Homework 1

Pranav Manjunath

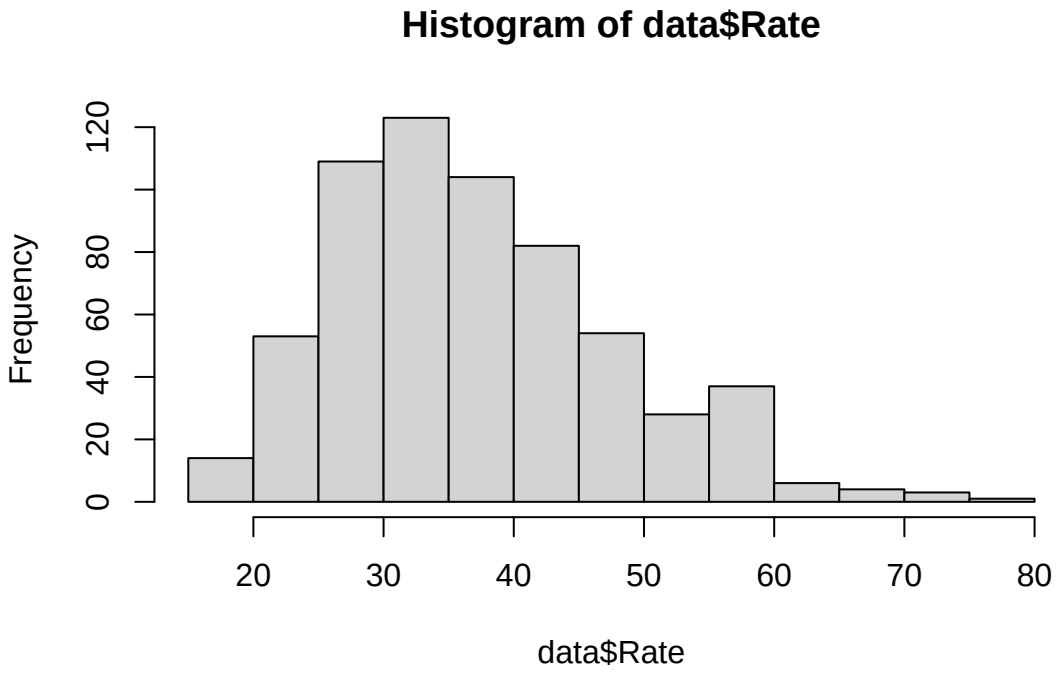
Question 1

(A)

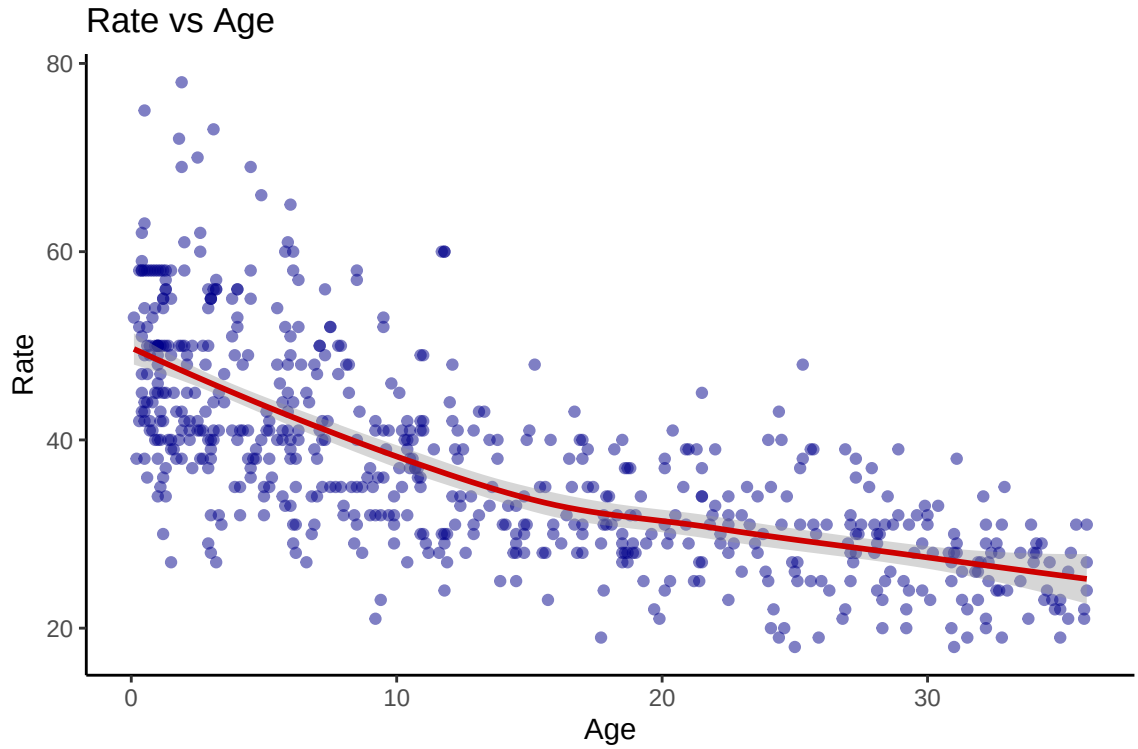
Do exploratory analysis on the data and include a useful plot that a physician could use to assess a "normal" range of respiratory rates for children of any age between 0 and 3.

The dataset consists of 618 observations across 3 columns, X, Age, and Rate. The mean of the age and respiratory rate is 13.39 months and 37.74 respectively. There is a negative correlation between Rate and Age (-0.6903627). The histogram of Rate is a skewed distribution and hence taking log(Rate) could be use to assess a "normal" range of respiratory rates for children of any age between 0 and 3.

