Joyce Woznica

jlwoznic@syr.edu

Joyce Woznica  
HW1

House Price Analysis

Contents

[Introduction 3](#_Toc5176739)

[The Data 3](#_Toc5176740)

[Data Categorization 3](#_Toc5176741)

[Data Visualization 4](#_Toc5176742)

[Correlation Analysis 4](#_Toc5176743)

[Quantitative Variables Only 4](#_Toc5176744)

[All Variables 5](#_Toc5176745)

[Initial Regression Analysis 6](#_Toc5176746)

[Prediction Model and Sensitivity Analysis 6](#_Toc5176747)

[Conclusions 6](#_Toc5176748)

List of Figures

[Figure 1: Average Housing Price by Construction Type and Neighborhood 3](#_Toc5176749)

[Figure 2: Average Square Foot by Construction Type and Neighborhood 3](#_Toc5176750)

[Figure 3: Average Housing Price and Square Feet by Construction Type and Neighborhood 4](#_Toc5176751)

[Figure 4: Average Housing Price and Square Feet by Neighborhood and Construction Type 4](#_Toc5176752)

[Figure 5: Initial Correlation of Housing Prices Data 5](#_Toc5176753)

[Figure 6: Housing Dataset with Type of Construction and Neighborhood as Quantitative Variables 5](#_Toc5176754)

[Figure 7: Correlation on Full Housing Dataset 6](#_Toc5176755)

# Introduction

In this exercise, we will be using the house prices data provided () to determine the factors which influence the price of a home. To do this we will be executing the following:

* Data Categorization
* Pivot Tables for Visualization
* Correlation Analysis
* Initial Regression Analysis
* Prediction Model and Sensitivity Analysis
* Conclusions

# The Data

Our house price data provides the following variables about each house data point collected:

* ID – a unique identifier
* Price – price of the home in dollars
* SqFt – square foot area of the home
* Bedrooms – number of bedrooms
* Bathrooms – number of bathrooms
* Offers – number of offers received on the home before the sales
* Brick – Yes/No on if brick construction
* Neighborhood – location of the home in east, west, north quadrants of the city

# Data Categorization

To initially review the data, I used Pivot Table to review the Price by type of construction (Brick: Yes or No) and Average Square Feet by designed neighborhood. The first table represents reviewing prices within certain neighborhoods (East, West, North) based on the construction type.

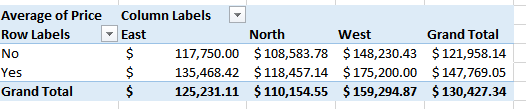


Figure 1: Average Housing Price by Construction Type and Neighborhood

|  |  |  |  |
| --- | --- | --- | --- |
| *Average Square Feet of House* | Brick | |  |
| Neighborhood | *No* | *Yes* | **Grand Total** |
| *East* | 2001.54 | 2031.05 | 2014.00 |
| *North* | 1928.11 | 1857.14 | 1916.82 |
| *West* | 2073.48 | 2091.25 | 2080.77 |
| Grand Total | **1989.19** | **2025.00** | **2000.94** |

Figure 2: Average Square Footage by Construction Type and Neighborhood

The next Pivot Table adds the average square feet into the table reflecting the average square feet of the homes in the particular neighborhood areas and their average price by construction type.

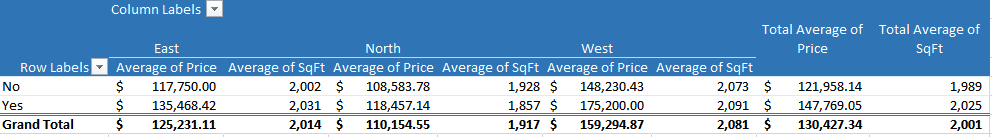


Figure 3: Average Housing Price and Square Feet by Construction Type and Neighborhood

The following table shows the same information as Figure 2, but listed by neighborhood on the left-hand side and then by Construction Type and the average house price and average square feet.

