R Data Analysis

For the questions(1-11), you will need to use R and the following dataset:
The dataset was collected from Airbnb with data on listings in the city of Asheville, NC. Here is the data provided for each listing:

- room id: A unique number identifying an Airbnb listing.
- host id: A unique number identifying an Airbnb host.
- room type: One of 'Entire home/apt', 'Private room', or 'Shared room'
- reviews: The number of reviews that a listing has received.
- overall satisfaction: The average rating (out of five) that the listing has received from those visitors who left a review.
- accommodates: The number of guests a listing can accommodate.
- bedrooms: The number of bedrooms a listing offers.
- price: The price (in USD) for a night stay. In early surveys, there may be some values that were recorded by month.

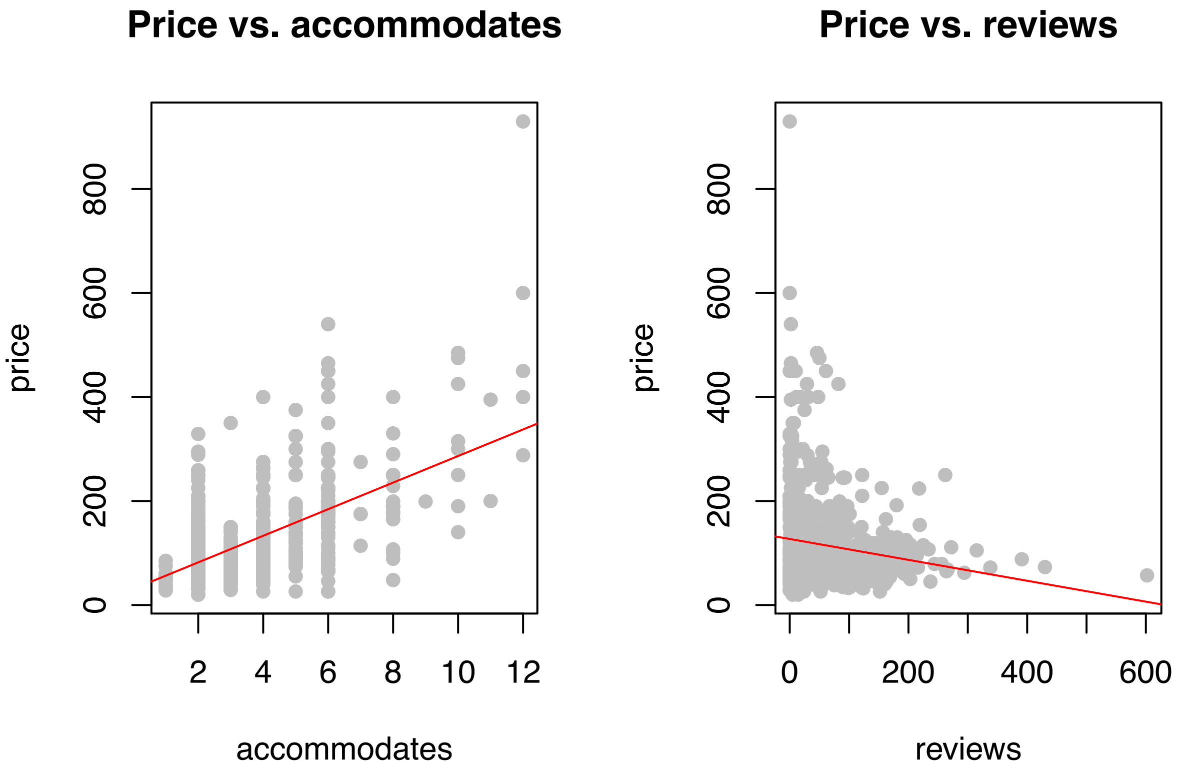
Fit a multiple linear regression model named 'model1' using price as the response variable and the following predicting variables: room type, reviews, overall satisfaction, accommodates, and bedrooms.

```r
# Read in the data
house = read.csv("tomslee_airbnb_asheville_1498_2017-07-20.csv", head = TRUE, sep = ",")
# Show the first few rows of data
head(house, 3)
```

```
##    room_id survey_id   host_id   room_type      city reviews
## 1 15771735      1498 101992409 Shared room Asheville       0
## 2 18284194      1498 126414164 Shared room Asheville      32
## 3 18091012      1498 122380971 Shared room Asheville       4
##   overall_satisfaction accommodates bedrooms price
## 1                  0.0            4        1    67
## 2                  5.0            4        1    76
## 3                  4.5            2        1    45
```

