Linear Regression Model of Maximum Heart Rate with Blood Pressure and Cholesterol.

Fahim Hoq

Content

Abstract

Cleveland dataset of 303 people was used to see if maximum heart rate is an association with resting blood pressure and if maximum heart rate has an association with serum cholesterol. Two different linear regression models was used to answer these questions. Other variables were also taken into both models as confounding and precision variables. In both cases, p-value was higher than 0.05, so at 5% level of significance, we did not have sufficient evidence to find an association for either models.

Introduction

"The Rhythm of the heart has not only fascinated cardiologists but also inspired poets and musicians." (Stauss H.M. 2000.) Heart rate variability is also related with cardiovascular disease, in this paper, we are modeling how the heart rate is effected by blood pressure and how its effected by

cholesterol. Since, blood pressure is an important predictors for cardiovascular mortality (Kikuya et. al 2000). Similarly, serum cholesterol having an effect on heart disease (Kannel, W. B. 1995). This paper tries to verify these claims using the dataset of 303 Cleveland heart patients. These questions are important as if there are association, then people might take more preventative measures in order to decrease their chance of developing heart failure by trying to better control their blood pressure, and cholesterol levels.

Data Description

The dataset was downloaded from kaggle. The data was collected by Robert Detrano, M.D., Ph.D of the Cleveland Clinic Foundation. The dataset has 14 variables. The response variable analyzed in this paper is the Maximum heart rate achieved. Sample size in the dataset is 303. Covariates of interest are Blood Pressure and Cholesterol. Other variables such as gender, age of patients, blood sugar, chest pain etc. were used as confounding and precision variables in the modeling. There were only six missing values, two for the variable nuclear stress test and four for the variable number of major vessels. Moreover, there are around twice as many male in the dataset, these could result in bias of our analysis with regards to the sex variable (R. Detrano 2020).

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
1	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
2	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
3	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
4	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
5	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0

Software Used for Analysis

In this paper, the data was analyzed using R-studio version 4.2.2. Library such as ggplot2 was used in plotting most of the figures in this paper. The library dplyr was used in cleaning and organizing the data. Moreover, the library car was used to find the variance inflation factor. Finally, libraries such as knitr, table1, htmlTable was used in making the table used in this paper. For more information please refer to the appendix and reference in this paper.

Explanatory Analysis

Histograms showed four out of the five continuous variables were approximately normal. Only, “depression induced by exercise relative to rest” variable was very right skewed. Bar plot of Gender showed that there are more than twice the number of male participants in the study compared to female. Scatter plot of hear rate with age showed a slightly downward slope. Which is slightly counter intuitive, perhaps there are confounding effect there. As we would naturally, expect older patients to have higher risk of cardiovascular disease, and hence have more maximum hear rate. Box plots showed all the categorical variables are distributed evenly with respect to maximum heart rate. Heat maps didn’t reveal any alarmingly high correlation between the variables. The variables “exercise relative to rest”, “exercise induced angina”, and “peak exercise segment” showed bit more correlation as expected.

Question of Interest

The questions of interest in the paper is if there is an association between blood pressure and maximum heart rate achieved and likewise, if there is an association between cholesterol and maximum heart rate achieved. The reason, these models are chosen are backed by literature review. According to Kikuya M. et. al blood pressure has an effect on heart rate variabilities (Kikiya M. et. al 2000). And, according to Kannel, W. B. serum cholesterol has an effect on developing coronary artery disease (Kannel, W. B. (1995). Age, is taken as a confounding variable as a older people is more vunerable to heart disease (Tsuji, H. et al. 1996). Similarly, chest pain. blood sugar have been taken as confounding

variable. Gender has been taken as a precision variable (Tan, Y. et al. 2010). Exercise induced angina has also been taken as a precision variable (D'Antono. et. al. 2006). For both my models, I have used Linear Regression models to answer both questions.

Modeling (2 models)

Model1: Does Blood Pressure has an effect on on maximum heart rate achieved.

$y_i \sim B_0 + B_1x_1 + B_2x_2 + B_3x_3 + B_4x_4 + B_5x_5 + B_6x_6 + B_7x_7 + \varepsilon_i$, $\varepsilon_i \sim N(0,\sigma^2)$

Maximum Heart Rate (yi) ~ Intercept + Blood Pressure(x_1) + age(x_2) + sex(x_3) + Blood sugar (x_4) + chest pain(x_5) + Exercise Relative Rest(x_6) + Exercise Induced Angina(x_7) + Peak Exercise (x_8) + Error Term.

Now we want to test the hypothesis H_0: $B_1 = 0$ v H_1: $B_1 \neq 0$

if B_1 is not equal to zero. Where, B_1 here is resting blood pressure in mm Hg.

Level of significance chosen is 5 %.

Likewise,

Model2: Does serum Cholesterol has an effect on maximum heart rate achieved.

$y_i \sim B_0 + B_1x_1 + B_2x_2 + B_3x_3 + B_4x_4 + B_5x_5 + B_6x_6 + B_7x_7 + \varepsilon_i$, $\varepsilon_i \sim N(0,\sigma^2)$

Maximum Heart Rate (y_i) ~ Intercept + Cholesterol(x_1) + age(x_2) + sex(x_3) + Blood sugar (x_4) + chest pain(x_5) + Exercise Relative Rest(x_6) + Exercise Induced Angina(x_7) + Peak Exercise (x_8) + Error Term.

Now we want to test the hypothesis: H_0: $B_1 = 0$ v H_1: $B_1 \neq 0$

B_1 is not equal to zero. Where, B_1 here is serum cholesterol in mg/dl

Level of significance chosen is 5%.

Assumptions and Diagnostic Plots

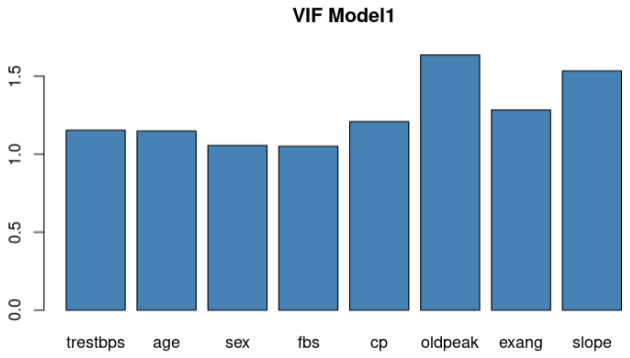
Assumptions of Linear models are the response or error term are independent and identically normally distributed with mean zero and constant variance. The diagnostic plot for both model shows the assumptions are satisfied. The residual verses fitted plot on both models indicates homoscedasticity as the points are spread almost evenly across. The quantile - quantile plot for both models are similar , they both show that the line is almost linear. The left side of the graph deviates slightly from straight but there is not a lot of data there. Hence, the normality assumption is satisfied. In the leverage plot, there does not appear to be any big outliers.

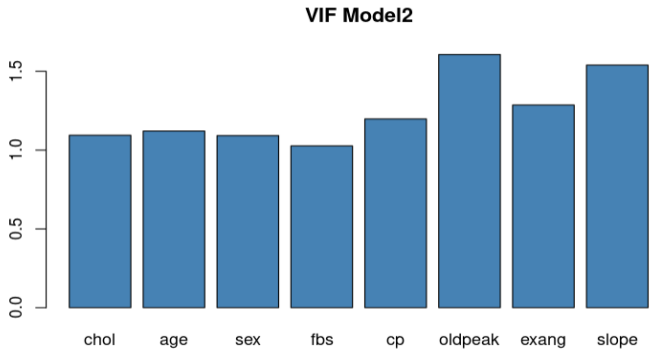
Multicollinearity

Variance Inflation Factor (VIF_i) was used to find multicollinearity.

