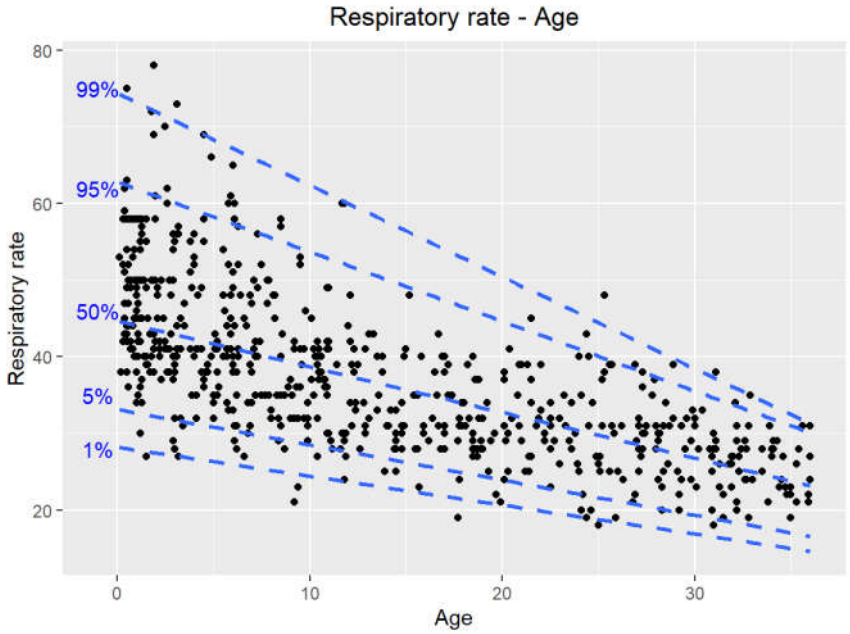
Data Analysis Assignment 1

Yuanjing Zhu

Respiratory rates for children

a. Do exploratory analysis on the data and include a useful plot that a physician could use to assess a "normal" range of respiratory rates for children of any age between 0 and 3.



b. Write down a regression model for predicting respiratory rates from age. Make sure to use the right mathematical notation.

$$Respiratory\ rate = \beta_0 + \beta_1 * Age + \epsilon_i;\ \epsilon_i \sim N(0, \sigma^2),\ i = 1, \ldots, n$$

c. Fit the model to the data. Include a table showing the output from the regression model including the estimated intercept, slope, residual standard error, and proportion of variation explained by the model.

SLR Model Regressing respiratory rate on age

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

[a] Multiple R-squared: 0.4766
[b] Adjusted R-squared: 0.4758

95% Confidence Interval

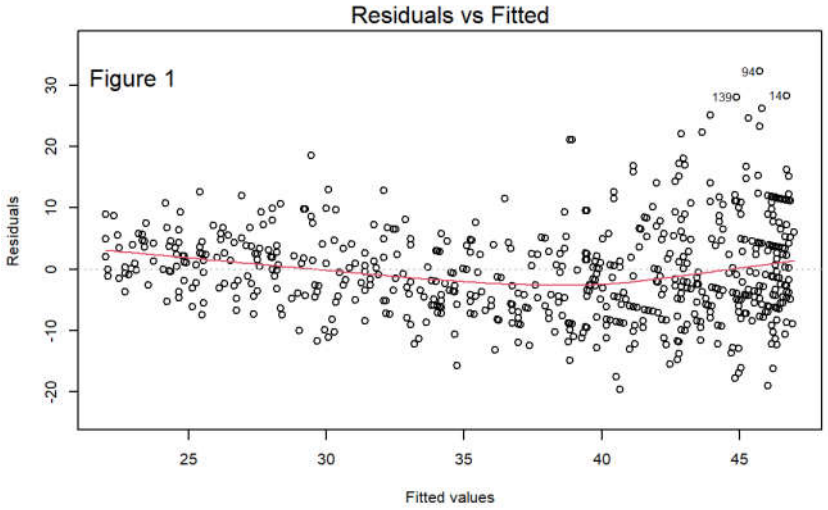
Predictor	2.5%	97.5%
Intercept	46.0620	48.0424
Age	-0.7534	-0.6380

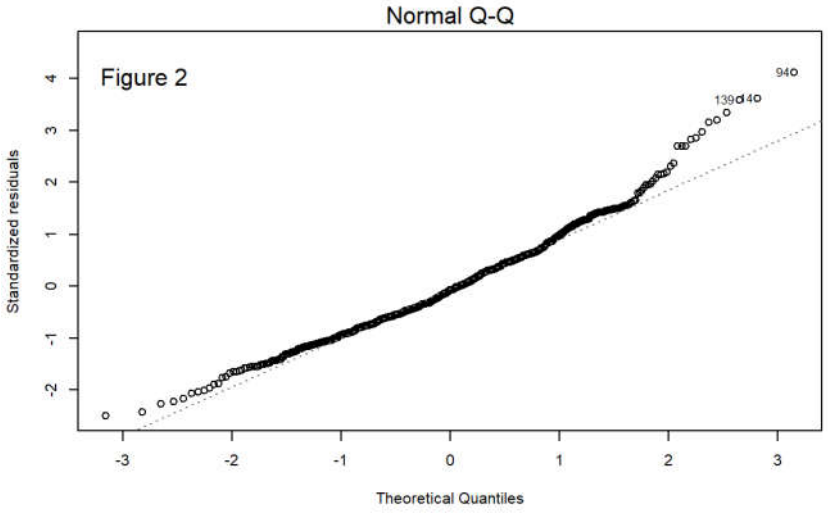
d. Interpret your results. In the context of the problem, what do you conclude? Your interpretation should mention an appropriate p-value, 95% confidence interval, and R2 value.

