BAN 5733
Individual Exercise 3 (10 Points)

Ms. Green is a real estate investor you have previously contracted with to complete a report on AirBNB in her area. She loved the report so much that she has asked you to do some additional analysis for her. She is still looking for a data-informed method to buy rental property to make available on AirBNB.

Now she wants to know if the combination of the housing variables makes a difference in the price. She is trying to decide if all the variables really are important or just some of them. With this information, Ms. Green will be able to make informed decisions about which properties she should buy. As Ms. Green's analyst, you have been asked to examine the **AIRBNB_BOSTON.sas7bdat** data with the following data variables.

Variable Name	Description	Type
Listing_ID	Listing ID is the ID given to each individual property available on AirBNB for booking.	Categorical
Accomodates	The number of guests the listing can accommodate on each stay	Continuous
Bathrooms	Number of bathrooms available for use by guests	Continuous
Bedrooms	The number of bedrooms available to guests	Continuous
Beds	The number of beds the guests can use	Continuous
Price	The nightly price for the listing	Continuous

Your goal is to determine how the set of predictors relate to price. She has asked you (as an analyst) to report your findings in an understandable way as she is neither a statistician nor an analyst. You need to explain the results of the tests in non-technical terms with references to the technical output. Interpreting your output into meaningful business decisions and options is important to Ms. Green to make good actionable decisions.

You will be creating a report that is no more than 5 pages long with all supporting tables, charts and graphs included in the appendix with references to the appendices in the main body of the report.

Tasks:

1. Conduct an analysis to determine if the set of predictors alters the nightly rental price.
    a. Variables to use:
        i. Dependent – PRICE
        ii. Independent – ACCOMODATES, BEDROOMS, BATHROOMS, BEDS
    b. Create a null hypothesis and alternative hypothesis (2 points)
        i. Write these out in a way that is accurate and understandable to Ms. Green
        ii. Be sure to include your criteria for decision making
        iii. You may include a more technical hypothesis as a reference
    c. Test your hypothesis with a multiple regression analysis. (2 points)
        i. Any supporting code or screen shots should be in an appendix
    d. Verify the assumptions have been met for the test (2 points)
        i. Make sure you include information to discuss each of the assumptions for regression.

2. Summarize the results from the multiple regression analysis so Ms. Green is able to make appropriate decisions during her property buying process. Include appropriate graphs and charts to support your summary in an appendix and reference within the text appropriately. Make sure you reference as necessary in your report. (4 points)

<u>Deliverables</u>:

- As you complete the exercise, create a report in Microsoft Word and in this report answer the questions in the exercise description.
- Copy and paste supporting tables/diagrams as needed to an appendix to justify any of your answers.
- Make sure you *print your name, student ID#, student email on the cover page* of the report and turn-in the report as communicated by your instructor.
- Please also put a running *header/footer with your name, on each page of your exercise* solution report.
- Failure to follow these instructions will result in deduction of points

Introduction

The data provide provided will be used in an attempt to see which variables are significant in predicting the rental prices for property that AirBnB customers are willing to pay for a rental. The data includes 3,273 observations for individual properties that will be used to predict the price of a rental. Price will be the response variable in the analysis. The data set provides four variables that will be tested as potential predictors for the response variable, price. The predictor variables are as follows:

- Accommodates: The number of guests the listing can accommodate on each stay
- Bathrooms: Number of bathrooms available for use by quest
- Bedrooms: The number of bedrooms available to quest
- Bed: The number of beds the guest can use

Statistical Testing

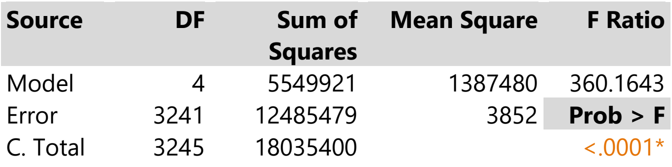
To determine the significance of a predictive equation for price, we will form a statistical test (hypothesis test) that will be used to determine the significance of the predictive variables effect on the response variable. Our null hypothesis, the test that we are trying to prove is incorrect, will be that the predictor variables do not significantly explain the variability in the response variable, price. The alternative hypothesis, the test which is opposite from the null, will be that at least one of the predictor variable explains a significant amount of variability in the estimation of price. A more technical way to state these test is as follows:

- Null Hypothesis: Price = β_0; *All non-intercept $\beta = 0$*
- Alternative Hypothesis: Price = $\beta_0 + \beta_1$*Accommodates + β_2*Bathrooms+ β_3*Bedroom + β_4*Bed; *At least one non-intercept $\beta \neq 0$*

Since we will be considering multiple variables, multiple linear regression will be used to determine a formula which best estimates the price an AirBnB customer is willing to pay based on the features of the property. A significance level of $\alpha = 0.05$ will be used as a cutoff for statistical significance in the multiple regression and hypothesis testing.

Results

To validate the result of the multiple regression model we will use the results in the Analysis of Variance table (Table 1). The first step is to confirm that the overall model has statistical significance and whether or not we will reject the null hypothesis.



Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	5549921	1387480	360.1643
Error	3241	12485479	3852	**Prob > F**
C. Total	3245	18035400		<.0001*

Table 1: A JMP output for the Analysis of Variance.

Referring to Table 1, the column of Pr > F, indicates the overall statistical significance of the model. The model has a probability of less than 0.0001 of selecting the F-value which is used to decide the statistical significance of the model. Since the value is less than our cutoff value of 0.05, we conclude that the model has statistical significance. Therefore, we would reject our null hypothesis and conclude that at least one of the predictor values significantly effects the variability of the response variable.

Testing the model for meeting regression assumptions is important. Referring to the Appendix 1, under the Diagnostic Plots for Price, we refer to the Residual plot and the Q-Q plot to validate our regression assumptions. From these plots we see that there may be some issues with normality of the data. There is an inherent pattern to the residuals. Looking at the distributions of all the independent variables shows a significant amount of skewness in the data (Figure 1). Correlation analysis also indicates some pretty significant correlations among the predictor variables (Table 2).

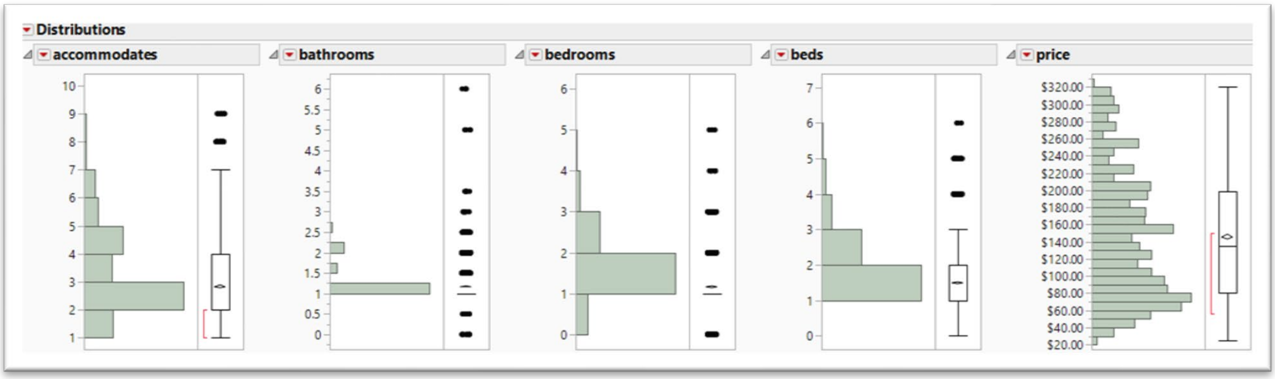
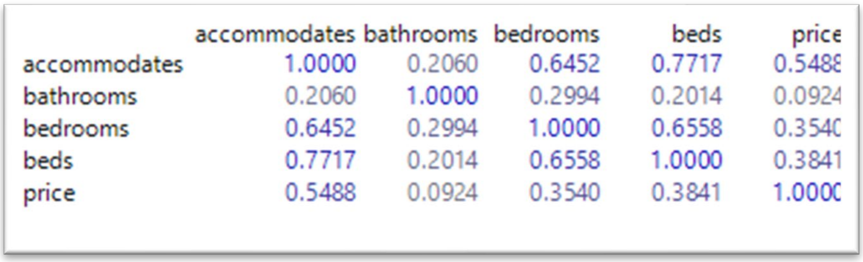