Where, $VIF_i = 1 / (1 - R_j^2)$, where R_j^2 is the adjusted R-squared.

The variance inflation factor for all the variables were under 3 for all the variables. As none of them exceeded 5, multicollinearity is not a problem for this dataset either.





Conclusions & Interpretation

For model1, Provided every covariate is constant, increase of one beats per minute of maximum heart rate leads to 0.1173 mm Hg increase of resting blood pressure. However, blood pressure was not significant at 5% level of significance, it was only significant at 10% level of significance. For model2, Provided every covariate is constant, increase of one beats per minute of maximum heart rate leads to

0.0396 mg/dl increase of serum cholesterol. However, cholesterol too was not significant at 5% level of significance, it was only significant at 10% level of significance. Since, we initially, stated 5 level of significance at the beginning of the study, therefore, in both cases, we have to conclude that there was not sufficient evidence to reject the null hypothesis. Therefore, there appears to be no association of blood pressure with maximum heart rate, and there appears to be no association between serum cholesterol and maximum heart rate.

Summary of the 1st model

	Est	ci95.lo	ci95.hi	t value	Pr(>|t|)
(Intercept)	212.3488	191.3737	233.3239	19.9244	0.0000
trestbps	0.1173	-0.0094	0.2440	1.8220	0.0695
age	-0.8608	-1.1069	-0.6146	-6.8814	0.0000
sex	-1.4862	-6.0524	3.0801	-0.6405	0.5223
fbs	1.9696	-4.0058	7.9451	0.6487	0.5170
cp	-3.9774	-6.3551	-1.5998	-3.2923	0.0011
oldpeak	-1.2509	-3.5380	1.0361	-1.0765	0.2826
exang	-10.1487	-15.1557	-5.1417	-3.9891	0.0001
slope	-8.3929	-12.5656	-4.2203	-3.9586	0.0001

Summary of the 2nd model

	Est	ci95.lo	ci95.hi	t value	Pr(>|t|)
(Intercept)	217.2644	199.4999	235.0289	24.0700	0.0000
chol	0.0396	-0.0023	0.0816	1.8603	0.0638
age	-0.8483	-1.0914	-0.6052	-6.8673	0.0000
sex	-0.9439	-5.5853	3.6975	-0.4002	0.6893
fbs	2.8061	-3.0980	8.7103	0.9354	0.3504
cp	-4.2655	-6.6317	-1.8993	-3.5478	0.0005
oldpeak	-1.0397	-3.3057	1.2262	-0.9030	0.3672
exang	-10.2372	-15.2477	-5.2267	-4.0211	0.0001
slope	-8.2043	-12.3840	-4.0247	-3.8632	0.0001

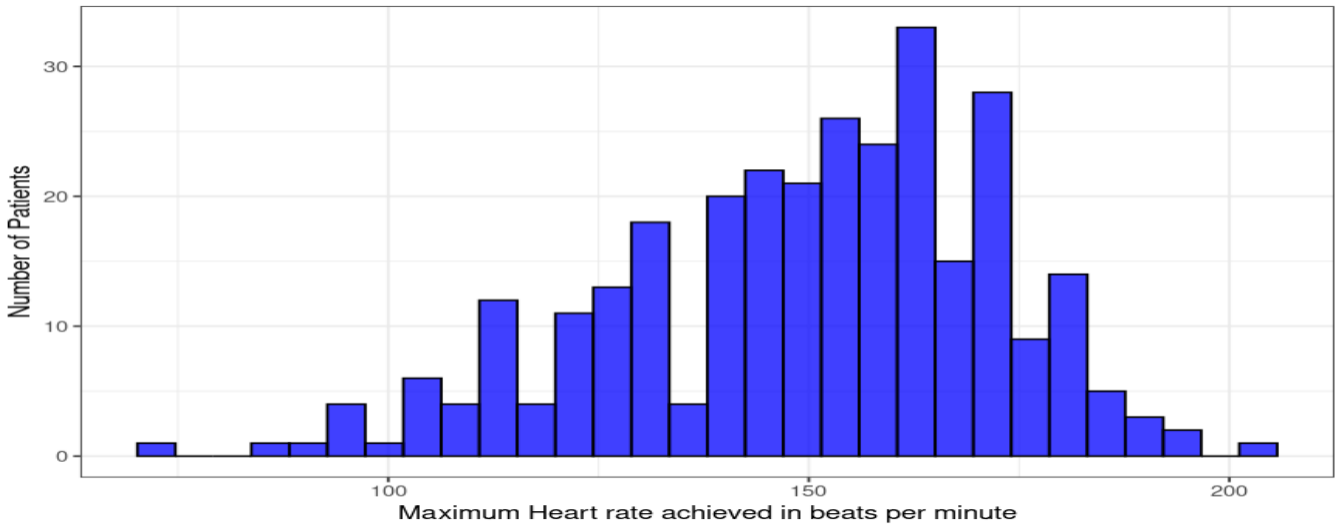
Limitations of this study and Future Study

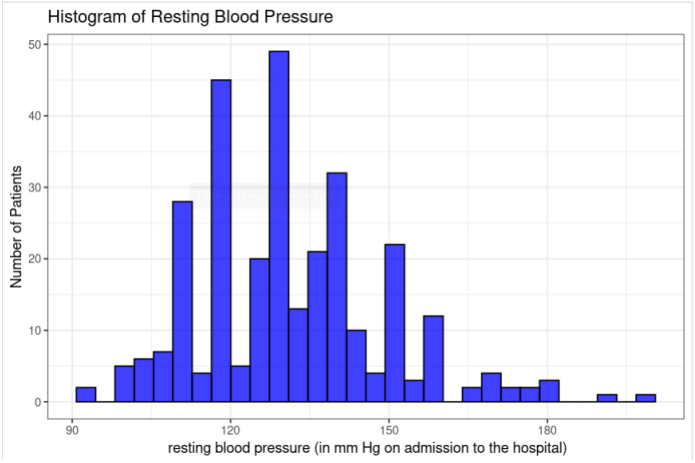
There are more than twice the number of male patients in the data compared to female. So even though gender was not significant on both models, it is hard to come to any conclusive statement. For future study, we might want to run the analysis again after collecting a dataset where the ratio of male and female participants is similar. Another limitation was that stylostic and diastolic blood pressure was not here in the data set. So even though, resting blood pressure did not indicate significance, perhaps reading for those would have indicated otherwise. Furthermore, sample size of 303 might not be low, but ideally, we might want higher power for 5% level of significance, hence bigger dataset would be nice. This is another thought for future study. Finally, we used a very simple linear model, for future study a more complex model with transformation for some covariates might also lead to interesting results.

