A Multivariate Data Analysis on Radiotherapy Data

Samya Debepsita Rupayan

April 9, 2023

Abstract

In this project we have analyzed radiotherapy data. The set contains data on people undergoing radiotherapy. Variables like number of symptoms, average appetite and average sleep (etc) throughout their treatment were considered. The dataset contained a categorical variable for skin reaction. The analysis includes EDA, normality checking and transformations, MANOVA, LDA, PCA and factor analysis. By studying the correlation structure of the data and Kendall's tau test we conclude that one of the variables is independent from others and consequently we drop that from our analysis. We checked whether the observations in different categories belong to the same population. From EDA and MANOVA we conclude that the observations indeed come from the same population.Consequently, discriminant analysis is not meaningful. This fact is supported by APER calculated using one-out cross validation method. From factor analysis we conclude that one variable is significantly different compared to others. Also among the other four, there are two factors affecting the variables. One factor is affecting appetite and food consumed. This factor could be due to the treatment's effect on mouth/digestive system. The other is affecting the number of symptoms and sleep. .

Part I
Introduction

In any data analysis, it is of paramount importance to know two things. One is domain knowledge. We must have atleast some understanding of what the variables are about to be able to make sound analysis. Secondly, we also need to have proper understanding of the background of the study from which the data was generated. We start with the domain knowledge.

Radiotherapy and its side effects

Radiation therapy or radiotherapy, often abbreviated RT, RTx, or XRT, is a therapy using ionizing radiation, generally provided as part of cancer treatment to control or kill malignant cells.Radiation therapy is in itself painless.It causes minimal or no side effects, although short-term pain flare-up can be experienced in the days following treatment.Most side effects are predictable and expected. Side effects from radiation are usually limited to the area of the patient's body that is under treatment.Some of the poential side effects are as follows :

1. Nausea and vomiting.
2. Epithelial surfaces may sustain damage from radiation therapy. Depending on the area being treated, this may include the skin, oral mucosa, pharyngeal, bowel mucosa and ureter. (Skin Reaction)
3. If the head and neck area is treated, temporary soreness and ulceration commonly occur in the mouth and throat. If severe, this can affect swallowing, and the patient may need painkillers and nutritional support/food supplements. The esophagus can also become sore if it is treated directly, or if, as commonly occurs, it receives a dose of collateral radiation during treatment of lung cancer.
4. Cognitive decline.

In this project we are interested to check whether we can identify different factors influencing different aspects like appetite and sleep of a patient throughout their treatment. We are also interested in checking whether we can classify different patients based on their skin reaction (to the treatment).

Background of the Data

Average ratings over the course of treatment for patients undergoing radiotherapy.

1. Col. 1: (V1) = number of symptoms
2. Col. 2: (V2) = amount of activity (1-5 scale)
3. Col. 3: (V3) = amount of sleep (1-5 scale)
4. Col. 4: (V4) = amount of food consumed (1-3 scale)

5. Col. 5: (V5) = appetite (1-5 scale))
6. Col. 6: (V6) = skin reaction (0, 1, 2 or 3)

As they have taken average of ratings, we have assumed that $V_1, .., V_5$ are continuous variables.

In this report Vi or V_i will be interchangably used for the following $(i = 1, .., 6)$:

- V1 : AVERAGE NUMBER OF SYMPTOMS THROUGHOUT THE TREATMENT
- V2: AVERAGE AMOUNT OF ACTIVITY THROUGHOUT THE TREATMENT
- V3: AVERAGE AMOUNT OF SLEEP THROUGHOUT THE TREATMENT
- V4: AVERAGE AMOUNT OF FOOD CONSUMED THROUGHOUT THE TREATMENT
- V5 : V2: AVERAGE AMOUNT OF APPETITE THROUGHOUT THE TREATMENT
- V2: SKIN REACTION AFTER THE TREATMENT

Organisation of the Report

The report has been divided into different parts. Each part signifies a particular important step in the analysis. We have started with EDA (Exploratory data analysis), followed by checking and correcting for univariate and multivariate normality. These have been followed by MANOVA, LDA,PCA and Factor Analysis respectively.

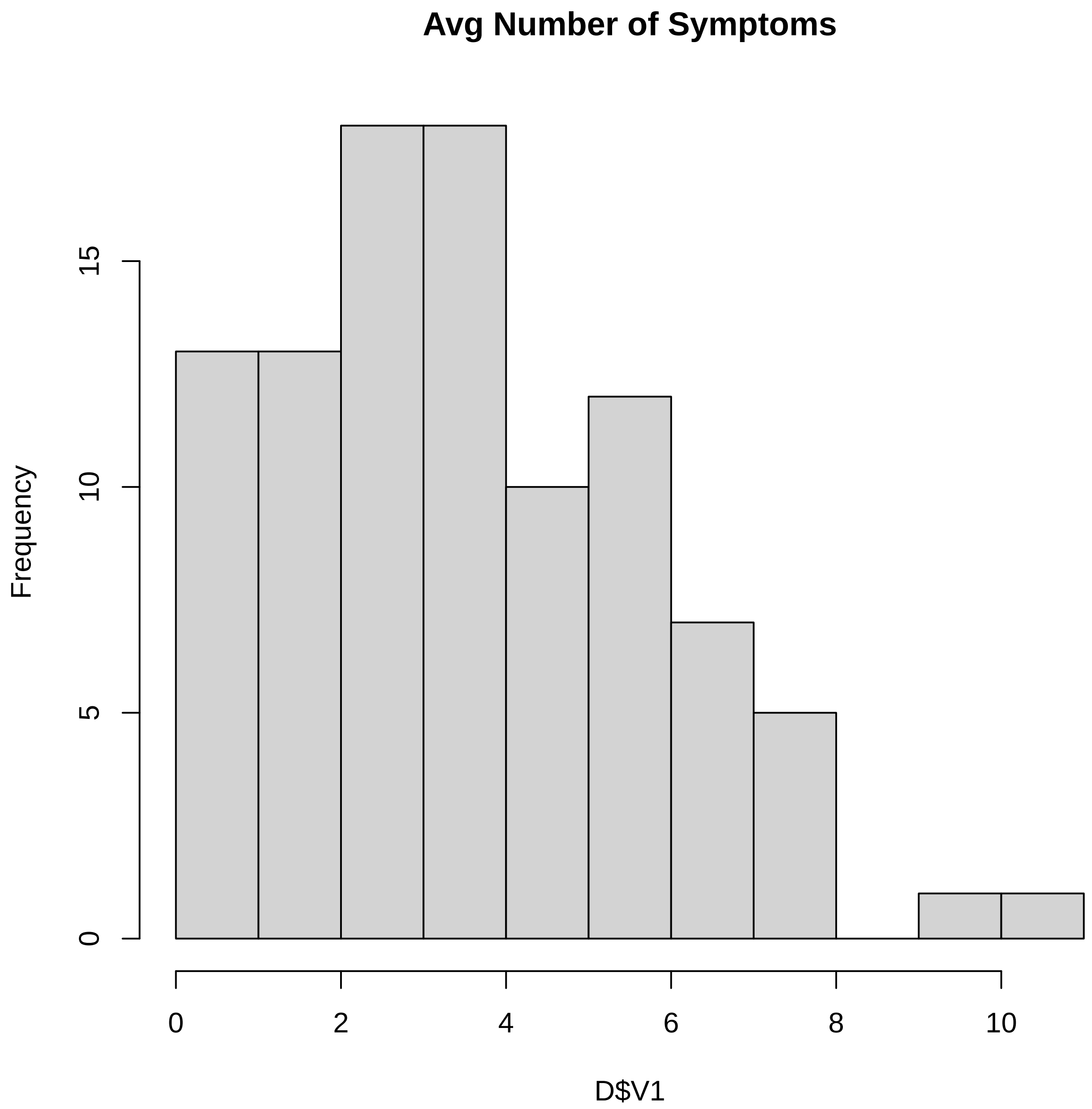
Part II
EDA [EXPLORATORY DATA ANALYSIS]

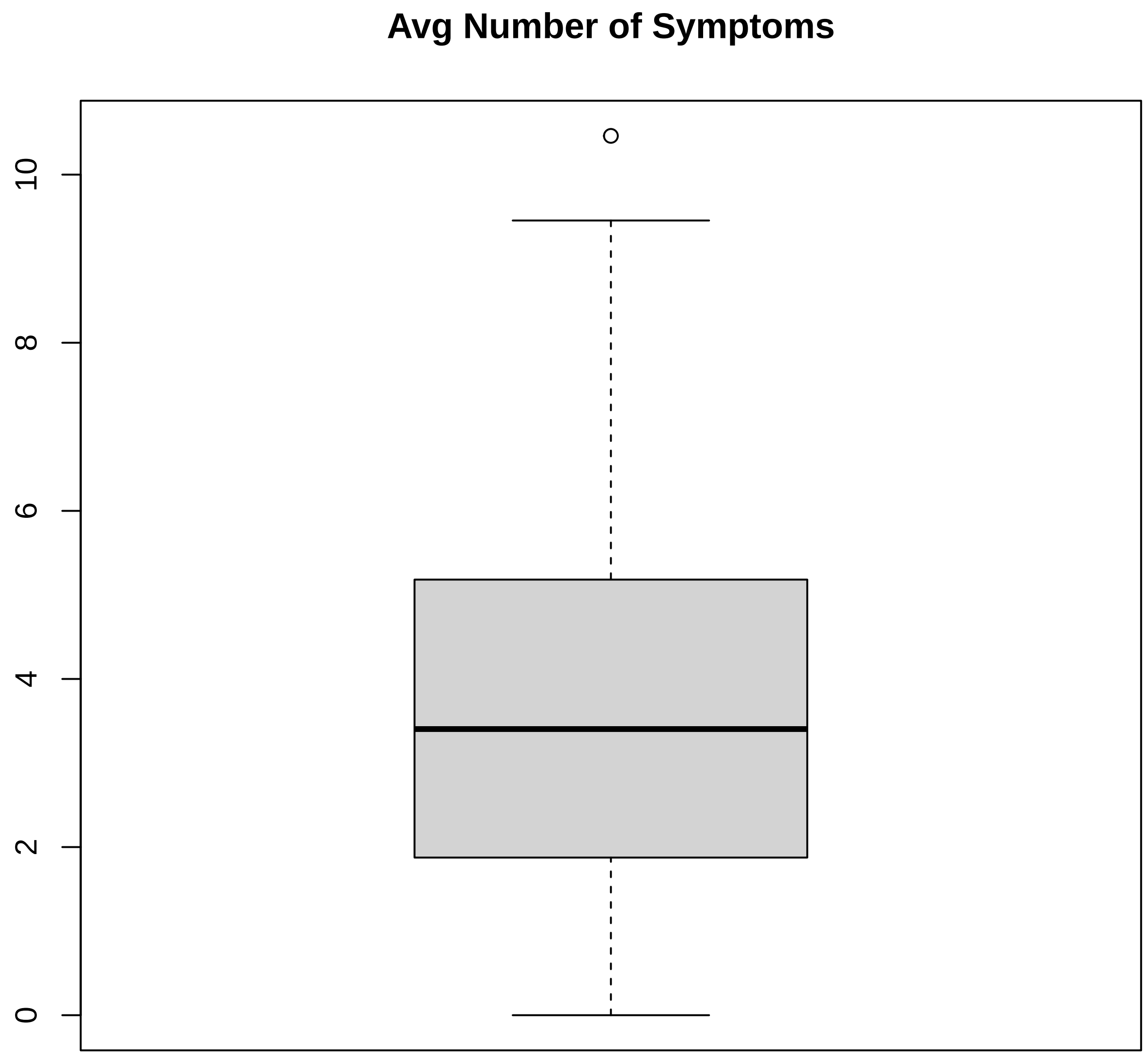
```
## Loading required package:  car
## Loading required package:  carData
## Loading required package:  broom
##
## Attaching package:  ’rgl’
## The following object is masked from ’package:heplots’:
##
##     arrow3d
## *** Package RVAideMemoire v 0.9-81-2 ***
##
## Attaching package:  ’RVAideMemoire’
## The following object is masked from ’package:broom’:
##
##     bootstrap
##
## Attaching package:  ’psych’
## The following objects are masked from ’package:ggplot2’:
##
##     %+%, alpha
## The following object is masked from ’package:car’:
##
##     logit
## Loading required package:  class
## Loaded mda 0.5-3
## --
## biotools version 4.2
##
## Attaching package:  ’biotools’
## The following object is masked from ’package:heplots’:
##
##     boxM
```

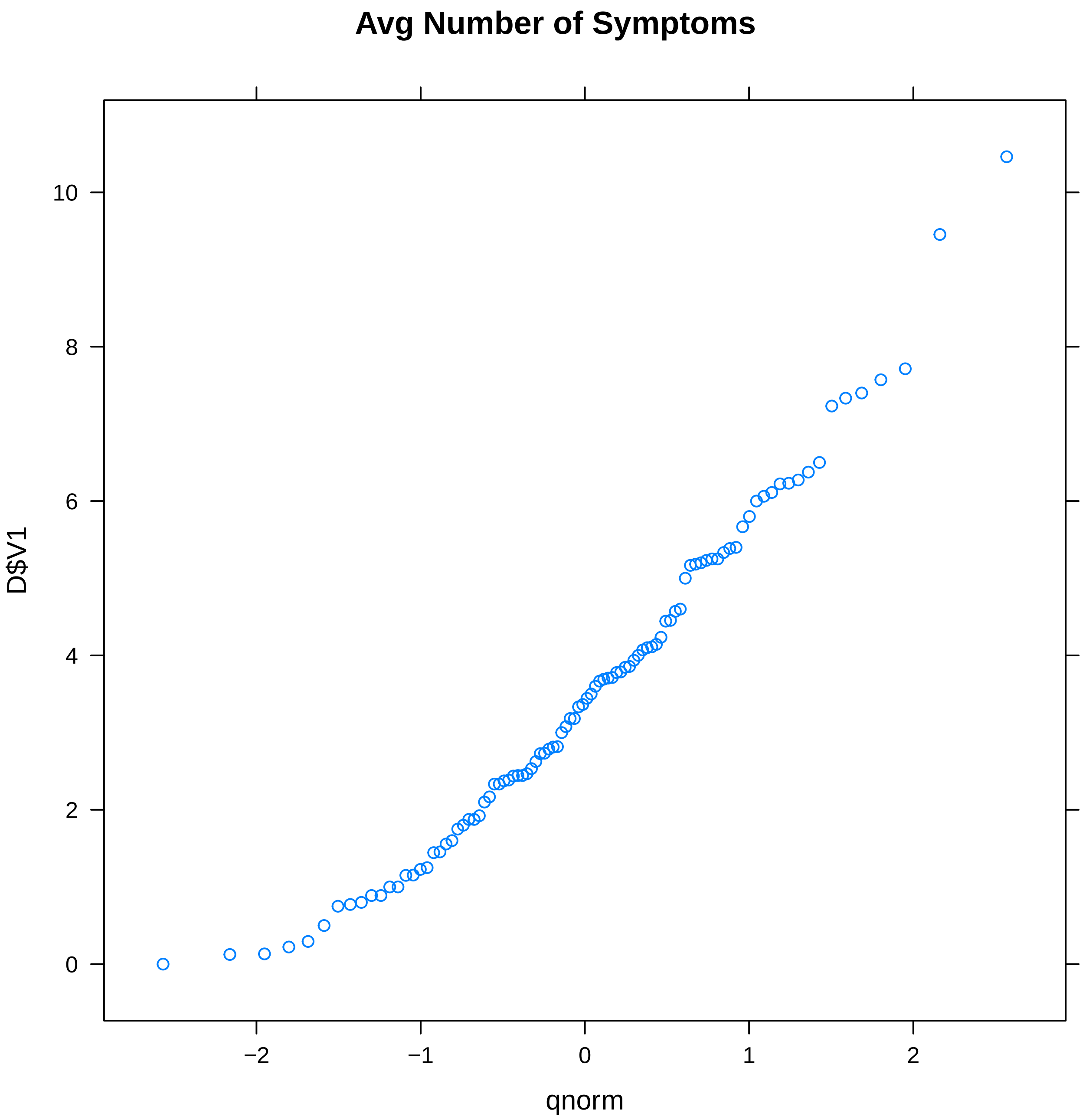
0.1 Univariate EDA

0.1.1 Average number of symptoms

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.887   3.404   3.542   5.178  10.461
```

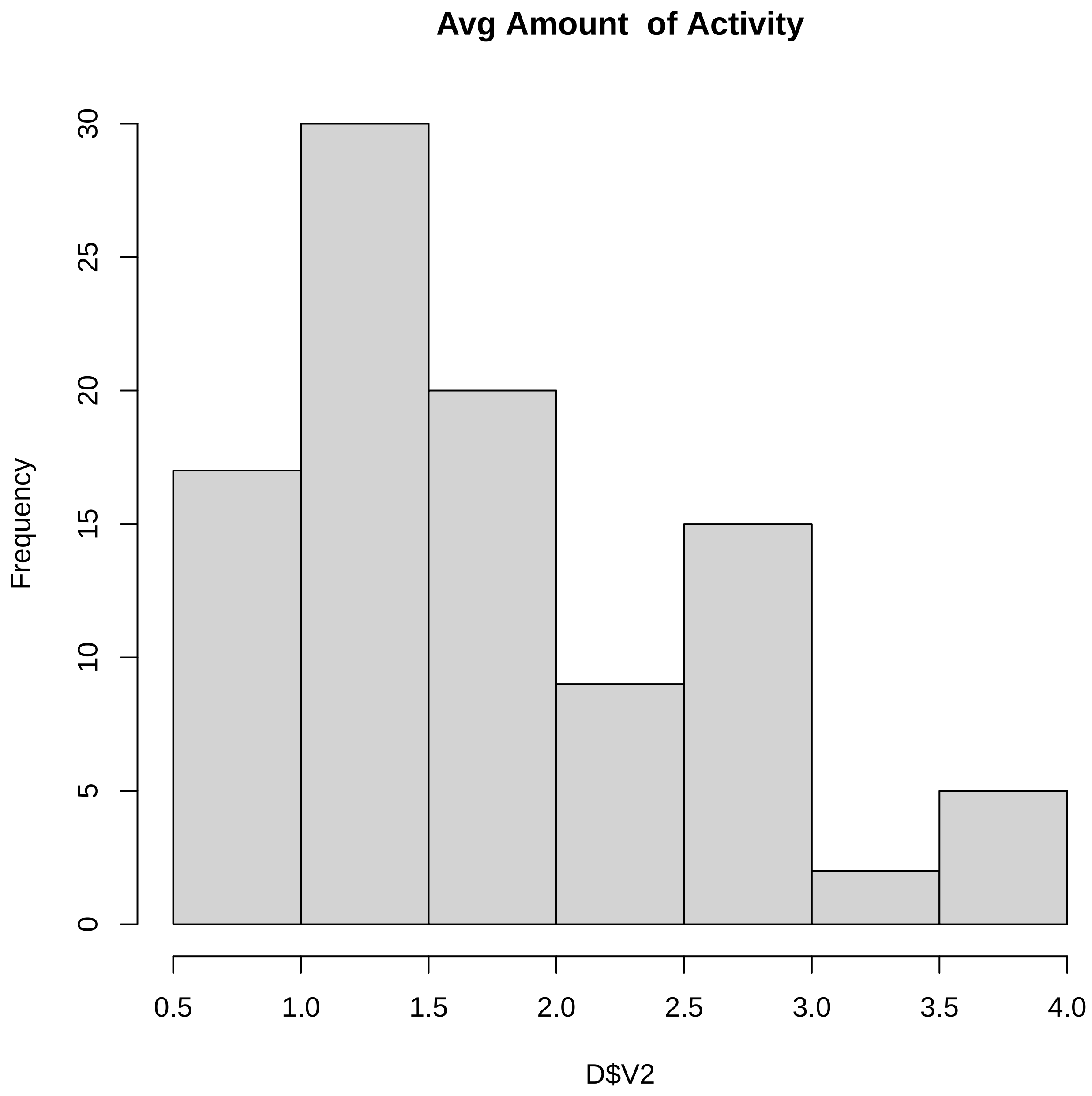

Avg Number of Symptoms
Frequency
0
5
10
15
0
2
4
6
8
10
D$V1

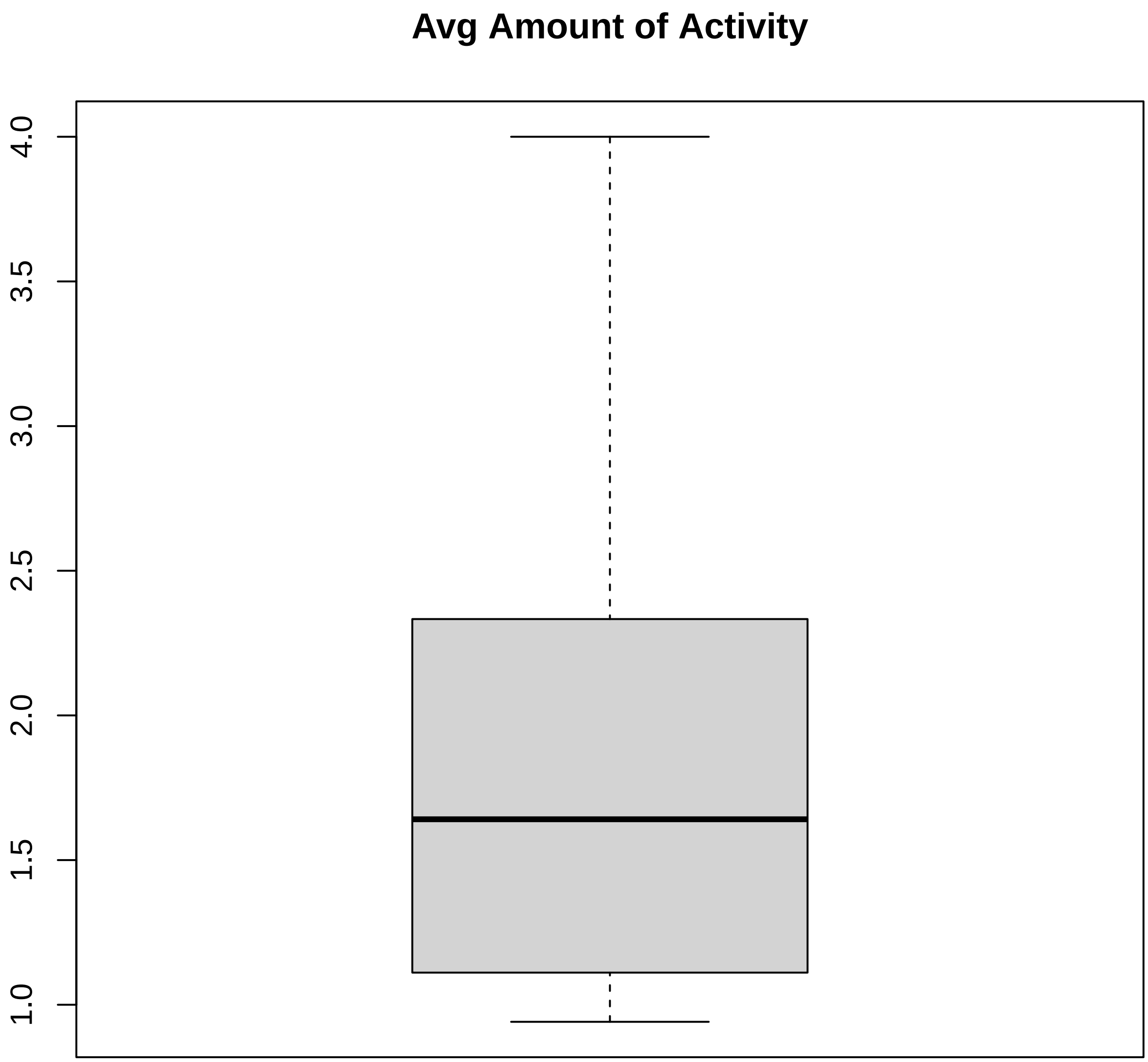

Avg Number of Symptoms
0
2
4
6
8
10

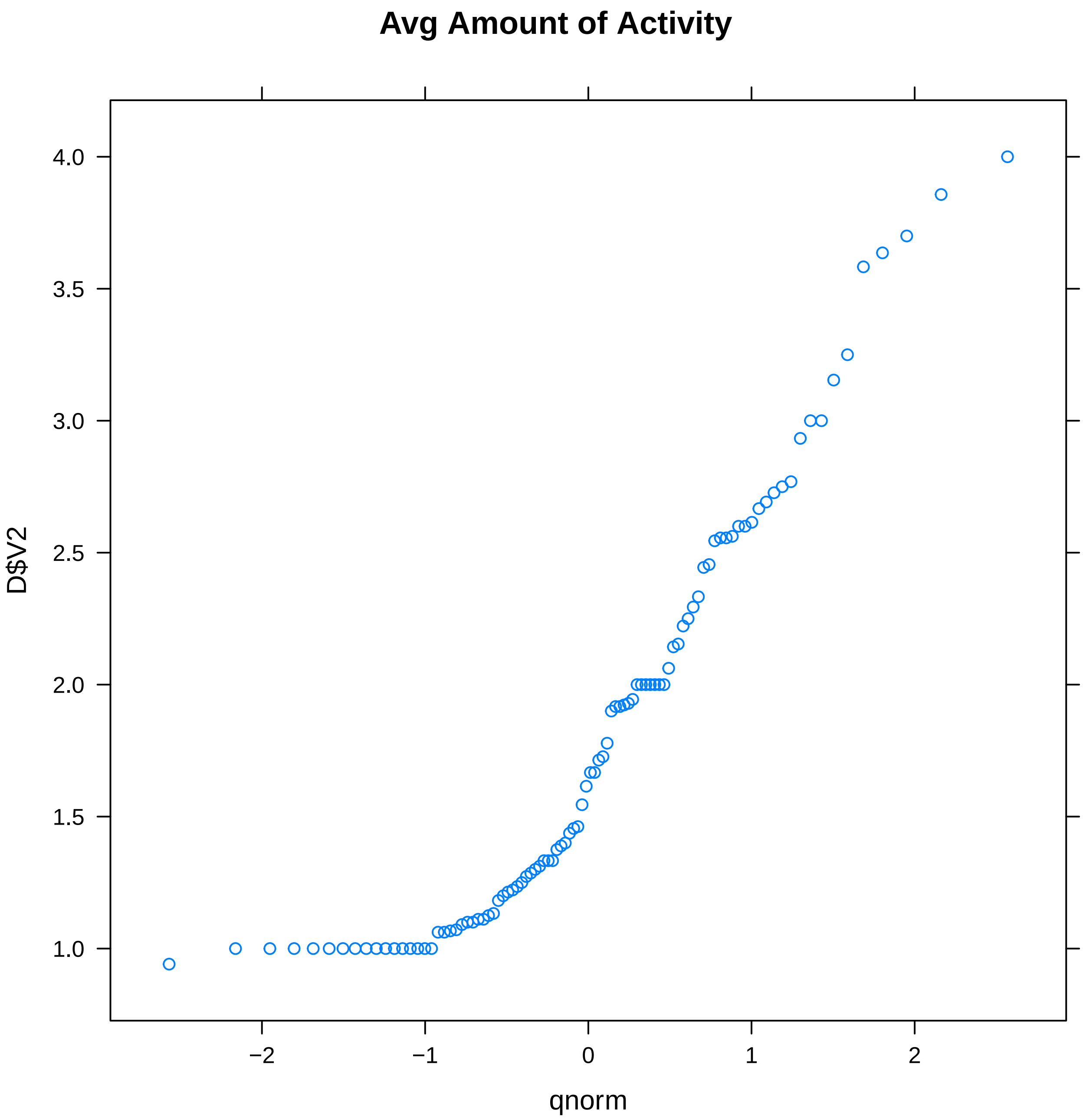


0.1.2 Average Amount of Activity

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.941   1.111   1.641   1.809   2.323   4.000
```

