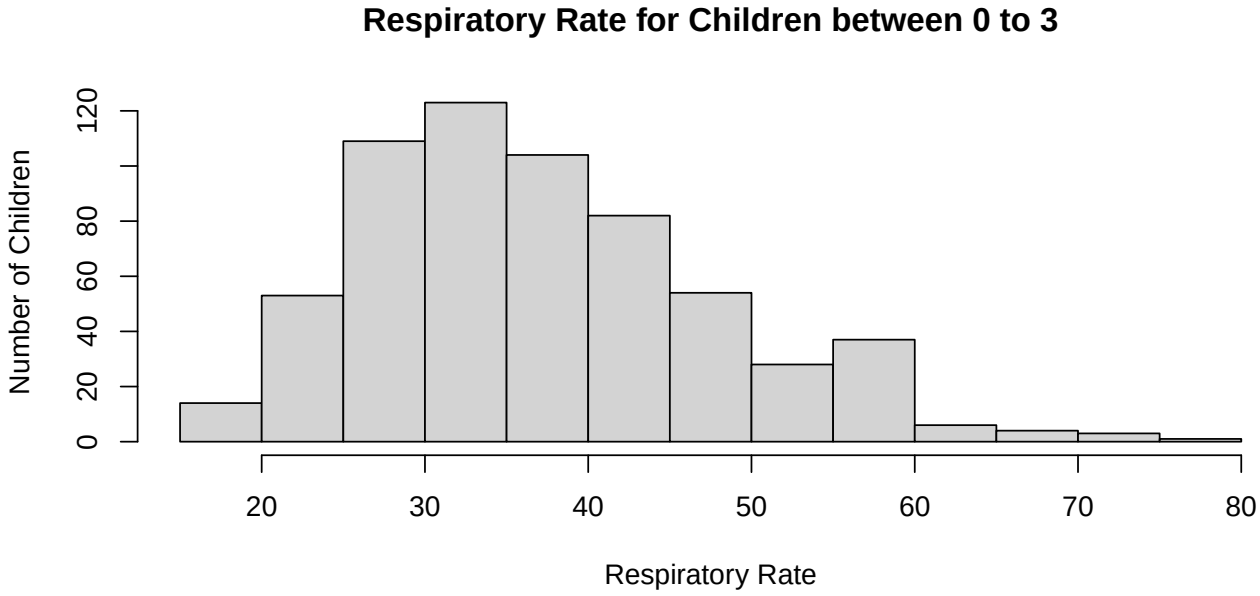
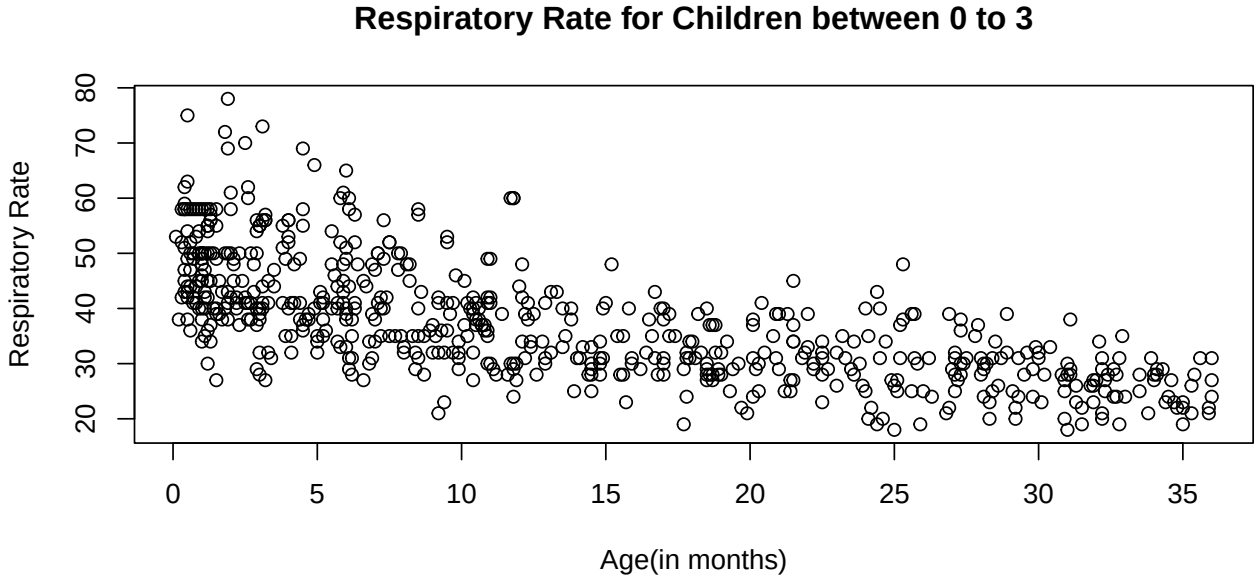
Data Analysis Homework 1