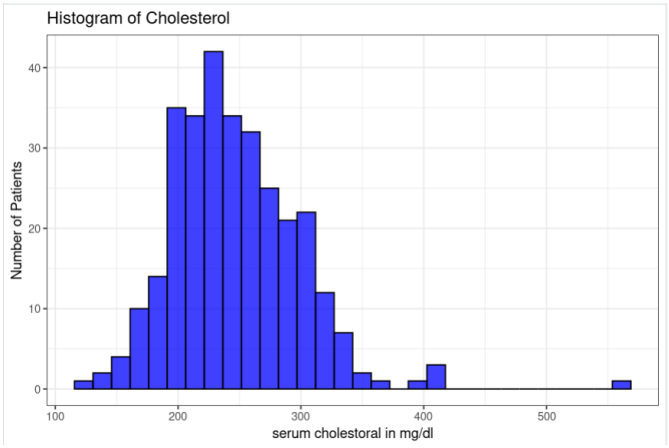
Moreover, Age was highly significant in both models. This is expected, Tsuji H et, al mentioned that age is related with cardiovascular disease (Tsuji et. al 1996). However the coefficient was negative. Which suggest that older patients is less likely to have maximum heart rate achieved. Another interesting find was that the peak exercise segment and exercise induced angina were highly significant on both model 1 and 2. However, those were not our original question of interest. Perhaps in a future study, we can do the analysis on how exercise effect maximum heart rate achieved. Furthermore, the dataset only has maximum heart rate achieved for patients, we need to also look at minimum heart rate achieved as well, as low heart rate per minute is also related to cardiovascular disease (Stauss H.M. 2000). Finally, this dataset did not have variables such as diabetes. So we might also want to include other variables in our future modeling for future study as these variables could have confounding effect.

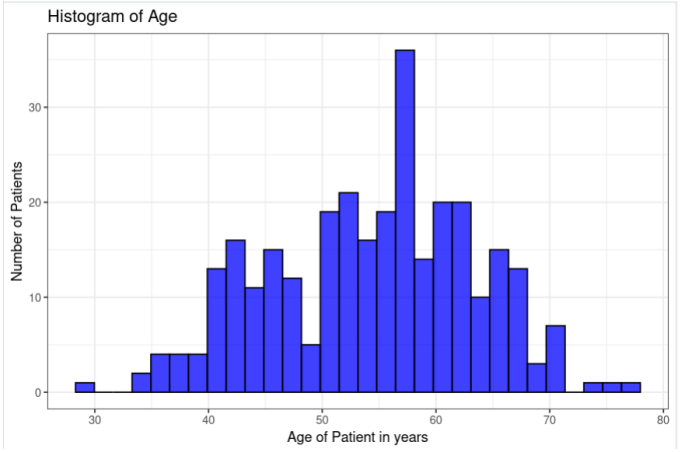
Appendix A (Plots)

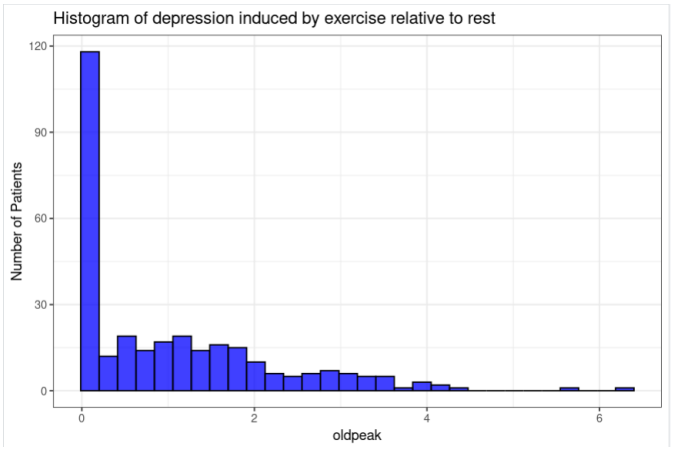
Histograms



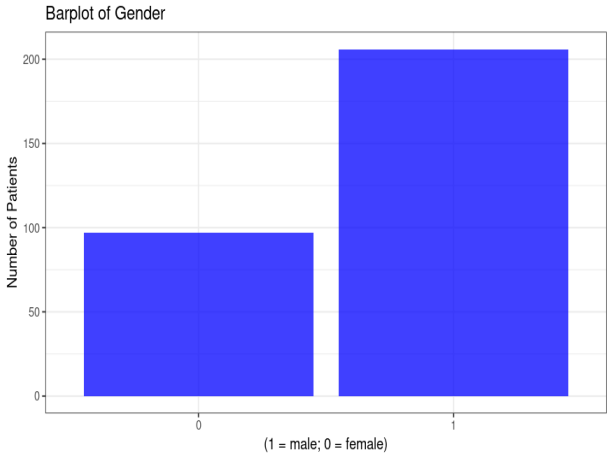


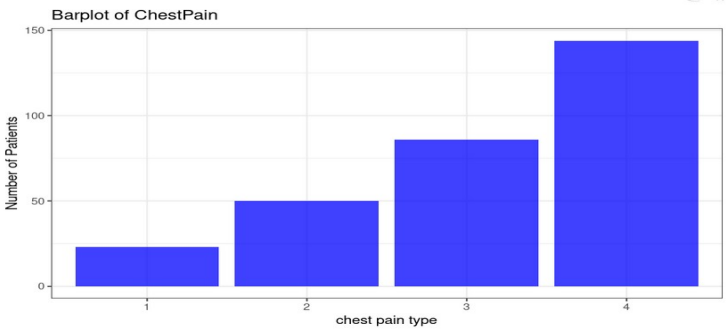


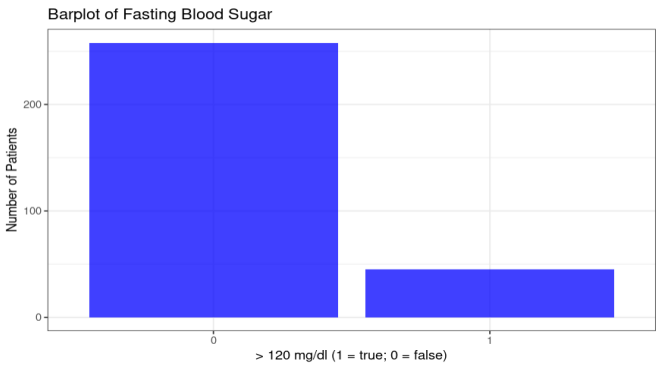


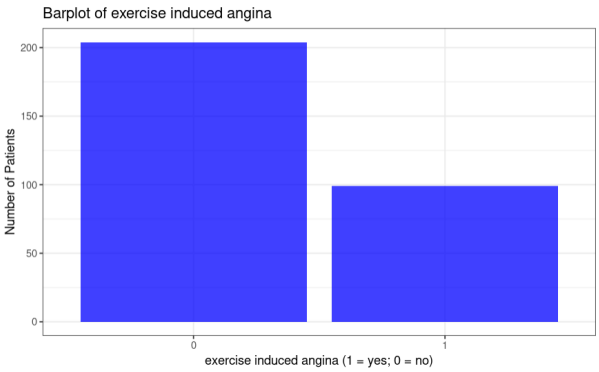


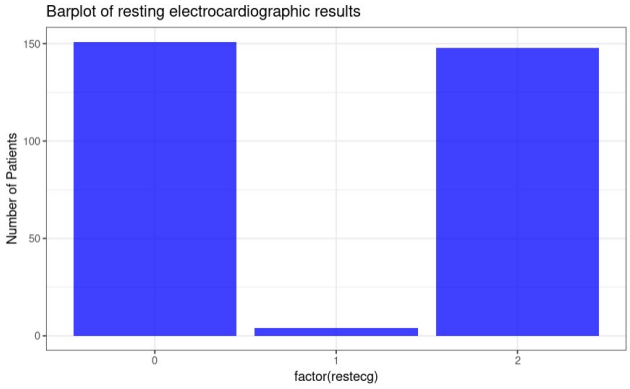
Bar Plots

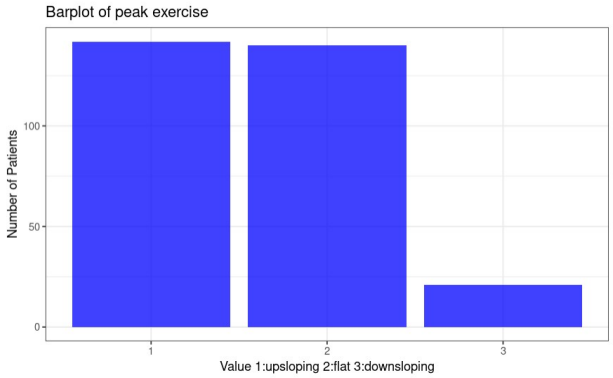

Barplot of Gender
Number of Patients
(1 = male; 0 = female)

