Data analytics report

Group members: Mingyang Zhang, Dongying Chen, Ruotong Xu

Data Exploration

In the dataset, there are 164 (54%) people who have heart disease 139 (46%) people who don't have heart disease.

Furthermore, by plotting charts, we found some relationship between other factor and probability of heart disease:

1. Heart disease usually develops between the ages of 50 and 60,
2. Heart disease is more prevalent in men.
3. People who have asymptomatic chest pain are most likely to have heart disease.
4. No obvious connection was found betweean resting blood pressure, serum cholesterol and probability of heart disease.
5. Higher fasting blood sugar is related to higher possibility of heart disease.
6. People having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) are most likely to have heart disease, followed by people showing probable or definite left ventricular hypertrophy by Estes' criteria.
7. Lower maximum heart rate achieved will cause a higher possibility of heart disease.
8. Exercise induced angina and higher ST depression induced by exercise relative to rest seem to be factors of influence on heart disease.
9. Flat and decreasing slope of the peak exercise ST segment would cause a higher probability of heart disease.
10 The more the number of major vessels (0-3) colored by fluoroscopy are, the higher the possibility of getting a heart disease is.
11. Reversable defect and fixed defect (6&7 in thal) could also be the influential factors.
12. Our group decide to explore our data by “Hierarchical Clustering”. Firstly, we use scale() command to make every variable equal important, and then we choose “complete” as our method to define similarity. In this case, we try to create four clusters and choose continuous variables such as age, trestbps, chol, thalach, and oldpeak in order to see what relationships among these variables. However, there are no obvious relationships among these variables when we use “Hierarchical Clustering”.
13. We first plot trestbps with other variables, and guess which variables have relationships with trestbps. Based on the plots, trestbps maybe has relationships with age, chol, thalach and oldpeak.

Hypotheses

From the data exploration phase, we plan to further clarify the relationships between age, sex, cp, fbs, Restecg, thalach, exang, oldpeak, slope, ca, thal and num with modeling. We also want to find out the relationships among all the predictors.

Hypothesis 1: According to the Exhibit 2, we find age might have some connection with num. What’s more, the graph shows that the probability of heart disease increase with the growth of age. And people aging from 60-70 years old are most likely to have heart disease.

Hypothesis 2: According to Exhibit 3, we guess that men have higher possibility to get heart disease than women.

Hypothesis 3: According to Exhibit 4, we find out that asymptomatic chest pain is closely related to heart disease, which we will focus on in the modeling.

Hypothesis 4: According to Exhibit 5 and 11, trestbps and oldpeak may have connections with num.
Hypothesis 5: The change in trestbps may lead to the changes on other variables.

Methodologies

1. Relationship between num and other variables

We used logistic regression with two coefficient estimation method to find relationship between num and other variables. The method to estimate coefficients and to select variables are traditional maximum likelihood method, and LASSO & cross validation method. The aim to use LASSO & cross validation is to avoid Multicollinearity.

(1) Maximum likelihood method

First, initial processing of data. We did the initial processing of data, removing all the observations with any NA and converting some numeric variables to factor variables.

Second, selection of independent variables. We ran a logistic regression with num as the dependent variable and all the other variables as independent variables. After reviewing the summary of result, we deleted some variables which had high p-values and left other variables with low p-value.

Third, establishment of final logistic model. In order to test the accuracy of our model, we evenly split the data into two subsets with a seed set, one was a training dataset, another was a testing dataset. Then, we ran the logistic regression with selected variables on the training data.

Fourth, testing the accuracy of model. We firstly predicted the probability of num=1 for the test data. Based on the assumption, that if the probability is beyond 50%, num=1, we computed the percentage of people the prediction was correct.

(2) LASSO & Cross validation

The correlation between independent variables are high, which we can see from Exhibit 32. Therefore, LASSO is a good method to avoid multicollinearity.

The first step of using LASSO & cross validation method were the same as using maximum likelihood method.

Second, splitting data to train subset and test subset. Like what we did in maximum likelihood method. We evenly split the data into two subsets with a seed set, one was a training dataset, another was a testing dataset.

Third, using cross validation to choose the best lambda of LASSO. We ran a logistic regression using cv.glmnet(), after which we could get a best lambda that minimized the MSE.

Fourth, using LASSO to select variables and estimate coefficients. We ran glm() with best lambda to get a best fitting model. Then the coefficient can be draw from coef().

