Homework 1

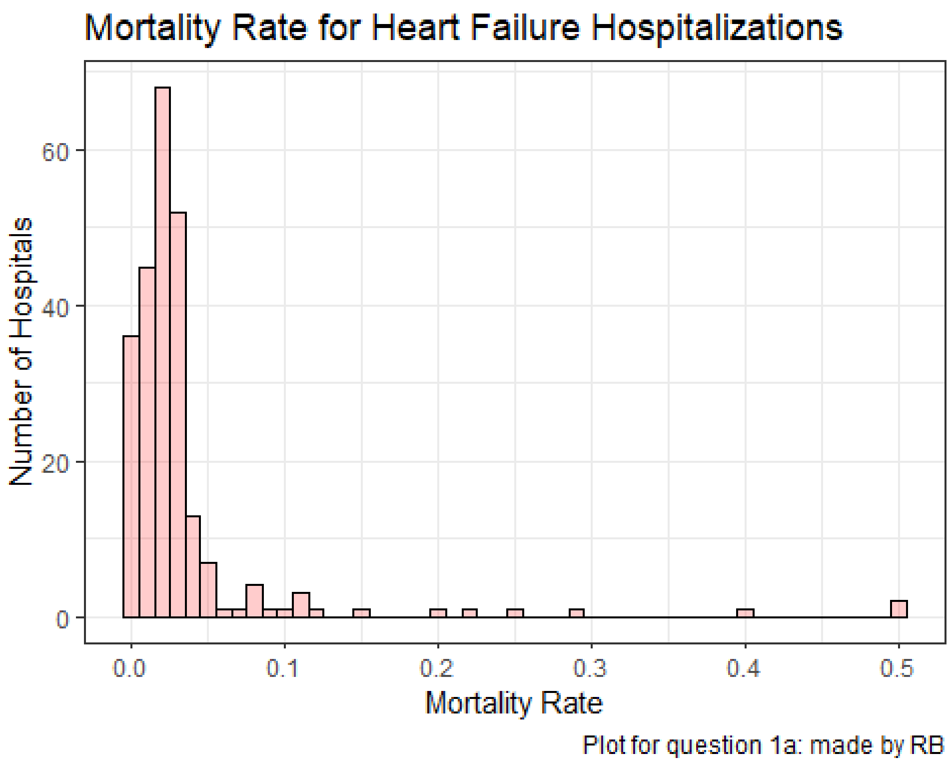
```
knitr::opts_chunk$set(echo = TRUE)

FL_2014_CORE <- readRDS("FL_2014_CORE.rds")
library(data.table)

## Warning: package 'data.table' was built under R version 3.6.2

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.6.2

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.2

names(FL_2014_CORE) <- tolower(names(FL_2014_CORE))

FL_2014 <- select(FL_2014_CORE,
                  age,atype,aweekend,died,drg,female,dqtr,dshospid,
                  los,nchronic,ndx,npr,pl_cbsa,race,visitlink,mdc,
                  zip,totchg,orproc,pay1)

FL_2014 <- FL_2014 %>%
  mutate(los = as.integer(los), totchg = as.integer(totchg), atype = as.facto
r(atype),
         dqtr = as.factor(dqtr), aweekend = as.factor(aweekend), pay1 = as.fa
ctor(pay1))

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion
```

```
q1 <- FL_2014 %>%
select(died, drg, dshospid)%>%
filter(., drg %in% 291:293) %>%     #Filter for heart failure patients
group_by(., dshospid) %>%
summarize(., Count = n(),TotalDeath = sum(died), MortalityRate = mean(died))%
>%
arrange(desc(MortalityRate))

# Printing top 10 Hospital IDs in terms of Highest Mortality Rate
q1 %>%
top_n(10,MortalityRate)
```

```
## # A tibble: 12 x 4
##    dshospid Count TotalDeath MortalityRate
##       <int> <int>      <int>         <dbl>
##   1   120001     2          1         0.5
##   2 23960082     2          1         0.5
##   3 23960074     5          2         0.4
##   4   100152     7          2         0.286
##   5   100134     4          1         0.25
##   6   100143     9          2         0.222
##   7 23960043     5          1         0.2
##   8 23960028    40          6         0.15
##   9   100120     8          1         0.125
## 10   100042     9          1         0.111
## 11   100138     9          1         0.111
## 12 23960011     9          1         0.111
```

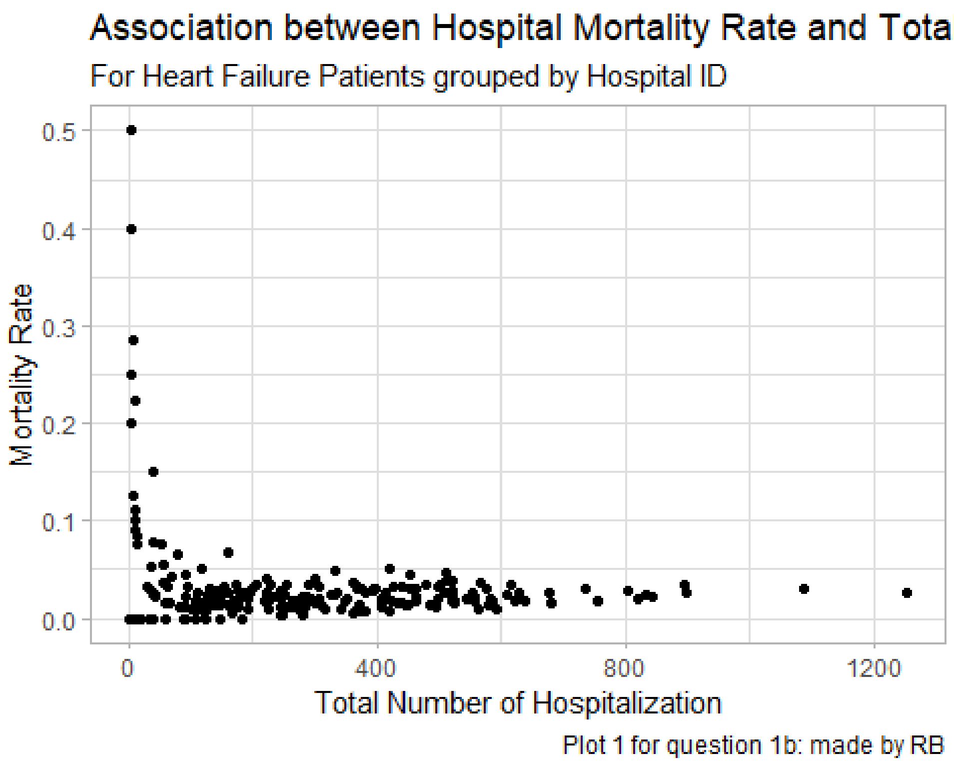
There are 12 hospitals with top 10 mortality rate since three hospitals are tied with a mortality rate of 0.111 at the bottom.

```
Graph0 <- ggplot(q1,aes(x=MortalityRate)) +
  geom_histogram(binwidth = 0.01,col="black",
                 fill="red",
                 alpha = .2)

Graph0 + theme_bw()+
  labs(title="Mortality Rate for Heart Failure Hospitalizations") +
  labs(x="Mortality Rate", y="Number of Hospitals",
       caption = " Plot for question 1a: made by RB")
```



The histogram is heavily skewed towards the right because with the increase in mortality rate, the number of hospitals is decreasing. There are a large number of hospitals who have mortality rate between 0.0 and 0.1 for heart failure hospitalizations.