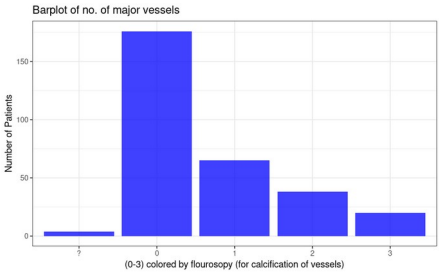

Barplot of ChestPain
Number of Patients
chest pain type

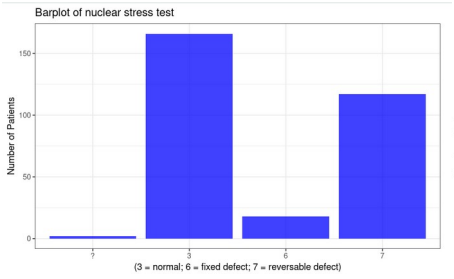

Barplot of Fasting Blood Sugar
Number of Patients
> 120 mg/dl (1 = true; 0 = false)

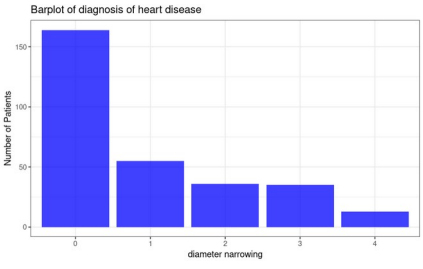

Barplot of exercise induced angina
Number of Patients
exercise induced angina (1 = yes; 0 = no)


Barplot of resting electrocardiographic results
Number of Patients
factor(restecg)

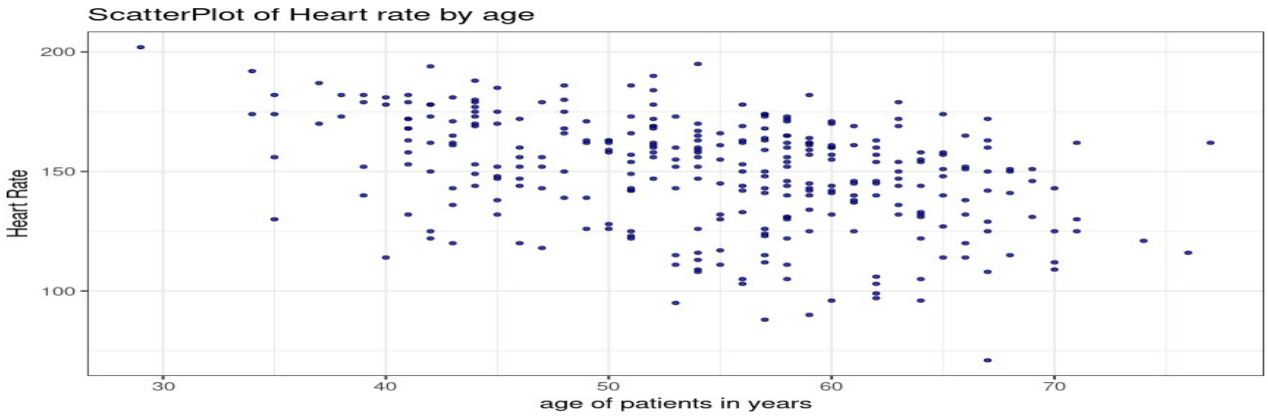

Barplot of peak exercise
Number of Patients
Value 1:upsloping 2:flat 3:downsloping

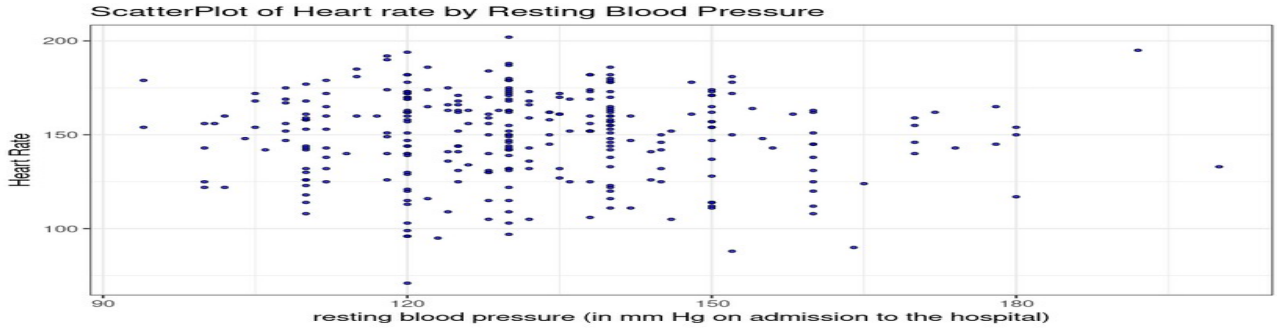

Barplot of no. of major vessels
Number of Patients
(0-3) colored by flourosopy (for calcification of vessels)

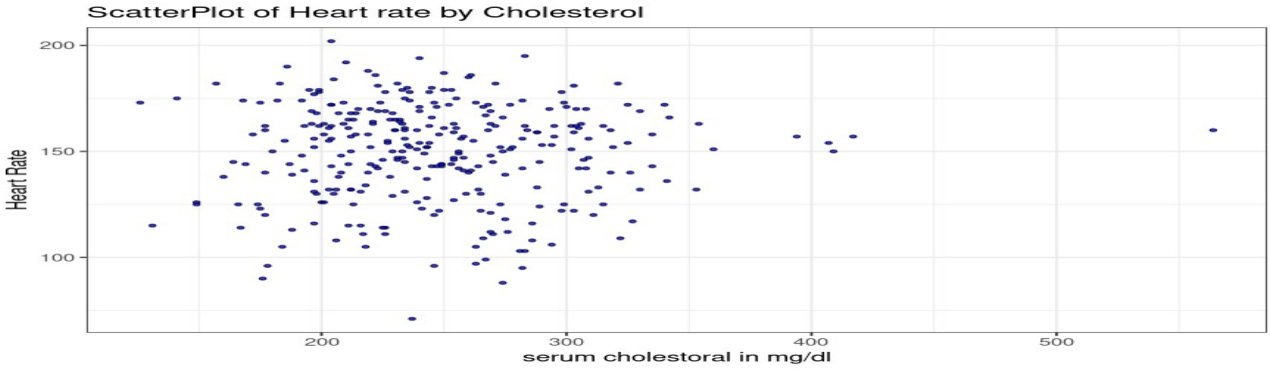

Barplot of nuclear stress test
Number of Patients
(3 = normal; 6 = fixed defect; 7 = reversable defect)