Fourth, testing the accuracy of model. As we did in maximum likelihood method, we firstly predicted the probability of num=1 for the test data, setting those larger than 0.5 to 1 and those smaller than 0.5 to 0. At last, we computed the percentage of people the prediction was correct.

2. Relationship between trestbps and other predictors.

Multiple Liner regression (LASSO, Cross Validation)

In this model, we would like to test relationship of trestbps with other variables, and we decide to use LASSO to build model. Firstly, we use cross validation to choose the best lambda of LASSO, and the best lambda is 1.400339. Then, we use glmnet() commend to run LASSO, and get the coefficients of best multiple liner regression model.

Results and Conclusion

(1) Logistic Regression Results（Maximum likelihood method）

Exhibit 15 displays the logistic regression result that sex, cp, thalach, exang, ca and thal have significant p-values, so we use these predictors and omit other ones to run the final logistic regression. According to Exhibit 16, we find out the results as follows. A unit increase in the men increases the odds of getting a heart disease by e^ 1.024767= 2.7864 times. A unit increase in the cp4 increases the odds of getting a heart disease by e^ 1.388905= 4.0106 times. A unit increase in the thalach increase the odds of getting a heart disease by e^ -0.024627= 0.9757 times. A unit increase in the exang1 increase the odds of getting a heart disease by e^ 0.876158= 2.4017 times. A unit increase in the ca1 increase the odds of getting a heart disease by e^ 1.706684= 5.5107 times. A unit increase in the ca2 increase the odds of getting a heart disease by e^ 2.639768 = 14.0100 times. A unit increase in the ca2 increase the odds of getting a heart disease by e^ 2.094642= 8.1225 times. A unit increase in the thal7 increase the odds of getting a heart disease by e^ 1.697189= 5.4586 times. We compute the percentage of time the prediction was correct. The mean is 0.8187919.

(2) Logistic Regression Results（LASSO & Cross validation）

Based on the Exhibit31, we notice that the probability of heart disease has a negative relationship with age. The odds of male who get heart disease are e^ (0.351) = 1.419 times higher than the odds of female who get heart disease. The odds of people who have cp3 are e^(-0.245) = 0.78 times higher than the odds of people who have cp1. The odds of people who have cp4 are e^(2.089) = 8.08 times higher than the odds of people who have cp1. The number of trestbps has a positive relationship with the probability of heart disease. The odds of people who have restecg2 are e^(0.568) = 1.7647 times higher than the odds of people who have restecg1. The probability of heart disease has a negative relationship with thalach and positive relationship with exang1. The odds of people who have ca1 are e^(2.17) = 8.758 times higher than the odds of people who have ca0. The odds of people who have ca2 are e^(3.01) = 20.08 times higher than the odds of people who have ca0. The odds of people who have ca3 are e^(1.8487) = 6.35 times higher than the odds of people who have ca0. The odds of people who have thal6 are e^(0.98) = 2.66 times higher than the odds of people who have thal3. The odds of people who have thal6 are e^(1.1249) = 3.08 times higher than the odds of people who have thal3. The mean is 0.8120805.

(3) Linear Regression Results (trestbps and other variables)

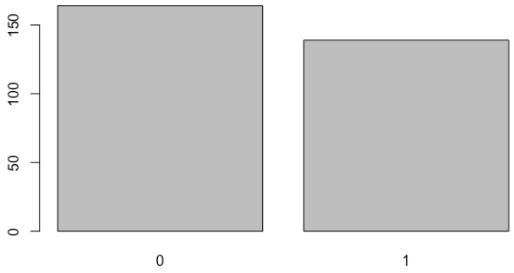
Then we do the modeling (multiple liner regression – lasso). Based on the summary of our lasso about the relationship among trestbps and other variables, we could find that there is a significant relationship among trestbps with age, chol, fbs, Restecg2, oldpeak, ca, and num. Specific relationship is that when age increase 1, trestbps will increase 0.36, when chol increase 1, trestbps will increase 0.0026, when Restecg is 2, trestbps will increase 0.084, when oldpeak increase 1, trestbps will increase 1.1, when ca increases 1, trestbps will decrease 0.58, and when num is 1, trestbps will increase 0.015.

(4) Conclusion

Through the model provided above, more than 80% of heart disease prediction accuracy can be achieved. The model can be used to make predictions and assist doctors in preventing and treating heart disease. The significant variables are age, sex1, cp3, cp4, trestbps, restecg2, thalach, exang1, slope2, slope3, ca1, ca2, ca3, thal6, thal7.