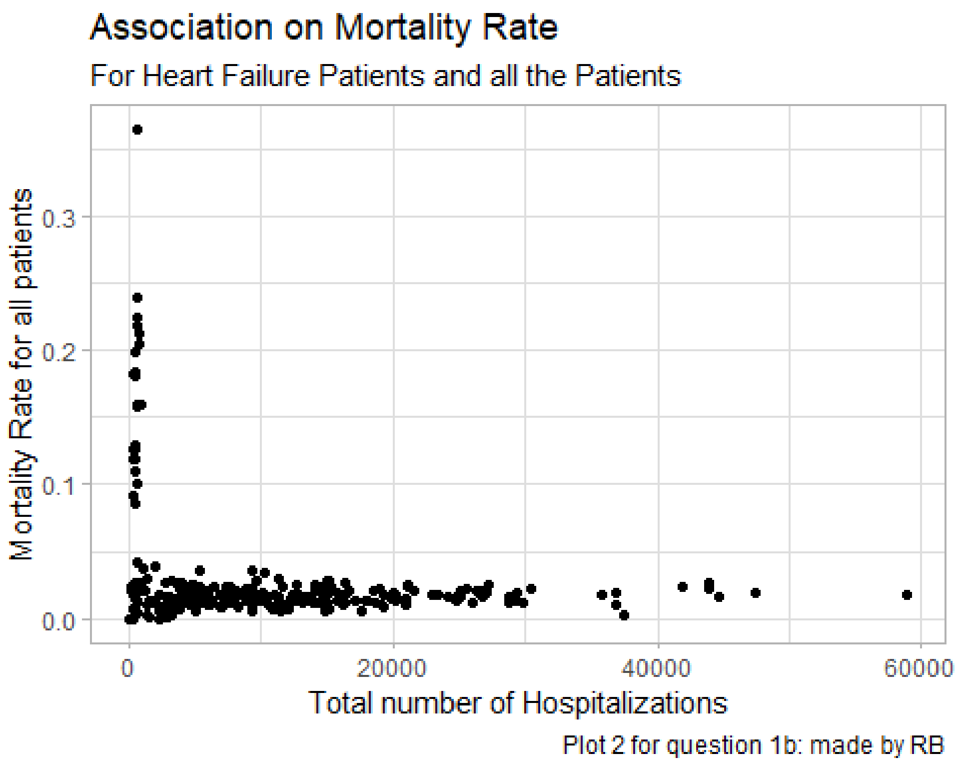
```
mygraph1 <- ggplot(q1, aes(x = q1$Count, y = q1$MortalityRate)) +
  geom_point()

mygraph1 +
  theme_light() +
  labs(
    x = "Total Number of Hospitalization",
    y = "Mortality Rate",
    color = "Green",
    title = "Association between Hospital Mortality Rate and Total Number Hos
pitalizations",
    subtitle = "For Heart Failure Patients grouped by Hospital ID",
    caption = "Plot 1 for question 1b: made by RB")
```



The mortality rate is high initially is it depends on the number of admissions. For example, if we have 2 admits in a hospital and one of them dies, the mortality rate will be considerably high. This is primary the reason why we see high mortality rates with low number of hospital admits. As we move forward in the number of hospitalization, the mortality rate considerably drops down below 0.1

```
q2 <- FL_2014 %>%
  select(died, drg, dshospid)%>%
  group_by(., dshospid) %>%
  summarize(., Count2 = n(),TotalDeath2 = sum(died), MortalityRate2 = mean(di
ed))%>%
  arrange(desc(MortalityRate2))

Merged <- merge(q1,q2)

mygraph2 <- ggplot(Merged, aes(x = Merged$Count2, y = Merged$MortalityRate2))
+
  geom_point()

mygraph2 +
  theme_light() +
  labs(
    x = "Total number of Hospitalizations",
    y = "Mortality Rate for all patients",
    color = "Green",
    title = "Association on Mortality Rate",
```

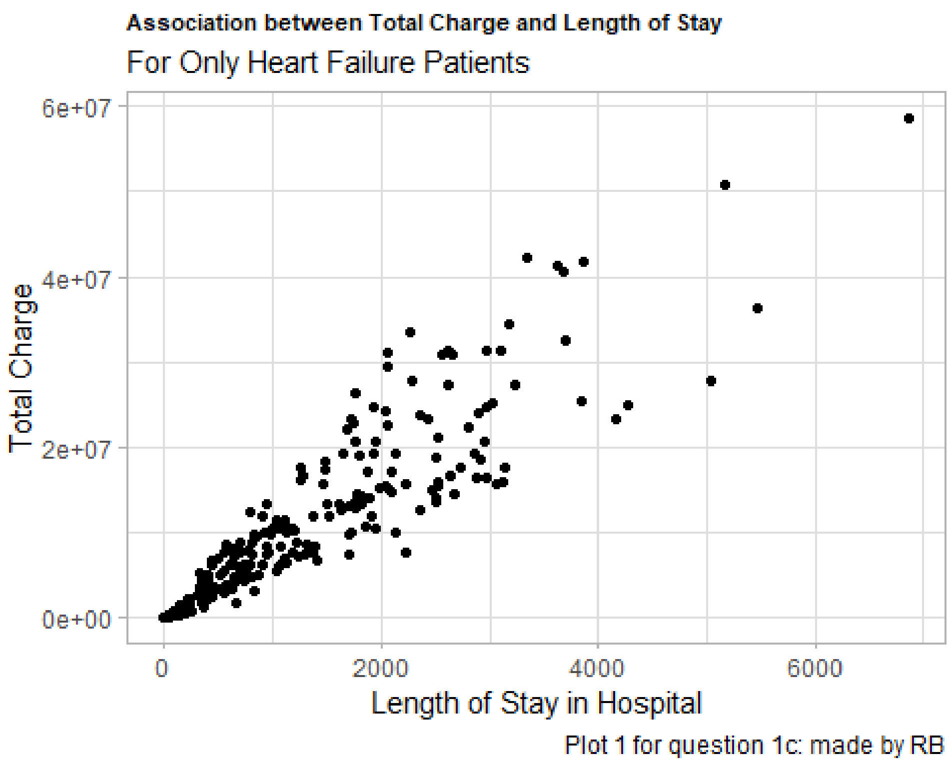
```
    subtitle = "For Heart Failure Patients and all the Patients",
    caption = "Plot 2 for question 1b: made by RB")
```



The plot doesn't provide sufficient evidence for association between mortality rate for all the patients and 'only' heart failure patients. Majority of the data points are condensed near the origin depitcing in addition to the correlation being 0.44 which is a weak correlation.