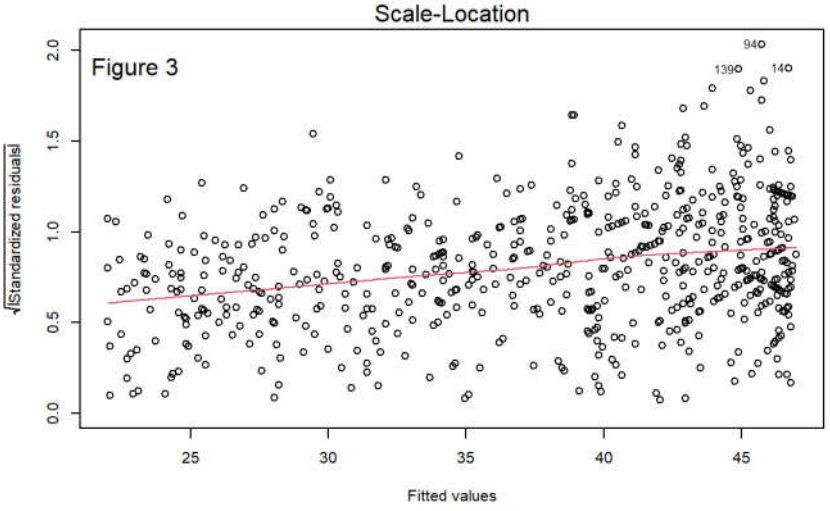
1. The p-value is pretty small(< 0.05), indicating that age is a significant predictor of respiratory rate at the $\alpha = 0.05$ significance level.
2. The respiratory rate is expected to decrease by 0.6957 when the value of age increases by one unit.

3. The confidence interval of slope is [-0.75, -0.64] while that of intercept is [46.06, 48.04]. It means that if we repeatedly draw random samples of the same size to fit the same model, about 95% of the time the confidence interval will capture the true value of these two coefficients.
4. R^2 value of this model is 0.4766, which means that 47.66% of the variance for respiratory rate can explained by the age. So our model doesn't fit the data very well.

e. Is there enough evidence that the model assumptions are reasonable for this data? Include appropriate plots in your answer.







1. The linearity assumption can be verified by Figure 1. It might be a potential quadratic trend in the dataset, so we need to further check the assumption.

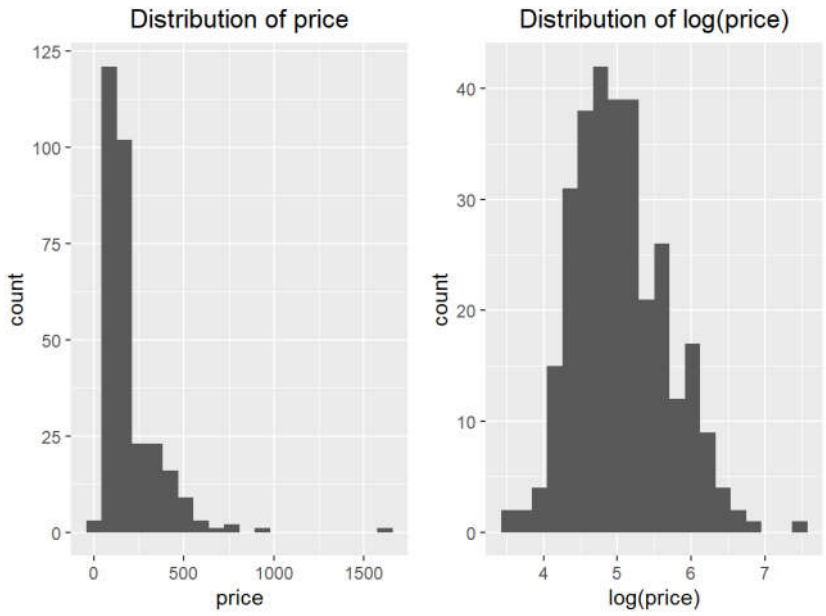
2. The normality assumption can be checked in Figure 2. Most of the points fall near the line of identity despite the fact that deviations occur at the higher end of the line. Overall, there is no any type like an S form, an exponential curve, so the normality assumption is satisfied and the residuals follow a normal distribution.
3. The constant variance assumption can be checked in Figure3. If the constant variance assumption is met, the spread of the points should be constant across the whole window and the LOESS curve should be a flat line. However, there seems to be more larger fitted values at the right end and the LOESS curve is tilted. So we need to check for potential violation further.
4. The independence assumption can also be checked in Figure 1. Because there is no observable pattern in the plot, the independence assumption seems plausible for this dataset.

Airbnb listing for Seattle, WA

a. Analyze the data by doing EDA, then model fitting, and model assessment. Consider transformations if needed.

