Data Analysis Assignment 1

Xiaoquan Liu

2022-09-11

Question 1

a

Exploratory Analysis

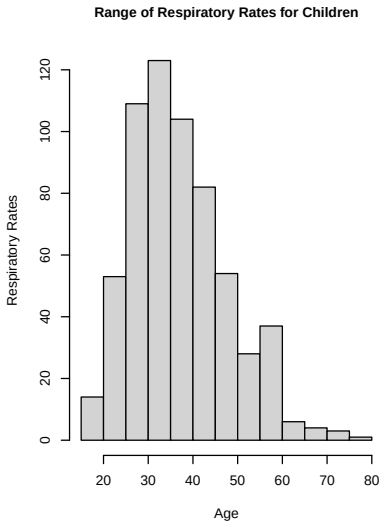
After simple exploratory data analysis, we can tell that this dataset contains 618 rows and 3 columns(variables). Data type for three variables are X(int), Age(num), and Rate(int). Tables below shows some statistic feature of variable **Rate**.

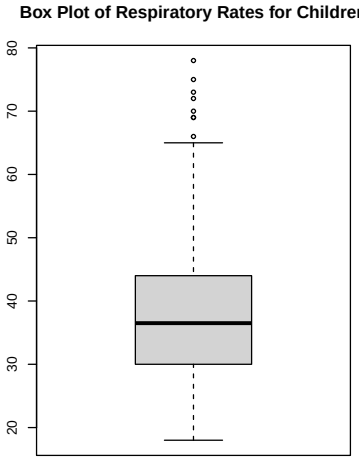
Min	first_Qu	Median	Mean	third_Qu	Max
18	30	36.5	37.74	44	78

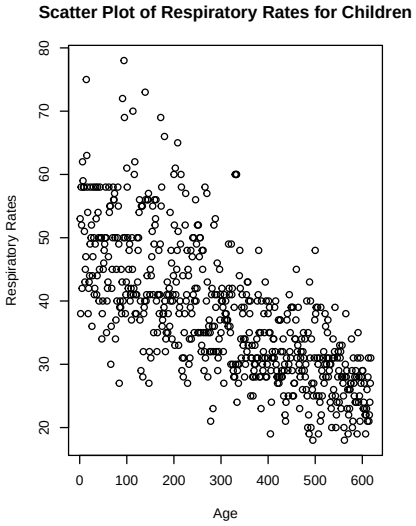
Rate	n	percent
18	2	0.003236246
19	6	0.009708738
20	6	0.009708738
21	7	0.011326861
22	8	0.012944984
23	10	0.016181230
24	12	0.019417476
25	16	0.025889968
26	8	0.012944984
27	19	0.030744337
28	34	0.055016181
29	25	0.040453074
30	23	0.037216828
31	39	0.063106796
32	26	0.042071197
33	10	0.016181230

34	21	0.033980583
35	27	0.043689320
36	10	0.016181230
37	17	0.027508091
38	25	0.040453074
39	21	0.033980583
40	31	0.050161812
41	30	0.048543689
42	16	0.025889968
43	13	0.021035599
44	9	0.014563107
45	14	0.022653722
46	3	0.004854369
47	6	0.009708738
48	13	0.021035599
49	11	0.017799353
50	21	0.033980583
51	3	0.004854369
52	8	0.012944984
53	4	0.006472492
54	5	0.008090615
55	8	0.012944984
56	9	0.014563107
57	4	0.006472492
58	17	0.027508091
59	1	0.001618123
60	6	0.009708738
61	2	0.003236246
62	2	0.003236246
63	1	0.001618123
65	1	0.001618123
66	1	0.001618123
69	2	0.003236246
70	1	0.001618123
72	1	0.001618123
73	1	0.001618123

75	1	0.001618123
78	1	0.001618123







	Quantile
0%	18.0
25%	30.0
50%	36.5
75%	44.0
100%	78.0

Statistically speaking, data inside IQR range could be the normal range of respiratory rates for children of any age between 0 and 3. That is, any number below 44 could be considered normal.

b

regression model for predicting respiratory rates from age

Rate $=\beta 0 + \beta 1$**Age** $+ \epsilon$

c

Fitted Model

$$\hat{\text{Rate}} = 47.05 - 0.70(\text{Age})$$

d

- The results of the linear model indicate that age is a significant predictor of children respiratory rates at the $\alpha = 0.05$ significance level ($p < 0.001$)

