Project Proposal

R Markdown

```
# Installing all required packages
install.packages("car")
```

```
## Installing package into '/opt/r'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(boot)
library(knitr)
library(broom) ## Needed to make the regression output 'tidy'
library(ggplot2)
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:boot':
##
##     logit
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
```

```
##
##     some
```

```
# Load in the heart attack data
 heart_attack <- read.csv("heart.csv")
glimpse(heart_attack)
```

```
## Rows: 303
## Columns: 14
## $ age      <int> 63, 37, 41, 56, 57, 57, 56, 44, 52, 57, 54, 48, 49, 64, 58, 5~
## $ sex      <int> 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1~
## $ cp       <int> 3, 2, 1, 1, 0, 0, 1, 1, 2, 2, 0, 2, 1, 3, 3, 2, 2, 3, 0, 3, 0~
## $ trtbps   <int> 145, 130, 130, 120, 120, 140, 140, 120, 172, 150, 140, 130, 1~
## $ chol     <int> 233, 250, 204, 236, 354, 192, 294, 263, 199, 168, 239, 275, 2~
## $ fbs      <int> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0~
## $ restecg  <int> 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1~
## $ thalachh <int> 150, 187, 172, 178, 163, 148, 153, 173, 162, 174, 160, 139, 1~
## $ exng     <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0~
## $ oldpeak  <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, 1.6, 1.2, 0.2, 0~
## $ slp      <int> 0, 0, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 0, 2, 2, 1~
## $ caa      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2~
## $ thall    <int> 1, 2, 2, 2, 2, 1, 2, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3~
## $ output   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

```
#Finding any NA values in the data set
colSums(is.na(heart_attack), na.rm = TRUE)
```

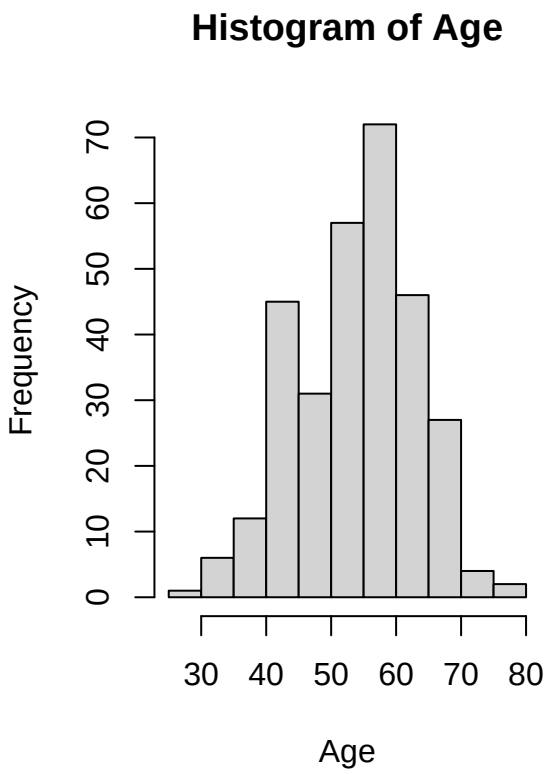
```
##      age      sex       cp   trtbps     chol      fbs  restecg thalachh
##        0        0        0        0        0        0        0        0
##     exng  oldpeak      slp      caa    thall   output
##        0        0        0        0        0        0
```

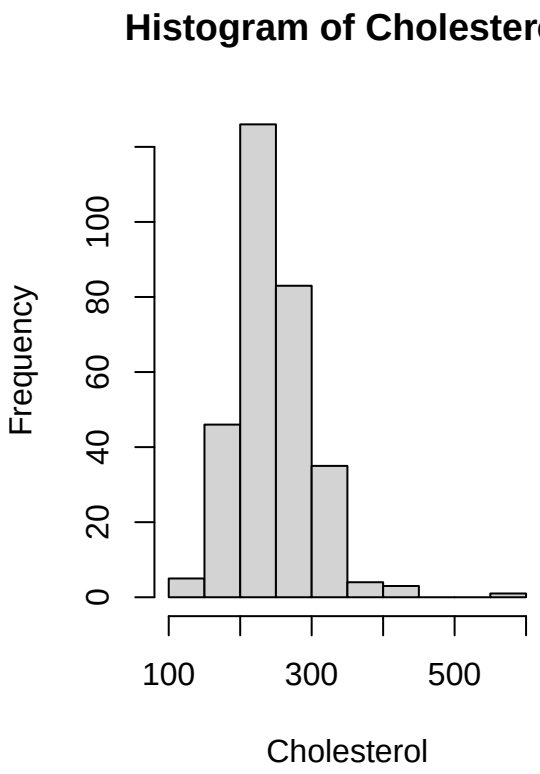
```
df <- as.data.frame(heart_attack)
heart_attack_data <- tibble::rowid_to_column(df, "index")
```