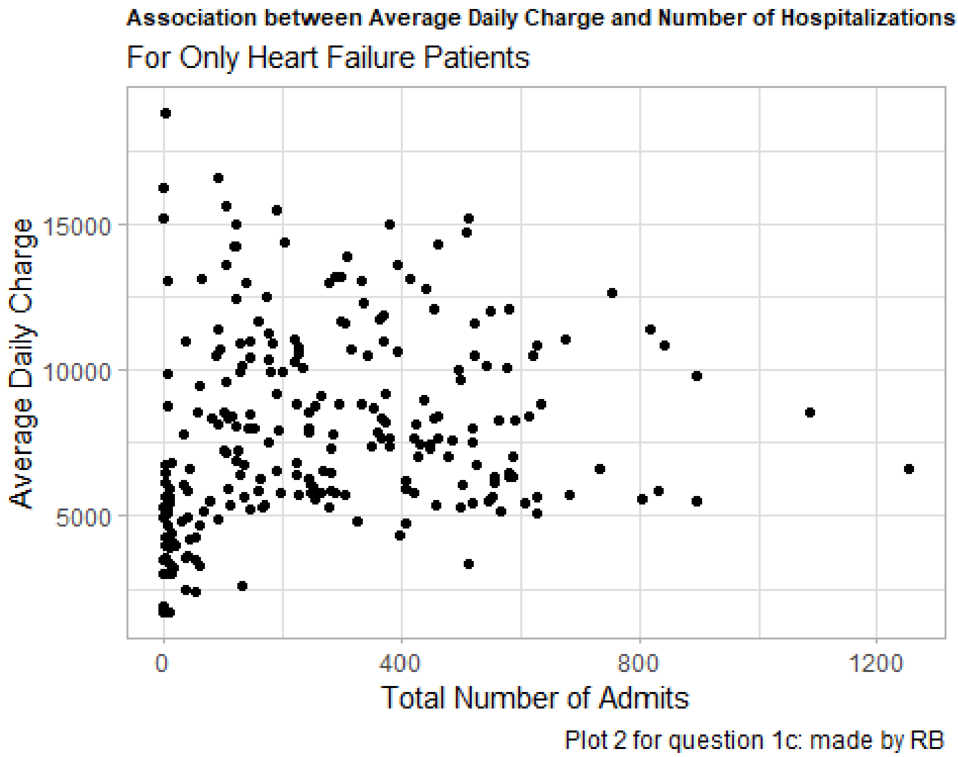
```
q3 <- FL_2014 %>%
  select(died, drg, dshospid,totchg,los)%>%
  filter(., drg %in% 291:293) %>%     #Filter for heart failure patients
  group_by(., dshospid) %>%
  summarize(., Count = n(),TotalDeath = sum(died), MortalityRate = mean(died)
,
             LOS = sum(los), TotalCharge = sum(totchg) , AverageDailyCharge =
sum(totchg)/sum(los))%>%
  arrange(desc(MortalityRate))

mygraph3 <- ggplot(q3, aes(y = q3$TotalCharge, x = q3$LOS)) +
  geom_point()

mygraph3 +
  theme_light() +
  labs(
    x = "Length of Stay in Hospital",
    y = "Total Charge",
    color = "Green",
    title = "Association between Total Charge and Length of Stay",
```

```
    subtitle = "For Only Heart Failure Patients",
    caption = "Plot 1 for question 1c: made by RB") +
  theme(plot.title = element_text(size = 8, face ='bold'))
```



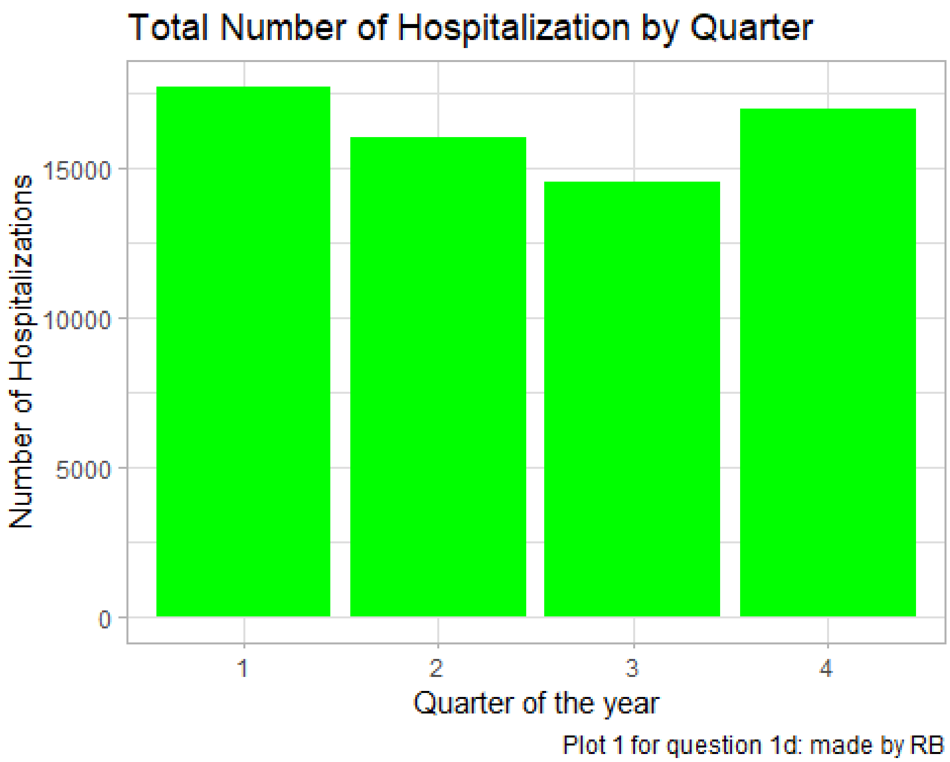
There is a strong association between length of stay in hospital and total charge a associated with it. As the duration of stay increases, the total cost of treatment increases. These two variables have a positive relationship. Additionally, total charge is dependent on the length of stay.

```
mygraph4 <- ggplot(q3, aes(y = q3$AverageDailyCharge, x = q3$Count)) +
  geom_point()

mygraph4 +
  theme_light() +
  labs(
    x = "Total Number of Admits",
    y = "Average Daily Charge",
    color = "Green",
    title = "Association between Average Daily Charge and Number of Hospitali
zations",
    subtitle = "For Only Heart Failure Patients",
    caption = "Plot 2 for question 1c: made by RB") +
    theme(plot.title = element_text(size = 8, face ='bold'))
```



Weak association between Average Daily Charge and Number of Total Admits in the hospital. There is not definite pattern (increase or decrease) with the increase in the number of admits.