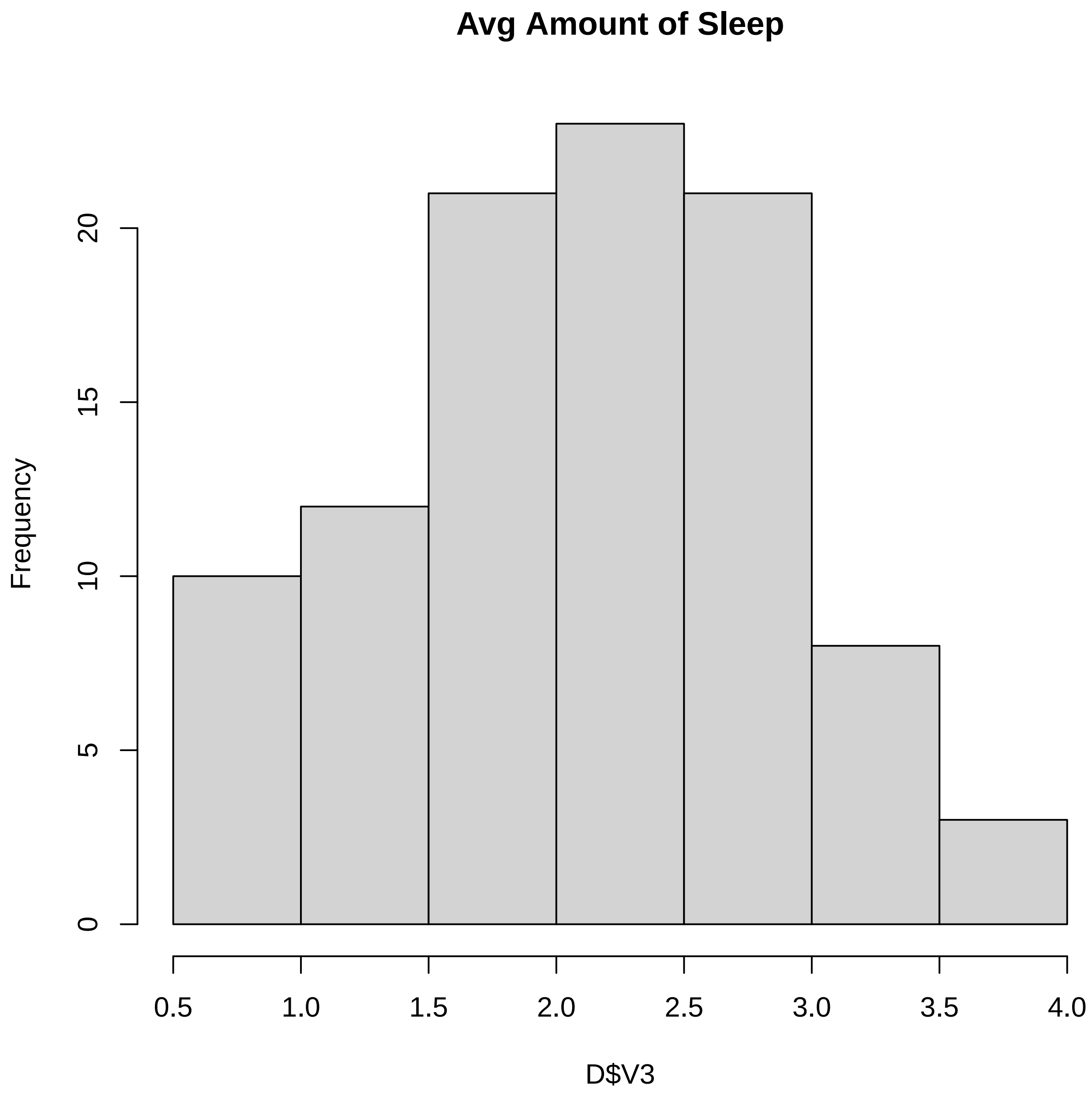

Avg Amount of Activity
Frequency
0
5
10
15
20
25
30
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
D$V2

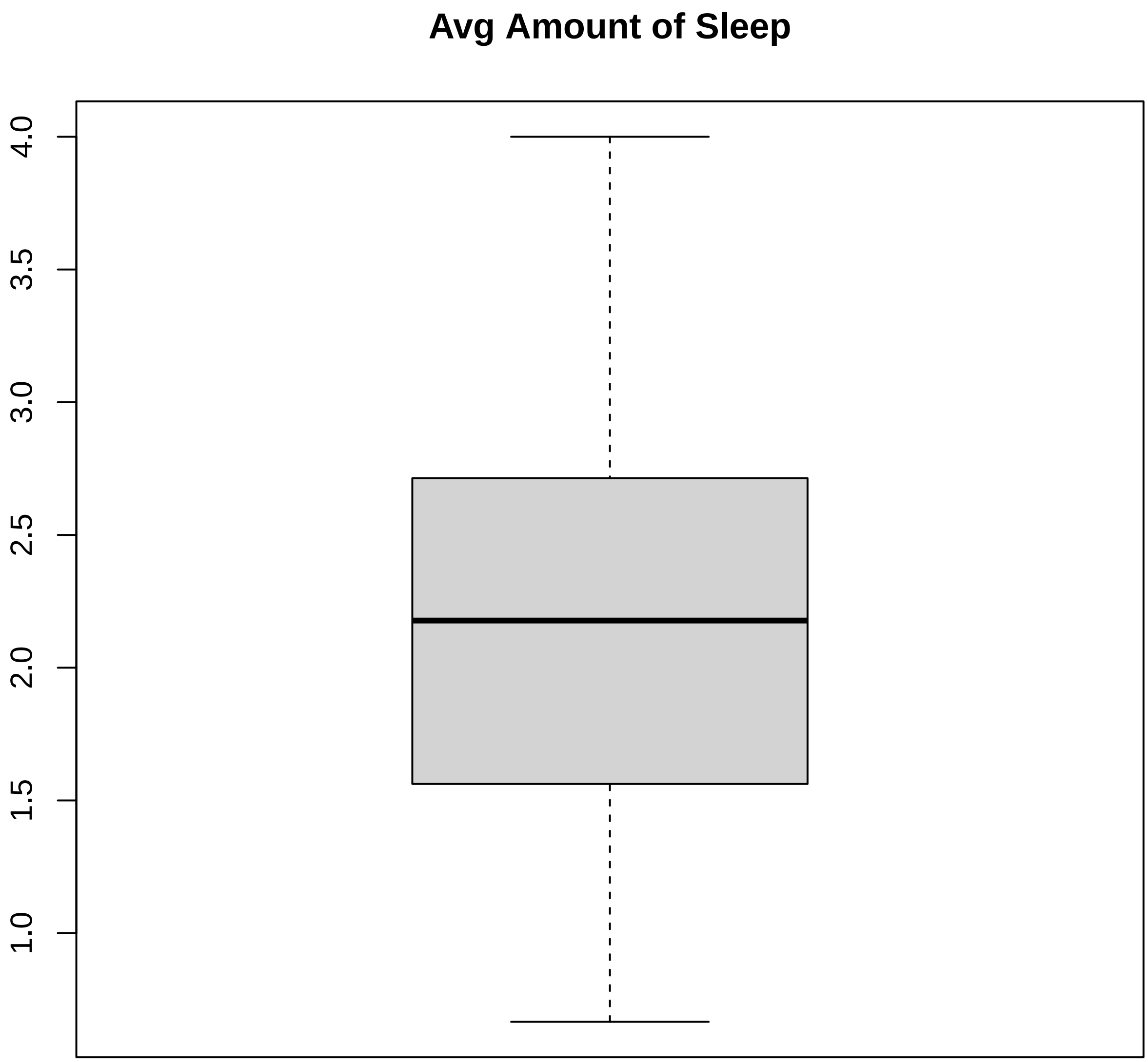

Avg Amount of Activity
1.0
1.5
2.0
2.5
3.0
3.5
4.0

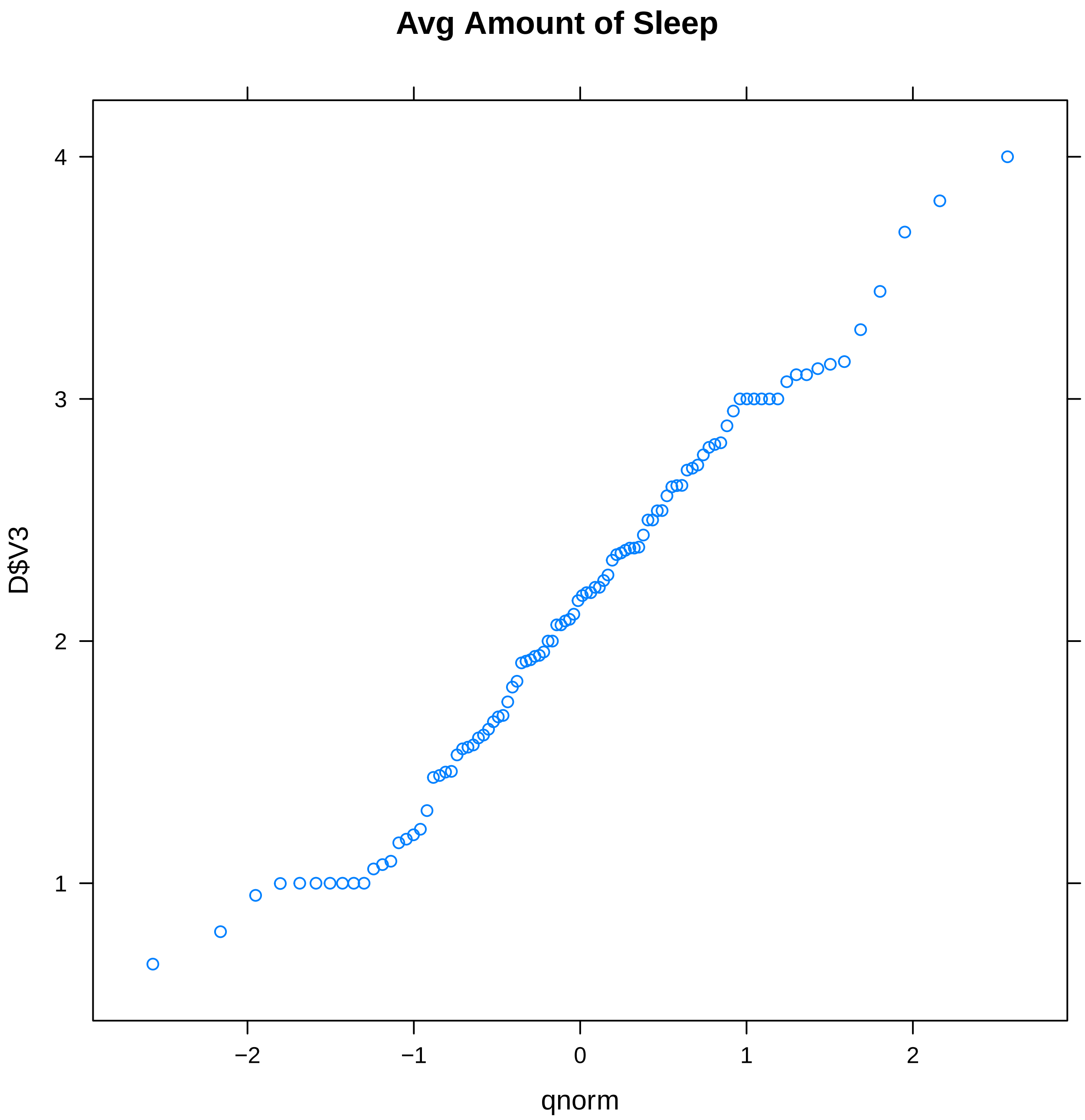


1.1.3 Average Amount of Sleep

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.666   1.564   2.178   2.138   2.712   4.000
```

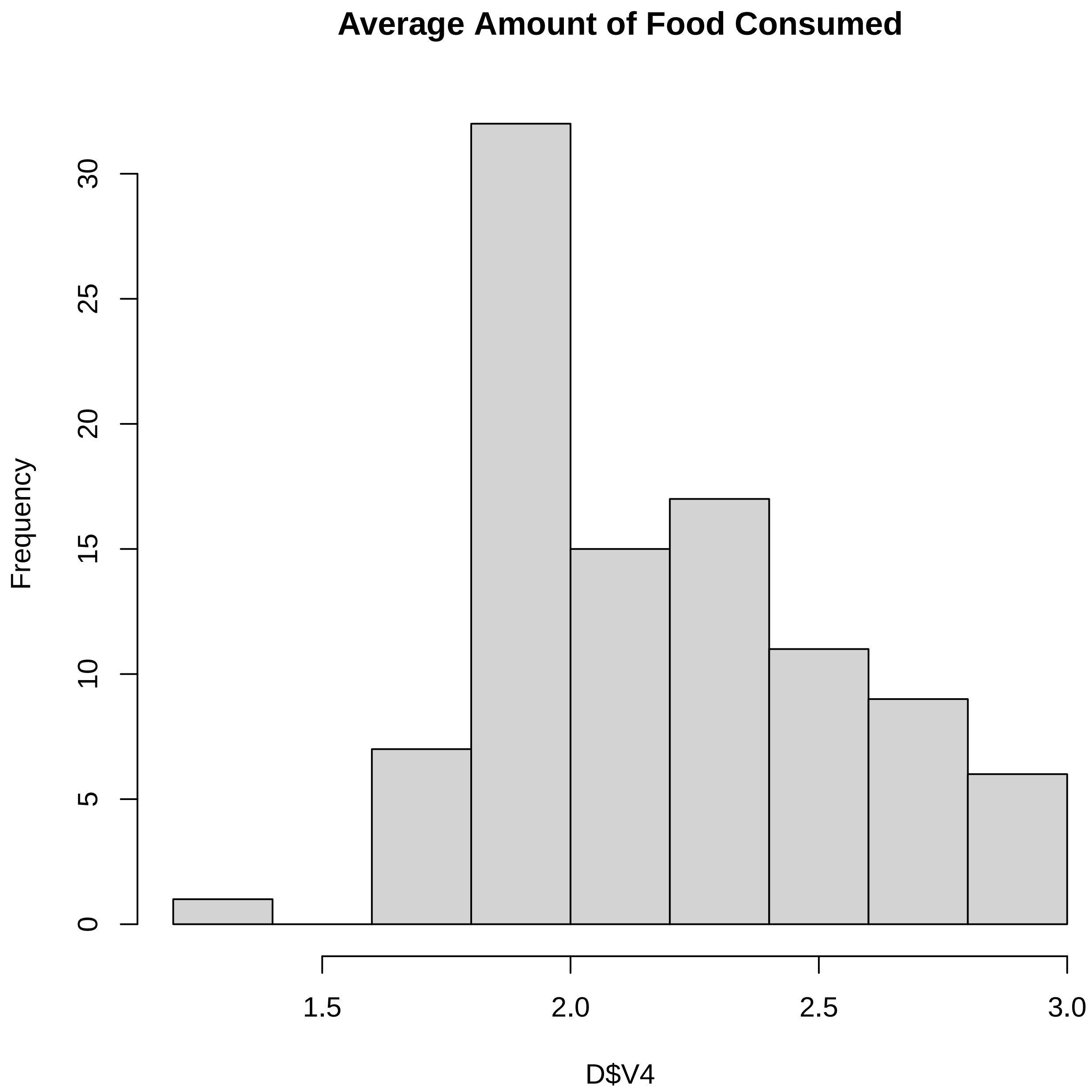

Avg Amount of Sleep
Frequency
0
5
10
15
20
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
D$V3

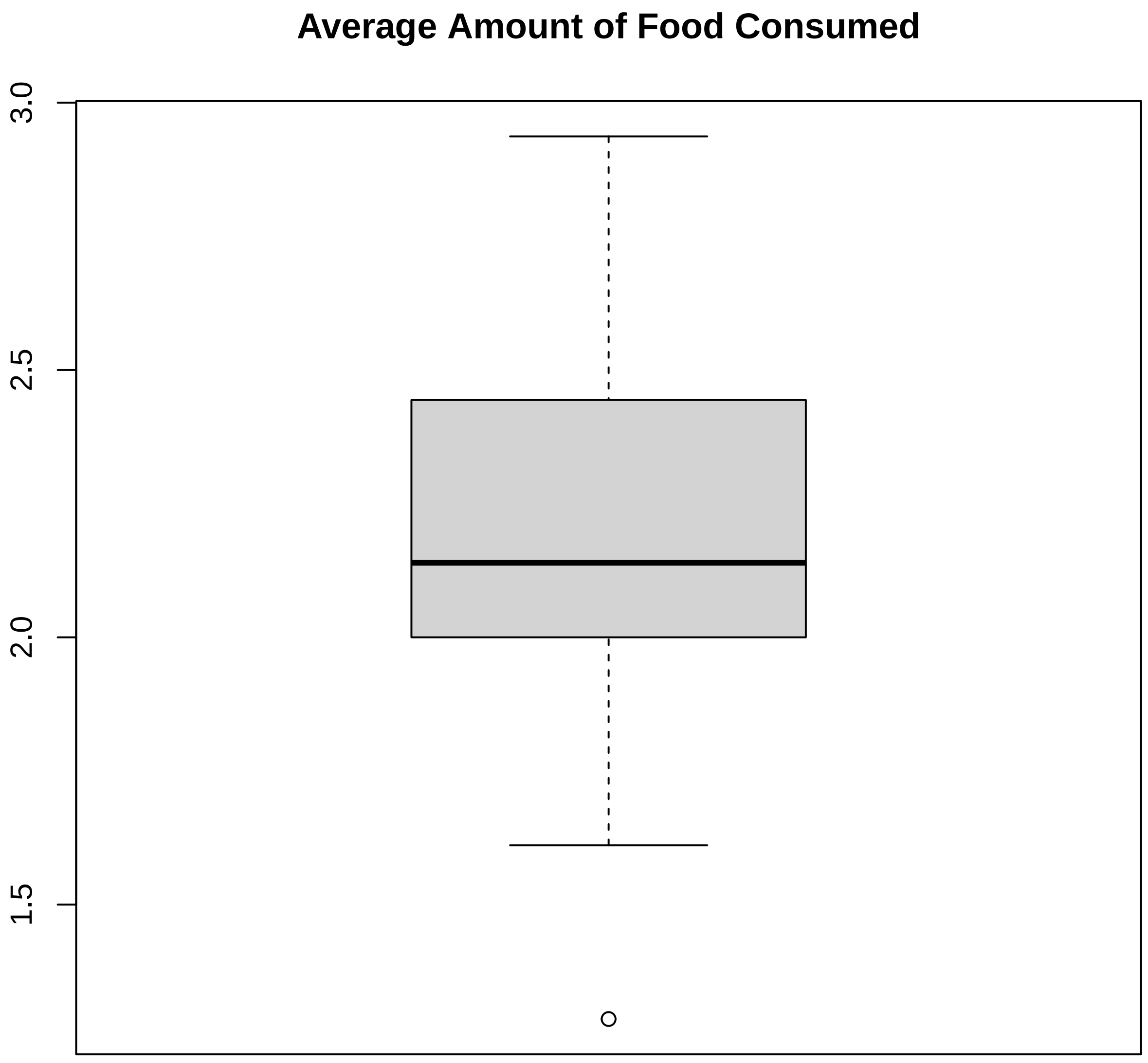

Avg Amount of Sleep
4.0
3.5
3.0
2.5
2.0
1.5
1.0

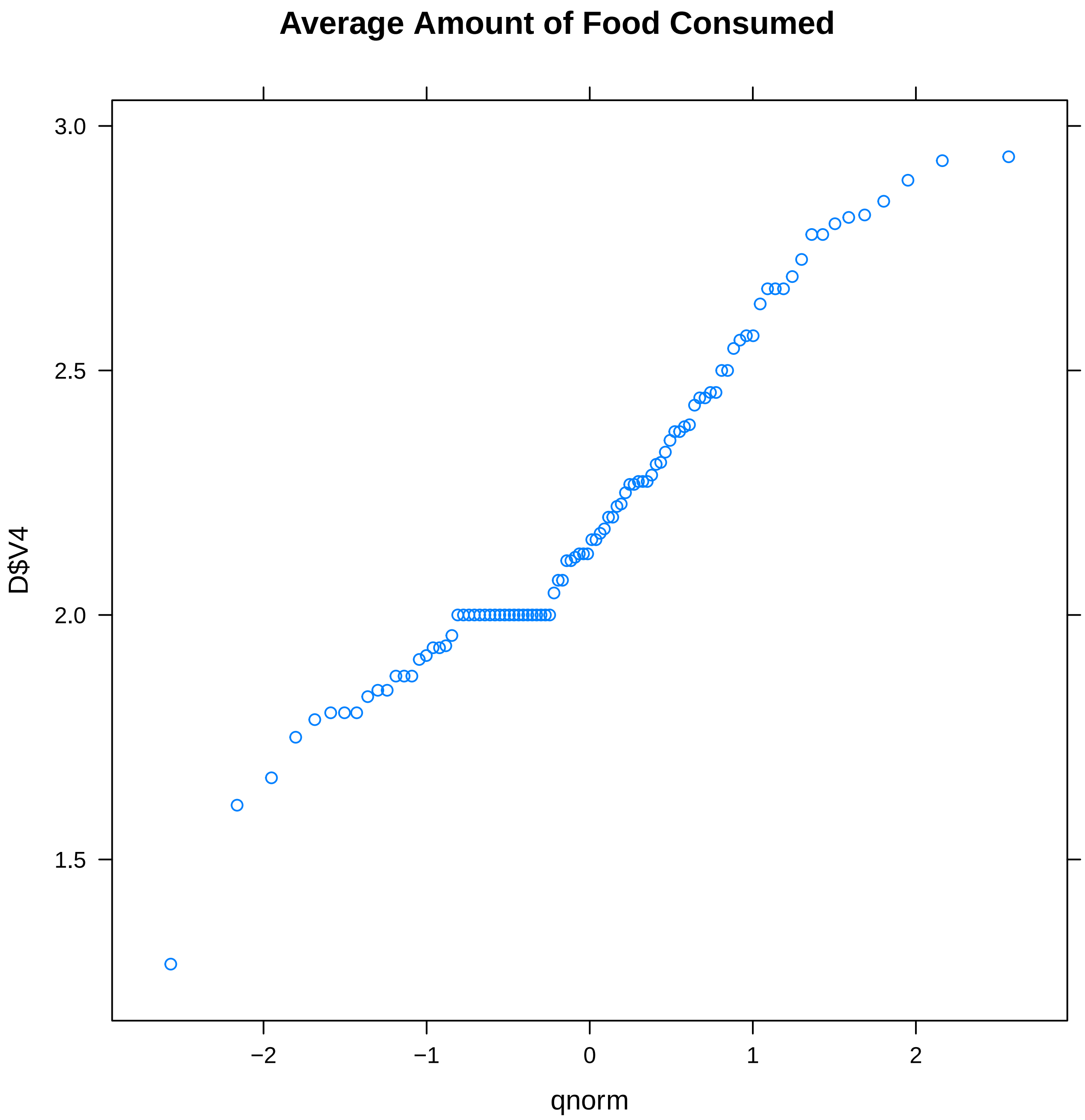


1.1.4 Average Amount of Food Consumed

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.286   2.000   2.139   2.209   2.440   2.937
```