Appendix

how many people have heart disease
Exhibit 1



the relationships between all the factors and heart disease
Exhibit 2

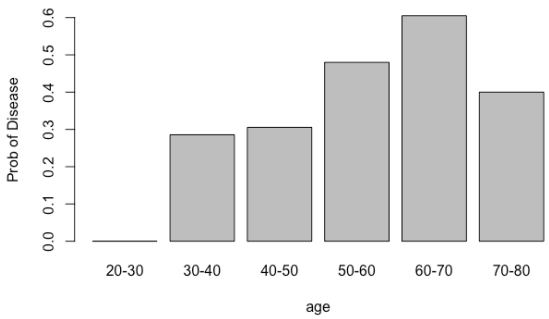


Exhibit 3

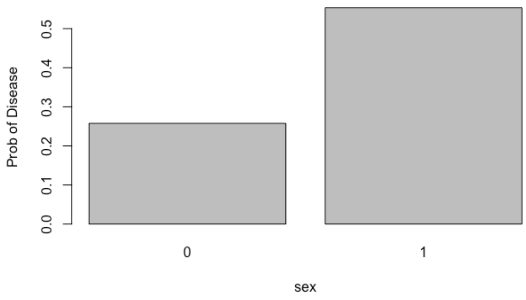


Exhibit 4

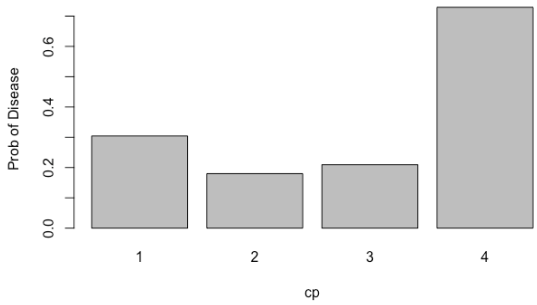


Exhibit 5

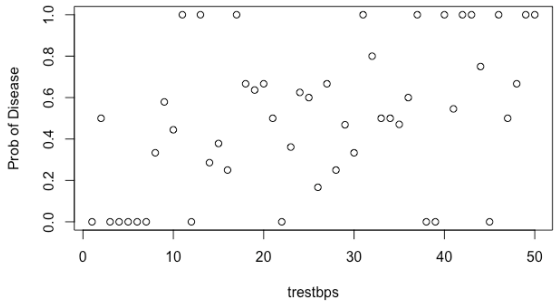


Exhibit 6

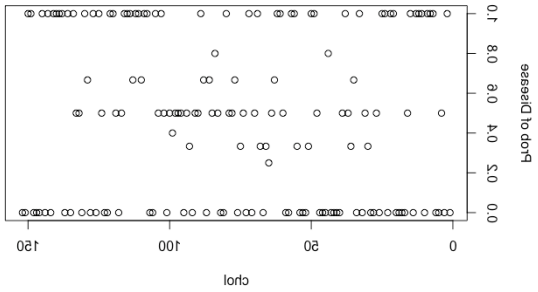


Exhibit 7

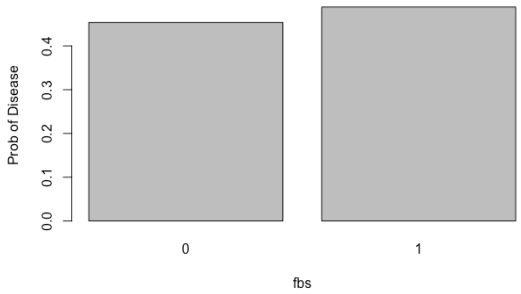


Exhibit 8

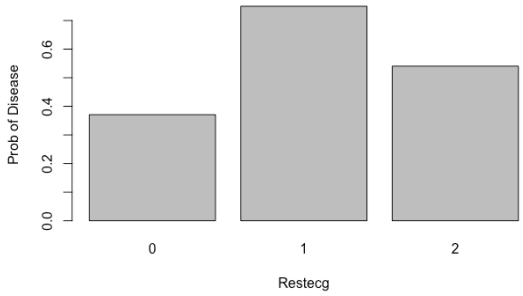


Exhibit 9

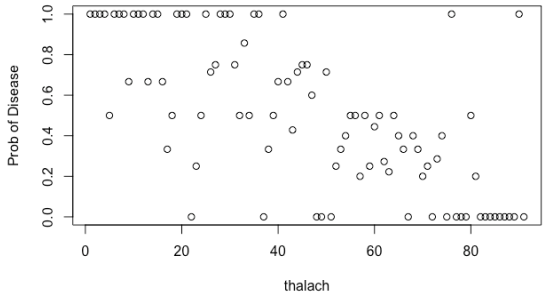


Exhibit 10

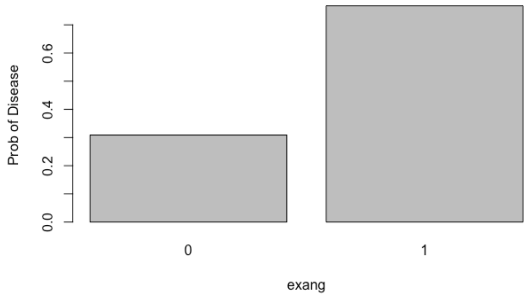


Exhibit 11

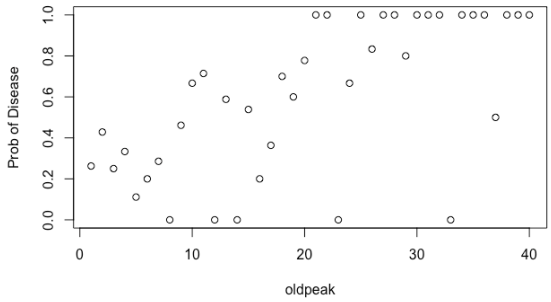


Exhibit 12

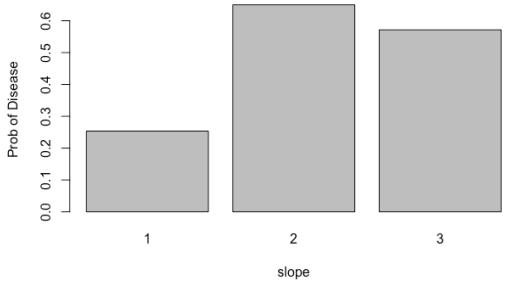


Exhibit 13

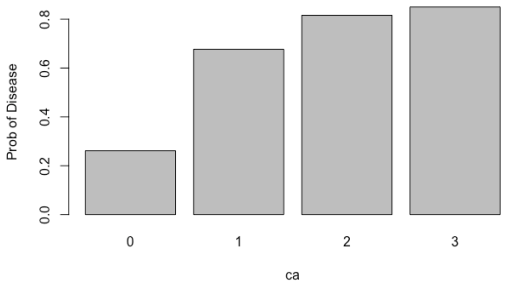


Exhibit 14

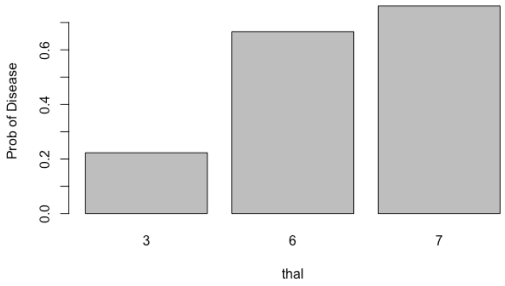


Exhibit 15

```
Call:
glm(formula = num ~ ., family = binomial, data = heart.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0490  -0.4847  -0.1213   0.3039   2.9086

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.253978   2.960399  -2.113 0.034640 *
age         -0.023508   0.025122  -0.936 0.349402