Question 1 Create plots of the response, *price*, against two quantitative predictors *accommodates*,and *reviews* . Describe the general trend (direction and form) of each plot

```r
# Grid the plots
par(mfrow=c(1,2))
# Plot price vs accommodates
plot(price~accommodates, data=house, main="Price vs. accommodates", col="grey", pch = 16)
abline(lm(price~accommodates, data=house), col="red")
# Plot price vs reviews
plot(price~reviews, data=house, main="Price vs. reviews",col="grey", pch = 16)
abline(lm(price~reviews, data=house), col="red")
```



Response to Question 1: General trend: There appears to be a positive and linear relationship between the response, price, and the predictor, accommodates. There appears to be a slight negative and linear relationship between the response, price, and the predictor, reviews. But we can also observe that there are lots of noise in these two scatters plots, so more analysis would need to be done to determine the strength of the relationships.

Question 2 What is the value of the correlation coefficient for each of the above pair of response and predictor variables? What does it tell you about your comments in Question 1?

```r
# Print the correlation coefficients between the predictors and the response
cat("cor(price, accommodates):", cor(house$price, house$accommodates)[1], end="\n")
```

```
## cor(price, accommodates): 0.5886389
```

```r
cat("cor(price, reviews):", cor(house$price, house$reviews)[1], end="\n")
```

```
## cor(price, reviews): -0.1532973
```

Response to Question 2: The correlation coefficient between price and accommodates (0.5886389) is the highest of the two groups. This isn't particularly high, but it does communicate that a moderate positive linear relationship between the two variables. The correlation coefficient between price and reviews (-0.1532973) shows a very slight negative linear relationship. These results reinforces that our comments about the general trend for the price vs. accommodates and price vs. reviews plots were correct.

Question 3 Use the *accommodates* as the predictor to build a simple linear regression model for predicting the *price*, named model1. What is the coefficient of *accommodates* in this model?