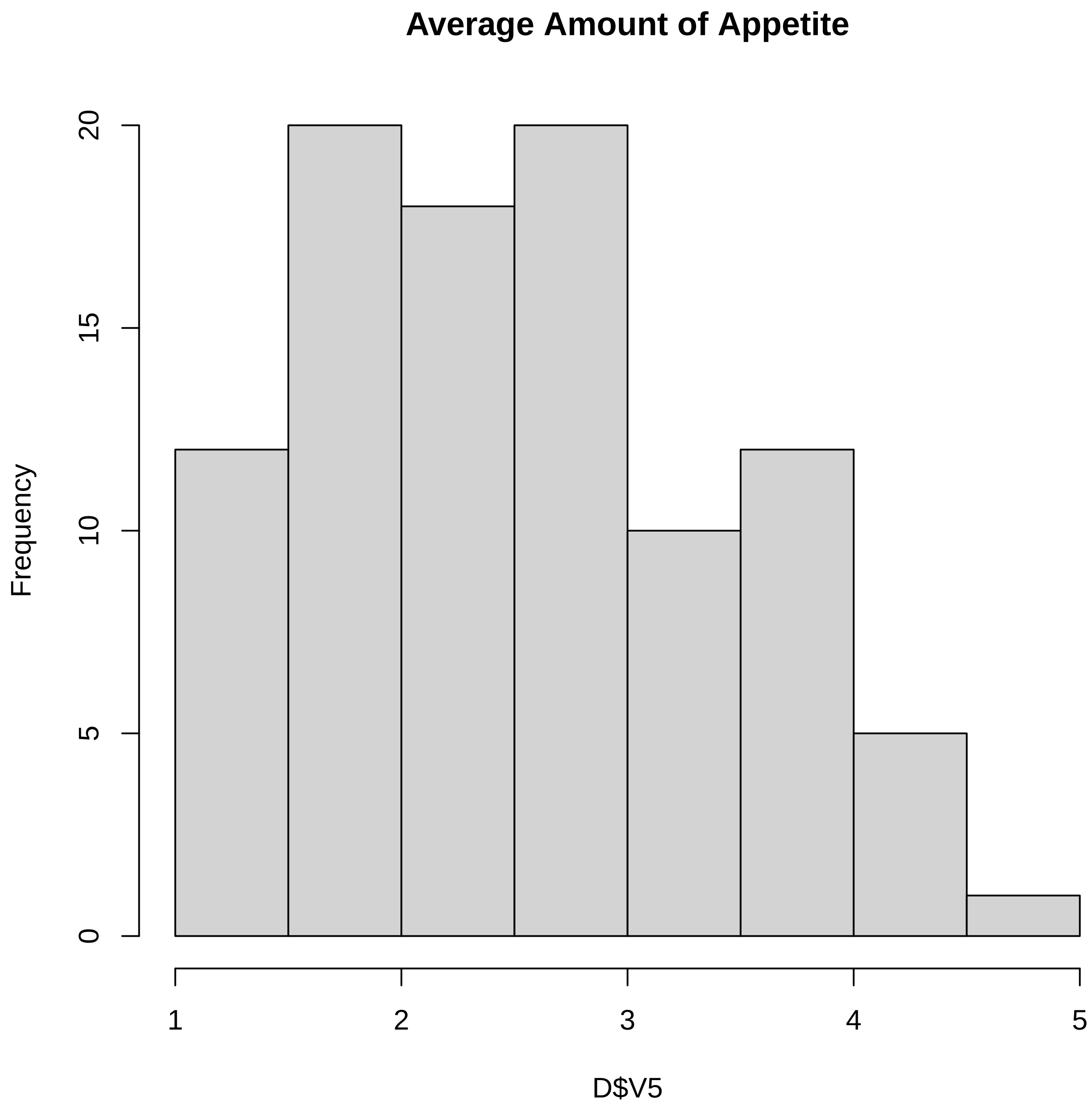

Average Amount of Food Consumed
Frequency
0
5
10
15
20
25
30
1.5
2.0
2.5
3.0
D$V4