Pomelo Wu

2022-09-14

R Data Analysis 1 Question 1 a.





R Data Analysis 1 Question 1 b.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$Y : RespiratoryRate$$

$$X : Age(inmonth)$$

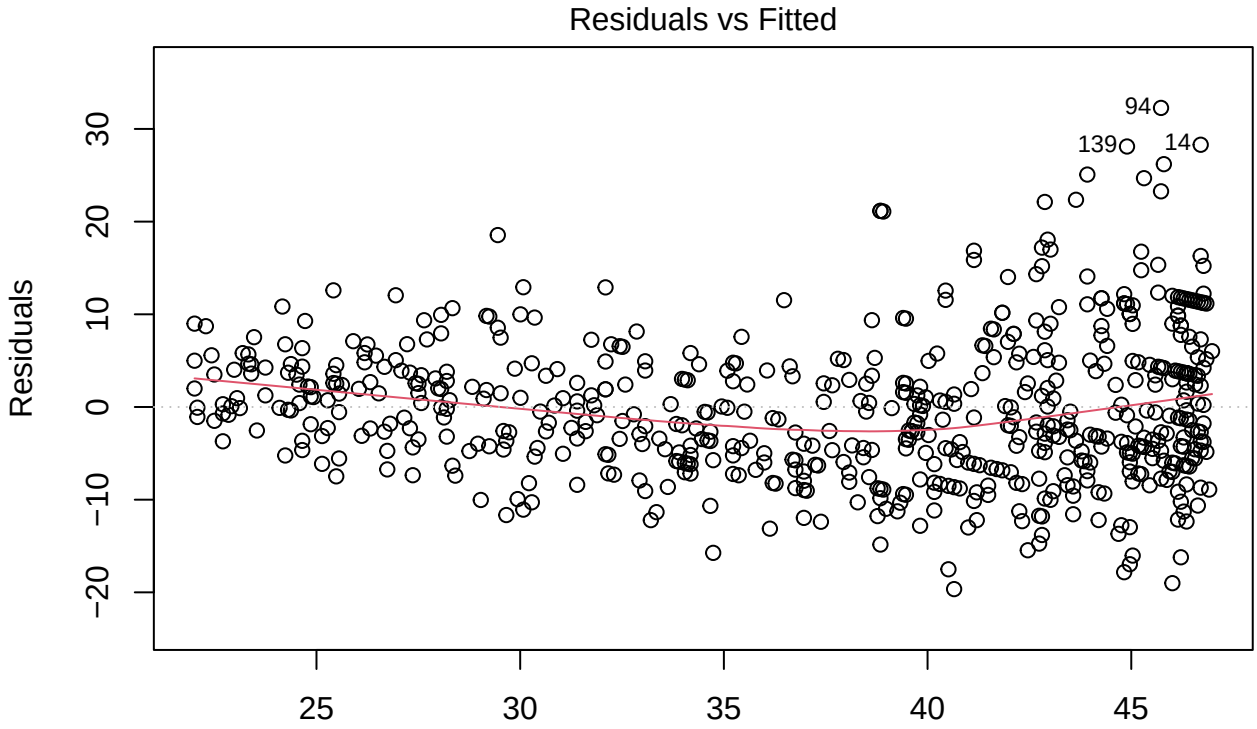
R Data Analysis 1 Question 1 c.