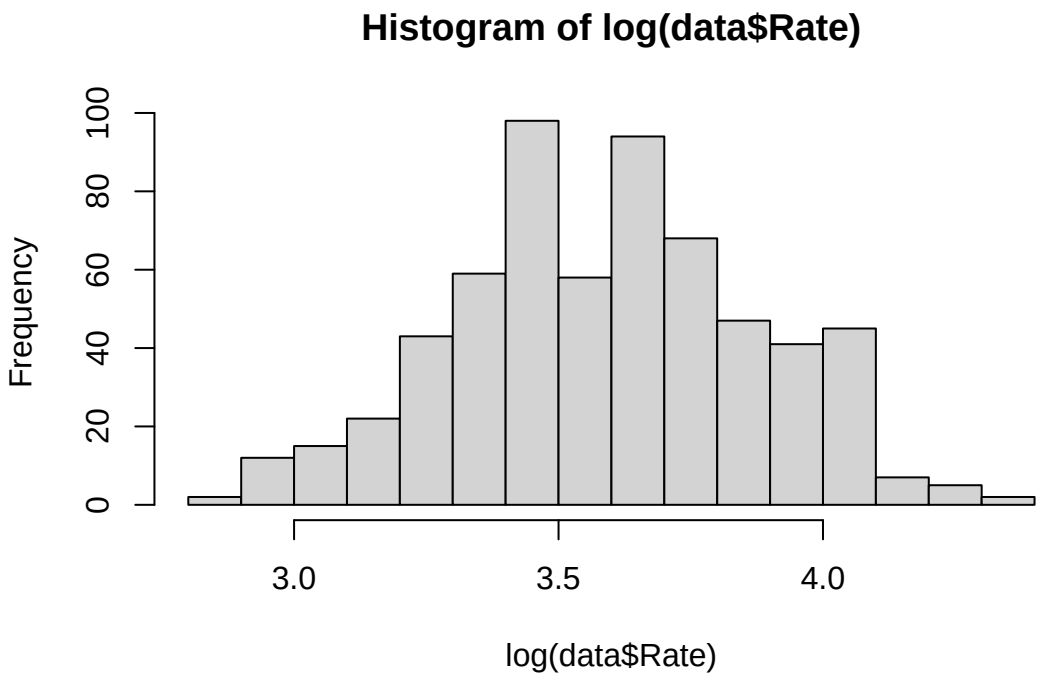
```
##        X              Age             Rate      
##  Min.   :  1.0   Min.   : 0.10   Min.   :18.00  
##  1st Qu.:155.2   1st Qu.: 3.80   1st Qu.:30.00  
##  Median :309.5   Median :10.55   Median :36.50  
##  Mean   :309.5   Mean   :13.39   Mean   :37.74  
##  3rd Qu.:463.8   3rd Qu.:22.00   3rd Qu.:44.00  
##  Max.   :618.0   Max.   :36.00   Max.   :78.00
```



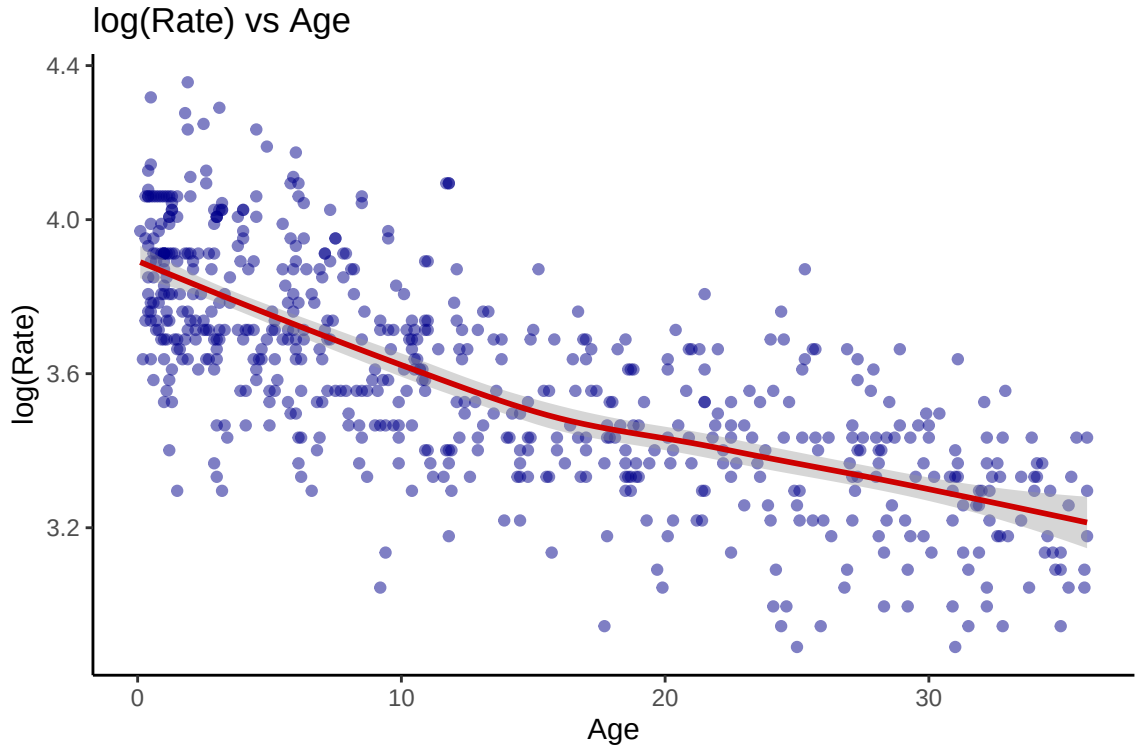

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



When fitting a line to the Rate vs Age scatter plot, we can see that there is a slight non linear relationship. As the histogram looks skewed, I have transformed the response variable to its log value.




```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The log(Rate) histogram has a normal distribution. I have used log(Rate) as the response variable for the linear regression model.

(B)

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

$\hat{y} = \hat{(}b_0) + \hat{(}b_1)$ * x

$log(\hat{Rate}) = 3.8451185 - 0.0190090 * Age$

(C)

Fit the model to the data and interpret your results.

The model has an intercept of 3.8451185 on the log(y) axis. As this data is not centered, the meaning of this intercept is if the age of a child is 0 (unpractical), the respiratory rate will be 46.76423 (log(3.8451185)). However, if the age is centered, the intercept indicates that the respiratory rate of a child of 13.39 months is 36.2551 (log(3.59058)). Age is statistically significant. The slope of Age is -0.0190090 on the log(rate) axis, which indicates that an increase of age by one month would lead to a 1.88295% decrease of the respiratory rate. The percent of the variability in respiration rate explained by the regression model (R-squared) is 52.01%.

(D)

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.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.845	0.01263	304.5	0
Age	-0.01901	0.0007357	-25.84	2.74e-100

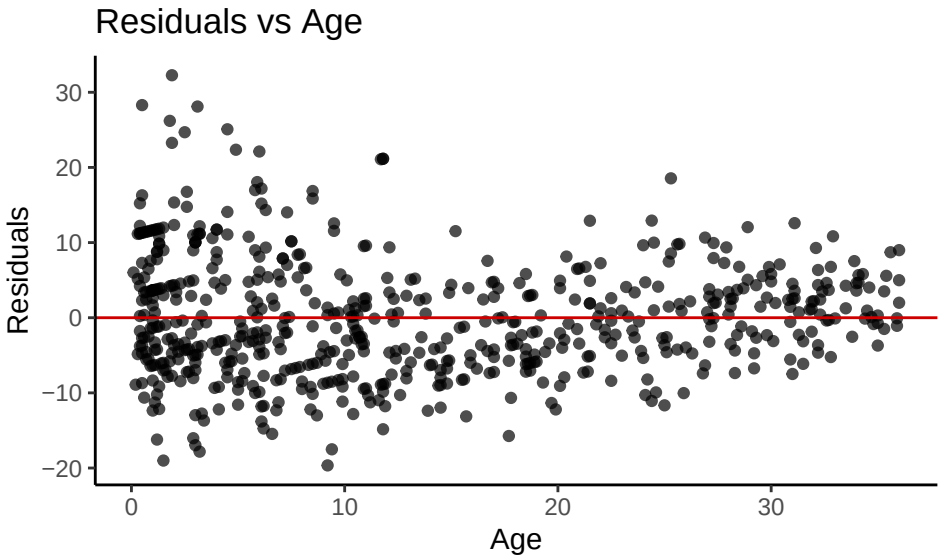
Table 2: Fitting linear model: log(Rate) ~ Age

Observations	Residual Std. Error	R^2	Adjusted R^2
618	0.1964	0.5201	0.5193

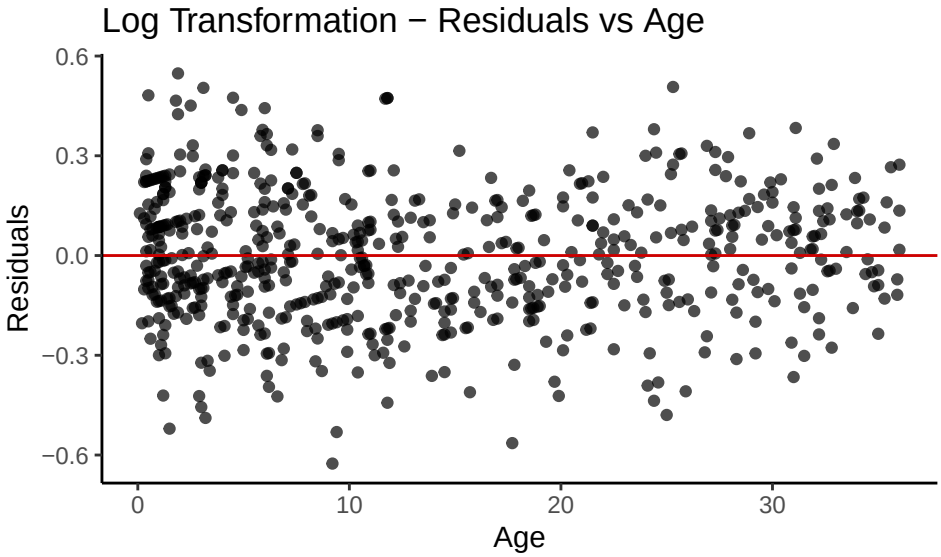
(E)

Is there enough evidence that the model assumptions are reasonable for this data? You should consider transformations (think log transformations, etc) if you think there's a violation of normality and/or linearity.