Figure 1: Histograms for All Variables



	accommodates	bathrooms	bedrooms	beds	price
accommodates	1.0000	0.2060	0.6452	0.7717	0.5488
bathrooms	0.2060	1.0000	0.2994	0.2014	0.0924
bedrooms	0.6452	0.2994	1.0000	0.6558	0.3540
beds	0.7717	0.2014	0.6558	1.0000	0.3841
price	0.5488	0.0924	0.3540	0.3841	1.0000

Table 2: Correlation Analysis of Variables

Summary

The multiple regression formula generated to predict the response variable is as follows:

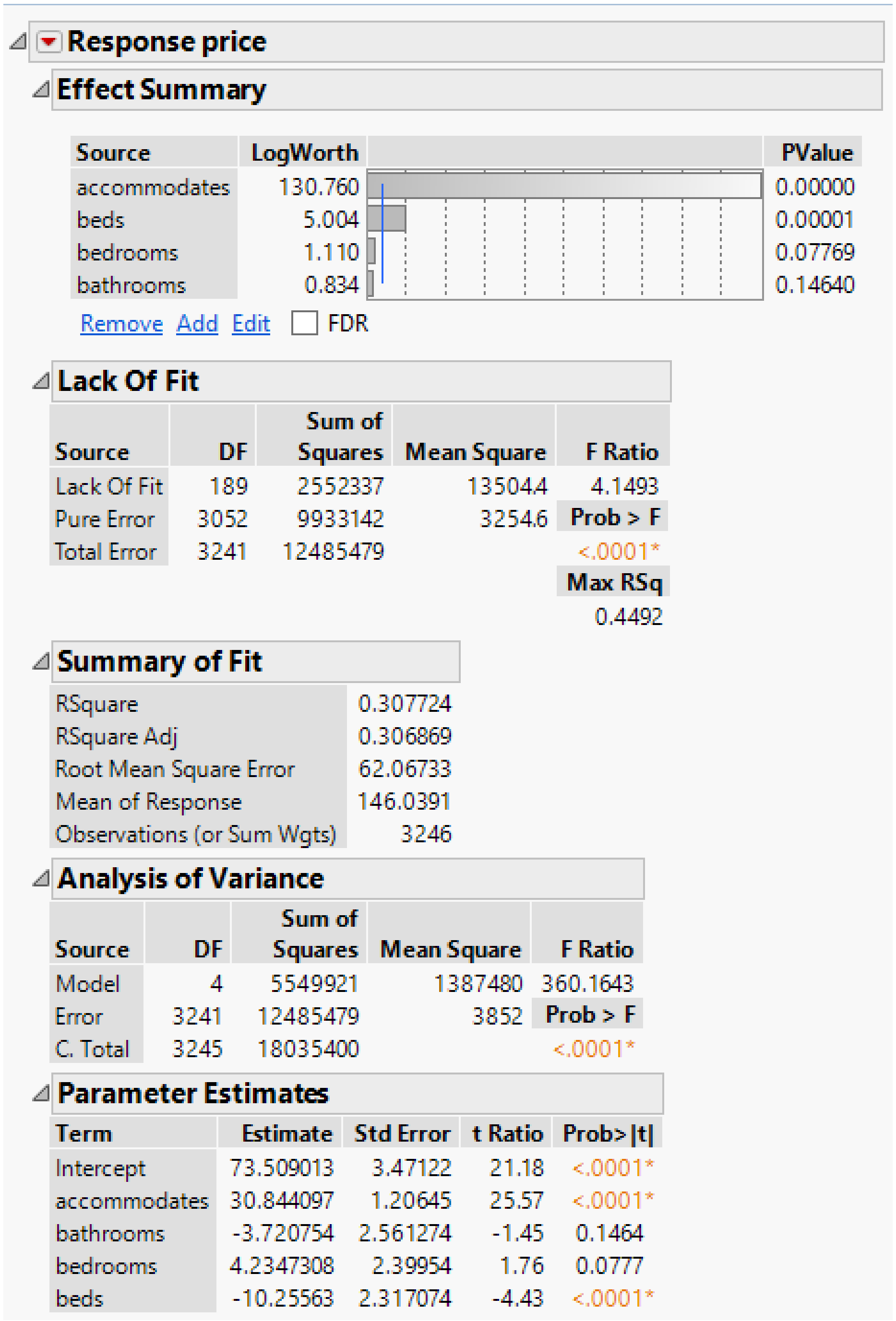
Estimated Rental Price = $73.51 + $30.84*Accommodates – $10.25*Beds