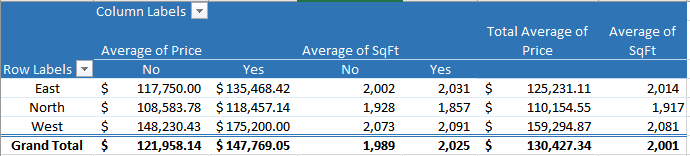


Figure 4: Average Housing Price and Square Feet by Neighborhood and Construction Type

# Data Visualization

Now that we can see the details about the housing prices in certain neighborhoods by both average square feet as well as type of construction.

# Correlation Analysis

It is important to review and determine if there are any correlations between the input quantitative variables (square feet, number of bedrooms, number of bathrooms, offers) and the dependent variable which is the housing price. We can make these determinations by looking at descriptive statistics on the data set provided.

Initially, I ran correlation statistics on only the quantitative (continuous) variables which were the following:

* Square Feet
* Number of Bedrooms
* Number of Bathrooms
* Offers
* Price

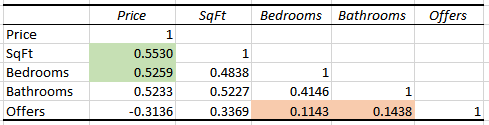


Figure 5: Initial Correlation of Housing Prices Data

As shown in the previous figure, the numbers (not the 1’s) in the table are our correlation coefficients. This information simply provides the relationship between two variables, but does not give us the statistical significance of that correlation (descriptive statistic) in the overall model.

To explain the data in this table, there is a largest magnitude correlation between square feet and price (0.5530) as well as number of bedrooms and price (0.5259). There is a smallest magnitude correlation between number of offers on the house and number of bedrooms (0.1143) as well as number of offers on the house and number of bathrooms (0.1438).

### All Variables

To show the influence and correlation of all variables, I converted the Neighborhood (East/West/North) to a quantitative variable and did the same for Type of Construction. For Neighborhood, a value of 1 was given for the neighborhood region (for example, if East, a 1 is entered in the East variable column and a 0 in the West and North columns. If North, a 1 in the North column and a 0 in the East and West columns, etc.). For Type of Construction, a 1 was given to Brick=Yes and a 0 was given to Brick=No. The data looked something like that shown in the following figure.

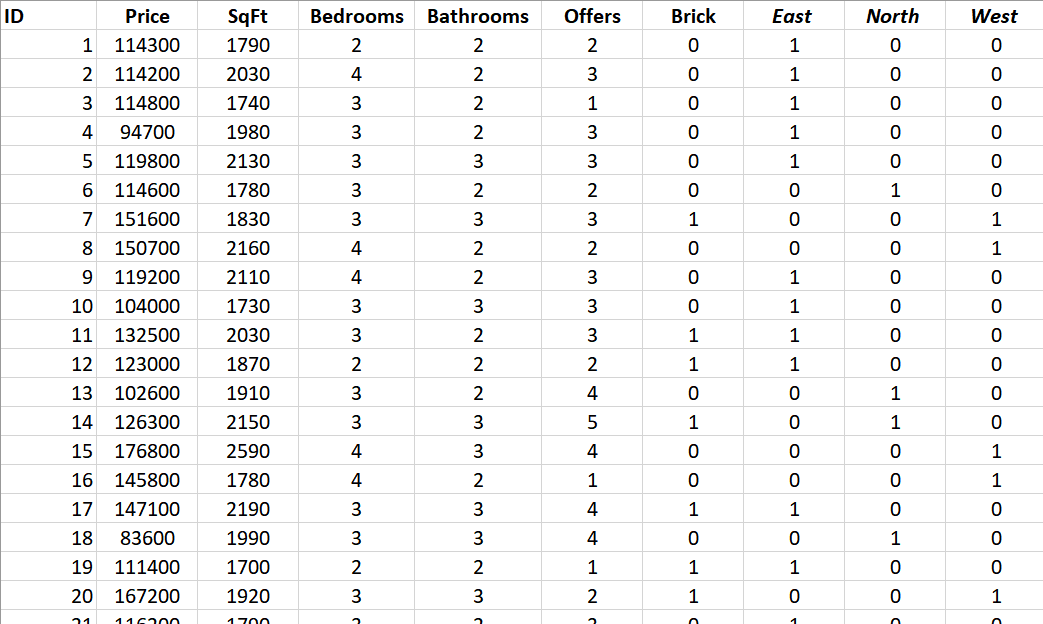


Figure 6: Housing Dataset with Type of Construction and Neighborhood as Quantitative Variables

This approach yielded different information as shown in the following figure.