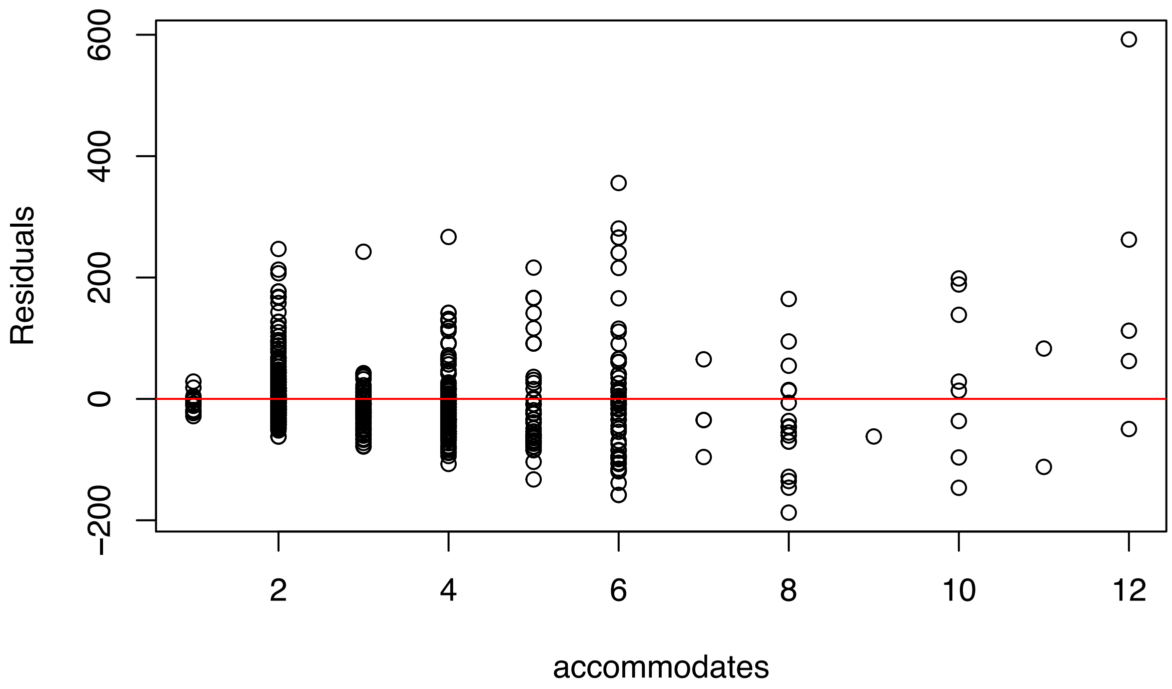
```r
model1=lm(price~accommodates, data = house);
summary(model1);
```

```
##
## Call:
## lm(formula = price ~ accommodates, data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -187.36  -34.80   -8.12   18.00  592.40
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    30.876      4.632   6.666 4.74e-11 ***
## accommodates   25.560      1.205  21.217  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.82 on 849 degrees of freedom
## Multiple R-squared:  0.3465, Adjusted R-squared:  0.3457
## F-statistic: 450.1 on 1 and 849 DF,  p-value: < 2.2e-16
```

Response to Question 3: The coefficient is 25.560.

Question 4: Assess whether the model assumptions hold, comment on whether there are any apparent departures from the assumptions of the linear regression model. Make sure that you state the model assumptions and assess each one. Each graph may be used to assess one or more model assumptions.

```r
## Residuals Vs Predictor Plots
library(MASS)
resids= residuals(model1)
plot(house[,8],resids,xlab="accommodates",ylab="Residuals")
abline(0,0,col="red")
```



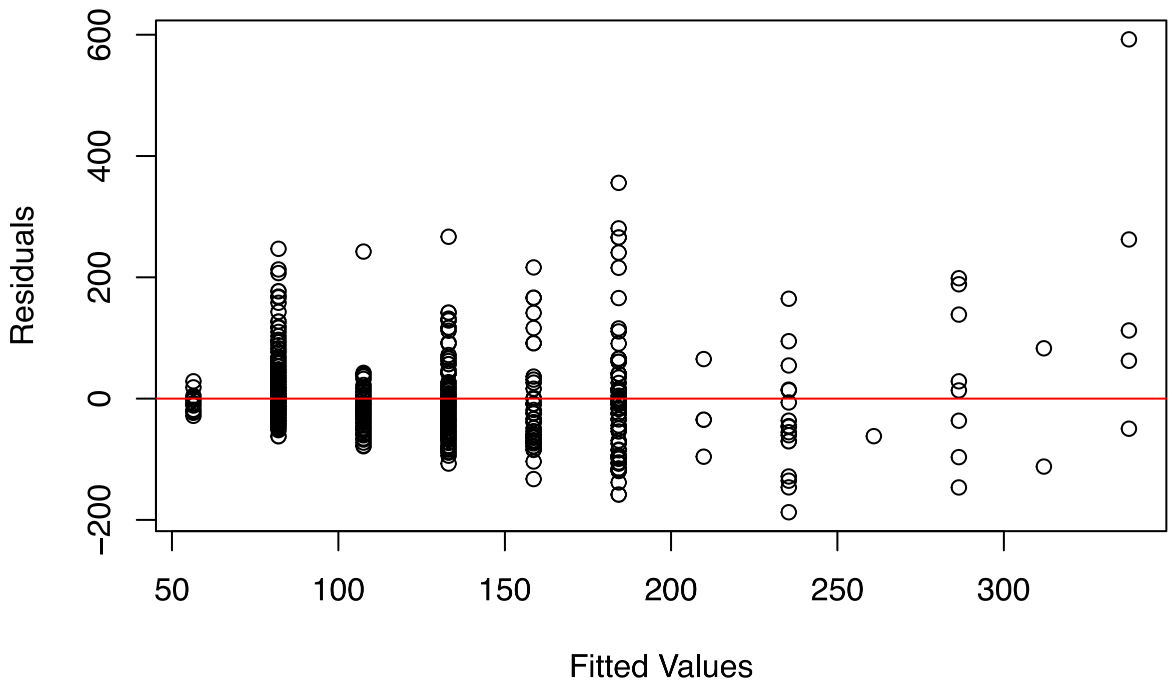
```
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.5
```

