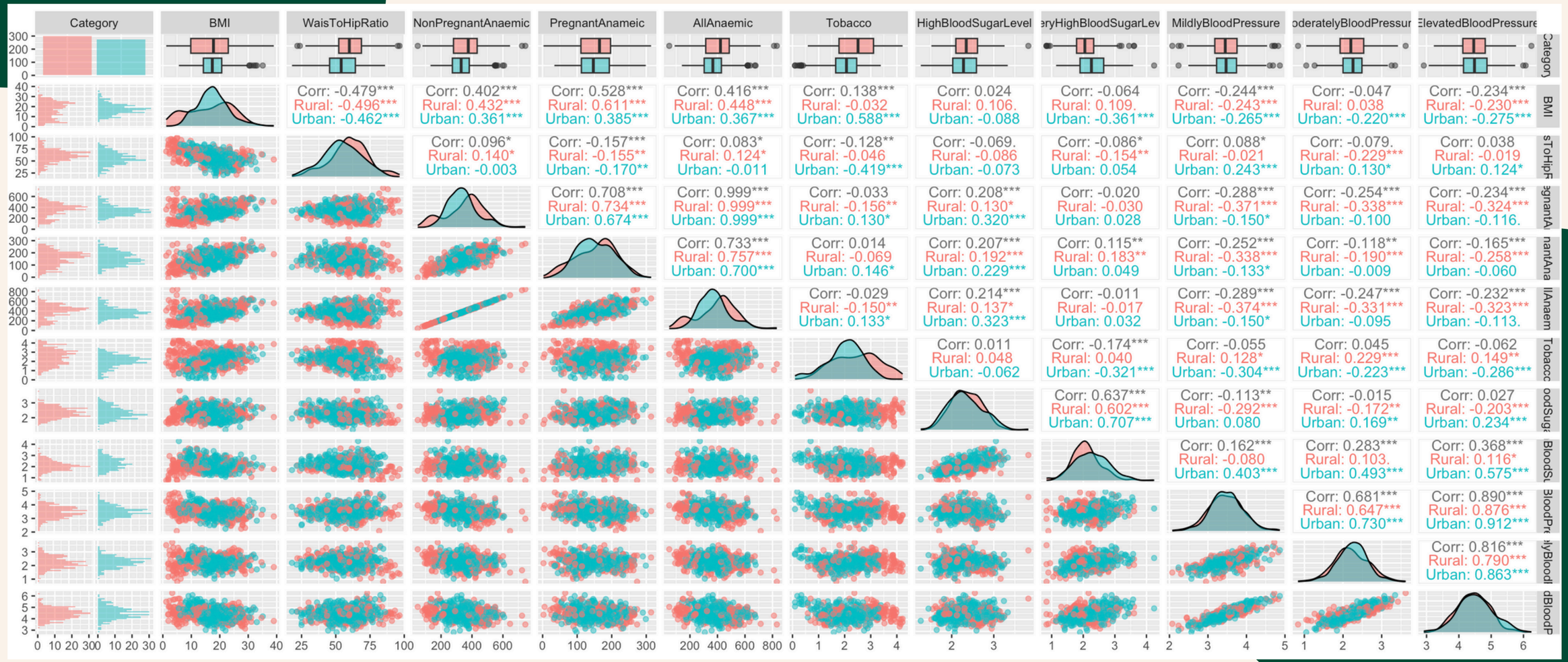
The low off-diagonal correlations confirm that the extracted factors are distinct and do not overlap significantly.

The independence of extracted factors confirms the success of the factor analysis and ensures robust insights.



LINEAR DISCRIMINANT ANALYSIS AND CLASSIFICATION

The Linear Discriminant Analysis was conducted to measure differences between Rural and Urban populations based on health-related variables.



