MA3505 Multivariate Statistics Project 1

April 29, 2016

1 Introduction and exploratory data analysis for the variables.

The heart disease data set is a study into heart disease diagnoses from four different locations:

1. Cleveland
2. Hungary
3. Long Beach
4. Switzerland

In this report we will be looking at how the means of certain variables are affected by exercise protocols; multivariate regression to estimate the coefficients for regressing variables against each other; dimension reduction see which of the variables affect the diagnosis of the heart disease and discriminant analysis/classification tools to distinguish patients based on their disease.

Each data set has 75 variables (76 as an indicator variable is added to tell the location) and all are numerical. Each location has the following amount of observations:

1. 282
2. 294
3. 200
4. 123

Which gives a total of 899 observations overall.
Using the ***summary()*** function on each loction firstly all were missing the variables **pncaden, restckm, exerckm, restef, restwm, exeref, exerwm, earlobe.**

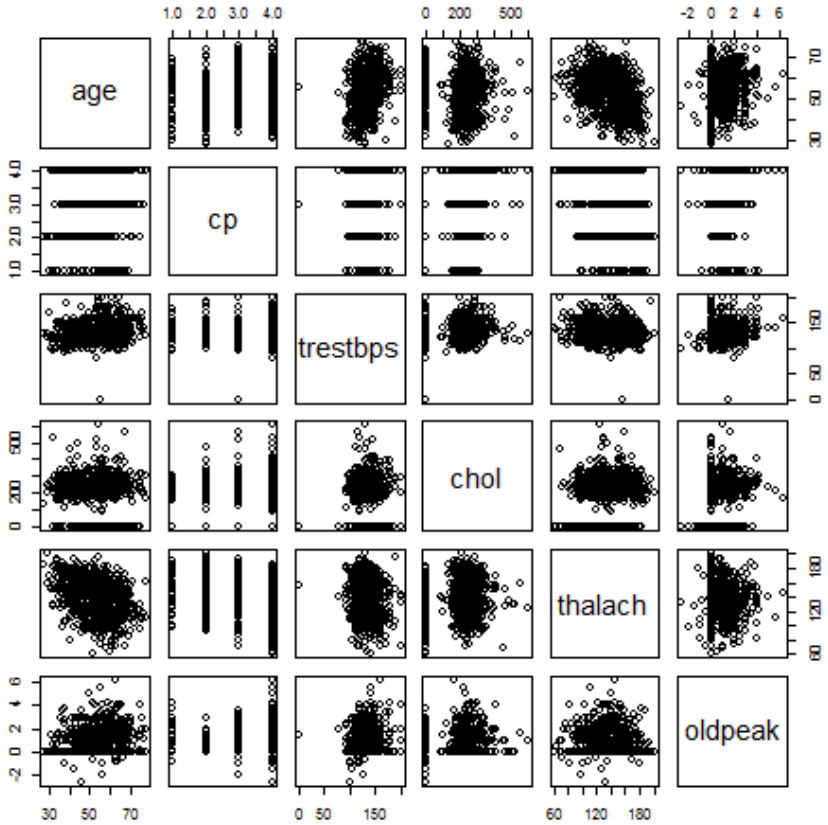
For the exploratory data analysis I will be looking at the following variables:

- **age**
- **sex**
- **cp**
- **trestbps**
- **chol**
- **fbs**
- **restecg**
- **thalech**
- **exang**
- **oldpeak**
- **slope**
- **num**

First we see the correlation matrix for these variables for the whole data set.

	age	sex	cp	trestbps	chol
age	1.00000000	0.04013501	0.19043423	0.24734252	−0.07139191
sex	0.04013501	1.00000000	0.16081260	0.01754740	−0.17804228
cp	0.19043423	0.16081260	1.00000000	0.02197548	−0.07592894
trestbps	0.24734252	0.01754740	0.02197548	1.00000000	0.06410291
chol	−0.07139191	−0.17804228	−0.07592894	0.06410291	1.00000000
fbs	0.22162138	0.07575863	0.01275136	0.14763192	0.03662188
restecg	0.20639095	−0.02138613	0.05094178	0.06837272	0.08643970
thalach	−0.36895303	−0.17341628	−0.35564428	−0.11877387	0.19819530
exang	0.23726947	0.20464501	0.43430919	0.18251305	−0.06214659
oldpeak	0.24953164	0.11323773	0.24103121	0.18546162	0.05980302
num	0.27829623	0.26020650	0.39818386	0.19109744	−0.08450473
	fbs	restecg	thalach	exang	oldpeak
age	0.22162138	0.20639095	−0.36895303	0.23726947	0.24953164
sex	0.07575863	−0.02138613	−0.17341628	0.20464501	0.11323773
cp	0.01275136	0.05094178	−0.35564428	0.43430919	0.24103121
trestbps	0.14763192	0.06837272	−0.11877387	0.18251305	0.18546162
chol	0.03662188	0.08643970	0.19819530	−0.06214659	0.05980302
fbs	1.00000000	0.09991019	−0.04765425	0.04438154	0.04279935
restecg	0.09991019	1.00000000	0.05346434	0.03965127	0.10585235
thalach	−0.04765425	0.05346434	1.00000000	−0.38887336	−0.18117607
exang	0.04438154	0.03965127	−0.38887336	1.00000000	0.41549843
oldpeak	0.04279935	0.10585235	−0.18117607	0.41549843	1.00000000
num	0.10921234	0.09396999	−0.38752927	0.47852707	0.48149175
	num				
age	0.27829623				
sex	0.26020650				
cp	0.39818386				
trestbps	0.19109744				
chol	−0.08450473				
fbs	0.10921234				
restecg	0.09396999				
thalach	−0.38752927				
exang	0.47852707				
oldpeak	0.48149175				
num	1.00000000				

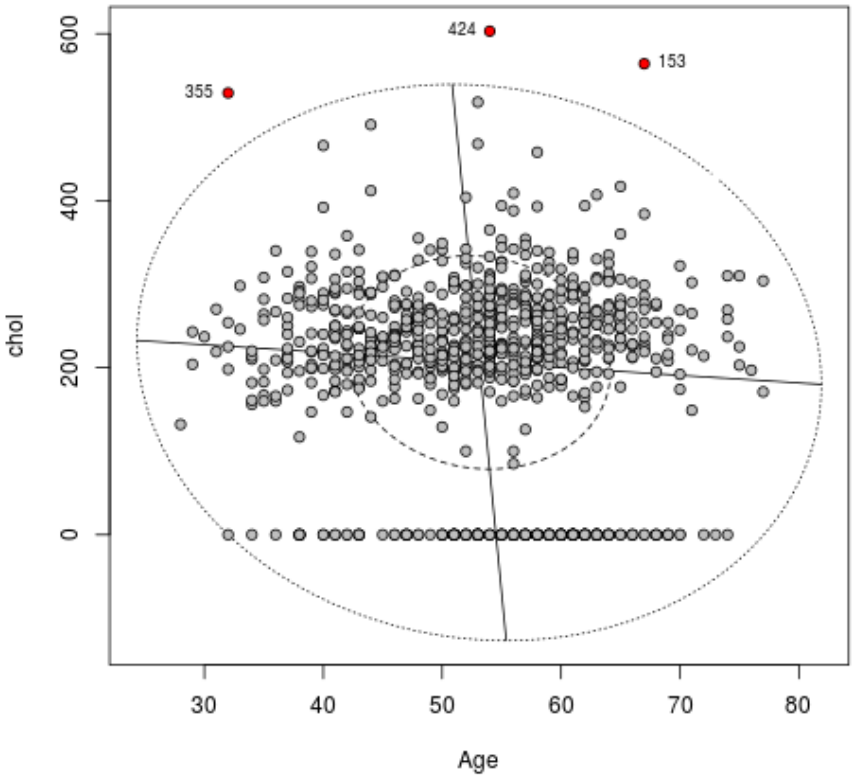
We can see hear that there are no highly correlated variables, naturally one would think age would have an effect on some of the variables.Below shows a large scatterplot matrix for 6 variables that take on more than 2 values.



files/intplot2.png

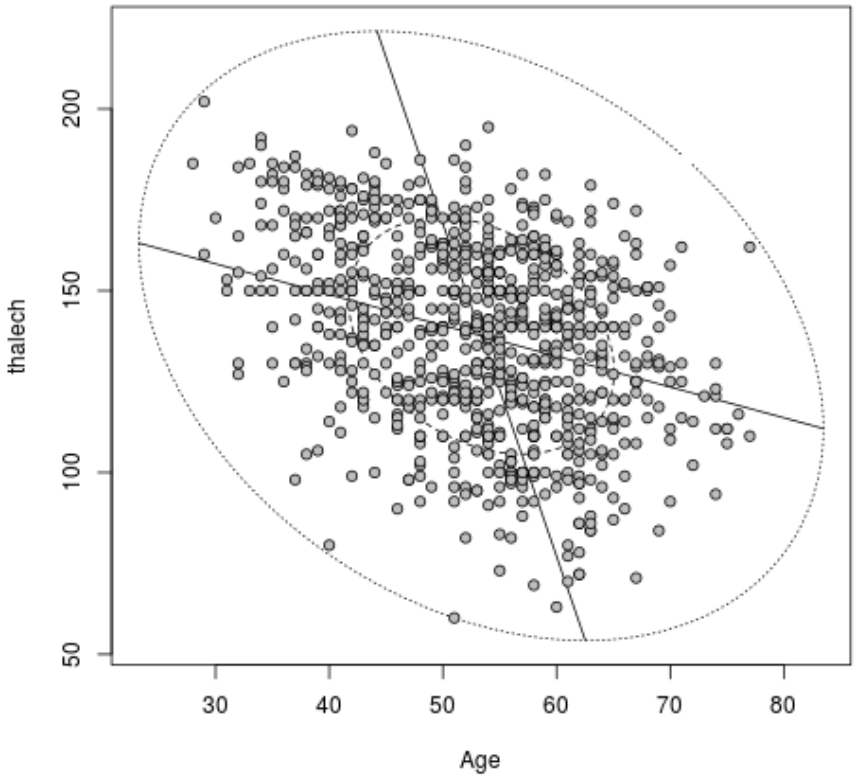
Figure 1: Scatterplot matrix

To show what the data looks like further below shows bivariate boxplots of **age** plotted against **chol**(cholestrol) and **thalech**(maximum heart rate achieved).



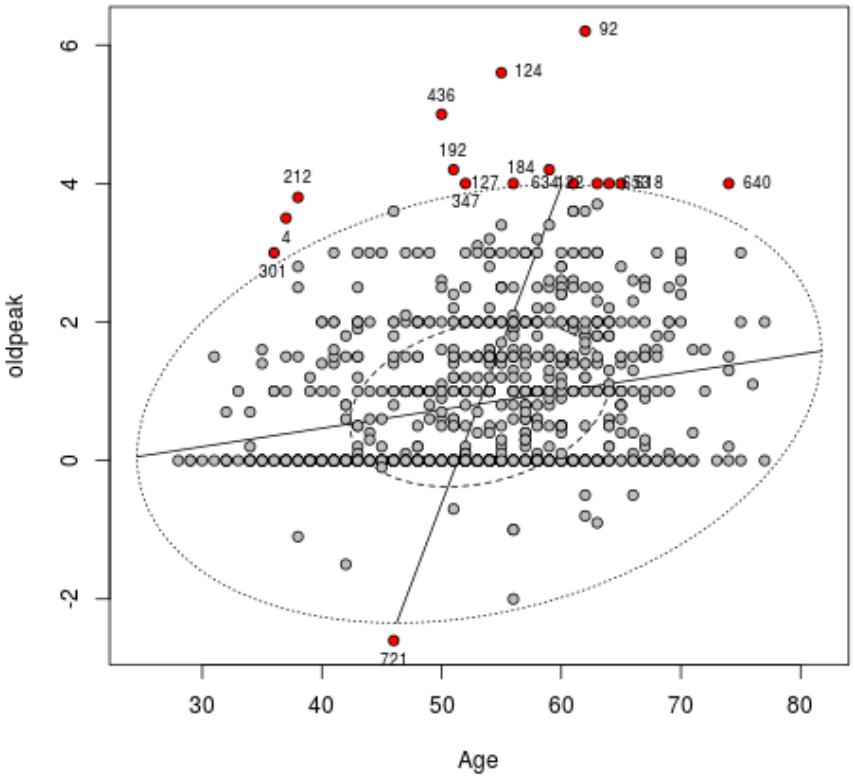
files/intplot3.png

Figure 2: Bivariate boxplot: Age against Chol



files/intplot4.png

Figure 3: Bivariate boxplot: Age against thalech



files/intplot5.png

Figure 4: Bivariate boxplot: Age against Oldpeak

In the next chapter we can find out how the variables relate to each other in more detail.

2 Analysis to answer each research question

2.1 Question 1

For this question since Cleveland and Switzerland only used one exercise protocol each rather than looking at the four data sets separately, it was looked at as one data set.Also, it appeared that the protocol values for Hungary did not match the description in the brief therefore the values were changed to:

$$150 = 7,$$
$$100 = 9,$$
$$50 = 11,$$
$$25 = 11,$$
$$175 = 7,$$
$$125 = 8,$$
$$75 = 10,$$
$$130 = 8,$$
$$200 = \text{NA}.$$

Each variable was assigned the number that represented the value to the nearest 25.

Next I created a subset of the data set with just the required variables, the variables rldv5, rldv5e and met were removed as the first had no values for Cleveland and Switzerland and the latter two had most values missing for Switzerland. So the code for the subset was

```
allproto1 = subset(all.df, select=c(proto,chol,
thaldur, thaltime, thalach, thalrest, tpeakbps, tpeakbpd, trestbpd,
oldpeak, indictor))
```

While running a couple of preliminary tests a lot of errors were occurring, therefore checking through each exercise protocol, protocols 4,6 and 7 had very few observations and protocol 12 had dependent columns, thus they were removed from the test.

So that the tests would run the protocol classes will be renamed to:

$$8 = 2,$$
$$9 = 3,$$
$$10 = 4,$$
$$11 = 5.$$

2.1.1 Comparison of covariance matrices

To compute the mean tests first need to compute a hypotheses test to test equality of covariance matrices.For each we are testing

$$H_0 : \Sigma_A = \Sigma_{p_i} \; vs \; H_A : not \; H_0$$

The tests give the following output

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

data:  allproto1cc[, 2:9]
Chi-Sq (approx.) = 309.63, df = 36, p-value < 2.2e-16
```

Exercise protocol 1(Bruce)
p-value: $< 2.2 \times 10^{-16}$
Decision: Reject the null
Conclusion: There is statistical significant evidence covariance matrices are differen.

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

data:  allproto1cc[, 2:9]
Chi-Sq (approx.) = 75.591, df = 36, p-value = 0.0001247
```

Exercise protocol 2(8)(bike 125 kpa min/min)
p-value: < 0.0001247
Decision: Reject the null
Conclusion: There is statistical significant evidence covariance matrices are different.