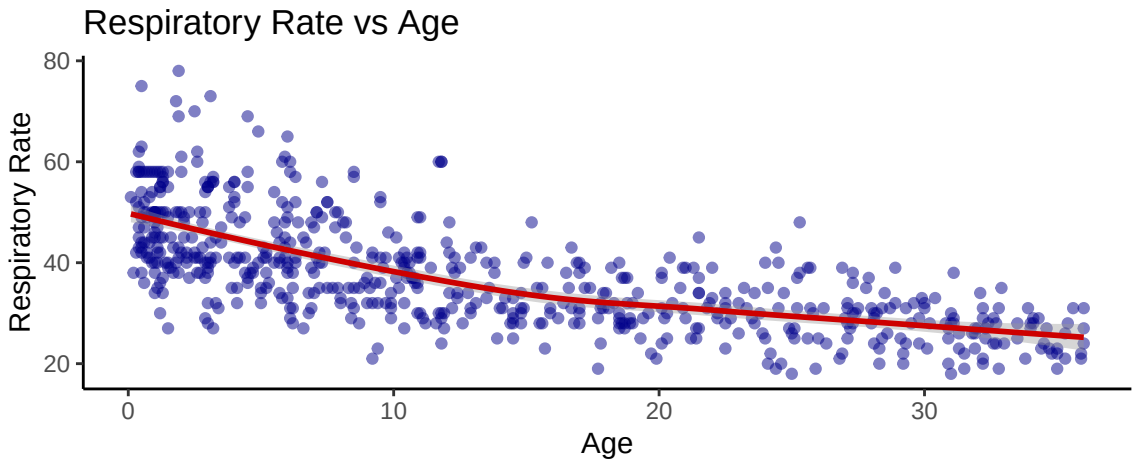
Table 3: SLR Model Regressing Respiratory on Age

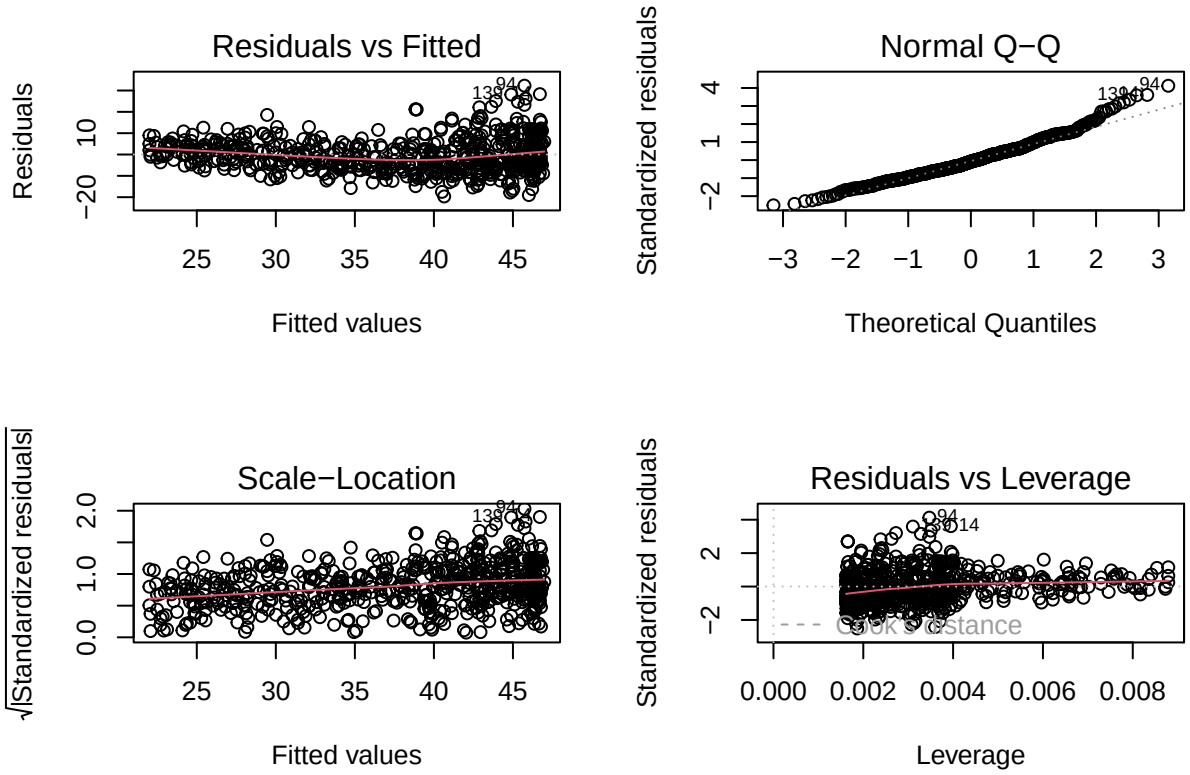
	Estimate	SE	t	p-value
Intercept	47.0522	0.5042	93.3173	<.001
Age	-0.6957	0.0294	-23.6838	<.001

- if a child is one year older, his/her respiratory rate is estimated to be 0.70% lower on average. When a children is 0 year old, his/her respiratory rate is 47.05 (since a children can't be 0 year old, the intercept has no practical value and is not meaningful). Since the p-value for both the estimated intercept and slope <2e-16, the result is quite significant.
- A 95% confidence interval for intercept is (46.06, 48.04), which is a set of possible values of the estimated intercept that we are 95% confident contains the true value of the intercept. A 95% confidence interval for age is (-0.75, -0.64), which is a set of possible values of the estimated slope that we are 95% confident contains the true value of the slope.
- The model fit produces an R^2 value of 0.48,meaning using children age to estimate their respiratory rate reduced the uncertainty in the estimate by explaining approximately 48% of the variability in the response.(48% of the variation in children respiratory rate is explained by age.)