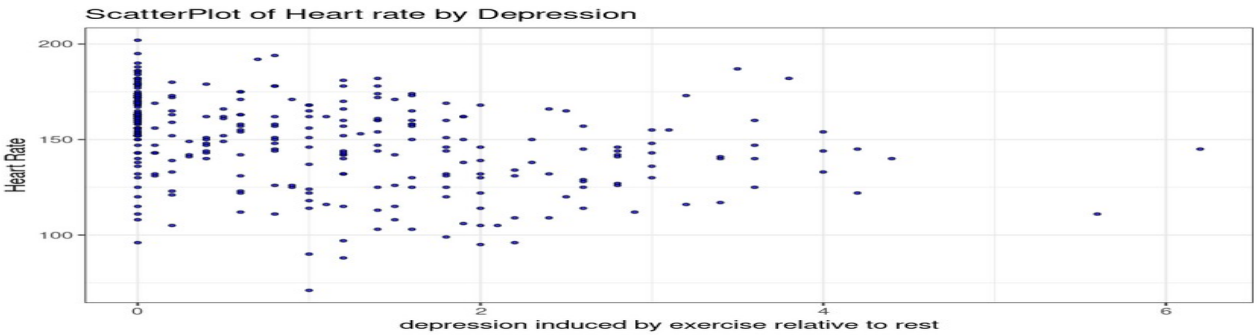

Barplot of diagnosis of heart disease
Number of Patients
diameter narrowing

Scatter Plots

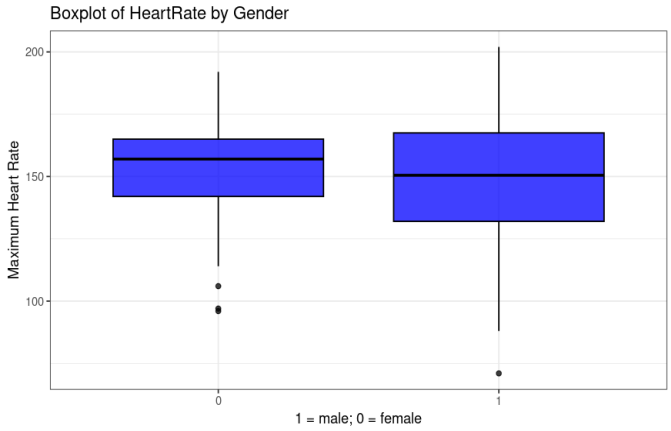


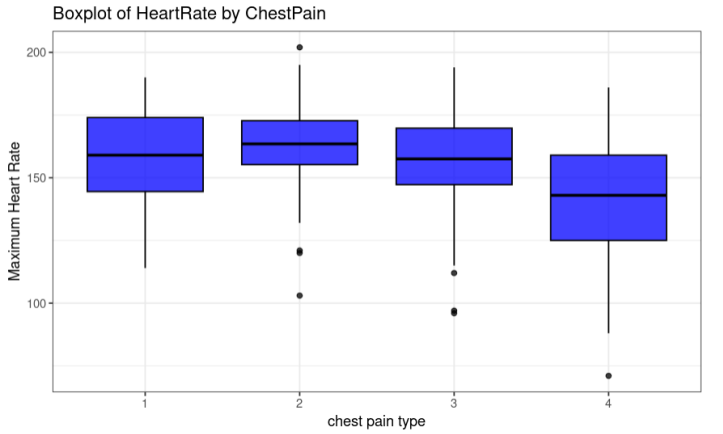


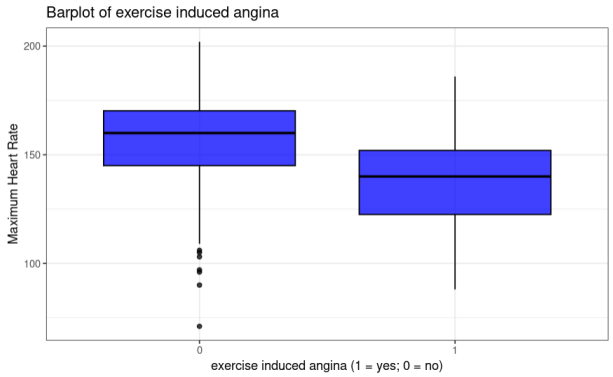


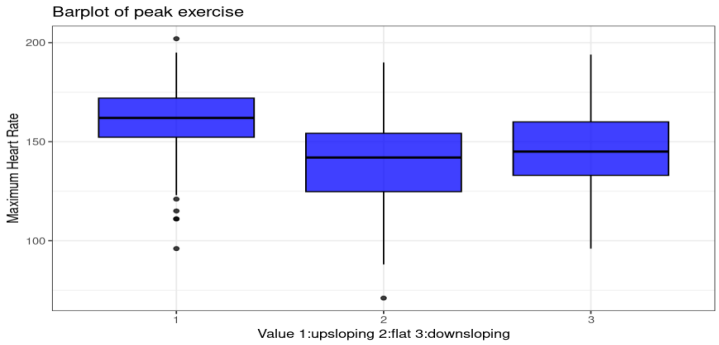


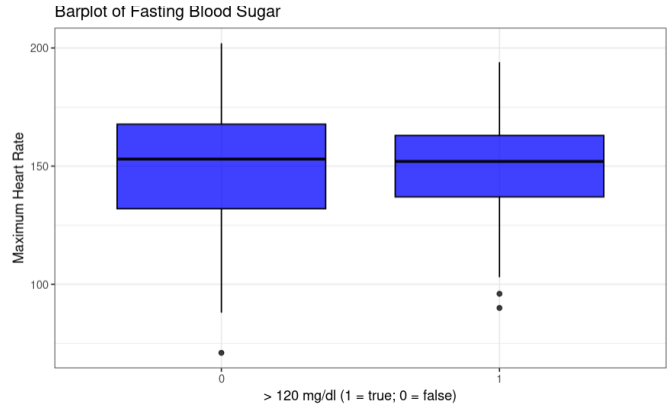
Box Plots

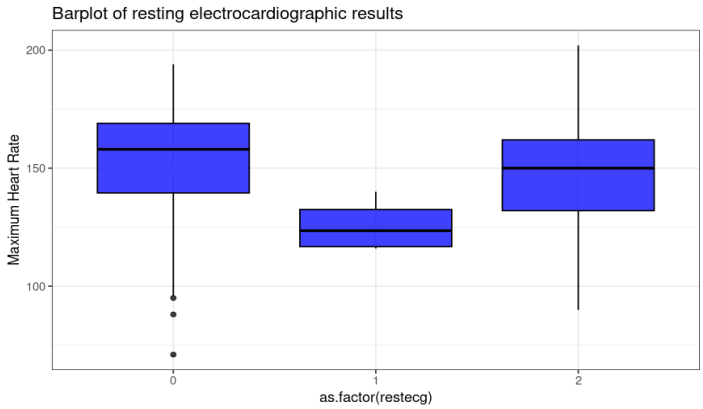

Boxplot of HeartRate by Gender
Maximum Heart Rate
200
150
100
0
1
1 = male; 0 = female

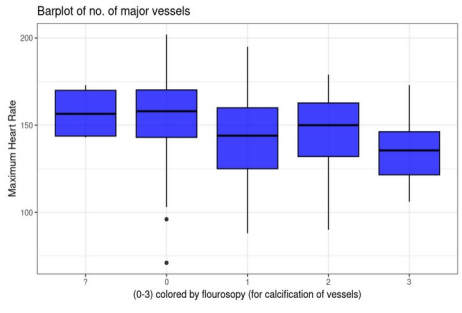

Boxplot of HeartRate by ChestPain
Maximum Heart Rate
200
150
100
1
2
3
4
chest pain type

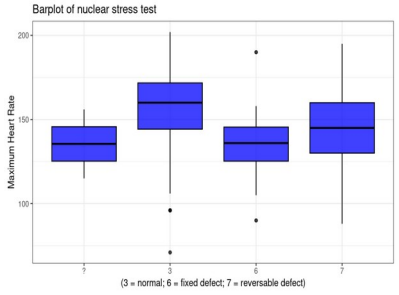

Barplot of exercise induced angina
Maximum Heart Rate
200
150
100
0
1
exercise induced angina (1 = yes; 0 = no)