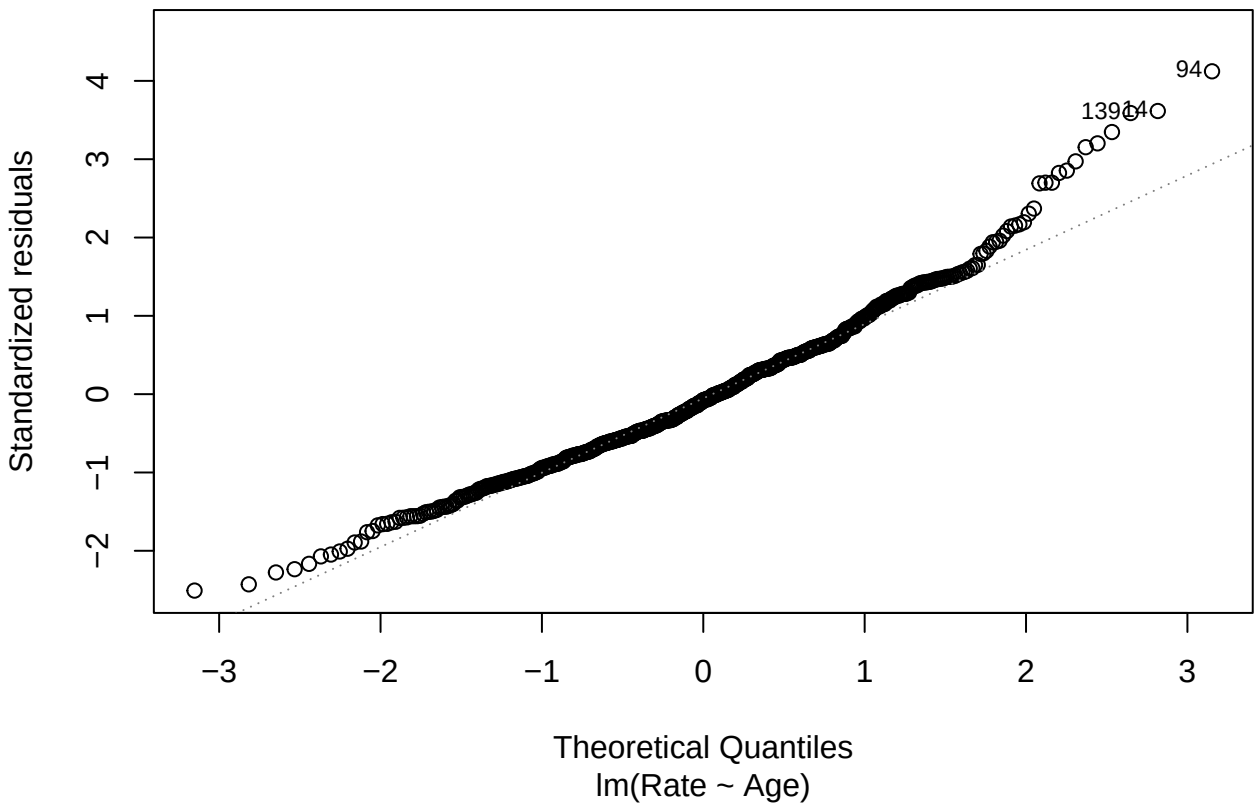
```
##
## Table1: Respiratory Rate of Children under 3 years old
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                  Rate
## ------------------------------------------------
## Age                           -0.6957***
##                           (-0.7533, -0.6381)
##
## Constant                      47.0522***
##                           (46.0639, 48.0404)
##
## ------------------------------------------------
## Observations                    618
## R2                             0.4766
## Adjusted R2                    0.4758
## Residual Std. Error       7.8422 (df = 616)
## F Statistic           560.9217*** (df = 1; 616)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

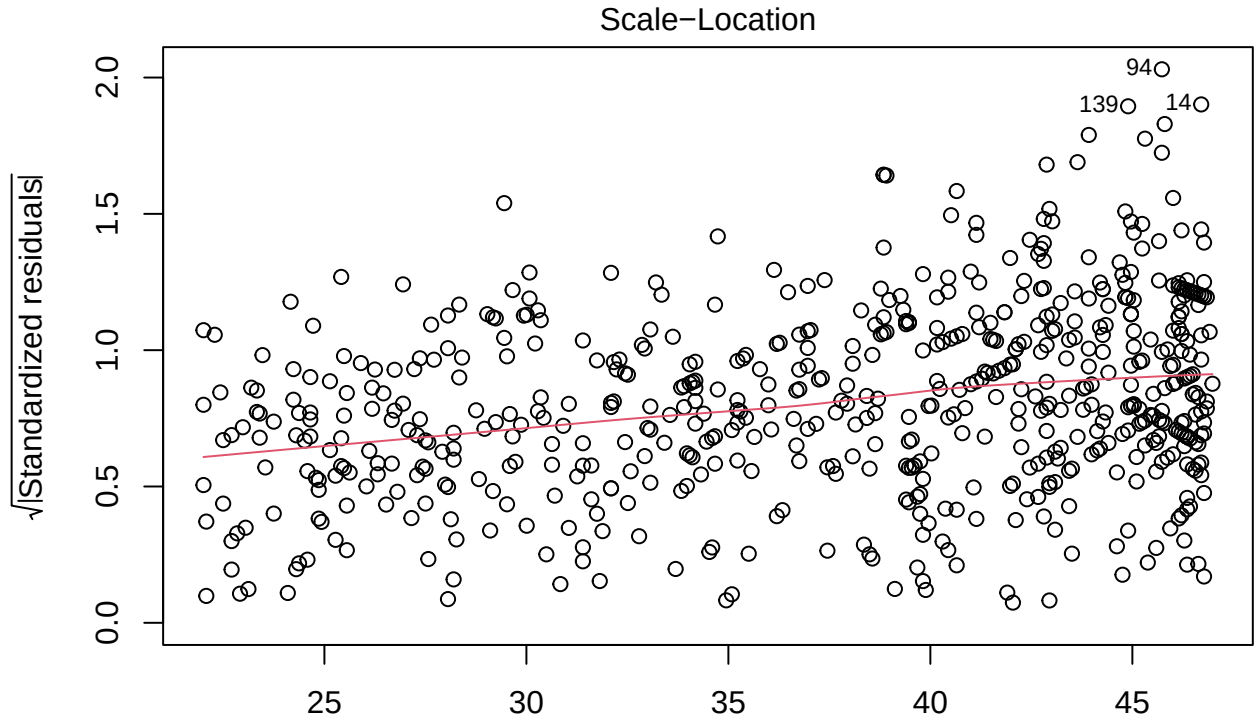
R Data Analysis 1 Question 1 d.

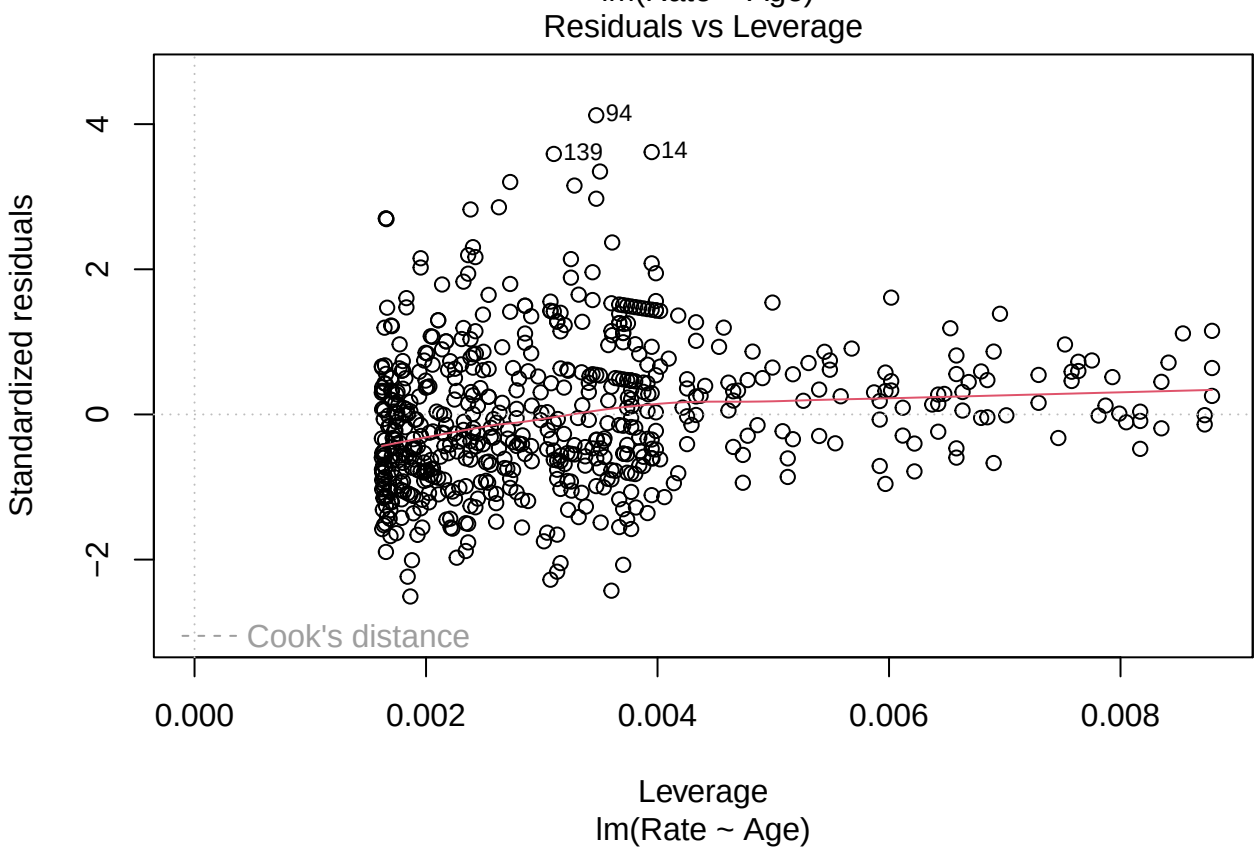
The fit of model is relatively low ($R^2 = 47.66\%$), meaning that his model can only explain about 47.66% variability of the respiratory rate in the regression model. The overall regression is statistically significant (p<0.01), however. The predictor variable, age, could statistically significantly predict the response variable, respiratory rate. On average, children under 3 years old have a respiratory rate of 47.05 with a 95% confidence interval[46.06, 48.04]. When age increases by 1, the respiratory rate on average would decrease by 0.6957.

R Data Analysis 1 Question 1 e.









Four assumptions of the linear model must be examined. The first is the linearity assumption. Based on

the plot, this model meets the assumption because the data is quite randomly scattered and does not have an obvious discernable pattern. The LOESS curve also appears quite linear. The second is the normality assumption. Based on the plot, this model largely meets the assumption because most data fall in the line of identity. The third is the constant variance assumption. This model also meets this assumption as data spread appears to be constant and the LOESS curve is mostly flat. The fourth is the independence assumption and since there is no obvious evidence that data falls outside Cook's distance, the assumption is not violated.

