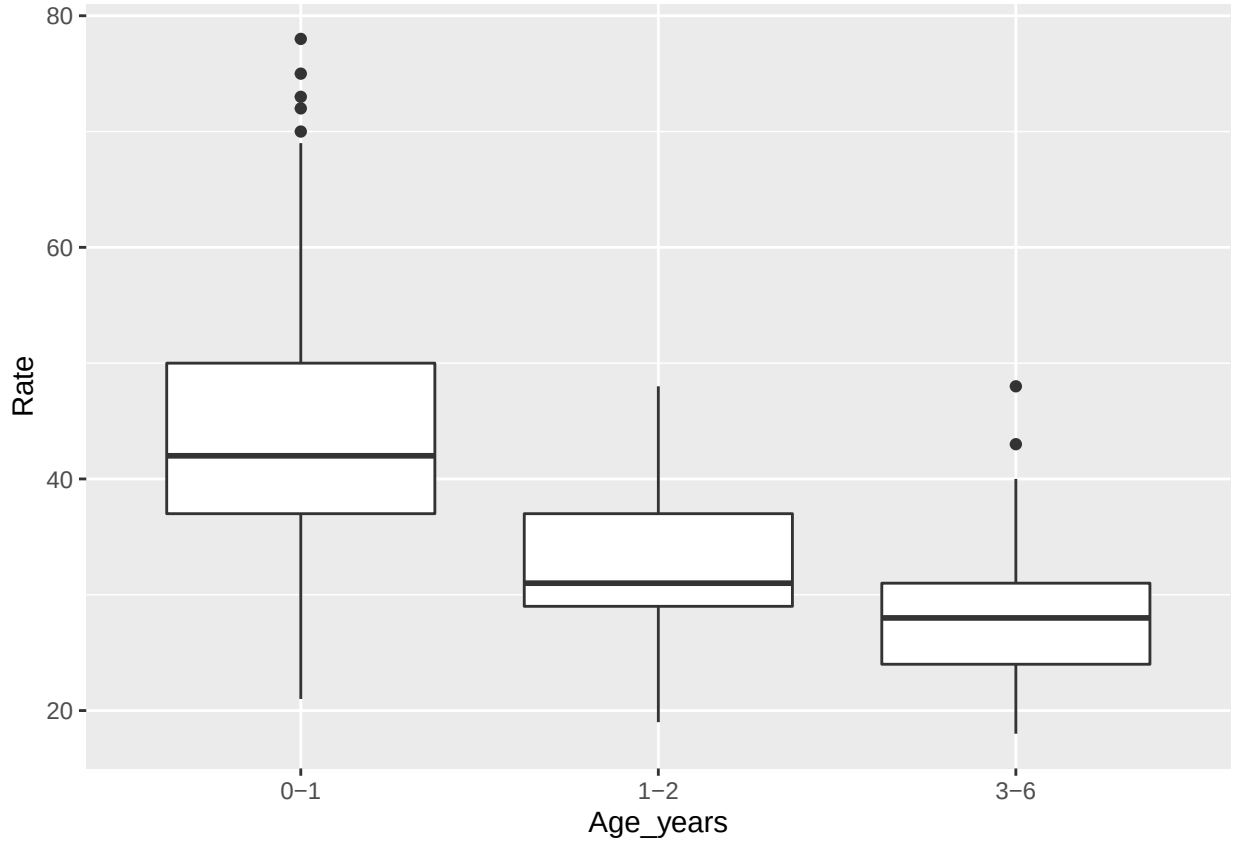
Assignment 1

Mohammad Anas

Question 1

Part1

When we conduct exploratory data analysis, we see that as age increases there is a a gradual decrease in the heart rate. This is clearly visible from the box plot shown.



Part 2

We will use a simple linear regression equation to predict rate from age.

$$Rate_t = \beta_0 + \beta_1 * Age_t. \tag{1}$$

Part 3

According to our model, an increase in 1 unit(month) of Age leads to a decrease of 0.6957 units on average in the heart rate. As the p-value for this coeffecient is small, the age is a statistically significant coeeficient.The intercept was found to be 47.05, which according to our model is the heart rate if the age is zero. As this is counter intuitive, I centered the data and ran the regression. This resulted in an intercept of 1.597e-15. This means a child with an age of 13.39(mean age) months will have the heart rate of 1.597e-15 on average. However as the p-value for the intercept coefficient was 1 (only in the case of centered data), we assume this is statistically insignificant.

Part 4

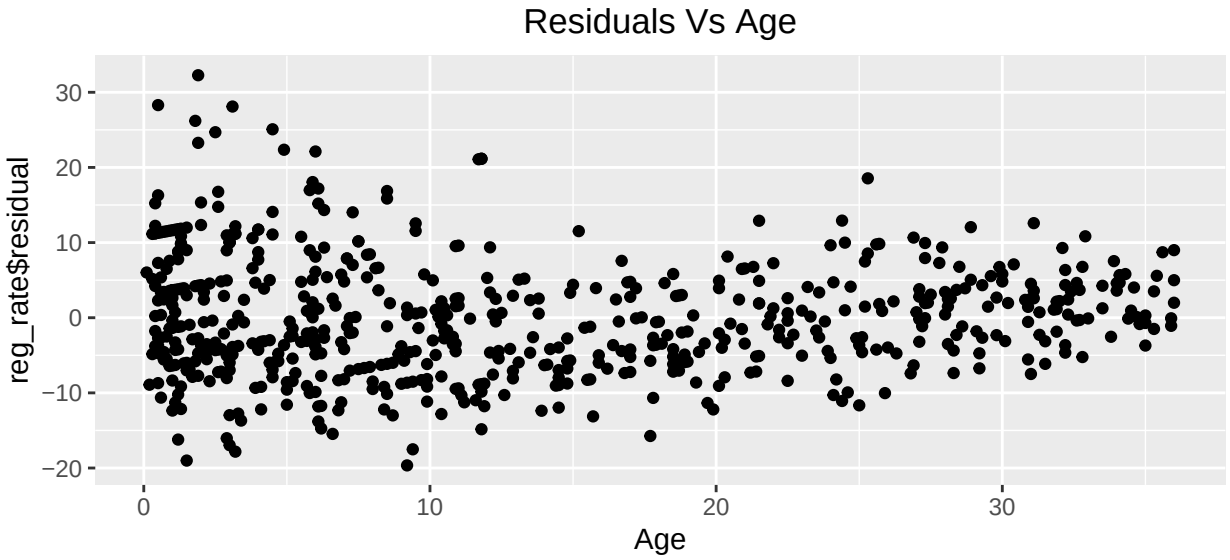
Here are the results of regression run on the non centered data.