sex1         1.670152   0.552486   3.023 0.002503 **
cp2          1.448396   0.809136   1.790 0.073446 .
cp3          0.393353   0.700338   0.562 0.574347
cp4          2.373287   0.709094   3.347 0.000817 ***
trestbps     0.027720   0.011748   2.359 0.018300 *
chol         0.004445   0.004091   1.087 0.277253
fbs1        -0.574079   0.592539  -0.969 0.332622
Restecg1     1.000887   2.638393   0.379 0.704424
Restecg2     0.486408   0.396327   1.227 0.219713
thalach     -0.019695   0.011717  -1.681 0.092781 .
exang1       0.653306   0.447445   1.460 0.144267
oldpeak      0.390679   0.239173   1.633 0.102373
slope2       1.302289   0.486197   2.679 0.007395 **
slope3       0.606760   0.939324   0.646 0.518309
ca1          2.237444   0.514770   4.346 1.38e-05 ***
ca2          3.271852   0.785123   4.167 3.08e-05 ***
ca3          2.188715   0.928644   2.357 0.018428 *
thal6       -0.168439   0.810310  -0.208 0.835331
thal7        1.433319   0.440567   3.253 0.001141 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 409.95  on 296  degrees of freedom
Residual deviance: 183.10  on 276  degrees of freedom
AIC: 225.1

Number of Fisher Scoring iterations: 6
```