The regression formula shows that only two of the four predictor variables are significant in their effect on rental prices. The two predictor variable, accommodates and bed, are the only two significant predicts that can be used to estimate rental prices. Referring to Appendix 1, above the parameter estimates, we see that the model has an Adjusted R-Squared of 0.31. This indicates that the given predictor variables account for approximately 31 percent of the variability of rental prices. This value is slightly low indicating that there could be other variables which effect the rental prices of properties.

An additional explanation of the weak results could be the violation of linear regression assumptions. Standardizing, removing outliers or reducing the number of redundant variables could solve the issues stated about multi-collinearity and distributions. There could also be interactions between the variables that may cause a poor fit without them being in the model.

Overall, the prediction value is slightly weaker than ideal. It is my opinion that this model and formula serves as a starting point for estimating rental prices. Going forward, further analysis should be conducted on some of the issues mentioned in relation to assumptions and Ms. Green should gather additional data using the same variables as well as new variables like quality and review information, which may serve to generate a more predictive model.

Appendix 1



Response price

Effect Summary

Source	LogWorth		PValue
accommodates	130.760		0.00000
beds	5.004		0.00001
bedrooms	1.110		0.07769
bathrooms	0.834		0.14640

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	189	2552337	13504.4	4.1493
Pure Error	3052	9933142	3254.6	Prob > F
Total Error	3241	12485479		<.0001*
				Max RSq
				0.4492

Summary of Fit

RSquare	0.307724
RSquare Adj	0.306869
Root Mean Square Error	62.06733
Mean of Response	146.0391
Observations (or Sum Wgts)	3246

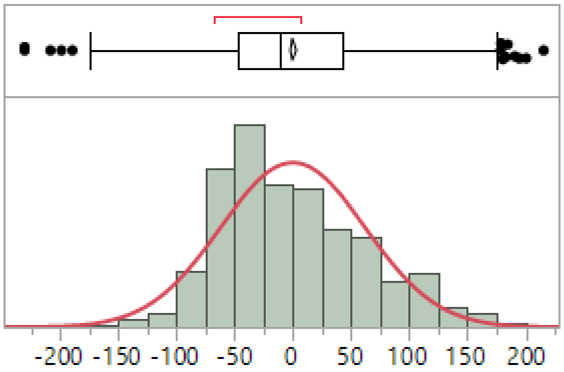
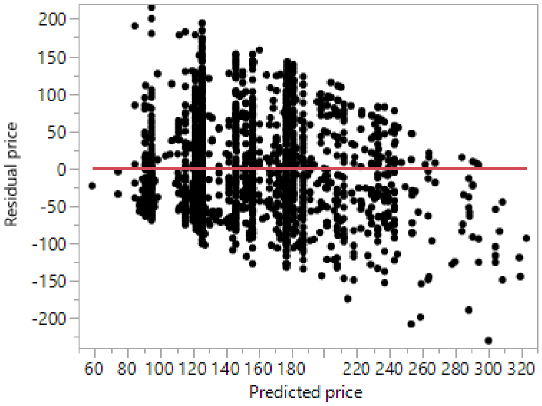
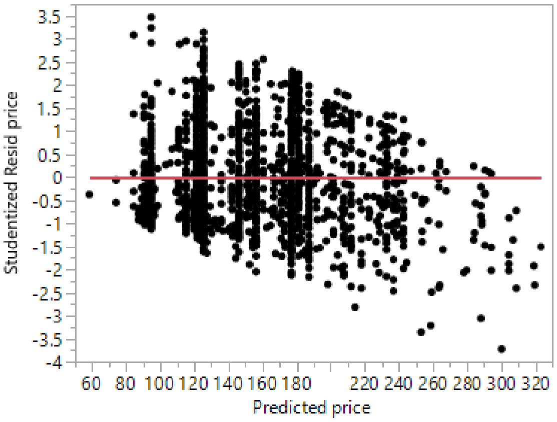
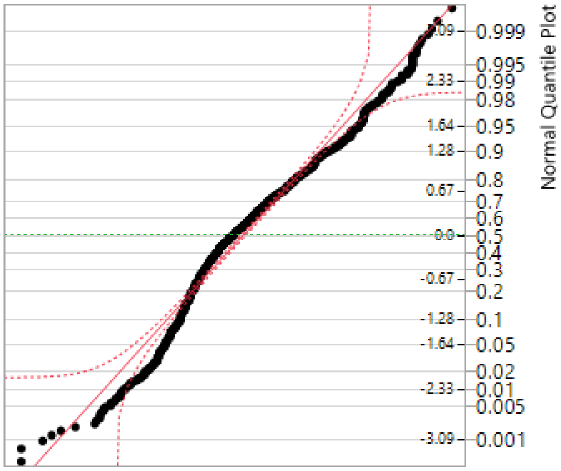
Analysis of Variance