e

Simple linear regression model requires four assumptions: **linearity, nearly normal residuals, constant variability, and independent observations.**





There is a linear relationship between these two variables, but there are several data points scatter around low age with higher respiratory rate.The Q-Q plot is a little bit skewed, but clustering of the points is still around the 45 degree, so normality assumption is not violated. In the residuals vs fitted values plot, the red line is roughly a flat line at 0. So there is no obvious pattern and nearly normal residuals assumption is not violated. The variance is also roughly constant. These are not time series observation, and can be viewed as independent observations. Therefore, the model assumptions are reasonable.

Question 2

a

EDA

After simple exploratory data analysis, we can tell that this dataset contains 305 rows (observations) and 8 columns(variables). Data type for eight variables are id(int), host_is_superhost(chr), host_identity_verified(chr), room_type(chr), accommodates(int), bathrooms(num), bedrooms(int),and price(int).

table for host$host_is_superhost

Flase	True
160	145

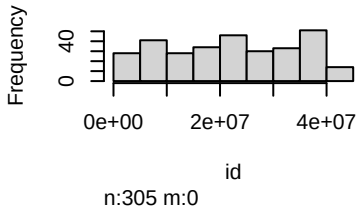
table for host$host_identity_verifies

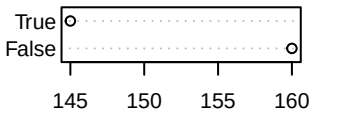
Flase	True
142	163

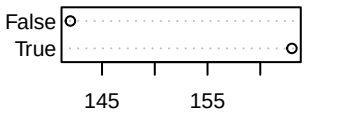
table for host$room_type

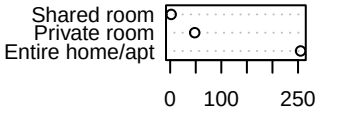
Entire_home_apt	Private_room	Shared_room
225	48	2

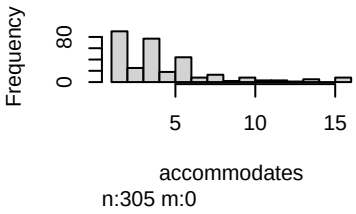
Here, I present some exploratory plots of each variables.

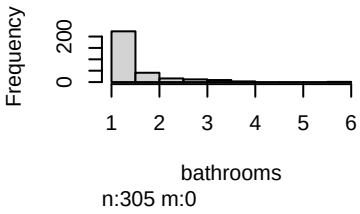


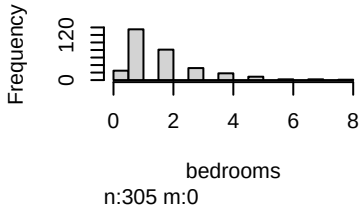












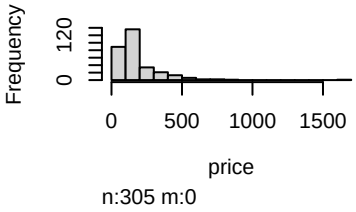
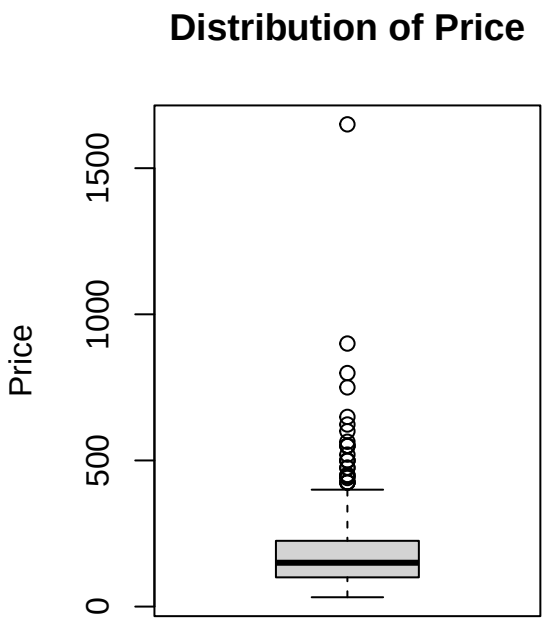