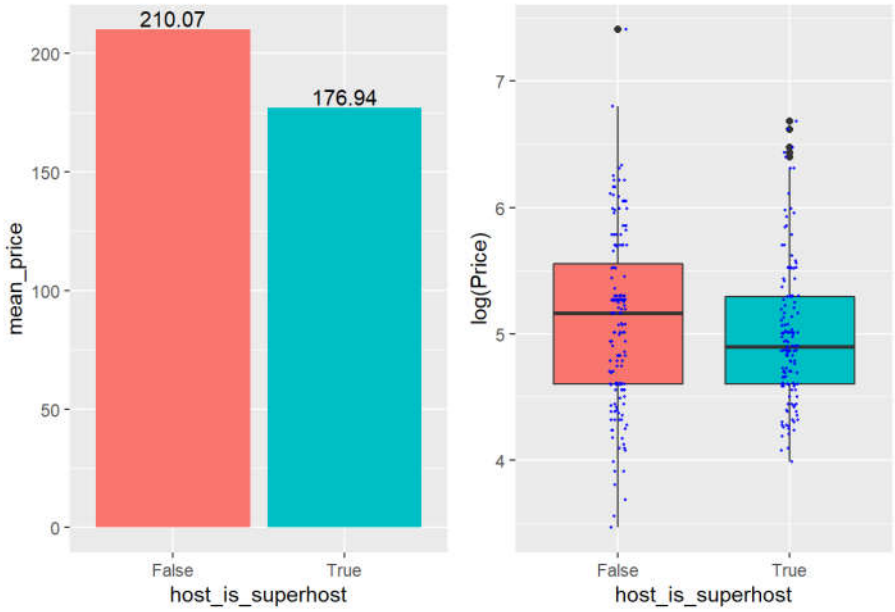
EDA

1. Distribution of price



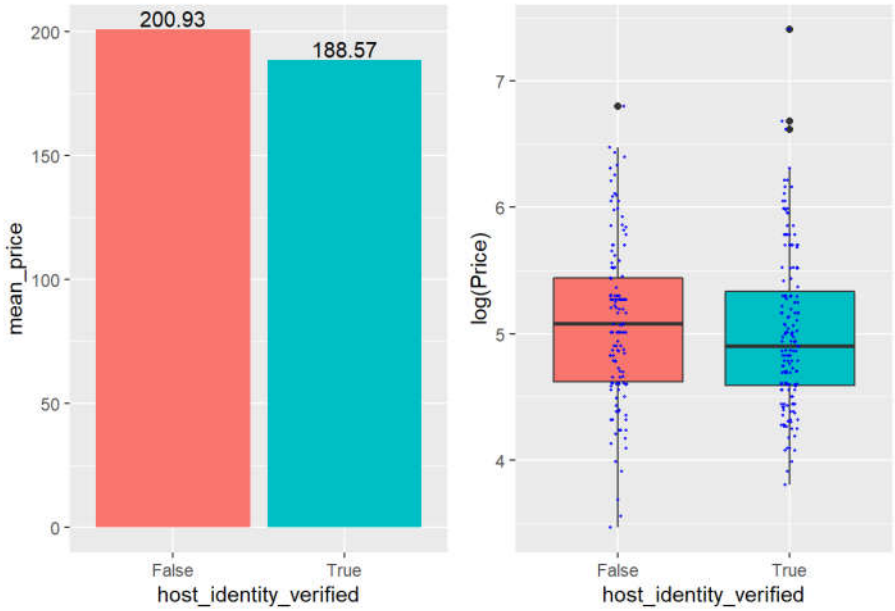
From the left histogram, we can see that the distribution of price is right-skewed. Most of the listing price is less than $500 while only a few are more than $500. So we consider using log transformation on price. The distribution of log(price) is shown in the right figure. It follows the bell curve.

2. log(price) – host_is_superhost



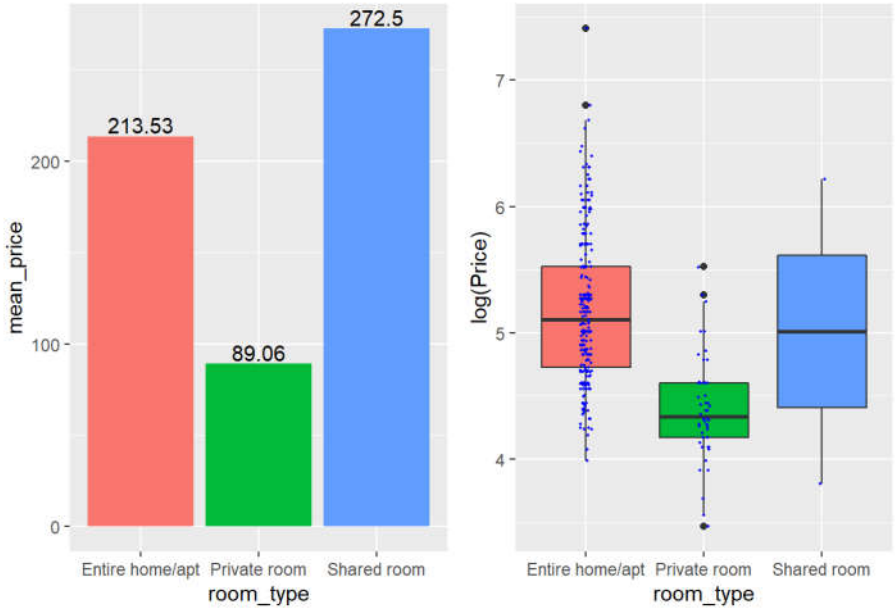
Average listing price is higher when host is not superhost. There are some outliers when the listing price is super high.

3. log(price) – host_identity_verified



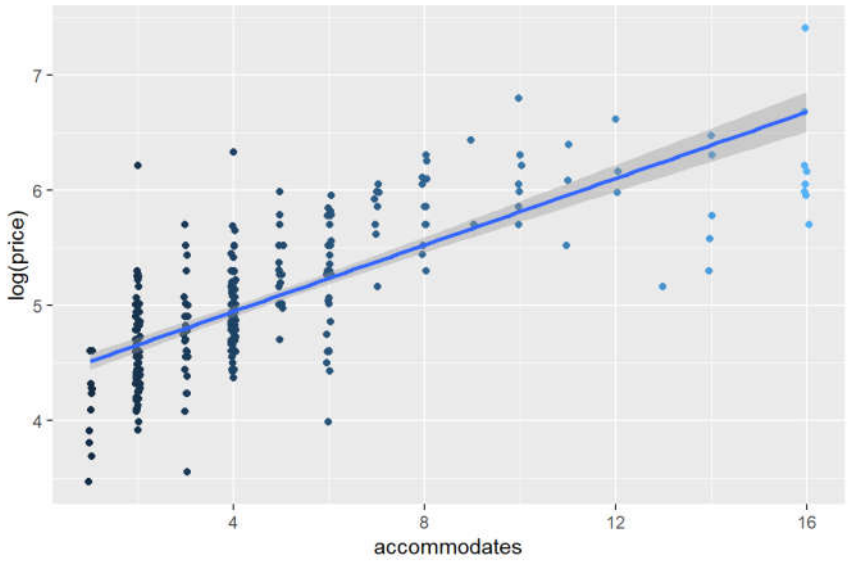
Listing price tends to be higher when the host hasn't verified their identity with Airbnb.