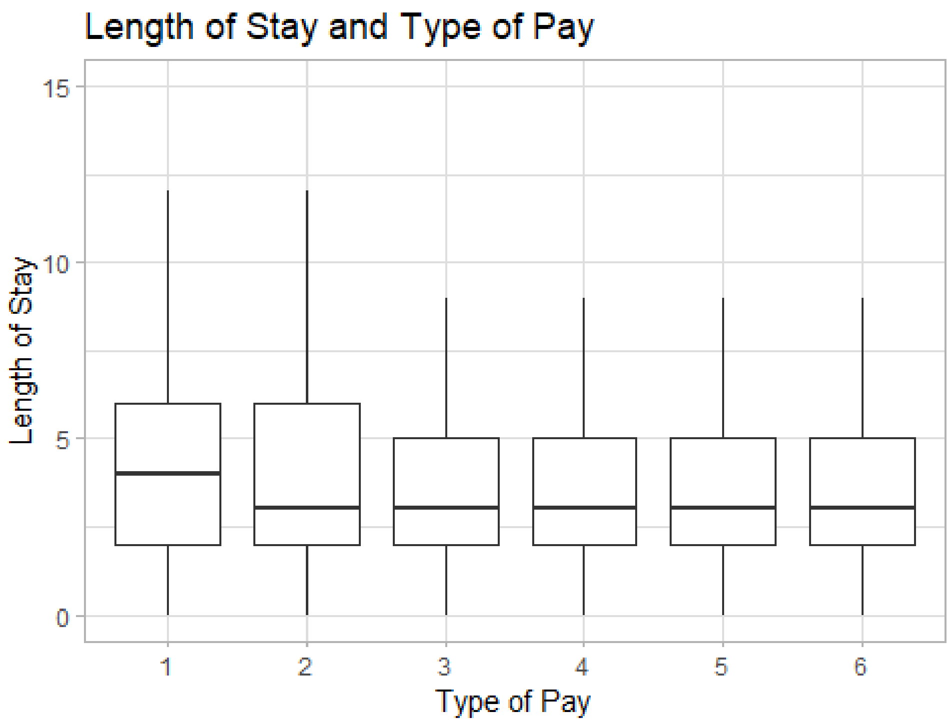
```
q4 <- FL_2014 %>%
  select(drg, dshospid, dqtr, totchg,)%>%
  filter(drg %in% 291:293)%>%            # Filter for heart failure patients
  group_by(dqtr)%>%
  summarise(N = n(), TotalCharges = mean(totchg))

mygraph5 <- ggplot(q4, aes(y= N, x= q4$dqtr))+
  geom_bar(stat="identity", fill="green")

mygraph5 + ggtitle("Total Number of Hospitalization by Quarter") +
  xlab("Quarter of the year") + ylab("Number of Hospitalizations")+
        labs(caption = "Plot 1 for question 1d: made by RB")+
    theme_light()
```



The number of hospital admits stays in the same range for all the quarters. It marginally varies between the four quarters. It is highest in the first quarter and lowest in the third quarter.

```
q5 <- FL_2014 %>%
    select(drg, dshospid, dqtr, totchg, pay1, los)%>%
    filter(drg %in% 291:293)%>%            # Filter for heart failure patient
s
    group_by(pay1)%>%
    mutate(DailyCharges = as.numeric(totchg) / as.numeric(los))

# 1 Medicare
# 2 Medicaid
# 3 Private insurance
# 4 Self-pay
# 5 No charge
# 6 Other

mygraph6a <- ggplot(q5, aes(x= pay1,y= los))+
  geom_boxplot(aes(group=pay1),outlier.shape = NA) + scale_y_continuous(limit
= c(0,15))

mygraph6a + ggtitle("Length of Stay and Type of Pay") +
  xlab("Type of Pay") + ylab("Length of Stay") +
  labs(caption = "Plot 2 for question 1e: made by RB") +
  theme_light()
```

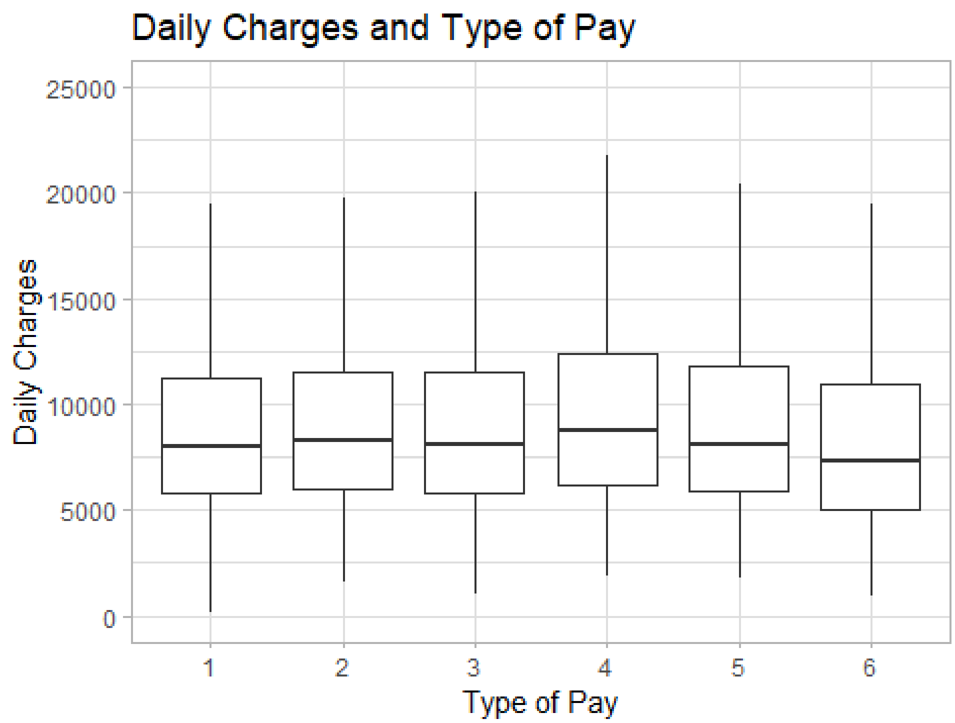
```
## Warning: Removed 1559 rows containing non-finite values (stat_boxplot).
```



Plot 2 for question 1e: made by RB

The length of stay has a median of less than five regardless of its pay type. The boxplots display a similar range and does not vary much amonth the different pay types. 1559 observations were ommited after setting the higher sclae of 15 on the Y-axis.

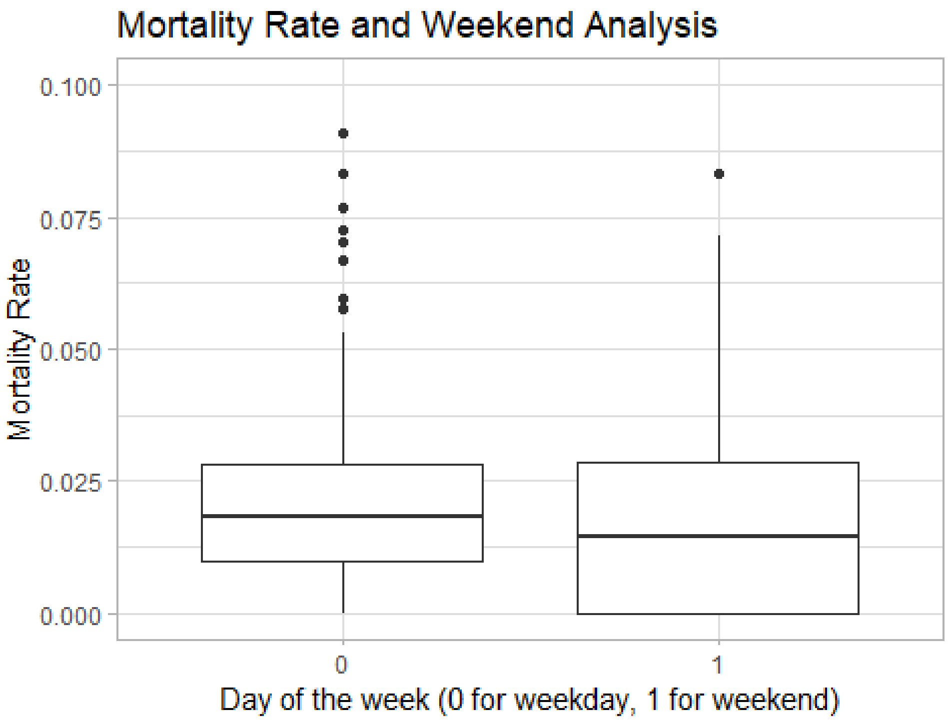
```
mygraph6 <- ggplot(q5, aes(x= pay1,y= DailyCharges))+
  geom_boxplot(aes(group=pay1),outlier.shape = NA) + scale_y_continuous(limit
= c(0,25000))

mygraph6 + ggtitle("Daily Charges and Type of Pay") +
xlab("Type of Pay") + ylab("Daily Charges") +
  labs(caption = "Plot 1 for question 1e: made by RB") +
theme_light()

## Warning: Removed 2417 rows containing non-finite values (stat_boxplot).
```



Plot 1 for question 1e: made by RB

The median for daily charges is around $7500 regardless of its pay type. There is not much variation among the different boxplots. 2417 rows were ommitted after setting an upper limit of 25000 on the Y-axis.