Testing for Linearity: To check whether the model (without log transformation) satisfies the linearity assumption, the Model Residual vs Age plot is shown below.

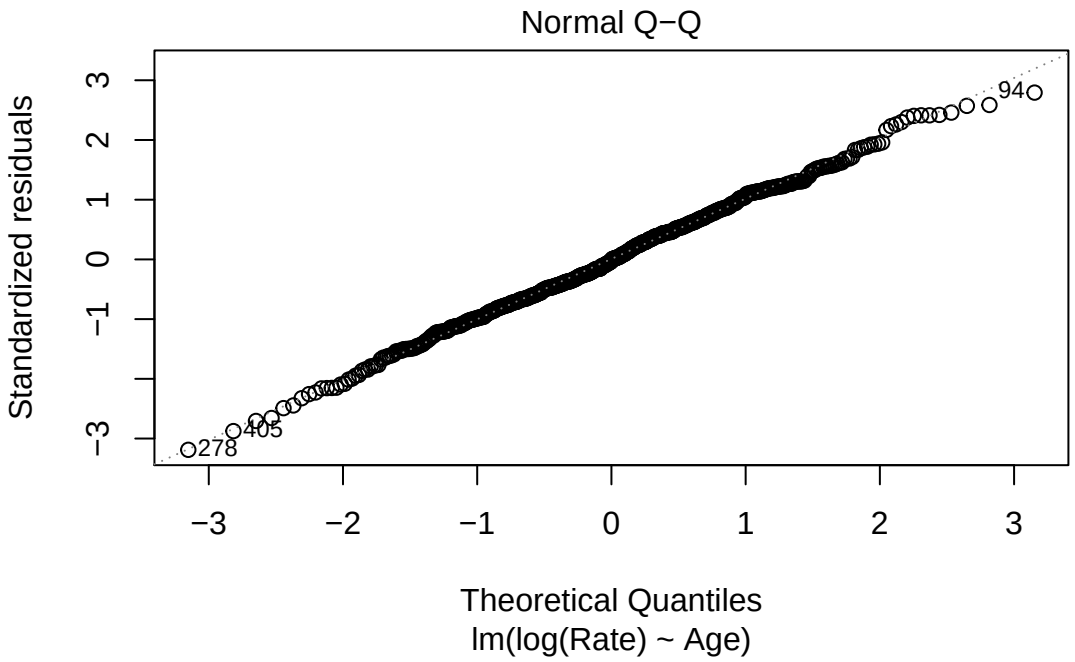


As shown above, there seems to be a quadratic curve pattern amongst data points in the plot. Hence, the log(Rate) is taken in consideration and the Model Residual vs Age plot (with log transformation) is plotted below.



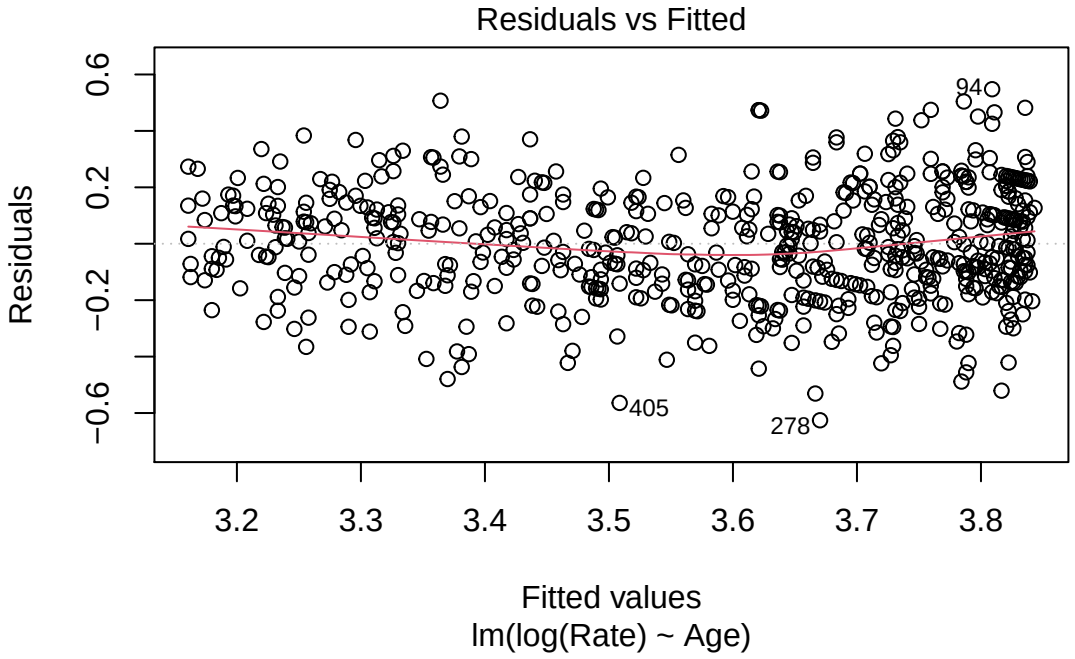
After transforming the y variable to its log, quadratic curve seems less evident and a higher degree of randomness can be observed. Hence, I used log(Rate) as the response variable for the regression model.

Testing for Normality: Using log(Rate) as the response, the Q-Q Plot is plotted below to identify Normality.



As the points lie on the 45 degree line, the normality assumption is not violated.

Testing for Independence and Equal Variance: Using log(Rate) as the response, the Residual vs Fitted Plot is plotted below to check if the independence and equal variance assumptions are not violated.



As the plots are spaced out throughout the X axis, and lie around the 0 residual mark on the Y axis, the

model is assumed to be independent and have equal variance.

(F)

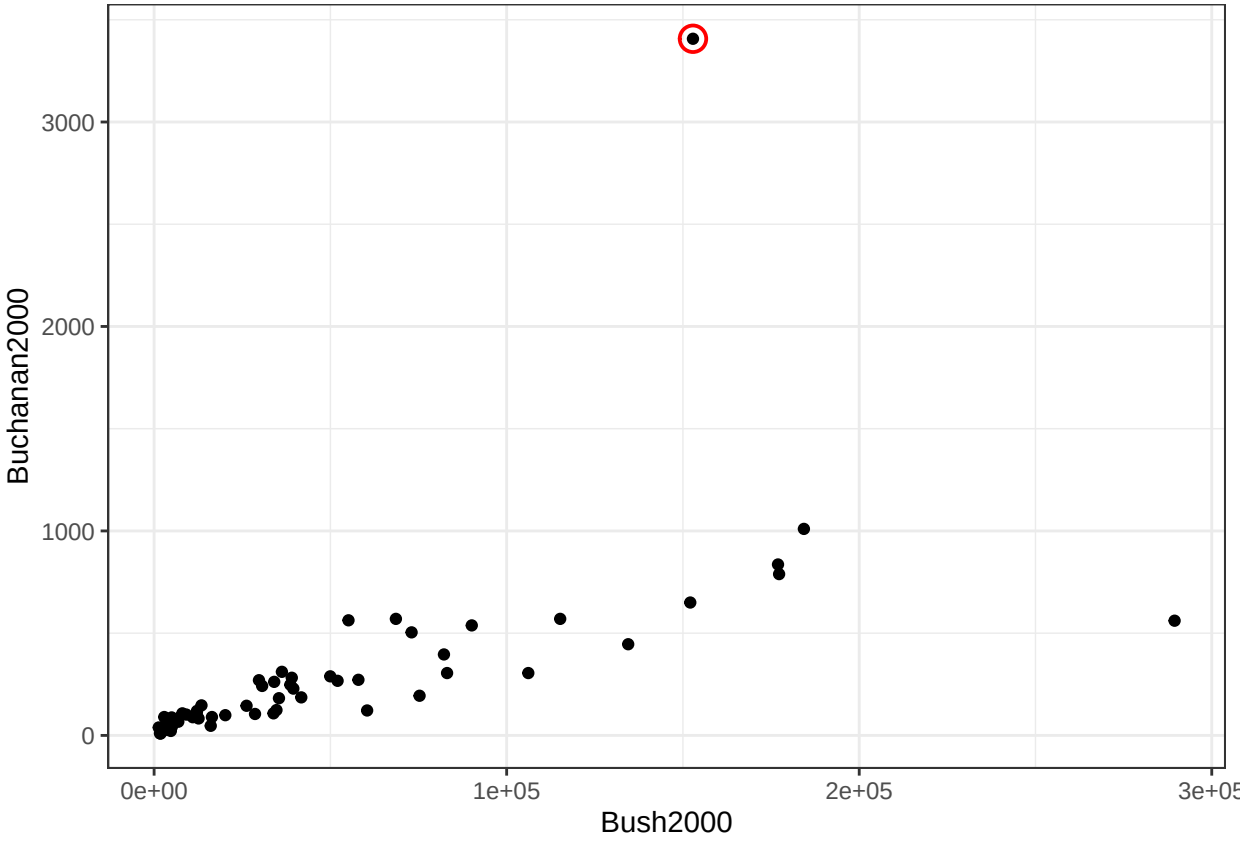
Demonstrate the usefulness of the model by providing 95% prediction intervals for the rate for three individual children: a 1 month old, an 18 months old, and a 29 months old.