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

data:  allproto1cc[, 2:9]
Chi-Sq (approx.) = 77.734, df = 36, p-value = 6.739e-05
```

Exercise protocol 3(9)(bike 100 kpa min/min)
p-value: $< 6.739 \times 10^{-5}$
Decision: Reject the null
Conclusion: There is statistical significant evidence covariance matrices are different.

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

data:  allproto1cc[, 2:9]
Chi-Sq (approx.) = 167.95, df = 36, p-value < 2.2e-16
```

Exercise protocol 4(10)(bike 75 kpa min/min)
p-value: $< 2.2 \times 10^{-16}$
Decision: Reject the null
Conclusion: There is statistical significant evidence covariance matrices are different.

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

data:  allproto1cc[, 2:9]
Chi-Sq (approx.) = 158.45, df = 36, p-value < 2.2e-16
```

Exercise protocol 5(10)(bike 50 kpa min/min)
p-value: $< 2.2 \times 10^{-16}$
Decision: Reject the null
Conclusion: There is statistical significant evidence covariance matrices are different.

All tests we reject the null so for the equality of mean tests we can use the James test.

2.1.2 Equality of mean tests

Running the Multivariate James returns:

```
$test
[1] 1972.425

$correction
[1] 1.421495

$corrected.critical.value
[1] 65.6649

$p.value
[1] 0
```

$H_0 : \mu_{p_1} = \mu_{p_2} = \mu_{p_3} = \mu_{p_4} = \mu_{p_5}$ *vs* H_A : *not* H_0
p-value: < 0
Decision: Reject the null
Conclusion: There is statistical significant evidence that the mean for each exercise protocol is different.

To take a closer look, I conducted some two sample James tests to see if the means for any of the exercise protocols were equal. From the results we see that again there is statistical significant evidence that no means are equal to each other but what is interesting to see firstly for the James test R =1 all tests had a zero p-value other than the protocols 2 and 3, and 3 and 4, (tests shown below). This is not surprising as these are the protocols performed on bikes (with the highest kpa's).

```
$test
[1] 65.9968

$correction
[1] 1.454695

$corrected.critical.value
[1] 22.55842

$p.value
[1] 3.133546e-07
```

$H_0 : \mu_{p_2} = \mu_{p_3}$ *vs* H_A : *not* H_0

```
$test
[1] 132.0527

$correction
[1] 1.406772

$corrected.critical.value
[1] 21.81525

$p.value
[1] 1.110223e-16
```

$H_0 : \mu_{p_3} = \mu_{p_4}$ *vs* H_A : *not* H_0

Finally, when conducting the R = 2 James test again all tests had p-value's that give evidence to reject the null but testing betweens protocols 1 and 5 gave an answer of zero.

```
$test
[1] 23.89011

$critical
[1] 2.042361

$df1
[1] 8

$df2
[1] 88.52957

$p.value
[1] 0
```

$H_0 : \mu_{p_1} = \mu_{p_5}$ *vs* H_A : *not* H_0

2.2 Question 2

```
model2 = lm(cbind(chol, thaldur, thaltime, met, thalach, thalrest, tpeakbps,
                  tpeakbpd, trestbpd, oldpeak, rldv5, rldv5e) ~ proto + restecg + dig
                  + prop + nitr + pro + diuretic, data=datall)
```

Using the above code I created the necessary multivariate regression model. I was able to use this model to get the following table of coefficients:

	chol	thaldur	thaltime	met	thalach
(Intercept)	2.182e+02	2.964e+00	1.941e+00	4.167e+00	1.216e+02
proto	5.933e−01	7.621e−02	7.395e−02	1.284e−02	1.241e−01
restecg	−1.965e+01	1.506e−01	4.325e−01	1.305e−01	−2.202e−01
dig	6.033e+00	3.130e+00	2.889e+00	1.519e+00	−8.249e+00
prop	1.800e+01	3.734e−01	6.787e−01	4.875e−02	−6.759e+00
nitr	−1.390e+01	−3.582e−01	−3.903e−01	3.416e−02	−5.949e+00
pro	−6.872e+01	1.142e+00	9.582e−01	4.637e−01	1.990e−02
diuretic	−4.914e+01	1.516e+00	6.732e−01	3.744e−01	1.610e+01
	thalrest	tpeakbps	tpeakbpd	trestbpd	oldpeak
(Intercept)	7.475e+01	1.607e+02	9.326e+01	8.488e+01	1.937e+00
proto	4.602e−02	2.114e−01	3.306e−02	1.262e−02	−3.094e−03
restecg	1.481e+00	3.762e+00	−1.245e+00	1.159e+00	−2.081e−01
dig	2.175e+00	−7.984e+00	−1.854e+01	−4.949e+00	4.202e−01
prop	−2.692e−01	5.788e−02	−1.988e+00	5.123e−01	−1.674e−02
nitr	−8.676e+00	−9.099e+00	−3.690e+00	−3.270e+00	2.621e−01
pro	2.958e+00	4.851e+00	7.011e+00	2.962e−01	−8.122e−01
diuretic	−8.346e−01	6.602e+00	2.153e+00	1.116e+00	−2.439e−02
	rldv5	rldv5e			
(Intercept)	1.487e+01	1.497e+01			
proto	−6.571e−04	−6.529e−03			
restecg	1.703e−01	2.203e−01			
dig	−2.153e+00	−2.219e+00			
prop	1.272e+00	1.175e+00			
nitr	6.343e−01	−6.043e−01			
pro	−1.583e+00	7.800e−01			
diuretic	−1.303e−01	3.239e+00			

However this is not very useful, so I used the **summary()** function to enable me to achieve a more detailed view of my analysis. Below I have tried my best to explain the detailed view for each response variable.

```
Response chol :
Call:
lm(formula = chol ~ proto + restecg + dig + prop + nitr + pro +
    diuretic, data = datall)

Residuals:
     Min       1Q   Median       3Q      Max
-221.153  -37.934   -0.852   55.190  310.650

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 218.1866    16.6717  13.087  < 2e-16 ***
proto         0.5933     0.1884   3.149  0.00207 **
restecg     -19.6467    15.5463  -1.264  0.20877
dig           6.0327    42.3211   0.143  0.88689
prop         17.9968    25.2562   0.713  0.47750
nitr        -13.8953    24.8506  -0.559  0.57710
pro         -68.7201    27.9126  -2.462  0.01524 *
diuretic    -49.1356    33.0596  -1.486  0.13983
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  1

Residual standard error: 85.12 on 120 degrees of freedom
  (771 observations deleted due to missingness)
Multiple R-squared:  0.2402,    Adjusted R-squared:  0.1958
F-statistic: 5.418 on 7 and 120 DF,  p-value: 2.026e-05
```

From the table above we can see that the predictor that had the most affect in the value of the **chol** response was **proto**. As *chol* refers to the amount of cholesterol in a person's system and *proto* refers to the type of exercise that they do, it is not a major surprise that this is the most important as in theory the higher the intensity of the your exercise program the lower your cholesterol will be. The second most important variable is **pro**; this is an indicator variable that tells us if someone uses *calcium channel blocker used during exercise* (it is used in cholesteryl ester hydrolysis which helps reduce cholesterol) during their exercise routine.

```
Response thaldur :
Call:
lm(formula = thaldur ~ proto + restecg + dig + prop + nitr +
    pro + diuretic, data = datall)

Residuals:
   Min     1Q Median     3Q    Max
-4.312 -1.681 -0.310  1.422  6.440

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.964440   0.443361   6.686 7.65e-10 ***
proto        0.076214   0.005011  15.209  < 2e-16 ***
restecg      0.150581   0.413433   0.364   0.7163
dig          3.129630   1.125473   2.781   0.0063 **
prop         0.373380   0.671654   0.556   0.5793
nitr        -0.358181   0.660868  -0.542   0.5888
pro          1.141536   0.742300   1.538   0.1267
diuretic     1.516199   0.879177   1.725   0.0872 .
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 2.264 on 120 degrees of freedom
  (771 observations deleted due to missingness)
Multiple R-squared:  0.6909,    Adjusted R-squared:  0.6728
F-statistic: 38.31 on 7 and 120 DF,  p-value: < 2.2e-16
```

The predictor variable in this instance is **thaldur** which represents the length of time a person spends on an exercise test, it is therefore no surprise that **proto** is the most important predictor as the harder the exercise test the less time you will be able to do it for. The second most significant predictor **dig** refers to whether or not the person is taking a drug called *digitails* during exercise. Studies have shown that the use of this drug during exercise increases blood flow which could allow someone to exercise for longer *(experts are not sure if it is a performance enhancing drug as trial results vary).*

```
Response thaltime :

Call:
lm(formula = thaltime ~ proto + restecg + dig + prop + nitr +
    pro + diuretic, data = datall)

Residuals:
   Min     1Q Median     3Q    Max
-4.469 -1.639 -0.139  1.053  7.352

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.941466   0.449229   4.322 3.21e-05 ***
proto        0.073951   0.005077  14.565  < 2e-16 ***
restecg      0.432490   0.418906   1.032   0.3039
dig          2.888715   1.140370   2.533   0.0126 *
prop         0.678710   0.680544   0.997   0.3206
nitr        -0.390289   0.669615  -0.583   0.5611
pro          0.958162   0.752125   1.274   0.2051
diuretic     0.673187   0.890814   0.756   0.4513
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1

Residual standard error: 2.294 on 120 degrees of freedom
  (771 observations deleted due to missingness)
Multiple R-squared:  0.6704,    Adjusted R-squared:  0.6511
F-statistic: 34.86 on 7 and 120 DF,  p-value: < 2.2e-16
```

thaltime refers to the time at which a person's ST depression was measured. It is therefore no surprise that **proto** has the highest effect as different exercises will take different amount of times to complete meaning that if *thaltime* is always measured at the end of the exercise test people who do different tests will have different times but those who take the same test should have very similar times. **dig** is the next significant variable which sort of makes sense as you most likely have to wait for the drug to leave your system before your ST depression can be measured.

```
Response met :
Call:
lm(formula = met ~ proto + restecg + dig + prop + nitr + pro +
    diuretic, data = datall)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7325 -1.0919 -0.1298  0.8792  5.8206

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.166578   0.336945  12.366   <2e-16 ***
proto       0.012843   0.003808   3.372   0.0010 **
restecg     0.130481   0.314201   0.415   0.6787
dig         1.518904   0.855338   1.776   0.0783 .
prop        0.048745   0.510444   0.095   0.9241
nitr        0.034160   0.502247   0.068   0.9459
pro         0.463725   0.564133   0.822   0.4127
diuretic    0.374352   0.668158   0.560   0.5763
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 1.72 on 120 degrees of freedom
  (771 observations deleted due to missingness)
Multiple R-squared:  0.1108,    Adjusted R-squared:  0.05892
F-statistic: 2.136 on 7 and 120 DF,  p-value: 0.04484
```

The predictor **met** refers to the *metabolic equivalent of resting oxygen consumption while sitting* and therefore it is not much of a surprise that the response **proto** is the most significant. It is also not that surprising that it is as significant as before, as the trial that produced these results most likely used people of varying athletic abilities for each test in order to make the results more accurate.

```
Response thalach :

Call:
lm(formula = thalach ~ proto + restecg + dig + prop + nitr +
    pro + diuretic, data = datall)

Residuals:
    Min      1Q  Median      3Q     Max
-42.497 -10.060  -0.925  13.735  53.075

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 121.6141     3.6895  32.962  < 2e-16 ***
proto         0.1241     0.0417   2.977  0.00352 **
restecg      -0.2202     3.4405  -0.064  0.94908
dig          -8.2489     9.3659  -0.881  0.38022
prop         -6.7587     5.5893  -1.209  0.22896
nitr         -5.9491     5.4996  -1.082  0.28154
pro           0.0199     6.1772   0.003  0.99743
diuretic     16.0994     7.3163   2.200  0.02969 *
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 18.84 on 120 degrees of freedom
  (771 observations deleted due to missingness)
Multiple R-squared:  0.1867,    Adjusted R-squared:  0.1392
F-statistic: 3.934 on 7 and 120 DF,  p-value: 0.0006723
```

The predictor **thalach** refers to the maximum heart rate that a person achieves during their exercise test and as such it is no surprise that the response variable that is the most significant when calculating it is **proto**. This is because the more intense the exercise test is the more oxygen your body is going to need thus you will have a higher heart rate. Again it is not surprising that *proto* is only a 2* rather than a 3* significance level as your maximum heart rate will depend on how athletic you are, the more athletic the lower your max heart rate will be. **diuretic** is the other significant response variable and it refers to whether or not the subject uses diuretic used during exercise. Diuretic is considered to be a performance enhancing drug so it is therefore no surprise that it only has a 1* significance level due to the fact that the analysis up to now has shown that there is a high probability that athletes are involved in this trial and would be band by WADA if they were caught using it.

```
Response thalrest :

Call:
lm(formula = thalrest ~ proto + restecg + dig + prop + nitr +
    pro + diuretic, data = datall)

Residuals:
    Min      1Q  Median      3Q     Max
-28.204  -8.542  -1.909   8.172  55.796

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 74.75201    2.60040  28.746   <2e-16 ***
proto        0.04602    0.02939   1.566    0.120
restecg      1.48140    2.42487   0.611    0.542
dig          2.17533    6.60112   0.330    0.742
prop        -0.26923    3.93939  -0.068    0.946
nitr        -8.67576    3.87612  -2.238    0.027 *
pro          2.95844    4.35374   0.680    0.498
diuretic    -0.83465    5.15655  -0.162    0.872
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1

Residual standard error: 13.28 on 120 degrees of freedom
  (771 observations deleted due to missingness)
Multiple R-squared:  0.09876,   Adjusted R-squared:  0.04619
F-statistic: 1.879 on 7 and 120 DF,  p-value: 0.07885
```

The **thalrest** variable refers to the subjects resting heart rate and the only variable that has any significant effect on the outcome of this result is **nitr** which tells us whether or not the subject uses nitrates used during their exercise. I am not quite sure what the use of nitrates has to do with the resting heart rates but I do know that they are added to 'unhealthy foods' such as *bacon, sandwich meats and salami* which could indicate that they are not very athletic but a high resting heart does not mean that someone is less athletic.

In this trial the subjects the measuring of their peak blood pressure was split into two different variables: **tpeakbps** and **tpeakbpd**, google wasn't able to explain why this is the case.