Table 1: Results

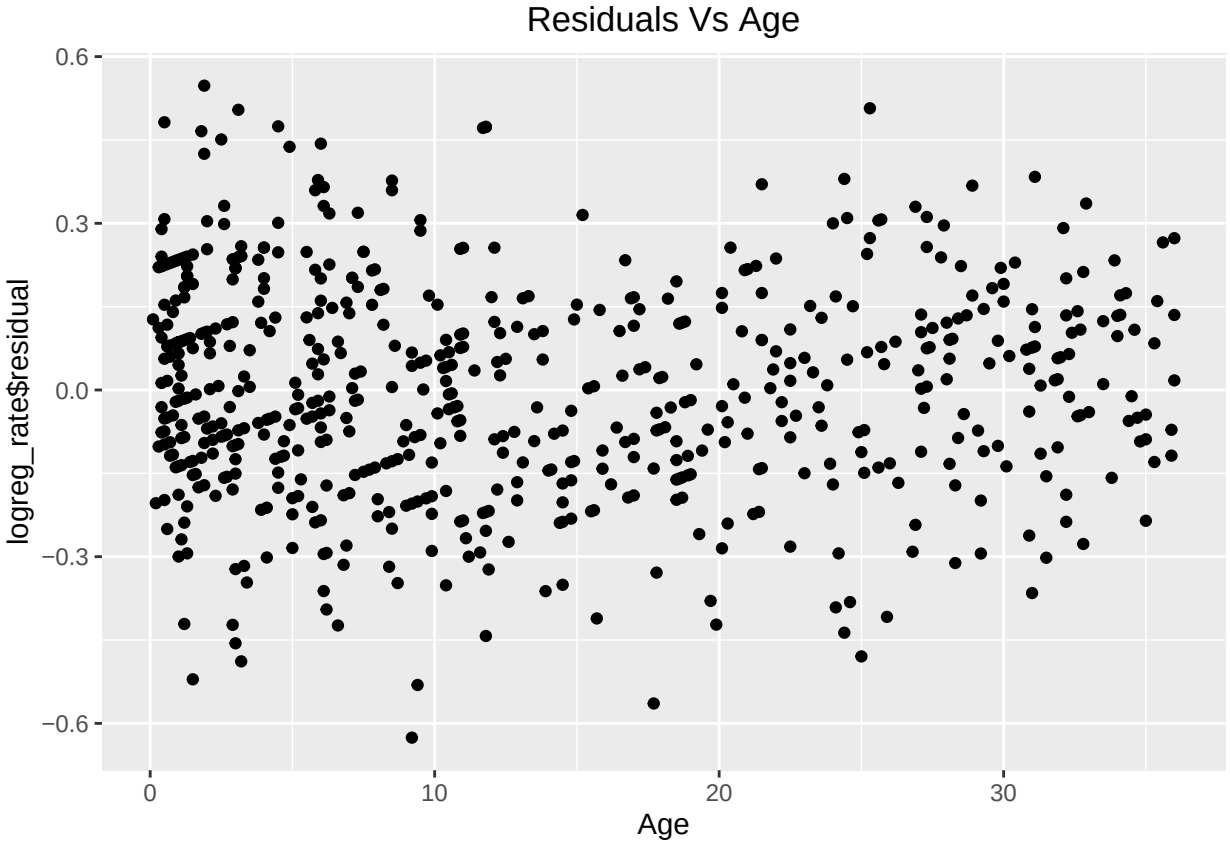
	Dependent variable:
	Rate
Age	-0.70^{***}
	(0.03)
Constant	47.05^{***}
	(0.50)
Observations	618
R^2	0.48
Adjusted R^2	0.48
Residual Std. Error	7.84 (df = 616)
F Statistic	560.92^{***} (df = 1; 616)
Note:	*p<0.1; **p<0.05; ***p<0.01