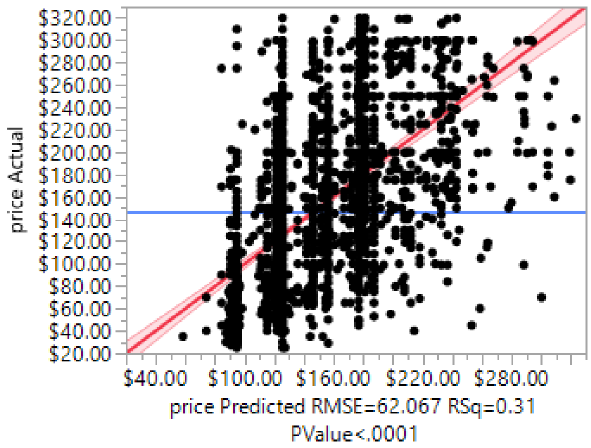
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	5549921	1387480	360.1643
Error	3241	12485479	3852	Prob > F
C. Total	3245	18035400		<.0001*

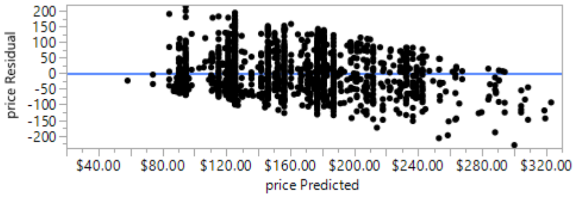
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	73.509013	3.47122	21.18	<.0001*
accommodates	30.844097	1.20645	25.57	<.0001*
bathrooms	-3.720754	2.561274	-1.45	0.1464
bedrooms	4.2347308	2.39954	1.76	0.0777
beds	-10.25563	2.317074	-4.43	<.0001*

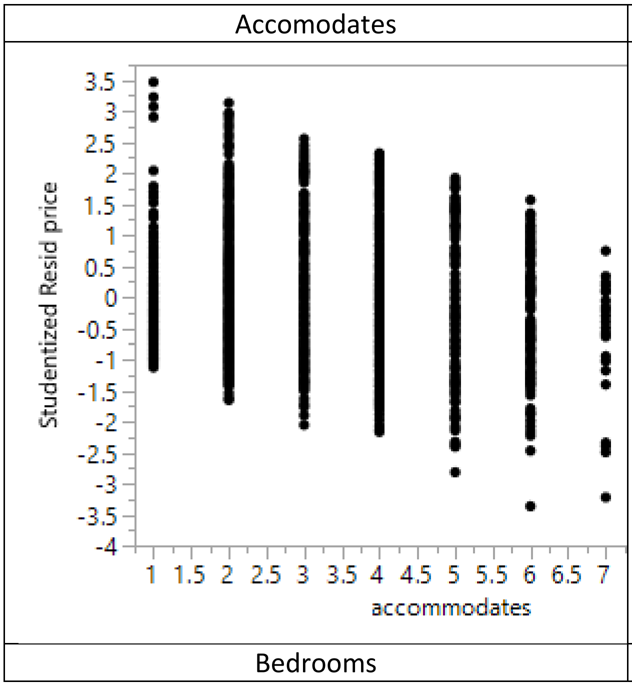
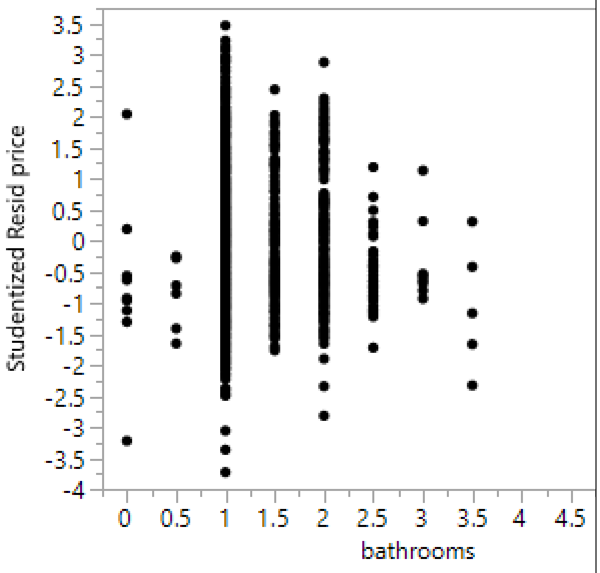
Diagnostic Plots for Price