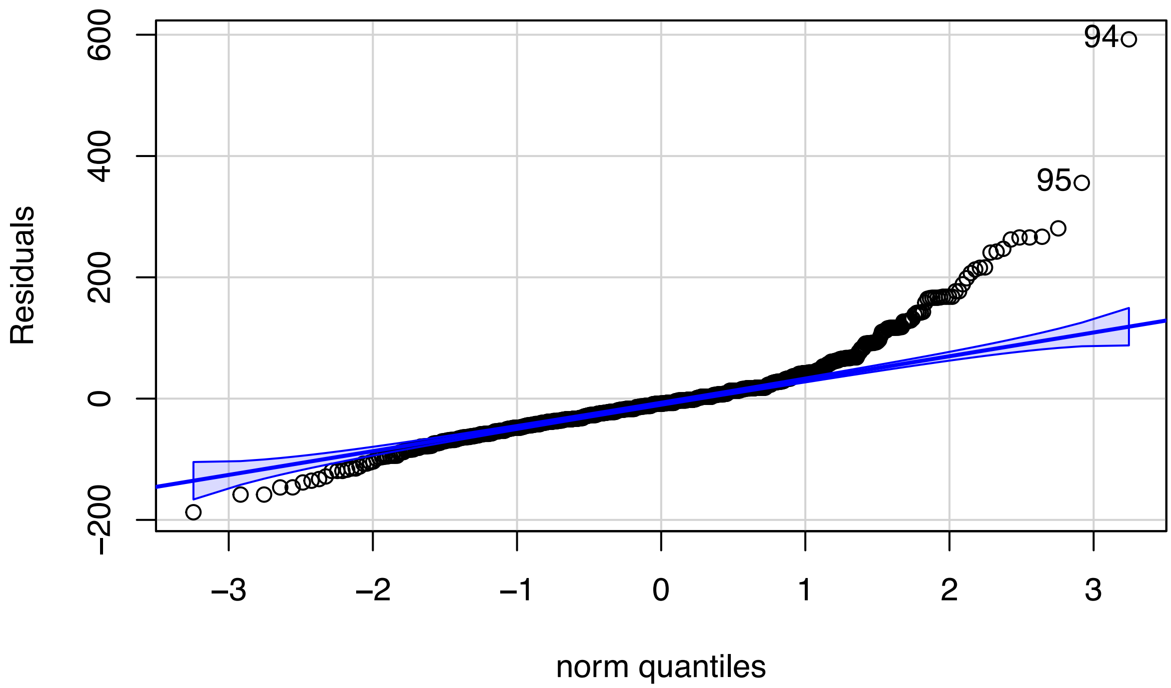
```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.0.5
```

```
# Residuals Vs Fitted and Q-Q plot
plot(predict(model1), resids, xlab="Fitted Values",ylab="Residuals")
abline(0,0,col="red")
```



```
qqPlot(resids, ylab="Residuals", main = "")
```