```
Response tpeakbps :

Call:
lm(formula = tpeakbps ~ proto + restecg + dig + prop + nitr +
    pro + diuretic, data = datall)

Residuals:
   Min     1Q Median     3Q    Max
-46.56 -15.18  -2.97  13.36  58.73

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 160.70295    4.33227  37.094  < 2e-16 ***
proto         0.21138    0.04897   4.317 3.27e-05 ***
restecg       3.76224    4.03984   0.931    0.354
dig          -7.98425   10.99749  -0.726    0.469
prop          0.05788    6.56303   0.009    0.993
nitr         -9.09857    6.45763  -1.409    0.161
pro           4.85054    7.25334   0.669    0.505
diuretic      6.60213    8.59083   0.769    0.444
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1

Residual standard error: 22.12 on 120 degrees of freedom
  (771 observations deleted due to missingness)
Multiple R-squared:  0.1992,    Adjusted R-squared:  0.1525
F-statistic: 4.266 on 7 and 120 DF,  p-value: 0.0003059
```

For the variable that had the most significant affect on **tpeakbps** was (as normal it seems in this trial) **proto**. This is most likely because of the fact that exercise can lower your blood pressure and therefore the subjects that are able to take the more intensive exercise tests were likely to have a lower peak blood pressure.

```
Response tpeakbpd :

Call:
lm(formula = tpeakbpd ~ proto + restecg + dig + prop + nitr +
    pro + diuretic, data = datall)

Residuals:
    Min      1Q  Median      3Q     Max
-60.687  -7.517  -0.023   8.638  36.329

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  93.26322    2.68207  34.773  < 2e-16 ***
proto         0.03306    0.03031   1.091  0.27768
restecg      -1.24519    2.50103  -0.498  0.61948
dig         -18.54131    6.80844  -2.723  0.00743 **
prop         -1.98759    4.06311  -0.489  0.62561
nitr         -3.69032    3.99786  -0.923  0.35782
pro           7.01102    4.49047   1.561  0.12108
diuretic      2.15344    5.31850   0.405  0.68627
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1

Residual standard error: 13.69 on 120 degrees of freedom
  (771 observations deleted due to missingness)
Multiple R-squared:  0.1321,    Adjusted R-squared:  0.0815
F-statistic:  2.61 on 7 and 120 DF,  p-value: 0.01525
```

The response variable that was most significant when working out the predictor **tpeakbpd** was **dig**. This makes sense as studies have shown that the use of the drug digitalis during exercise lowers a person's blood pressure.

```
Response trestbpd :
Call:
lm(formula = trestbpd ~ proto + restecg + dig + prop + nitr +
    pro + diuretic, data = datall)

Residuals:
    Min      1Q  Median      3Q     Max
-35.510  -6.141  -1.298   5.543  24.175

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 84.87854    1.88379  45.057   <2e-16 ***
proto        0.01262    0.02129   0.593    0.554
restecg      1.15852    1.75663   0.660    0.511
dig         -4.94866    4.78200  -1.035    0.303
prop         0.51233    2.85378   0.180    0.858
nitr        -3.27042    2.80795  -1.165    0.246
pro          0.29623    3.15394   0.094    0.925
diuretic     1.11644    3.73552   0.299    0.766
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 9.618 on 120 degrees of freedom
  (771 observations deleted due to missingness)
Multiple R-squared:  0.0366,    Adjusted R-squared:  -0.01959
F-statistic: 0.6514 on 7 and 120 DF,  p-value: 0.7126
```

The predictor variable **trestbpd** refers to the subjects resting blood pressure. As this must be taken before any exercise is started it makes sense that none of the responses are significant in determining what this value shall be due to them being manly related to the exercise test the subject takes.

```
Response oldpeak :

Call:
lm(formula = oldpeak ~ proto + restecg + dig + prop + nitr +
    pro + diuretic, data = datall)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9343 -0.6280 -0.0506  0.3642  3.5801

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.937386   0.169250  11.447  < 2e-16 ***
proto       -0.003094   0.001913  -1.617  0.10841
restecg     -0.208133   0.157825  -1.319  0.18976
dig          0.420187   0.429642   0.978  0.33004
prop        -0.016737   0.256399  -0.065  0.94806
nitr         0.262057   0.252282   1.039  0.30101
pro         -0.812187   0.283368  -2.866  0.00491 **
diuretic    -0.024390   0.335620  -0.073  0.94219
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1

Residual standard error: 0.8642 on 120 degrees of freedom
  (771 observations deleted due to missingness)
Multiple R-squared:  0.1081,    Adjusted R-squared:  0.05603
F-statistic: 2.077 on 7 and 120 DF,  p-value: 0.0511
```

The predictor variable **oldpeak** refers to *ST depression induced by exercise relative to rest* (which I understand from google to be a fancy way of saying that the subject gets a small heart attack during exercise). It makes sense then that the most significant variable in deciding what the value of which if it is high can cause heart attacks. *oldpeak* is going to be is **pro** as helps to lower cholesterol

The next two predictors, **rldv5** and **rldv5e**, refer to *height at rest* and *height at peak exercise*. I don't know what *height* they are referring to (I am assuming it is not just how tall they are as that would be dull to measure at rest and during peak exercise as it would not change) and luckily none of the response variables are significant in working out what the values of the variables will be.

```
Response rldv5 :

Call:
lm(formula = rldv5 ~ proto + restecg + dig + prop + nitr + pro +
    diuretic, data = datall)

Residuals:
     Min       1Q   Median       3Q      Max
-10.7927  -3.3161  -0.7927   3.0914  16.1580

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.8748145  1.0543301  14.108   <2e-16 ***
proto       -0.0006571  0.0119165  -0.055    0.956
restecg      0.1703404  0.9831619   0.173    0.863
dig         -2.1530704  2.6764216  -0.804    0.423
prop         1.2718279  1.5972216   0.796    0.427
nitr         0.6342939  1.5715709   0.404    0.687
pro         -1.5831511  1.7652196  -0.897    0.372
diuretic    -0.1302669  2.0907202  -0.062    0.950
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1

Residual standard error: 5.383 on 120 degrees of freedom
  (771 observations deleted due to missingness)
Multiple R-squared:  0.016,     Adjusted R-squared:  -0.0414
F-statistic: 0.2787 on 7 and 120 DF,  p-value: 0.9612
```

```
Response rldv5e :

Call:
lm(formula = rldv5e ~ proto + restecg + dig + prop + nitr + pro +
    diuretic, data = datall)

Residuals:
     Min       1Q   Median       3Q      Max
-11.1533  -3.5409  -0.4798   2.7664  14.0371

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.969428   1.048021  14.284   <2e-16 ***
proto       -0.006529   0.011845  -0.551    0.583
restecg      0.220336   0.977279   0.225    0.822
dig         -2.219224   2.660406  -0.834    0.406
prop         1.174742   1.587664   0.740    0.461
nitr        -0.604272   1.562167  -0.387    0.700
pro          0.779970   1.754657   0.445    0.657
diuretic     3.238914   2.078210   1.559    0.122
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1