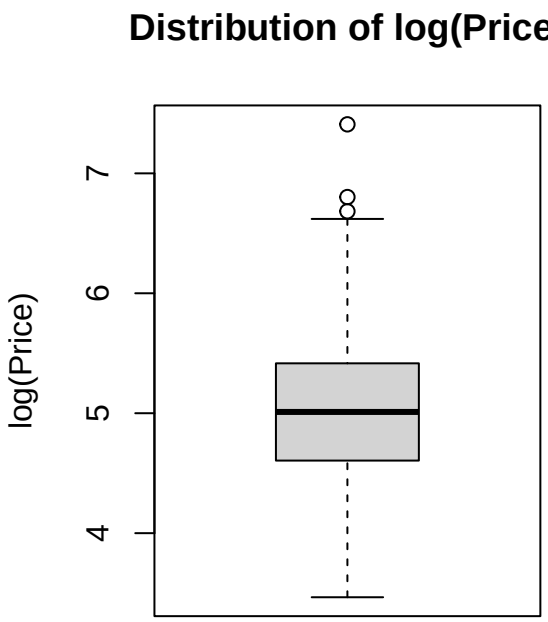
table for host$Price

Min	first_Qu	Median	Mean	third_Qu	Max
32	100	150	194.3	225	1,650

Distribution plot for host*Priceandlog*(*host***Price)**

We can tell that the distribution of the response variable **Price** is not normal. Therefore, we may need to do transformation on the response variable later. By transforming price to log(Price), the distribution of response variable is more normal than before.





model fitting

First of all, I tried to include all the variables, including host_is_superhost, host_identity_verified, room_type, accommodates, bathrooms, and bedroom in my linear regression model, but this one doesn't meet the assumptions for multiple linear regression, like the linearity between each predictor and the response variable, multicollinearity between accommodates and bedroom, etc. Also, results for host_is_superhost, host_identity_verified, room_type, and accommodates are not significant enough.

Therefore, after trying to add different predictors into the linear regression model and doing natural log transformation, my final regression model is: $\mathbf{log(price)} = \beta 0 + \beta 1\mathbf{bathrooms} + \beta 2\mathbf{bedrooms} + \epsilon$

model assessment

MLR assumptions are as follows:

1.Linear relationship between reach x and y

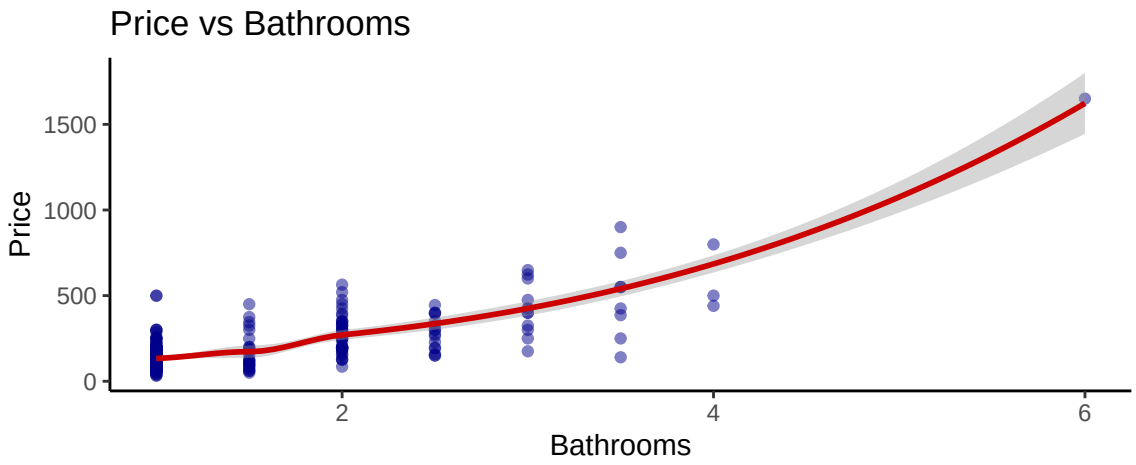
2.Independence of errors

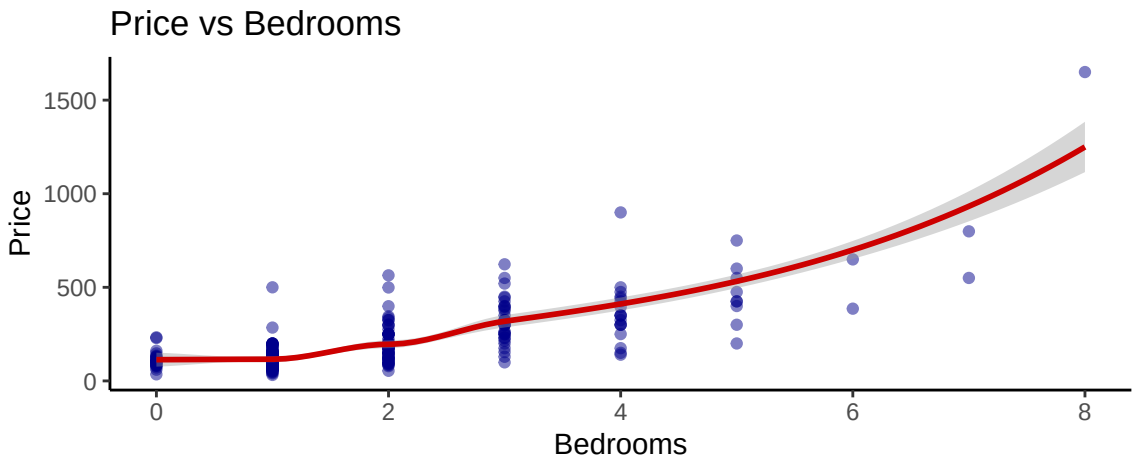
3.Equal variance of errors

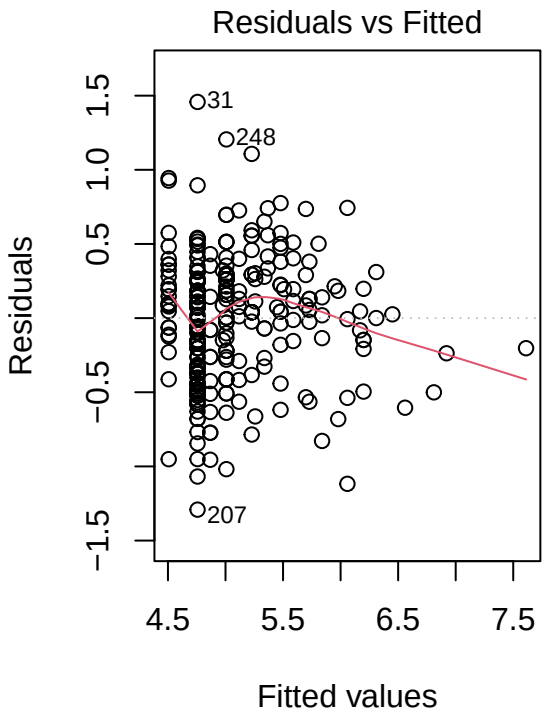
4.Normality of errors

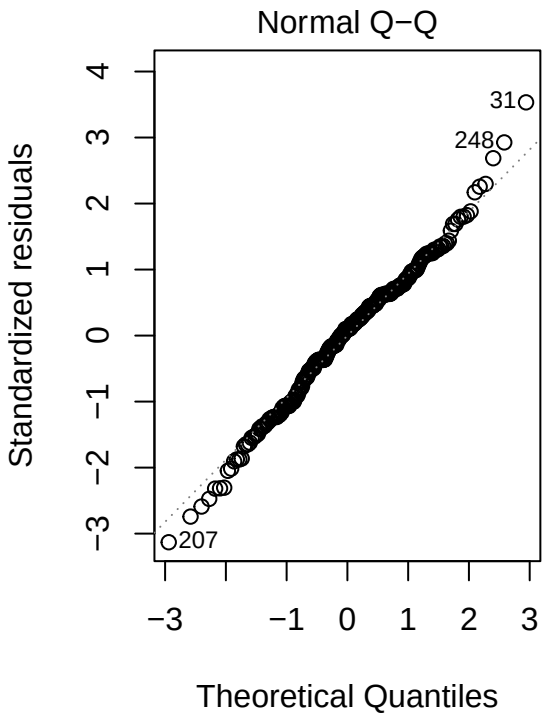
5.No multicollinearity

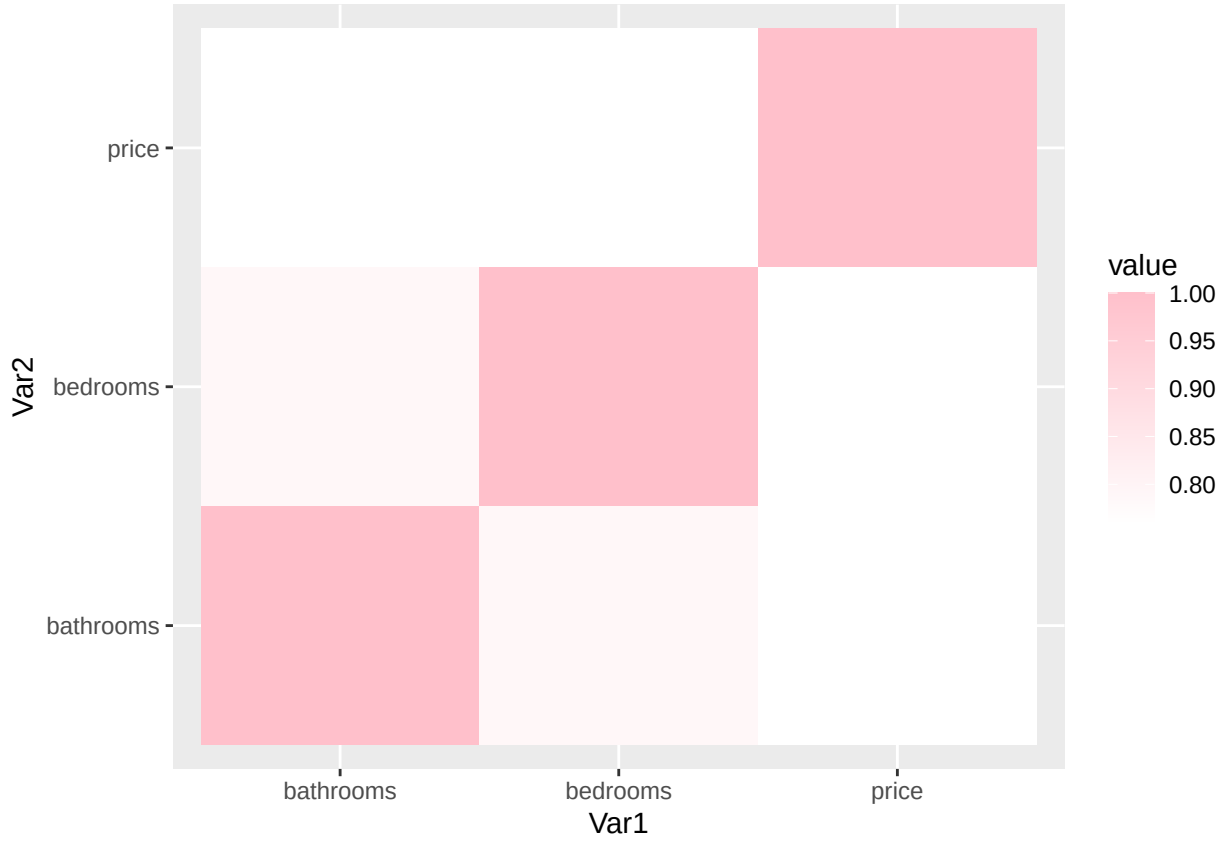
There is a linear relationship between these each predictor and the response variable. In the residuals vs fitted values plot, though the red line is not strictly a flat line at 0, there is no obvious pattern. Also, there is independence of errors and roughly equal variance of errors. The Q-Q plot is a little bit skewed, but clustering of the points is still around the 45 degree, so normality assumption is not violated. Since there is no value higher than 0.8 in the heatmap, the model meets the no multicollinearity assumption.