4. log(price) – room_type



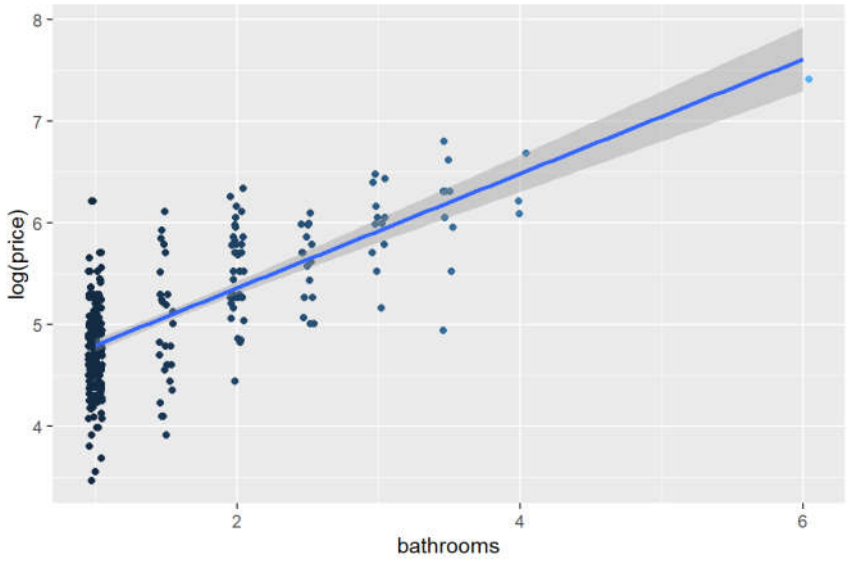
Most of the rooms are entire home or apartment while some are private room. Only two are shared room. In general, the listing price is higher when the room type is entire home/apt.

5. log(price) – accommodates



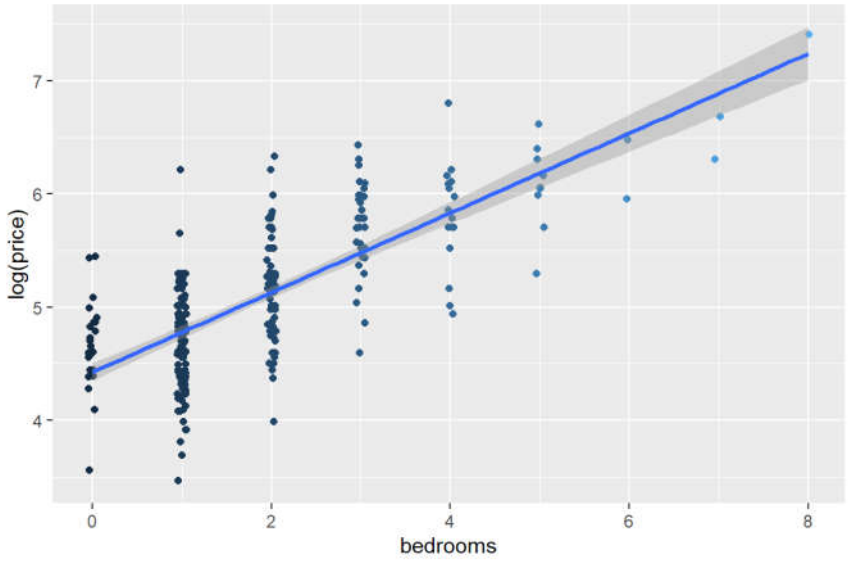
The more people the room can accommodates, the higher the listing price in spite of some data points when the room can accommodate more than 12 people.
Concern: Some high-leverage points might affect model fitting.

6. log(price) – bathrooms



Most rooms have only one bathroom. The more bathroom a room has, the higher the listing price.
Concern: Some high-leverage points might affect model fitting.

7. log(price) – bedrooms



Rooms with more bedrooms have higher listing price.
Concern: Some high-leverage points might affect model fitting.