Age	fit	lwr	upr
1	45.88	31.18	67.53
18	33.21	22.58	48.86
29	26.95	18.31	39.67

Question 2

(A)

Make a scatterplot of the variables Buchanan2000 and Bush2000. What evidence is there in the scatterplot that Buchanan received more votes than expected in Palm Beach County?



As shown in the above scatter plot, the circled data point identifies that Buchanan received more votes than expected in Palm Beach County. As the circled point is clearly an outlier, illustrating that is the only county where the votes are past 3000.

(B)

Fit a linear regression model to the data to predict Buchanan votes from Bush votes, without using Palm Beach County results. You should consider transformations for both variables if you think there's a violation of normality and/or linearity.

$\hat{(}y) = \hat{(}B_0) + \hat{(}B_1)$x

$\hat{(}log(Buchanan2000)) =$ -2.34149 + 0.73096 * log(Bush2000)

(C)

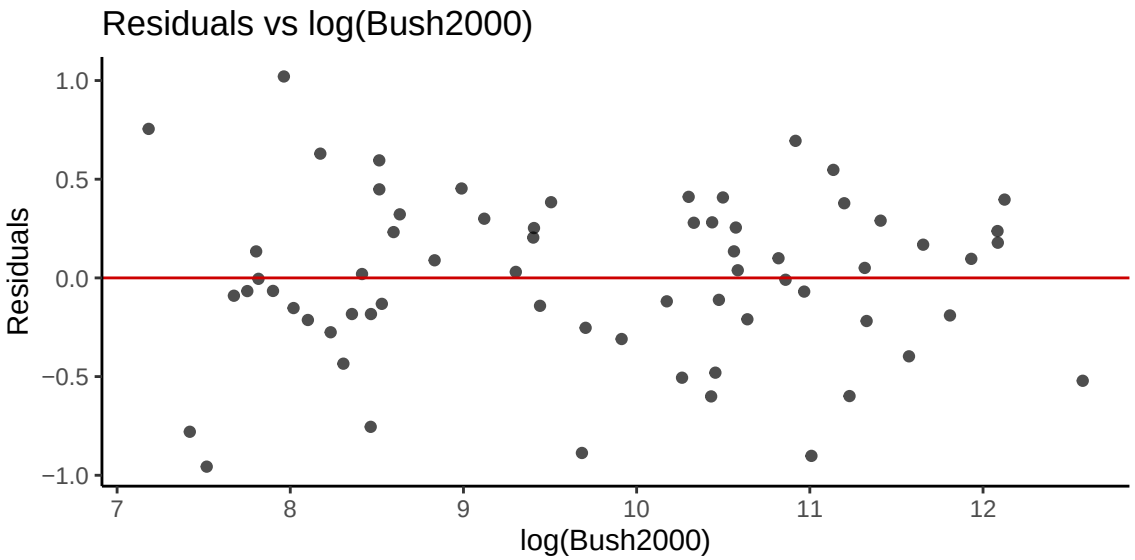
Include the output from the final regression model that you used, as well as evidence that the model fits the assumptions reasonably well.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.341	0.3544	-6.607	9.071e-09
log(Bush2000)	0.731	0.03597	20.32	1.294e-29

Table 5: Fitting linear model: log(Buchanan2000) ~ log(Bush2000)

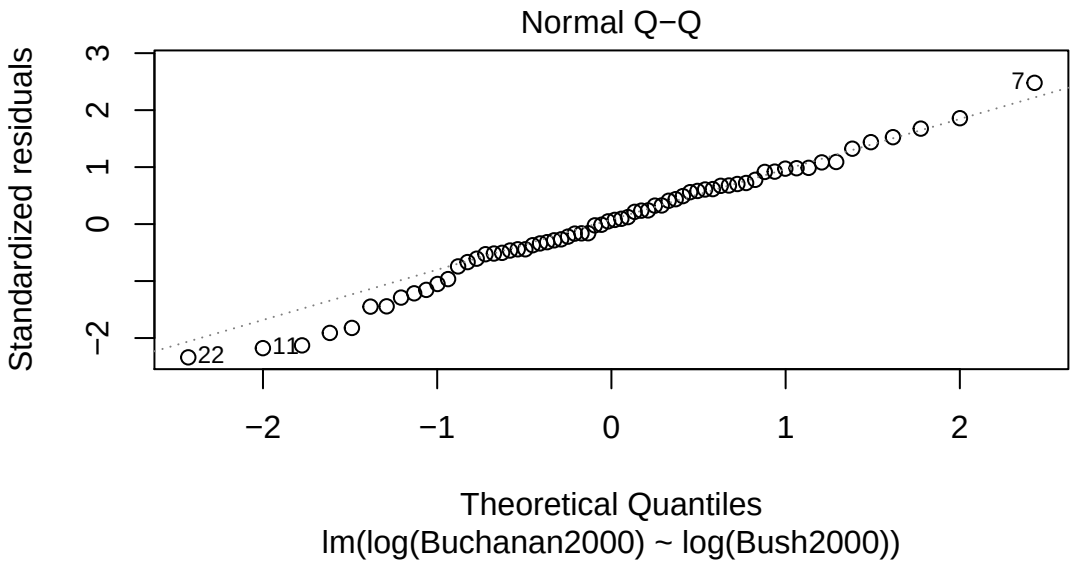
Observations	Residual Std. Error	R^2	Adjusted R^2
66	0.4198	0.8658	0.8637

Test for Linearity



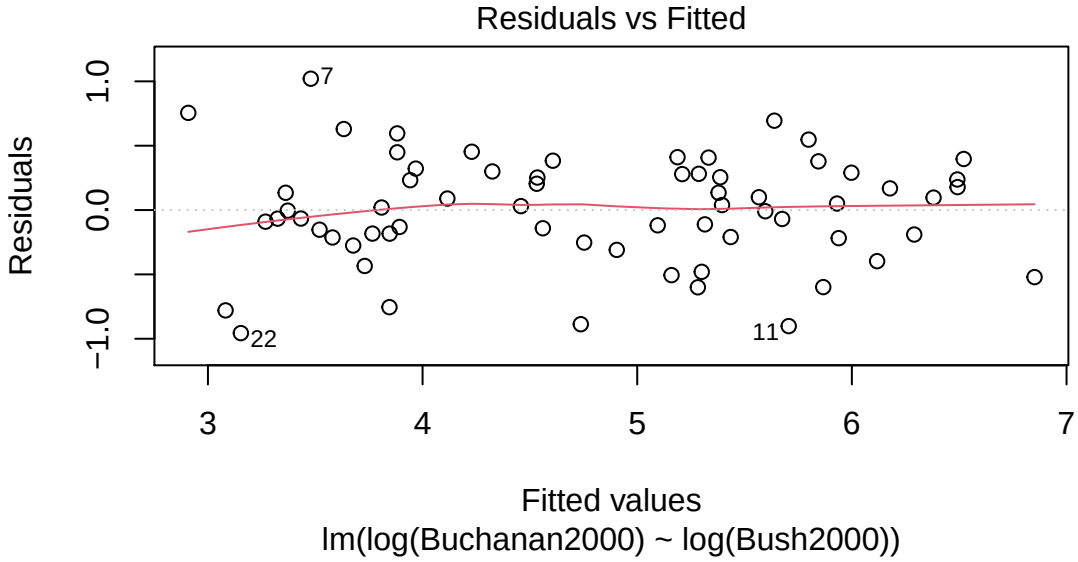
As there exists no distinct pattern, the model linearity assumption holds true.