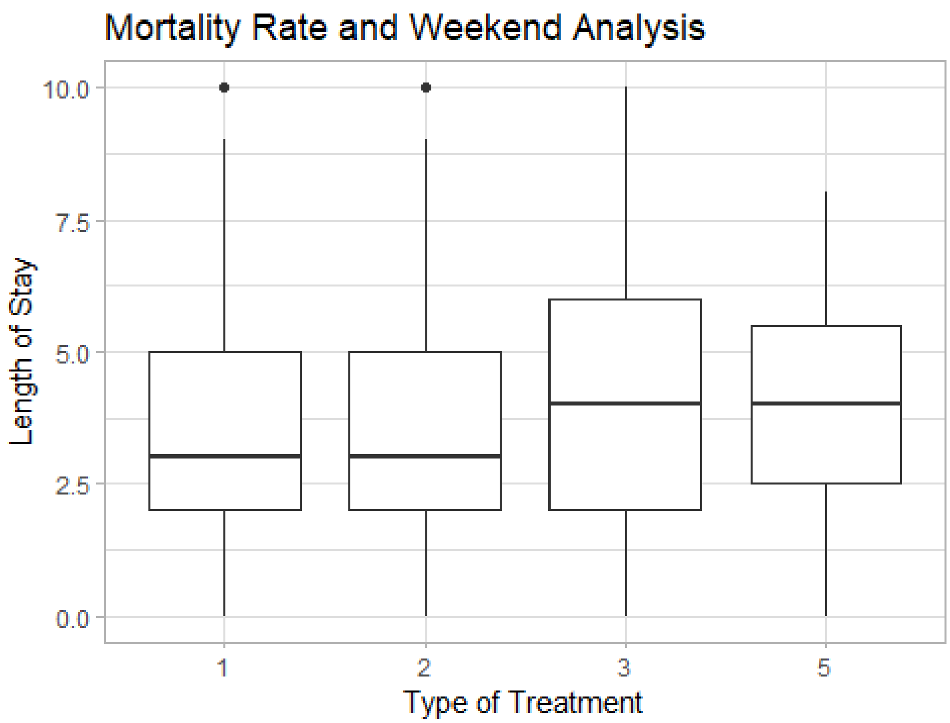
```
q6 <- FL_2014 %>%
  select(died, drg, dshospid, aweekend)%>%
  filter(., drg %in% 291:293) %>%     #Filter for heart failure patients
  group_by(., dshospid, aweekend) %>%
  summarize(., Count = n(),TotalDeath = sum(died), MortalityRate = mean(died)
)%>%
  arrange(desc(MortalityRate))

# '0'   Admitted Monday-Friday
# '1'   Admitted Saturday-Sunday

mygraph7 <- ggplot(q6, aes(x = aweekend, y = MortalityRate, group = aweekend)
) +
  geom_boxplot() + scale_y_continuous(limit = c(0,.1)) +
  theme_light()

  mygraph7 + ggtitle("Mortality Rate and Weekend Analysis") +
  xlab("Day of the week (0 for weekday, 1 for weekend)") + ylab("Mortality Ra
te") +
    labs(caption = "Plot 1 for question 1f: made by RB")

## Warning: Removed 20 rows containing non-finite values (stat_boxplot).
```



Plot 1 for question 1f: made by RB

There is no significant different on mortality rate based on the day of the week. The median for both the cases is above 1.25% but less than 2.5%.

```
q7 <- FL_2014 %>%
  select(died, drg, dshospid, atype, los)%>%
  filter(., drg %in% 291:293) %>%     #Filter for heart failure patients
  group_by(., dshospid)

mygraph8 <- ggplot(q7, aes(x = atype, y = los, group = atype)) +
  geom_boxplot() + scale_y_continuous(limit = c(0,10)) +
  theme_light()

mygraph8 + ggtitle("Mortality Rate and Weekend Analysis") +
  xlab("Type of Treatment") + ylab("Length of Stay") +
  labs(caption = "Plot 1 for question 1g: made by RB")

## Warning: Removed 4584 rows containing non-finite values (stat_boxplot).
```



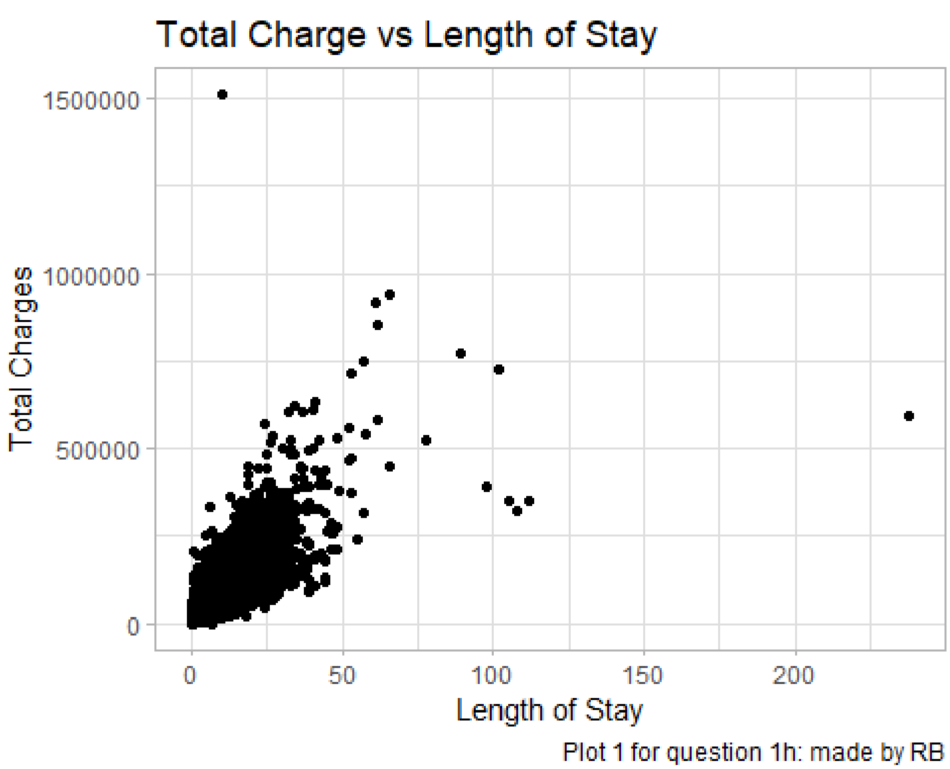
The median number for length of stay is higher for type 3 (elective) and type 5(Delivery). There is no length of stay of newborn (type 4). The hiher limit for length of stay on the Y-axis was set at 10, which ommitted 4584 rows.

```
linearMod <- lm(los ~ atype, data=q7)
linearMod

##
## Call:
## lm(formula = los ~ atype, data = q7)
##
## Coefficients:
## (Intercept)       atype2       atype3       atype5
##     4.73498      0.11865      1.14828     -0.06831

summary(linearMod)

##
## Call:
## lm(formula = los ~ atype, data = q7)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -5.883  -2.735  -0.735   1.265 233.265
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  4.73498    0.01749 270.720   <2e-16 ***
## atype2       0.11865    0.06723   1.765   0.0776 .
## atype3       1.14828    0.09615  11.943   <2e-16 ***
## atype5      -0.06831    1.22464  -0.056   0.9555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.242 on 65109 degrees of freedom
## Multiple R-squared:  0.002208,    Adjusted R-squared:  0.002162
## F-statistic: 48.03 on 3 and 65109 DF,  p-value: < 2.2e-16
```