Exhibit 16

```
Call:
glm(formula = num ~ sex + cp + thalach + exang + ca + thal, family = binomial,
    data = heart.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0142  -0.5266  -0.2105   0.4304   2.3808

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.424111   1.529896   0.277 0.781615
sex1         1.024767   0.441806   2.319 0.020368 *
cp2          0.100994   0.678154   0.149 0.881613
cp3         -0.331350   0.631348  -0.525 0.599702
cp4          1.388905   0.589992   2.354 0.018567 *
thalach     -0.024627   0.009266  -2.658 0.007868 **
exang1       0.876158   0.411613   2.129 0.033288 *
ca1          1.706684   0.445860   3.828 0.000129 ***
ca2          2.639768   0.606714   4.351 1.36e-05 ***
ca3          2.094642   0.774492   2.705 0.006840 **
thal6        0.461450   0.724834   0.637 0.524366
thal7        1.697189   0.396950   4.276 1.91e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 409.95  on 296  degrees of freedom
Residual deviance: 211.62  on 285  degrees of freedom
AIC: 235.62

Number of Fisher Scoring iterations: 5
```

Exhibit 17

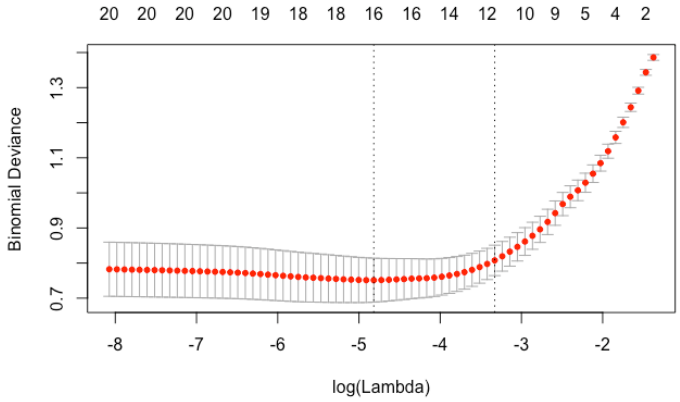


Exhibit 18

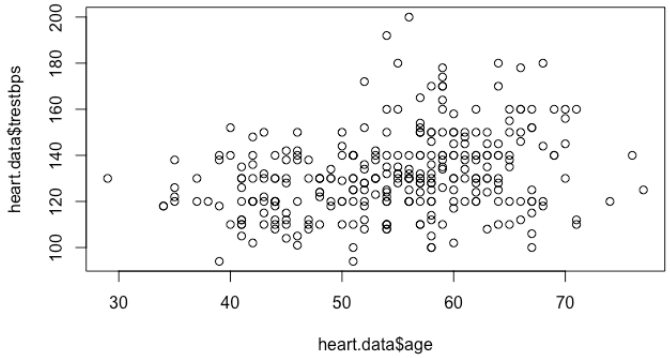


Exhibit 19

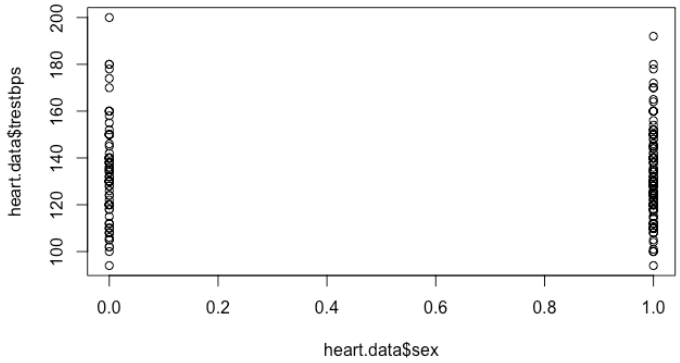


Exhibit 20

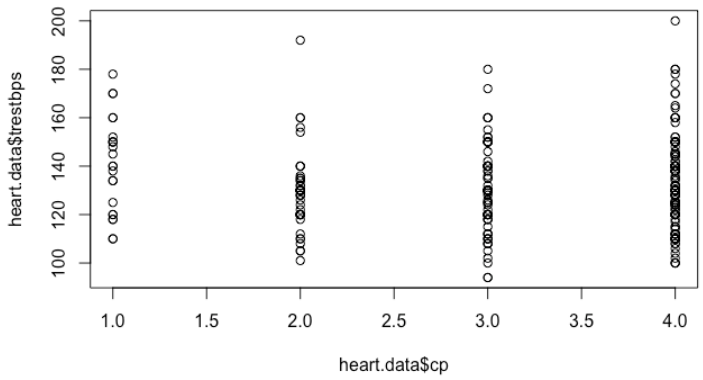


Exhibit 21

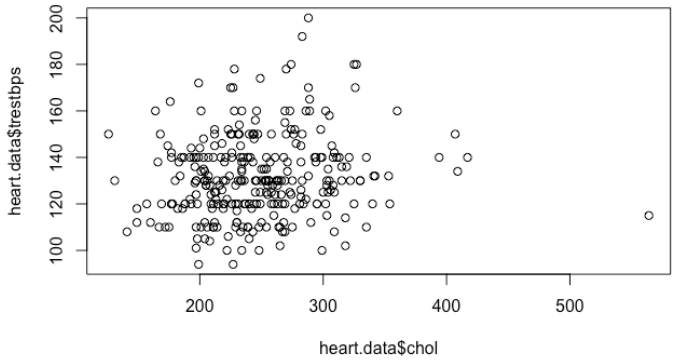


Exhibit 22

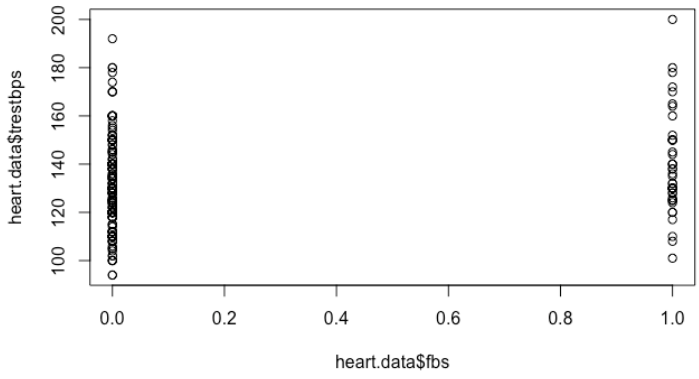


Exhibit 23

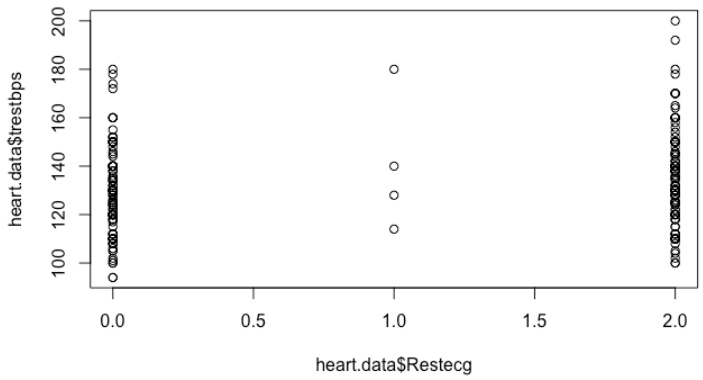


Exhibit 24

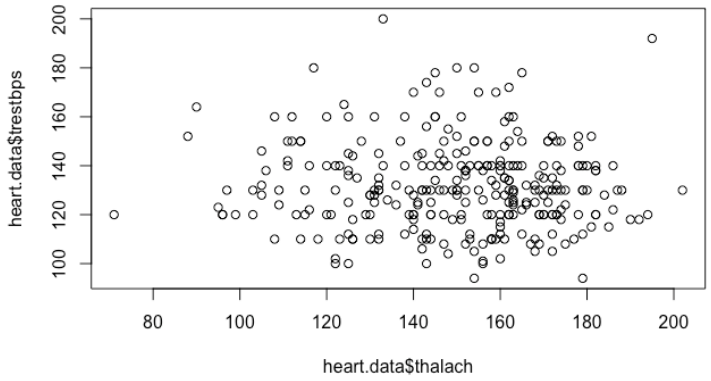