Test for Normality



As most of data points lie on the 45 degree line, the model normality assumption holds true.

Test for Independence and Equal Variance



As the data points are spread throughout the X axis and are around the 0 residual mark on the Y axis, the model independence and equal variance holds true.

(D)

Obtain a 95% prediction interval for the number of Buchanan votes in Palm Beach from this result, assuming the relationship is the same in this county as in the others. If it is assumed that Buchanan's actual count contains a number of votes intended for Gore, what can be said about the likely size of this number from the prediction interval?

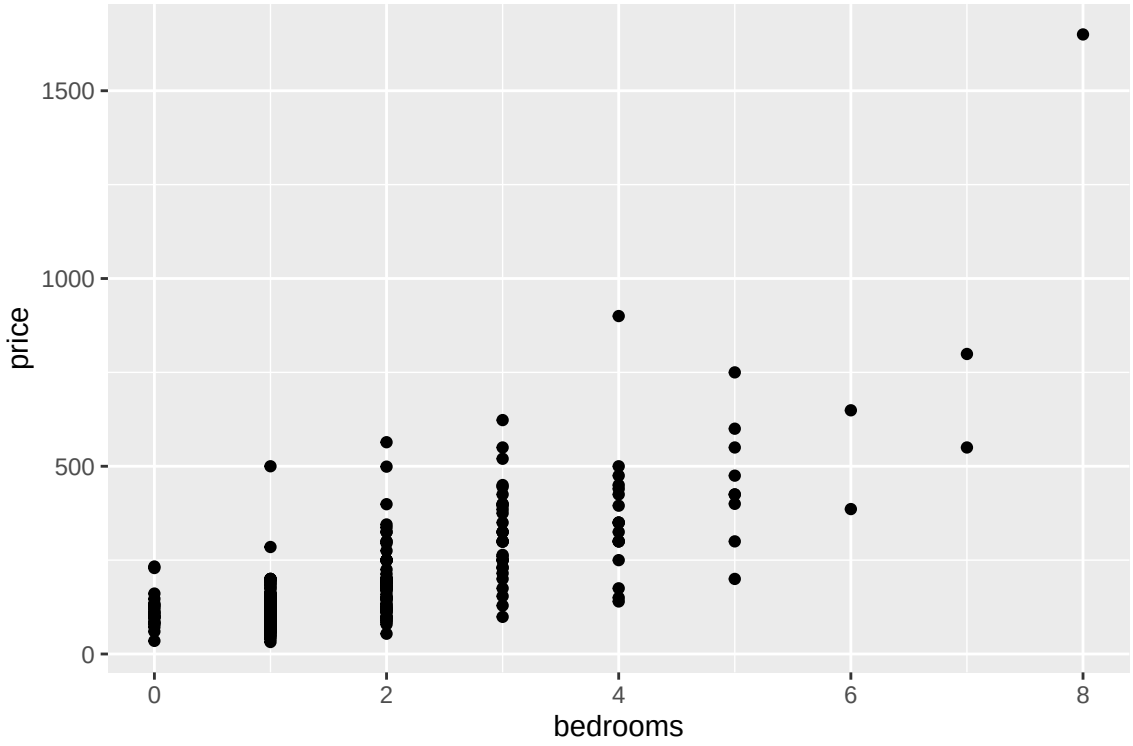
Bush2000	fit	lwr	upr
152846	592.4	250.8	1399

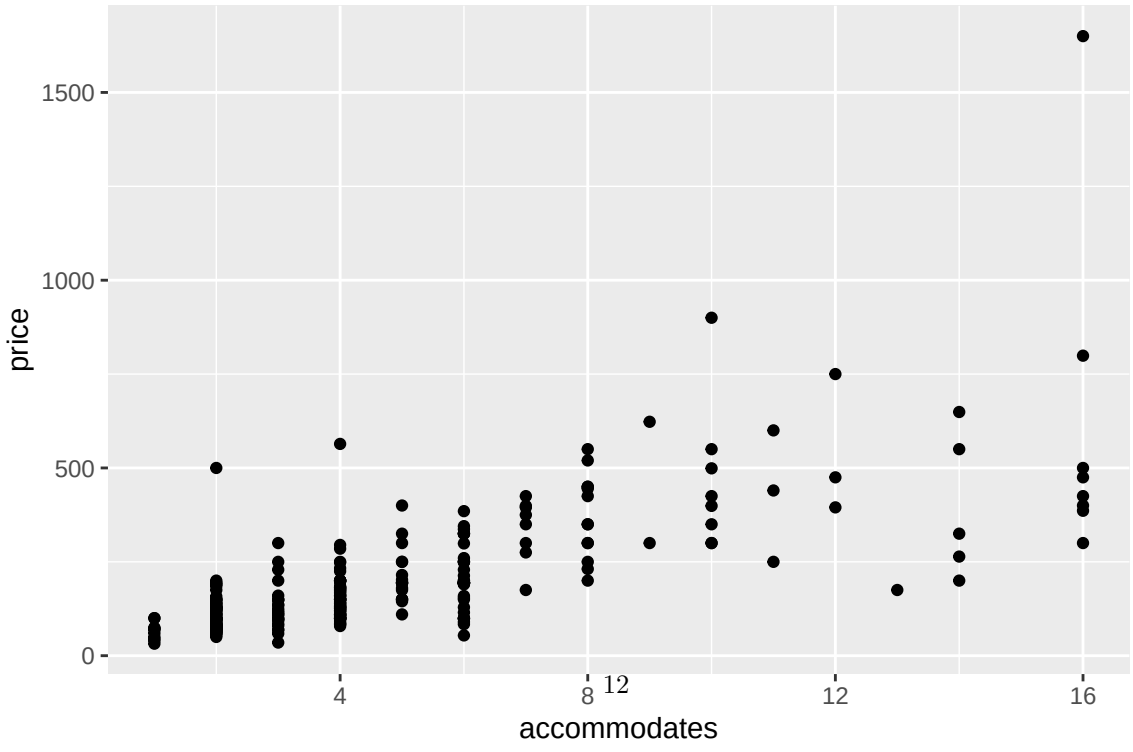
Initially, the votes for Buchanan at Palm Beach County were 3407. When the linear regression model was fitted after removing this point, for the same number of votes for Bush at Palm Beach County (153846), the model predicts the number of votes for Buchanan to be ~592 votes (difference of 2815 votes from actual number of votes). At 95% prediction interval, the lower limit is 250.8 and the upper limit is 1399. Even though the prediction intervals are wider than confidence intervals, the upper limit is still significantly lower than the actual votes for Buchanan (3407). This concludes that Buchanan's actual count contains a number of votes intended for Gore.