Residual standard error: 5.351 on 120 degrees of freedom
  (771 observations deleted due to missingness)
Multiple R-squared:  0.03959,    Adjusted R-squared:  -0.01643
F-statistic: 0.7067 on 7 and 120 DF,  p-value: 0.6663
```

2.3 Question 3

Due to that each dataset is missing different variables from the data, we have decided that in order to maximise the amount of variables we have, we are going to be using each dataset independent of the others.

For each dataset we removed the dummy variables and variables that were missing at least a percentage of data. This percent was different for each data set and we were aiming for approximate at least double the number of observations to the number of variables.

In addition when listing the PCA loadings for each components, we have only included the listings that were required for identify the variables in the appropriate number of principle components.

2.3.1 Cleveland

After removing dummy variables and variables with at least 90% NA data, we are left with 45 variables and 201 observations.

Cleveland variance inflation factor

age	sex	cp	trestbps	htn	chol	cigs	years
2.070591	2.379469	1.683710	2.935706	1.734144	1.326342	2.346224	2.315459
fbs	famhist	restecg	ekgmo	ekgday	ekgyr	dig	prop
1.281244	1.291443	1.338021	14.903816	3.357399	78.992867	1.296383	1.679766
nitr	pro	diuretic	thaldur	thaltime	met	thalach	thalrest
1.546570	1.415979	1.480903	9.549788	1.422540	10.328475	2.868773	1.713892
tpeakbps	tpeakbpd	trestbpd	exang	xhypo	oldpeak	slope	rldv5e
2.829387	2.173463	2.785971	1.734917	1.870852	2.831028	2.291928	1.557587
ca	thal	cmo	cday	cyr	lmt	ladprox	laddist
1.841289	2.051953	15.389866	3.413846	80.511913	1.401270	1.496650	1.526869
cxmain	om1	rcaprox	rcadist				
1.543251	1.789705	1.764053	1.835745				

From the variance inflation factor we see the variables **ekgmo**, **ekgyr**, **cmo** and **cyr** are highly collinear with other variables in the model.

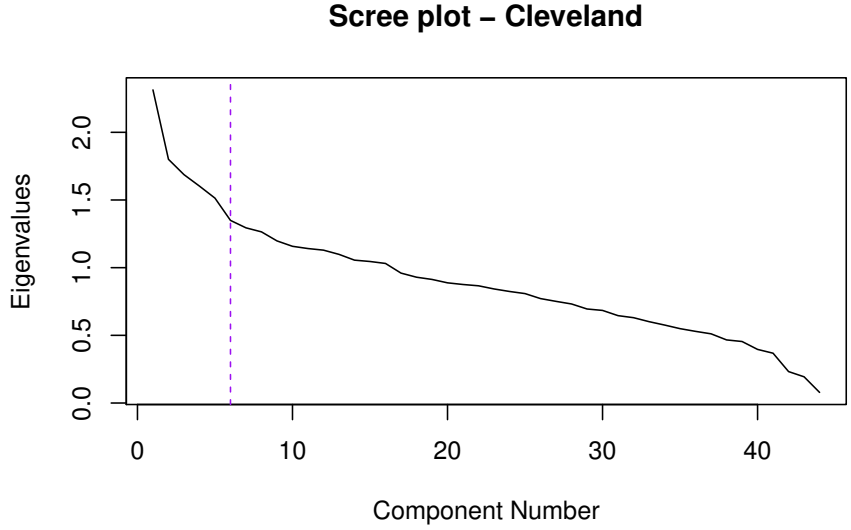


Figure 5: Scree plot for PCA of Cleveland

From the scree plot in Figure 5 we see that we keep 6 components.

We have the loadings of each components as follows.

Cleveland PCA loadings

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
age	0.192		−0.196			0.167		−0.122	0.163	0.372
sex		−0.195	0.306	0.193					−0.303	
cp	0.208							0.384		−0.116
trestbps	0.133	−0.144	−0.297	0.222	0.107	0.119		−0.149		
htn			0.222		−0.189	0.390		0.117		
chol			−0.184					0.184	0.213	0.222
cigs		−0.200	0.181	0.231	−0.292		−0.128		−0.214	
years		−0.189	0.145	0.223	−0.330	0.138			−0.156	
fbs			−0.128	0.143		−0.214	0.132			0.129
famhist					0.123	0.140		0.133	0.162	−0.136
restecg		−0.103	−0.132				−0.128	0.238		
ekgmo		−0.244		−0.433	−0.109		0.220	−0.161	−0.144	
ekgday					0.384	0.326	0.255	0.298	−0.109	
ekgyr		0.414		0.193			0.268		−0.129	−0.212
dig		0.105				0.195	−0.112	−0.150	−0.230	
prop	0.102	0.107			0.162	−0.263	−0.247	0.105		0.173
nitr	0.142	0.107		−0.128			−0.141		−0.180	0.131
pro		0.236				−0.115		0.154	−0.222	
diuretic					0.128	−0.417			−0.136	0.199
thaldur	−0.301	−0.109	0.237	0.125	0.184	−0.149				
thaltime			0.154		0.153			−0.191	0.359	−0.189
met	−0.295	−0.137	0.228	0.135	0.181	−0.167				
thalach	−0.298	−0.172					0.130	0.101	0.106	−0.141
thalrest			−0.229		−0.254		0.221	0.145	0.194	−0.113
tpeakbps		−0.211	−0.236	0.297				−0.211		
tpeakbpd		−0.167	−0.330	0.142			−0.128	0.161	−0.148	−0.159
trestbpd		−0.222	−0.314	0.112	0.130		−0.106		−0.167	−0.194
exang	0.224			−0.100					−0.207	−0.157
xhypo	0.104	0.153		−0.229		−0.102			−0.113	
oldpeak	0.280				0.185			−0.219	0.150	−0.159
slope	0.232				0.230		−0.120	−0.217		−0.263
rldv5e				0.126	0.127	0.166		−0.295		
ca	0.213			0.113	−0.124		0.283		0.211	0.251
thal	0.231	−0.163	0.167	0.102					−0.143	
cmo		−0.243		−0.433	−0.116		0.209	−0.162	−0.122	
cday					0.391	0.294	0.280	0.243	−0.135	
cyr		0.415		0.195			0.261		−0.130	−0.218
lmt	0.130	−0.106					0.132			−0.126
ladprox	0.183		0.147				−0.234	0.150	0.132	
laddist	0.206		0.107			−0.114	0.254			
cxmain	0.189		0.150	0.104					0.111	0.201
om1	0.249					−0.108	0.202			
rcaprox	0.191					−0.237	0.162	0.151	0.181	−0.291
rcadist	0.196		0.103				0.183	−0.130		0.250

We see that the first principle component is mostly formed of **thaldur**, **thalach**, **met** and **oldpeak** variables.

The second principle component is mostly formed of **cyr** and **ekgyr** variables.

The third principle component is mostly formed of **tpeakbpd**, **trestbpd**, **sex** and **trestbps** variables.

The fourth principle component is mostly formed of **ekgmo** and **cmo** variables.

The fifth principle component is mostly formed of **cday**, **ekgday**, **years** and **cigs** variables.

The sixth principle component is mostly formed of **diuretic**, **htn**, **ekgday**

2.3.2 Hungary

After removing dummy variables and variables with at least 79% NA data, we are left with 36 variables and 88 observations.

Hungary variance inflation factor

age	sex	painloc	painexer	relrest	cp	trestbps
2.434590	2.080449	2.654746	10.046548	6.565678	19.690793	4.275217
htn	chol	fbs	restecg	ekgmo	ekgday	ekgyr
1.861599	2.510843	2.038199	1.494730	26.201141	3.202566	179.069024
prop	nitr	pro	diuretic	proto	thaldur	thaltime
4.687494	8.190981	9.216954	3.508473	40.100317	160.858151	159.083733
met	thalach	thalrest	tpeakbps	tpeakbpd	trestbpd	exang
8.256667	3.592979	1.974234	3.419070	3.548859	3.556353	3.194635
oldpeak	slope	rldv5	rldv5e	cmo	cday	cyr
2.059364	2.928694	9.159900	8.276601	26.609872	2.619470	173.894964

From the variance inflation factor we see that the variables **painexer**, **cp**, **ekgmo**, **ekgyr**, **proto**, **thaldur**, **thaltime**, **cmo** and cyr are highly collinear with other variables in the model.

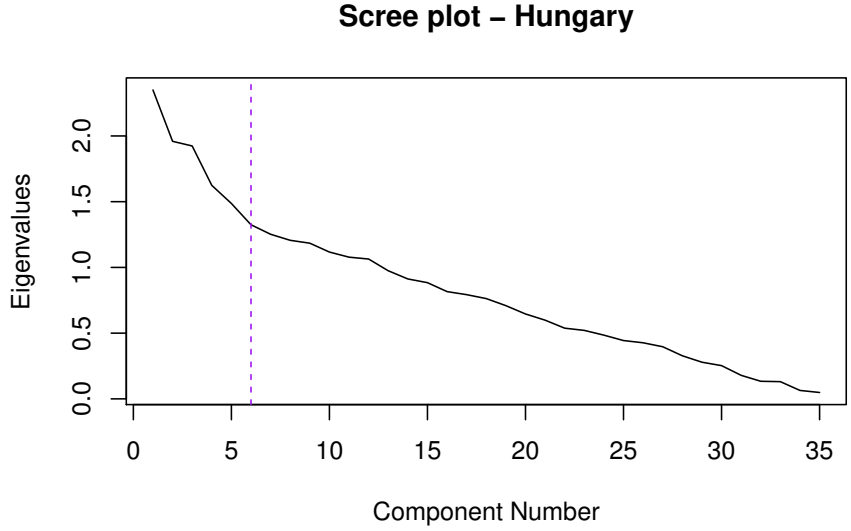


Figure 6: Scree plot for PCA of Hungary

From the scree plot in Figure 6 we see that we keep 6 components.

We have the loadings of each components as follows.

Hungary PCA loadings

Loadings:										
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
age	0.162		0.184		−0.253		−0.105	−0.185	−0.114	0.137
sex		−0.144	−0.164	−0.236					0.232	
painloc	0.143		−0.261	−0.112	0.116					0.214
painexer	0.212		−0.335			−0.146				
relrest	0.228		−0.332					−0.159		−0.108
cp	0.229		−0.357			−0.163		−0.115		
trestbps	0.213		0.179	−0.290	0.113			−0.214		0.198
htn		−0.160	0.101			0.247		0.125	0.230	0.214
chol				−0.139	−0.207	−0.105		0.186	−0.529	
fbs		−0.170		−0.127	−0.193	0.168		0.265	−0.304	−0.195
restecg									−0.135	0.304

ekgmo	−0.191		−0.177		−0.276		0.344	−0.380		
ekgday			−0.108		−0.254	0.479	−0.110		0.255	
ekgyr	0.126	−0.312	0.189			−0.181	0.237		0.101	−0.326
prop	0.132	−0.253		0.301	0.117		0.112			0.194
nitr		−0.286		0.402						0.163
pro		−0.309		0.355					−0.105	0.119
diuretic			0.129	−0.101	0.149		0.275	0.123	0.377	0.281
proto	−0.312	−0.277	−0.136	−0.121						
thaldur	−0.305	−0.277	−0.135	−0.130						
thaltime	−0.303	−0.270	−0.138	−0.128			−0.111			0.113
met	−0.306	−0.227					−0.192			
thalach	−0.259			−0.135	0.126		0.376	0.177	−0.142	
thalrest				−0.126			0.539	0.233	−0.141	0.269
tpeakbps		−0.225		−0.292		0.170		−0.264		
tpeakbpd		−0.191	0.271	−0.231		0.134		−0.199		
trestbpd	0.157		0.155	−0.301	0.137			−0.207	−0.150	0.256
exang	0.237		−0.216	−0.140						−0.175
oldpeak		0.113		−0.168	0.217	0.146			0.259	−0.200
slope	0.156		−0.245			0.280		0.287		
rldv5	−0.147	0.126			0.439	0.283		−0.130	−0.232	−0.109
rldv5e	−0.128			0.116	0.444	0.289		−0.179	−0.172	−0.129
cmo	−0.175		−0.194		−0.284	0.103	0.297	−0.395		
cday		−0.123			−0.229	0.387	0.133	0.142		−0.211
cyr	0.130	−0.316	0.179			−0.181	0.222			−0.331

We see that the first principle component is mostly formed of **proto**, **met**, **thaldur** and **thaltime** variables.

The second principle component is mostly formed of **cyr**, **ekgyr** and **pro** variables.

The third principle component is mostly formed of **cp**, **painexer** and **relrest** variables.

The fourth principle component is mostly formed of **nitr** and **pro** variables.

The fifth principle component is mostly formed of **rldv5e** and **rldv5** variables.

The sixth principle component is mostly formed of **ekgday** and **cday** variables.

2.3.3 Longbeach

After removing dummy variables and variables with at least 50% NA data, we are left with 50 variables and 94 observations.

Longbeach variance inflation factor

age	sex	painloc	painexer	relrest	cp	trestbps	htn
3.228090	1.931427	2.577184	7.718893	6.621863	14.044802	3.617851	2.921354
chol	smoke	cigs	years	fbs	famhist	restecg	ekgmo
2.045184	4.098390	3.361309	4.794619	3.248168	2.369061	2.192354	3.658305
ekgday	ekgyr	dig	prop	nitr	pro	diuretic	proto
2.449554	39.950248	2.509731	2.528256	1.901292	1.732049	2.476101	4.533802
thaldur	met	thalach	thalrest	tpeakbps	tpeakbpd	trestbpd	exang
18.058001	16.434757	3.931213	2.659439	3.784661	2.735284	3.161172	2.222555
xhypo	oldpeak	rldv5	rldv5e	cmo	cday	cyr	lmt
2.096734	4.015243	7.661567	6.930116	4.934749	2.272059	43.093455	1.678378
ladprox	laddist	diag	cxmain	ramus	om1	om2	rcaprox
2.126046	1.837113	1.766507	1.847673	2.269851	2.527439	2.913807	2.246981
rcadist							
2.142703							

From the variance inflation factor we see that the variables **cp**, **ekgyr**, **thaldur**, **met** and **cyr** are highly collinear with other variables in the model.

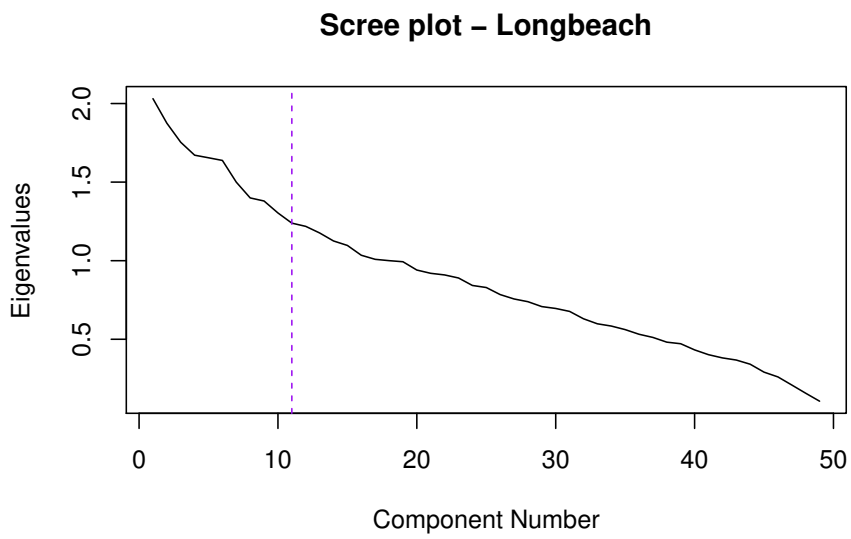


Figure 7: Scree plot for PCA of Longbeach

From the scree plot in Figure 7 we see that we keep 11 components.

We have the loadings of each components as follows.

Longbeach PCA loadings

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
age	−0.204		0.197			−0.182		−0.205		
sex			−0.106	−0.215			0.147			0.146
painloc		−0.133	0.209			0.255		0.136	−0.105	
painexer	−0.192	−0.223			0.288	0.188	−0.162			
relrest	−0.181	−0.203			0.114	0.314		0.197		
cp	−0.181	−0.276			0.216	0.288	−0.128	0.184		
trestbps	−0.196		0.310	−0.133		−0.202				
htn			0.313	−0.134	−0.219					0.106
chol			0.118	0.166	0.175	−0.156	−0.232			−0.194
smoke	0.160		−0.191	−0.334						−0.186
cigs			−0.250	−0.320						0.240
years	0.155	−0.117	−0.133	−0.344					0.103	
fbs		0.111	0.196			−0.168	−0.154			0.315
famhist				−0.124	−0.235			0.159		−0.316
restecg	0.125		0.132					−0.128	0.257	0.229
ekgmo		−0.189	−0.134				0.178		−0.289	
ekgday			−0.149				0.357		0.264	0.166
ekgyr	−0.357	0.130	−0.161	−0.127			−0.195			
dig	0.166				−0.111	−0.241		0.111	0.227	−0.183
prop			0.137	−0.106		0.115	0.179	−0.179	0.135	−0.116
nitr			0.123		−0.215	0.181		−0.255	−0.139	
pro				−0.177	−0.253			0.170	−0.112	
diuretic		−0.140	0.162	−0.221			0.106		0.187	−0.109
proto	−0.288	0.135	−0.240		−0.102			−0.171		
thaldur		0.402					0.102	0.153	−0.190	−0.113
met		0.353			0.126		0.131	0.237	−0.172	
thalach		0.151		−0.129	0.323		0.279	0.172		
thalrest					0.349		0.169		0.119	
tpeakbps		0.264	0.229	−0.167			0.154			0.158
tpeakbpd	0.139		0.223		0.124	0.141	0.235	0.144		
trestbpd			0.200	−0.181		−0.154		0.135		

exang	−0.105	−0.260				−0.141	0.135		−0.110	−0.174
xhypo	−0.128				0.153	−0.177			−0.148	
oldpeak	−0.250				0.204	−0.208	0.116			−0.274
rldv5	−0.238			0.274	−0.118		0.280		0.249	−0.137
rldv5e	−0.238			0.262	−0.191		0.255		0.224	−0.153
cmo		−0.244	−0.131	0.117		−0.124	0.210		−0.291	
cday					0.135	0.129	0.160	−0.310		0.299
cyr	−0.352	0.156	−0.160	−0.134			−0.185			
lmt	−0.104						−0.149	0.166		
ladprox		−0.103		−0.127	0.229	−0.220				0.140
laddist						0.123	0.102		−0.233	
diag	−0.105				−0.109			0.384		0.174
cxmain		−0.126				−0.233			−0.118	
ramus		0.103	0.113	−0.139	0.140		−0.114	−0.232		−0.116
om1		−0.105	0.100	−0.144					−0.347	−0.161
om2			0.249	−0.217	0.100	0.133		−0.310		−0.142
rcaprox		−0.139		−0.118		−0.347		−0.113		
rcadist	−0.196				−0.192		0.131		−0.177	0.172

	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15	Comp.16	Comp.17	Comp.18	Comp.19
age	−0.163		−0.136	0.110		0.110		0.185	−0.230
sex		−0.264		0.125			−0.206	−0.150	−0.363
painloc				0.210	0.397			−0.138	−0.199
painexer						−0.246			
relrest	−0.147						0.229	0.114	
cp						−0.138			
trestbps	0.112		−0.129	−0.101					
htn	−0.123		0.241	−0.196	0.100	0.102			
chol		0.180	−0.233				0.113	−0.105	−0.178
smoke	−0.107						0.203		
cigs			−0.160		0.131			−0.118	0.198
years			−0.264				0.156	0.211	
fbs	−0.182	0.233		0.139	0.176			0.135	0.250
famhist		0.269		0.199	0.119		−0.113		0.209
restecg	−0.169					−0.216	−0.139		−0.238
ekgmo	−0.384	0.110	−0.184	−0.205	−0.258				
ekgday	0.168	0.147		0.172	0.161				
ekgyr		0.133				0.168	−0.111		
dig			0.121	−0.146		−0.255	−0.247		−0.129
prop		0.242	−0.265	−0.255	−0.141	−0.265		−0.185	−0.188
nitr				0.322	−0.116			−0.135	
pro		0.168	0.127	0.208	−0.186	−0.188	−0.190		
diuretic	−0.194		0.170	−0.138	0.170	0.322	0.248	0.107	−0.126
proto							−0.156		
thaldur	−0.152				0.114	−0.138	0.111	−0.183	−0.152
met	−0.139				0.171	−0.230	0.215	−0.143	−0.166
thalach				0.135		0.102		0.290	0.111
thalrest		0.183			−0.220	0.151	−0.344		−0.102
tpeakbps	−0.110	−0.145		0.166	−0.242			0.162	−0.112
tpeakbpd	0.102	−0.105		0.146	−0.169	0.353	−0.157		
trestbpd	0.371	0.130	−0.147	−0.210	−0.110		−0.134	−0.124	0.114
exang	0.206	−0.116	−0.215	0.208	0.156	−0.106			
xhypo		0.349		−0.156	0.253	0.155		0.123	
oldpeak		−0.159	0.101		0.100			0.143	0.175
rldv5		−0.105							0.119
rldv5e					−0.108		0.104		
cmo	−0.337		−0.142				−0.116		
cday	0.209	0.127			0.122		0.258	−0.119	
cyr		0.146				0.164			
lmt		−0.226	−0.455	−0.245		0.126			
ladprox		−0.137	0.224				0.117	−0.310	0.236

laddist	0.331	−0.207		−0.191	0.182	−0.167	−0.168	0.349	−0.174
diag		−0.262						−0.327	0.139
cxmain	0.104	0.130		0.137	−0.356	−0.184	0.350	0.191	−0.105
ramus	−0.196	−0.260		0.165	0.104	−0.133	−0.169	0.115	0.279
om1			0.313	−0.280		0.122		−0.129	
om2						−0.257	−0.104		0.209
rcaprox				0.184	0.156			−0.297	
rcadist		0.124	0.214				−0.106		

We see that the first principle component is mostly formed of **ekgyr**, **cyr** and **proto** variables.

The second principle component is mostly formed of **thaldur**, **met** variables.

The third principle component is mostly formed of **htn**, **trestbps** and **cigs** variables.

The fourth principle component is mostly formed of **years**, **smoke** and **cigs** variables.

The fifth principle component is mostly formed of **thalrest**, **thalach** and **painexer** variables.

The sixth principle component is mostly formed of **rcaprox**, **relrest** and **cp** variables.

The seventh principle component is mostly formed of **ekgday**, **rldv5** and **thalach** variables.

The eight principle component is mostly formed of **diag**, **cday** and **om2** variables.

The ninth principle component is mostly formed of **om1**, **cmo** and **ekgmo** variables.

The tenth principle component is mostly formed of **famhist**, **fbs** and **cday** variables.

The eleventh principle component is mostly formed of **ekgmo**, **trestbpd**, **cmo** and **laddist** variables.

2.3.4 Switzerland

After removing dummy variables and variables with at least 13% NA data, we are left with 39 variables and 101 observations.

Switzerland variance inflation factor

age	sex	painloc	painexer	relrest	cp	trestbps	restecg
2.369562	1.738028	3.014841	5.301607	5.348634	5.703978	3.512376	2.391685
ekgmo	ekgday	ekgyr	dig	prop	nitr	pro	diuretic
15.412883	4.698930	11.069307	1.660195	2.140460	2.363016	2.250616	1.810968
thaldur	thalach	thalrest	tpeakbps	tpeakbpd	trestbpd	exang	xhypo
4.680438	4.923162	3.031050	4.382267	2.124042	2.928830	1.982196	2.170784
oldpeak	cmo	cday	cyr	lmt	ladprox	laddist	diag
2.282318	17.422769	4.254917	6.008334	1.696866	2.296964	1.815151	1.777471
cxmain	ramus	om1	om2	rcaprox	rcadist		
2.269139	2.183169	2.849526	1.660434	1.868517	1.905493		

From the variance inflation factor we see that the variables **ekgmo**, **ekgyr** and **cmo** are highly collinear with other variables in the model.

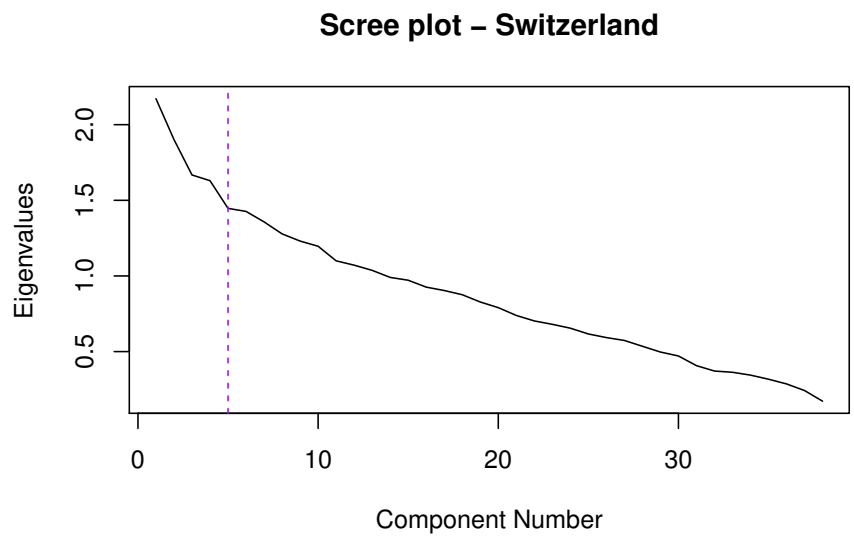


Figure 8: Scree plot for PCA of Switzerland

From the scree plot in Figure 8 we see that we keep 5 components.

We have the loadings of each components as follows.

Switzerland PCA loadings

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
age			0.366		0.115		−0.106	0.158	0.220	−0.138
sex		−0.107			−0.238			−0.326	0.118	
painloc	−0.209	−0.265	−0.145	−0.197						0.160
painexer	−0.245	−0.238	−0.183	−0.211			0.160			
relrest	−0.215	−0.193	−0.187	−0.264			0.147	0.144		
cp	−0.214	−0.292	−0.203	−0.250						
trestbps	−0.154	0.126	0.369	−0.192		0.114		0.169	0.116	
restecg		0.110	0.170		−0.182			0.101	0.202	−0.451
ekgmo	−0.337	0.199			−0.120	−0.168				
ekgday		−0.143	−0.219		−0.172	0.420	−0.178	0.222		
ekgyr	0.316	−0.219	0.142	−0.179	0.137		0.171			
dig				0.140		0.167	0.281	0.187		0.342
prop	0.111	−0.227		0.129		−0.201	−0.174		0.204	0.143
nitr	0.137	−0.214				−0.307		0.181	0.332	
pro		−0.111	0.106		−0.198	−0.331	−0.201	0.325		0.104
diuretic	−0.105	−0.119	0.133	0.113	−0.114	−0.201	−0.162	0.173	−0.162	
thaldur	0.231			−0.107	−0.355	−0.161	0.192	−0.166		0.214
thalach	0.184	0.172	−0.191	−0.187	−0.242		0.247	0.232	−0.236	
thalrest		0.269	−0.156				0.222	0.411	−0.151	−0.160
tpeakbps		0.215	0.117	−0.372	−0.176			0.197		
tpeakbpd	−0.116	0.125	0.148	−0.352		0.100			0.189	
trestbpd	−0.206	0.106	0.199	−0.179		0.214	−0.123		0.163	0.152
exang	−0.188					0.153	−0.315		−0.251	−0.117
xhypo			0.114	0.170	0.226	0.229	0.290	0.269		0.182
oldpeak		−0.110		−0.298			0.151		−0.205	−0.187
cmo	−0.345	0.185			−0.128	−0.128		−0.109		
cday		−0.129	−0.189		−0.188	0.381	−0.231	0.228		
cyr	0.275	−0.194	0.136	−0.154	0.202		0.144	−0.138		
lmt		−0.103	−0.142	0.123	0.119	−0.116	0.135		0.166	−0.333
ladprox	−0.150	−0.113		0.186	0.132			0.122		−0.195
laddist		−0.150	0.117		−0.348					0.198

diag		−0.187	0.196		−0.179				−0.323	
cxmain	−0.102	−0.128			−0.190	0.219		−0.208		−0.366
ramus	−0.139	−0.107	0.237	0.129			0.108		−0.311	
om1		−0.186	0.154	0.153	−0.247		0.315		0.124	−0.149
om2		−0.161	0.212						−0.338	
rcaprox	−0.189		0.119		0.159		0.133		0.116	0.156
rcadist	−0.106			0.137	−0.255		0.277		0.220	

We see that the first principle component is mostly formed of **cmo**, **ekgmo**, **ekgyr** and **cyr** variables.

The second principle component is mostly formed of **cp**, **thalrest**, **painloc**, and **painexer** variables.

The third principle component is mostly formed of **trestbps** and **age** variables.

The fourth principle component is mostly formed of **tpeakbps**, **tpeakbpd** and **oldpeak** variables.

The fifth principle component is mostly formed of **thaldur** and **laddist** variables.

2.4 Question 4

In order to distinguish patients based on their disease through discriminant analysis, I had to create four models, one for each of the datasets. As the variable *num* has 5 different types it could be: *0, 1, 2, 3 & 4*, I had to use multiclass LDA to perform my analysis. I have included the output of the multiclass LDA even though I will be performing my analysis upon the related historgrams.

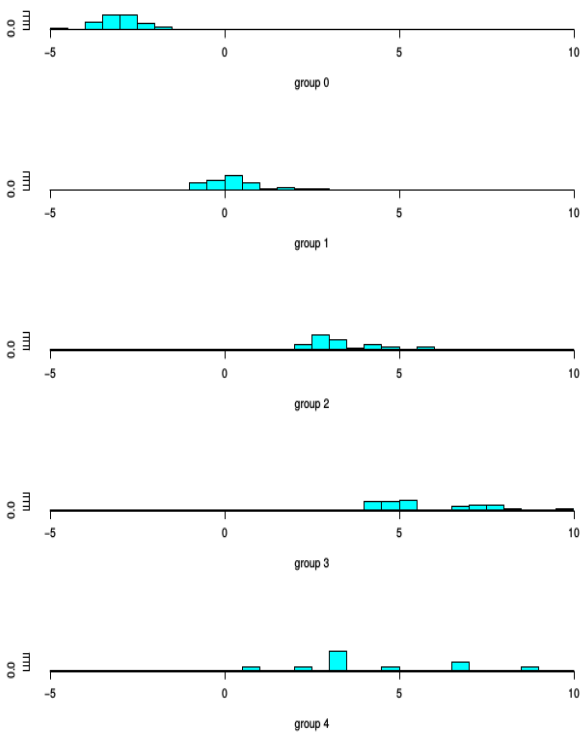
2.4.1 Cleveland