Price vs Bathrooms
Price
1500
1000
500
0
2
4
6
Bathrooms


Price vs Bedrooms
Price
1500
1000
500
0
0
2
4
6
8
Bedrooms


Residuals vs Fitted
Residuals
Fitted values
31
248
207
−1.5
−0.5
0.5
1.0
1.5
4.5
5.5
6.5
7.5


Normal Q−Q
Standardized residuals
Theoretical Quantiles
31
248
207
−3
−2
−1
0
1
2
3
4



b

Fitted Model

$$\hat{\log \text{Price}} = 4.29 + 0.22(\text{bathrooms}) + 0.25(\text{bedrooms})$$

output	Intercept	bathrooms	bedrooms
estimated coefficients	4.29	0.22	0.25
SE	0.05	0.05	0.03
p-values	<2e-16	2.21e-05	1.35e-15
confidece intervals	(4.19,4.39)	(0.12,0.32)	(0.19,0.31)

c

1. The estimated coefficient of bathrooms is 0.22. For each additional bathroom in the listing, we expect log(price) to increase by about 0.22 (positive difference of approximately 12.46% in price), holding all other variables constant.
2. The estimated coefficient of bedrooms is 0.25. For each additional bedroom in the listing, we expect log(price) to increase by about 0.25 (positive difference of approximately 12.84% in price), holding all other variables constant.
3. The estimated intercept is 4.29. When the number of bedrooms and the number of bathrooms are both 0 in the listing, we expect log(price) to be 4.29(which is not meaningful enough)