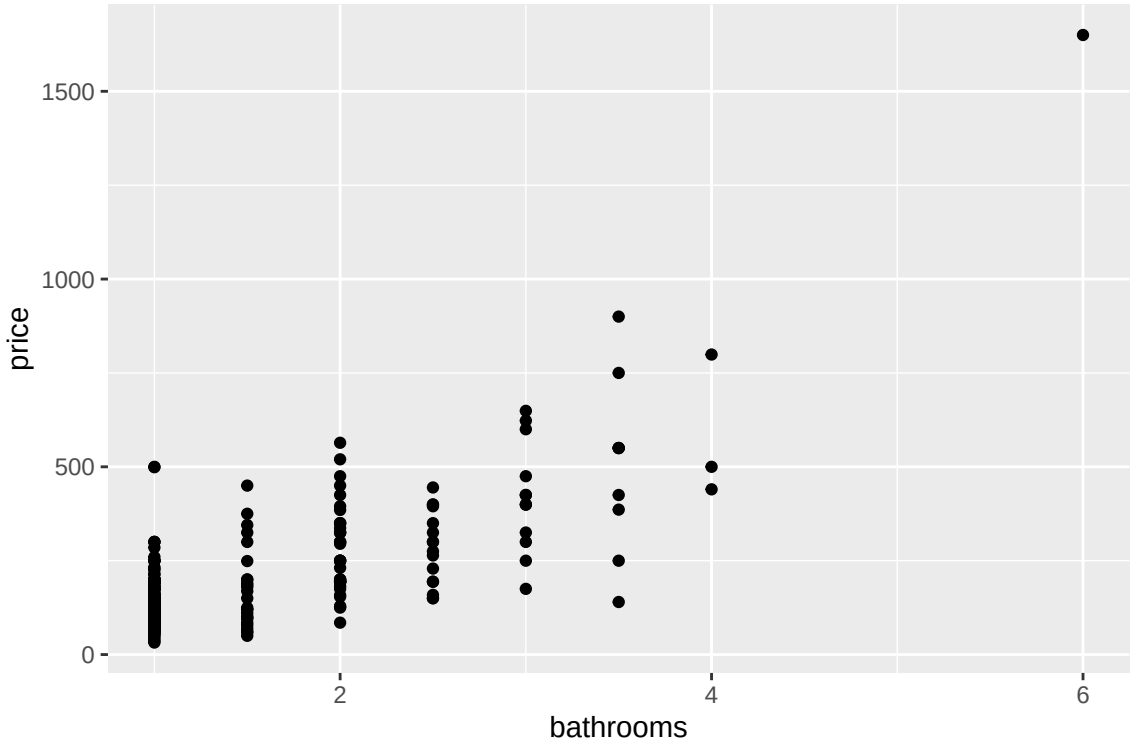
Question 3

(A)

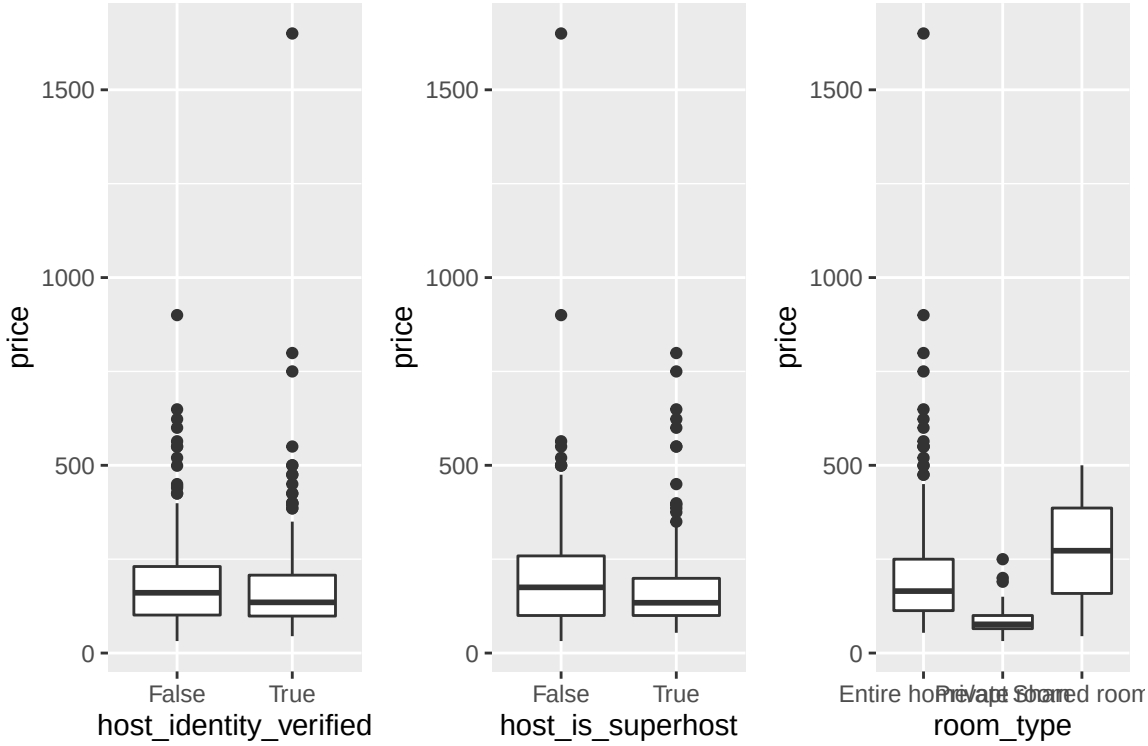
Analyze the data using host_is_superhost,host_identity_verified, room_type, accommodates, bathrooms and bedrooms as predictors. You should start by doing EDA, then model fitting, and model assessment. You should consider transformations if needed.





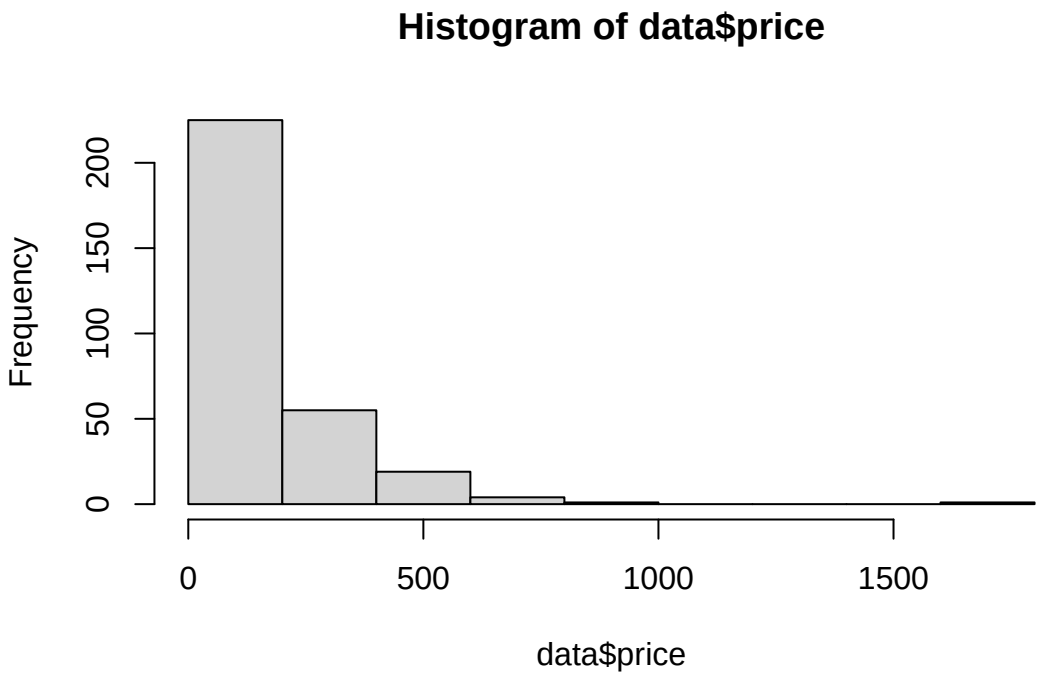


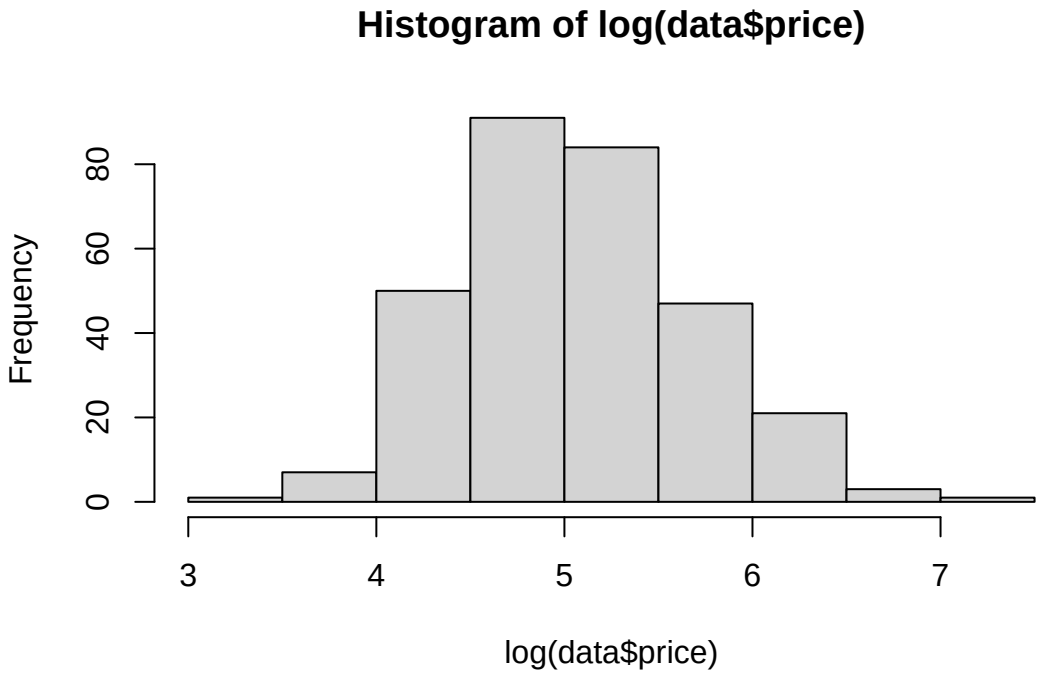
Looking at the three scatter plots above, we tend to see a positive linear relationship between bedrooms and price, bathrooms and price, and accommodates and price.