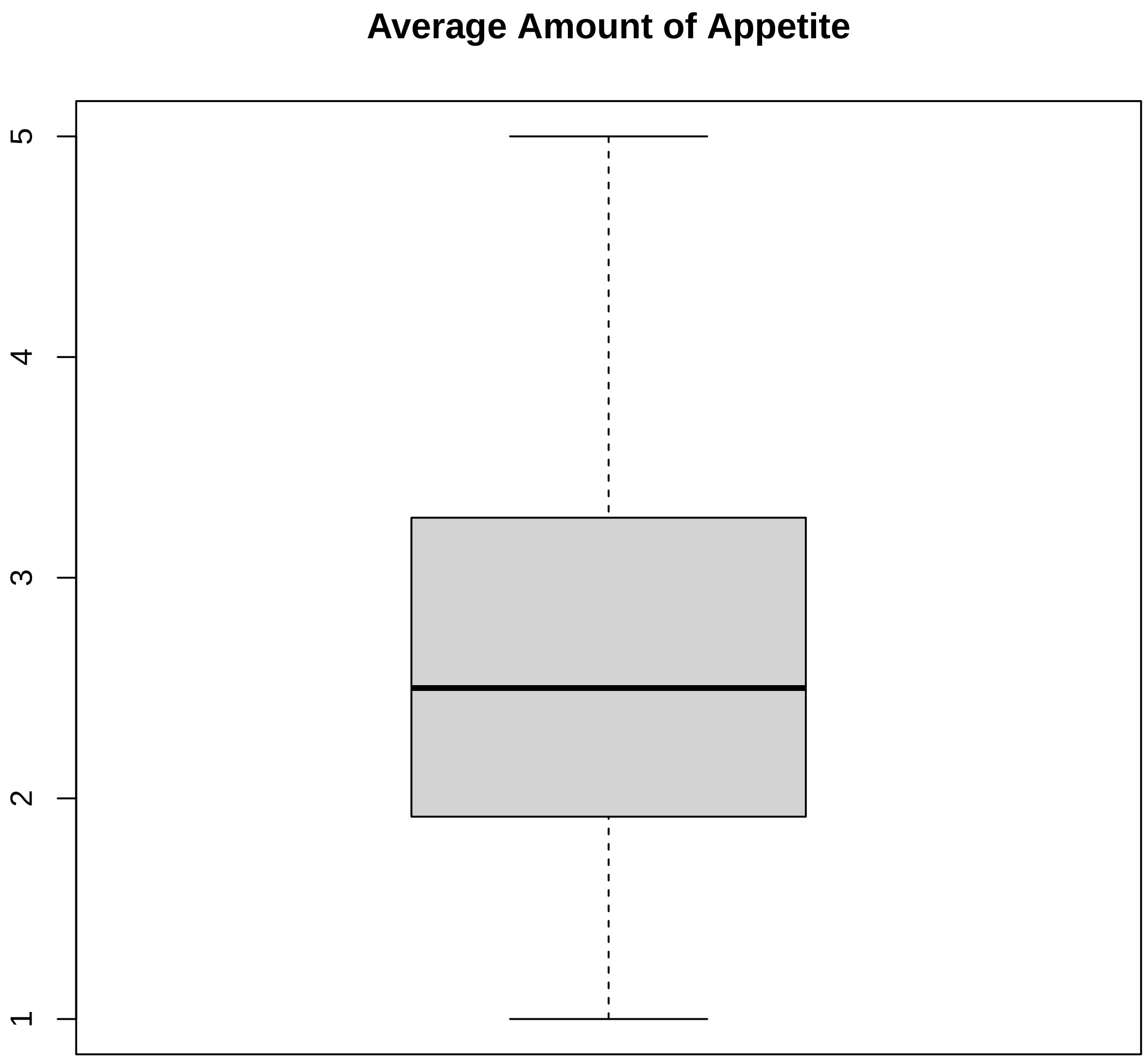

Average Amount of Food Consumed
3.0
2.5
2.0
1.5

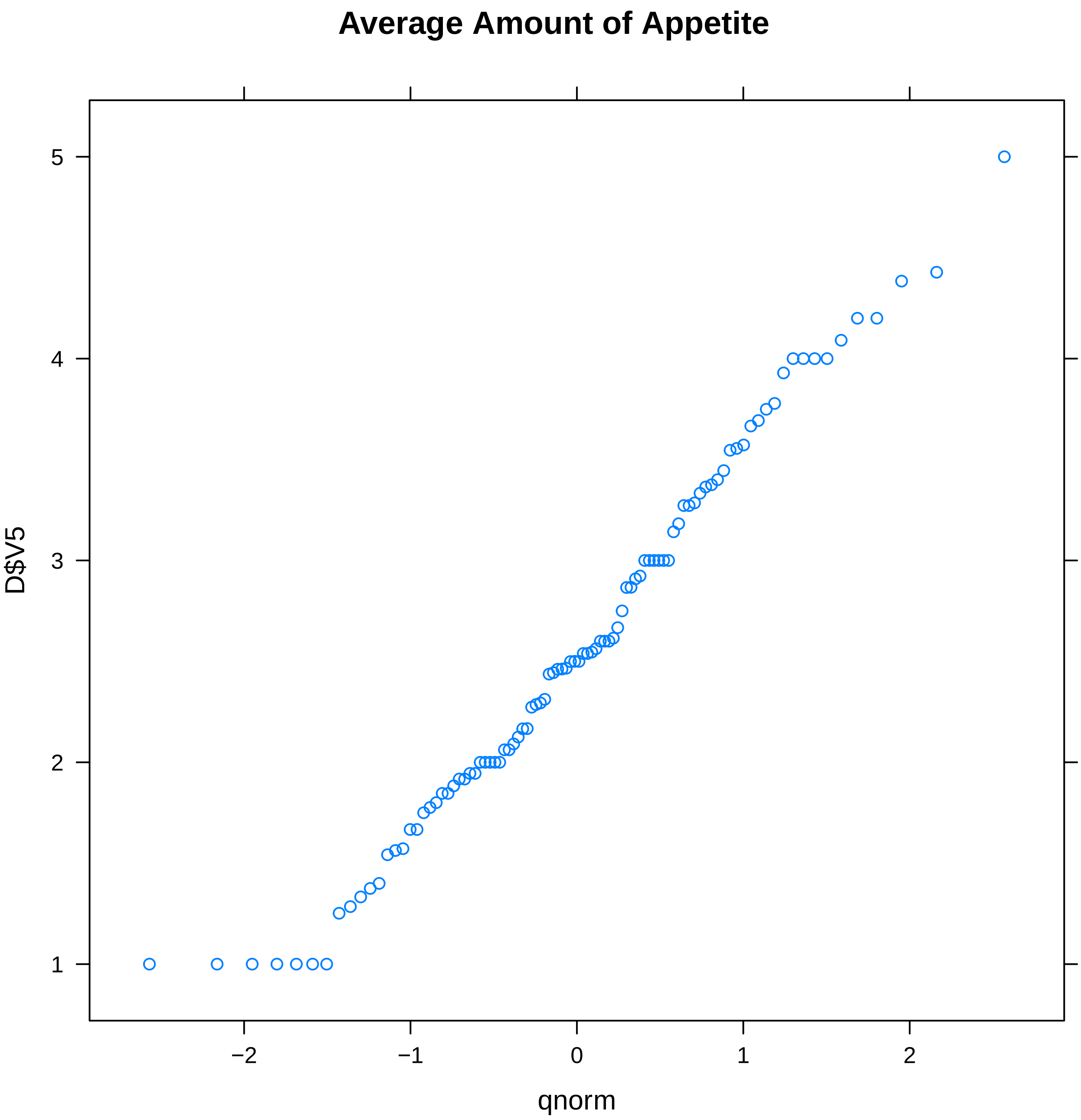


1.1.5 Average Amount of Appetite

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.924   2.500   2.575   3.272   5.000
```


Average Amount of Appetite
Frequency
0
5
10
15
20
1
2
3
4
5
D$V5


Average Amount of Appetite
5
4
3
2
1



- Observations :

1. V1 ,V2 and V4 are positively skewed. V2 is particular is more positively skewed than others. Thse will not (potentially) follow normal distribution. Square root transformation is an option.

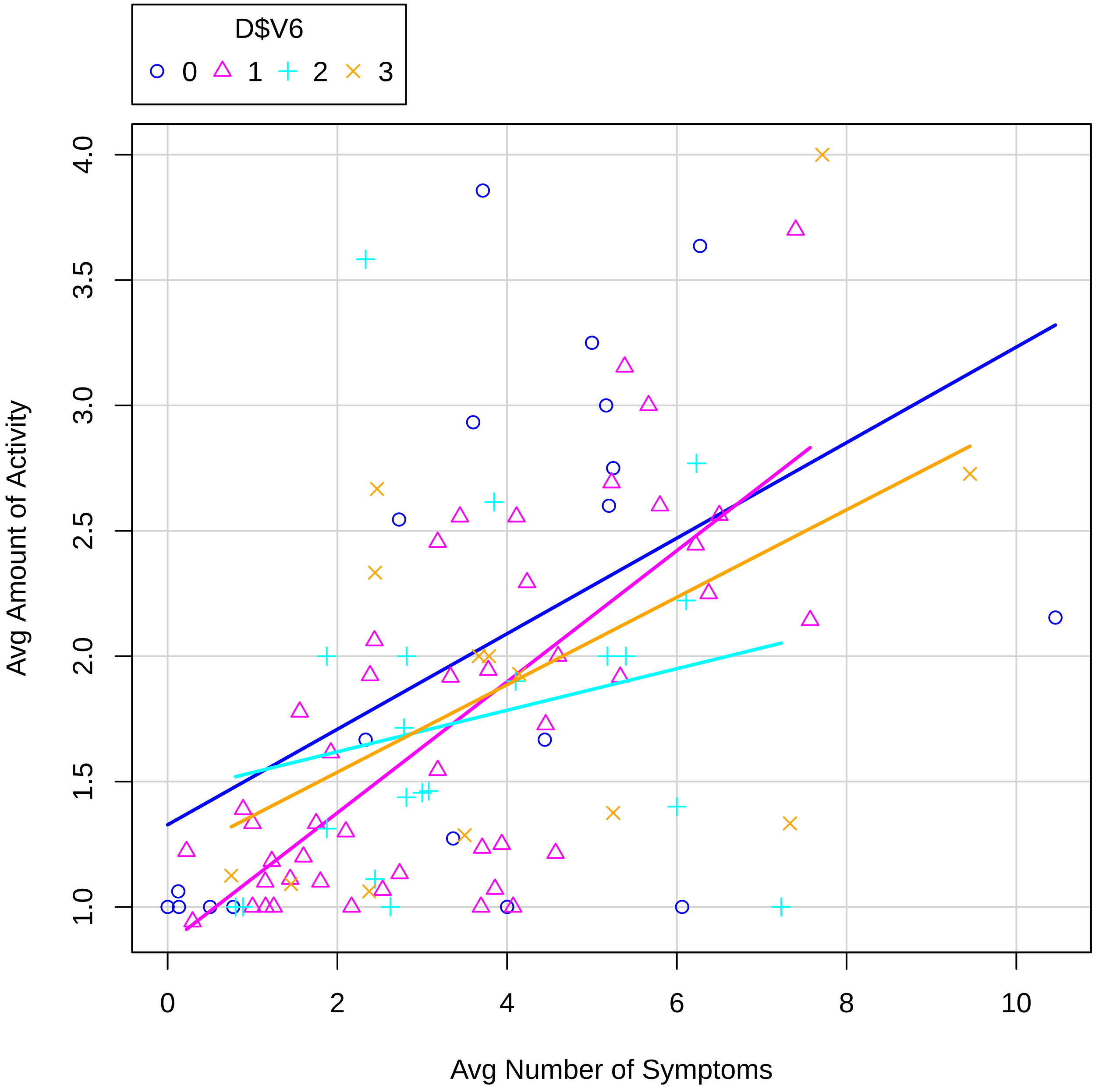
2. There aren't any prominent outliers in the data.

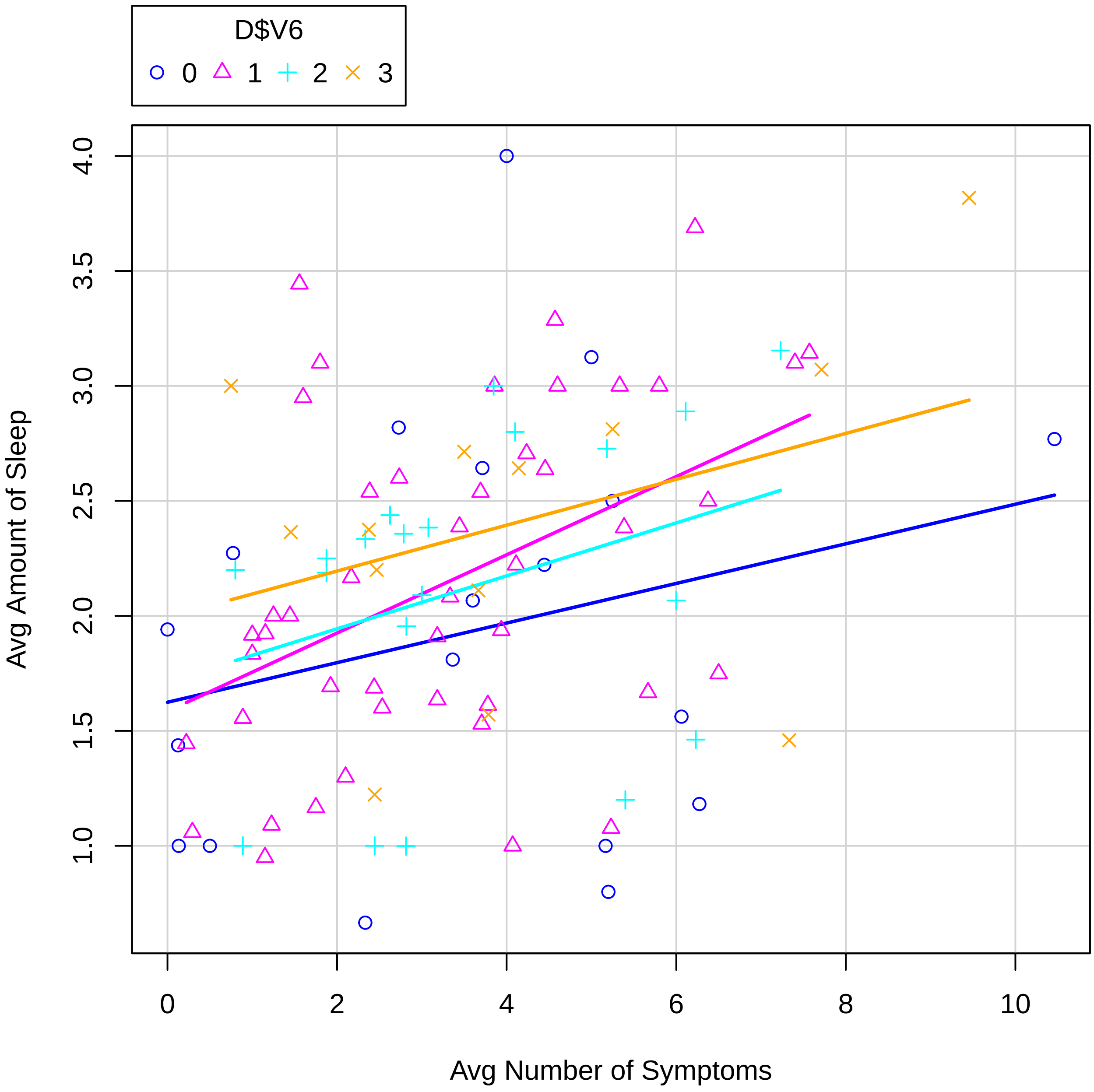
3. V3 and V5 seem to follow normal from the qq plot.

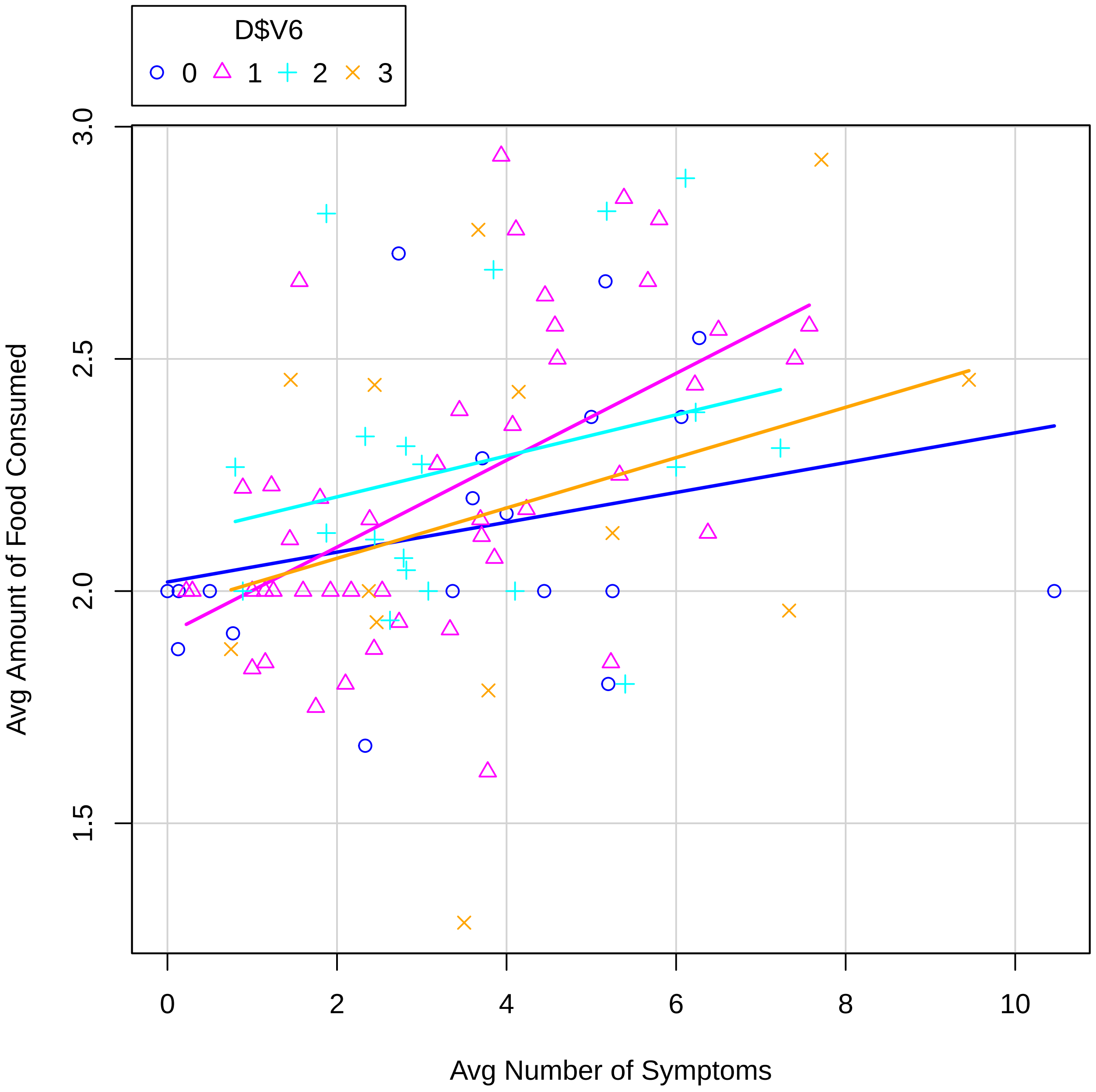
Bivariate EDA

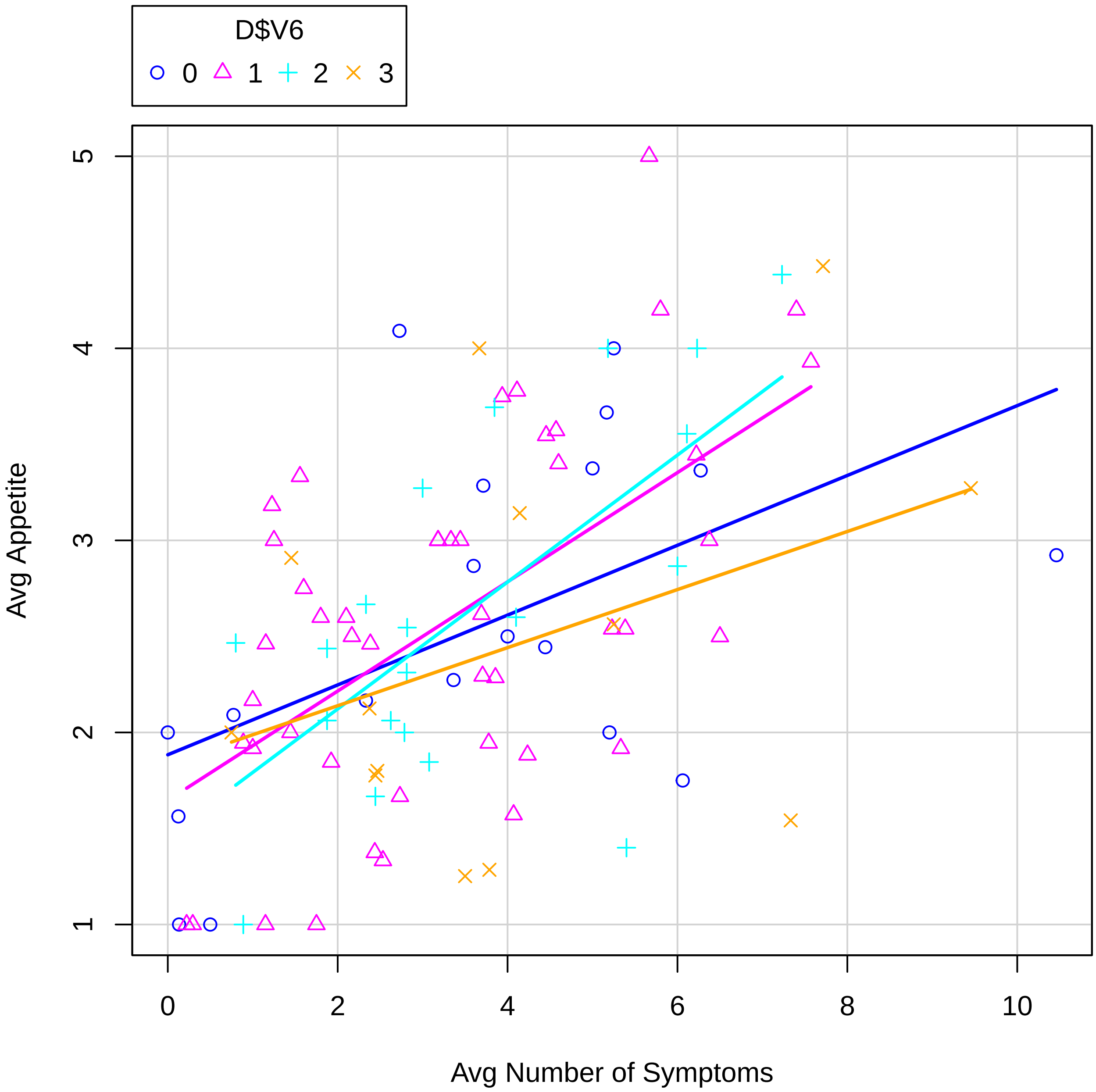
Scatter Plots

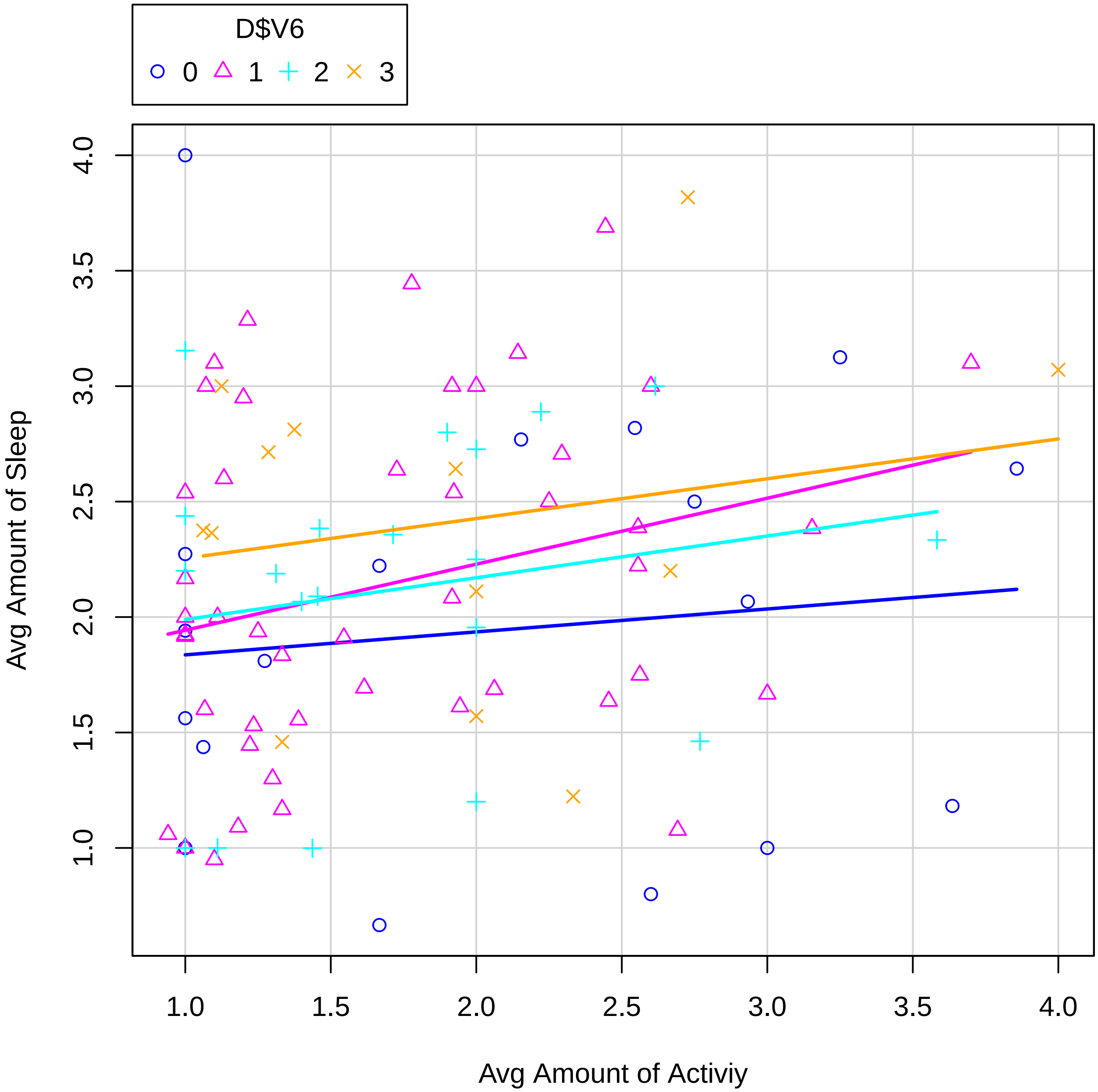
Here we plot the scatterplots, along with the regression lines. As is evident in all the plots, groupwise, the populations are well mixed. They do not seem to come from different populations. The regression lines help understand the general trend.

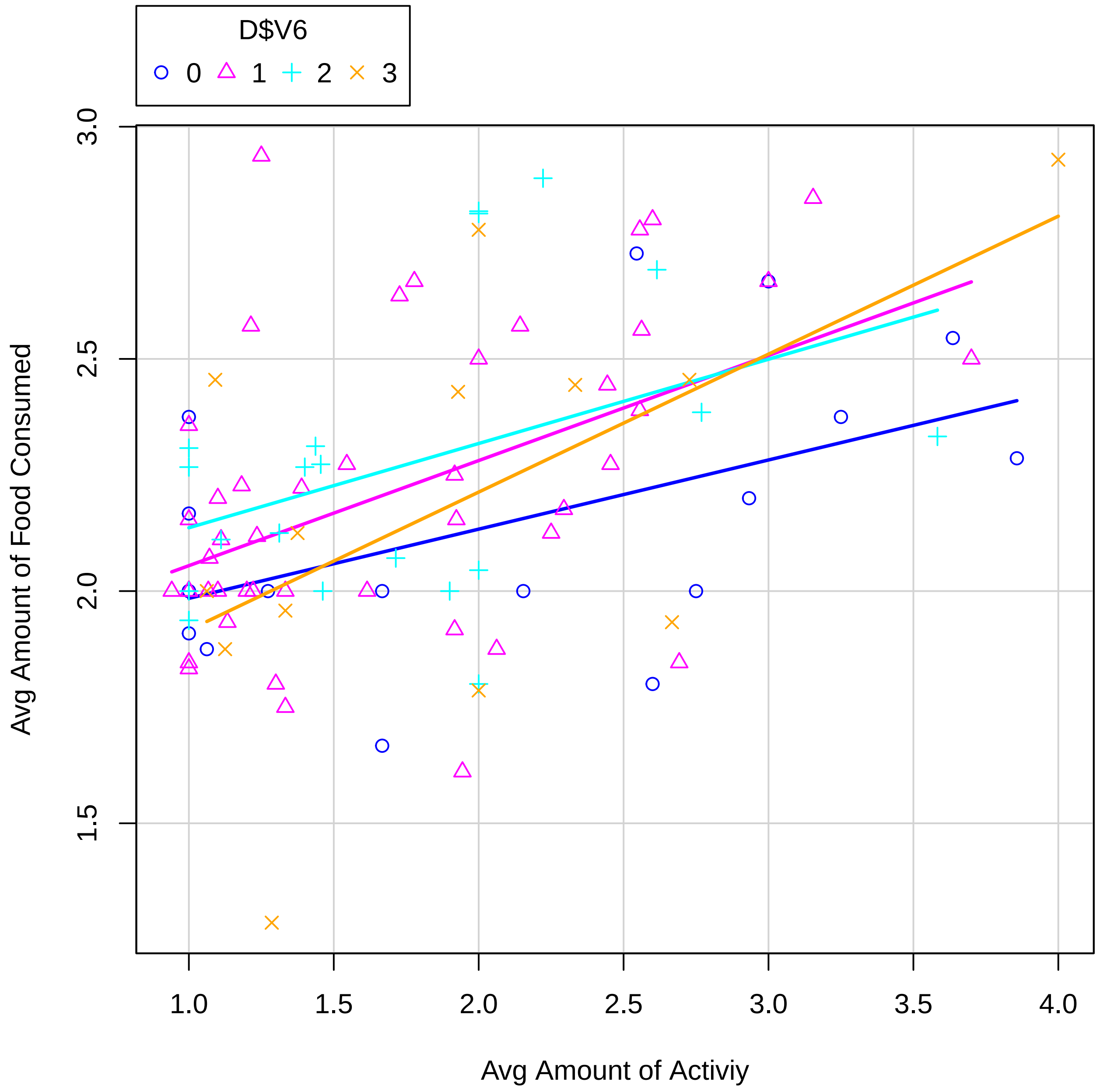


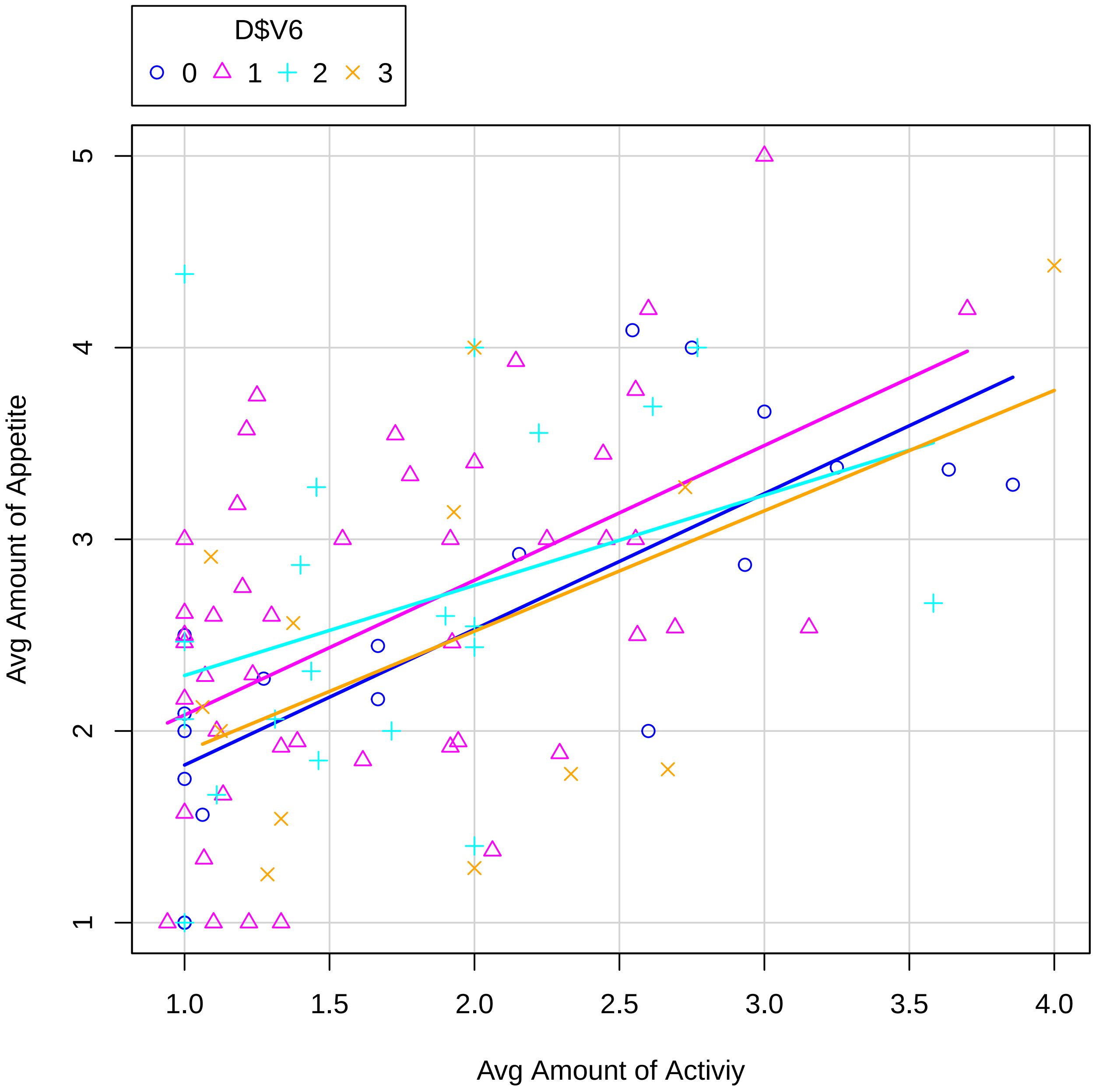

D$V6
○ 0
△ 1
+ 2
× 3
Avg Amount of Sleep
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0
2
4
6
8
10
Avg Number of Symptoms

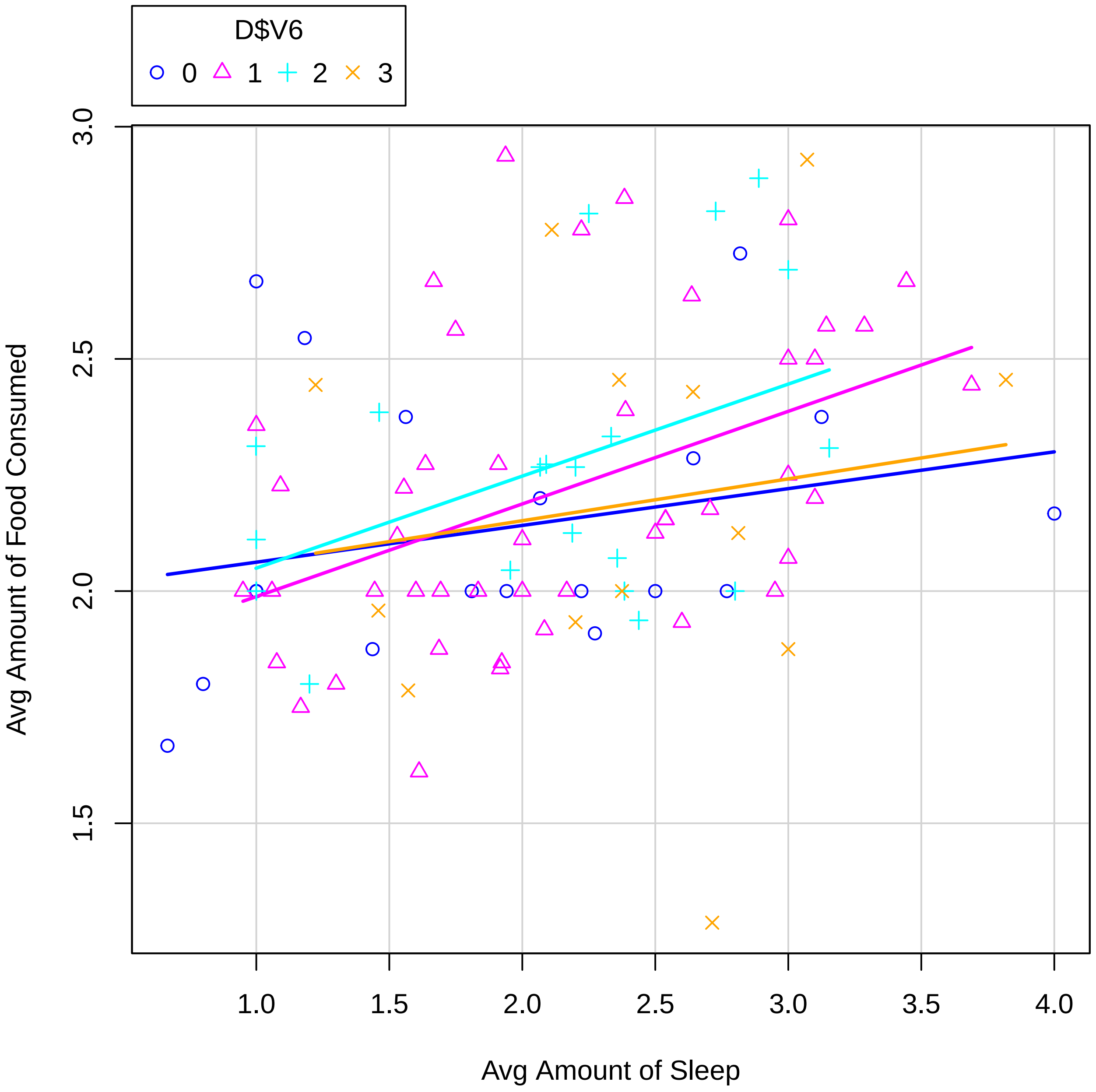

D$V6
○ 0
△ 1
+ 2
× 3
3.0
2.5
2.0
1.5
Avg Amount of Food Consumed
0
2
4
6
8
10
Avg Number of Symptoms

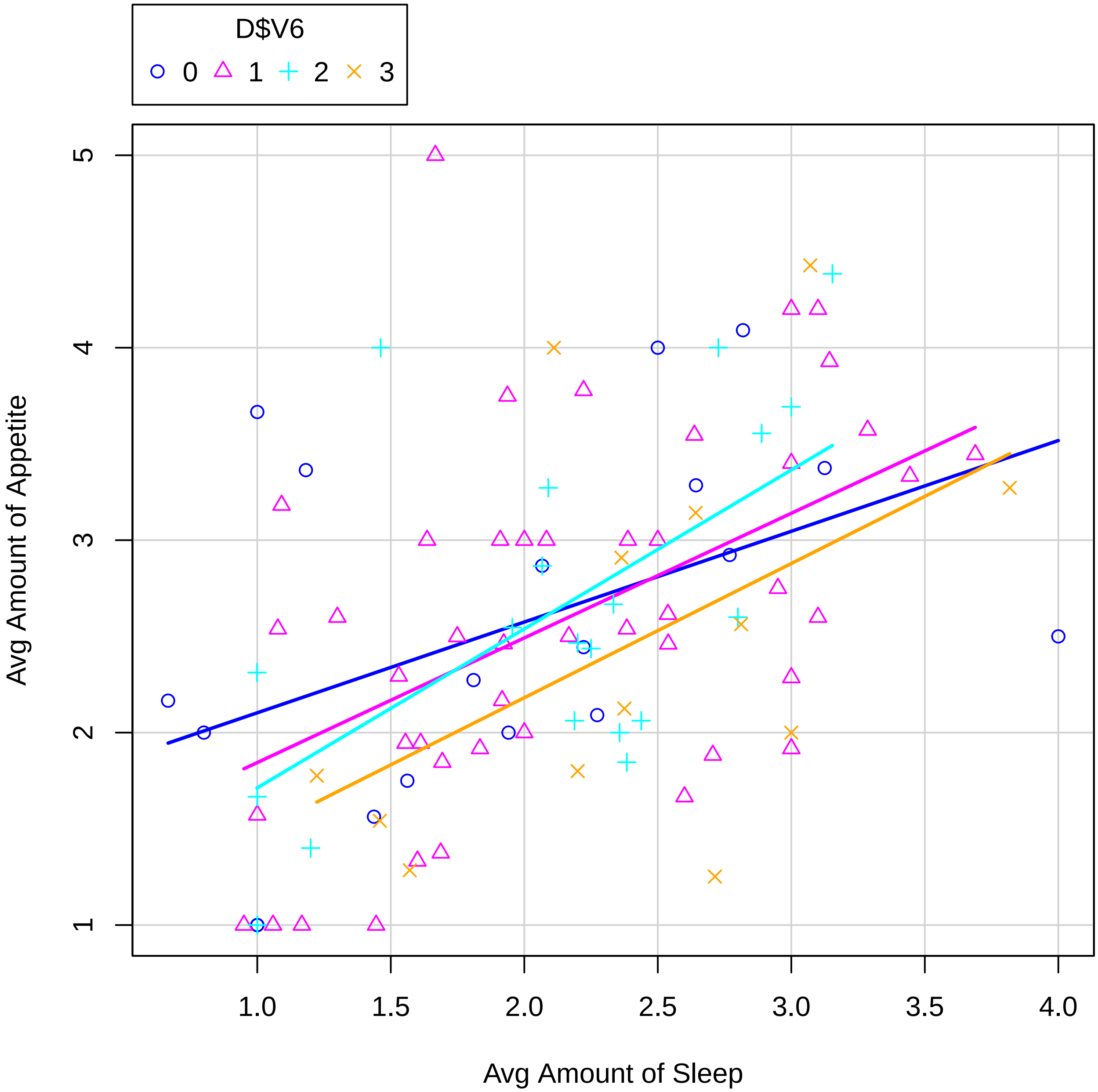

D$V6
○ 0
△ 1
+ 2
× 3