```
Call:
lda(num ~ ., data = clevdata)

Prior probabilities of groups:
        0         1         2         3         4
0.50248756 0.18407960 0.12935323 0.12935323 0.05472637

Group means:
       age       sex       cp trestbps       htn     chol     cigs     years
0 53.52475 0.5247525 2.841584 128.7921 0.5544554 249.3465 14.52475 13.77228
1 55.86486 0.8918919 3.459459 132.7838 0.7027027 252.4324 21.67568 17.27027
2 58.00000 0.8076923 3.692308 133.0385 0.7307692 274.0000 15.30769 15.15385
3 56.03846 0.8076923 3.807692 136.7308 0.4230769 252.1923 15.88462 12.73077
4 59.54545 0.9090909 3.636364 140.0000 0.6363636 233.5455 26.27273 22.63636
         fbs    famhist    restecg     ekgmo    ekgday     ekgyr        dig
0 0.14851485 0.5445545 1.0000000 5.900990 14.89109 82.30693 0.06930693
1 0.08108108 0.6756757 1.1351351 6.756757 13.29730 82.02703 0.00000000
2 0.19230769 0.8076923 0.9230769 5.192308 12.34615 82.38462 0.00000000
3 0.23076923 0.5384615 1.1538462 5.692308 16.80769 82.03846 0.00000000
4 0.09090909 0.5454545 1.6363636 7.636364 14.18182 82.09091 0.00000000
       prop       nitr        pro   diuretic  thaldur thaltime       met  thalach
0 0.2772277 0.2079208 0.11881188 0.1386139 8.836634 4.271287 10.227723 157.2376
1 0.3783784 0.3513514 0.10810811 0.1081081 8.294595 6.027027  9.837838 145.3784
2 0.5769231 0.3076923 0.15384615 0.1153846 7.873077 6.215385  8.807692 132.3846
3 0.3846154 0.4230769 0.03846154 0.1923077 6.707692 4.642308  8.076923 132.9231
4 0.3636364 0.2727273 0.00000000 0.0000000 6.027273 4.863636  7.363636 139.4545
  thalrest tpeakbps tpeakbpd trestbpd     exang      xhypo   oldpeak    slope
0 77.28713 170.0891 78.96040 84.23762 0.1782178 0.00990099 0.7435644 1.504950
1 72.37838 172.3243 77.59459 85.05405 0.5405405 0.02702703 1.2864865 1.702703
2 69.76923 158.4615 81.69231 85.26923 0.5769231 0.07692308 1.7653846 1.961538
3 75.11538 160.2692 79.26923 87.65385 0.6538462 0.07692308 2.2884615 2.038462
4 74.63636 165.4545 84.09091 86.63636 0.5454545 0.00000000 2.0272727 2.000000
    rldv5e        ca     thal      cmo     cday      cyr  ladprox  laddist
0 122.9010 0.3267327 3.792079 5.811881 15.22772 82.30693 1.000000 1.000000
1 131.3243 0.8648649 5.432432 6.756757 14.67568 82.02703 1.270270 1.324324
2 127.0000 1.3076923 6.115385 5.153846 13.61538 82.42308 1.423077 1.384615
3 138.9231 1.4230769 6.538462 5.923077 16.50000 82.07692 1.461538 1.730769
4 107.8182 1.4545455 6.181818 7.636364 15.36364 82.00000 1.272727 1.545455
    cxmain      om1  rcaprox  rcadist
0 1.000000 1.000000 1.000000 1.000000
1 1.135135 1.081081 1.216216 1.108108
2 1.384615 1.346154 1.576923 1.230769
3 1.500000 1.730769 1.461538 1.576923
4 1.545455 1.545455 1.454545 1.272727