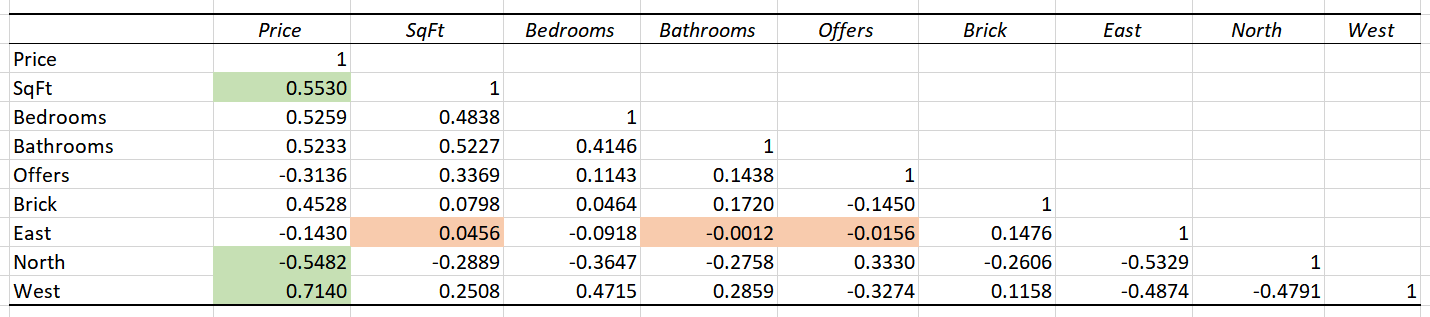


Figure 7: Correlation on Full Housing Dataset

To explain the data in this table, there is a largest magnitude correlation between square feet and price (0.5530) which is expected; however, larger magnitude correlations than that for number of bedrooms and price (0.5259) were found. These were the location of house in the North neighborhood and its relationship to price (-0.5482) and the location in the West neighborhood (0.7140). This would imply (without doing further analysis on the statistical significance) that the North neighborhood has less expensive homes (negative correlation) and that the West has more expensive homes.

There is a smallest magnitude correlation between number of offers on the house and number of bathrooms in the East neighborhood (-0.0012) and then the number of offers on the house in the East neighborhood (-0.0156). The next smallest magnitude correlation was the square footage of a home in the East neighborhood (0.0456).

As indicated, the largest magnitude correlations imply that there is a strong relationship between that variable and the other – for example, square feet and price. As the square footage increases, so does the price of the home as indicated by a large positive correlation coefficient of 0.5530.

There are several negative correlations which imply that the correlation between this variable the other decreases the values. For example, in the figure above, if the home is in the North neighborhood, this implies a decrease in the price as indicated by the -0.5482 correlation coefficient.

### Overall Correlation Conclusions

The information found as the result of running these correlations indicate that square footage and neighborhood has a large correlation on the price of the home. This, in my opinion, is intuitive, as we would expect to see that the higher the square footage, the more expensive the home. It is also intuitive that certain neighborhoods are more desirable which is also indicated by the correlations.

# Regression Analysis

Now that we have done some initial correlation and descriptive statistics on this housing data, we need to look at the statistical significance of certain independent variables and how that affects the dependent variable (price).

### Quantitative Variables Only

As before, the first regression was completed on the already quantitative variables. This is noted in the following figure.