```
# Calculating numerical summaries of necessary variables in the dataset.
heart_attack_1 <- heart_attack_data %>% select(age, chol, trtbps, thalachh)
summary(heart_attack_1) %>% kable()
```

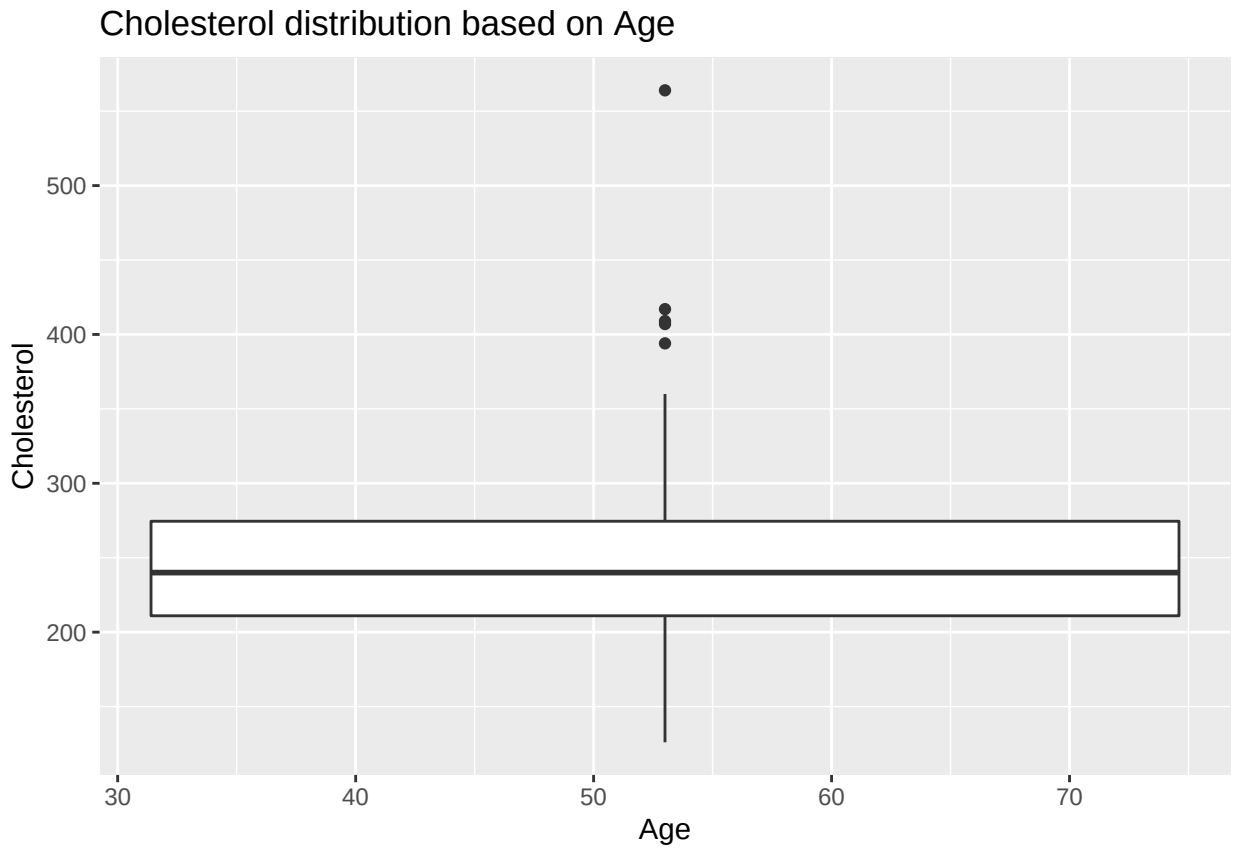
age	chol	trtbps	thalachh
Min. :29.00	Min. :126.0	Min. : 94.0	Min. : 71.0
1st Qu.:47.50	1st Qu.:211.0	1st Qu.:120.0	1st Qu.:133.5
Median :55.00	Median :240.0	Median :130.0	Median :153.0
Mean :54.37	Mean :246.3	Mean :131.6	Mean :149.6
3rd Qu.:61.00	3rd Qu.:274.5	3rd Qu.:140.0	3rd Qu.:166.0
Max. :77.00	Max. :564.0	Max. :200.0	Max. :202.0

```
# Histogram of Age and Cholesterol Levels.
par(mfrow=c(1,2))
hgA <- hist(x=heart_attack_data$age, xlab= "Age", main= "Histogram of Age")
hgC<-hist(x=heart_attack$chol, xlab= "Cholesterol",
main= "Histogram of Cholesterol" )
```






```
ggplot(heart_attack_data)+ geom_boxplot(aes(x=age, y=chol,group=1)) + labs(x= "Age", y= "Cholesterol",
title= "Cholesterol distribution based on Age")
```