Coefficients of linear discriminants:
                    LD1           LD2          LD3           LD4
age       -0.0064021152 -0.0425071101  0.017806402 -1.388222e-02
sex        0.2188951081 -0.8385787717  0.966289676 -1.080167e-01
cp         0.0836865201 -0.0684255751  0.213747047 -2.619771e-01
trestbps   0.0086722361 -0.0110214859  0.009667569  7.267368e-03
```

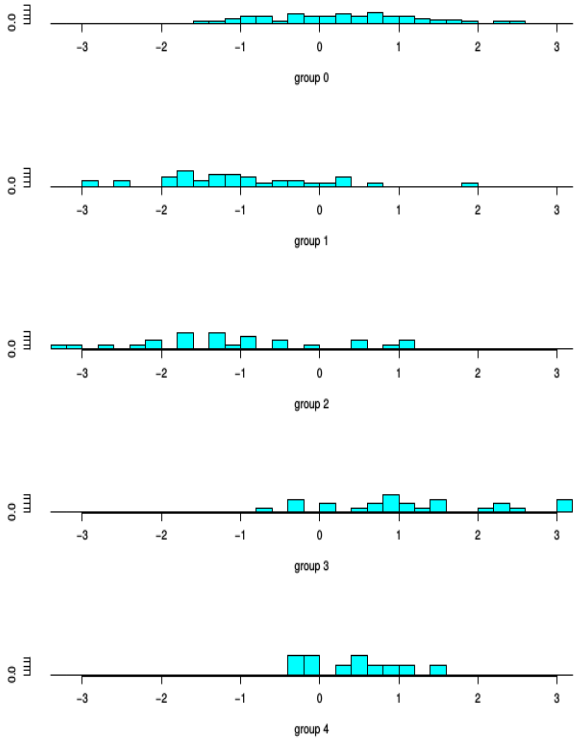
```
htn          0.1339883778  -0.5852997536  -0.309526408   3.116354e-01
chol        -0.0001853615  -0.0026943630  -0.007365030   1.378357e-03
cigs         0.0014297509  -0.0040736758   0.023693179  -5.710203e-03
years       -0.0058677471   0.0064698922  -0.010415602   2.424284e-02
fbs         -0.0677560105   0.4462514441  -0.809205331   1.160757e-01
famhist     -0.1536120593  -0.6331094392  -0.046988498  -7.834152e-03
restecg      0.0020724640   0.0306173919   0.343424560   1.242487e-01
ekgmo       -0.0333931677  -0.1894708072   0.003215433   1.790893e-01
ekgday       0.0307768259   0.0202459475   0.004269374  -1.967420e-02
ekgyr        0.2849909378  -0.1910807911   3.244719860   3.043414e+00
dig         -0.5880796669   1.2356661900  -0.935588867   4.891422e-01
prop        -0.2044468618  -0.4228151708   0.314419800   2.952549e-01
nitr        -0.0500314492  -0.1132297678  -0.008236496  -6.704832e-01
pro         -0.0422771375  -0.3051736743  -0.212490269  -7.934840e-02
diuretic     0.0903449725  -0.1880557316  -0.424110167  -3.147873e-01
thaldur      0.0123720933  -0.0940568662  -0.296638460   2.257168e-01
thaltime     0.0352787900  -0.1062383655   0.033833393   2.174663e-03
met         -0.0227104963   0.0333032838   0.052992717  -2.510038e-01
thalach      0.0004083536   0.0016079361   0.014067133  -2.553563e-05
thalrest    -0.0143560702   0.0142803750   0.001611765  -7.030841e-03
tpeakbps    -0.0071065254  -0.0069409621   0.008450720  -1.586376e-02
tpeakbpd     0.0126391705   0.0036640956  -0.017755491   4.596684e-02
trestbpd    -0.0092676913   0.0131619284  -0.014903990  -3.778290e-02
exang        0.3742483643  -0.6846868868  -0.162847360  -3.792109e-01
xhypo       -0.3070163922  -0.9550240951  -0.837879117  -9.084565e-01
oldpeak     -0.0507070406   0.3552810935  -0.016595052   7.428472e-02
slope        0.2158763397  -0.3990518860   0.152776941   1.517427e-01
rldv5e      -0.0016341882  -0.0009474799  -0.004657443  -6.008307e-03
ca           0.0604185429  -0.1322843859   0.044212892   3.194839e-02
thal         0.1790094224  -0.0163313489  -0.036193049  -1.153858e-01
cmo          0.0652759054   0.1570432167   0.099289729  -1.540472e-01
cday        -0.0202700434   0.0109527280   0.006883157   1.056202e-02
cyr         -0.2318889441   0.0552735975  -3.287530952  -2.820889e+00
ladprox      2.1449671081  -0.7520705684  -0.456454141  -3.504782e-01
laddist      2.2474742138   0.3244757867   0.480884470  -6.849638e-01
cxmain       2.4577588717   0.6761926661   0.110008133   9.377253e-01
om1          2.3487229073   1.7069044343  -0.588196080   6.720648e-01
rcaprox      2.5240275251  -0.9443404888  -0.769549935   8.529602e-01
rcadist      1.8594484941   0.6017481881  -0.073459411  -4.100481e-01

Proportion of trace:
   LD1    LD2    LD3    LD4
0.9090 0.0477 0.0268 0.0164
```

We can see from the above output that LDA has produced 4 classifiers and plotting their histograms 9 shows us that only the **first LDA classifier** is any good as it was able to clearly seperate 4 of the five types, with the final type (group) spread between types *3 & 4*. The **third and fourth LDA classifiers** are the worst as they are seemingly unable to seperate any of the types from each other. The **second LDA classifier** is it seems is only able to sperate types *0, 3 & 4* from types *1 & 2* it seems so is more helpful to us that the third and fourth but it is nowhere near as helpful as the first LDA classifier.

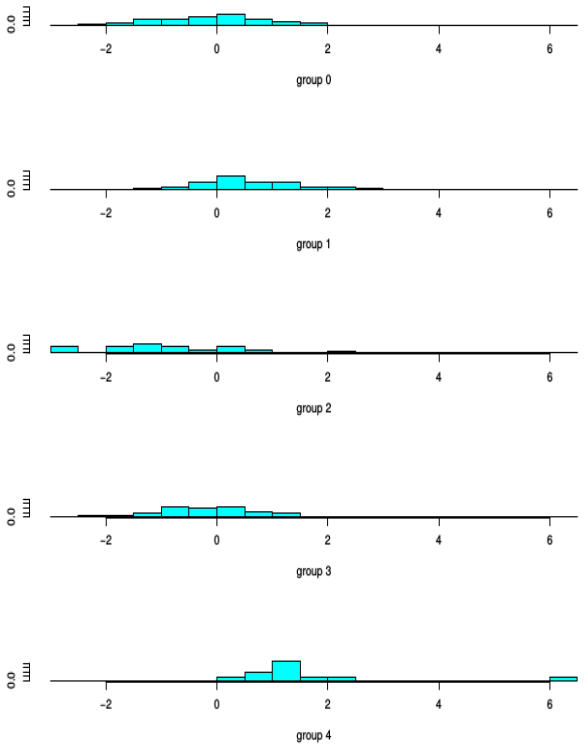
(a) First LDA classifier



(b) Second LDA classifier



(c) Third LDA classifier



(d) Fourth LDA classifier

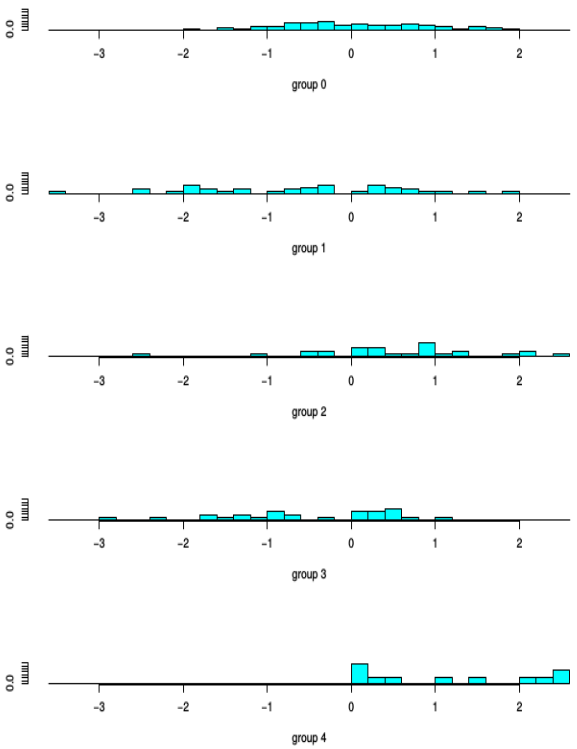


Figure 9: Histergram for discriminant analysis of Cleveland

2.4.2 Hungary

```
Call:
lda(num ~ ., data = ex3dathun)

Prior probabilities of groups:
        0         1         2         3         4
0.2954545 0.2045455 0.1590909 0.2272727 0.1136364

Group means:
       age       sex   painloc  painexer   relrest       cp trestbps       htn
0 51.53846 0.4615385 0.8461538 0.4230769 0.4230769 2.884615 131.6923 0.3076923
1 47.72222 0.8333333 1.0000000 0.8333333 0.8888889 3.722222 130.9444 0.5000000
2 51.07143 0.9285714 1.0000000 1.0000000 0.9285714 4.000000 145.8571 0.7142857
3 51.15000 0.9000000 0.9500000 0.7000000 0.7000000 3.400000 140.8000 0.4000000
4 50.40000 0.8000000 1.0000000 0.9000000 1.0000000 3.900000 138.5000 0.5000000
      chol       fbs   restecg    ekgmo   ekgday    ekgyr       prop       nitr
0 246.9615 0.0000000 0.2307692 6.192308 14.96154 85.03846 0.07692308 0.11538462
1 246.0000 0.1111111 0.1111111 5.333333 16.61111 85.55556 0.11111111 0.22222222
2 288.4286 0.2142857 0.1428571 4.642857 16.07143 85.57143 0.14285714 0.07142857
3 250.4500 0.1500000 0.3000000 5.750000 17.60000 85.35000 0.05000000 0.25000000
4 275.0000 0.1000000 0.4000000 7.200000 16.80000 84.90000 0.20000000 0.20000000
         pro   diuretic     proto  thaldur thaltime      met  thalach thalrest
0 0.03846154 0.00000000 83.65385 9.038462 8.038462 5.269231 136.1923 81.26923
1 0.22222222 0.00000000 81.94444 9.055556 7.972222 5.111111 129.8333 76.77778
2 0.07142857 0.07142857 75.00000 8.428571 7.285714 4.235714 119.0714 82.57143
3 0.20000000 0.05000000 76.25000 7.975000 6.800000 4.820000 121.5000 71.95000
4 0.20000000 0.00000000 60.00000 5.900000 4.950000 3.800000 127.0000 82.30000
  tpeakbps  tpeakbpd trestbpd     exang  oldpeak    slope    rldv5   rldv5e
0 176.1538  95.15385 84.96154 0.5000000 1.242308 1.653846 15.53846 15.46154
1 171.4444  96.94444 83.55556 0.8333333 1.611111 2.000000 15.00000 14.94444
2 181.1429 100.35714 90.71429 0.9285714 1.785714 2.071429 14.00000 13.57143
3 182.0000  98.95000 89.05000 0.7500000 1.750000 2.000000 12.90000 13.55000
4 171.0000  94.20000 89.40000 0.9000000 2.300000 2.000000 17.80000 16.50000
       cmo      cday       cyr
0 6.307692 15.30769 85.03846
1 5.722222 16.11111 85.55556
2 4.642857 20.28571 85.57143
3 6.300000 18.10000 85.30000
4 7.200000 21.60000 85.00000

Coefficients of linear discriminants:
                  LD1           LD2          LD3          LD4
age       -0.0319039718  0.0949747891  0.013319391  0.025531882
sex        1.0915165611 -1.0511591741 -0.128089043 -0.217640165
painloc    0.1592508146 -0.5982393519  0.751516162 -1.356165886
painexer   1.0443921480 -1.8274651017  0.423915039 -0.129822932
relrest   -0.5353907890  0.3080857927  1.366855699 -1.292874896
cp         0.2051091438  1.0860954298 -0.347952792  0.582142933
trestbps  -0.0291801216 -0.0210943388 -0.005915740  0.010631108
htn        0.7977232482  0.7125540372  0.687763392  0.090163649
chol       0.0059740613  0.0003533943 -0.008706924  0.001824229
fbs        0.3020817737 -1.2745529471  0.340669290 -0.124606147
restecg   -0.0034817198  0.0287925187 -0.860566466 -0.477512401
ekgmo      0.1766896834  0.2064361753 -0.235624384 -0.163174965
ekgday    -0.0008617825 -0.0371982064 -0.025145255 -0.019070970
ekgyr     -2.1415512648 -2.4433469890  1.742048061  0.200436115
prop      -1.7298024975  2.4134058221  0.644332191  0.279975382
nitr      -1.3114932699 -1.8866584974 -1.084697545  0.526330742
pro        2.2930978213 -0.2622377510 -0.638507586 -1.800873718
```

```
diuretic   4.0807841243 -1.4825164759 -3.921468835  0.528639304
proto     -0.0268455617  0.0109964005 -0.052095246 -0.002381793
thaldur   -0.1915704888 -0.2807526053  0.116598448  0.766573761
thaltime   0.2397031414  0.3449238438  0.509998503 -0.660418467
met        0.1538161967 -0.1558001805 -0.188012336  0.045169327
thalach   -0.0113911821 -0.0075009574  0.012289493 -0.030730660
thalrest  -0.0148107903  0.0404184418  0.014743907  0.032907539
tpeakbps   0.0055246183 -0.0024652818 -0.014276027  0.015577249
tpeakbpd   0.0147301260  0.0121621412  0.034171915 -0.027803617
trestbpd   0.0447580846 -0.0017554695 -0.059802256  0.041413630
exang      0.3026614220 -0.0396448017  0.081711393 -0.223888568
oldpeak    0.5557096564  0.6919891057 -0.374908659 -0.319812279
slope      1.5871575130 -0.1881557988  0.947157542  0.617202485
rldv5      0.0870335285  0.2322222300  0.056690032 -0.001700644
rldv5e    -0.0568400417 -0.1650279852 -0.053695184 -0.020264457
cmo       -0.1012481889 -0.1914088336  0.070107909  0.059942757
cday       0.0266986647  0.0502210848 -0.025879692  0.029755031
cyr        2.4408858944  2.4327046963 -1.459350228 -0.307030741