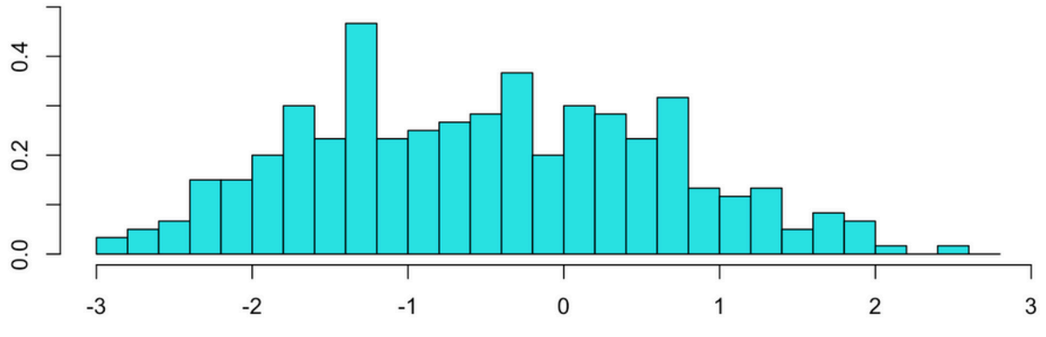
The dataset contains two groups (Rural and Urban), and the analysis aimed to identify key predictors that differentiate these categories.

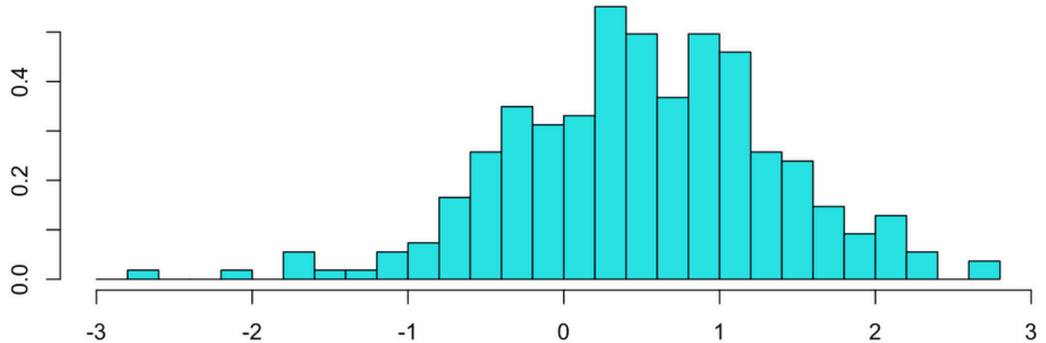
```
Prior probabilities of groups:
    Rural     Urban
0.5244755 0.4755245
```

```
Coefficients of linear discriminants:
                                    LD1
BMI                          0.02709971
WaisToHipRatio              -0.03489783
NonPregnantAnaemic          -0.08960737
PregnantAnameic             -0.01093365
AllAnaemic                   0.08182960
Tobacco                     -0.84359136
HighBloodSugarLevel         -1.64089264
VeryHighBloodSugarLevel      1.55633787
MildlyBloodPressure          0.10076814
ModeratelyBloodPressure.    -0.70308435
ElevatedBloodPressure        0.12155177
```

LDA Model Equation:

LD1=0.0271 · BMI−0.0349 · WaisToHipRatio−0.0896 · Non PregnantAnaemic−0.0109 · PregnantAnameic+0.0818 · All Anaemic−0.8436 · Tobacco..... −0.7031 · ModeratelyBloodPressure+0.1216 · ElevatedBloo dPressure



