**Homework 3. Splitting Training/Validation Data**

Author: Rohit V. Akole

UConn ID: 3133252

University of Connecticut

OPIM5604 Predictive Modeling

Professor: Pankaj Prakash

Abstract

This homework is about the West Roxbury Housing Raw Dataset and Lending Case Study. SAS JMP is used as statistical analytics software. Splitting dataset into Training and Validation dataset. Build a multi-linear regression plot to check the integrity of the model and use the final model to validate the data in the validation model.

Table of Contents

[Question 1 4](#_Toc145770402)

[A. Total Value by Tax 4](#_Toc145770403)

[B. Total Value by Gross Area 5](#_Toc145770404)

[C. Total Value by Living Area 5](#_Toc145770405)

[D. Total Value by Bedrooms 6](#_Toc145770406)

[Question 2 7](#_Toc145770407)

[A. All iterations of the Multiple Linear Regression Model 7](#_Toc145770408)

[B. Bedroom vs. Total Value 12](#_Toc145770409)

[C. Best predictor of Total Value 13](#_Toc145770410)

[D. Remodeling vs. Total Value 13](#_Toc145770411)

[Question 3 14](#_Toc145770412)

# Question 1

Q1. A scatterplot is an x-y graph that is used to investigate the relationship between two quantitative variables, one on each axis. From the West Roxbury Housing Raw Dataset, describe the relationship of 4 different variables (on the X-axis) with the Total Value of the property (On the Y-axis) using the 4 different scatter plots created in JMP.

## Total Value by Tax

|  |  |
| --- | --- |
| Figure 1. Total Value vs. Tax | Figure 2. Total Value vs. Tax Heatmap |

There is a positive collinearity between Total Value and Tax. As the total value increases, tax increases as well. There are no outliers in Total Value vs. Tax. As per the heatmap description, most of the houses are between 300k-400k Total value, and Tax for them is between 4000-5000. There is a strong upward relationship between these 2 variables.

## Total Value by Gross Area

|  |  |
| --- | --- |
| Figure 3. Total Value vs. Gross Area | Figure 4. Total Value vs. Gross Area Heatmap |

In the above scatter plot, the trend is upward. As the gross area increases, the house's total value is also increasing. There are some outliers, but one is way off the normal trend between 7000 – 7500. As we can see from the heat map, the biggest cluster of houses is between the gross area of 2000 – 3000 and the total value for those houses is between 300k – 400k.

## Total Value by Living Area

|  |  |
| --- | --- |
| Figure 5. Total Value vs. Living Area | Figure 6. Total Value vs. Living Area Heatmap |

The trend for Total Value by Living Area is like the total value by gross area. The scatter plot suggests a little more inclination as compared to Figure 3. The data points are also more clustered in Figure 5.

The heatmap results are also like Total value vs. Gross Area heatmap. But the Heatmap in Figure 6 is more dispersed between Living Area 1000 – 2000 and Total Value between 200k – 500k. Unlike the Figure 4, there are outliers in Figure 6 both above and below the regression line between the living area 4000 – 5000.

## Total Value by Bedrooms

|  |  |
| --- | --- |
| Figure 7. Total Value vs. Bedrooms | Figure 8. Total Value vs. Living Area Heatmap |

As per the scatter plot, the trend between Total Value by Bedrooms is positively (upward) correlated. As the number of bedrooms increases the total value of the houses also increases. As per the heatmap, the greatest number of houses in this dataset has 3 bedrooms and the total value of those big clusters is between 300 – 400. The total value of the houses is more spread at the 4-bedroom houses. The outlier lies in the 5-bedroom house at a little over 1200k.

# Question 2

Q2. From the West Roxbury Housing Raw Dataset – use the process below to create a train / validation set (use seed 888).

## All iterations of the Multiple Linear Regression Model

|  |  |
| --- | --- |
| Figure 9. Multi-Linear Model 1 | Figure 10. Multi-Linear Model 2 |

At the 1st model in Figure 9, the model is overfitting. The R2 is 1 at the model 1. We can see that only the Tax variable has the greatest log-worth resulting in overfitting of the model. So in model 2, I removed Tax and the R2 is at 0.812936 and the Adjusted R2 is at 0.812094. For the Training vs. Validation dataset, the RASE is at 42.578 vs. 43.141 respectively after splitting the dataset in 50% - 50% and keeping the seed at 888. We can see that the p-value for 3 variables is above 0.05, in which, the bedroom has the highest p-value. So, we will remove the bedrooms in the next model.

|  |  |
| --- | --- |
| Figure 11. Multi-Linear Model 3 | Figure 12. Multi-Linear Model 4 |

After removing the bedrooms, the p-value for the Rooms increased significantly in Model 3. R2 and R2 Adjusted reduced slightly and RMSE increased slightly.