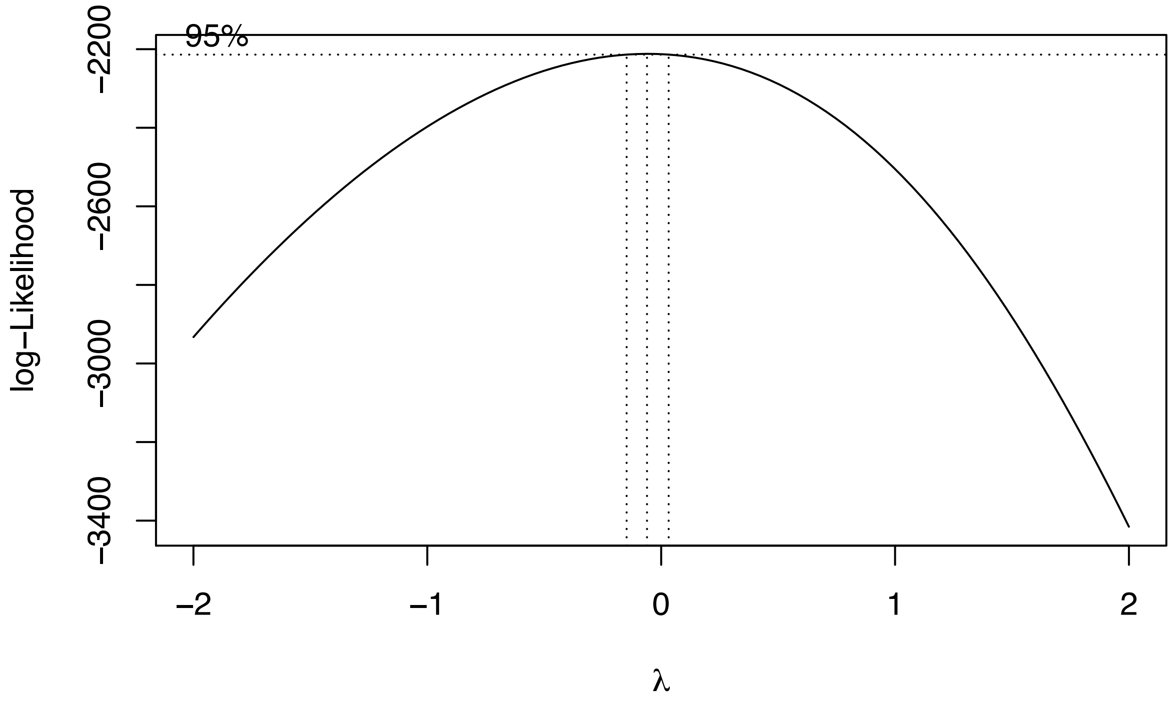


```
## [1] 94 95
```

Response to Question 4: From the residuals/predictor plot, the linearity/mean zero assumption appears to hold reasonably well. Data appears to be symmetrical about the zero line.
From the residuals/fitted plot, the constant variance assumption does not appear to hold. Lower values have smaller variance than higher values.
From the residuals/fitted plot, the uncorrelated error assumption holds, there are no apparent clusters of residuals.
From the qq plot, the normality assumption does not appear to hold. The data appears to be skewed to the right.

Question 5 For improving the fitness, we can use a box-cox transformation. Find the optimal lambda value rounded to the nearest half integer. Report this best lambda and corresponding transformation.

```
bc = boxcox(model1);
```