```
#Simple linear regression model
lmHeart <- lm(age~chol, data= heart_attack_data)
#Summary of our linear regression model
lmHeart %>%
  tidy() %>%
  kable(caption = "The summary from the simple linear regression model", digits = 4)
```

Table 2: The summary from the simple linear regression model

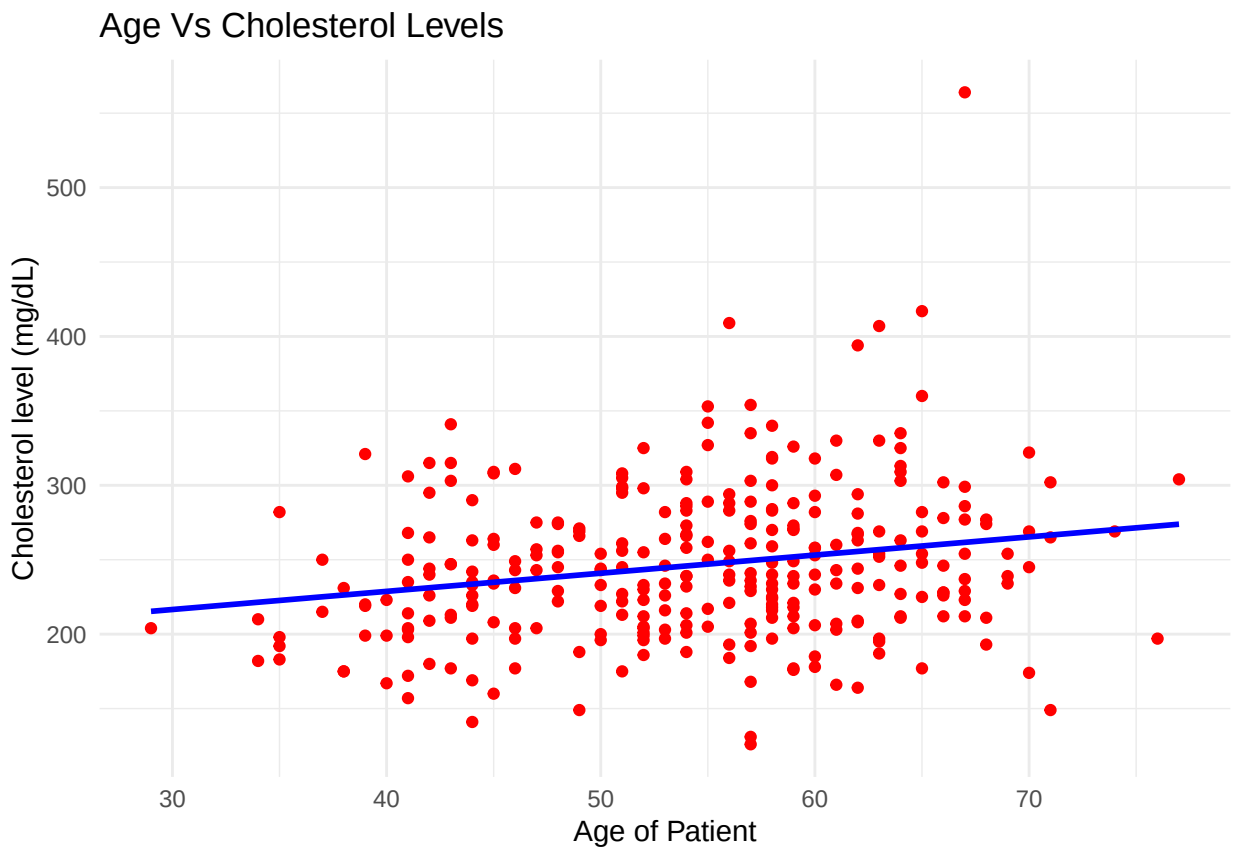
term	estimate	std.error	statistic	p.value
(Intercept)	45.1457	2.4828	18.1830	0e+00
chol	0.0374	0.0099	3.7948	2e-04

```
#Checking correlation between cholesterol levels and age of patient
cor(heart_attack_data$age, heart_attack_data$chol)
```