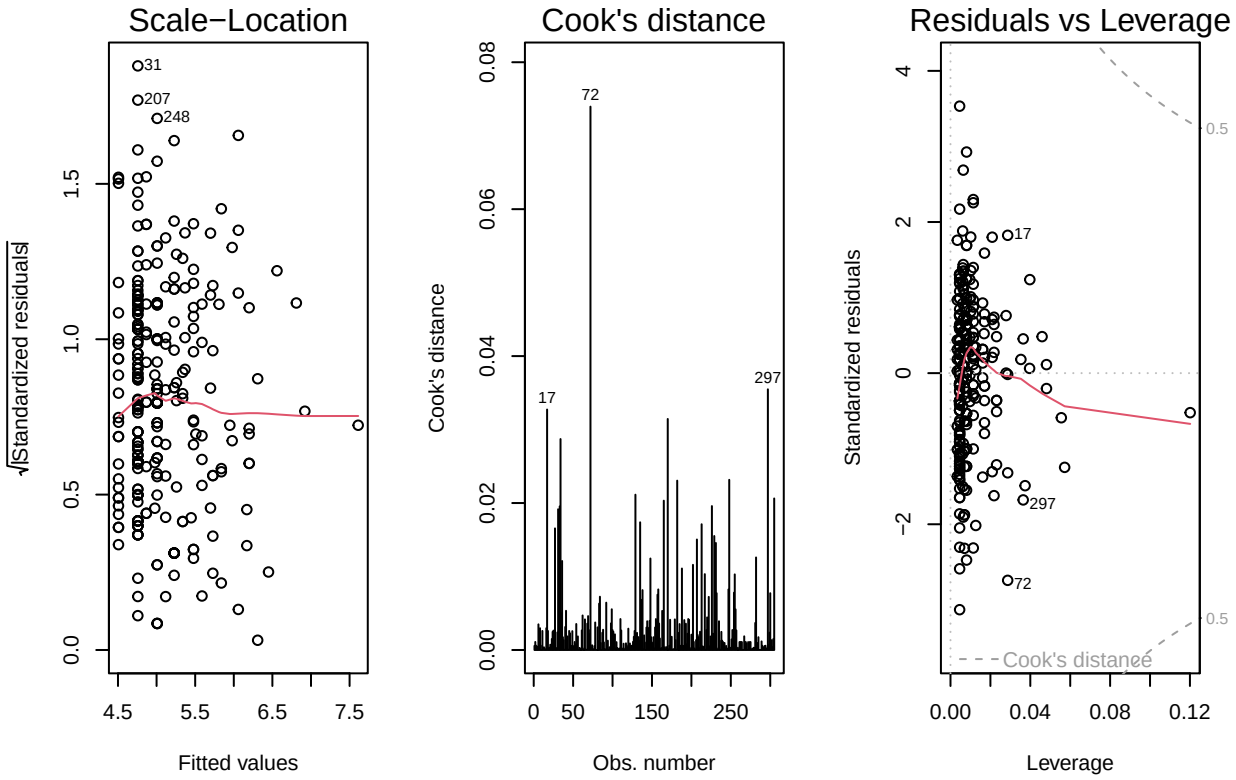
Table 8: Results

	Dependent variable:
	log(price)
bathrooms	0.22***
	(0.12,0.32)
bedrooms	0.25***
	(0.19,0.31)
Constant	4.29***
	(4.19,4.39)
Observations	305
R^2	0.57
Adjusted R^2	0.57
Residual Std. Error	0.41 (df = 302)
F Statistic	199.08*** (df = 2; 302)
Note:	*p<0.1; **p<0.05; ***p<0.01

4. Adjusted R-squared is 0.57, indicating that this model reduced the uncertainty in the estimate by explaining approximately 57% of the variability in the response.(57% of the variation in price is explained by this model.)

d

Looking at the right plot below, we can tell that there is two outliers (observation 31 and 207) that falls outside the (-3,3) standardized residuals. However, their cook's distance are lower than 0.5, which indicate that they are possible not necessarily influential on the regression line.



Exclude the two points and fit the model.

Since those two points' Cook's distance is lower than 0.5, they don't have much influence on the regression model. Coefficients and SEs, and confidence intervals of the model almost remain the same as before. But after excluding the outliers, the t-value for each coefficient increases (p-value decreases), indicating an increase in the significance of predictors. Also, the adjusted R-squared increases from 0.57 to 0.58, indicating the increase in explanatory power of the model.

output	Intercept	bathrooms	bedrooms
estimated coefficients	4.29	0.22	0.25
SE	0.05	0.05	0.03
p-values	<2e-16	1.12e-05	<2e-16
confidece intervals	(4.19,4.38)	(0.12,0.32)	(0.19,0.31)

Table 11: Results for New Model without Outliers

	Dependent variable:
	log(price)
bathrooms	0.22***
	(0.12,0.32)
bedrooms	0.25***
	(0.19,0.31)
Constant	4.29***
	(4.19,4.38)
Observations	303
R^2	0.59
Adjusted R^2	0.58
Residual Std. Error	0.40 (df = 300)
F Statistic	213.20*** (df = 2; 300)
Note:	*p<0.1; **p<0.05; ***p<0.01

e

1. The Adjusted R-squared is 0.57 for my original regression model and 0.58 for my model without outliers, which means the model can only explain 57%/58% of the variance of the response variable(log(price)). The model needs to be improved.
2. The intercept now is not meaningful or realistic, maybe we can do something like centering(mean-center continuous predictors) to improve interpretation of the intercept.