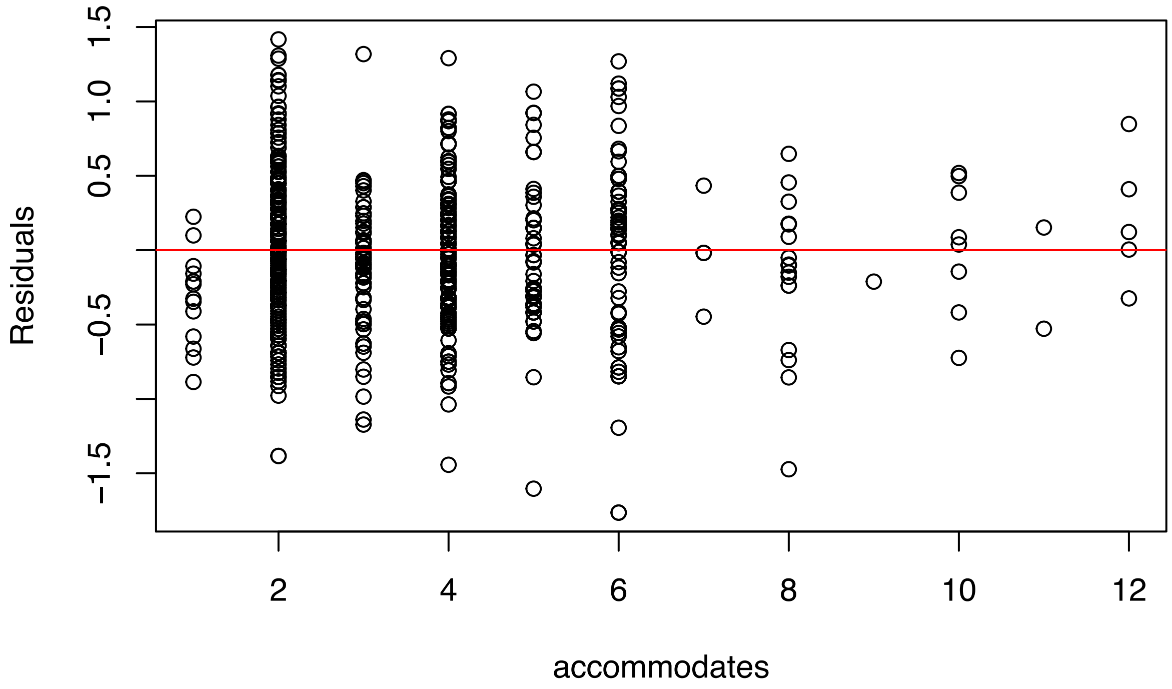
```
opt.lambda = bc$x[which.max(bc$y)]
cat("Optimal lambda:", round(opt.lambda/0.5)*0.5, end="\n")
```

```
## Optimal lambda: 0
```

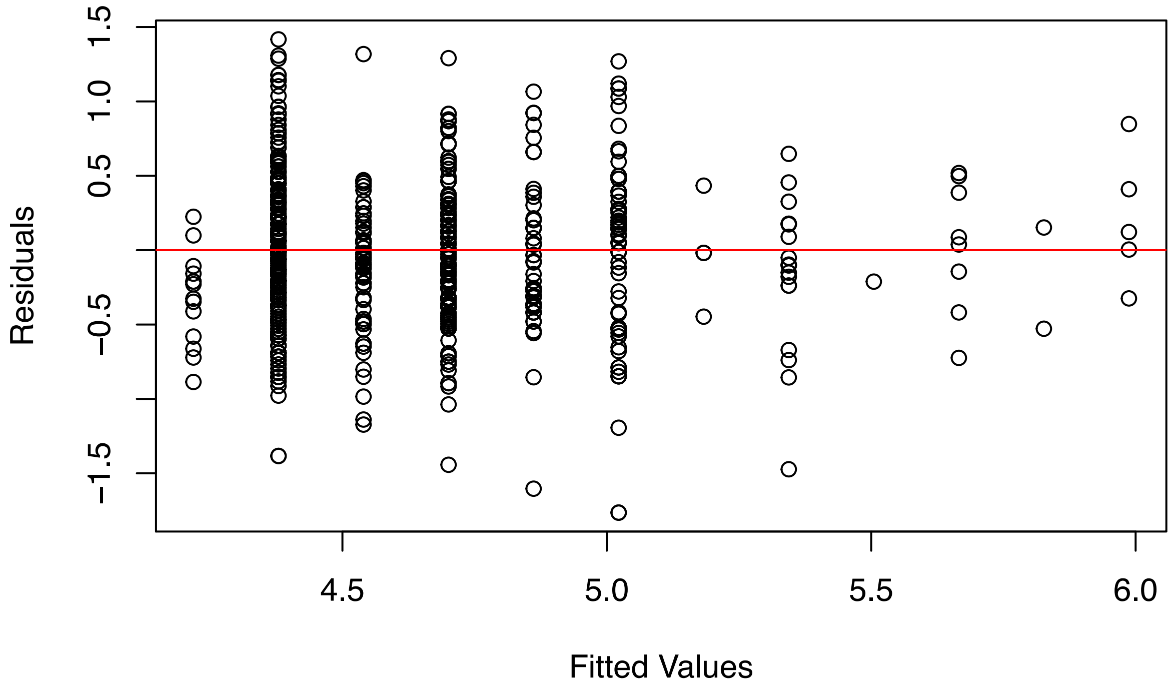
Response to Question 5 The optimal value of lambda should be 0. The optimal lambda value is zero, suggesting that the log of the response may improve constant variance and the normality.