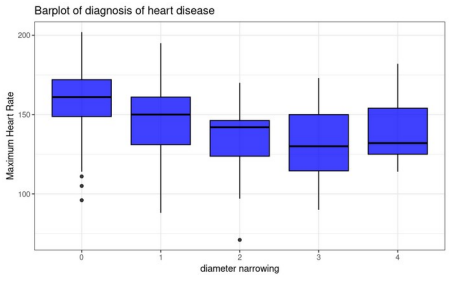

Barplot of peak exercise
Maximum Heart Rate
200
150
100
1
2
3
Value 1:upsloping 2:flat 3:downsloping


Barplot of Fasting Blood Sugar
Maximum Heart Rate
200
150
100
0
1
> 120 mg/dl (1 = true; 0 = false)

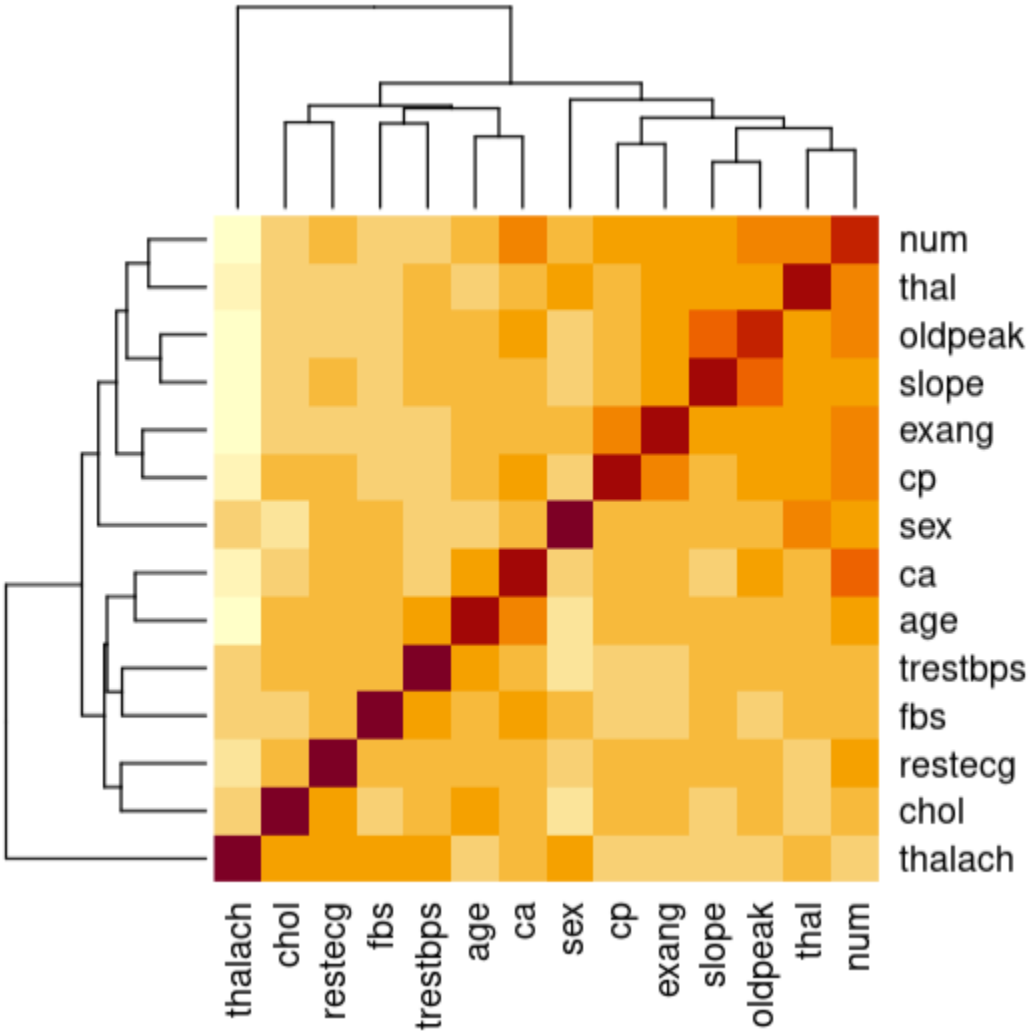

Barplot of resting electrocardiographic results
Maximum Heart Rate
200
150
100
0
1
2
as.factor(restecg)


Barplot of no. of major vessels
Maximum Heart Rate
200
150
100
0
1
2
3
(0-3) colored by flourosopy (for calcification of vessels)

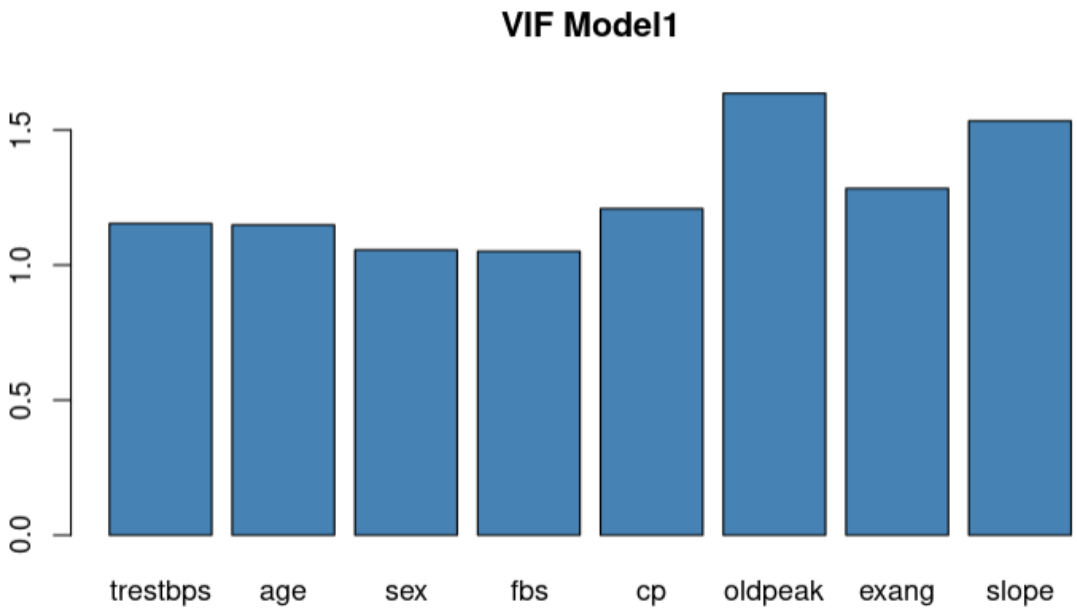

Barplot of nuclear stress test
Maximum Heart Rate
200
150
100
3
6
7
(3 = normal; 6 = fixed defect; 7 = reversable defect)