Avg Appetite
1
2
3
4
5
Avg Number of Symptoms
0
2
4
6
8
10

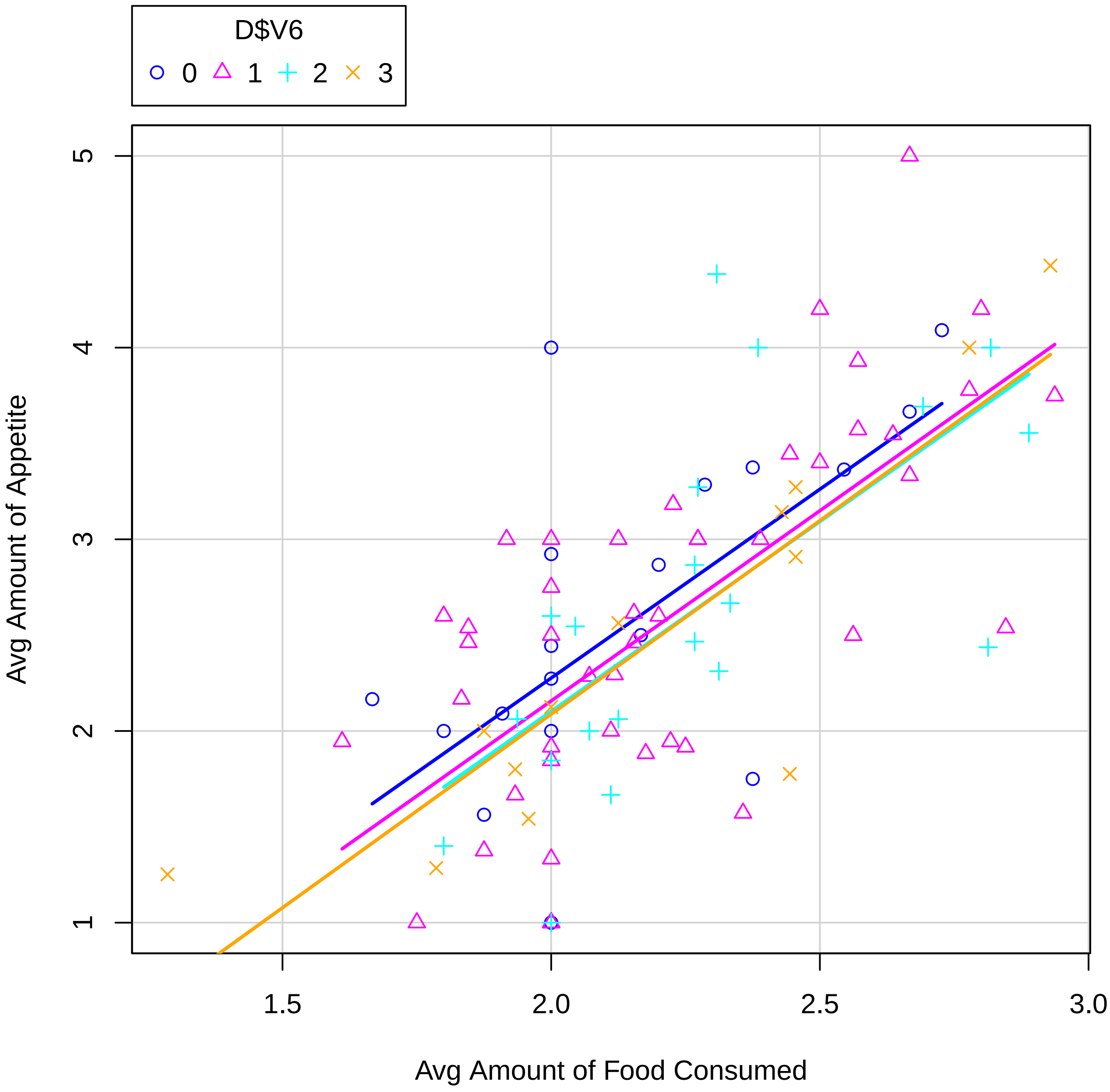

D$V6
○ 0
△ 1
+ 2
× 3
Avg Amount of Sleep
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Avg Amount of Activiy
1.0
1.5
2.0
2.5
3.0
3.5
4.0


D$V6
○ 0
△ 1
+ 2
× 3
Avg Amount of Food Consumed
1.5
2.0
2.5
3.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Avg Amount of Activiy


D$V6
○ 0
△ 1
+ 2
× 3
Avg Amount of Appetite
1
2
3
4
5
Avg Amount of Activiy
1.0
1.5
2.0
2.5
3.0
3.5
4.0

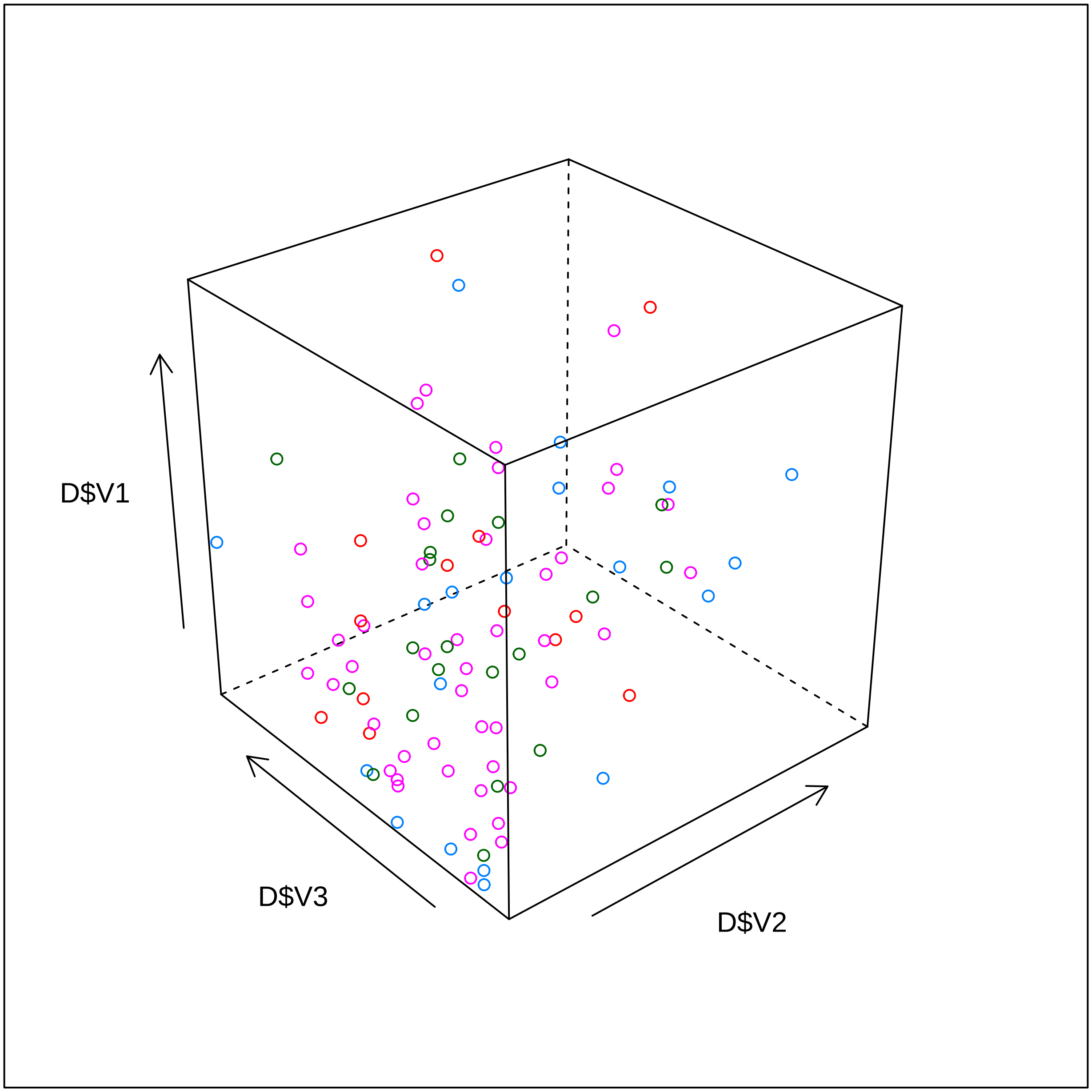

D$V6
○ 0 △ 1 + 2 × 3
Avg Amount of Food Consumed
Avg Amount of Sleep
1.5
2.0
2.5
3.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0

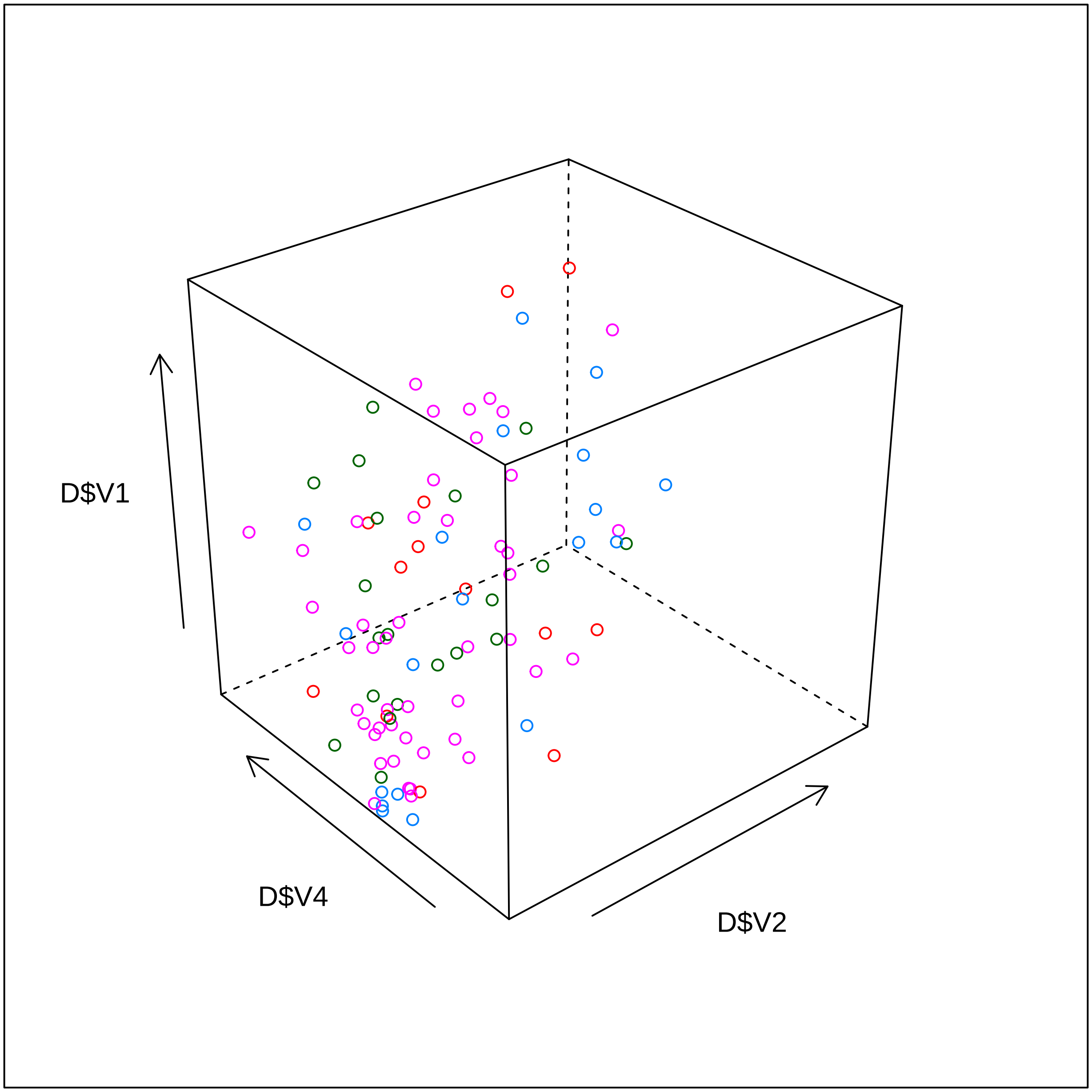

D$V6
○ 0
△ 1
+ 2
× 3
Avg Amount of Appetite
1
2
3
4
5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Avg Amount of Sleep

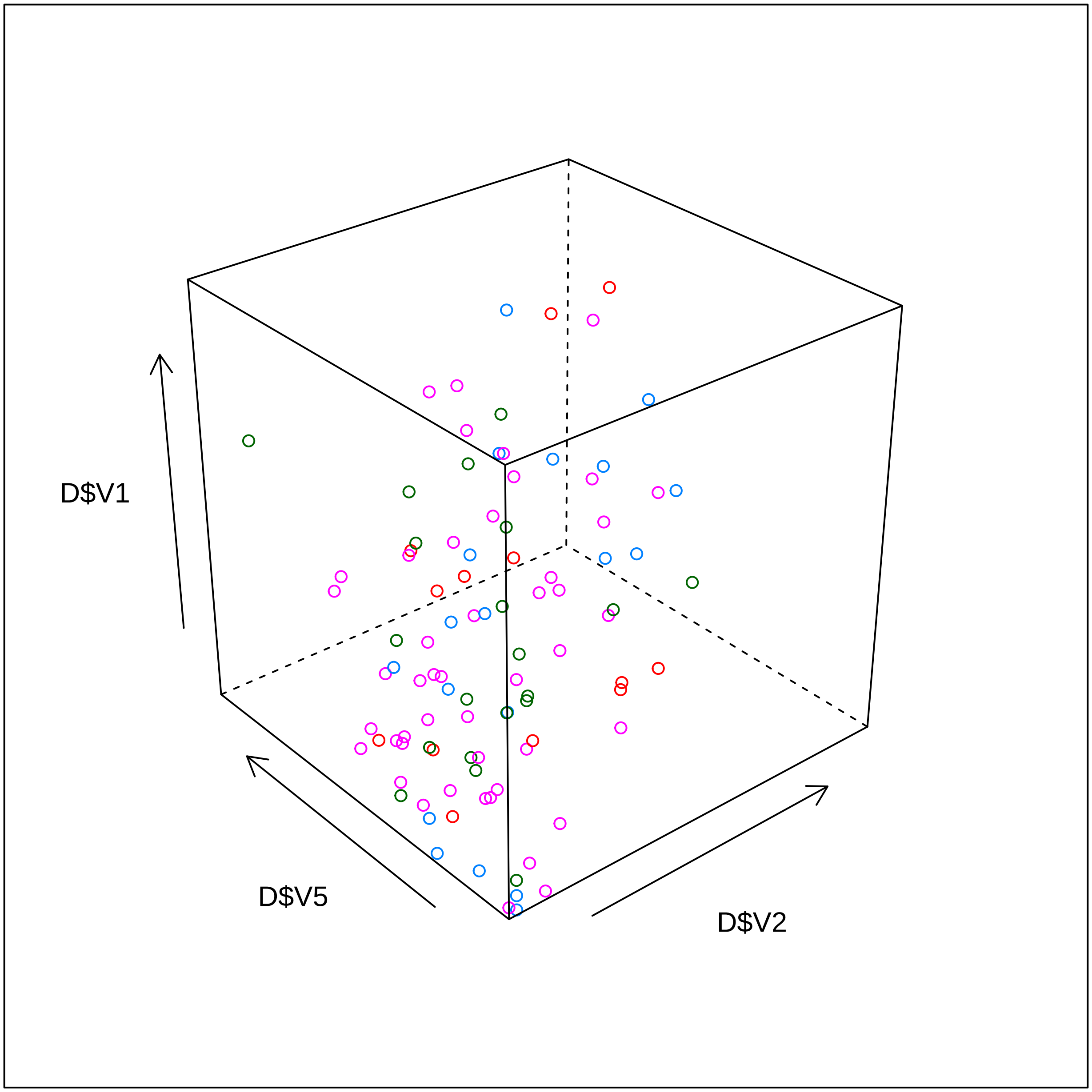


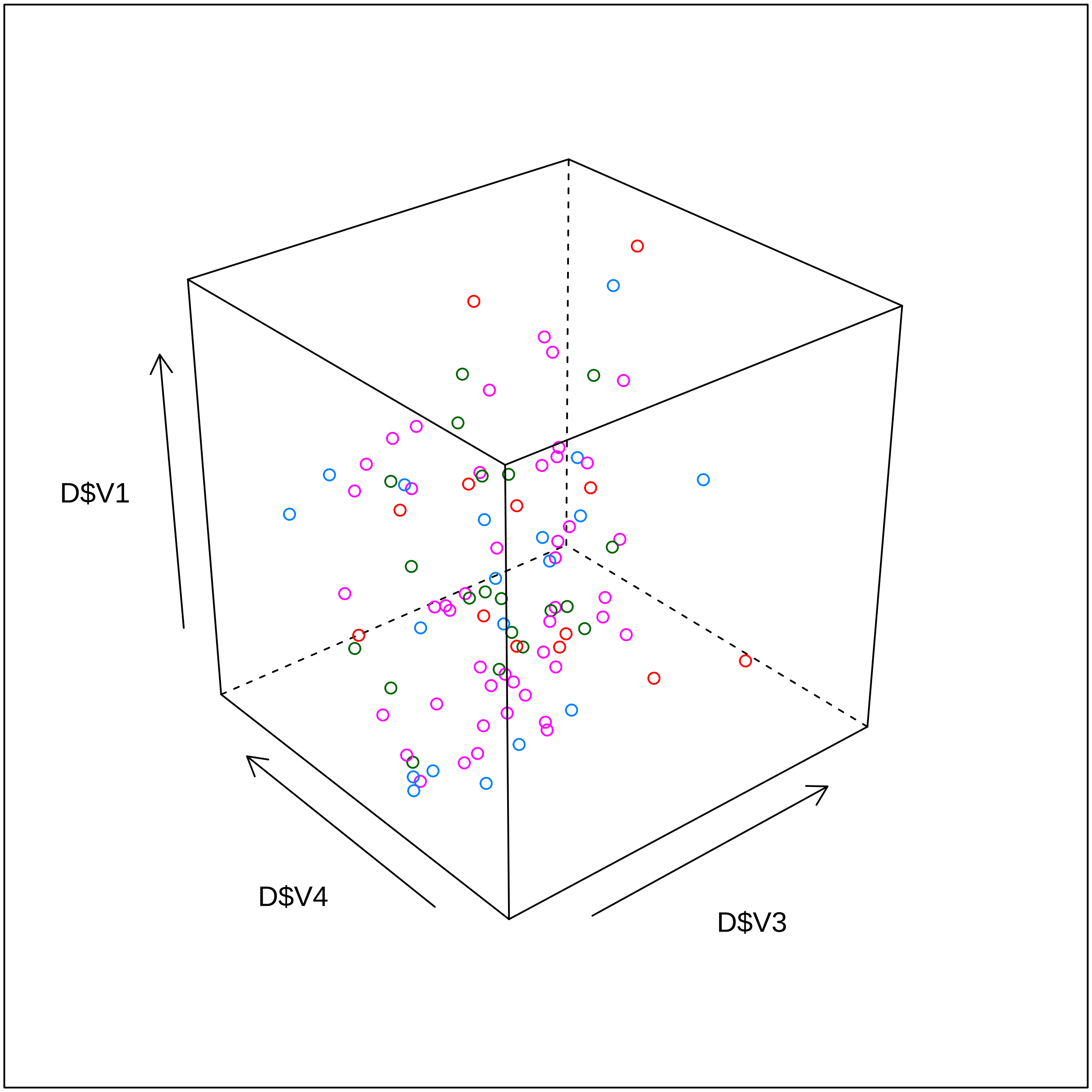
Trivariate EDA

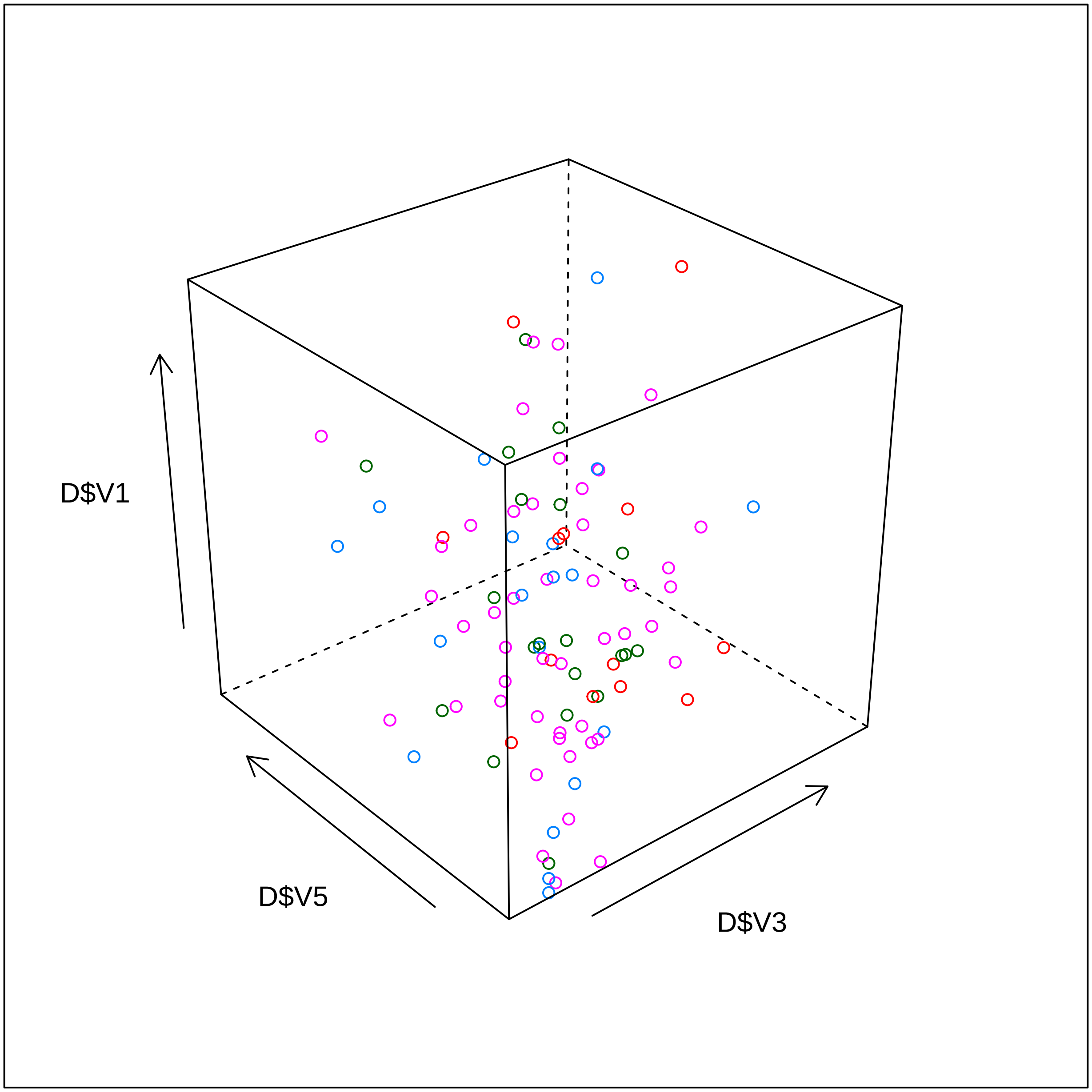
Here we take all 3- variable combinations and plot them. Here also, they seem to be quite well mixed. They do not seem to come from different populations.

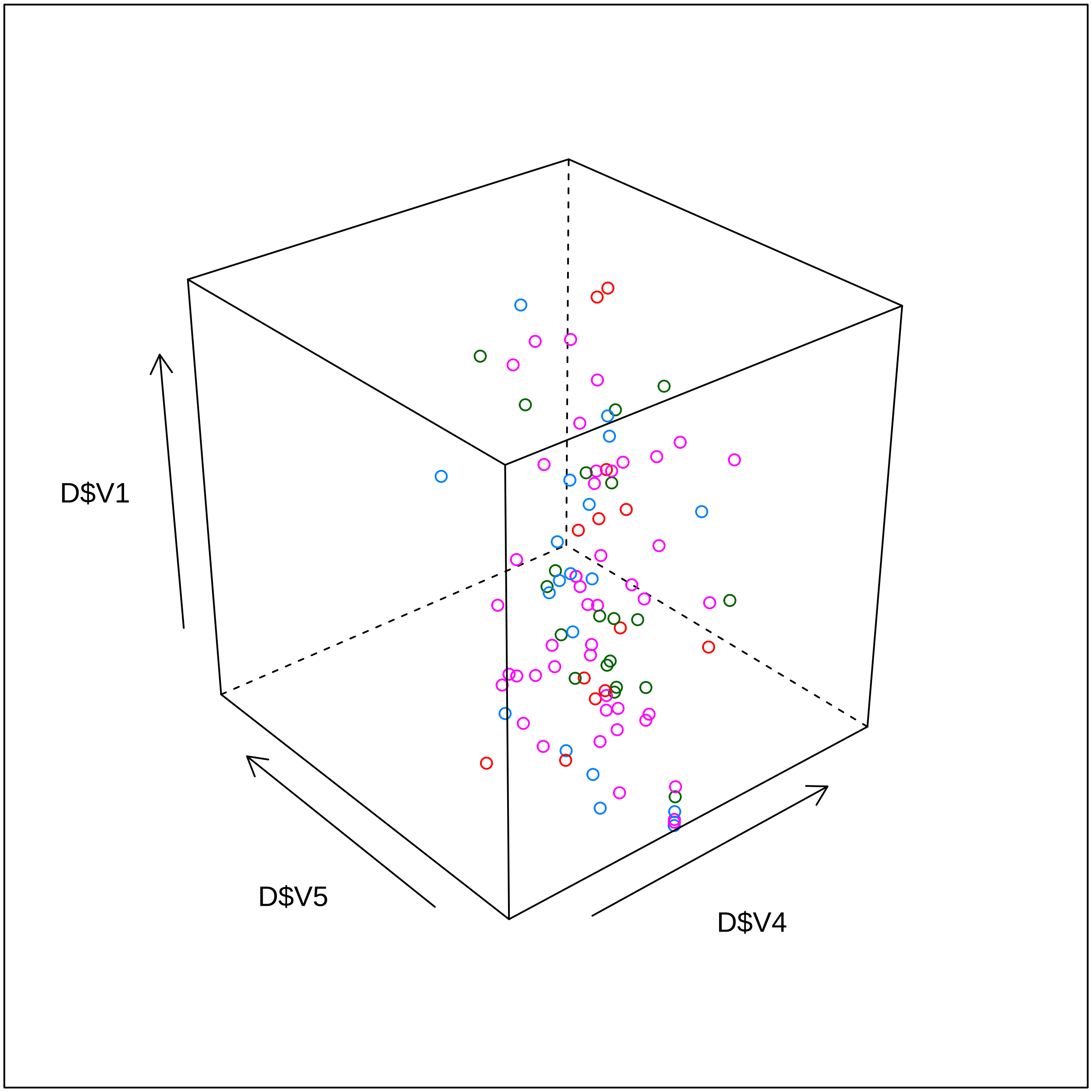

D$V1
D$V3
D$V2

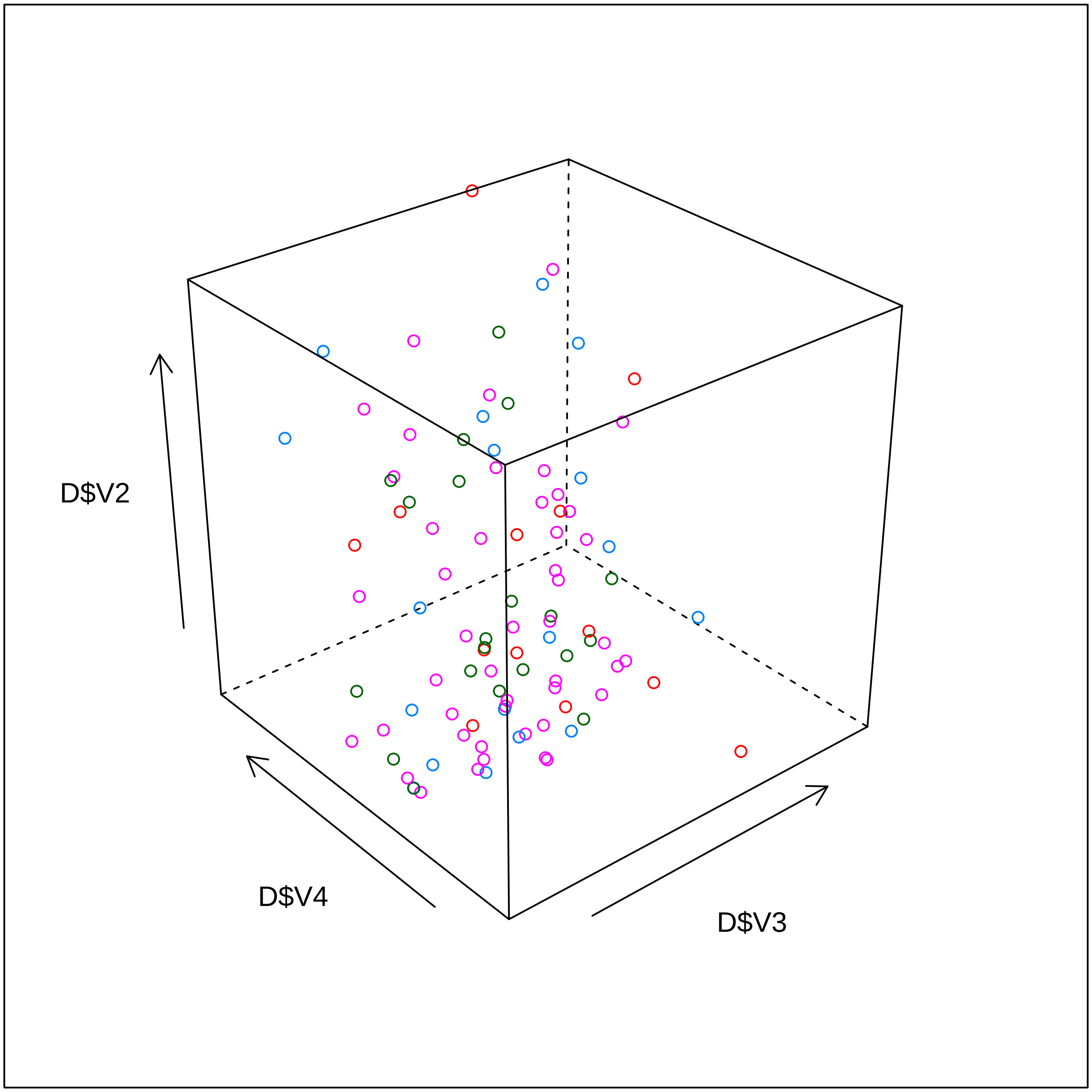

D$V1
D$V4
D$V2

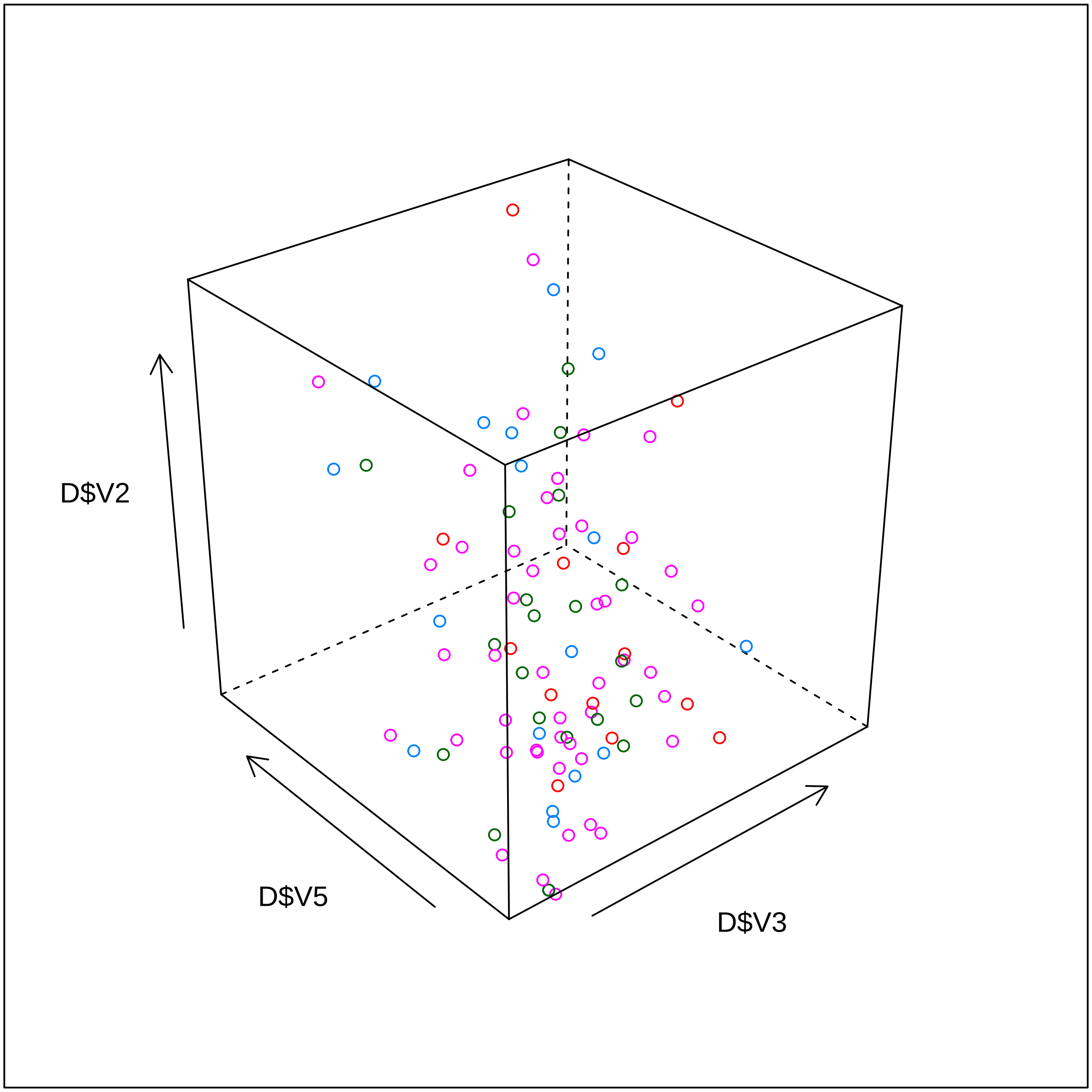

D$V1
D$V5
D$V2

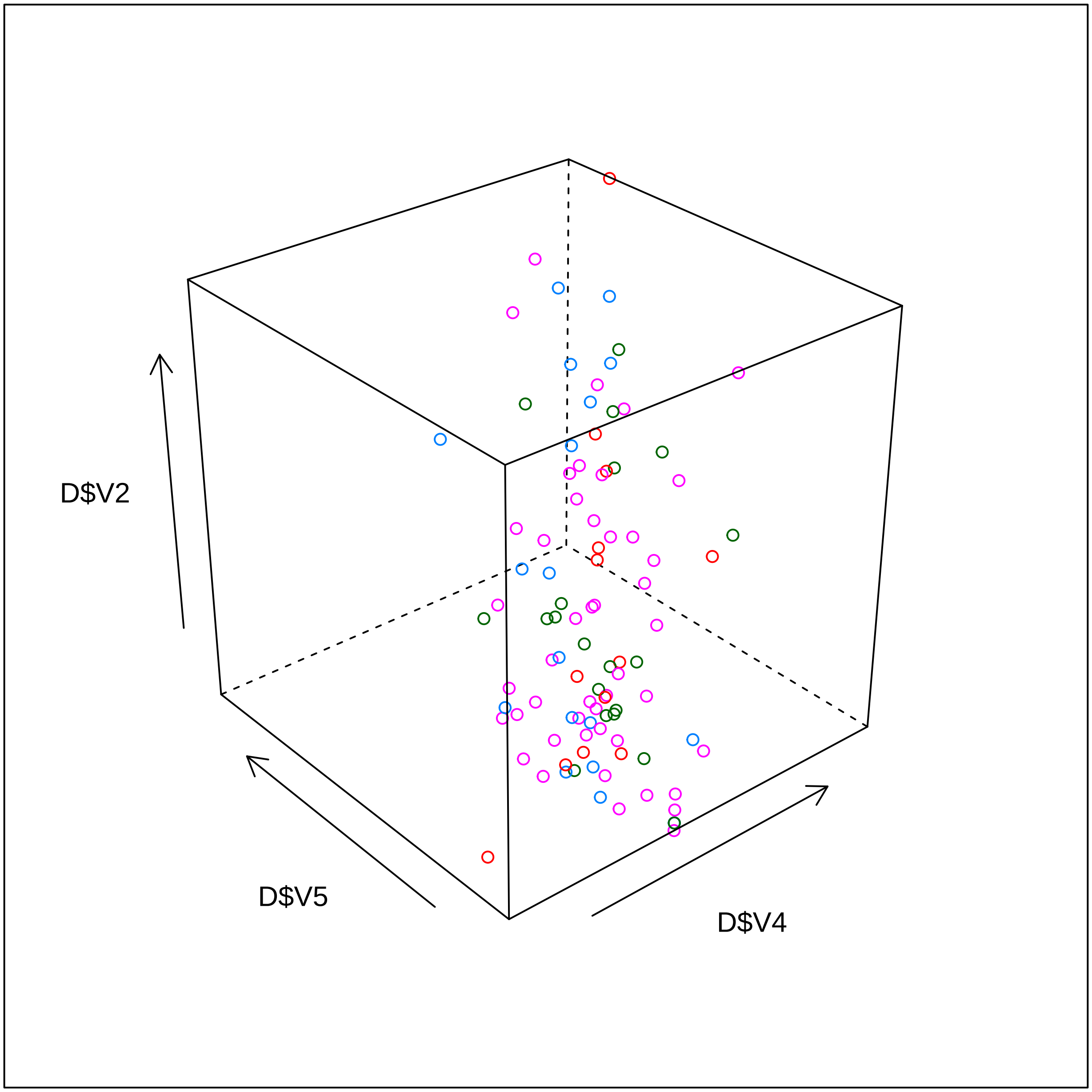

D$V1
D$V4
D$V3

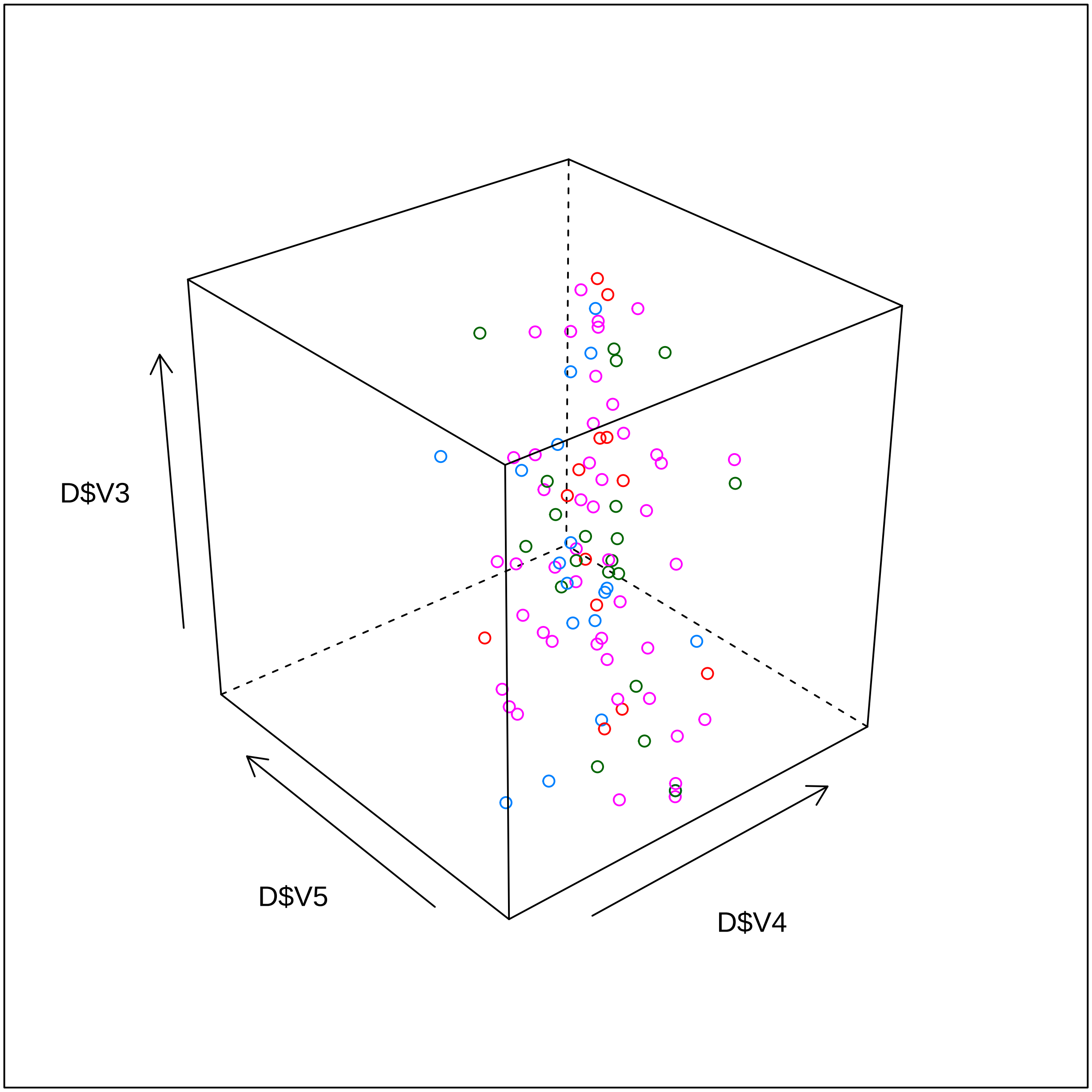

D$V1
D$V5
D$V3


D$V1
D$V5
D$V4


D$V2
D$V4
D$V3


D$V2
D$V5
D$V3


D$V2
D$V5
D$V4


D$V3
D$V5
D$V4

Part III
Test Of (Univariate) Normality

Kolmogorov-Smirnoff Test

Testing Normality For V1