```
## [1] 0.213678
```

```
#Simple Linear Regression Analysis for Age and Cholesterol Levels.
p <- ggplot(heart_attack_data, aes(x=age, y=chol)) + geom_point(col = 'red')
            p + scale_color_brewer(palette="Dark2") + theme_minimal() +
  geom_smooth(method=lm, se=FALSE, col = "blue") + ggtitle("Age Vs Cholesterol Levels") +
              xlab("Age of Patient") + ylab("Cholesterol level (mg/dL)")
```

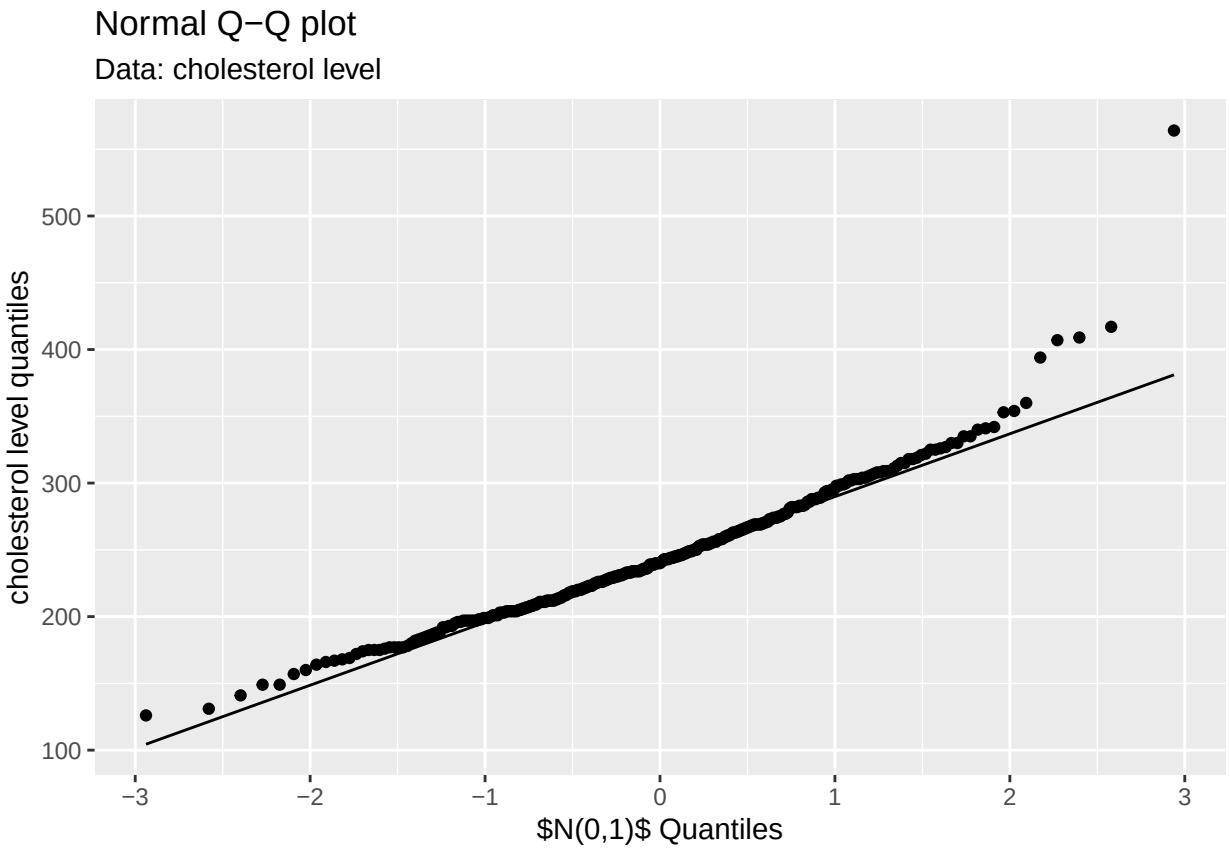
```
## `geom_smooth()` using formula 'y ~ x'
```



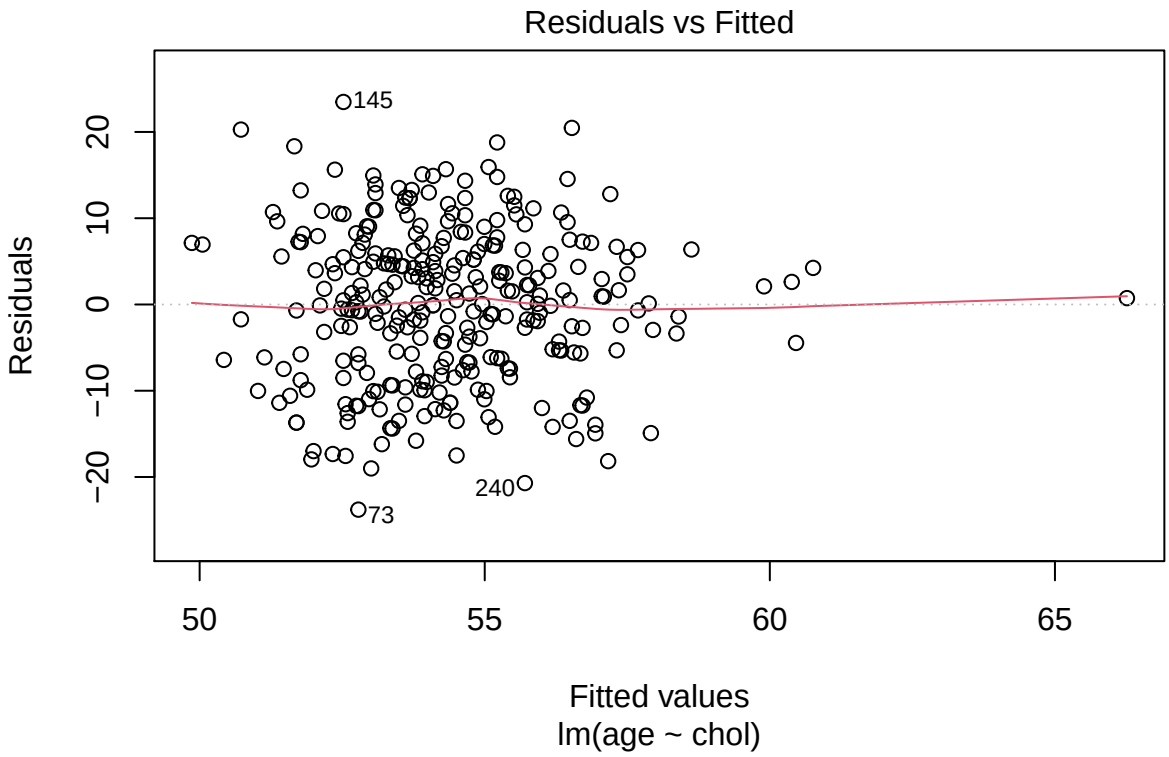