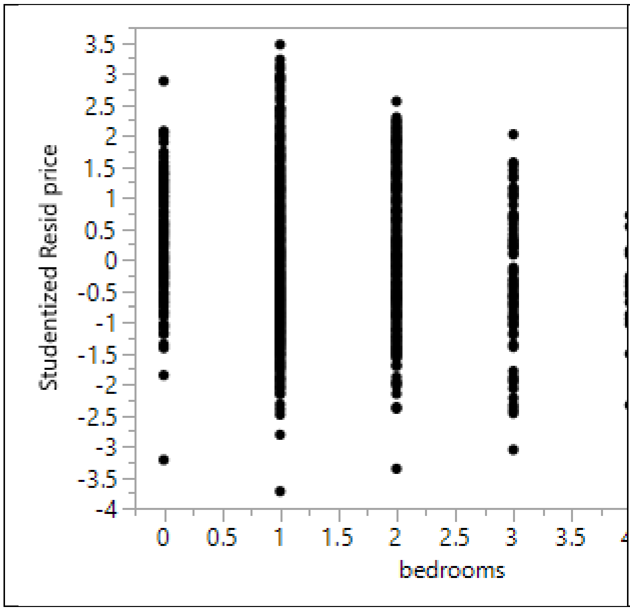
Distribution of Residuals for Price	Residual by Predicted for Price
	
Studentized Residuals by Predicted for Price	Q-Q Plot of Residuals for Price
	
Predicted Price by Price	Predicted Price by Residual

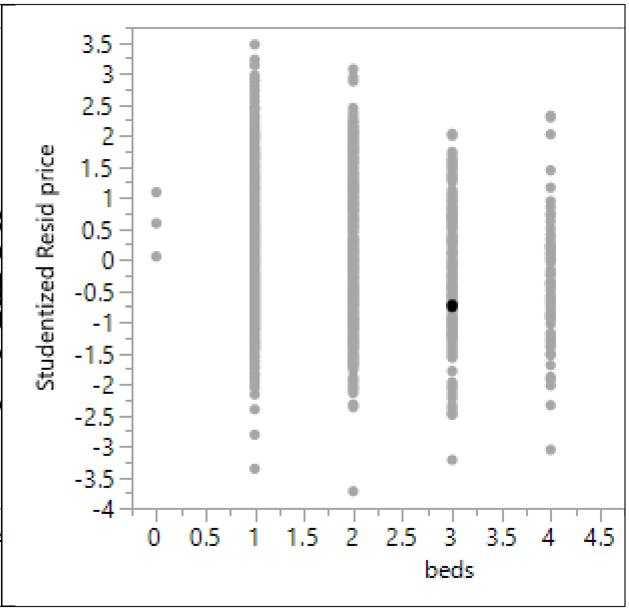




Residuals by Regressors for Price

Accomodates	Bathrooms
	
Bedrooms	Beds


Studentized Resid price
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
0
0.5
1
1.5
2
2.5
3
3.5
4
bedrooms


Studentized Resid price
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
beds