MA3505 Multivariate Statistics Project 1

April 27, 2016

1 Introduction and exploratory data analysis for the variables.

2 Analysis to answer each research question

2.1 Question 1

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 nessarcery 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** reponse was **proto**. As *chol* refers to the amount of cholestoral 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 importent as in theory the higher the intensity of the your exercise program the lower your chloestrol will be. The secound most importent 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 cholestrol) 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 importent 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 theuse if 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 supprise 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 likley 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 varing athaletic abilities for each test inorder 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 achives during their exercise test and as such it is no surprise that the response variable that is the most signicant 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 surprissing that *proto* is only a 2* rather than a 3* signifcance level as your maximum heart rate will depend on how athletic you are, the more atheletic the lower your max heart rate will be. **diuretic** is the other significant response variable and it refers to whether or not the subiect uses diuretic used used during exercise. Diuretic is considered to be a performance enhancing drug so it is therefore no surprise that it only has a 1* signicance level due to the fact that the analysis up to now has shown that there is a high probability that atheletes 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, sandich meats and salami* which could indercate that they are not very atheletic 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 digitails 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 cholestrol

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 refering 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 luckly 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

2.3.1 Cleveland

From running variance inflation factor we get the following

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				

Here we see the variables, ekgmo, ekgyr, cmo and cyr are collinear with other variables in the model.

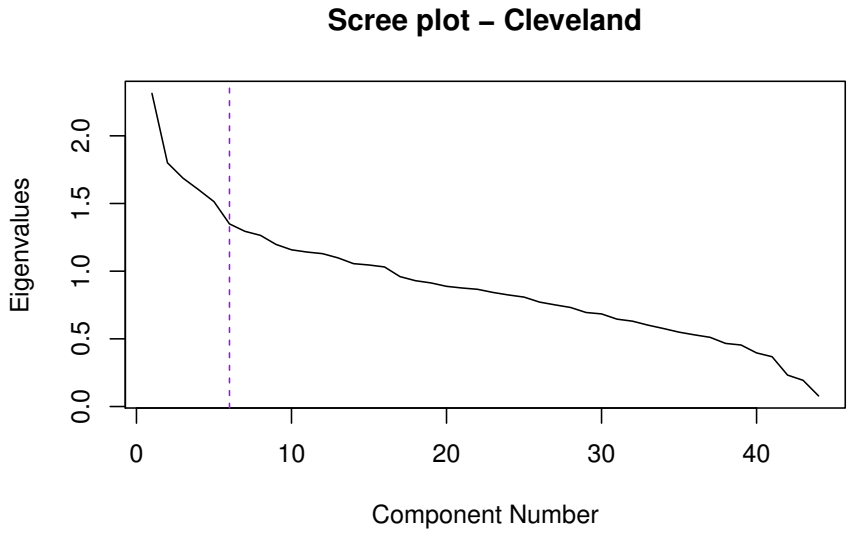

Figure 1: Scree plot for PCA of Cleveland

From the scree plot in Figure 1 we see that we keep 6 components.

We have the loadings of each components as follows.

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

2.3.2 Hungary

Figure 2: Scree plot for PCA of Hungary

From the scree plot in Figure 2 we see that we keep 6 components.

We have the loadings of each components as follows.

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

2.3.3 Longbeach

Figure 3: Scree plot for PCA of Longbeach

From the scree plot in Figure 3 we see that we keep 11 components.

We have the loadings of each components as follows.

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		

2.3.4 Switzerland

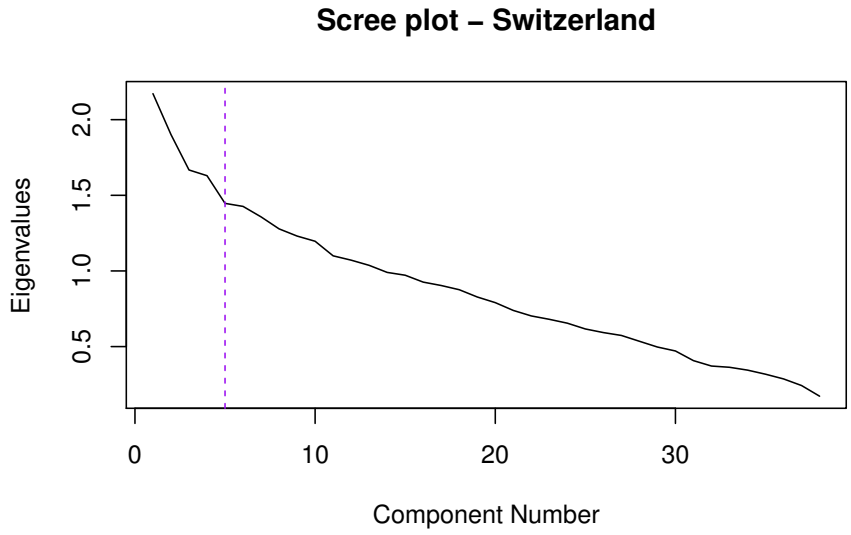

Figure 4: Scree plot for PCA of Switzerland

From the scree plot in Figure 4 we see that we keep 5 components.

We have the loadings of each components as follows.

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	

2.4 Question 4

2.4.1 Cleveland

2.4.2 Hungary

2.4.3 Longbeach

2.4.4 Switzerland

3 Summary