R Data Analysis 1 Question 2 a.

###See complete codes in the appendix

For EDA, I examined the histogram and boxplot of each numeric variable (accommodates, bathrooms, and bedrooms) and the summary of the categorical variable (host_is_superhost, identity_verified, and room_type) to observe its distribution. All three numeric variables are right skewed and the three categorical variables are not evenly distributed. Therefore, I knew that I need to be careful about extreme values. As for model fitting, I first adopted the linear model and regressed all the predictor variables directly on the response variable, price. This model seems to satisfy three assumptions (linearity assumption, normality assumption, and constant variance assumption), but violates the independence assumption. I also checked the summary statistics. Out of seven variables, four variables only have a significant level of 0.5 while two variables are not statistically significant. Next, I chose to log-transform the response variable to see if the log model could generate better results. The assumption check of the log-transformed model is similar to the linear model but the summary statistics seem to be better as more variables are statistically significant and the significance level of two other variables (accommodates and private room type) are higher. However, merely doing the y-value transformation decreases the interpretability of the model. Therefore, I used the linear model for this data set (see b.for detailed reasoning).

R Data Analysis 1 Question 2 b.

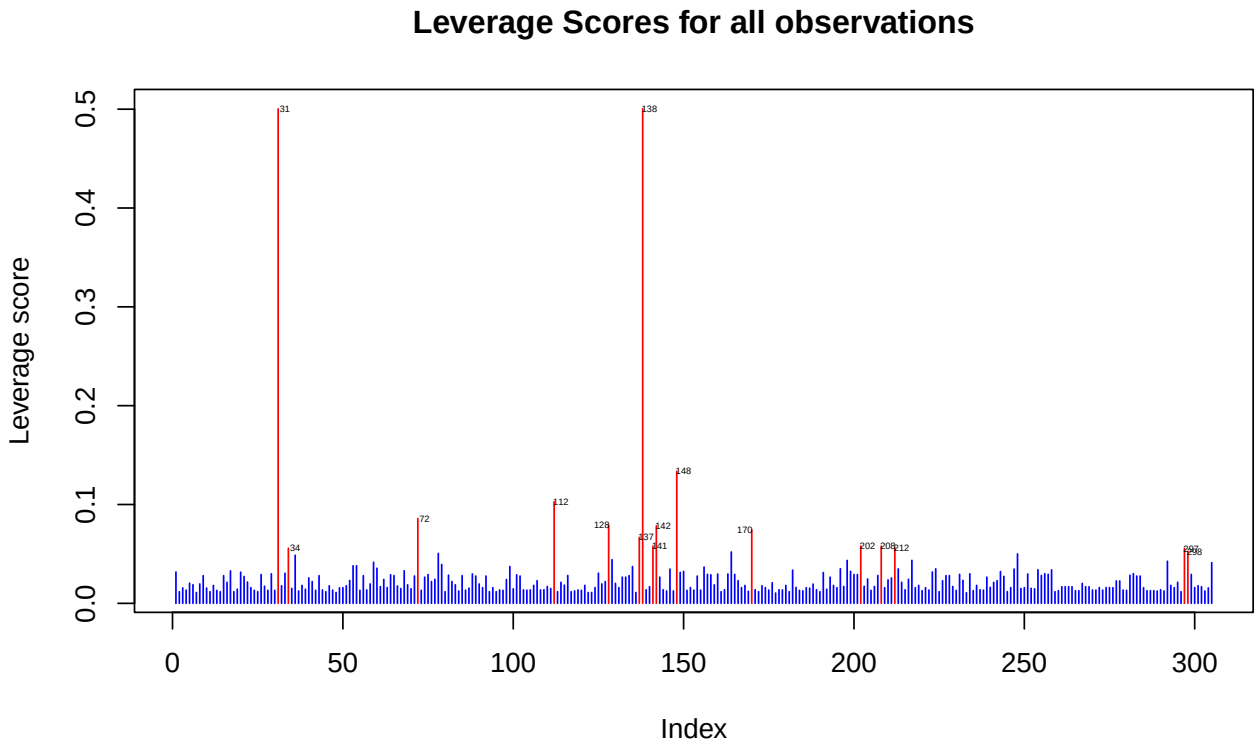
I chose the linear model because it does not make sense to describe the percent change of price in terms of percent change of bathroom, bedroom, and accommodates. Since the linear model and the log-transformed both satisfies three assumptions and the (R^2) are about the same, plus that the linear model is easier to interpret, I adopted the linear model in the following analysis.