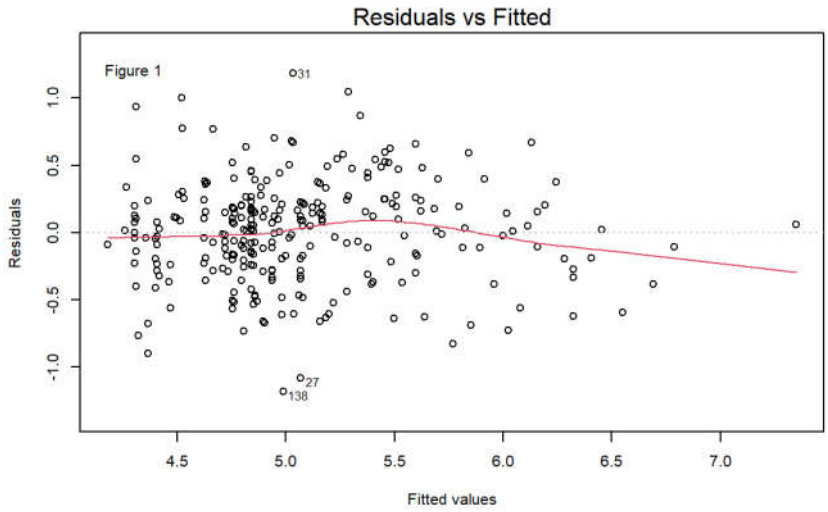
Model fitting

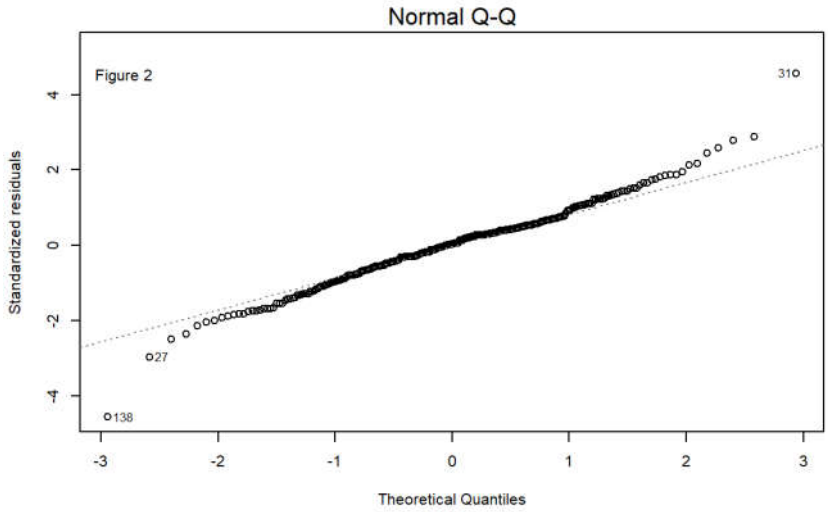
MLR:

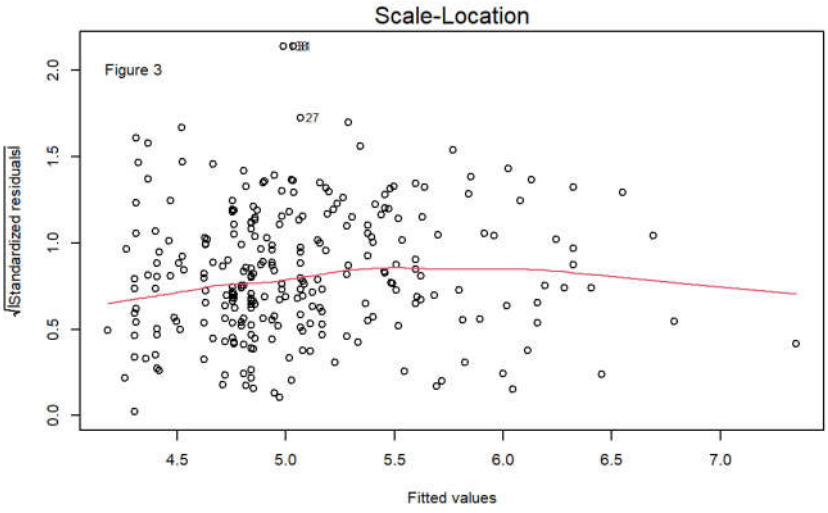
$$\begin{aligned} log(price) = \beta_0 &+ \beta_1 * host\_is\_superhost \\ &+ \beta_2 * host\_identity\_verified \\ &+ \beta_3 * room\_type \\ &+ \beta_4 * accommodates \\ &+ \beta_5 * bathrooms \\ &+ \beta_6 * bedrooms \\ &+ \epsilon_i;\ \epsilon_i \sim N(0, \sigma^2),\ i = 1, \ldots, n \end{aligned}$$

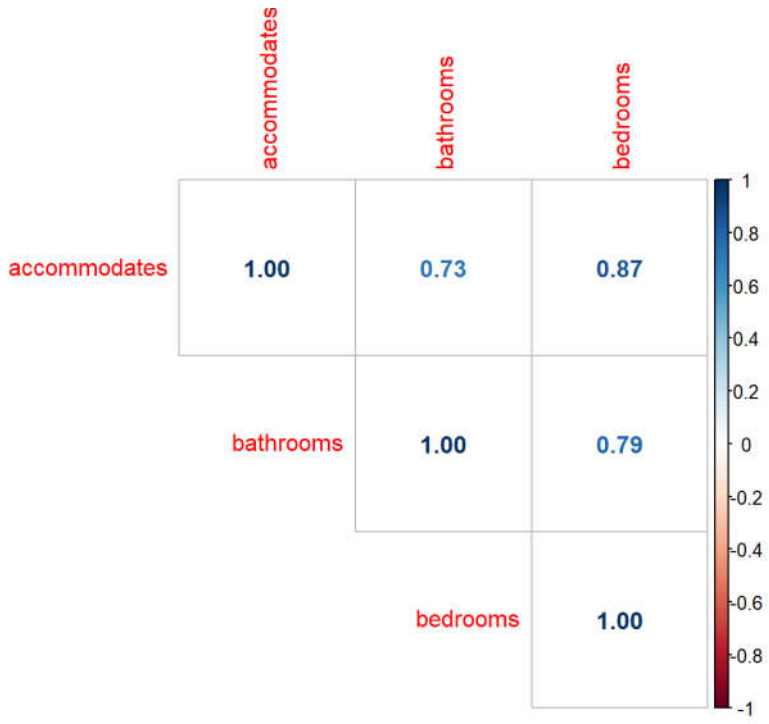
where host_is_superhost, host_identity_verified are dummy variables and accommodates, bathrooms and bedrooms are numeric variables.