So, in Model 4, I removed Rooms as the p-value for the rooms was highest and Log-Worth was lowest. Now the only variable above 0.05 p-value left is YR Built. Other values for Model 4 didn’t change much. However, the p-value for the Intercept is still higher than 0.05. So, from Model 4, I removed the YR Built.

|  |  |
| --- | --- |
| Figure 13. Multi-Linear Model 5 | Figure 14. Multi-Linear Model 6 |

After removing the YR Built from Model 4, now all the p-values are under 0.05 in Model 5. The p-value for the Kitchen is close to but still under 0.05. After taking a glance at VIF in Model 5, it is greater than 4 for Living Area and Gross Area, 7.697 and 6.145 respectively. Since the Living area has a greater VIF, I removed it from the model. Which stabilized the VIF for Gross Area in Model 6. But the p-value for the kitchen increased to 0.576 which is way more than 0.05. Since that’s the case, I removed the Kitchen variable from the Model 6.

|  |  |
| --- | --- |
| Figure 15. Multi-Linear Model 7 | Figure 16. Multi-Linear Model 8 |

In the Model 7, I removed Kitchen, resulting stable model with p-values for all the variables under 0.05, and VIF is under 4 for all variables as well. After playing around a little bit more, I found out I could add back Rooms without affecting the Model 7. In Model 8, it increased the R2 a little more as compared to Model 7 and RMSE also lowered. Once we add that, this is the final model we get for the West Roxbury Housing dataset. Now, we can change the Remodel from categorical variable into indicator columns everything else being equal, this will give us in-depth insights for the different remodeling indicators.

|  |  |
| --- | --- |
| Figure 17. Multi-Linear Final Model | Figure 18. Prediction Expression |

For the final model, I have created the dummy variables/indicator columns for the categorical variable Remodel. We also got the Prediction Expression for the final model as seen in Figure 18. All the variables are stable in the final model in Figure 17. The difference between R2 is also nominal in cross-validation. Also, the RASE difference is acceptable.

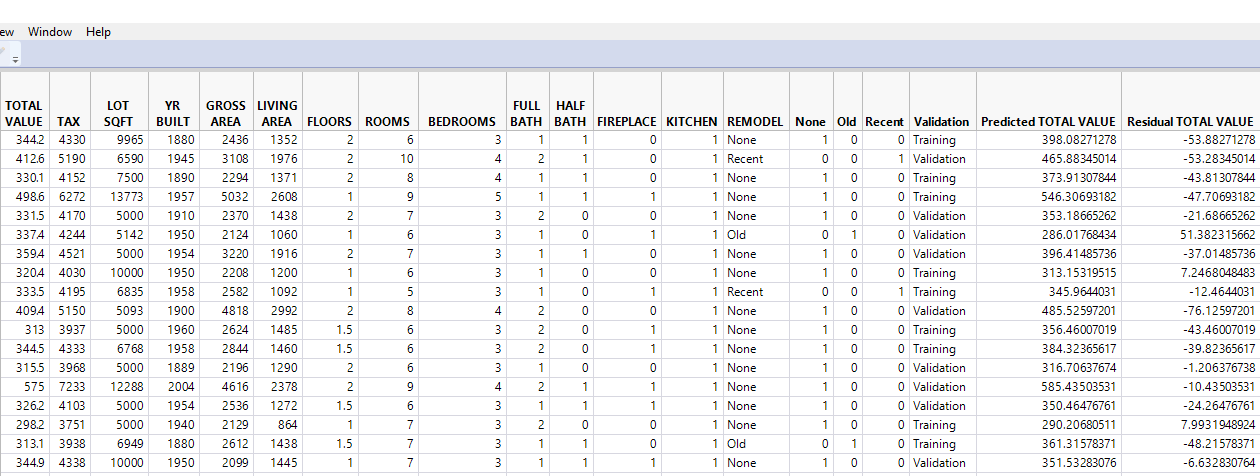


Figure 19. West Roxbury Dataset including Predicted and Residual for Total Value from Final Model

Using the Residual Total Value, we can see the difference in Actual Total values and predicted Total Value.