Only 0.2% of the variation is explained by the variable 'atype'. Hence, the model does not provide a good estimate for the length of stay.

```
q8 <- FL_2014 %>%
  select(died, drg, dshospid, totchg, los)%>%
  filter(., drg %in% 291:293) %>%     #Filter for heart failure patients
  group_by(., dshospid)

mygraph9 <- ggplot(q8, aes(x = los, y = totchg, group = los)) +
  geom_point() +
  theme_light()

mygraph9 + ggtitle("Total Charge vs Length of Stay") +
  xlab("Length of Stay") + ylab("Total Charges")+
  labs(caption = "Plot 1 for question 1h: made by RB")
```



```
linearMod <- lm(totchg ~ los, data=q8)
linearMod
```

```
##
## Call:
## lm(formula = totchg ~ los, data = q8)
##
## Coefficients:
## (Intercept)          los
##        7930         6770
```

```
summary(linearMod)
```

```
##
## Call:
## lm(formula = totchg ~ los, data = q8)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1027295   -11385    -2937     8183  1435238
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7930.08     139.59   56.81   <2e-16 ***
## los          6769.71      21.84  310.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23660 on 65111 degrees of freedom
## Multiple R-squared:  0.5961, Adjusted R-squared:  0.5961
## F-statistic: 9.611e+04 on 1 and 65111 DF,  p-value: < 2.2e-16
```

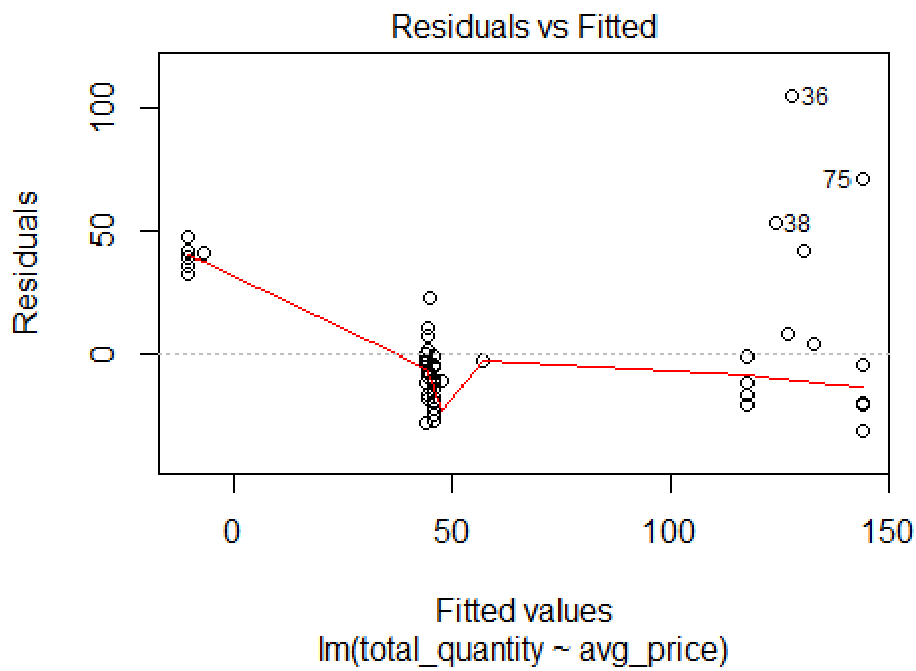
Slope is 6769.71 - An increase in stay at the hospital by each additional day will increase the cost by $6770. R-sqaure is approximately 60% which is extremely good, as one variable out of the 302 variables in the data explain about 60% variation in the model. RMSE is 23660, which portrays the typical error when we predict the total charge on the basis of length of stay.

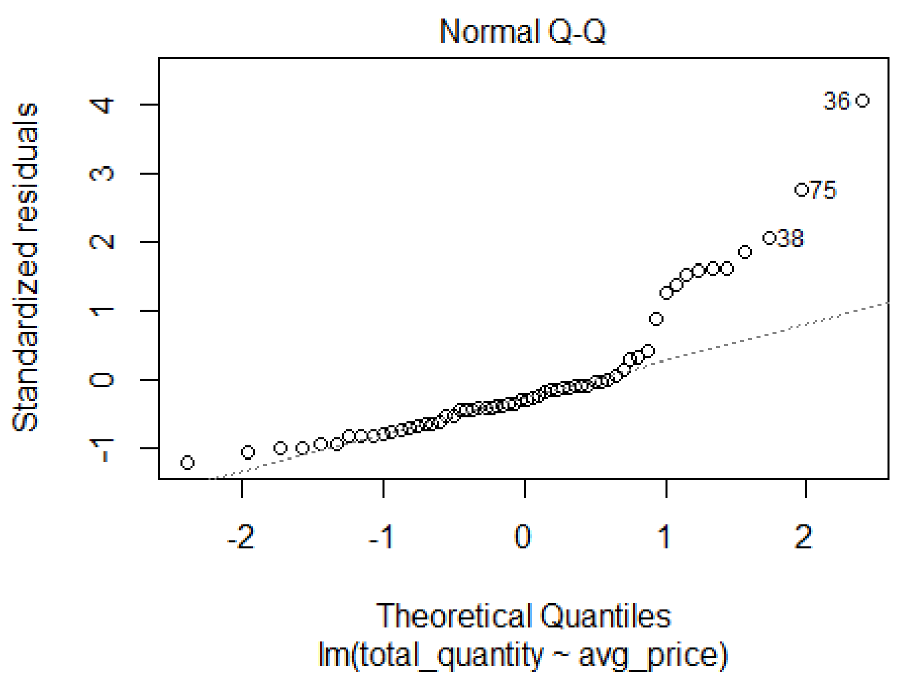
```
data <- read.csv("HW1PROB8.csv")
data <- data[-c(1:21,82:102),]
```