Model assessment


Residuals vs Fitted
Figure 1
31
27
138
Residuals
1.0
0.5
0.0
-0.5
-1.0
4.5
5.0
5.5
6.0
6.5
7.0
Fitted values


Normal Q-Q
Figure 2
310
27
138
Standardized residuals
4
2
0
-2
-4
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Scale-Location
Figure 3
27
√|Standardized residuals|
2.0
1.5
1.0
0.5
0.0
4.5
5.0
5.5
6.0
6.5
7.0
Fitted values



VIF values for each predictor

	VIF
host_is_superhost	1.0490
host_identity_verified	1.0205
room_type	1.2100
accommodates	4.7145
bathrooms	2.8021
bedrooms	5.3795

1. **Linearity**: Figure 1 is the plot of residuals vs fitted values. There is no discernable pattern so the linearity assumption is satisfied.
2. **Independence of errors**: The scatter points in Figure 1 seem random, so the error terms are independent.
3. **Equal variance of errors**: The LOESS curve in Figure 1 is primarily a flat line around zero, so heteroscedasticity assumption is met.
4. **Normality of errors**: In Figure 2, most points are clustering around the 45° line, which implies normality assumption is not violated.
5. **No multicollinearity**: From the last figure, we can see that the numeric variables: accommodates, bathrooms and bedrooms are correlated. That makes sense because the more bathrooms and bedrooms an apartment has, more people it can accommodate. Also the VIF of accommodates, bathrooms and bedrooms are relatively high. This could be problematic and need further inspection.

b. Include the output from the final regression model that you used, as well as evidence that the model fits the assumptions reasonably well. Your regression output should includes a table with coefficients and SEs, p-values, and confidence intervals.

MLR Model Regressing listing price

Predictor	Estimate	SE	t	p-value
(Intercept)	4.4299	0.0613	72.2604	0.0000
host_is_superhostTrue	-0.0087	0.0429	-0.2036	0.8388

[a] Multiple R-squared: 0.6682
[b] Adjusted R-squared: 0.6604

Predictor	Estimate	SE	t	p-value
host_identity_verifiedTrue	-0.0968	0.0424	-2.2832	0.0231
room_typePrivate room	-0.4502	0.0626	-7.1957	0.0000
room_typeShared room	0.2698	0.2632	1.0251	0.3061
accommodates	0.0441	0.0141	3.1223	0.0020
bathrooms	0.2111	0.0460	4.5861	0.0000
bedrooms	0.1305	0.0371	3.5214	0.0005

[a] Multiple R-squared: 0.6682
[b] Adjusted R-squared: 0.6604

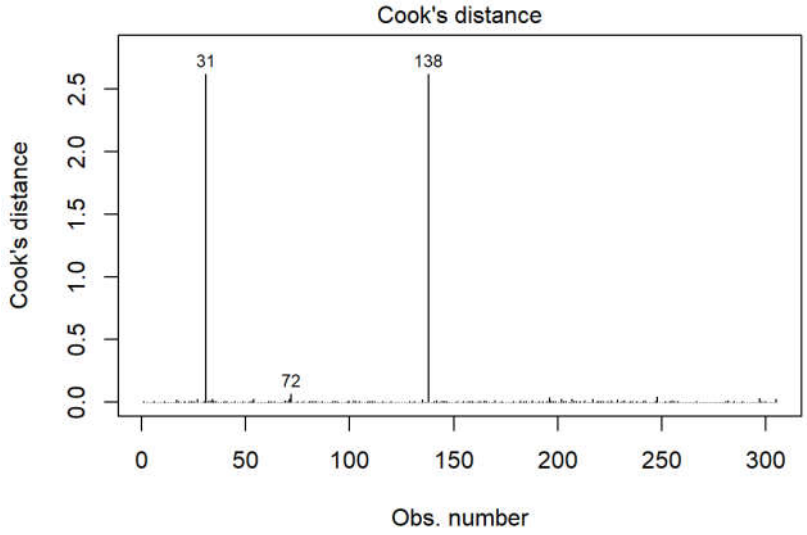
95% Confidence Interval

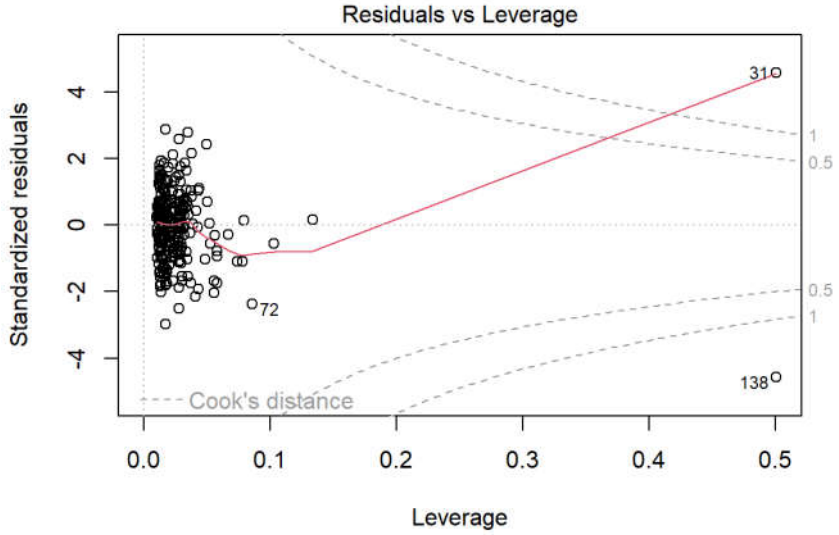
	2.5 %	97.5 %
(Intercept)	4.3092	4.5505
host_is_superhostTrue	-0.0932	0.0758
host_identity_verifiedTrue	-0.1803	-0.0134
room_typePrivate room	-0.5734	-0.3271
room_typeShared room	-0.2482	0.7878
accommodates	0.0163	0.0719
bathrooms	0.1205	0.3017
bedrooms	0.0576	0.2034

c. Interpret the results of your fitted model in the context of the data.