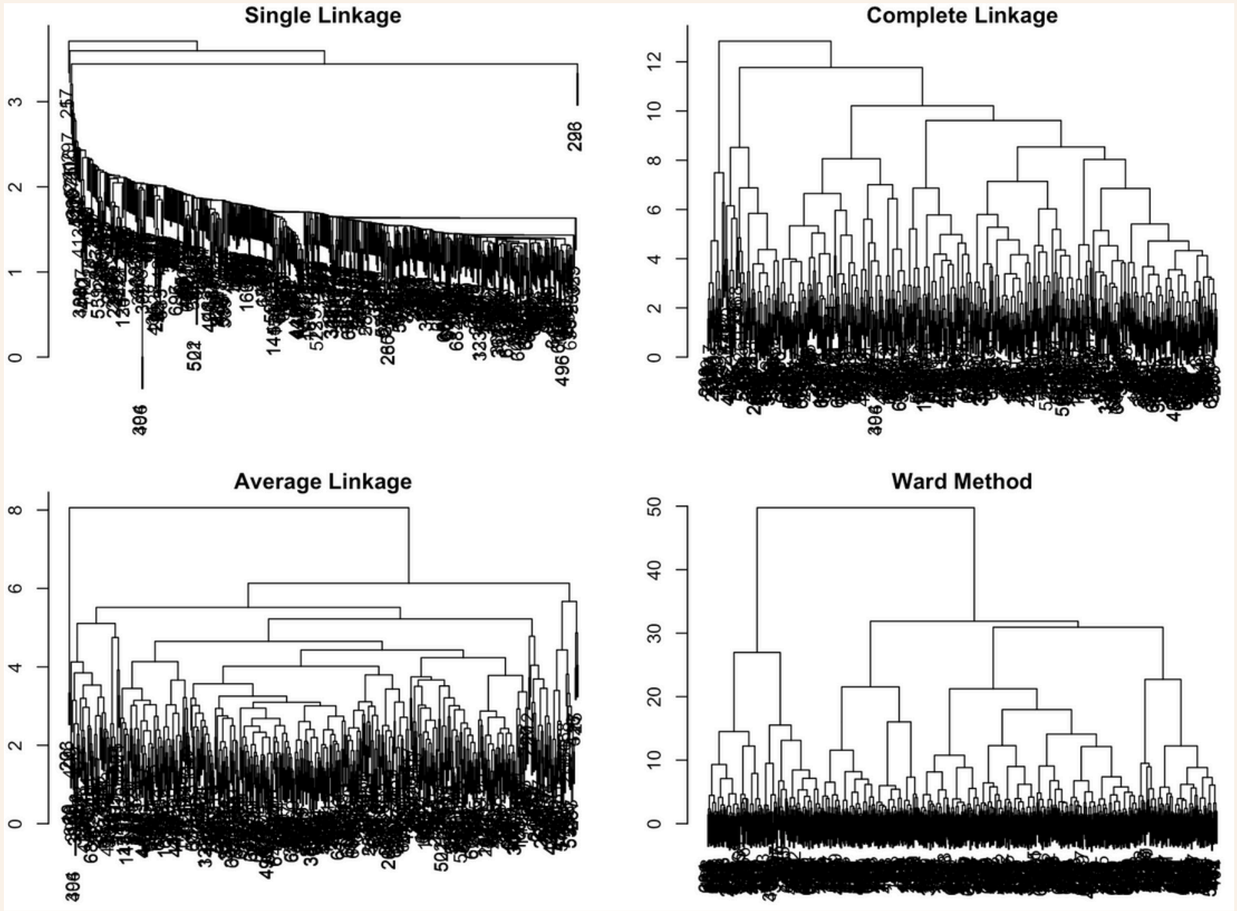


PARTITION PLOT

LINEAR DISCRIMINANT ANALYSIS AND CLASSIFICATION

To uncover patterns and groupings within the dataset using clustering techniques.

As a key variables we used numerical such as BMI, Blood Presure Levels and others