The mean price of a shared room listing is the highest among the different room types. The mean listing price of a verified host is slightly less than the mean listing price of a non verified host. Similarly, the mean

listing price of a superhost is less than the mean listing of a non superhost.





The Price variable (first histogram) does not follow the normal distribution. However, the log(data$Price) variable follows the normal distribution and will use this as the response variable for the regression model.

(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, and p-values or confidence intervals.

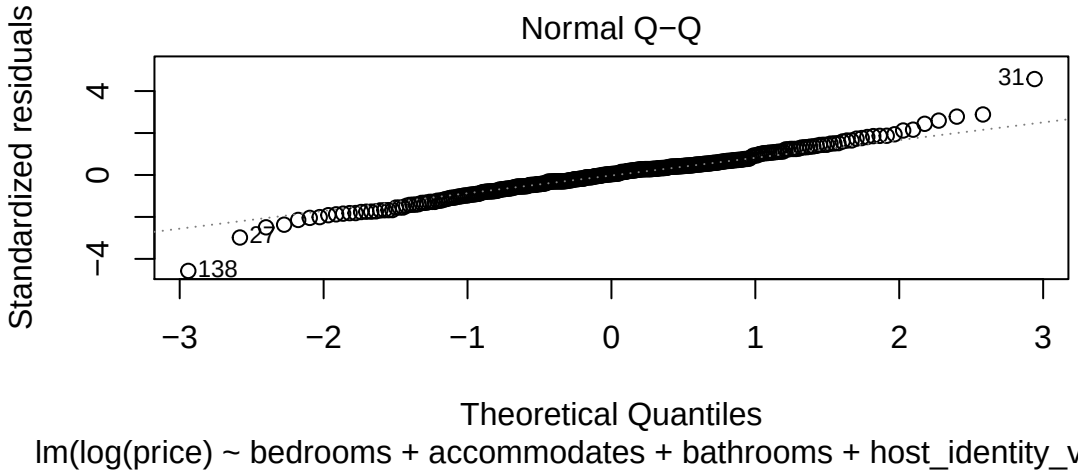
Regression Model Output

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.43	0.0613	72.26	1.661e-190
bedrooms	0.1305	0.03706	3.521	0.0004967
accommodates	0.0441	0.01413	3.122	0.001971
bathrooms	0.2111	0.04603	4.586	6.664e-06
host_identity_verifiedTrue	-0.09682	0.0424	-2.283	0.02312
host_is_superhostTrue	-0.008744	0.04294	-0.2036	0.8388
room_typePrivate room	-0.4502	0.06257	-7.196	5.097e-12
room_typeShared room	0.2698	0.2632	1.025	0.3061

Table 8: Fitting linear model: log(price) ~ bedrooms + accommodates + bathrooms + host_identity_verified + host_is_superhost + room_type

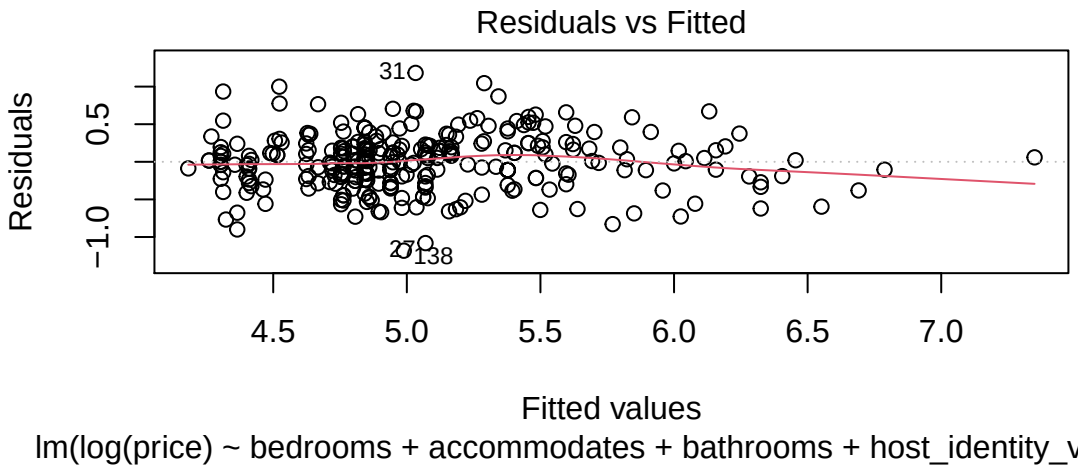
Observations	Residual Std. Error	R^2	Adjusted R^2
305	0.3657	0.6682	0.6604

Normality Assumption



As majority of the data points are on the 45 degree line, the model does not violate the normality assumption.

Independence and Equal Variance



As the points are random and spread throughout the X axis, the model does not violate the independence and equal variance assumption.

(C)

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

The model intercept is 4.43 which means that if a listing has no bedrooms, 0 accommodates, 0 bathrooms, host_identify_verified and host_is_superhost is False, and room type = 'Entire home/apt' the price of the house is $83.93142 (log(4.43)). The percent of the variability in price explained by the regression model (R-squared) is 66.82%.