Question 6 Use this optimal lambda value to transform the response variable. Build a new simple linear regression model, named model2, with the transformed response and the predictor *accommodates*. Similar to Question 4, check the model assumptions. Does this model seem a better fit?

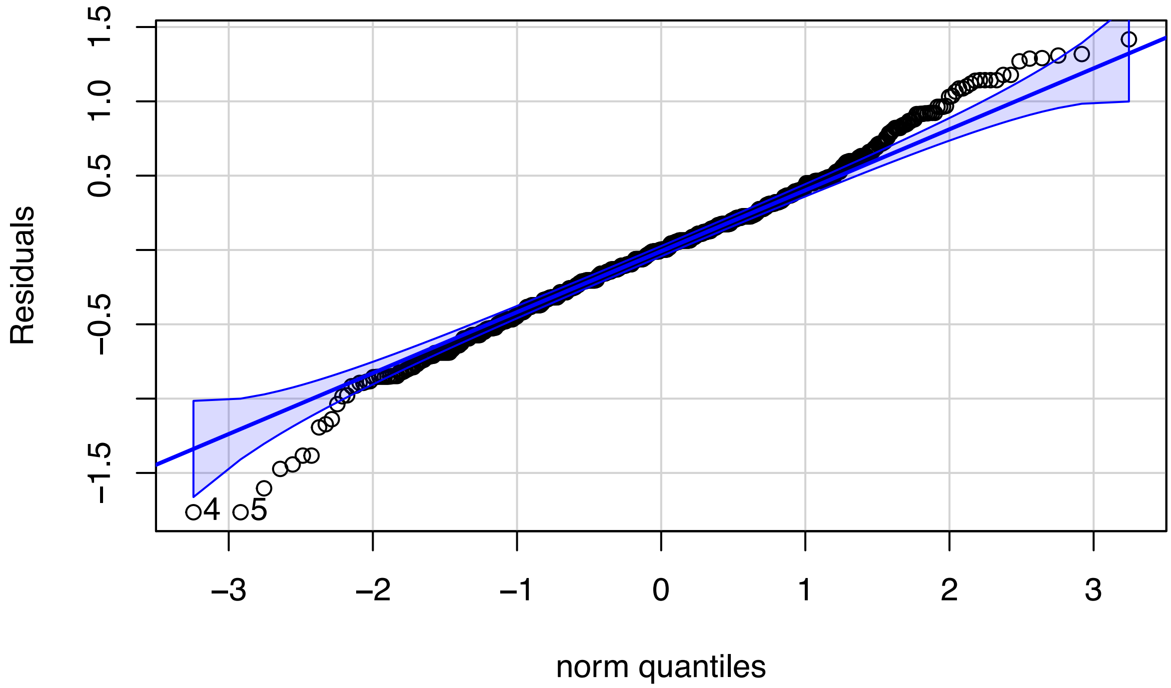
```
model2 = lm(log(price)~accommodates, data = house);
resids= residuals(model2)
plot(house[,8],resids,xlab="accommodates",ylab="Residuals")
abline(0,0,col="red")
```



```
plot(predict(model2), resids, xlab="Fitted Values",ylab="Residuals")
abline(0,0,col="red")
```



```
qqPlot(resids, ylab="Residuals", main = "")
```



```
## [1] 4 5
```