Part 5

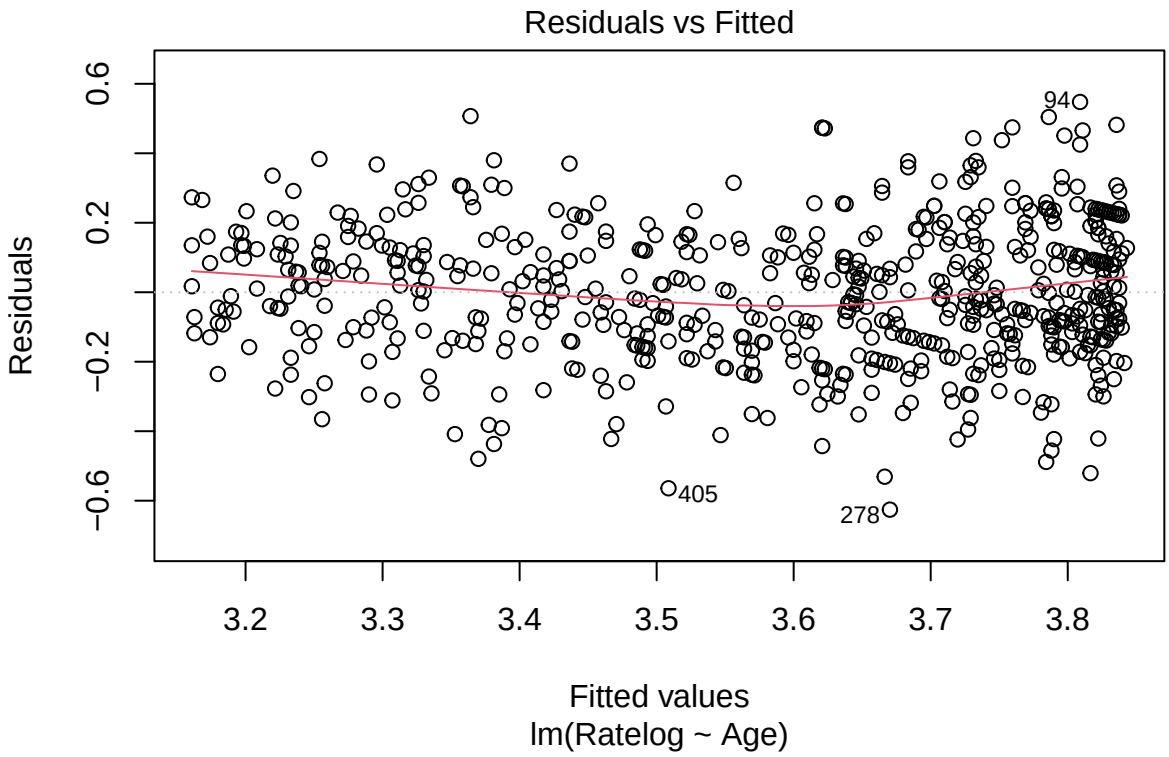
When assessing the model we notice that the linearity assumption has been violated as we see somewhat increasing trend in the “Residuals vs Age” plot.



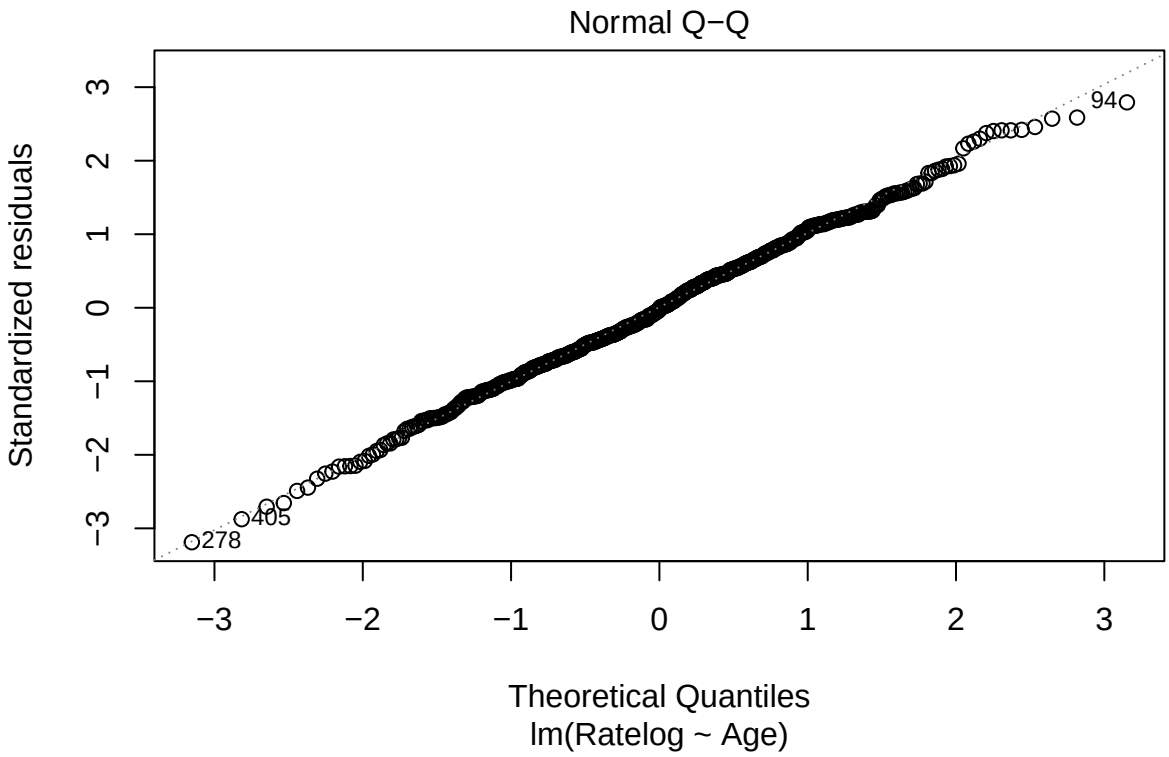
To improve our model, I took the log of the Rate variable and then tested the model for the assumption of linearity. We do see that the data points are more randomly scaterred and we seem good with the linearity assumption.



We notice below by looking at the "Residuals vs Fitted" plot that the variance of residuals are almost constant across all fitted values. Secondly, we see some randomness in the plot, so it is safe to say that the constant variance assumption and the independence assumption have been satisfied by our model.



By observing the QQ-Plot we notice that the normality assumption has also not been violated. The QQ-Plot can be seen below.