Keep the other variables constant,

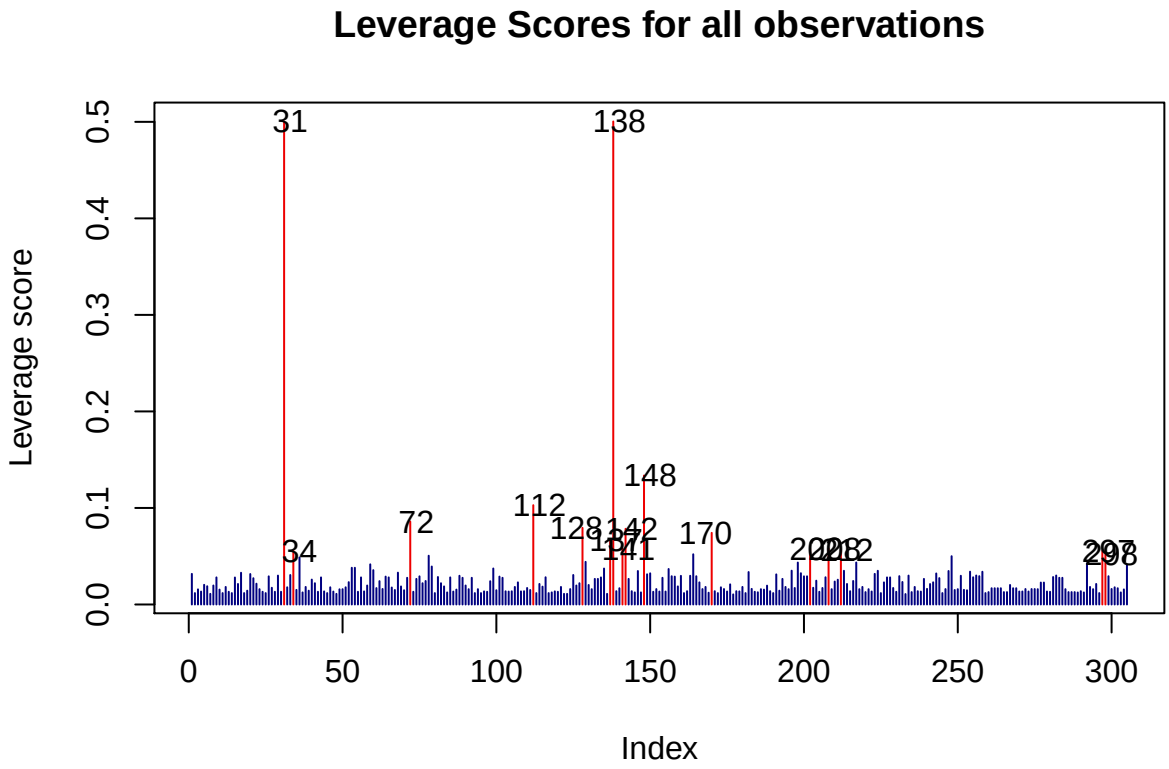
- A unit increase in bedrooms results in a price relative increase of 1.139398
- A unit increase in bedrooms results in a price relative increase of 1.235036
- A unit increase in accommodates results in a price relative increase of 1.045087
- Verified Host Identity by AirBnB results in a price relative decrease of 0.9077194
- A super host results in a price relative decrease of 0.9912941
- A Private Room listing results in the price relative decrease of 0.6375006
- A Shared Room listing results in the price relative increase of 1.309702

(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.

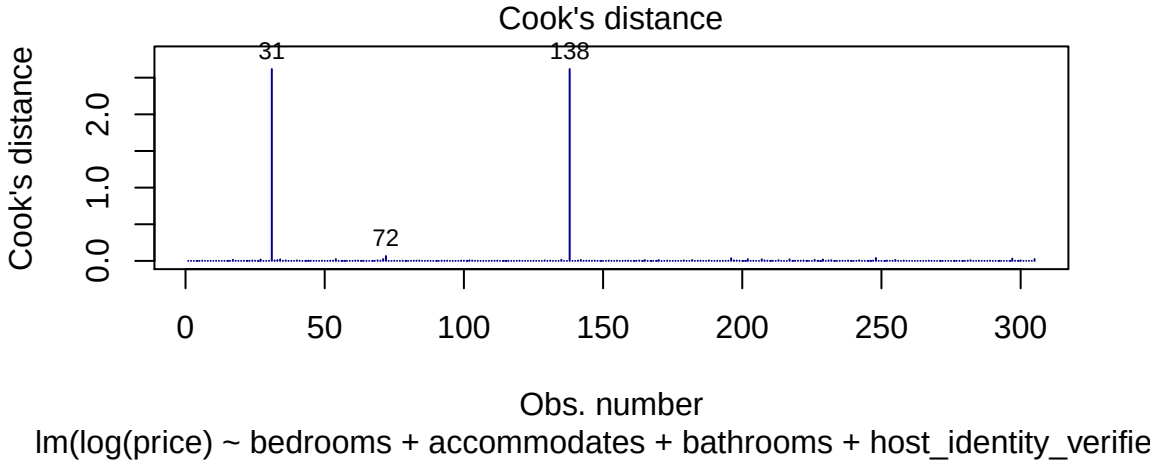
Identifying Leverage Points with leverage score = 2(p+1)/n

```
## Warning in c(1:n)[lev_scores > (2 * p/n)] + c(rep(2, 4), -2, 2): longer object
## length is not a multiple of shorter object length
```



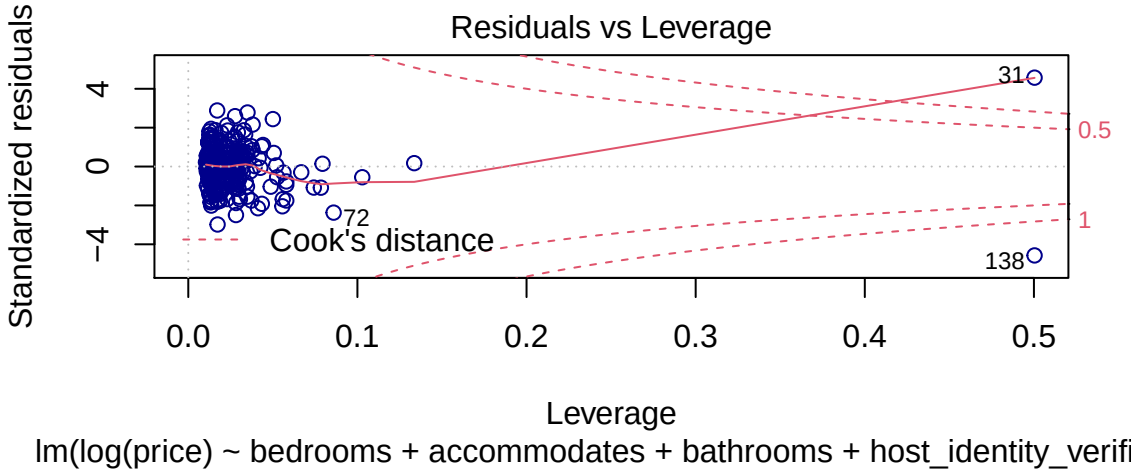
There are two leverage points in the data, data point 31 and data point 138.

Identifying Influence Points having Cook's distance>0.5



Data Points 138 and 31 are high influence points as their Cook's distance is greater than 0.5.

Identifying Outliers



Data Points 138 and 31 seem to also be outliers as their standardized residuals are at 4 and -4. Hence, points 138 and 31 are removed from the data and the updated data is fitted again.

After removing the high influence points and outliers from the data, the value "Shared Room" room type is removed from the dataset and also removed from the linear regression model output.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.432	0.05921	74.85	2.268e-194
bedrooms	0.1337	0.0358	3.735	0.000225
accommodates	0.04234	0.01365	3.102	0.002105
bathrooms	0.212	0.04445	4.768	2.922e-06
host_identity_verifiedTrue	-0.09706	0.04095	-2.37	0.01842
host_is_superhostTrue	-0.008812	0.04147	-0.2125	0.8319
room_typePrivate room	-0.4526	0.06043	-7.49	7.998e-13

Table 10: Fitting linear model: log(price) ~ bedrooms + accommodates + bathrooms + host_identity_verified + host_is_superhost + room_type

Observations	Residual Std. Error	R^2	Adjusted R^2
303	0.3531	0.6839	0.6775

(E)

Overall, are there any potential limitations of this analysis? If yes, what are two potential limitations?

There are potential limitations of this analysis. The data used in this analysis is a very small subset of the AirBnB listing in Queen Anne, Seattle, WA. Hence, the model (trained on the small subset of data) cannot accurately predict listing prices in the area. Only 2 data points in the entire dataset had a room type of "Shared Room". They were treated as outliers and influential points. In order to make accurate predictions, it is important to extract more Shared Room data points.