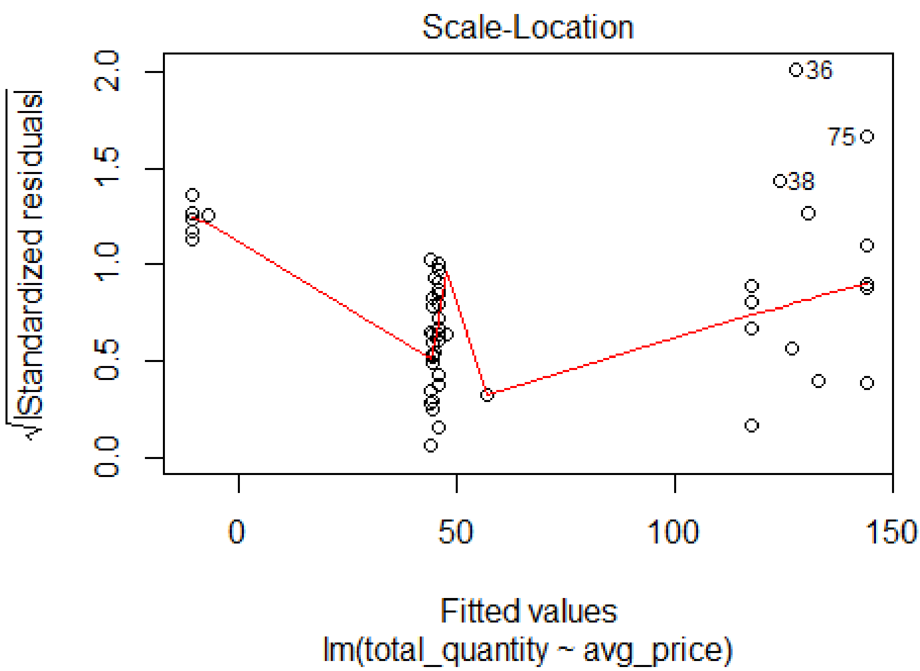
```
lm1 <- lm(total_quantity ~ avg_price, data = data)
lm1
```

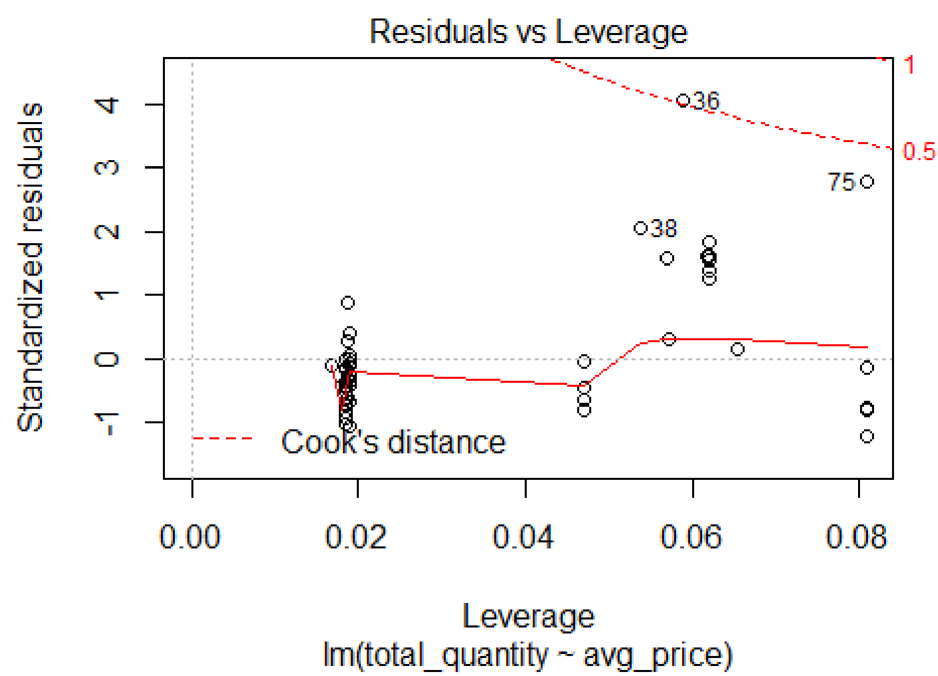
```
##
## Call:
## lm(formula = total_quantity ~ avg_price, data = data)
##
## Coefficients:
## (Intercept)    avg_price
##       335.6       -217.9
```

```
summary(lm1)

##
## Call:
## lm(formula = total_quantity ~ avg_price, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.896 -16.772  -7.644   2.262 105.011
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   335.64      22.49   14.92   <2e-16 ***
## avg_price    -217.90      17.56  -12.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.74 on 58 degrees of freedom
## Multiple R-squared:  0.7265, Adjusted R-squared:  0.7217
## F-statistic:   154 on 1 and 58 DF,  p-value: < 2.2e-16

plot(lm1)
```


Residuals vs Fitted
Residuals
100
50
0
0
50
100
150
Fitted values
lm(total_quantity ~ avg_price)
36
75
38


Normal Q-Q
Standardized residuals
4
3
2
1
0
-1
-2
-1
0
1
2
Theoretical Quantiles
lm(total_quantity ~ avg_price)
36
75
38





Response: R square is 72.65% which states that average price is able to estimate approximately 73% of variation in the total quantity. However, there are other variables which estimate the rest of the variation for total quantity.

Slope is -217.90 which means that one dollar increase in the average price will decrease the total quantity by approximately 218 units.

RMSE is 26.74 – It is the standard deviation of the residuals and measures how far the residuals are from the regression line. Hence, if we make a prediction for total quantity, we can expect an error of about 27 units.

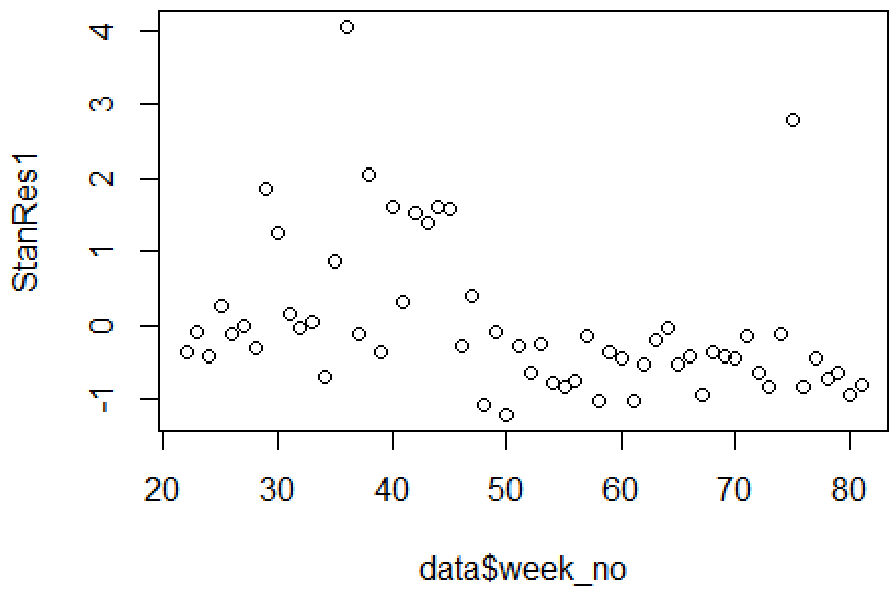
```
(leverage1 <- hatvalues(lm1))

##         22         23         24         25         26         27
## 0.01884780 0.01890956 0.01890956 0.01878691 0.01890956 0.01890956
##         28         29         30         31         32         33
## 0.01872687 0.06195467 0.06195467 0.06529279 0.01849537 0.01884780
##         34         35         36         37         38         39
## 0.01884780 0.01872687 0.05886103 0.01890956 0.05389182 0.01884780
##         40         41         42         43         44         45
## 0.06188019 0.05725818 0.06195467 0.06195467 0.06195467 0.05706440
##         46         47         48         49         50         51
## 0.01878691 0.01884780 0.01890956 0.01884780 0.08084824 0.01884780
##         52         53         54         55         56         57
## 0.01878691 0.01884780 0.08084824 0.01843964 0.01843964 0.01843964
##         58         59         60         61         62         63
## 0.01843964 0.01843964 0.01843964 0.01843964 0.01843964 0.01843964
##         64         65         66         67         68         69
## 0.04713604 0.01843964 0.01843964 0.01843964 0.01843964 0.01797695
##         70         71         72         73         74         75
## 0.04713604 0.08084824 0.04713604 0.08084824 0.01675277 0.08084824
##         76         77         78         79         80         81
## 0.01843964 0.01843964 0.01843964 0.01843964 0.01843964 0.04713604

(StanRes1 <- rstandard(lm1))

##           22           23           24           25           26
## -0.351748615 -0.079249477 -0.419042633  0.281827574 -0.117004273
##           27           28           29           30           31
## -0.003739887 -0.292676132  1.846041929  1.266872023  0.154686906
##           32           33           34           35           36
## -0.023566613  0.063541061 -0.691531077  0.877613565  4.047934716
##           37           38           39           40           41
## -0.117004273  2.043458807 -0.351748615  1.606703415  0.320372823
##           42           43           44           45           46
##  1.537151312  1.382706004  1.614373967  1.574669711 -0.284458957
##           47           48           49           50           51
##  0.403323523 -1.060874150 -0.087473367 -1.205122473 -0.276241401
##           52           53           54           55           56
## -0.624230876 -0.238487794 -0.776054881 -0.824451508 -0.748959994
##           57           58           59           60           61
## -0.145027887 -1.013180291 -0.371502427 -0.446993941 -1.013180291
##           62           63           64           65           66
## -0.522485454 -0.182773644 -0.028660192 -0.522485454 -0.409248184
##           67           68           69           70           71
```

```
## -0.937688778 -0.371502427 -0.407682648 -0.450069259 -0.151956565
##           72           73           74           75           76
## -0.641618836 -0.815061026 -0.103223783  2.773504291 -0.824451508
##           77           78           79           80           81
## -0.446993941 -0.711214237 -0.635722724 -0.937688778 -0.794858497
```

```
(residual1 <- lm1$residuals)
```