```
##
## Table2: Airbnb Price and Contributory Factors
## ============================================================
##                                    Dependent variable:
##                                ---------------------------
##                                           price
## ------------------------------------------------------------
## accommodates                              7.2127**
##                                        (0.1964, 14.2291)
##
## bathrooms                                 88.2053***
##                                       (65.3422, 111.0683)
##
## bedrooms                                  33.4446***
##                                       (15.0358, 51.8533)
##
## host_is_superhostFalse                     -2.3324
##                                       (-23.6622, 18.9975)
##
## host_identity_verifiedFalse                14.7551
##                                       (-6.3079, 35.8181)
##
## room_typePrivate room                     -41.0729**
##                                       (-72.1532, -9.9926)
##
## room_typeShared room                      170.3975**
##                                       (39.6550, 301.1401)
##
## Constant                                  -28.0341**
##                                       (-55.2859, -0.7823)
##
## ------------------------------------------------------------
## Observations                                 305
## R2                                         0.6680
## Adjusted R2                                0.6602
## Residual Std. Error                   92.6729 (df = 297)
## F Statistic                    85.3763*** (df = 7; 297)
## ============================================================
## Note:                          *p<0.1; **p<0.05; ***p<0.01
```

R Data Analysis 1 Question 2 c.

The fit of model is relatively high ($R^2 = 66.8\%$), meaning that this linear model can explain about 66.8% variability in the price. The overall regression model is statistically significant ($p<0.01$). Two predictor variables, bedrooms and bathrooms, are highly significant ($p<0.01$). One unit increase in bedroom is related to 33.45 unit increase in the price with a 95% confidence level [15.04, 51.85], and one unit change in bathroom corresponds to 88.21 unit change in the price with a 95% confidence level [65.34, 111.07]. Accommodates and room type are statistically significant as well($p<0.05$). One more accommodate is associated with 7.21 increase in price with a 95% confidence level [0.20, 14.23]. Compared to the base case of entire home/apartment, private room is associated with 41.07 decrease in price with a 95% confidence level[-72.15, -9.99]; meanwhile, compared to the entire home/apartment, shared room is related to 170.40 increase in price with a 95% confidence level[39.66, 301.14].

R Data Analysis 1 Question 2 d.



Based on exploratory data analysis in question a, there are some obvious outliers that influence the distributions and mean values of the three quantitative variables. By plotting to observe the cook's distance and conducted a calculation on leverage, it becomes clear that there are some influential points and high leverage points.

Excluding Outliers, influential points, and leverage points