Appendix: All code for this report

```
knitr::opts_chunk$set(echo = TRUE)
library(ISLR2)
library(knitr)
library(ggplot2)
library(kableExtra)
library(lattice)
library(dplyr)
library(stargazer)
library(gt)
library(janitor)
library(flextable)
library(magrittr)
library(Hmisc)
re <- read.csv("Respiratory.csv")
#dim(re)
#head(re)
#str(re)
#summary(re$Rate)
df<-data.frame(Min=18.00,first_Qu=30.00,Median=36.50,Mean=37.74,third_Qu=44.00,Max=78.00)
df %>% regulartable%>% autofit()
re %>% janitor::tabyl(Rate) %>%
  gt::gt()
par(mfrow = c(1,3))
hist(re$Rate, main = "Range of Respiratory Rates for Children",xlab ="Age", ylab ="Respiratory Rates",c
boxplot(re$Rate, main = "Box Plot of Respiratory Rates for Children")
plot(re$Rate, main = "Scatter Plot of Respiratory Rates for Children",xlab ="Age", ylab ="Respiratory Ra
kable(quantile(re$Rate), col.names = "Quantile")
lm_re <- lm(Rate ~ Age, data = re)
#summary(lm_re)
modmat <- matrix(c(round(summary(lm_re)$coefficients[,1:3],4),"<.001","<.001"),nrow=2,ncol=4)
rownames(modmat) <- c("Intercept","Age")
kable(modmat,row.names=TRUE,col.names=c("Estimate","SE","t","p-value"),format="latex",booktabs=T,captio
#confint(lm_re, level = 0.95)
#par(mfrow = c(1,3))
#plot(re$Age, re$Rate)
#plot(re$Age, lm_re$residuals)
#hist(lm_re$residuals)
p1 <-ggplot(re,aes(x=Age, y=Rate)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(col="red3") + theme_classic() +
  labs(title="Respiratory Rate vs Age",x="Age",y="Respiratory Rate")
p1
```

```
par(mfrow = c(2,2))
plot(lm_re)
host <- read.table('Listings_QueenAnne.txt', head = TRUE)
#head(host)
#str(host)
#summary(host[,2:8])
#table(host$host_is_superhost)
df<-data.frame(Flase=160,True=145)
df %>% regulartable%>% autofit()
#table(host$host_identity_verified)
df<-data.frame(Flase=142,True=163)
df %>% regulartable%>% autofit()
#table(host$room_type)
df<-data.frame(Entire_home_apt = 225,Private_room=48, Shared_room = 2)
df %>% regulartable%>% autofit()
hist.data.frame(host)
df<-data.frame(Min=32.00,first_Qu=100.00,Median=150.00,Mean=194.3,third_Qu=225.0,Max=1650.0)
df %>% regulartable%>% autofit()
par(mfrow = c(1,2))
boxplot(host$price,ylab = "Price",main ="Distribution of Price")
boxplot(log(host$price),ylab = "log(Price)", main ="Distribution of log(Price)")
par(mfrow = c(1,2))
p1 <-ggplot(host,aes(x=bathrooms, y=price)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(col="red3") + theme_classic() +
  labs(title="Price vs Bathrooms",x="Bathrooms",y="Price")

p2 <- ggplot(host,aes(x=bedrooms, y=price)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(col="red3") + theme_classic() +
  labs(title="Price vs Bedrooms",x="Bedrooms",y="Price")
p1
p2
#model0 <- lm(price ~ host_is_superhost + host_identity_verified + room_type + accommodates + bathrooms
#summary(model0)

#model1 <- lm(price~room_type + accommodates + bathrooms + bedrooms, data = host)
#summary(model1)

#model2 <- lm(price~accommodates + bathrooms + bedrooms, data = host)
#summary(model2)

#model2.1 <- lm(log(price)~accommodates + bathrooms + bedrooms, data = host)
#summary(model2.1)

#model3 <- lm(price~ bathrooms + bedrooms, data = host)
#summary(model3)
model3.1 <- lm(log(price)~ bathrooms + bedrooms, data = host)
par(mfrow = c(1,2))
plot(model3.1,1)
plot(model3.1,2)
#plot for multicollinearity
library(reshape2)
```

```r
# creating correlation matrix
corr_mat <- round(cor(as.matrix(host[,6:8])),2)

# reduce the size of correlation matrix
melted_corr_mat <- melt(corr_mat)
# head(melted_corr_mat)

# plotting the correlation heatmap
library(ggplot2)
ggplot(data = melted_corr_mat, aes(x=Var1, y=Var2,
                                   fill=value)) +
geom_tile()+
scale_fill_gradient(low="white", high="pink")

stargazer(model3.1,title="Results", align=TRUE,header=FALSE,ci = TRUE,digits = 2)
#summary(model3.1)
#confint(model3.1, level = 0.95)
output <- c('estimated coefficients', 'SE','p-values','confidece intervals' )
Intercept <- c('4.29', '0.05', '<2e-16', '(4.19,4.39)')
bathrooms <- c('0.22','0.05','2.21e-05','(0.12,0.32)')
bedrooms <- c('0.25','0.03','1.35e-15','(0.19,0.31)')
df <- data.frame(output, Intercept, bathrooms, bedrooms)
df %>% regulartable() %>% autofit()
par(mfrow = c(1,3))
plot(model3.1, 3)
plot(model3.1, 4)
plot(model3.1,5)
host_new <- host[-c(31,207),]
model3.1_new <- lm(log(price)~ bathrooms + bedrooms, data = host_new)
#summary(model3.1_new)
#confint(model3.1_new)
output <- c('estimated coefficients', 'SE','p-values','confidece intervals' )
Intercept <- c('4.29', '0.05', '<2e-16', '(4.19,4.38)')
bathrooms <- c('0.22','0.05','1.12e-05','(0.12,0.32)')
bedrooms <- c('0.25','0.03','<2e-16','(0.19,0.31)')
df <- data.frame(output, Intercept, bathrooms, bedrooms)
df %>% regulartable() %>% autofit()
stargazer(model3.1_new,title="Results for New Model without Outliers", align=TRUE,header=FALSE,ci = TRU
```