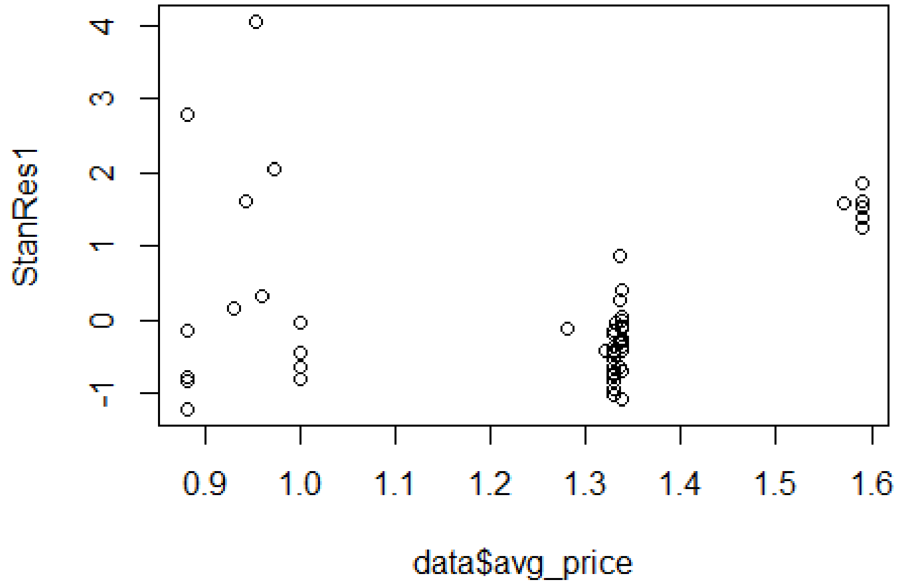
```
##           22           23           24           25           26
##  -9.31695390  -2.09905728 -11.09905728   7.46514948  -3.09905728
##           27           28           29           30           31
##  -0.09905728  -7.75274713  47.81089049  32.81089049   3.99912252
##           32           33           34           35           36
##  -0.62433361   1.68304610 -18.31695390  23.24725287 105.01074474
##           37           38           39           40           41
##  -3.09905728  53.15078049  -9.31695390  41.61388194   8.31812445
##           42           43           44           45           46
##  39.81089049  35.81089049  41.81089049  40.88875136  -7.53485052
##           47           48           49           50           51
##  10.68304610 -28.09905728  -2.31695390 -30.89570838  -7.31695390
##           52           53           54           55           56
## -16.53485052  -6.31695390 -19.89570838 -21.84223022 -19.84223022
##           57           58           59           60           61
##  -3.84223022 -26.84223022  -9.84223022 -11.84223022 -26.84223022
##           62           63           64           65           66
## -13.84223022  -4.84223022  -0.74811421 -13.84223022 -10.84223022
##           67           68           69           70           71
## -24.84223022  -9.84223022 -10.80329979 -11.74811421  -3.89570838
##           72           73           74           75           76
## -16.74811421 -20.89570838  -2.73706113  71.10429162 -21.84223022
##           77           78           79           80           81
## -11.84223022 -18.84223022 -16.84223022 -24.84223022 -20.74811421
```

```
(plot <- plot(data$week_no, StanRes1))
```



Response: No problem of non-linearity, non-constant variance around extreme values, presence of outliers, there is minimal auto-correlation, deviation in QQ-plot at extreme values, depicting data is not a normal distribution.

```
## NULL

predict(lm1, newdata = (avg_price= 1.59),interval="confidence",level=0.95)

##          fit       lwr      upr
## 1 -10.81089 -24.13423 2.512451

(plot <- plot(data$avg_price, StanRes1))
```



Response: Confidence interval is [-24.13, 2.51]. The confidence interval is not significant because it ranges from negative to positive quantity (The range covers large proportion of negative values). In reality, it is not plausible to have negative amount for total quantity. Additionally, a confidence interval with zero is not significant.

The model will under-predict by a large amount when average price is 1.59 because the standardized residuals is close to 2, portraying a large error.

```
## NULL
```

```r
lm2 <- lm(log(total_quantity) ~ log(avg_price), data = data)
lm2
```

```
##
## Call:
## lm(formula = log(total_quantity) ~ log(avg_price), data = data)
##
## Coefficients:
##    (Intercept)  log(avg_price)
##          4.605          -3.464
```

```r
summary(lm2)
```

```
##
## Call:
## lm(formula = log(total_quantity) ~ log(avg_price), data = data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82390 -0.17061  0.00956  0.18877  0.67901
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.60525    0.07063   65.20   <2e-16 ***
## log(avg_price) -3.46442    0.25339  -13.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3298 on 58 degrees of freedom
## Multiple R-squared:  0.7632, Adjusted R-squared:  0.7591
## F-statistic: 186.9 on 1 and 58 DF,  p-value: < 2.2e-16
```

```r
predict(lm2, newdata = (avg_price= 1.59),interval="confidence",level=0.95)
```

```
##        fit      lwr      upr
## 1 2.998679 2.849512 3.147846
```

Response: Taking the logarithmic scale removes the outliers (extreme values) and condenses the data. Now, we have a better confidence interval of [2.85,3.15] which provides a better estimate for total quantity. The values in the confidence interval is plausible which can be used with the model.