```
##      V1    V2    V3    V4    V5 V6
## 1 0.889 1.389 1.555 2.222 1.945  1
## 2 2.813 1.437 0.999 2.312 2.312  2
## 3 1.454 1.091 2.364 2.455 2.909  3
## 4 0.294 0.941 1.059 2.000 1.000  1
## 5 2.727 2.545 2.819 2.727 4.091  0
## 6 3.937 1.250 1.937 2.937 3.749  1
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  D$V1
## D = 0.75216, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Testing Normality For V2

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  D$V2
## D = 0.83114, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Testing Normality For V3

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  D$V3
## D = 0.81049, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Testing Normality For V4

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  D$V4
## D = 0.93621, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Testing Normality For V5

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  D$V5
## D = 0.84134, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Shapiro Wilk's Test

Testing Normality For V1

```
##
##  Shapiro-Wilk normality test
##
## data:  D$V1
## W = 0.96854, p-value = 0.01885
```

Testing Normality For V2

```
##
##  Shapiro-Wilk normality test
##
## data:  D$V2
## W = 0.89033, p-value = 6.332e-07
```

Testing Normality For V3

```
##
##  Shapiro-Wilk normality test
##
## data:  D$V3
## W = 0.97906, p-value = 0.1198
```

Testing Normality For V4

```
##
##  Shapiro-Wilk normality test
##
## data:  D$V4
## W = 0.96171, p-value = 0.005985
```

Testing Normality For V5

```
##
##  Shapiro-Wilk normality test
##
## data:  D$V5
## W = 0.97826, p-value = 0.1041
```

- Observations :

1. V3 and V5 exhibit normality. The other three do not.
2. V2 strongly rejects normality.

Part IV

Transformation to Convert Non-Normal Variables To Normal

From the previous tests we can observe that V1,V2 and V4 are not Normal random variables. We apply the following transformations to make them normal.

0.2 V1 (Average number of symptoms throughout the treatment)

As the data is positively skewed (of a positive variable), we apply the square root transformation to check whether it works. This is because the square root transformation is a standard transformation for such cases. We apply the shapiro Wilk's test. The result is as follows :

```
##
##  Shapiro-Wilk normality test
##
## data:  sqrt(D$V1)
## W = 0.98805, p-value = 0.5267
```

We note that the test fails to reject normality at 5 % level of significance.

0.3 V4 (Average amount of through Food consumed throughout the treatment)

As the data is positively skewed, we apply the square root transformation.

The result is as follows :

```
##
##  Shapiro-Wilk normality test
##
## data:  sqrt(D$V4)
## W = 0.96699, p-value = 0.01447
```

We note that the test fails to reject normality at 1 % level of significance.

0.4 V2 (Average amount of "Activity" throughout the treatment)

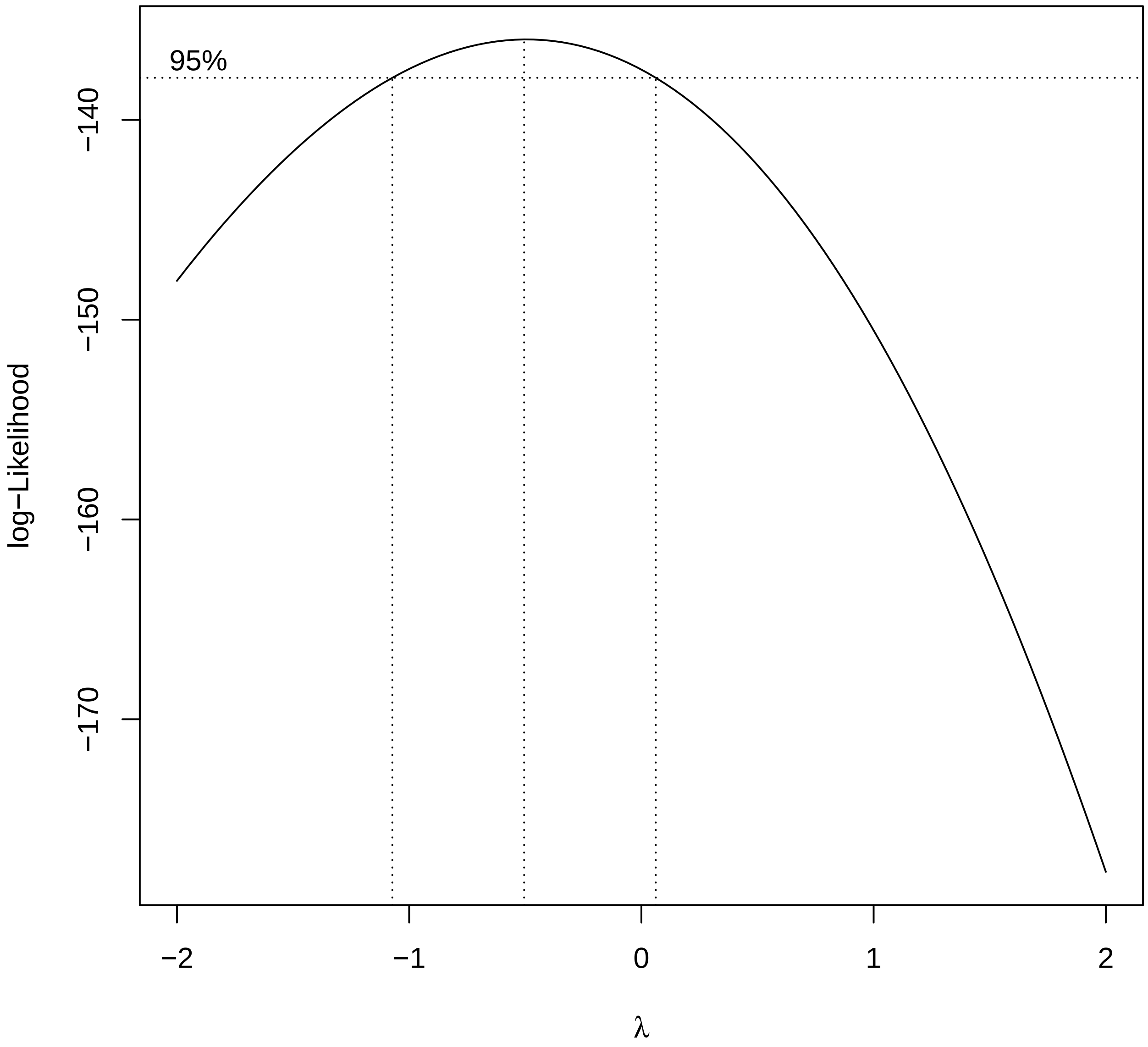
- As the data is positively skewed, we apply the square root transformation.

The result is as follows :

```
##
##  Shapiro-Wilk normality test
##
## data:  sqrt(D$V2)
## W = 0.91434, p-value = 8.643e-06
```

We note that the test strongly rejects normality.

- As the square-root transformation does not work, we apply the Box-Cox transformation and check how it performs.

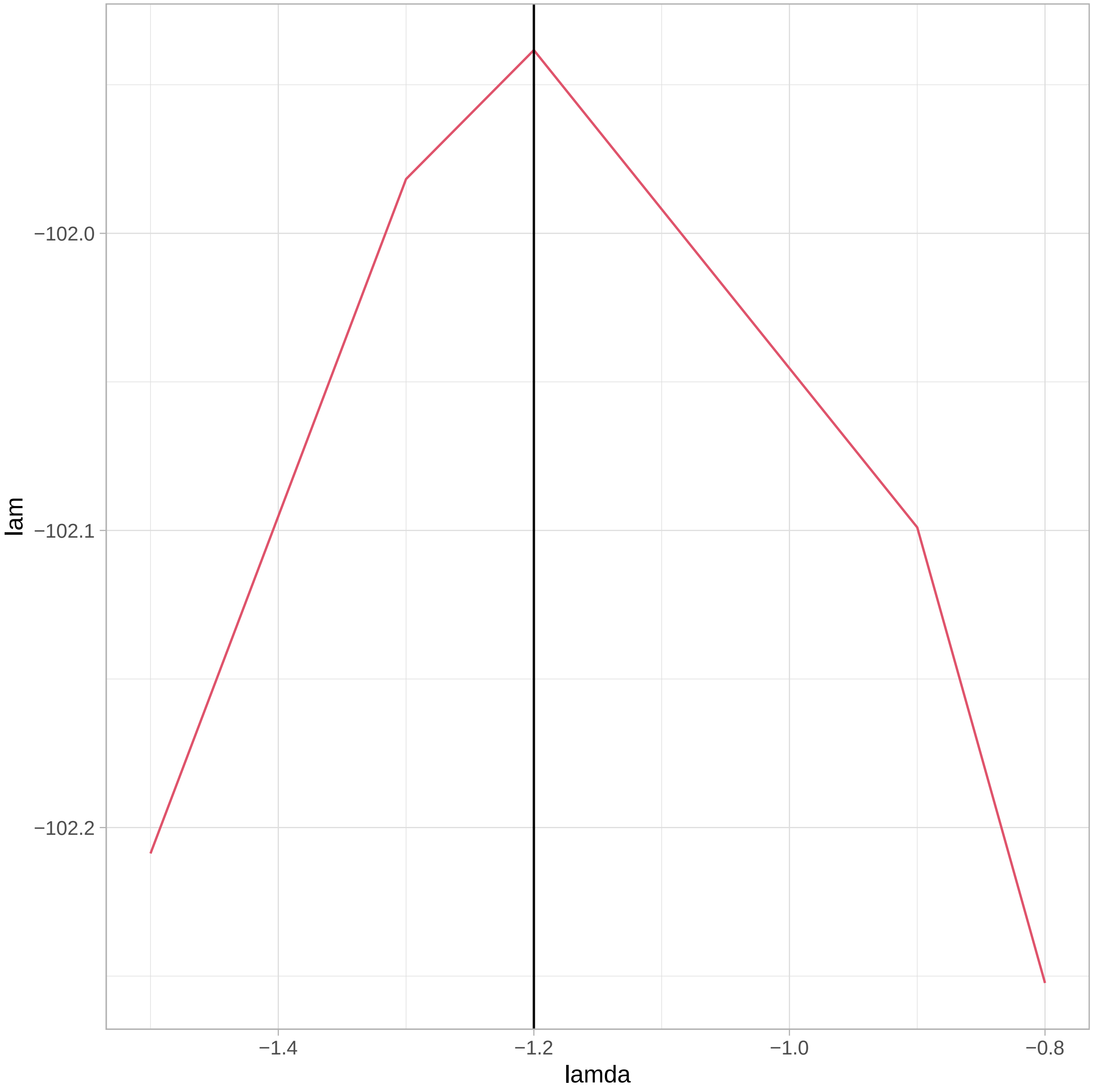


We see that the value of λ is approximately -0.5. We check for normality :

```
##
##  Shapiro-Wilk normality test
##
## data:  (1 - (D$V2)^(-0.5))/0.5
## W = 0.92445, p-value = 2.922e-05
```

We note that the test strongly rejects normality.

- As the Box-Cox transformation does not work, we apply the Johnson transformation and check how it performs.



We note that the value of the λ comes out to be -1.2. We apply this transformation and check for normality. The results are as follows :

```
##
##  Shapiro-Wilk normality test
##
## data:  X_2
## W = 0.92136, p-value = 1.995e-05
```

We note that the test strongly rejects normality.

Therefore, we conclude that the distribution of this variable greatly deviates from that of normal, and general transformations do not work.