```
##
## Table3: Comparing Models
## ================================================================================
##                                          Dependent variable:
##                                 ---------------------------------------------
##                                                    price
##                                        (1)                       (2)
## --------------------------------------------------------------------------------
## accommodates                          7.2127**                  7.9738**
##                                  (0.1964, 14.2291)         (1.8283, 14.1192)
##
## bathrooms                             88.2053***                70.3310***
##                                  (65.3422, 111.0683)       (50.3955, 90.2665)
##
## bedrooms                              33.4446***                30.7142***
##                                  (15.0358, 51.8533)        (14.9435, 46.4849)
##
## host_is_superhostFalse                -2.3324                   -1.7110
##                                  (-23.6622, 18.9975)       (-19.3238, 15.9017)
##
## host_identity_verifiedFalse           14.7551                   18.3674**
##                                  (-6.3079, 35.8181)         (0.8917, 35.8431)
##
## room_typePrivate room                 -41.0729**                -45.4167***
##                                  (-72.1532, -9.9926)       (-71.1804, -19.6529)
##
## room_typeShared room                  170.3975**
##                                  (39.6550, 301.1401)
##
## Constant                              -28.0341**                -3.9649
##                                  (-55.2859, -0.7823)       (-27.3368, 19.4069)
##
## --------------------------------------------------------------------------------
## Observations                             305                       300
## R2                                      0.6680                    0.6796
## Adjusted R2                             0.6602                    0.6731
## Residual Std. Error              92.6729 (df = 297)        76.4173 (df = 293)
## F Statistic                  85.3763*** (df = 7; 297) 103.5933*** (df = 6; 293)
## ================================================================================
## Note:                                              *p<0.1; **p<0.05; ***p<0.01
```

I excluded points beyond the cook's distance scope as well as significant high leverage points after performing calculations on leverage points. The model fit increases slightly from ($R^2 = 66.8\%$) to ($R^2 = 67.96\%$), but the overall model is still highly statistically significant (p<0.01). Though the predictor variables, bathroom and bedroom, are statistically significant, their coefficient values change. On average, 1 unit increase in bathroom decreases from 88.21 unit increase in price to 70.33 unit increase in price. Correspondingly, the 95% confidence interval also narrows down, changing from [65.34, 111.07] to [50.40, 90.27]. On average, 1 unit increase in bedroom no longer relates to a 33.45 change in price; instead, it is associated with 30.71 increase in price. The 95% confidence interval narrows down from [15.04, 51.85] to [14.94, 46.49]. However, the impact of private room becomes more statistically significant(p<0.01) compared to the previous linear model(p<0.05). The shared room is excluded out under the outlier-exclusion process. Therefore, only the private room appears. On average, the change from entire room/apartment to private room would experience a higher decrease from 41.07 to 45.42 in price. Host identity verified also becomes statistically significant

($p<0.05$). This means that on average, the verified host is related to 18.37 increase in price compared to unverified host.

R Data Analysis 1 Question 2 e.

Yes, there are some significant limitations for this analysis flow. First, we do not include a research question, such that we fail to focus our analysis on variables of interest. Instead of including all variables in one single analysis, we should have a question and conduct our analysis accordingly to that specific question. This is very important because it is not only the statistically significance matters, but the practical importance/significance also matters. Secondly, using plot to visualize each predictor variable on the response variable might be helpful (we haven't learned about this though; after searching online, I believe a matrix of variable might help). Moreover, the linear model and log transformation of the response variable might not be the most fit model at all. Some other possibilities, such as interactions between variables might need discussions and to be ruled out. Last but not the least, though some leverage points exist, that may be due to lack of data. It seems that during the data exclusion process, the shared room type is entirely excluded. However, this room type may still be valuable in data analysis. Instead of directly excluding these data points, collecting more data may be necessary for a comprehensive analysis.

Appendix: Code for Data Analysis Assignment #1