Part 6

Below provided are the prediction intervals and the predicted values of Rate, as estimated by our model. The prediction interval for each age tells us that if we take a large number of babies with the same age, 95% of the time, the heart rate for the babies will fall within that prediction interval. The row of the table below present the prediction intervals in the order of 1,18,29.

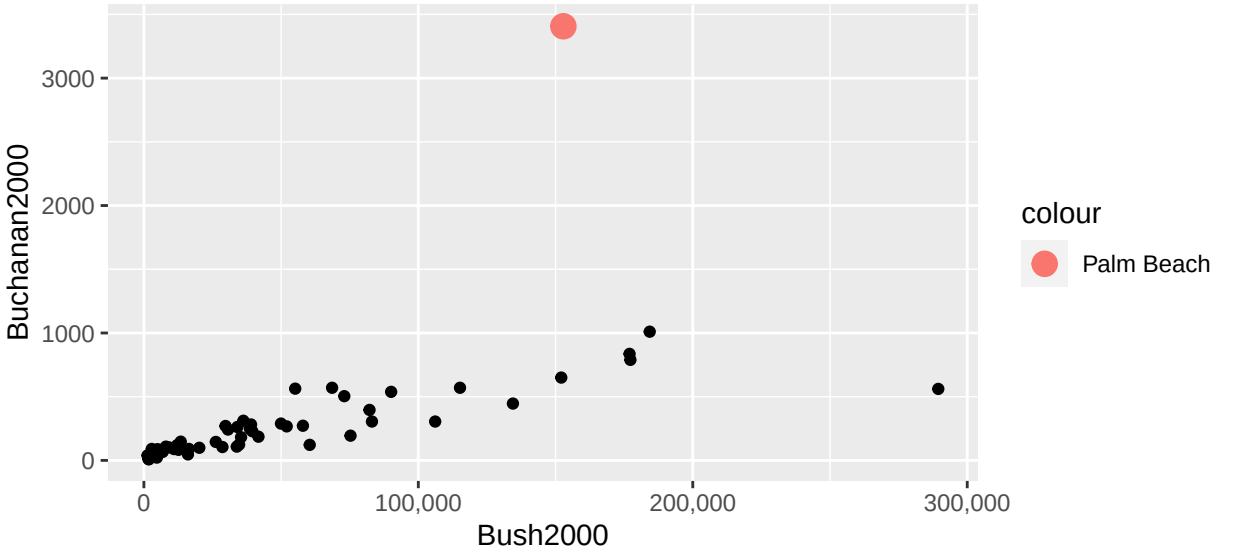
Table 2: Prediction Intervals

	fit	lwr	upr
1	45.88	31.18	67.53
2	33.21	22.58	48.86
3	26.95	18.31	39.67

Question 2

Part 1

The data point corresponding to the county Palm Beach is clearly an outlier. The number of votes received in that county are significantly greater than other counties. The number is clearly way more than expected and serves as an evidence that something was wrong.