Exhibit 25

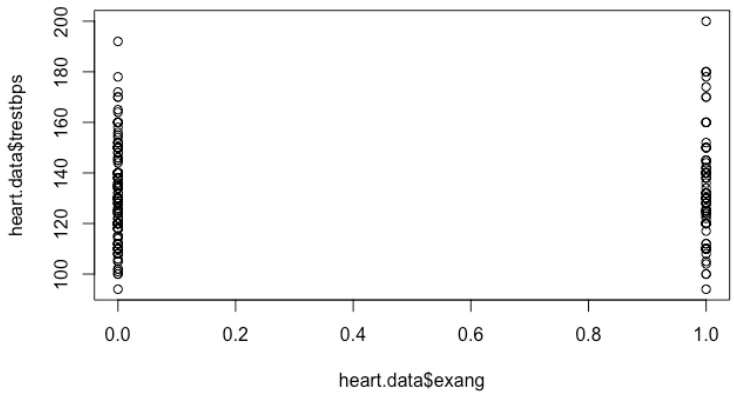


Exhibit 26

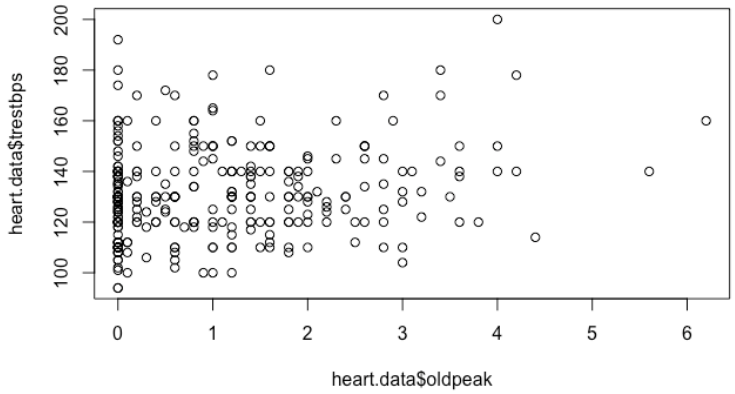


Exhibit 27

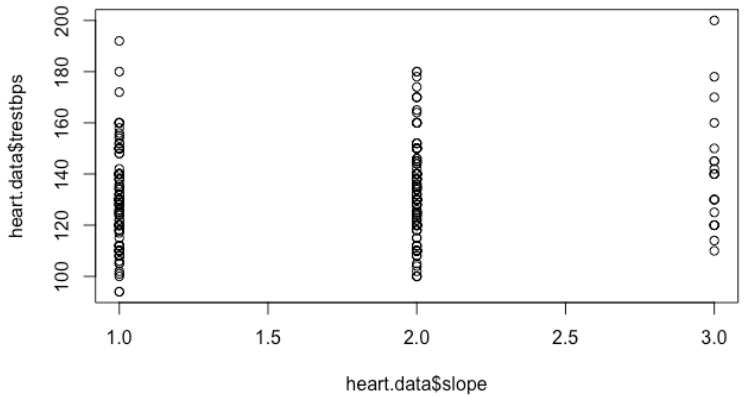


Exhibit 28

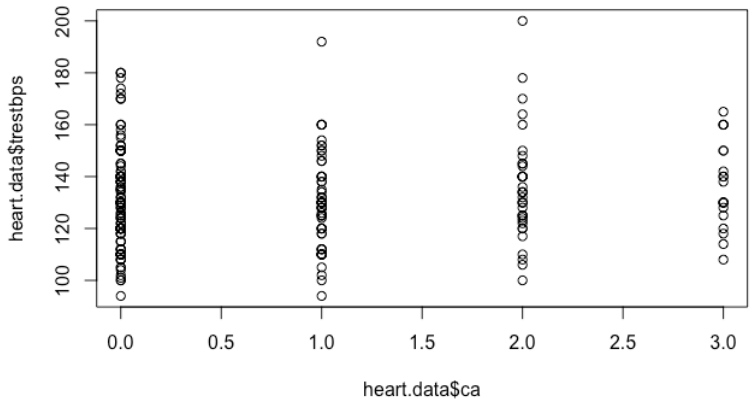


Exhibit 29

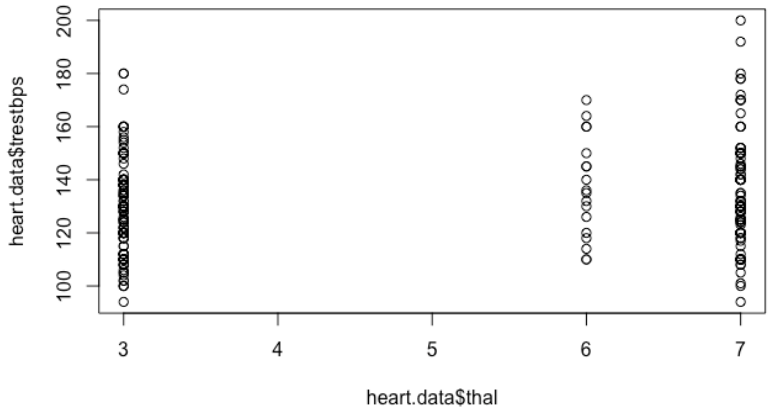


Exhibit 30

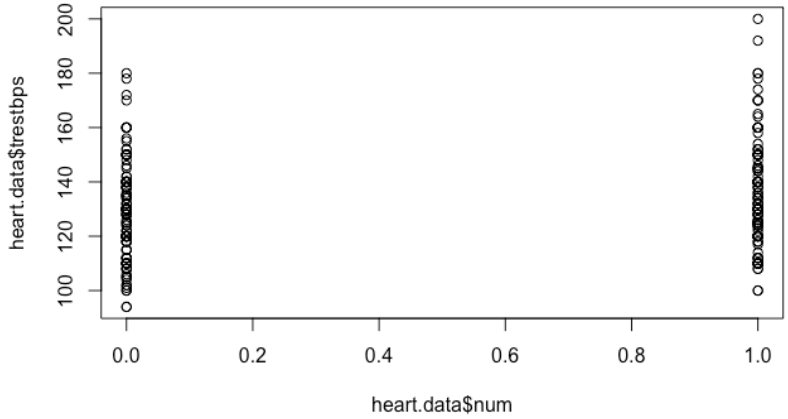


Exhibit 31

21 x 1 sparse Matrix of class "dgCMatrix"

	s0
(Intercept)	-1.12963528
age	-0.04392177
sex1	0.35112660
cp2	.
cp3	-0.24460656
cp4	2.08975457
trestbps	0.02663212
chol	.
fbs1	.
restecg1	.
restecg2	0.56800415
thalach	-0.02537498
exang1	0.25026809
oldpeak	.
slope2	1.27278074
slope3	0.20717447
ca1	2.17374687
ca2	3.01256765
ca3	1.84870682
thal6	0.98280734
thal7	1.12493177

Exhibit 32 Correlation Matrix