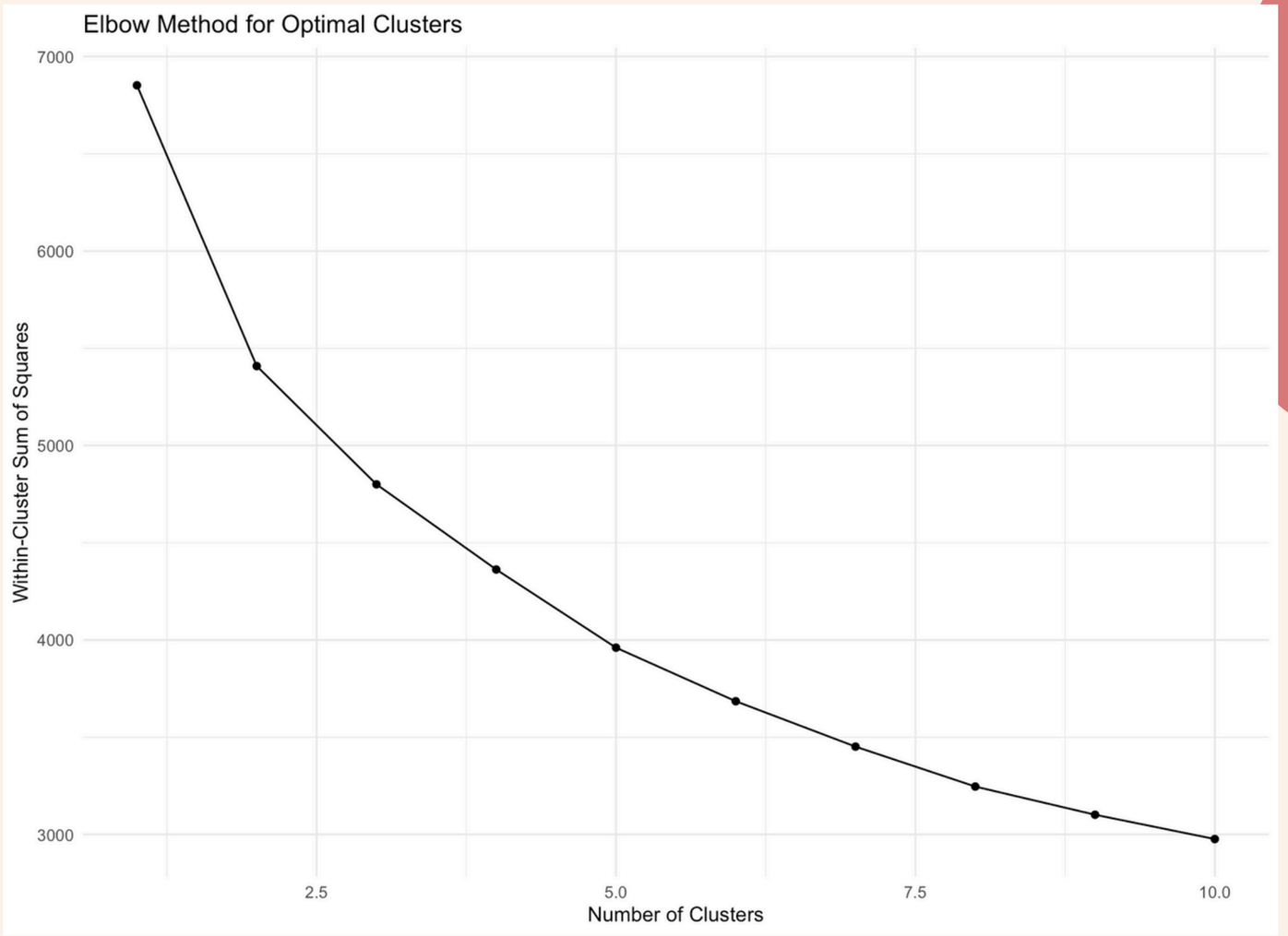


Dendrograms

Ward's method provides the most interpretable results.

Elow Method

WSS decreases sharply from 1 to 3 clusters, indicating that adding these clusters significantly improves the grouping of the data.



Elbow is observed at 3 clusters, suggesting this is the optimal number of clusters for the dataset.

Clusters based on the K-Means Cluster Centers

	BMI	WaisToHipRatio	NonPregnantAnaemic	PregnantAnameic	AllAnaemic	Tobacco
1	10.86258	249.833032	2.020894	1.018509	14.07157	137.622435
2	36.73756	539.877590	1.155629	1.232199	35.01262	461.423114
3	389.82134	3.010238	2.225162	2.196439	453.02264	2.377122

	Obese	HighBloodSugarLevel	VeryHighBloodSugarLevel	MildlyBloodPressure
1	5.505834	2.565366	27.21871	430.374620
2	1.918069	2.355841	43.72270	945.796936
3	2.649308	3.331021	470.78563	4.851125

	ModeratelyBloodPressure.	ElevatedBloodPressure
1	4.827743	4.182198
2	1.940401	3.588118
3	2.614690	4.324044

Cluster 1:
This group appears to have a balanced health profile with no extreme health risks but moderate use of tobacco.

Cluster 2:
This group shows unhealthy lifestyle behaviors, such as high tobacco use, and potential risks related to metabolism

Cluster 3:
group represents individuals with critical health concerns, including severe metabolic and blood sugar issues.

CONCLUSION

REFERENCES

Johnson, R. A., & Wichern, D. W. (2018). Applied Multivariate Statistical Analysis.

Kumar, P., & Anand, S. (2020). Multivariate statistical analysis for health outcomes: Case study on anemia. Journal of Public Health Analytics, 7(3), 150–162.

Tesfaye, T. S., Tessema, F., & Jarso, H. (2020). Prevalence of anemia and associated factors among "apparently healthy" urban and rural residents in Ethiopia: A comparative cross-sectional study. Journal of Blood Medicine, 11, 89–96. https://pmc.ncbi.nlm.nih.gov/articles/PMC7073428/pdf/jbm-11-89.pdf

THANK YOU FOR LISTENING US