```r
#Performing t test on our data set
t.test(heart_attack_data$chol, heart_attack_data$age)
```

```
##
##  Welch Two Sample t-test
##
## data:  heart_attack_data$chol and heart_attack_data$age
## t = 63.48, df = 320.53, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  185.9503 197.8451
## sample estimates:
## mean of x mean of y
## 246.26403  54.36634
```

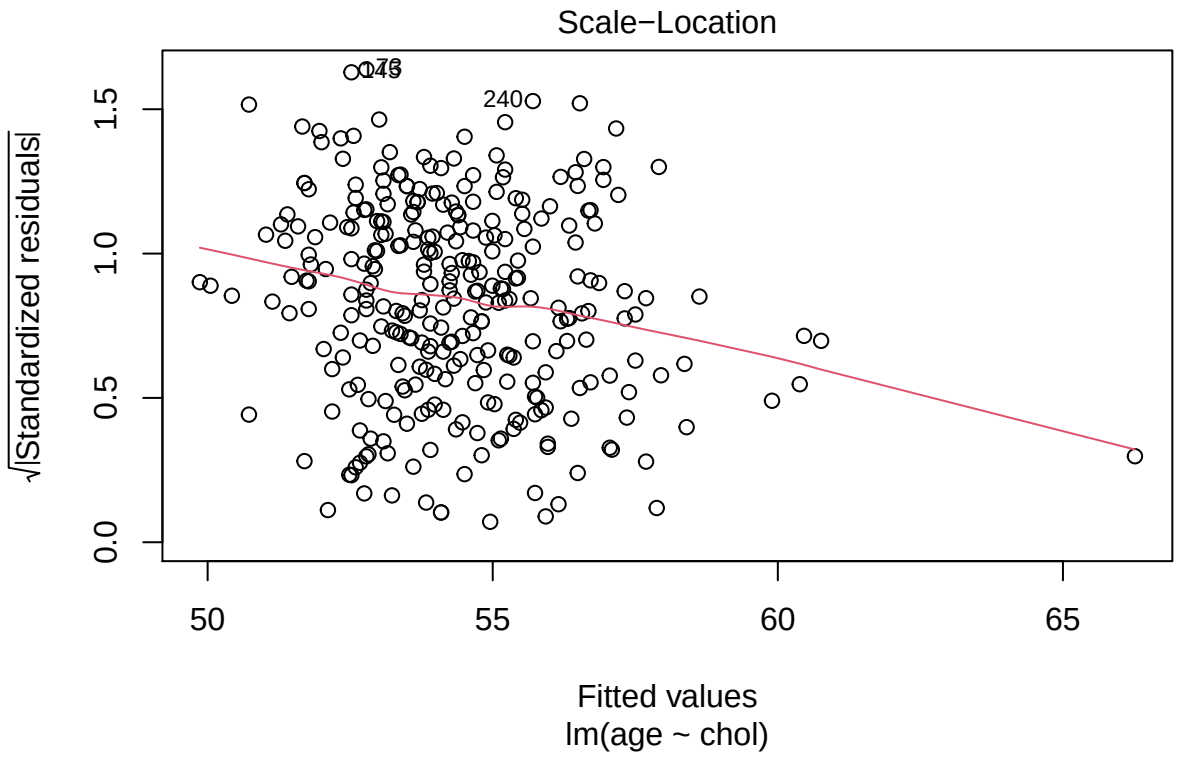
```r
#Verifying our model satisfies the normality assumption
x_bar <- mean(heart_attack_data$chol)
heart_attack_data %>%
ggplot(aes(sample = chol))+ geom_qq()+
geom_qq_line()+
labs(x = "$N(0,1)$ Quantiles", y = "cholesterol level quantiles", title = "Normal Q-Q plot",
subtitle = "Data: cholesterol level")
```



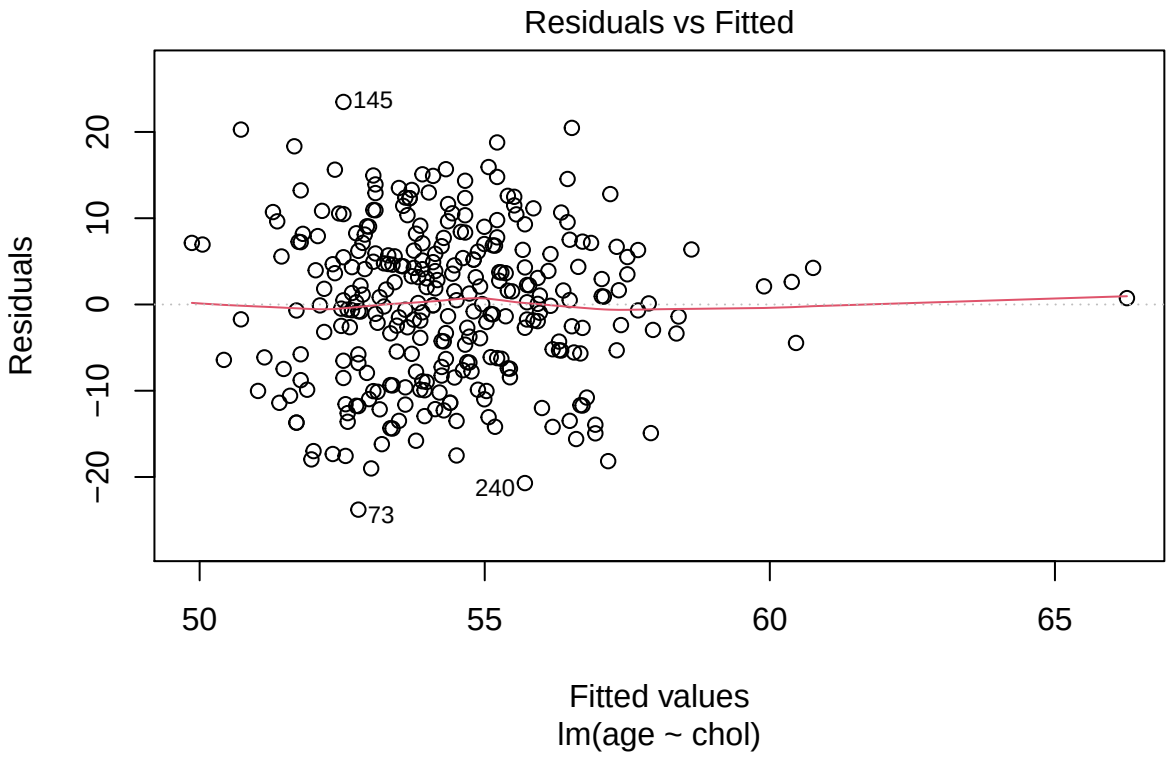