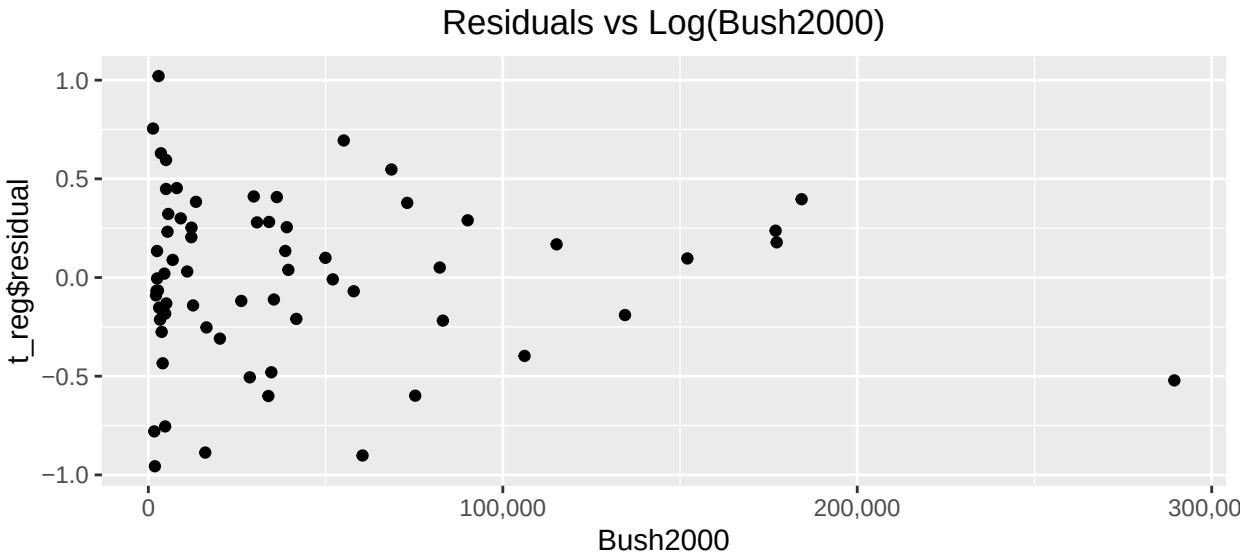


Part 2

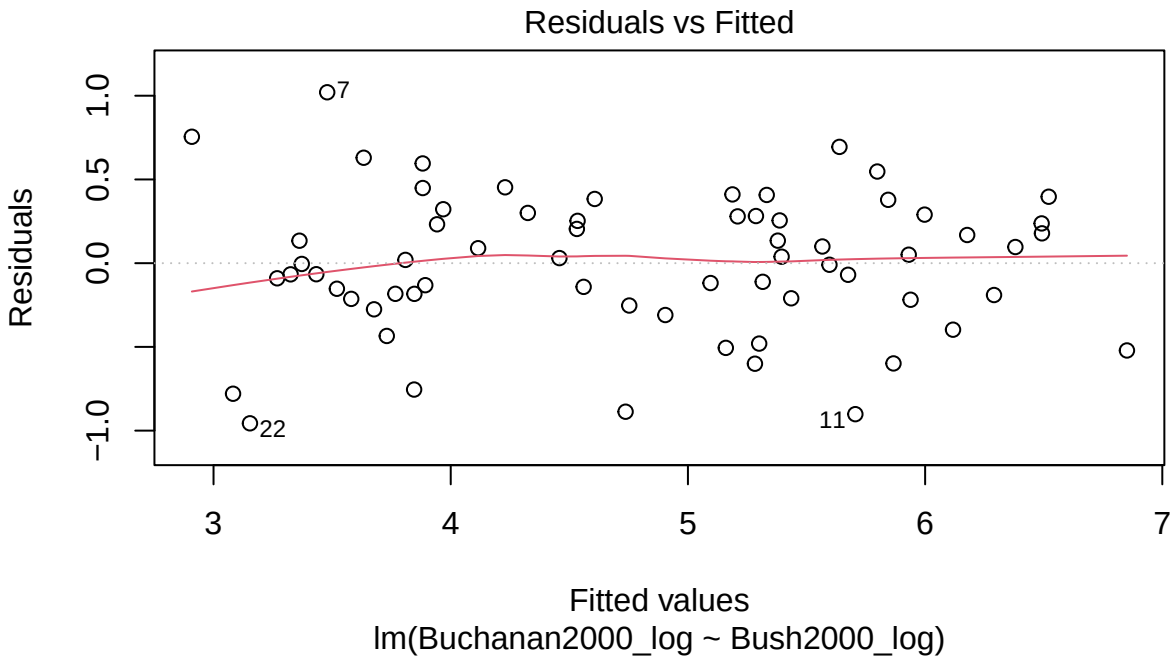
The double log linear regression model was used to fit the data as it resulted in a lower Root Mean Square Error. Both x and y variables were transformed to the natural log scale as using them without transformation of these variables, we see violation of the linearity assumption.

Part 3

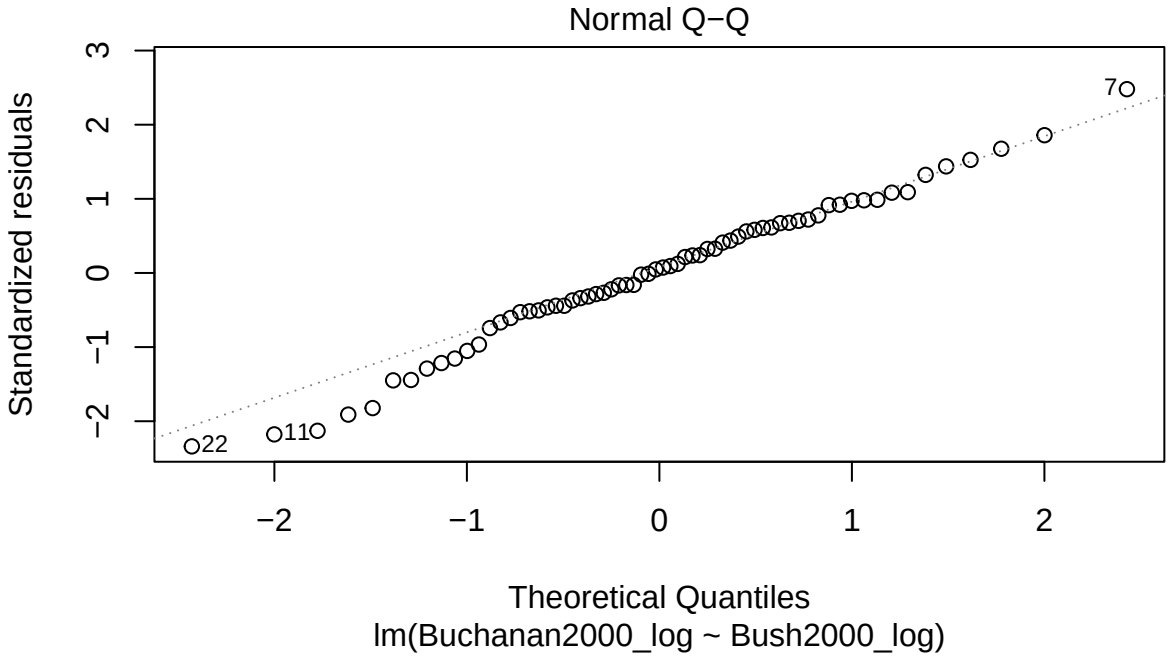
The transformation of both variables to a logarithmic scale resulted in a better fit for the model. If we see at the "Residuals vs Log(Bush2000)" graph below, we are unable to find a clear pattern. This model ssems to satisfy the linearity assumption better than the model without the tranformation of y variable to log.



We also notice that the residuals' variance remains somewhat constant and as scatter plot "Residuals vs Fitted" shows the points are randomly distributed, the independence assumption also seems to be not violated. However, we do need more data points to better asses these assumptions



Lastly, the QQ Plot below indicated that the normality assumption was not violated as well.



The table below indicated the results of the regression model. The coefficient suggest that a one percent increase in the votes received by Bush leads to 0.73% increase in the votes received by Buchanan.It can also be noted that the adjusted R-squared and F-statistic for the model are high.