## Bedroom vs. Total Value

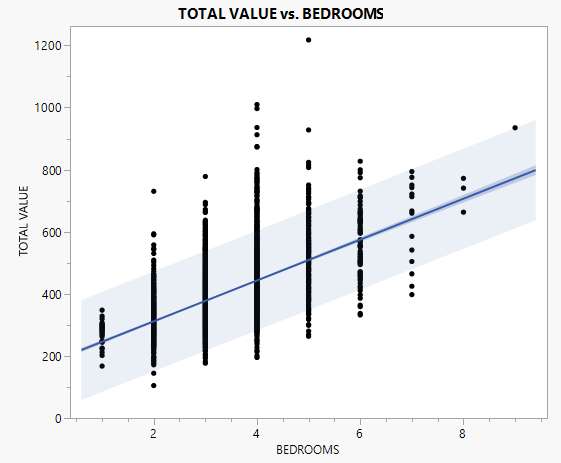


Figure 20. Bedroom Vs. Total Value

The relation between Bedroom vs. Total value shows an upward trend. As the number of bedrooms is increasing, the value of the houses is also increasing. Also, with each bedroom increase, the Total value spread is also larger. Also, the number of samples are reducing as the 5+ bedroom. As per the Multilinear regression Model, for every increase in a bedroom, the total value of the house decreases by 2,089. So, it was inversely related to our model. But the Scatter plot shows it differently where the increase in the number of bedrooms also increases the price.

## Best predictor of Total Value

As per the final model, the Gross Area of the house is the best predictor with a Log-Worth of 247.989, followed by Lot SQFT with 138.448 Log-Worth. For each gross floor SQFT increase, the house value is increasing by $52. Similarly, with each SQFT increase in Lot SQFT, house value increases by $9.

## Remodeling vs. Total Value

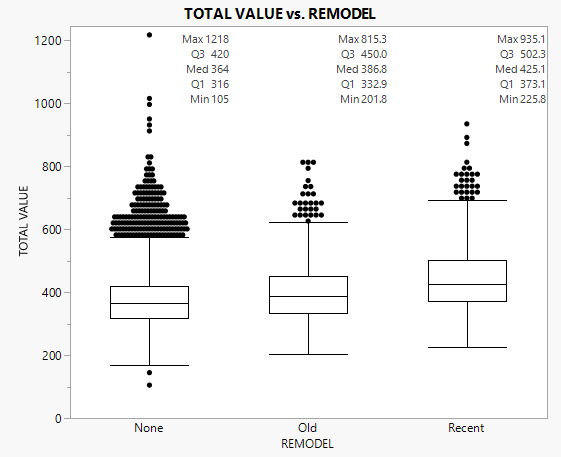


Figure 21. Remodeling vs. Total Value Box plot

Looking at the box plot, the total value for the houses with no remodeling is less than recently remodeled. With the above observation, no remodeling hurts the value of the houses as compared to recently remodeled and old remodeled houses. The median value of the houses with no remodeling is 364k vs. the median value of recently remodeled houses is 425.1k. Also as per the final model, the houses with no remodeling have $27,287 less total value than the houses that are recently remodeled.

# Question 3

Q3. From the Lending Case Study provided with this homework, answer the following questions. Answers to these questions should reflect your reasoned opinion.

## What decision will Jasmin need to make?

The first basic question will be whether she should make an investment in peer-to-peer lending. If yes, then Jasmin will have to decide which platform to use. Once she decides on the platform, then she will have to decide how much money she should invest in peer – to – peer lending as she wants to diversify her savings. Jasmin will have to decide whether to split data 2 ways or 3 ways depending on the size of the data. She will have to decide how much data she wants to keep for training, testing, and holding.

## What is Jasmin's objective when making these decisions?

Jasmin wants to diversify her portfolio to other than traditional investments like stocks, bonds, etc. She wants to incur high returns on this investment. Make a model which will predict the trend for each grading scheme, comparing it to all the relevant variables to get the final model with p-value less than 0.05 and VIF less than 4.

## What kind of metrics Jasmin might want to predict from this dataset?

Once Jasmin decides to invest in it, she will have to decide on predictors she should look at to decide which grading scheme will give her greater returns with less chance of the loan getting charged off. She will use the Grading Scheme column to identify the best grade to get higher returns as well as minimize the risk and compare it with Loan status, Interest Rate, Borrower attributes like home ownership, annual income, current income to debt ratio (if available), previous loan defaults (if available). She should be looking at the grading scheme for each number of claims where debts are currently in a 121-day period.

## How would you categorize the variables describing the loans in the dataset?

I would look at the Grading scheme as a categorical variable. I would compare it with the average interest rate per grading scheme. Number of claims charged off vs. fully paid in each grading scheme. The number of claims is already in 121-day period for each grading scheme. I would look for the data types of the variables to get the appropriate information from the dataset. I would analyze the data for the data integrity issues as well.