Proportion of trace:
   LD1    LD2    LD3    LD4
0.5086 0.2406 0.1662 0.0846
```

From the above output from the LDA of my Hungary model we that we have again produced 4 LDA classifiers which I then produced the historgrams 10 for. Sadly, they are able to tell us much about the data as it seems as though non of them were able to differentiate between any of the different types.

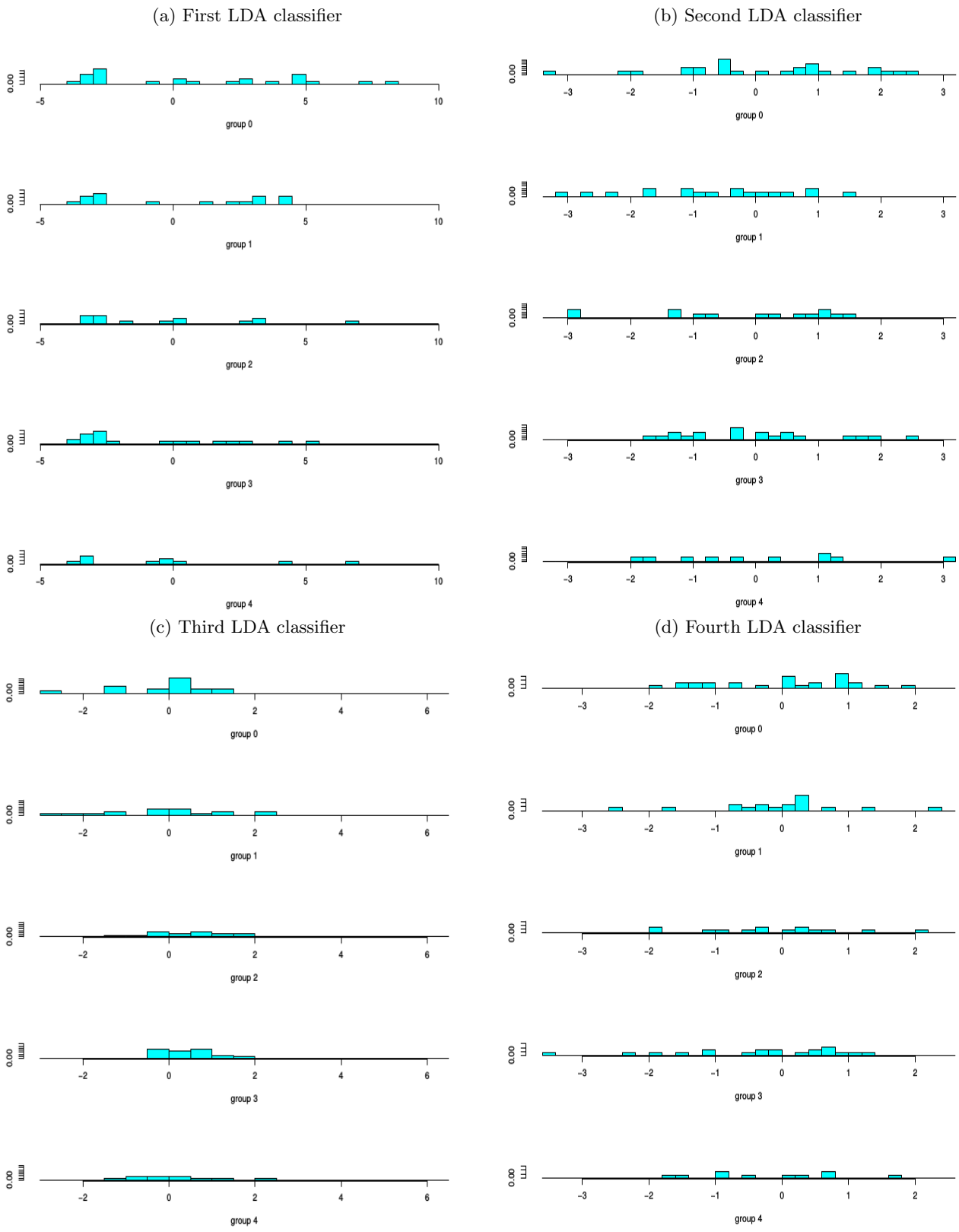


Figure 10: Histergram for discriminant analysis of Hungary

2.4.3 Longbeach

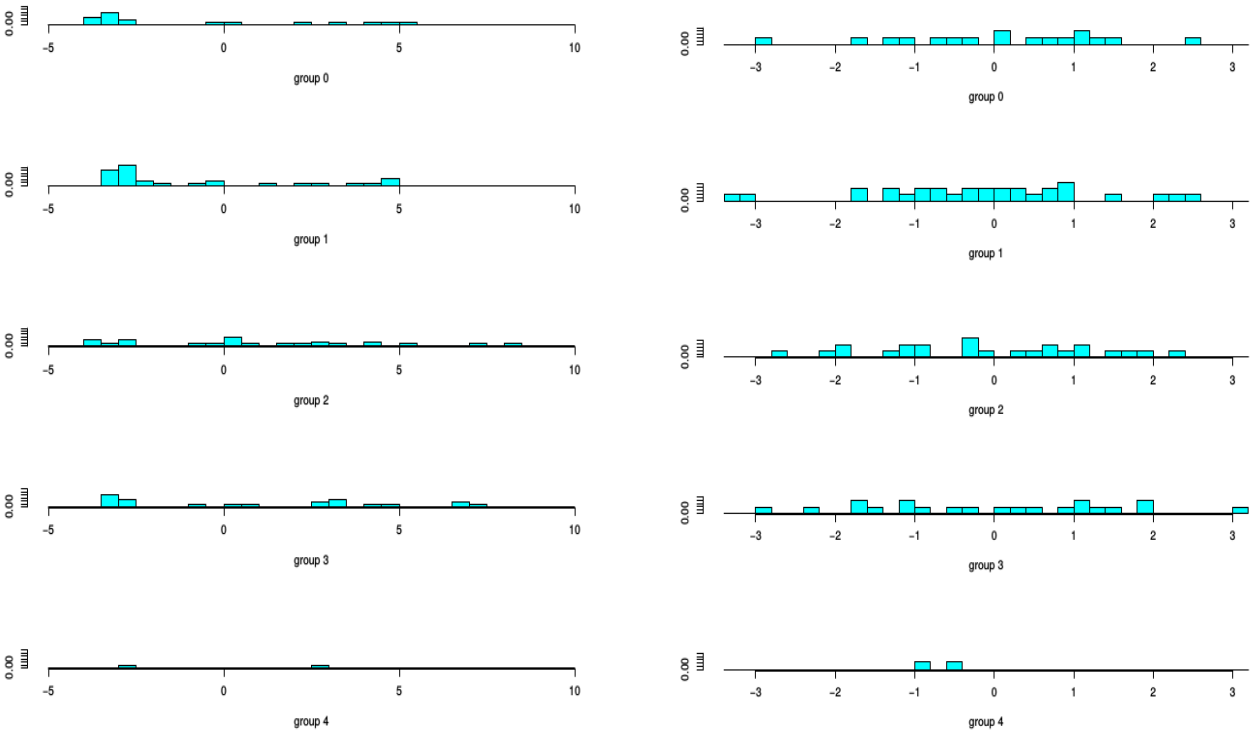
```
Call:
lda(num ~ ., data = ex3datlon)

Prior probabilities of groups:
        0         1         2         3         4
0.1808511 0.3191489 0.2553191 0.2234043 0.0212766

Group means:
       age       sex   painloc  painexer   relrest       cp trestbps       htn
0 58.88235 0.8823529 1.0000000 0.5882353 0.8235294 3.470588 124.5882 0.5294118
1 56.63333 0.9333333 0.9666667 0.8666667 0.9333333 3.833333 131.4667 0.5333333
2 60.58333 1.0000000 0.9583333 0.7916667 0.8333333 3.625000 136.6250 0.6666667
3 62.95238 1.0000000 0.9047619 0.8571429 0.9047619 3.714286 140.9048 0.7142857
4 63.50000 1.0000000 1.0000000 1.0000000 1.0000000 4.000000 148.0000 0.5000000
      chol     smoke      cigs     years       fbs   famhist   restecg     ekgmo
0 198.9412 0.5294118 24.70588 27.23529 0.4117647 0.7058824 0.7647059 5.647059
1 194.0000 0.4333333 19.90000 17.00000 0.3666667 0.4666667 0.7000000 6.700000
2 174.0417 0.5833333 18.33333 24.41667 0.2500000 0.4166667 0.7083333 6.208333
3 217.4286 0.5714286 23.80952 27.76190 0.3333333 0.3809524 0.6666667 7.000000
4 251.0000 1.0000000 30.00000 38.50000 0.5000000 1.0000000 0.5000000 6.000000
    ekgday    ekgyr        dig       prop      nitr       pro   diuretic     proto
0 15.05882 84.58824 0.11764706 0.4117647 0.7058824 0.1764706 0.3529412 3.588235
1 16.13333 84.40000 0.06666667 1.0666667 0.5666667 0.2000000 0.1666667 3.566667
2 13.33333 84.83333 0.12500000 0.4583333 0.5416667 0.2916667 0.3750000 4.000000
3 15.28571 84.42857 0.09523810 0.4285714 0.6190476 0.2380952 0.3809524 3.904762
4 26.50000 86.50000 0.50000000 0.0000000 1.0000000 0.5000000 0.5000000 5.000000
   thaldur      met  thalach thalrest tpeakbps tpeakbpd trestbpd     exang
0 5.858824 5.876471 115.3529 63.94118 159.7059 88.35294 78.23529 0.4705882
1 6.758333 6.740000 121.9667 69.70000 162.9667 91.66667 80.63333 0.6000000
2 6.504167 6.375000 122.0000 69.62500 172.0417 90.62500 83.29167 0.7083333
3 5.730952 5.619048 116.8571 67.04762 159.4286 83.66667 83.47619 0.9047619
4 4.250000 3.500000 124.0000 76.50000 170.0000 84.00000 79.00000 1.0000000
      xhypo   oldpeak    rldv5   rldv5e      cmo     cday      cyr      lmt
0 0.00000000 0.7529412 15.88235 17.52941 6.000000 10.52941 84.64706 10.47059
1 0.00000000 0.9166667 16.70000 16.06667 6.333333 16.13333 84.53333  1.00000
2 0.04166667 1.3791667 14.75000 14.83333 6.791667 14.58333 84.91667  1.00000
3 0.04761905 1.7142857 15.42857 15.47619 7.952381 14.09524 84.52381  1.00000
4 0.00000000 3.5000000 24.00000 19.50000 7.000000  6.50000 86.50000  2.00000
   ladprox  laddist     diag   cxmain ramus      om1      om2  rcaprox  rcadist
0 1.000000 1.000000 1.117647 1.000000 1.000 1.000000 1.000000 1.000000 1.000000
1 1.166667 1.133333 1.133333 1.333333 1.000 1.100000 1.033333 1.233333 1.100000
2 1.458333 1.166667 1.125000 1.458333 1.125 1.291667 1.083333 1.708333 1.083333
3 1.761905 1.238095 1.238095 1.857143 1.000 1.285714 1.095238 1.857143 1.238095
4 1.500000 1.000000 1.000000 2.000000 1.000 1.000000 1.000000 2.000000 1.000000

Coefficients of linear discriminants:
                  LD1           LD2           LD3           LD4
age        0.0990101557  0.035457144 -0.029404453 -0.037368451
sex       -0.7001377631 -1.223212780  2.472604131 -0.385002333
painloc    1.3527943877  2.689379586 -1.442324304  0.925270543
painexer   1.6779571229  1.095644137  1.132485802  0.132422919
relrest    1.0061607957  2.683724569  1.044784632 -0.315270284
cp        -1.3452114777 -2.179502234 -0.074000882  0.276431196
trestbps   0.0152057624  0.043147540  0.019492777  0.004725401
htn        0.2007845282 -0.213314404 -0.049905568  0.516879641
chol       0.0022668552  0.003074371  0.000609473 -0.004283149
smoke      1.3848919525  0.579322247  0.717647892 -0.203604359
```

```
cigs       -0.0021092781  0.040230996 -0.017470605 -0.046310142
years      -0.0189346337 -0.009935330 -0.025142432  0.020697154
fbs        -0.5602845218 -0.600564294  0.917260811 -0.583766088
famhist     0.2368404852  0.633387492  0.062625332 -0.435654580
restecg     0.3148426505  0.054613753  0.230598184 -0.620653987
ekgmo       0.1066755233  0.146072767  0.174751517  0.051019305
ekgday      0.0302154060  0.042708813  0.006610623  0.007431638
ekgyr      -0.6058539073  1.011772666 -0.482944430  1.028674481
dig         0.1934068922  1.829523966  0.420059417  1.532752361
prop       -0.0258580607 -0.128530397 -0.013717278 -0.004001685
nitr       -0.2979245919  0.890522089  0.221162357 -0.115880411
pro         0.2277745757  0.495507245 -0.293334921  0.886596619
diuretic   -0.5265872971 -1.108015873 -0.622279383 -0.614113919
proto       0.0389888030 -0.304394784 -0.045447788  0.044131177
thaldur     0.0974633776  0.347965964  0.471833171 -0.506561997
met         0.0520777936 -0.547177132 -0.234459599  0.387387794
thalach     0.0012019440 -0.009656180  0.001428743  0.009332465
thalrest    0.0349744126  0.037022030  0.013999657  0.005532030
tpeakbps   -0.0005518933  0.012532493 -0.016658042  0.026948500
tpeakbpd    0.0110131580  0.013873221  0.019296564  0.018578019
trestbpd   -0.0101443057 -0.111258996  0.003785584  0.004558456
exang      -0.2885342130 -0.491316817  0.711632300 -0.290403026
xhypo      -2.0890734776 -1.546304470 -2.533162072  2.585953825
oldpeak     0.0562423983  0.729631495 -0.150881175 -0.454416331
rldv5       0.1212482697  0.141480974  0.244757211  0.074758659
rldv5e     -0.1098458488 -0.119370257 -0.225942614  0.008272690
cmo        -0.0044982969 -0.215097791 -0.184654932  0.011105799
cday        0.0154098050 -0.018575403  0.045237609  0.014481709
cyr         0.7728103647 -0.622781492  0.438416369 -0.517516020
lmt        -0.0076644410  0.016199837 -0.011799002 -0.006409811
ladprox     5.1429956452 -0.236848734 -0.756165644 -0.403482082
laddist     5.4700303913 -0.538647061 -0.111674357 -0.040224894
diag        0.1674878034 -0.867356837 -0.233108173 -0.424424671
cxmain      4.9690488611  0.288788245  0.491785610 -0.752048882
ramus       2.2794956046 -2.518189778 -1.110123699  4.143654026
om1         2.7419747798 -0.569683759  0.382398821  0.098133650
om2         1.3014833954  0.573435941 -1.109721994 -1.917714264
rcaprox     4.3481392232  0.005190937 -0.578608990  1.523789635
rcadist     1.9383005198  0.312683826  0.793005347 -2.408552088