Table 3: Double Log Linear Regression Results

	Dependent variable:
	Buchanan2000_log
Bush2000_log	0.73^{***}
	(0.04)
Constant	-2.34^{***}
	(0.35)
Observations	66
R^2	0.87
Adjusted R^2	0.86
Residual Std. Error	0.42 (df = 64)
F Statistic	413.02^{***} (df = 1; 64)
Note:	*p<0.1; **p<0.05; ***p<0.01

Part 4

The predicted interval corresponds to the fact that if a large sample of counties was selected and within those counties Bush received the same number of votes he received in Palm County, than in 95% of the counties the number of votes received by Buchanan would be within the prediction interval below.

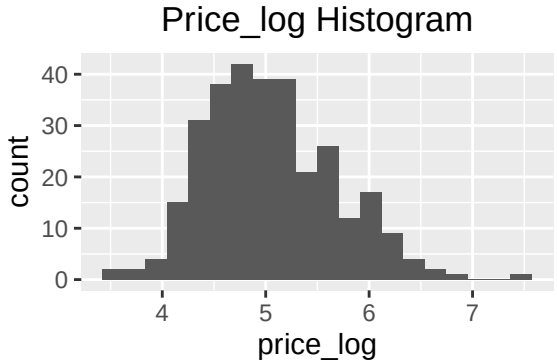
Table 4: Double Log Linear Regression Results

fit	lwr	upr
592.38	250.80	1,399.16

Question 3

Part 1

During the EDA, it was observed that the price variable was not normally distributed. Therefore, I took the log of price which had a normal distribution.

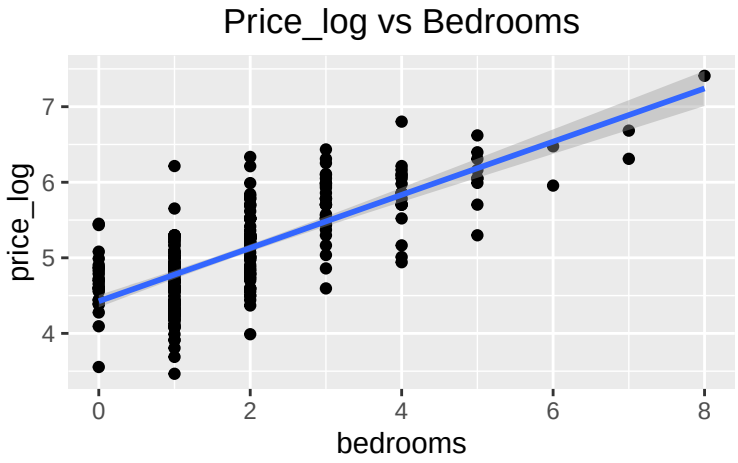


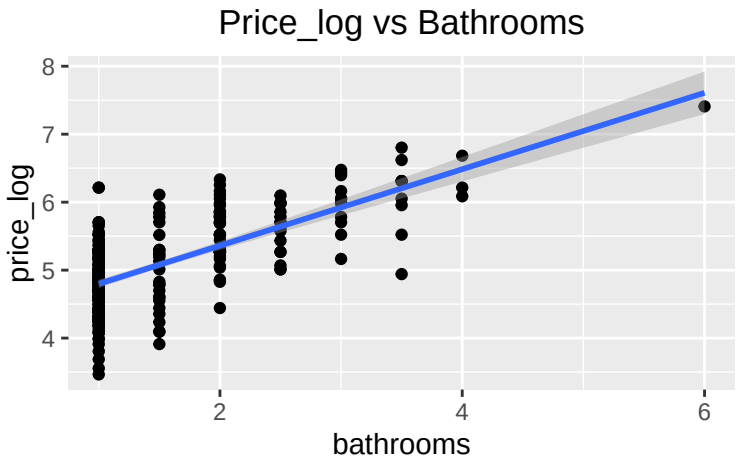
After that, we run a multiple linear regression on the log_price variable. The results of which are shown below.

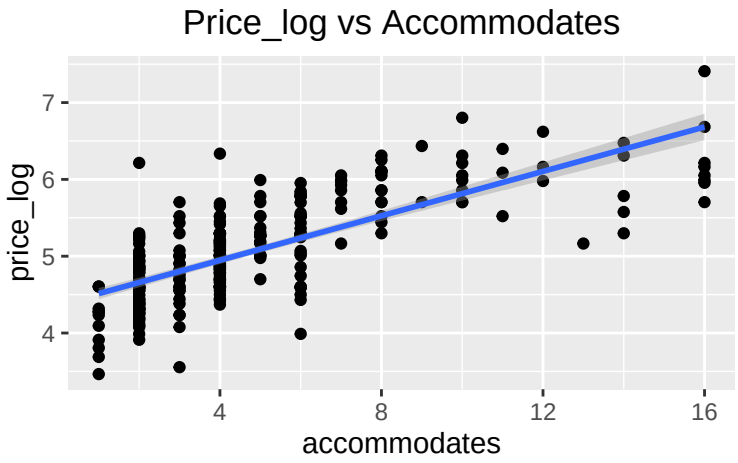
Table 5: log-Linear Regression Results

	Dependent variable:
	price_log
host_is_superhostTrue	−0.01
	(0.04)
host_identity_verifiedTrue	−0.10**
	(0.04)
room_typePrivate room	−0.45***
	(0.06)
room_typeShared room	0.27
	(0.26)
accommodates	0.04***
	(0.01)
bathrooms	0.21***
	(0.05)
bedrooms	0.13***
	(0.04)
Constant	4.43***
	(0.06)
Observations	305
R^2	0.67
Adjusted R^2	0.66
Residual Std. Error	0.37 (df = 297)
F Statistic	85.46*** (df = 7; 297)
Note:	*p<0.1; **p<0.05; ***p<0.01