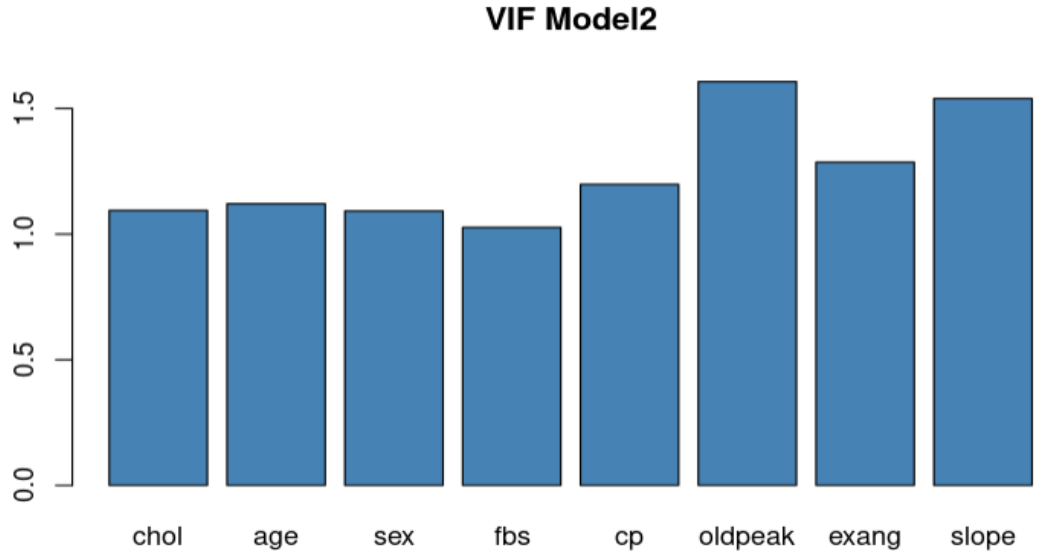

Barplot of diagnosis of heart disease
Maximum Heart Rate
200
150
100
0
1
2
3
4
diameter narrowing

Heat Maps


num
thal
oldpeak
slope
exang
cp
sex
ca
age
trestbps
fbs
restecg
chol
thalach
thalach
chol
restecg
fbs
trestbps
age
ca
sex
cp
exang
slope
oldpeak
thal
num

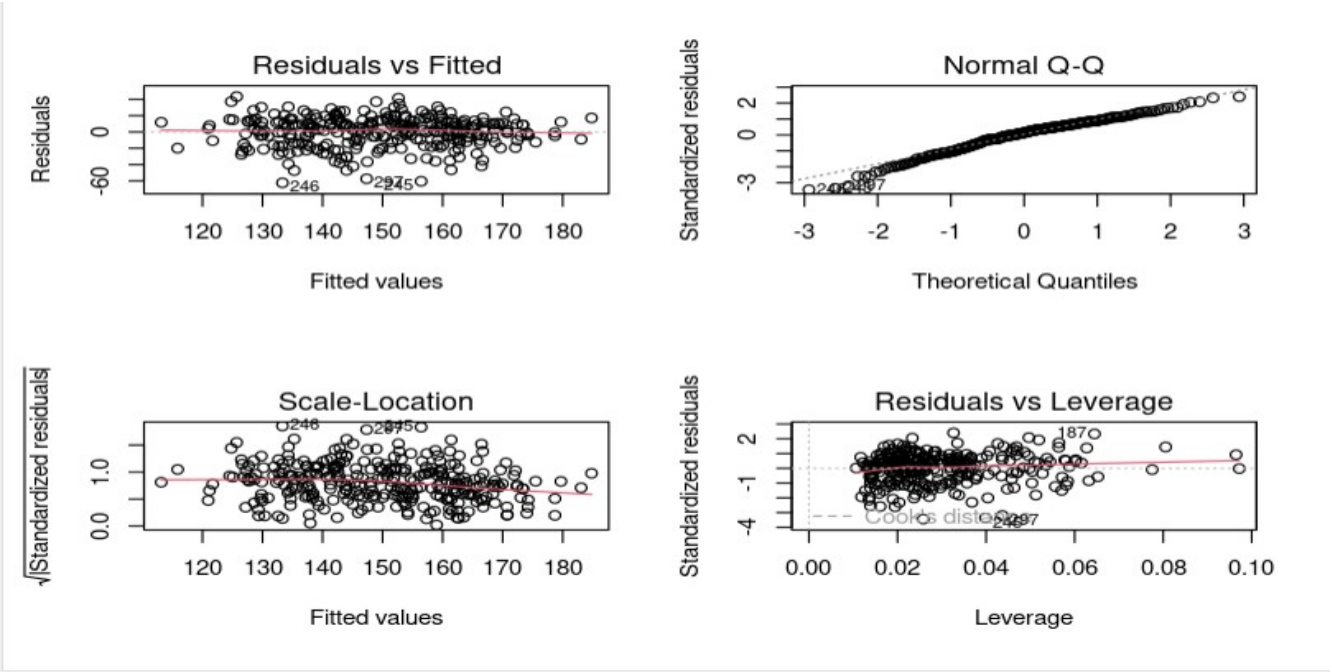
Variance Inflation Factor



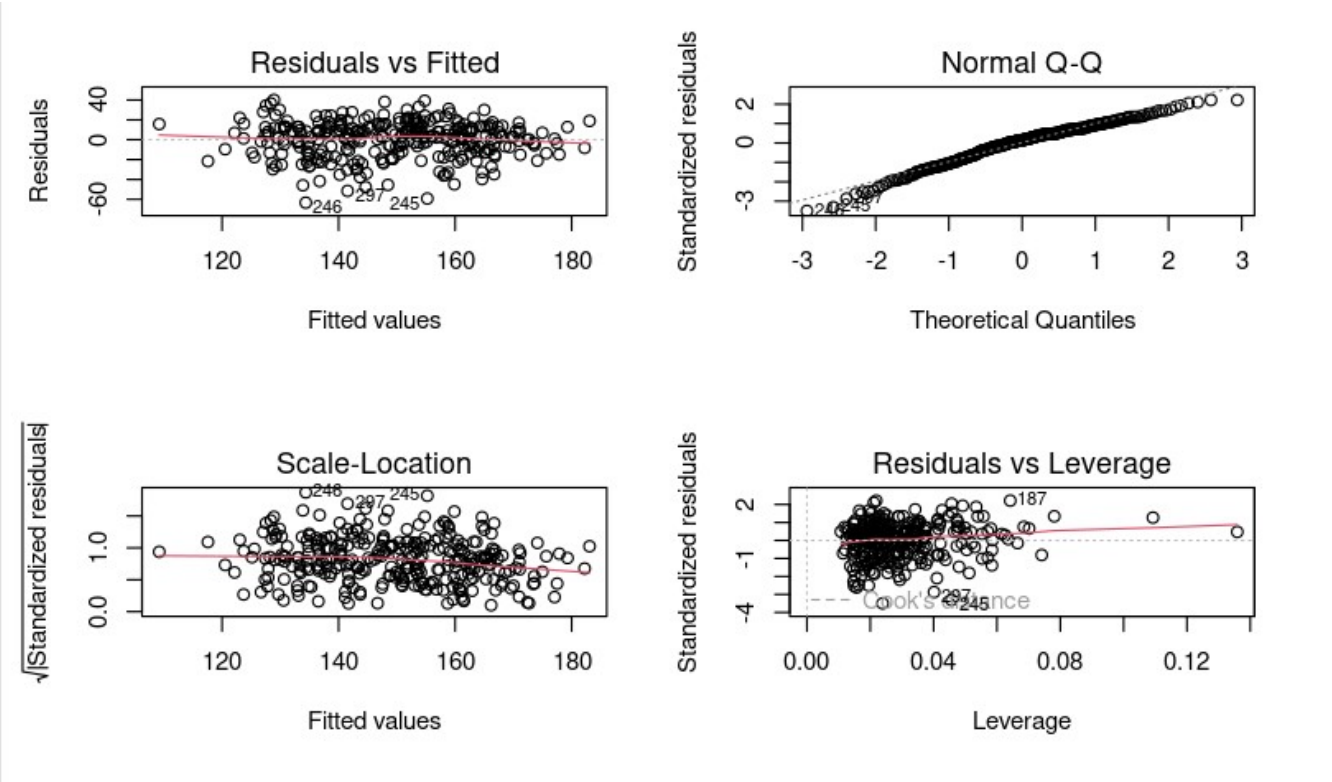


Diagnostic Plots

Model 1:



Model 2:



Appendix B (Tables)

Description of covariates

No.	Variable	Description
1	age	Age of Patents in years
2	sex	Gender of Patients
3	cp	Chest Pain Type (1: typical angina , 2: atypical angina, 3: non-anginal pain 4: asymptomatic)
4	trestbps	resting blood pressure (in mm Hg on admission to the hospital)
5	chol	serum cholestoral in mg/dl
6	fbs	fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7	restecg	resting electrocardiographic results (0: normal, 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), 2: showing probable or definite left ventricular hypertrophy by Estes' criteria)
8	thalach	maximum heart rate achieved
9	exang	exercise induced angina (1 = yes; 0 = no)
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	slope of the peak exercise ST segment (1: upsloping, 2: flat, 3: downsloping)
12	ca	no. of major vessels (0-3) colored by flourosopy (for calcification of vessels)
13	thal	results of nuclear stress test (3: normal; 6: fixed defect; 7: reversable defect)
14	num	target variable representing diagnosis of heart disease (angiographic disease status) in any major vessel (0: < 50% diameter narrowing, 1: > 50% diameter narrowing)

Descriptive Statistics (Continuous Variables)

Summary of Descriptive Statistics

	Total (N=303)
Maximum Heart Rate achieved	
Mean (SD)	150 (22.9)
Median [Min, Max]	153 [71.0, 202]
Age of Patients	
Mean (SD)	54.4 (9.04)
Median [Min, Max]	56.0 [29.0, 77.0]
Resting Blood Pressure in mm Hg	
Mean (SD)	132 (17.6)
Median [Min, Max]	130 [94.0, 200]
serum Cholesterol in mg/dl	
Mean (SD)	247 (51.8)
Median [Min, Max]	241 [126, 564]
depression induced by exercise	
Mean (SD)	1.04 (1.16)
Median [Min, Max]	0.800 [0, 6.20]

Descriptive Statistics (Categorical Variables)

Summary of Descriptive Statistics

	Total (N=303)
Gender	
Female	97 (32.0%)
Male	206 (68.0%)
Chest Pain	
typical angina	23 (7.6%)
atypical angina	50 (16.5%)
non-anginal pain	86 (28.4%)
asymptomatic	144 (47.5%)
Fasting Blood Sugar	
< 120 mg/dl	258 (85.1%)
> 120 mg/dl	45 (14.9%)
resting electrocardiographic results	
normal	151 (49.8%)
having ST-T wave abnormality	4 (1.3%)
probable or definite left ventricular hypertrophy	148 (48.8%)
exercise induced angina	
No	204 (67.3%)
Yes	99 (32.7%)
slope of the peak exercise	
upsloping	142 (46.9%)
flat	140 (46.2%)
downsloping	21 (6.9%)
number of major vessels	
Major vessels:0	176 (58.1%)
Major vessels:1	65 (21.5%)
Major vessels:2	38 (12.5%)
Major vessels:3	20 (6.6%)
Missing	4 (1.3%)
results of nuclear stress test	
normal	166 (54.8%)
fixed defect	18 (5.9%)
reversable defect	117 (38.6%)
Missing	2 (0.7%)

Linear Regression

Model 1:

Summary of the 1st model

	Est	ci95.lo	ci95.hi	t value	Pr(>|t|)
(Intercept)	212.3488	191.3737	233.3239	19.9244	0.0000
trestbps	0.1173	-0.0094	0.2440	1.8220	0.0695
age	-0.8608	-1.1069	-0.6146	-6.8814	0.0000
sex	-1.4862	-6.0524	3.0801	-0.6405	0.5223
fbs	1.9696	-4.0058	7.9451	0.6487	0.5170
cp	-3.9774	-6.3551	-1.5998	-3.2923	0.0011
oldpeak	-1.2509	-3.5380	1.0361	-1.0765	0.2826
exang	-10.1487	-15.1557	-5.1417	-3.9891	0.0001
slope	-8.3929	-12.5656	-4.2203	-3.9586	0.0001

Model 2:

Summary of the 2nd model

	Est	ci95.lo	ci95.hi	t value	Pr(>|t|)
(Intercept)	217.2644	199.4999	235.0289	24.0700	0.0000
chol	0.0396	-0.0023	0.0816	1.8603	0.0638
age	-0.8483	-1.0914	-0.6052	-6.8673	0.0000
sex	-0.9439	-5.5853	3.6975	-0.4002	0.6893
fbs	2.8061	-3.0980	8.7103	0.9354	0.3504
cp	-4.2655	-6.6317	-1.8993	-3.5478	0.0005
oldpeak	-1.0397	-3.3057	1.2262	-0.9030	0.3672
exang	-10.2372	-15.2477	-5.2267	-4.0211	0.0001
slope	-8.2043	-12.3840	-4.0247	-3.8632	0.0001

Reference

Malik, M., & Camm, A. J. (1990). Heart rate variability. *Clinical cardiology*, *13*(8), 570-576.

Tsuji, H., Venditti Jr, F. J., Manders, E. S., Evans, J. C., Larson, M. G., Feldman, C. L., & Levy, D. (1996). Determinants of heart rate variability. *Journal of the American College of Cardiology*, *28*(6), 1539-1546.

Kikuya, M., Hozawa, A., Ohokubo, T., Tsuji, I., Michimata, M., Matsubara, M., ... & Imai, Y. (2000). Prognostic significance of blood pressure and heart rate variabilities: the Ohasama study. *Hypertension*, *36*(5), 901-906.

Stauss, H. M. (2003). Heart rate variability. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *285*(5), R927-R931.

D'Antono, B., Dupuis, G., Fortin, C., Arsenault, A., & Burelle, D. (2006). Angina symptoms in men and women with stable coronary artery disease and evidence of exercise-induced myocardial perfusion defects. *American heart journal*, *151*(4), 813-819.

Kannel, W. B. (1995). Range of serum cholesterol values in the population developing coronary artery disease. *The American journal of cardiology*, *76*(9), 69C-77C.

Tan, Y. Y., Gast, G. C. M., & van der Schouw, Y. T. (2010). Gender differences in risk factors for coronary heart disease. *Maturitas*, *65*(2), 149-160.

Detrano, R. (2020, March 24). *Cleveland Clinic Heart Disease Dataset*. Kaggle. https://www.kaggle.com/datasets/aavigan/cleveland-clinic-heart-disease-dataset/data

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Wickham H, François R, Henry L, Müller K, Vaughan D (2023). _dplyr: A Grammar of Data Manipulation_. R package version 1.1.4, <https://CRAN.R-project.org/package=dplyr>.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Fox J, Weisberg S (2019). _An R Companion to Applied Regression_, Third edition. Sage, Thousand Oaks CA. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>.

Gordon M, Gragg S, Konings P (2023). _htmlTable: Advanced Tables for Markdown/HTML_. R package version 2.4.2, <https://CRAN.R-project.org/package=htmlTable>.

Rich B (2023). _table1: Tables of Descriptive Statistics in HTML_. R package version 1.4.3, <https://CRAN.R-project.org/package=table1>.

Xie Y (2024). _knitr: A General-Purpose Package for Dynamic Report Generation in R_. R package version 1.46, <https://yihui.org/knitr/>.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595