```
install.packages("stargazer",repos = "http://cran.us.r-project.org")
library(stargazer)
knitr::opts_chunk$set(echo = TRUE)

#Read the CSV file
res <- read.csv(file = "Respiratory.csv",
                stringsAsFactors = FALSE, sep = ",",
                 dec="," , nrows=618)
#Brief read through the data set
head(res)
summary(res)
#Ensure variables are in the correct format
res$Age <- as.numeric(res$Age)
res$Rate <- as.numeric(res$Rate)
#EDA on variables
boxplot(res$Age)
boxplot(res$Rate)
hist(res$Rate, main= "Respiratory Rate for Children between 0 to 3",
     xlab="Respiratory Rate", ylab="Number of Children")
#Scatterplot to observe their relationship
plot(x=res$Age,y=res$Rate,main= "Respiratory Rate for Children between 0 to 3",
     xlab="Age(in months)", ylab="Respiratory Rate")
mod <- lm(Rate~Age,data=res)
summary(mod)
confint(mod, level=0.95)
stargazer(mod,type="text",
          title="Table1: Respiratory Rate of Children under 3 years old",
          ci=TRUE,digits=4)
plot(mod)
#Read the txt file
airbnb <- read.table(file = "airbnb.txt", header=TRUE)
head(airbnb)
summary(airbnb)