```
#Verifying our data satisfies the Homeoscedasticity Assumption
plot(lmHeart, 1)
```

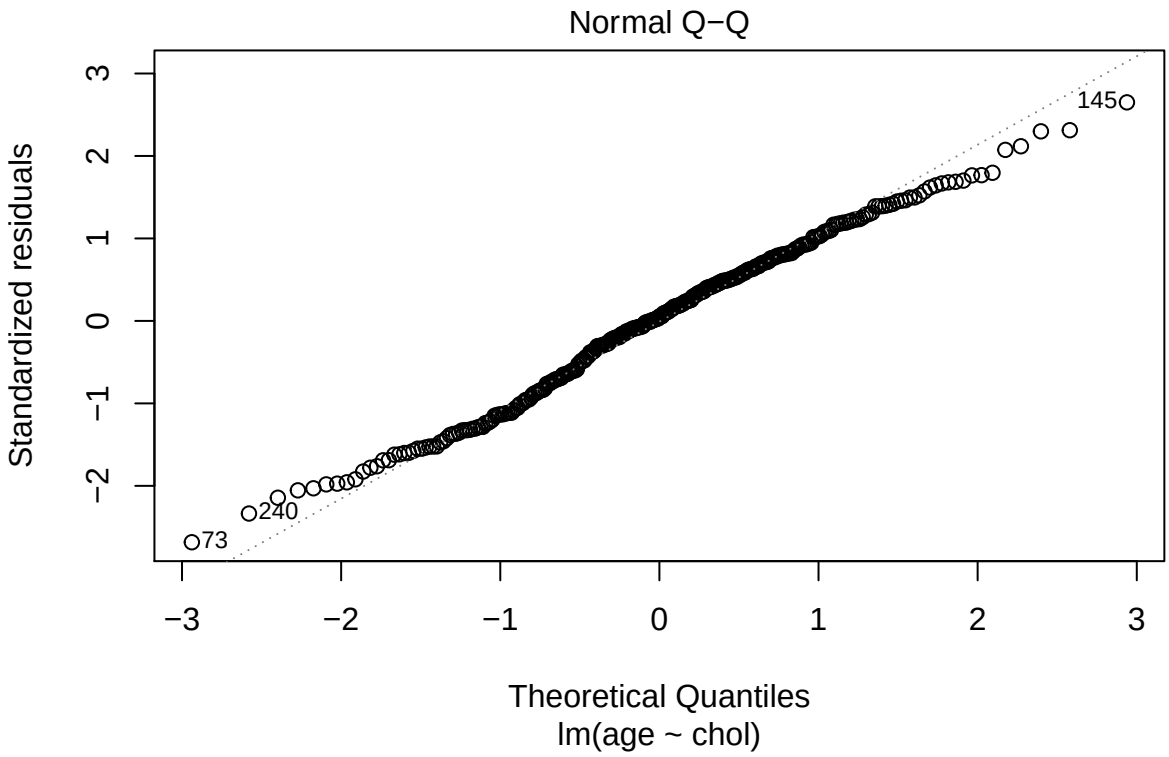


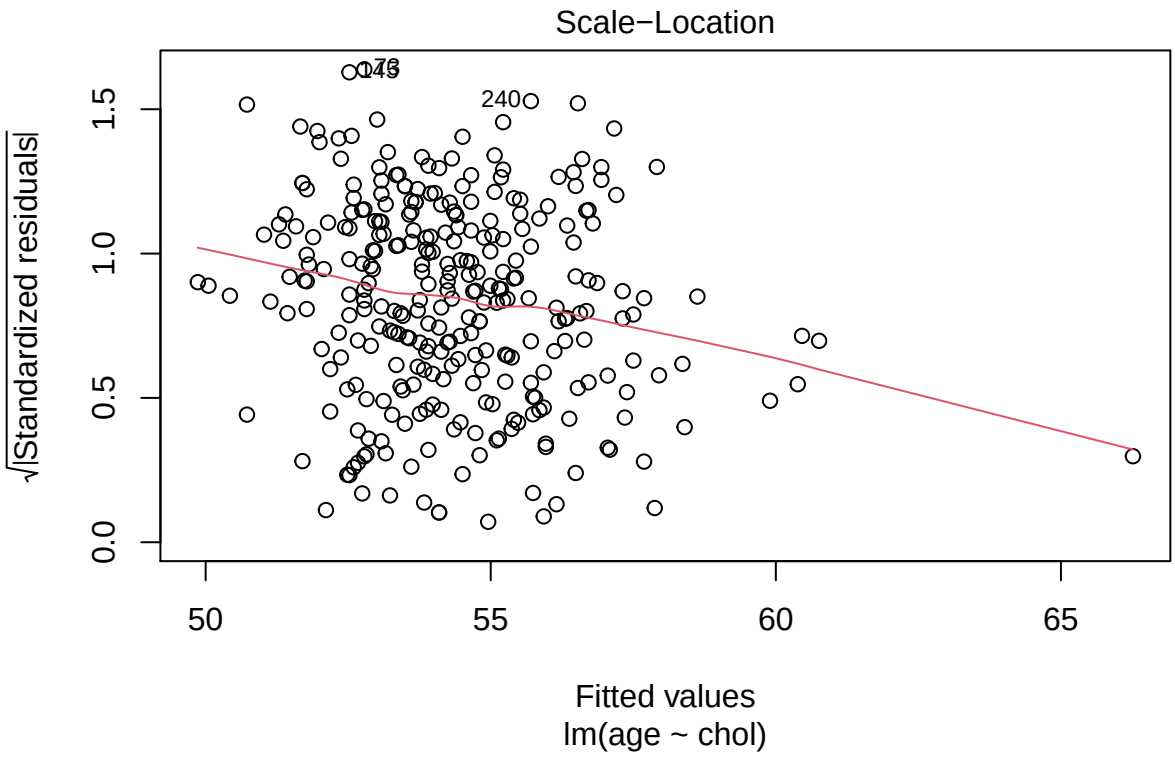