Proportion of trace:
   LD1    LD2    LD3    LD4
0.9072 0.0418 0.0262 0.0248
```

Via the above output from the LDA of my Longbeach model we see that 4 LDA classifiers have been created, for which I have produced historgrams 11 for. Sadly, they are able to tell us much about the model data as it not one of the classeifiers were able to differentiate between the different types, even less so than with my Hungary model.

(a) First LDA classifier

(b) Second LDA classifier



(c) Third LDA classifier

(d) Fourth LDA classifier

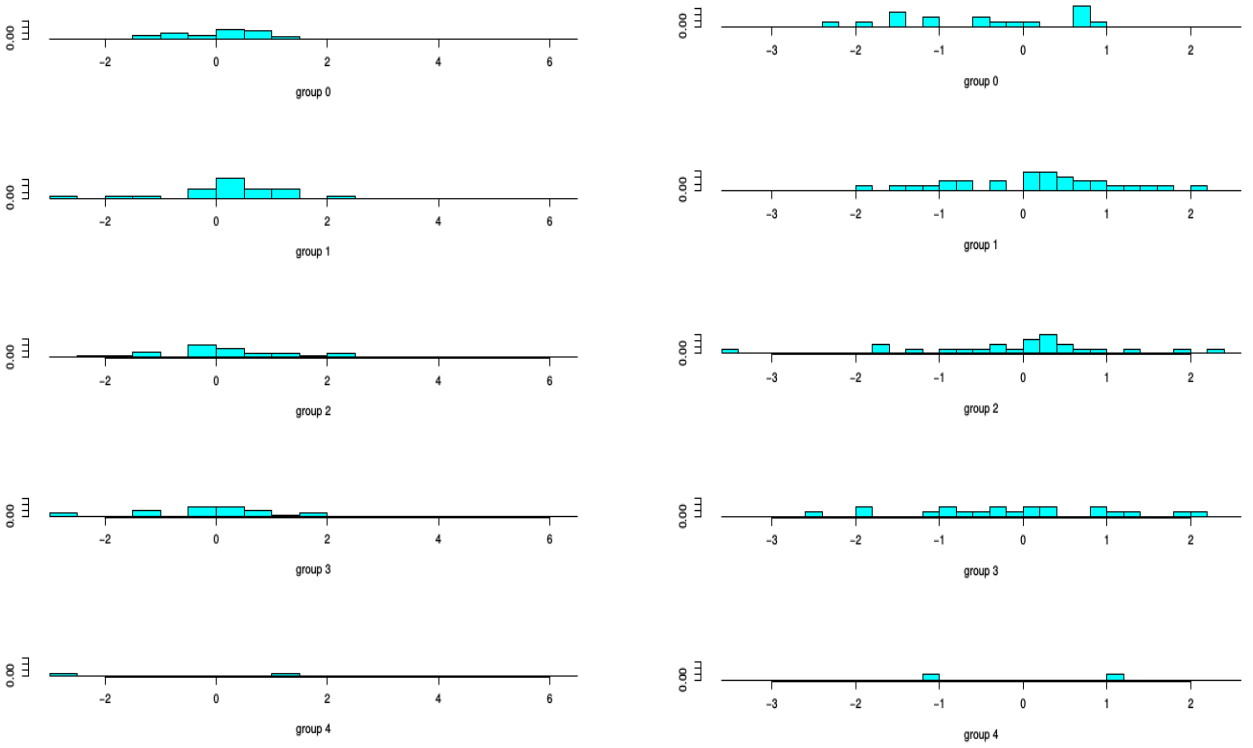


Figure 11: Histergram for discriminant analysis of Longbeach

2.4.4 Switzerland

```
Call:
lda(num ~ ., data = swidata)

Prior probabilities of groups:
         0          1          2          3          4
0.05940594 0.39603960 0.26732673 0.22772277 0.04950495

Group means:
       age       sex   painloc  painexer   relrest       cp trestbps   restecg
0 52.00000 1.0000000 0.6666667 0.3333333 0.3333333 3.166667 117.5000 0.1666667
1 56.80000 0.8750000 0.9500000 0.8500000 0.8000000 3.675000 129.3750 0.3250000
2 53.59259 0.9259259 0.8888889 0.8888889 0.8888889 3.740741 132.5926 0.4074074
3 58.65217 0.9565217 1.0000000 0.9130435 0.7826087 3.782609 140.6522 0.3478261
4 52.00000 1.0000000 1.0000000 1.0000000 1.0000000 4.000000 120.0000 0.4000000
     ekgmo   ekgday    ekgyr        dig       prop      nitr       pro  diuretic
0 2.500000 21.33333 84.83333 0.00000000 0.3333333 0.3333333 0.5000000 0.0000000
1 4.100000 16.27500 84.65000 0.00000000 0.4250000 0.4000000 0.4000000 0.1000000
2 5.814815 13.66667 84.48148 0.03703704 0.3703704 0.2962963 0.4814815 0.2222222
3 5.695652 13.95652 84.52174 0.00000000 0.3913043 0.2608696 0.4347826 0.4347826
4 5.800000 18.60000 84.60000 0.00000000 0.6000000 1.0000000 0.4000000 0.2000000
    thaldur  thalach thalrest tpeakbps tpeakbpd trestbpd      exang      xhypo
0 10.000000 143.1667 68.00000 175.8333 82.50000 75.00000 0.1666667 0.00000000
1  7.557500 123.7500 69.52500 171.7500 86.37500 81.62500 0.4750000 0.00000000
2  7.518519 125.3333 72.11111 169.8148 88.14815 83.51852 0.4074074 0.03703704
3  6.947826 104.2174 64.52174 160.0000 86.08696 87.17391 0.6086957 0.13043478
4  6.900000 109.8000 74.20000 138.0000 78.00000 76.00000 0.4000000 0.20000000
    oldpeak      cmo     cday      cyr  ladprox  laddist     diag   cxmain
0 0.3333333 2.333333 21.33333 84.83333 1.000000 1.000000 1.166667 1.000000
1 0.3675000 4.075000 17.12500 84.70000 1.300000 1.175000 1.075000 1.100000
2 0.7555556 5.814815 13.77778 84.37037 1.592593 1.481481 1.259259 1.333333
3 0.8608696 5.739130 14.21739 84.52174 1.782609 1.521739 1.478261 1.521739
4 0.9600000 6.000000 14.20000 84.60000 1.800000 1.200000 1.000000 1.400000
     ramus      om1      om2  rcaprox  rcadist
0 1.000000 1.000000 1.000000 1.000000 1.000000
1 1.025000 1.000000 1.000000 1.275000 1.125000
2 1.185185 1.000000 1.148148 1.518519 1.222222
3 1.521739 1.347826 1.130435 1.739130 1.521739
4 1.200000 1.200000 1.000000 1.400000 1.200000

Coefficients of linear discriminants:
                 LD1           LD2          LD3           LD4
age      -0.0007858366  0.0628048943 -0.071526268 -0.0417811598
sex       0.8186870033  0.2412555376  0.363652833  0.3287857513
painloc  -0.3837830485  2.2727914494 -1.184109094 -3.4202311632
painexer  0.9118113566 -1.4374997954  0.056616312 -2.5683124594
relrest  -1.2647469151 -1.2871470954  0.420021842 -0.3980500354
cp        0.3521125467  0.6077931365  0.237965504  1.4684042267
trestbps  0.0052216770 -0.0003397642  0.022936687 -0.0002354636
restecg  -0.2637513280 -0.3449011590  0.449630173 -0.2538746008
ekgmo     0.0417657735 -0.0734676135  0.099042462  0.2295703277
ekgday   -0.0056874619  0.0230409744  0.050349943  0.0416943072
ekgyr     0.5069361444  0.5686400550  0.296652266  0.3918472529
dig      -2.5341631270 -2.9792700230 -1.502320680  0.4888530492
prop     -0.2807770793 -0.0863581677  0.267336756 -0.1333181334
nitr      0.4315496561  0.2855985294  1.759176079 -0.4000743686
pro      -0.0005018530 -0.7611546405 -0.194037649  0.5060712017
diuretic  0.7636514352  0.7979700757 -0.147302771  0.0096484846
thaldur  -0.0033692288 -0.0929494992  0.081758599  0.0702982571
```

```
thalach   −0.0087387451  −0.0094572112  −0.009259432  −0.0050115022
thalrest   0.0504091039   0.0242042236   0.022939061  −0.0254912923
tpeakbps  −0.0081181662  −0.0013325295  −0.020142259   0.0032913902
tpeakbpd  −0.0111369375  −0.0140911285  −0.002362359   0.0210386963
trestbpd   0.0361591182  −0.0007605452  −0.024619507  −0.0320079275
exang     −0.1520069720   0.6009664626  −0.429942733  −0.4604614290
xhypo     −0.3341779201   1.5596277090   1.378414692   1.2645700698
oldpeak    0.1908935156  −0.0967478433   0.388405128   0.2164803514
cmo        0.0277780256   0.2412607985   0.073061630  −0.1960612646
cday      −0.0151970362  −0.0148062415  −0.026759876  −0.0152737194
cyr        0.2318645756   0.6265312268   0.095369608  −0.6515047003
ladprox    1.7277441803  −1.2217191986   0.314767813  −0.1033884127
laddist    1.1672585854  −0.8125308678  −0.278617954  −0.6830410921
diag       0.2504224578  −0.3244353820  −0.673472037   0.3895221999
cxmain     1.7362037233  −0.3507318707  −0.017925974   0.4765159079
ramus      1.1291841912   0.2856883362   0.286684835   1.0561701368
om1       −0.1893535303   3.8891176666  −0.704398542   0.5598622966
om2        0.4071110870  −2.5992153764  −0.676202349  −0.4854881569
rcaprox    1.4316751371  −0.7666014034  −0.317212280  −0.2971793102
rcadist    1.0688251201  −0.2259274588  −0.629453212   0.1276147609

Proportion of trace:
   LD1    LD2    LD3    LD4
0.6579 0.1599 0.1086 0.0736
```

As per usual, the LDA analysis of my Switzerland model produced 4 different LDA classifiers which I once again turned into historgrams 12 so that I could better analysis them. Straight away we can see that the **first LDA classifier** was able to seperate types *0 & 1* from types *2, 3 & 4* while the **second LDA classifier** was able to seperate type *4* from types *0, 1, 2 & 3*. The **third LDA classifier** is also able to seperate type *4* from the rest whilst sadly the **fourth LDA classifier** is seemingly unable to differentiate between any of the 5 types.

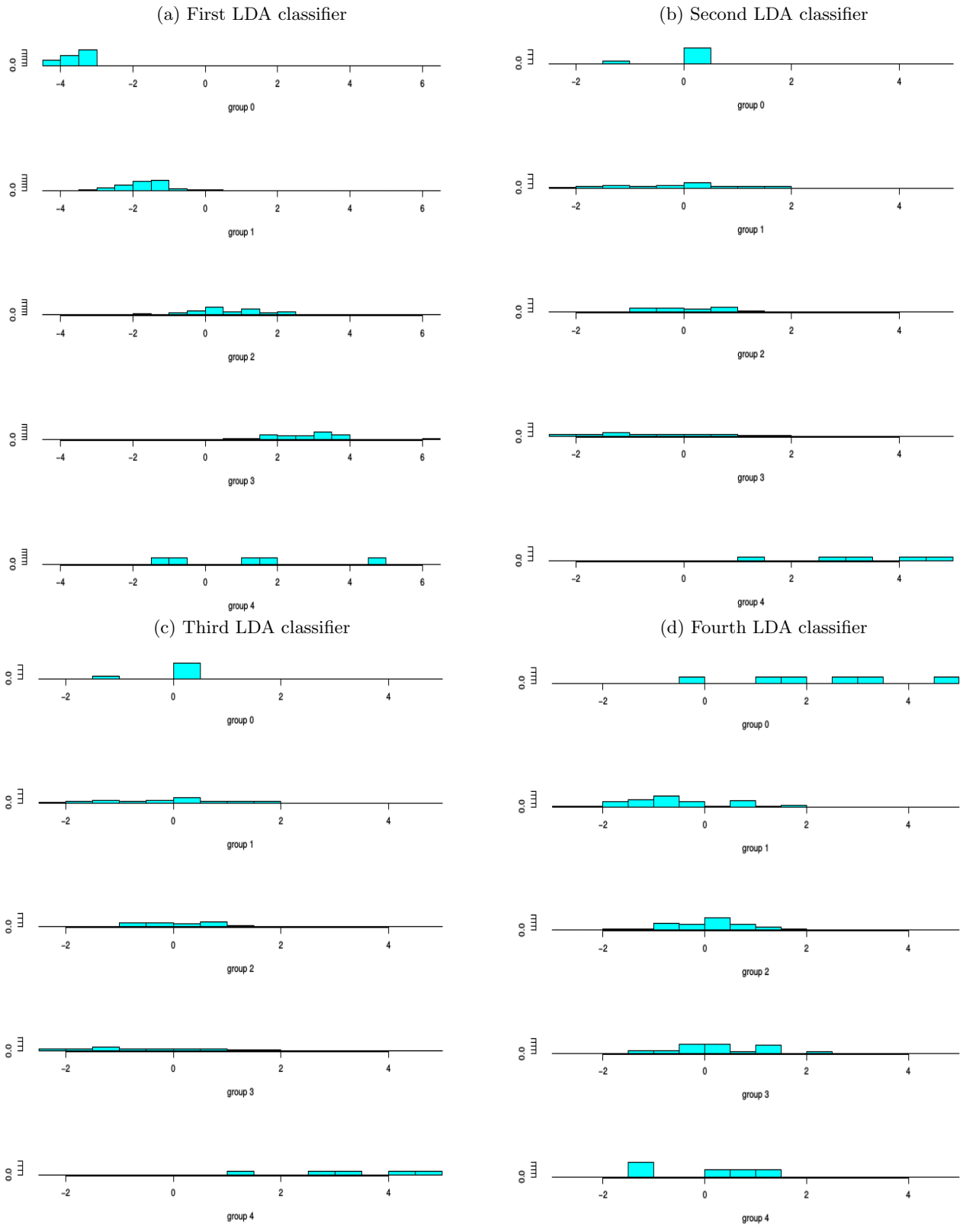


Figure 12: Histergrams for discriminant analysis of Switzerland

3 Summary

Overall we believe that this trial has shown that the type and intensity of the exercise that you do has more of an effect on you getting a heart disease than whether or not you take performance enhancing drugs. Although there were 4 datasets to analysis, a lot of the observations where NA which limited the amount of observations we were able to use in our research which makes it more difficult to draw any meaningful conclusions from our results. For instance, in **question four** we had to remove the *lmt* form our models for **Cleveland** and **Switzerland** but not **Hungary** or **Longbeach** in order for the LDA to work. This could be the reason why we were unable to draw any meaniful conclusions from their historgrams 10 & 11 however it could also be the fact that there were not that many observations avalible to use.