#Assign correct class to each variable
airbnb$bathrooms <- as.numeric(airbnb$bathrooms)
airbnb$bedrooms <- as.numeric(airbnb$bedrooms)
airbnb$accommodates <- as.numeric(airbnb$accommodates)
airbnb$price <- as.numeric(airbnb$price)
airbnb$host_is_superhost<- factor(airbnb$host_is_superhost,
                                  levels=c("True","False"))
airbnb$host_identity_verified <- factor(airbnb$host_identity_verified,levels=c("True","False"))
airbnb$room_type<- factor(airbnb$room_type)

#Exploratory Data Analysis
hist(airbnb$accommodates)
hist(airbnb$bathrooms)
hist(airbnb$bedrooms)
boxplot(airbnb$accommodates)
boxplot(airbnb$bathrooms)
boxplot(airbnb$bedrooms)
```

```
summary(airbnb$host_is_superhost)
summary(airbnb$host_identity_verified)
summary(airbnb$room_type)

#Multi-linear Model
lmmod <- lm(price ~ accommodates+bathrooms+bedrooms+host_is_superhost+
              host_identity_verified+room_type, data=airbnb)
summary(lmmod)
plot(lmmod)

#Transform Data
new_price <- log(airbnb$price)
##Log-transformed outcome variable model
lmmod2 <- lm(new_price ~ accommodates+bathrooms+bedrooms+host_is_superhost+
               host_identity_verified+room_type, data=airbnb)
summary(lmmod2)
plot(lmmod2)
lmmod <- lm(price ~ accommodates+bathrooms+bedrooms+host_is_superhost+
              host_identity_verified+room_type, data=airbnb)
summary(lmmod)
plot(lmmod)
confint(lmmod, level=0.95)
stargazer(lmmod,type="text",title="Table2: Airbnb Price and Contributory Factors",
          digits=4,ci=TRUE)
m <- nrow(model.matrix(lmmod))
p <- ncol(model.matrix(lmmod))
standard <- 2*p/m
leverage_value <-hatvalues(lmmod)
plot(leverage_value, col=ifelse(leverage_value > standard, 'red','blue'), type='h',
     ylab="Leverage score",
     xlab="Index",
     main="Leverage Scores for all observations")
text(x=c(1:m)[leverage_value > standard]+c(rep(2,4),-2,2),
     y=leverage_value[leverage_value > standard],
labels=c(1:m)[leverage_value > standard],cex=0.3)
library(dplyr)
new_airbnb <- airbnb %>% slice(-c(31,138,148,72,112))
summary(new_airbnb)
lmmod_new <- lm(price~accommodates+bathrooms+bedrooms+host_is_superhost+
                  room_type+host_identity_verified, data=new_airbnb)
summary(lmmod_new)
plot(lmmod_new)
confint(lmmod_new, level=0.95)
stargazer(lmmod,lmmod_new,type="text",title="Table3: Comparing Models",digits=4,ci=TRUE)
```