```
plot(lmHeart, 3)
```

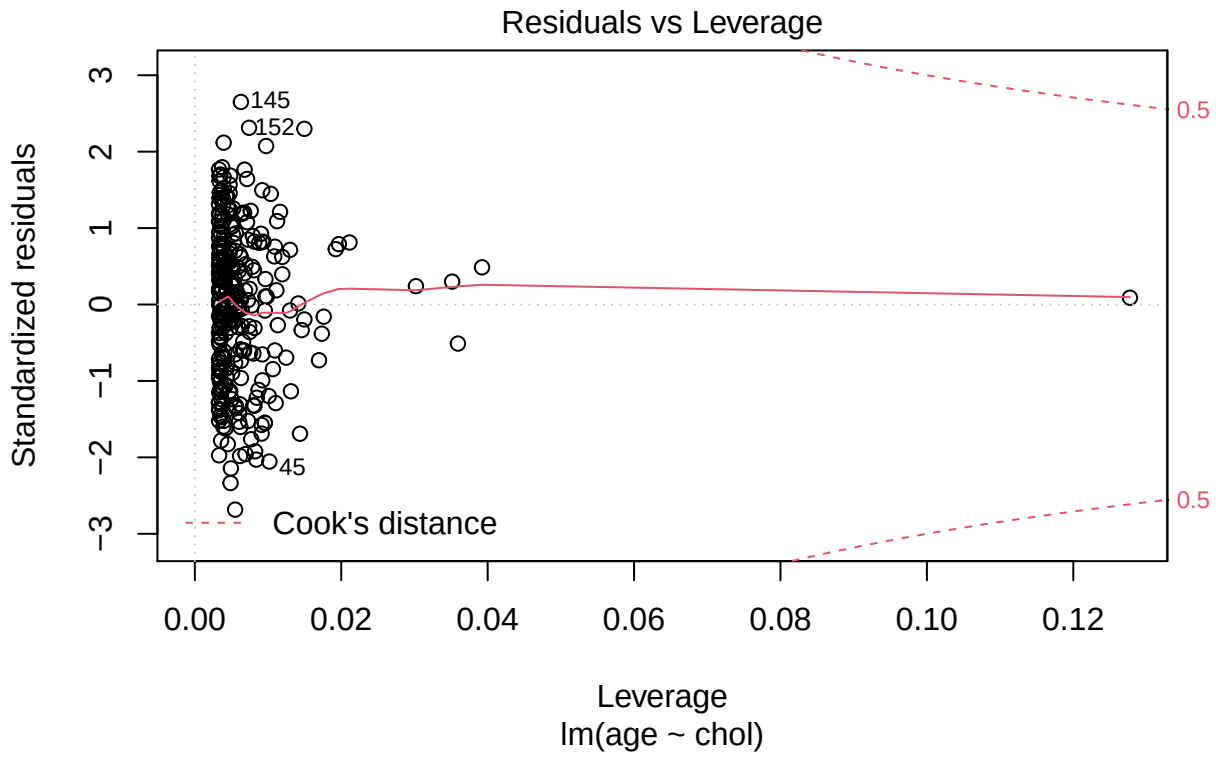



```
plot(lmHeart)
```

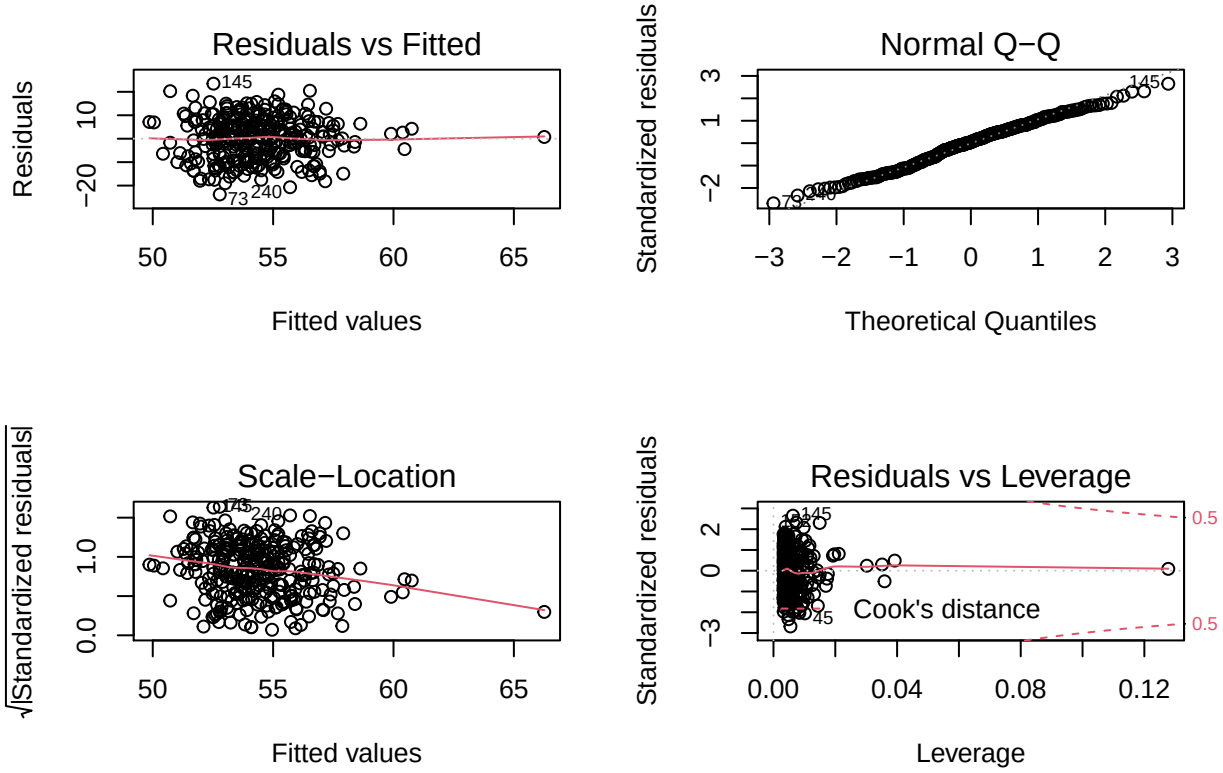

Residuals vs Fitted
Residuals
Fitted values
lm(age ~ chol)
145
240
73
20
10
0
−10
−20
50
55
60
65


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(age ~ chol)
73
240
145
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
50
55
60
65
Fitted values
lm(age ~ chol)
240




```
par(mfrow=c(2,2))
plot(lmHeart)
```




```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
lmtest::bptest(lmHeart)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lmHeart
## BP = 6.3962, df = 1, p-value = 0.01144
```

```
dwtest(heart_attack_data$age ~ heart_attack_data$chol)
```

```
##
##  Durbin-Watson test
##
## data:  heart_attack_data$age ~ heart_attack_data$chol
## DW = 1.9342, p-value = 0.2831
## alternative hypothesis: true autocorrelation is greater than 0
```