1. For numeric variables, coefficient of each predictor represents the difference in log(Price) for each on-unit difference in the predictor when other predictors remain constant. For example, for bathrooms, $\beta = 0.2111$, $e^{\beta} = 1.235$, so a difference of one unit in bathroom will lead to about 23.5% increase in listing price.
2. For categorical variables, coefficient is the average difference in log(Price) between category. For example, the average difference in price is 9% between host_is_superhostTrue and host_is_superhostFalse.
3. p-value of each predictor shows that room_typePrivate room, accommodates, bathrooms and bedrooms are significant.
4. Adjucted R^2 is 0.6604, which means that about 66.04% of variation in the response variable can be explained by variation in the predictors.
5. Overall p-value based on F-statistic is less than 0.05, indicating the model is significant.

d. Are there any (potential) outliers, leverage points or influential points? Provide evidence to support your response. Also, if there are influential points and/or outliers, exclude the points, fit your model without them, and report the changes in your overall conclusions.





This graph displays a scatterplot of the standardized residuals vs a leverage indicator. It also shows a LOESS curve and contours for Cook's distances of 0.5 and 1. Two points (observation 31 and 138) fall outside the boundary of Cook's contours, so they are potentially influential points.

Information of observation 31 and 138 is shown as follows:

95% Confidence Interval

	id	host_is_superhost	host_identity_verified	room_type	accommodates	bathrooms	bedrooms	price
31	5143477	False	True	Shared room	2	1	1	500
138	20481127	False	True	Shared room	1	1	1	45

They are the only two data points whose room type is shared room.
After we exclude them, we fit the model again and here are the results:

MLR Model without hight leverage points Regressing listing price

Predictor	Estimate	SE	t	p-value
(Intercept)	4.4318	0.0592	74.8516	0.0000
host_is_superhostTrue	-0.0088	0.0415	-0.2125	0.8319
host_identity_verifiedTrue	-0.0971	0.0410	-2.3701	0.0184
room_typePrivate room	-0.4526	0.0604	-7.4896	0.0000
accommodates	0.0423	0.0136	3.1024	0.0021
bathrooms	0.2120	0.0445	4.7681	0.0000
bedrooms	0.1337	0.0358	3.7351	0.0002

[a] Multiple R-squared: 0.6839
[b] Adjusted R-squared: 0.6775

The adjusted R^2 increased to 0.6775 from the original 0.6604 and the F-statistic increased from 85.46 to 106.7, so removing the influential points can improve our model.