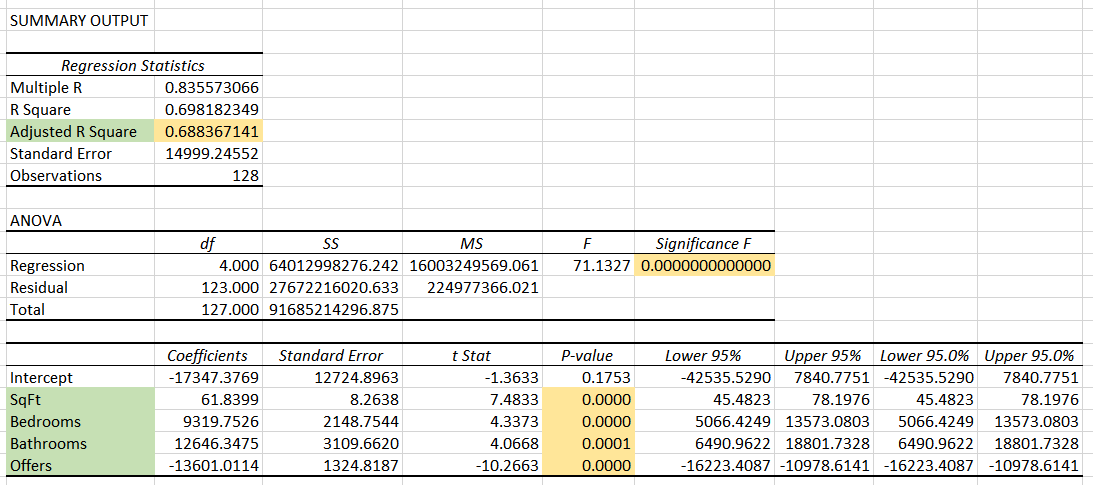


Figure 8: Regression on Quantitative Variables Only

Reviewing just the quantitative variables, we see a fairly strong Adjusted R Square (0.68836714) which is the appropriate R square to use if you have multiple independent variables. However, it is not as strong as I would like to see in a model. You will also see that the model shows a very small *Significance F* (well below an α = 0.05) in the ANOVA model which represents that the model is statistically significant.

Since all of the independent variables show a p-value of below the desired alpha of 0.05, this would indicate that all variables are significant in the prediction of price. The number represented in the Coefficients column shows how the price is adjusted (up or down) based on the value of that variable. For example, previous offers on the house brings the overall price down where higher numbers of bathrooms, bedrooms and higher square footage increase the housing price. I believe these coefficients are very intuitive.

### All Variables

Unfortunately, I ran into some errors with Excel around using all variables due to the p-values for both North and West neighborhoods being too small to display, as shown in the following figure.

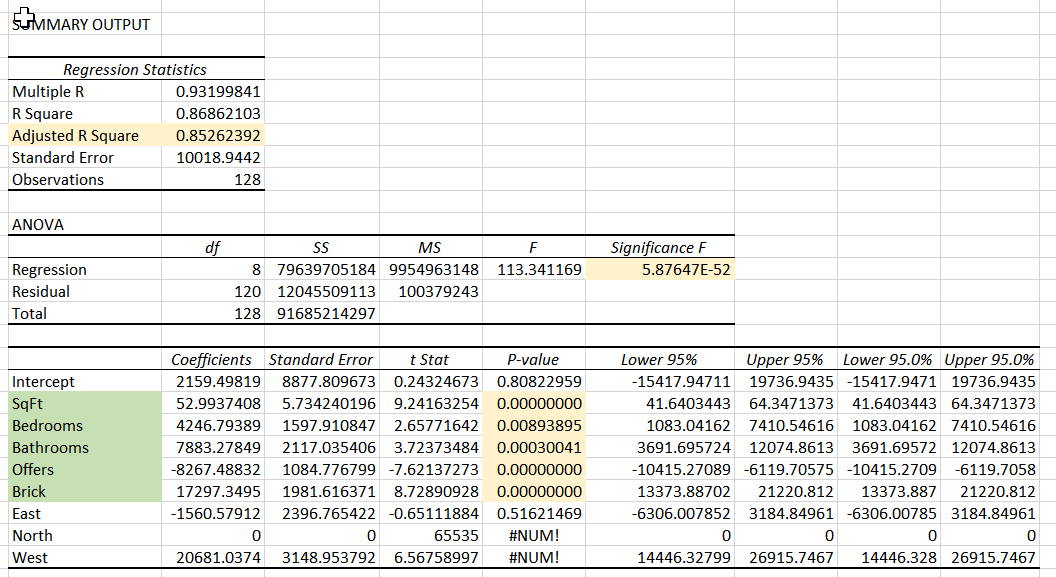


Figure 9: Regression on All Variables

I reran the model as follows:

* Removed North and West variables
* Removed North only
* Removed West only

# Prediction Model and Sensitivity Analysis

### Prediction Model

### Sensitivity Analysis

# Conclusions