	sex1	cp2	cp3	cp4	thalach	exang1	ca1	ca2	ca3	thal6	thal7
sex1	1.00	-0.04	-0.12	0.09	-0.06	0.14	0.10	-0.02	0.07	0.15	0.33
cp2	-0.04	1.00	-0.28	-0.43	0.26	-0.23	-0.06	-0.09	-0.08	-0.04	-0.20
cp3	-0.12	-0.28	1.00	-0.60	0.16	-0.26	0.03	-0.19	-0.02	-0.10	-0.16
cp4	0.09	-0.43	-0.60	1.00	-0.38	0.45	0.05	0.22	0.12	0.10	0.30
thalach	-0.06	0.26	0.16	-0.38	1.00	-0.38	-0.20	-0.06	-0.18	-0.16	-0.21
exang1	0.14	-0.23	-0.26	0.45	-0.38	1.00	0.15	0.10	0.01	0.06	0.30
ca1	0.10	-0.06	0.03	0.05	-0.20	0.15	1.00	-0.20	-0.14	0.00	0.11
ca2	-0.02	-0.09	-0.19	0.22	-0.06	0.10	-0.20	1.00	-0.10	0.07	0.11
ca3	0.07	-0.08	-0.02	0.12	-0.18	0.01	-0.14	-0.10	1.00	0.04	0.12
thal6	0.15	-0.04	-0.10	0.10	-0.16	0.06	0.00	0.07	0.04	1.00	-0.20
thal7	0.33	-0.20	-0.16	0.30	-0.21	0.30	0.11	0.11	0.12	-0.20	1.00