e. Overall, are there any potential limitations of this analysis? If yes, what are two potential limitations?
Potential limitations:
1. There is multicollinearity among accommodates, bathrooms and bedrooms.
2. We haven't taken quadratic terms or interaction terms into consideration.
3. There may be other variables that will affect listing price.

```
h1, h4 {
  text-align: center;
}

knitr::opts_chunk$set(warning = FALSE, message = FALSE)
library(dplyr)
library(tidyr)
library(ggplot2)
library(brew)
library(stargazer)
library(patchwork)
library(corrplot)
library(kableExtra)
library(broom)
library(car)

df <- read.csv('Respiratory.csv')

ggplot(df) +
    geom_point((aes(x = Age, y = Rate))) +
    geom_quantile(quantiles = c(0.01, 0.05, 0.5, 0.95, 0.99),
                  size = 1,
                  aes(x = Age, y = Rate),
                  linetype = "dashed",
                  show_legend = FALSE) +
    annotate("text", x = -1, y = 28, label = '1%', color = 'blue') +
    annotate("text", x = -1, y = 35, label = '5%', color = 'blue') +
    annotate("text", x = -1, y = 46, label = '50%', color = 'blue') +
    annotate("text", x = -1, y = 62, label = '95%', color = 'blue') +
    annotate("text", x = -1, y = 75, label = '99%', color = 'blue') +
    ggtitle("Respiratory rate - Age") +
    xlab("Age") +
    ylab("Respiratory rate") +
    theme(plot.title = element_text(hjust = 0.5))
    theme_bw() +
    theme(panel.grid = element_blank())
lm <- lm(df$Rate ~ df$Age)
lm %>%
  tidy() %>%
  mutate(p.value = c("<.001", "<.001"), term = c("Intercept", "Age")) %>%
  kable(caption = "SLR Model Regressing respiratory rate on age",
        col.names = c("Predictor", "Estimate", "SE", "t", "p-value"),
        digits = c(4, 4, 4, 4, 4),
        align = "l") %>%
  add_footnote(c("Multiple R-squared: 0.4766","Adjusted R-squared: 0.4758")) %>%
  kable_styling(position="center",full_width = T)
ci_1 <- confint(lm, level = 0.95)
ci_df <- data.frame(ci_1)
rownames(ci_df) <- NULL
Predictor <- c("Intercept", "Age")
cbind(data.frame(Predictor), ci_df) %>%
  kable( caption="95% Confidence Interval",
         col.names = c("Predictor", "2.5%", "97.5%"),
         digits = c(4, 4),
         align = "l") %>%
  kable_styling(position="center") %>%
  kable_styling(position="center", full_width = T)
par(mfrow = c(3, 1))
plot(lm, which = 1)
text(23, 31, 'Figure 1', cex = 1.5)
plot(lm, which = 2)
text(-2.8, 4, 'Figure 2', cex = 1.5)
plot(lm, which = 3)
text(23, 1.9, 'Figure 3', cex = 1.5)

abb <- read.table('Listings_QueenAnne.txt', header = 1, stringsAsFactors = TRUE)
gg1 <- ggplot(abb, aes(x = price)) +
      geom_histogram(bins = 20) +
      ggtitle("Distribution of price") +
      xlab("price") +
      ylab("count") +
      theme(plot.title = element_text(hjust = 0.5))

gg2 <- ggplot(abb, aes(x = log(price))) +
      geom_histogram(bins = 20) +
```

```
      ggtitle("Distribution of log(price)") +
      xlab("log(price)") +
      ylab("count") +
      theme(plot.title = element_text(hjust = 0.5))
gg1 + gg2

## host_is_superhost
his_p <- abb %>% select(host_is_superhost, price) %>% group_by(host_is_superhost) %>% summarise(mean_price = round(mean(pric
e),2))

gg3 <- ggplot(data = his_p, aes(x = host_is_superhost, y = mean_price)) +
          geom_col(aes(fill = host_is_superhost)) +
          geom_text(aes(label=mean_price, vjust = -0.2)) +
          theme(legend.position = 'none')

gg4 <- ggplot(data = abb, aes(x = host_is_superhost,
                              y = log(price),
                              fill = host_is_superhost)) +
            geom_boxplot() +
            geom_jitter(color = 'blue', width = 0.05, size = 0.4, alpha = 0.8) +
            theme(legend.position='none') +
            ylab("log(Price)")

gg3 + gg4


## host_identity_verified
hiv_p <- abb %>% select(host_identity_verified, price) %>% group_by(host_identity_verified) %>% summarise(mean_price = round
(mean(price),2))

gg5 <- ggplot(data = hiv_p, aes(x = host_identity_verified, y = mean_price)) +
          geom_col(aes(fill = host_identity_verified)) +
          geom_text(aes(label=mean_price, vjust = -0.2)) +
          theme(legend.position = 'none')

gg6 <- ggplot(data = abb, aes(x = host_identity_verified,
                          y = log(price),
                          fill = host_identity_verified)) +
            geom_boxplot() +
            geom_jitter(color = 'blue', width = 0.05, size = 0.4, alpha = 0.8) +
            theme(legend.position='none') +
            ylab("log(Price)")

gg5 + gg6


## room_type
rt_p <- abb %>% select(room_type, price) %>% group_by(room_type) %>% summarise(mean_price = round(mean(price),2))

gg7 <- ggplot(data = rt_p, aes(x = room_type, y = mean_price)) +
      geom_col(aes(fill = room_type)) +
      geom_text(aes(label=mean_price, vjust = -0.2)) +
      theme(legend.position="none")

gg8 <- ggplot(data = abb, aes(x = room_type, y = log(price), fill = room_type)) +
      geom_boxplot() +
      geom_jitter(color = 'blue', width = 0.05, size = 0.4, alpha = 0.8) +
      theme(legend.position="none") +
      ylab("log(Price)")

gg7 + gg8

## accommodates

ggplot(abb, aes(accommodates, log(price))) +
  geom_jitter(aes(colour = accommodates), width = 0.05) +
  geom_smooth(method = "lm") +
  theme(legend.position="none")

## bathrooms

ggplot(abb, aes(bathrooms, log(price))) +
  geom_jitter(aes(colour = bathrooms), width = 0.05) +
  geom_smooth(method = "lm") +
  theme(legend.position="none")
```

```
## bedrooms

ggplot(abb, aes(bedrooms, log(price))) +
  geom_jitter(aes(colour = bedrooms), width = 0.05) +
  geom_smooth(method = "lm") +
  theme(legend.position="none")

mlr = lm(log(price) ~ host_is_superhost + host_identity_verified + room_type + accommodates + bathrooms + bedrooms, data = a
bb)

## Model assessment
par(mfrow = c(3,1))
plot(mlr, which = 1)
text(4.3, 1.2, 'Figure 1')
plot(mlr, which = 2)
text(-2.8, 4.5, 'Figure 2')
plot(mlr, which = 3)
text(4.3, 2.0, 'Figure 3')
dat <- abb[, c( "accommodates", "bathrooms",  "bedrooms")]
corrplot(cor(dat), method = "number", type = "upper")
vif(mlr)[,1] %>%
  kable(caption = "VIF values for each predictor",
        col.names = c("VIF"),
        digits = c(4),
        align = "l") %>%
  kable_styling(position="center", full_width = T)
mlr %>%
  tidy() %>%
  kable(caption = "MLR Model Regressing listing price",
        col.names = c("Predictor", "Estimate", "SE", "t", "p-value"),
        digits = c(4, 4, 4, 4, 4),
        align = "l") %>%
  add_footnote(c("Multiple R-squared: 0.6682","Adjusted R-squared: 0.6604"))%>%
  kable_styling(position="center", full_width = T)
confint(mlr, level = 0.95) %>%
  kable(caption="95% Confidence Interval",
         digits = c(4, 4),
         align = "l") %>%
  kable_styling(position="center", full_width = T)
par(mfrow = c(2,1))
plot(mlr, which = 4)
plot(mlr, which = 5)
#axis(1, c(0.0,0.1,0.2,0.3,0.4,0.5))
cooksd <- cooks.distance(mlr)
# find the high-leverage points
influential <- as.numeric(names(cooksd)[(cooksd > 4*mean(cooksd, na.rm=T))])  # influential row numbers
head(abb[influential, ]) %>%
  kable(caption="95% Confidence Interval",
         align = "l") %>%
  kable_styling(position="center", full_width = T)

abb2 <- abb %>% filter(id != 5143477 & id != 20481127)  # remove high leverage point
# fit model
mlr2 <- lm(log(price) ~ host_is_superhost + host_identity_verified + room_type + accommodates + bathrooms + bedrooms, data =
abb2)
mlr2 %>%
  tidy() %>%
  kable(caption = "MLR Model without hight leverage points Regressing listing price",
        col.names = c("Predictor", "Estimate", "SE", "t", "p-value"),
        digits = c(4, 4, 4, 4, 4),
        align = "l") %>%
  add_footnote(c("Multiple R-squared: 0.6839","Adjusted R-squared: 0.6775"))%>%
  kable_styling(position="center", full_width = T)
```