Part V

Tests of (Multivariate) Normality

Till now in our analysis, we have tranformed V1 and V4. We have also dropped V2 because of the above mentioned reasons. So now, let the variables be as follows :

1. $Y_1 = \sqrt{V_1}$
2. $Y_2 = V_3$
3. $Y_3 = \sqrt{V_4}$
4. $Y_4 = V_5$

Therefore, now we have multivariate vectors **X**$= (Y_1, Y_2, Y_3, Y_4)$. Now we are interested in testing the multivariate normality of **X.** There are several ways (both graphical and theoretical) to check for multivariate normality. We apply the following :

- Graphical Methods :

1. Contour Plots (For Bivariate Normality).
2. Perspective Plots (For Trivariate Normality).
3. Gamma Plots.

- Tests :

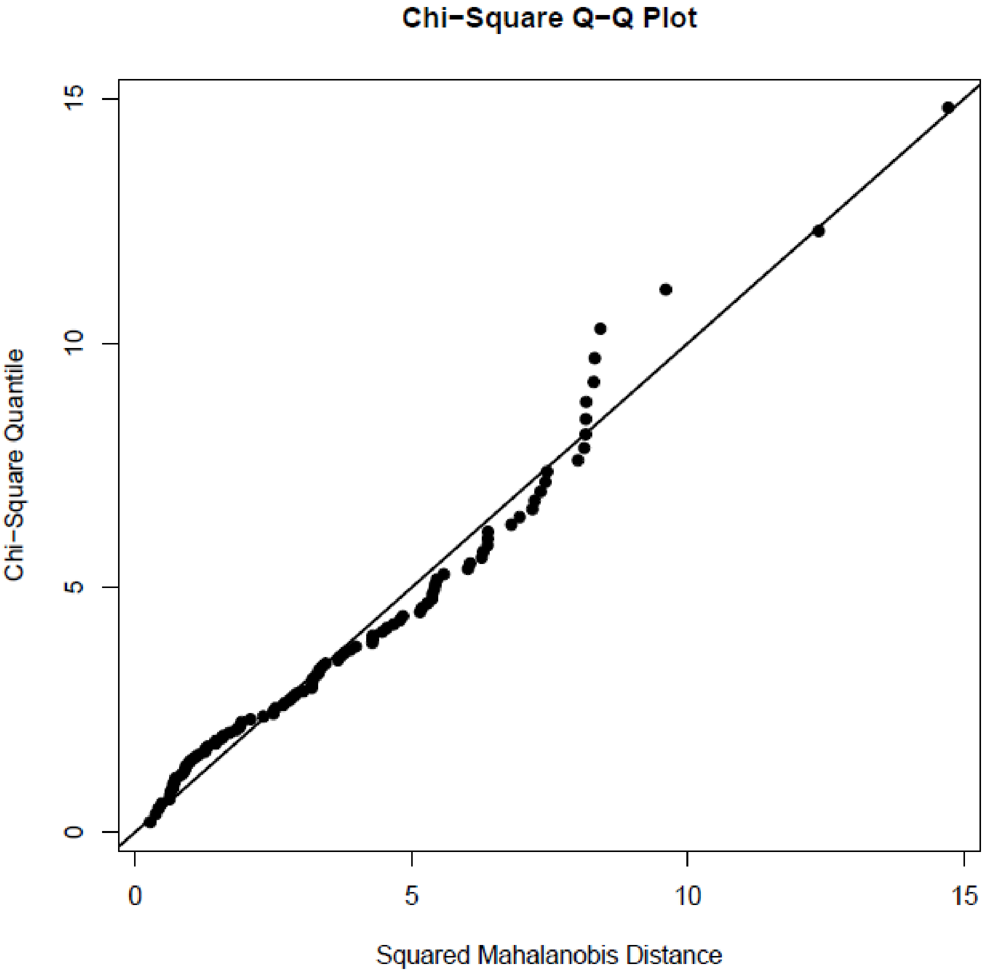
1. Royston's test
2. Henze-Zirkler's test
3. Characterization of Multivariate Normality and Bonferonni's correction.

Graphical Methods

We shall first look into graphical methods.

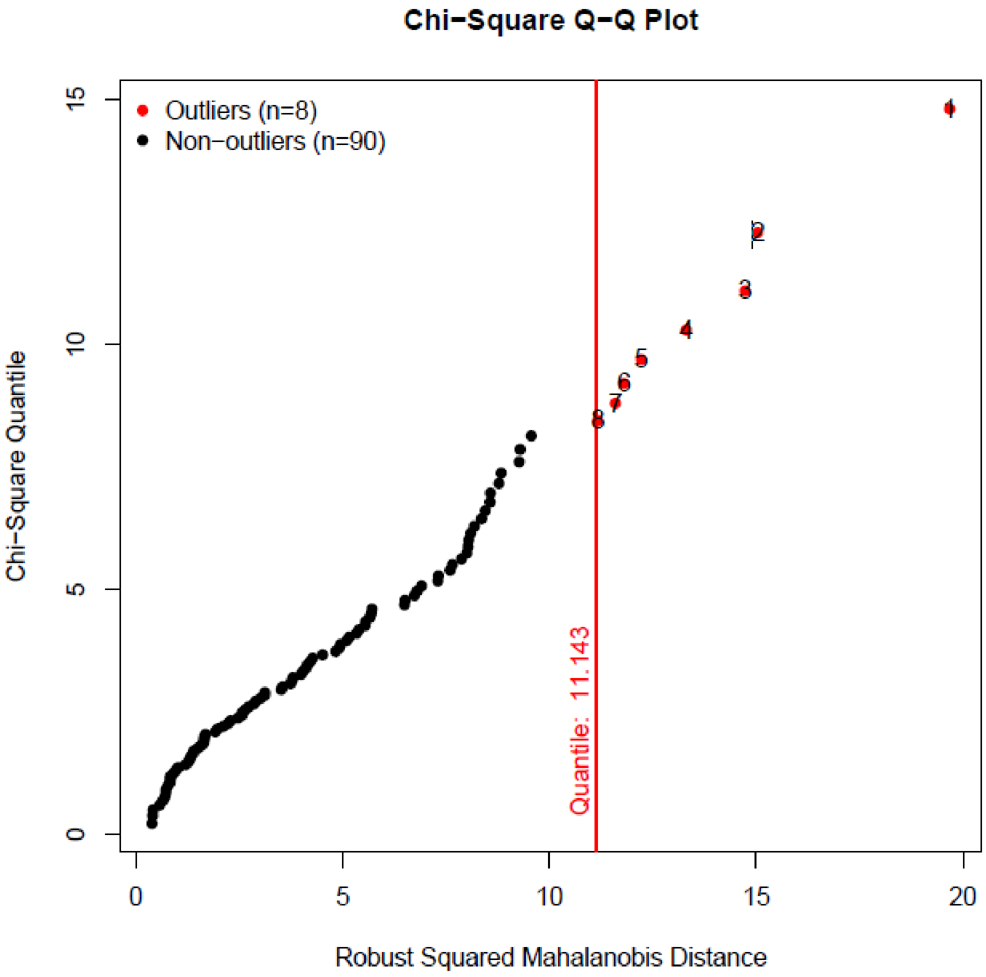
Gamma Plot

For Gamma plot, we calculate first calculate Mahalanobis's distance and plot them along the quantiles of $\chi(p)$, where p is the number of variables. The plot is as follows :



Comment : From the Gamma Plot it seems that most points (except 7-8 points) fall along the line. Now , our dataset contains 98 observations. So, a few observations may behave erratically. If the number of outliers is comparatively (very) small, we may assume multuvariate normality assumption to be true. Now let us check for outliers :

- Checking outliers :



Comment : As observed before, 8 observations are deviating from multivariate normality. As this value is comparatively small w.r.t 100, we choose to accept multivariate normality.

Tests

Now we apply some tests to check for multivariate normality.

Royston's test

```
## $multivariateNormality
##      Test        H    p value MVN
## 1 Royston 11.41584 0.02211722  NO
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    Y1        0.3655    0.4292    YES
## 2 Anderson-Darling    Y2        0.5126    0.1898    YES
## 3 Anderson-Darling    Y3        1.2325    0.0031    NO
## 4 Anderson-Darling    Y4        0.4783    0.2309    YES
```

```
##
## $Descriptives
##     n     Mean   Std.Dev   Median      Min      Max     25th     75th
## Y1 98 1.775344 0.6281126 1.844961 0.000000 3.234347 1.373660 2.275576
## Y2 98 2.137602 0.7559291 2.177500 0.666000 4.000000 1.564250 2.712000
## Y3 98 1.482131 0.1114231 1.462695 1.134019 1.713768 1.414214 1.562129
## Y4 98 2.574827 0.9285324 2.500000 1.000000 5.000000 1.924000 3.272000
##           Skew    Kurtosis
## Y1 -0.34993968 -0.09364816
## Y2  0.08460178 -0.73821013
## Y3  0.12367833 -0.12766913
## Y4  0.24483504 -0.62899658
```

We observe the following :

1. The test rejects the assumption of multivariate normality at 5% level of significance. However, the test fails to reject the assumption of multivariate normality at 1% level of significance.
2. We may accept the assumption of multivariate normality if we choose $\alpha = 0.01$.

However, there is a possibility that the data follows multivariate normal group wise.

- For Group 0

```
## $multivariateNormality
##      Test        H   p value MVN
## 1 Royston 5.391579 0.2610785 YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    V1        0.4359    0.2672    YES
## 2 Anderson-Darling    V3        0.2726    0.6282    YES
## 3 Anderson-Darling    V4        0.7416    0.0441    NO
## 4 Anderson-Darling    V5        0.2254    0.7900    YES
##
## $Descriptives
##     n     Mean   Std.Dev Median   Min    Max  25th   75th       Skew
## V1 19 3.638211 2.6657918  3.714 0.000 10.461 1.553 5.1835 0.48003593
## V3 19 1.937684 0.9031670  1.941 0.666  4.000 1.091 2.5715 0.40585598
## V4 19 2.136474 0.2908127  2.000 1.667  2.727 2.000 2.3305 0.56750134
## V5 19 2.545158 0.9224974  2.444 1.000  4.091 2.000 3.3245 0.03166006
##       Kurtosis
## V1  0.04125762
## V3 -0.74223631
## V4 -0.74205578
## V5 -1.11051003
```

Comment : Multivariable Normal is observed in this group.

- For Group 1

```
## $multivariateNormality
##      Test        H    p value MVN
## 1 Royston 8.116155 0.09104617 YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    V1        0.5254    0.1714    YES
## 2 Anderson-Darling    V3        0.5418    0.1556    YES
## 3 Anderson-Darling    V4        0.8178    0.0320    NO
## 4 Anderson-Darling    V5        0.2335    0.7844    YES
##
## $Descriptives
##     n     Mean   Std.Dev Median   Min   Max   25th    75th      Skew   Kurtosis
## V1 46 3.309804 1.9390133 3.2575 0.222 7.571 1.6375 4.54175 0.3784095 -0.8529747
## V3 46 2.148152 0.7322082 2.0000 0.950 3.689 1.6180 2.68875 0.1876929 -1.0564515
## V4 46 2.217174 0.3245969 2.1540 1.611 2.937 2.0000 2.48600 0.4456041 -0.7757405
## V5 46 2.587957 0.9351466 2.5390 1.000 5.000 1.9240 3.13650 0.2200974 -0.4103211
```

Comment : Multivariable Normal is observed in this group.

- For Group 2

```
## $multivariateNormality
##      Test        H    p value MVN
## 1 Royston 8.987751 0.06570336 YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    V1        0.6174    0.0930    YES
## 2 Anderson-Darling    V3        0.6189    0.0922    YES
## 3 Anderson-Darling    V4        0.7714    0.0373    NO
## 4 Anderson-Darling    V5        0.3093    0.5274    YES
##
## $Descriptives
##     n    Mean   Std.Dev Median   Min   Max    25th    75th       Skew
## V1 20 3.57180 1.8605152  2.909 0.800 7.231 2.41625 5.23650  0.4093321
## V3 20 2.12470 0.6744109  2.225 0.999 3.154 1.83175 2.51025 -0.4160068
## V4 20 2.27230 0.3131959  2.267 1.800 2.889 2.03375 2.34600  0.6508381
## V5 20 2.64175 0.9236534  2.506 1.000 4.384 2.04650 3.34275  0.2544680
##      Kurtosis
## V1 -1.0930361
## V3 -1.0324954
## V4 -0.7429230
## V5 -0.9765935
```

Comment : Multivariable Normal is observed in this group.

- For Group 3

```
## $multivariateNormality
##      Test        H   p value MVN
## 1 Royston 2.079578 0.6621113 YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    V1        0.4793    0.1935    YES
## 2 Anderson-Darling    V3        0.2026    0.8433    YES
## 3 Anderson-Darling    V4        0.3140    0.5057    YES
## 4 Anderson-Darling    V5        0.3467    0.4226    YES
##
## $Descriptives
##     n     Mean   Std.Dev Median   Min   Max  25th  75th       Skew   Kurtosis
## V1 13 4.179769 2.5928553  3.667 0.750 9.454 2.444 5.250  0.6451586 -0.8688624
## V3 13 2.412308 0.7204971  2.375 1.223 3.818 2.111 2.812  0.0388006 -0.8344643
## V4 13 2.188692 0.4466342  2.125 1.286 2.929 1.933 2.455 -0.1631988 -0.7828205
## V5 13 2.468769 1.0202204  2.125 1.252 4.428 1.776 3.142  0.5062718 -1.1342029
```

Comment : Multivariable Normal is observed in this group.

Conclusion : We conclude that groupwise , multivariate normality is achieved.

Part VI

Dropping V_2 variable.

In our study of MANOVA and discriminant analysis, we have decided to drop the V_2variable. This variable corresponds to the average amount of "activity" of a patient throughout their treatment. We have already noted that there is ambiguity in the meaning of activity. What is it measuring ? We do not know.

We had assumed that activity means physical activity and it is measured on a scale of 1-5 "by some means". However, that interpretation is not working due to the following reasons.

- If by activity we mean physical activity, it should have strong association with the average amount of sleep. Let us calculate their correlation :

```
## [1] 0.1874625
```

This value is surprisingly low.

- If by activity we mean physical activity, it certainly should not be independent from average amount of sleep. We apply Kendall's τ test. We note that as it is a non-parametric test, we do not need the assumption of normality. We do however need these variables to be continuous, which we have already assumed.

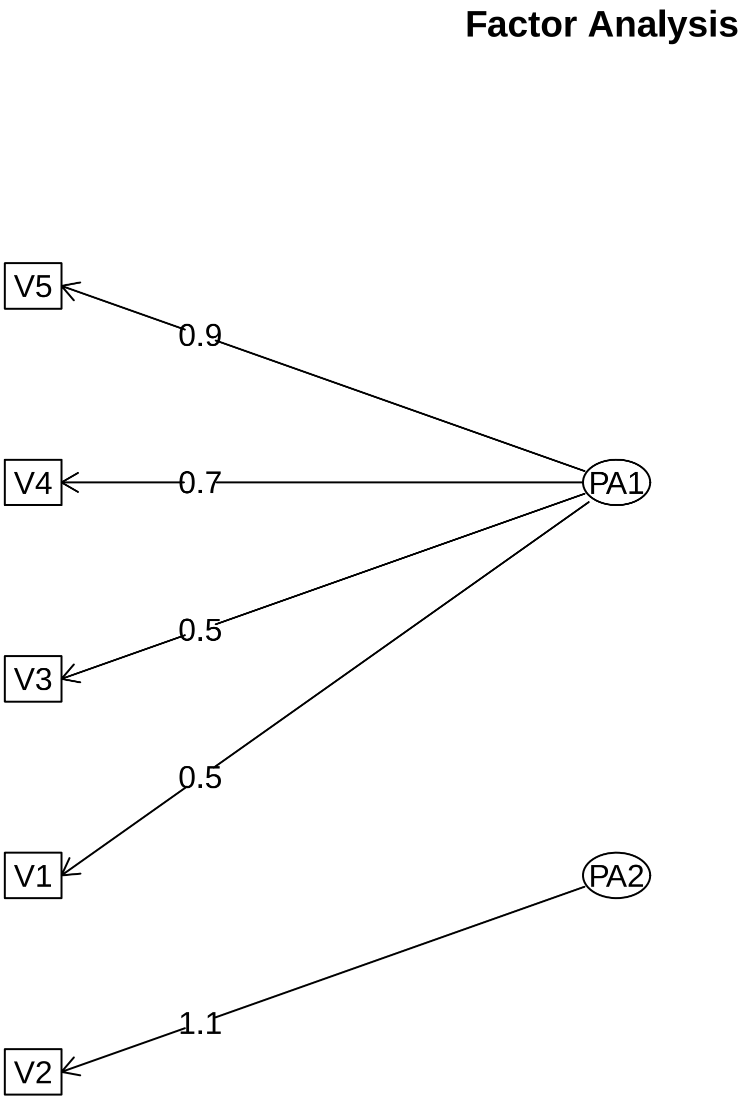
```
##
##  Kendall's rank correlation tau
##
## data:  D$V2 and D$V3
## z = 2.0501, p-value = 0.9798
## alternative hypothesis: true tau is less than 0
## sample estimates:
##       tau
## 0.1429754
```

We observe that the test strongly accepts the assumption of indepence.

- We have already seen that this variable strongly deviates from normality. Square-root transformation, Box-cox transformation and Johnson's transformation were applied but none worked. Hence, we CANNOT apply Box's M test, MANOVA or linear discriminant analysis. In principle we may try to apply Fisher's linear discriminants as normality assumption is not required. However, we need to check the assumption of homoscedasticity . We do that using Box's M test. However, that in turn requires the assumption of normality.

- We will later see that Factor Analysis (discussed in part X) suggests that there is one single factor affecting/influencing only this variable, independent of another factor affecting all the others. We shall discuss this in great detail in part X.

```
## maximum iteration exceeded
```



Part VII

Testing Equality of Population Cov Matrices and means

Box’s M test

Previously, we’ve seen that observations belonging to different catagories are coming from multivariate normal populations.Now, we are eager to know whether their variance-covariance matrix are same or not. Keeping that motive, we are conducting Box’s M-test and from this test we can conclude that we failed to reject Null hypothesis(i.e, variance-covariance matrices are same.) at level 0.05.

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  cbind(D$V1, D$V3, D$V4, D$V5)
## Chi-Sq (approx.) = 31.537, df = 30, p-value = 0.3894
```

Comment : We fail to reject H_0 at 5% level of significance. In other words, we accept the assumpion of Homoscedasticity.

MANOVA

From Box’s M-Test, we concluded that observations of different catagories are coming from homoscedastic multivariate normal populations. Now, we are interested to know whether the observations of different catagories are coming from same multivariate normal distribution or not(Basically, here we are testing the equality of population mean vectors of the observations belonging to different catagorties). From the Test of equality of means we can conclude that null hypothesis(i.e,population mean vectors of the observations of different catagories are equal) is accepted at level 0.05.

```
##           Df   Pillai approx F num Df den Df Pr(>F)
## V6         1 0.052783   1.2956      4     93 0.2775
## Residuals 96
```

Comment : We fail to reject H_0 at 5% level of significance. In other words, we accept the assumpion of equality of means.

- We note that we have failed to reject both Box’s M test as well as MANOVA. This implies that all the groups follow same distribution. The EDA had already suggested this.

Part VIII

LDA (Linear Discriminant Analysis)

Here we observed that observations belonging to different catagories are coming homoscedastic multivariate populations. So the assumptions required to apply LDA are satisfied . Hence we are applying LDA to discriminate the observations which are coming from different parent population.

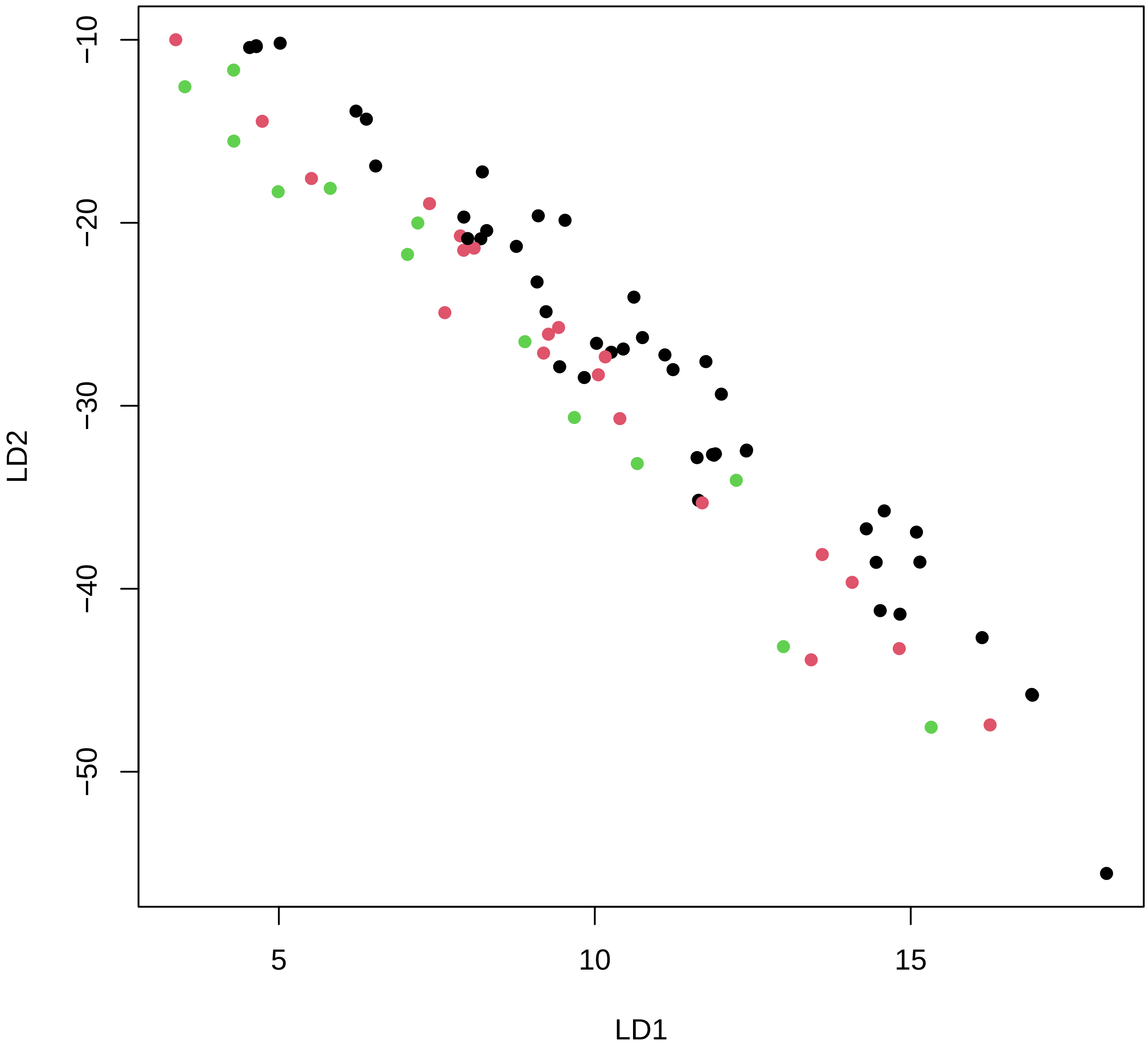
We apply the method of (One - Out) Cross-Validation to assess how well LDA performs.

The APER is :

```
## [1] 0.5510204
```

Confusion Matrix

```
##          true
## predicted  0  1  2  3
##         0  0 17  0  2
##         1  3 41  1  1
##         2  0 20  0  0
##         3  0 13  0  0
```



From the confusion matrix and APER(calculated by hold-out procedure) that error rate is high. So, we can conclude that LDA is working badly here (which is obvious because observations of different catagories are coming from same parent population.)

Part IX

PCA [Principal Component Analysis]

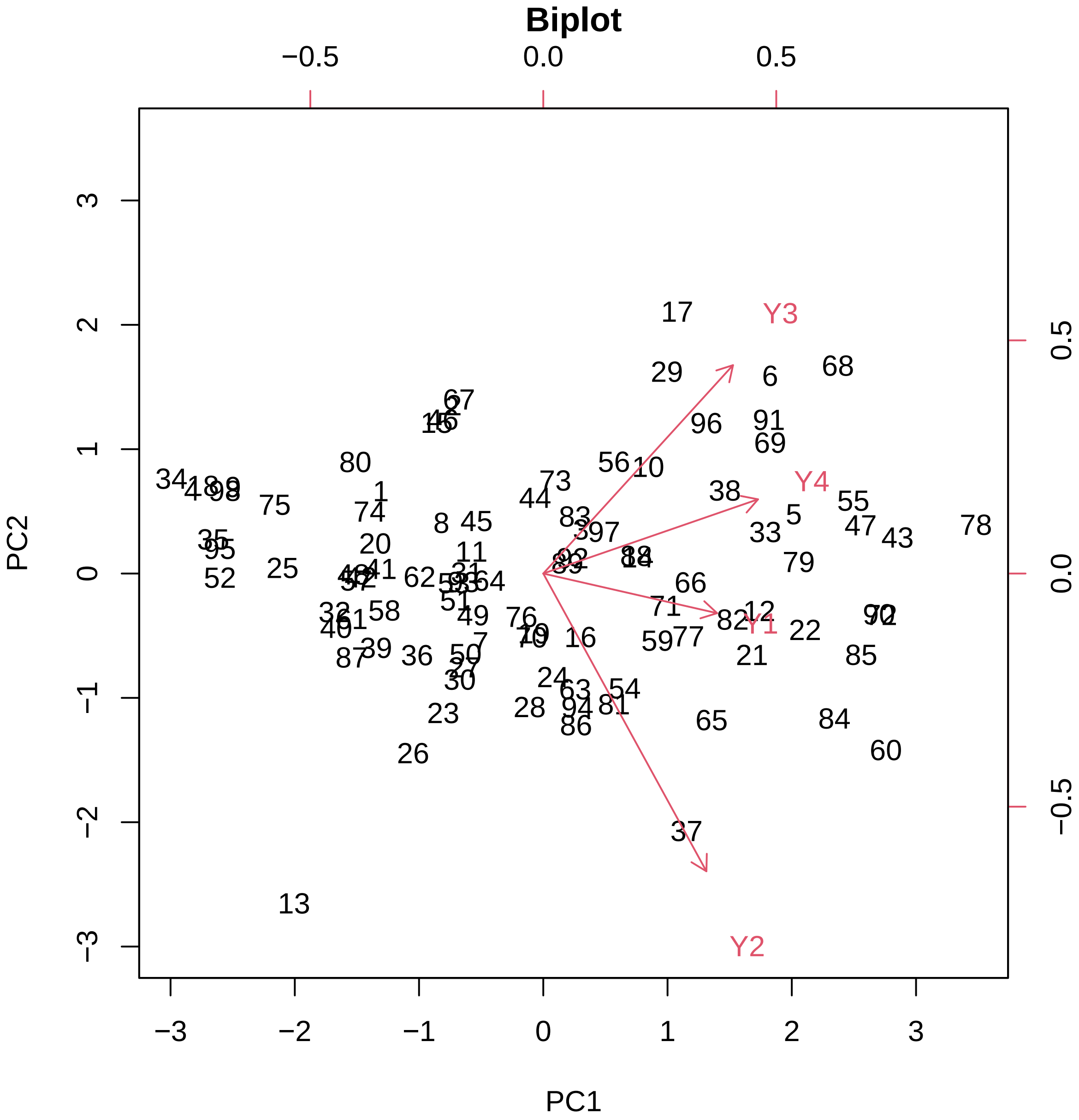
We apply PCA to study the correlation structure of the variables.We note that PCA is primarily used for dimensionality reduction purpose. As our dataset only has 4 variables, applying PCA may not be much meaningful. We first check how it performs.If one or two variables explain more than 90% of the variablity then we may apply (that is also debatable as we have only 4variables and we will be losing interpretability). If not, we will choose not to transform our variables.

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

- After applying PCA, the summary is :