Response to Question 6 From the residuals/predictor plot, the linearity/mean zero assumption appears to hold reasonably well. Data appears to be symmetrical about the zero line.
From the residuals/fitted plot, the constant variance assumption appear to hold. Lower values and higher values have the same variance.
From the residuals/fitted plot, the uncorrelated error assumption holds, there are no apparent clusters of residuals.
From the qq plot, the normality assumption appear to hold.
So we can conclude that, after the transformation, the model assumptions can be better met.

Question 7 Fit a multiple linear regression model, using price as the response variable and the following predicting variables: room type, reviews, overall satisfaction, accommodates, and bedrooms. Which coefficients (including intercept) are statistically significant at the 99% confidence level?

```
# Build the model
model3 = lm(price ~ room_type + reviews + overall_satisfaction + accommodates + bedrooms, data = house)
# Show the 99% confidence intervals
confint(model3,level = 0.99)
```

```
##                             0.5 %       99.5 %
## (Intercept)             57.2982887  99.19648106
## room_typePrivate room  -45.6753949 -20.88675705
## room_typeShared room  -149.2845561 -37.38566358
## reviews                   -0.1599122   0.02653578
## overall_satisfaction     -8.7732808  -2.25692842
## accommodates              7.1044650  16.97542961
## bedrooms                 17.6616970  40.27329765
```

Response to Question 7:
Coefficient | 0.5 % | 99.5 % |
————|———|——|
(Intercept) | 57.2982887| 99.19648106
room_typePrivate room | -45.6753949 |-20.88675705
room_typeShared room | -149.2845561 | -37.38566358
reviews | -0.1599122 | 0.02653578 |
overall_satisfaction | -8.7732808 | -2.25692842
accommodates | 7.1044650 | 16.97542961
bedrooms | 17.6616970 | 40.27329765
☑ Intercept
☑ room type of Private room
☑ room type of Shared room
□ reviews
☑ overall satisfaction
☑ accomodates
☑ bedrooms

Question 8 What is the estimated coefficient for room type = "Private Room" in this MLR model?

```r
model3$coefficients['room_typePrivate room']
```

```
## room_typePrivate room
##             -33.28108
```

Response to Question 8: room_typePrivate room -33.28108

Question 9 What is the interpretation for the estimated coefficient for room type = "Private Room"?
Response to Question 9 A listing for a private room has an estimated cost of 33.28 USD less than an entire home/apt, holding all other variables constant.

Question 10 Report the coefficient of determination for your MLR model and give a concise interpretation of this value.

```r
# Extract R^2
cat("R^2:",summary(model3)$r.squared)
```

```
## R^2: 0.4353298
```

Response to Question 10: R^2 is 0.4353298 or 43.53%. We can interpret this as 43.53% of the variation in the response is explained by the predictors in the model.

Question 11 Using your MLR model, make a prediction for a listing on Airbnb in Asheville with the following factors:
bedrooms = 1, accommodates = 2, reviews = 92, overall_satisfaction = 3.5, and room_type= 'Private

room'.
What is your predicted price for such a listing and the corresponding 95% prediction interval?

```
new_data = data.frame(bedrooms=1, accommodates=2, reviews=92,
overall_satisfaction = 3.5, room_type= 'Private room')
predict(model3, new_data, interval="prediction", level=0.95)
```

```
##        fit       lwr      upr
## 1 72.57552 -46.28007 191.4311
```

Response to Question 11:
fit | lwr | upr |
—-|—–|—–|
72.57552 | -46.28007 | 191.4311