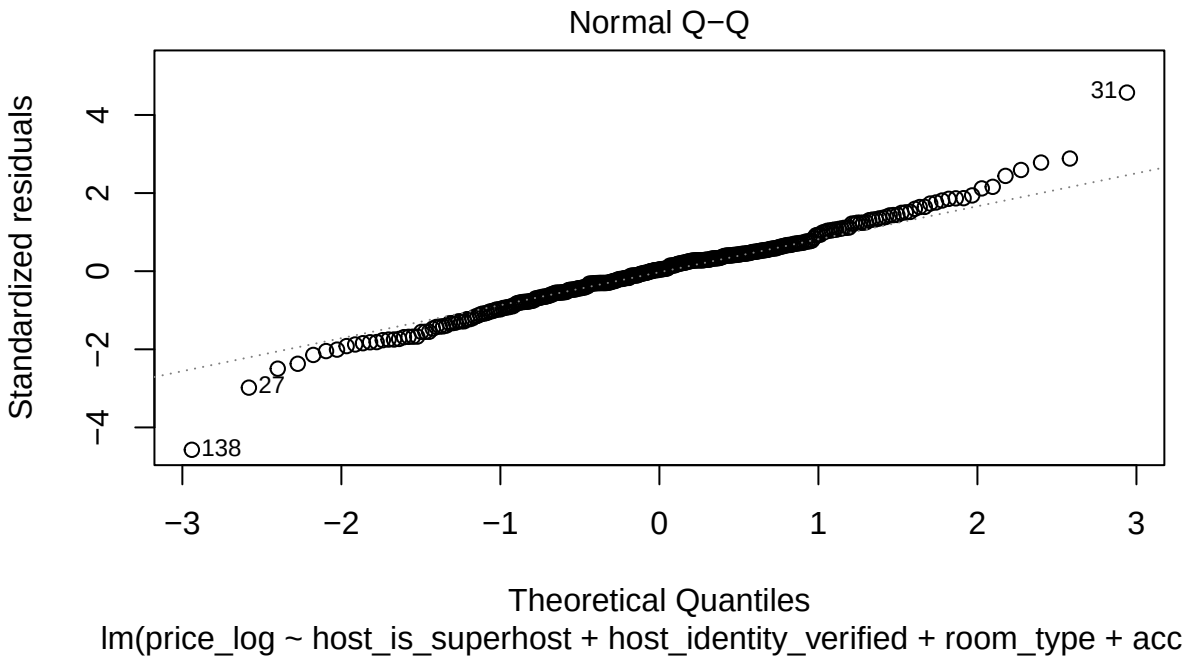
To test the linearity assumption, we create a scatter plot of the non-categorical predictors against log of Price which can be seen below. As integer variables behave like categorical variables, It is difficult to assess linearity using the "Residuals vs Predictor" plot. Therefore, we choose to use a normal scatter plot of predictors against our dependant variable.



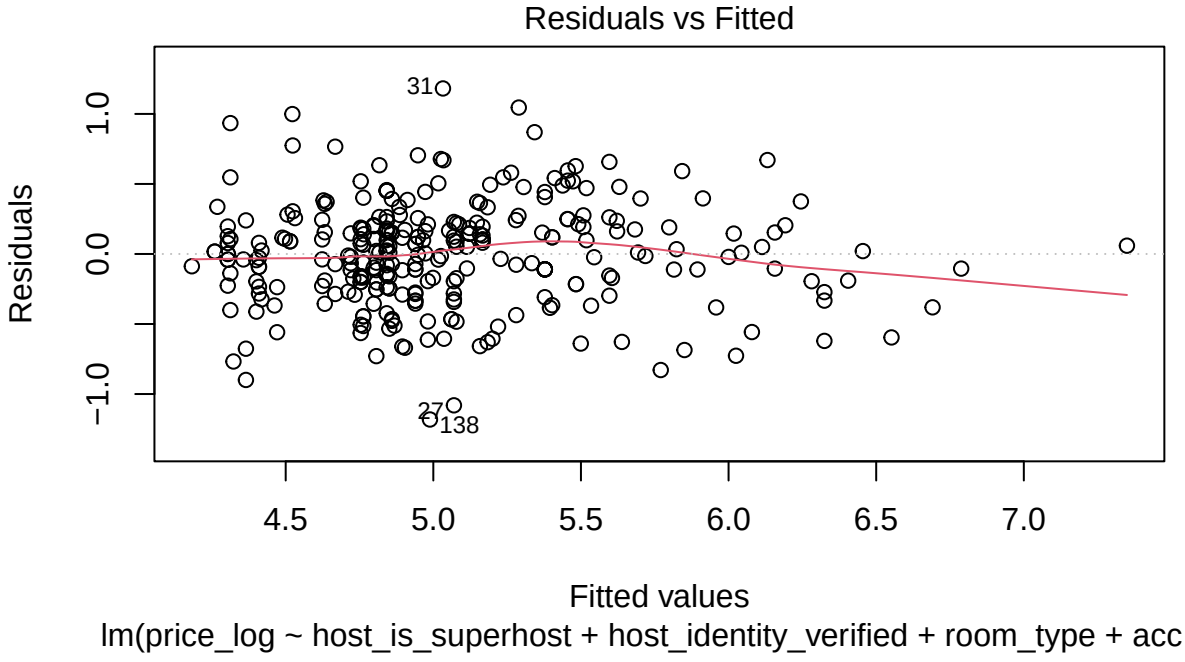




The above graphs indicate that the linearity assumption is somewhat satisfied. We use a QQ-Plot to check for the normality assumptions. The plot below indicates that the normality assumption is satisfied as well.