```
## Importance of components:
##                           PC1    PC2    PC3    PC4
## Standard deviation     1.5626 0.8268 0.7879 0.5040
## Proportion of Variance 0.6104 0.1709 0.1552 0.0635
## Cumulative Proportion  0.6104 0.7813 0.9365 1.0000
```

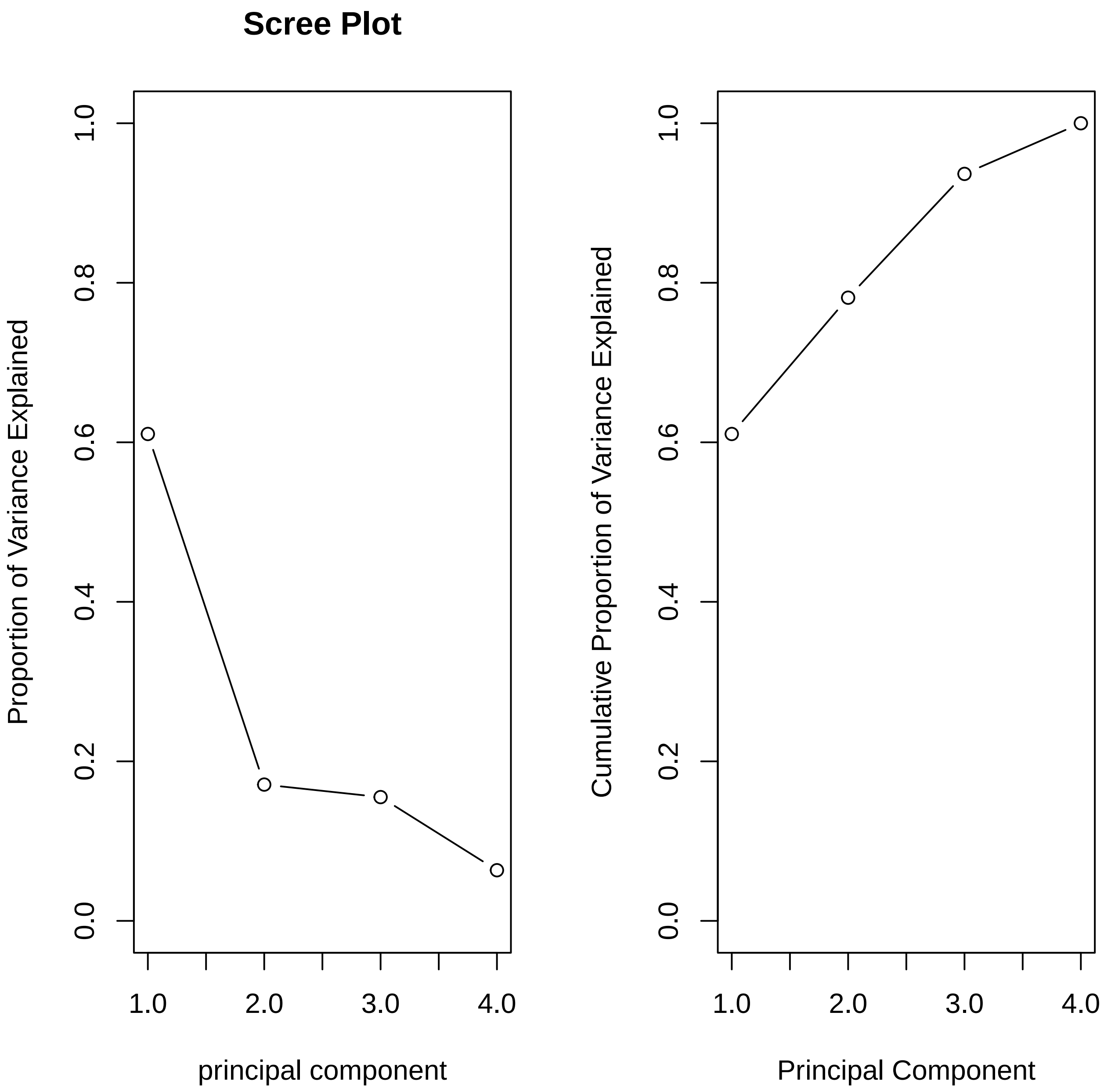
- The Biplot is as follows :



We note that :

1. The average amount of sleep strongly influences PCA 2.
2. The average number of symptoms and the average amount of appetite strongly influences PCA 1.
3. The angle between Y1 is less with all three other variables indicating its association with he others.

- The plots showing proportion of variance explained by each principal component and cumulative proportion of variance explained by each component are as follows :



- Our observations are as follows :

1. It takes 3 out of 4 principal components to explain more than 90% of variability in the data.
2. The first pc explains more than 60% of variability in the data, but the rest do not have a significant effect.

We conclude that it is not logical to apply principal component transformation and use the principal components, in place of the original variables.

Part X

Factor Analysis

We apply factor analysis in two models. In the first case we use the full model using all five variables. As this does not follow multivariate normality, we shall only use the PCA method. Then we drop V_2 and apply factor analysis by using both PCA method and method of maximum likelihood.

Full Model

Bartlett's test

We first apply Bartlett's test to check whether the correlation matrix is factorable..

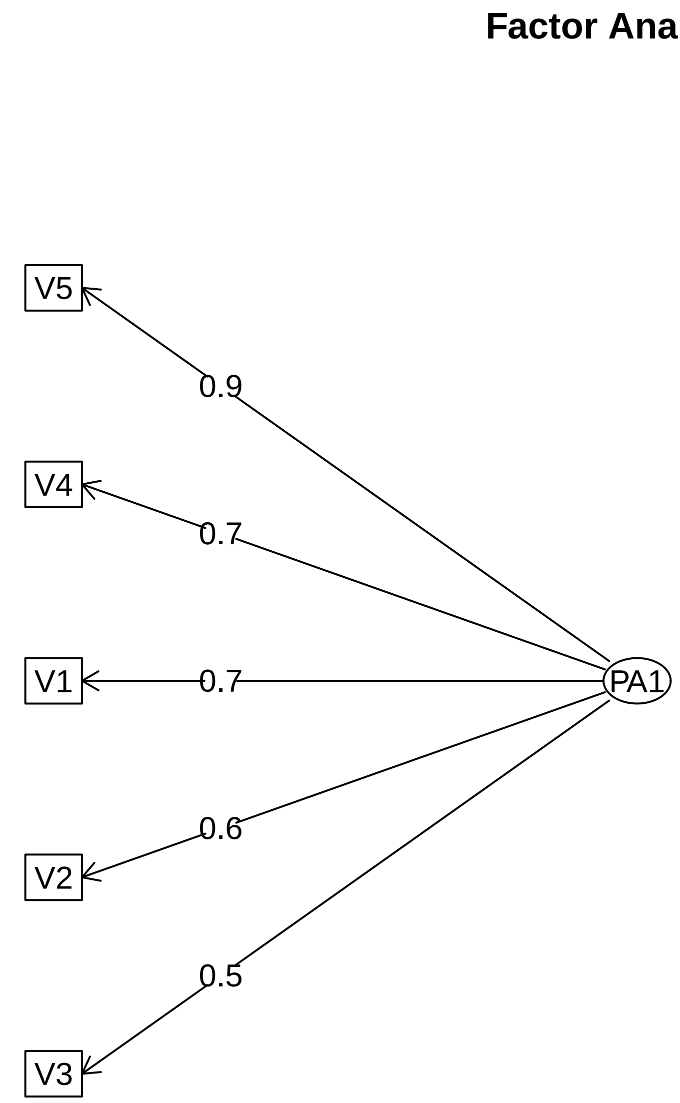
```
## R was not square, finding R from data

## $chisq
## [1] 180.4333
##
## $p.value
## [1] 1.944077e-30
##
## $df
## [1] 15
```

Hence, we strongly reject the null hypothesis. Consequently we conclude that the correlation matrix is factorable.

- We first check with one factor.

```
## Factor Analysis using method =  pa
## Call: fa(r = D[, -6], nfactors = 1, rotate = "varimax", scores = "regression",
##     max.iter = 100, fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##     PA1   h2   u2 com
## V1 0.65 0.43 0.57   1
## V2 0.63 0.40 0.60   1
## V3 0.49 0.24 0.76   1
## V4 0.71 0.51 0.49   1
## V5 0.91 0.83 0.17   1
##
##                 PA1
## SS loadings    2.40
## Proportion Var 0.48
##
```

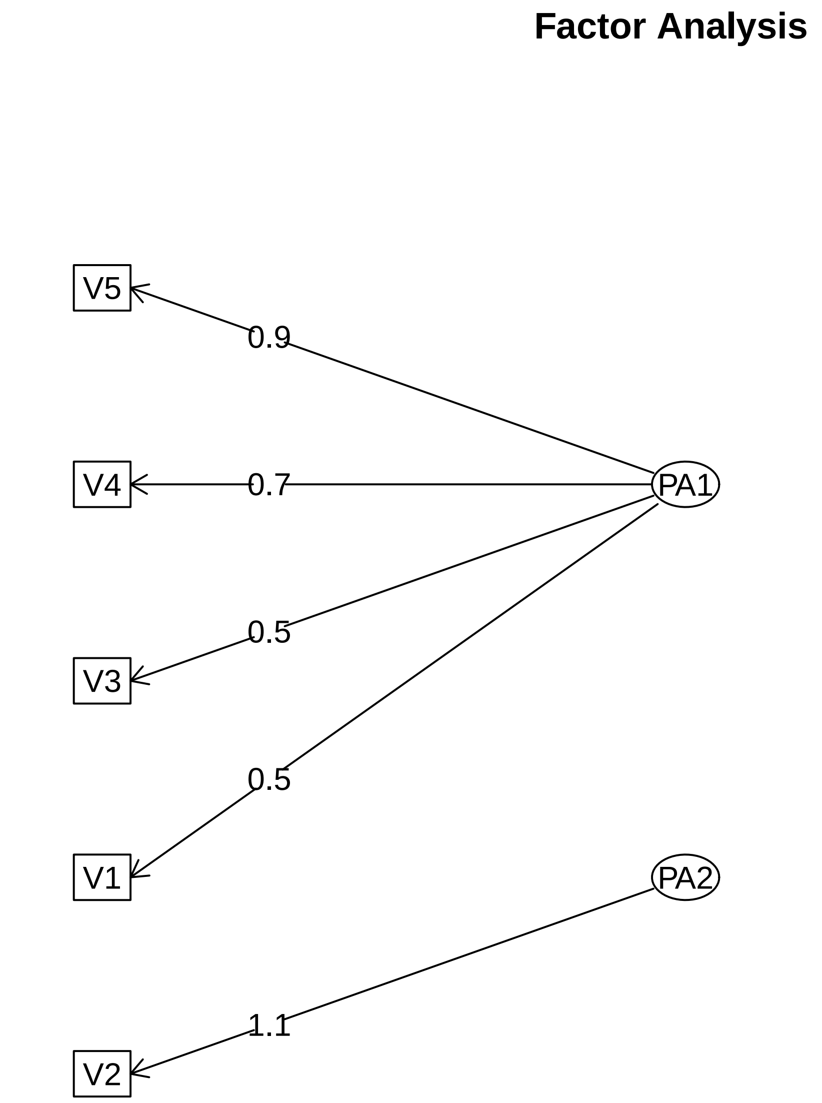
```
## Mean item complexity =  1
## Test of the hypothesis that 1 factor is sufficient.
##
## df null model =  10  with the objective function =  1.86 with Chi Square =  175.54
## df of  the model are 5  and the objective function was  0.19
##
## The root mean square of the residuals (RMSR) is  0.07
## The df corrected root mean square of the residuals is  0.1
##
## The harmonic n.obs is  98 with the empirical chi square  10.34  with prob <  0.066
## The total n.obs was  98  with Likelihood Chi Square =  17.98  with prob <  0.003
##
## Tucker Lewis Index of factoring reliability =  0.842
## RMSEA index =  0.162  and the 90 % confidence intervals are  0.086 0.248
## BIC =  -4.94
## Fit based upon off diagonal values = 0.98
## Measures of factor score adequacy
##                                                    PA1
## Correlation of (regression) scores with factors   0.94
## Multiple R square of scores with factors          0.89
## Minimum correlation of possible factor scores     0.77
```



Observation : The test rejects the hypothesis that one factor is sufficient.

- We now check with two factors.

```
## maximum iteration exceeded

## Factor Analysis using method =  pa
## Call: fa(r = D[, -6], nfactors = 2, rotate = "varimax", scores = "regression",
##     max.iter = 100, fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##     PA1  PA2   h2     u2 com
## V1 0.49 0.38 0.38  0.616 1.9
## V2 0.24 1.12 1.32 -0.317 1.1
## V3 0.54 0.07 0.30  0.705 1.0
## V4 0.65 0.26 0.50  0.504 1.3
## V5 0.92 0.28 0.92  0.076 1.2
##
##                        PA1  PA2
## SS loadings           1.86 1.56
## Proportion Var        0.37 0.31
## Cumulative Var        0.37 0.68
## Proportion Explained  0.54 0.46
## Cumulative Proportion 0.54 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 factors are sufficient.
##
## df null model =  10  with the objective function =  1.86 with Chi Square =  175.54
## df of  the model are 1  and the objective function was  0.03
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.09
##
## The harmonic n.obs is  98 with the empirical chi square  1.74  with prob <  0.19
## The total n.obs was  98  with Likelihood Chi Square =  3.14  with prob <  0.076
##
## Tucker Lewis Index of factoring reliability =  0.869
## RMSEA index =  0.147  and the 90 % confidence intervals are  0 0.347
## BIC =  -1.45
## Fit based upon off diagonal values = 1
```



Observation : V_2 has a different factor compared to others.Via studying correlation structure and Kendall's tau test we had suspected that this variable does not allign with the rest. This plot supports our decision to drop the variable.

Model after dropping V2

PCA METHOD

We first apply the PC method.

- We first check for one factor :

```
##
## Call:
## factanal(x = D.Y, factors = 1, scores = c("regression"), rotation = "varimax")
##
## Uniquenesses:
##    Y1    Y2    Y3    Y4
## 0.681 0.741 0.488 0.045
##
## Loadings:
##    Factor1
## Y1 0.565
## Y2 0.509
## Y3 0.715
## Y4 0.977
##
##                Factor1
## SS loadings      2.045
## Proportion Var   0.511
##
## Test of the hypothesis that 1 factor is sufficient.
## The chi square statistic is 1.31 on 2 degrees of freedom.
## The p-value is 0.52
```

We note that the proportional variance is 0.511.

- We now check for two factor :

```
## Error in factanal(D.Y, factors = 2, scores = c("regression"), rotation = "varimax"):
2 factors are too many for 4 variables
```

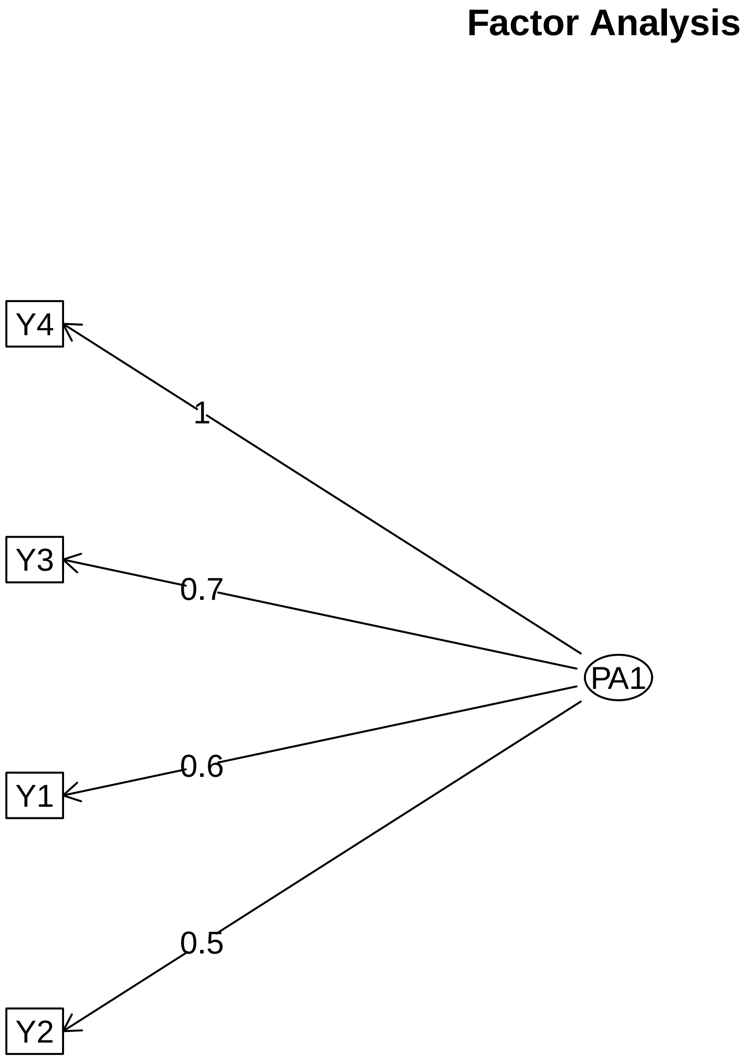
```
##
## Call:
## factanal(x = D.Y, factors = 1, scores = c("regression"), rotation = "varimax")
##
## Uniquenesses:
##    Y1    Y2    Y3    Y4
## 0.681 0.741 0.488 0.045
##
## Loadings:
##    Factor1
## Y1 0.565
## Y2 0.509
## Y3 0.715
## Y4 0.977
##
##                Factor1
## SS loadings      2.045
## Proportion Var   0.511
##
## Test of the hypothesis that 1 factor is sufficient.
## The chi square statistic is 1.31 on 2 degrees of freedom.
## The p-value is 0.52
```

We note that 2 factors will be considered a lot. Hence we stick with one factor.

Maximum Likelihood Method

- We first check for one factor :

```
## Factor Analysis using method =  pa
## Call: fa(r = cbind(Y1, Y2, Y3, Y4), nfactors = 1, rotate = "varimax",
##     scores = "regression", max.iter = 100, fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##     PA1   h2    u2 com
## Y1 0.58 0.34 0.659   1
## Y2 0.53 0.28 0.720   1
## Y3 0.70 0.49 0.514   1
## Y4 0.97 0.94 0.064   1
##
##                 PA1
## SS loadings    2.04
## Proportion Var 0.51
##
## Mean item complexity =  1
## Test of the hypothesis that 1 factor is sufficient.
##
## df null model =  6  with the objective function =  1.34 with Chi Square =  126.61
## df of  the model are 2  and the objective function was  0.02
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.05
##
## The harmonic n.obs is  98 with the empirical chi square  1.03  with prob <  0.6
## The total n.obs was  98  with Likelihood Chi Square =  1.71  with prob <  0.42
##
## Tucker Lewis Index of factoring reliability =  1.007
## RMSEA index =  0  and the 90 % confidence intervals are  0 0.192
## BIC =  -7.46
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                    PA1
## Correlation of (regression) scores with factors   0.97
## Multiple R square of scores with factors          0.94
## Minimum correlation of possible factor scores     0.89
```

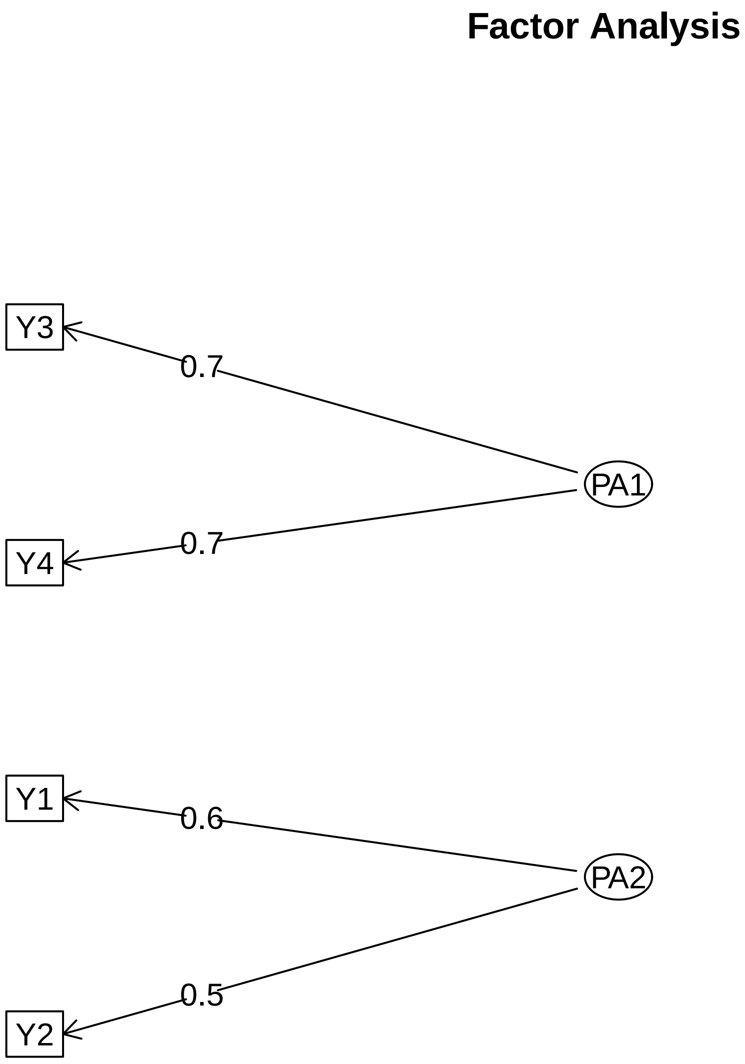


Observations :

1. The proportional variance explained is 0.51

- We now check for two factor :

```
## Factor Analysis using method =  pa
## Call: fa(r = cbind(Y1, Y2, Y3, Y4), nfactors = 2, rotate = "varimax",
##     scores = "regression", max.iter = 100, fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##     PA1  PA2   h2   u2 com
## Y1 0.29 0.56 0.40 0.60 1.5
## Y2 0.26 0.51 0.33 0.67 1.5
## Y3 0.72 0.31 0.61 0.39 1.4
## Y4 0.71 0.61 0.88 0.12 2.0
##
##                        PA1  PA2
## SS loadings           1.17 1.06
## Proportion Var        0.29 0.26
## Cumulative Var        0.29 0.56
## Proportion Explained  0.53 0.47
## Cumulative Proportion 0.53 1.00
##
## Mean item complexity =  1.6
## Test of the hypothesis that 2 factors are sufficient.
##
## df null model =  6  with the objective function =  1.34 with Chi Square =  126.61
## df of  the model are -1  and the objective function was  0
##
## The root mean square of the residuals (RMSR) is  0
## The df corrected root mean square of the residuals is  NA
##
## The harmonic n.obs is  98 with the empirical chi square  0  with prob <  NA
## The total n.obs was  98  with Likelihood Chi Square =  0  with prob <  NA
##
## Tucker Lewis Index of factoring reliability =  1.051
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                    PA1  PA2
## Correlation of (regression) scores with factors   0.79 0.72
## Multiple R square of scores with factors          0.62 0.52
## Minimum correlation of possible factor scores     0.24 0.04
```



Observations :

1. The cumulative variance explained is 1.
2. The first factor contains measures of appetite and amount of food consumed. This indicates that this is some kind of “ Metabolic/Digestive” factor in the radiotherapy treatment.
3. The second factor contains measures of number of symptoms and average amount of sleep.

- Conclusion : We have applied both maximum likelihood method and pc method. We observe that two factors obtained via maximum likelihood explain well.

Conclusion

We conclude the following :

1. Both Box's M test and MANOVA are being accepted. This implies that groupwise they belong to the same distribution. This was also evident from exploratory data analysis. Therefore, classification does not make much sense for his data. Every type of classification method will perform poorly in this dataset.

2. There is a high chance that "activity" which was being measured in this study was not physical activity. One of the side effects of radiotherapy is cognitive decline. Then it will make sense why it is independent of sleep and why another factor is affecting it compared to others.

3. Factor analysis has yielded that there is one factor which only affects activity and as discussed above this may be the "cognitive decline" factor. Among the other variables, one factor is the "Digestive" factor (one of the side effects of radion is ulceration and stomach sores). There is another factor affecting sleep and number of symptoms. This may be the "patient" factor - how each patient is responding to treatment.

Acknowledgement

We would like to thank our professor Dr. Swagata Nandi for suggesting this dataset and guiding us through this project. We would also like to thank our batchmates for their advices and criticisms.

References

1. Applied Multivariate Statistical Analysis - Richard Johnson
2. A new family of power transformations to improve normality or symmetry - IN-KWON YEO and RICHARD A. JOHNSON.