Lastly, we check for the independence and constant variance assumption by taking a look at the "Residuals vs Fitted" plot. Given that the points are randomly scattered, the assumption for independence has not been violated.



Part 3

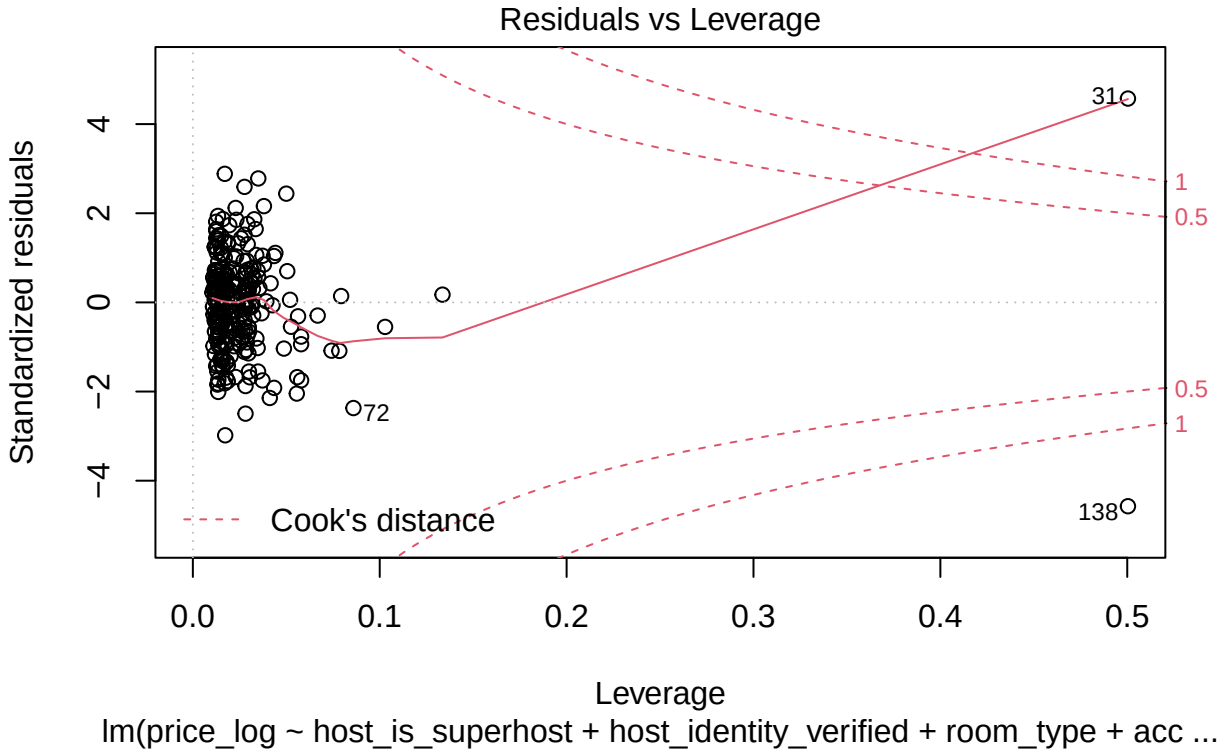
When host is a super host the price of the house on average seems to decrease by 0.9%. According to our model the price also decreases by 9.6% when the host's identity has been verified. As compared to houses with room type as 'entire home', the house with private rooms seem to be 45% cheaper and houses with shared rooms are 27% expensive. These seems to be counter intuitive results. One unit increase in accommodates leads to 4% increase in the price. Increase in one bathroom is associated with 13% increase in price and increase in one bedroom leads to a 21% increase in price. We assume that all other variable are held constant while making interpretation of each coefficient. The p-values for the variables host_is_super host and room_type (shared_room) are large and are thus statistically insignificant.

Part 4

The model does contain influential points, outliers and leverage points. The point with leverage greater than 0.045 are the leverage points. This threshold has been calculated by using the formula below where p is the number of predictors and n is the number of data observations:

$$Thresh - hold = \frac{2 * (p + 1)}{n} \tag{2}$$

We observe the "Residuals vs Leverage" plot to identify these points.By looking at the the plot we see that there are several leverage points in our data that are not influential points. The plot indicates that point 31, 72 and 138 are outliers, however, the only the points 31 and 138 are influential points as they have a Cook's distance greater than 1. The plot is shown below.



Now we exclude the influential points from the data and re-run the regression.The results are as follows:

Table 6: log-Linear Regression Results

	Dependent variable:
	price_log
host_is_superhostTrue	−0.01
	(0.04)
host_identity_verifiedTrue	−0.09**
	(0.04)
room_typePrivate room	−0.46***
	(0.06)
accommodates	0.03**
	(0.01)
bathrooms	0.23***
	(0.04)
bedrooms	0.15***
	(0.04)
Constant	4.43***
	(0.06)
Observations	302
R^2	0.69
Adjusted R^2	0.68
Residual Std. Error	0.35 (df = 295)
F Statistic	109.68*** (df = 6; 295)
Note:	*p<0.1; **p<0.05; ***p<0.01

The major changes we note are in the p-values which have decreased a little bit. However, if we assume a 95% confidence interval, the variables host_is_superhost, house identity verified and room_type (shared) are still statistically insignificant. The p-value for the variable accommodates has increased as well. There are no major changes in the values of the beta coefficients after removal of outlines and influential points.

Part 5

There are 2 major limitations:

1. One Major limitation of the model here is that even the integer predictor variables here behave as categorical variables. This makes it difficult to test the assumption of linearity.
2. When we observe the "Residuals vs Fitted", we notice that the constant variance assumption has been violated by our model.