House Price Analysis using correlation and regression  
  
Team 2

BB, RS, NK, Jennifer Mead

Contents

[Background 1](#_Toc505324666)

[1. Data categorization 1](#_Toc505324667)

[2. Pivot Charts 1](#_Toc505324668)

[3. Correlation analysis 2](#_Toc505324669)

[4. Regression Analysis 4](#_Toc505324670)

[5. Two-way sensitivity 5](#_Toc505324671)

[6. Conclusion 6](#_Toc505324672)

[Glossary 8](#_Toc505324673)

# Background

Using the House Prices data, the following analyses determine the factors which influence the price of a home based on the House prices dataset.

# Data categorization

* **Develop a categorization of your data using pivot tables. Develop two pivot tables of average price and average square feet by type of construction (brick) and neighborhood.**

The House average price and average square feet by construction type and neighborhood have been categorized using the pivot tables

|  |  |  |
| --- | --- | --- |
| **Average of Price** | **Brick** |  |
| **Neighborhood** | **No** | **Yes** |
| East | $117,750.00 | $135,468.42 |
| North | $108,583.78 | $118,457.14 |
| West | $148,230.43 | $175,200.00 |
| **Grand Total** | **$121,958.14** | **$147,769.05** |

**Interpretation**: The above pivot table shows the average prices based on the neighborhood location and based on the house construction (using brick or not). The 2 data points are good to compare as brick homes typically cost more for a given area and here we can view this trend. Also, West neighborhood has the highest prices for brick and non-brick houses.

|  |  |  |
| --- | --- | --- |
| **Average of SqFt** | **Brick** |  |
| **Neighborhood** | **No** | **Yes** |
| East | 2,001.54 | 2,031.05 |
| North | 1,928.11 | 1,857.14 |
| West | 2,073.48 | 2,091.25 |
| **Grand Total** | **1,989.19** | **2,025.00** |

**Interpretation**: The above pivot table shows the average sqft versus neighborhood and type of construction. Here you can view that the type of construction may not be a good comparison to sqft as brick homes showed to have less sqft in some neighborhoods. However, similar to home prices table above, West neighborhood has the highest sqft.

# Pivot Charts

* **Using the two pivot tables above, generate pivot charts for average price and average square feet by type of construction (brick) and neighborhood.**

Average price and average square feet by construction type and neighborhood are displayed in the charts below.

**Interpretation**: The highest prices are in the West neighborhood. Brick houses costs more than non-brick houses.

**Interpretation:** The largest houses tend to be in the West neighborhood. Brick houses are bigger than non-brick houses in the West and East neighborhoods, but smaller in the North neighborhood.

# Correlation analysis

* **Perform a correlation analysis of all quantitative variables except ID**

Correlation analysis of all quantitative variables (Price, SqFt, Bedrooms, Bathrooms, and Offers) from the House Prices dataset is presented in the figure below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Price** | **SqFt** | **Bedrooms** | **Bathrooms** | **Offers** |
| Price | 1.00 |  |  |  |  |
| SqFt | 0.55 | 1.00 |  |  |  |
| Bedrooms | 0.53 | 0.48 | 1.00 |  |  |
| Bathrooms | 0.52 | 0.52 | 0.41 | 1.00 |  |
| Offers | (0.31) | 0.34 | 0.11 | 0.14 | 1.00 |
| highest correlation |  |  |  |  |  |
| lowest correlation |  |  |  |  |  |

|  |  |
| --- | --- |
| **Question** | **Answer** |
| *Which two variables have the largest magnitude correlation?* | Square footage and Price have the largest magnitude correlation. |
| *Which two variables have the smallest magnitude correlation?* | Offers and number of bedrooms have the smallest magnitude correlation. |
| *What does the largest magnitude imply if we perform a regression analysis next?* | We would expect Square footage to be an influential variable in the regression equation. |
| *Are there any negative correlations?* | Number of offers has a negative correlation with price. When one goes up the other goes down. |
| *Are these correlations intuitive? If not, why not?* | * **Square footage and price with a positive correlation**: The positive sign is expected as bigger houses tend to cost more. * **Offers and price with a negative correlation**: This result is non-intuitive. More square feet might affect the price but more offers don’t make the price go down and the opposite is true. The relationship is somewhat non-intuitive as one would expect the larger number of bids should translate into higher home prices in an auction scenario aka bidding war (California sales would be a good example, where the seller in a hot market will choose the bidder offering the highest price. Therefore, we would expect to see the Price increases as the number of offers increases). * However, in a recession scenario where the seller initially is asking unreasonably high price compare to the market price than the buyers are willing to pay would most likely result in multiple offers (moving by for example 10K increments). The seller might agree to lower the price if the bids are coming low and there is some urgency to sell the home. |

# Regression Analysis

* **Perform an initial regression analysis of the quantitative variables excluding the ID.**

The results of an initial regression analysis of the quantitative variables excluding the ID is shown in the figure below.



* **Which variables are statistically significant?**

SqFt, Bedrooms, Bathrooms, and Offers all have p-values below 0.05. In fact, they are all nearly zero. They are all statistically significant.

* **What does each coefficient mean in a real world sense?**

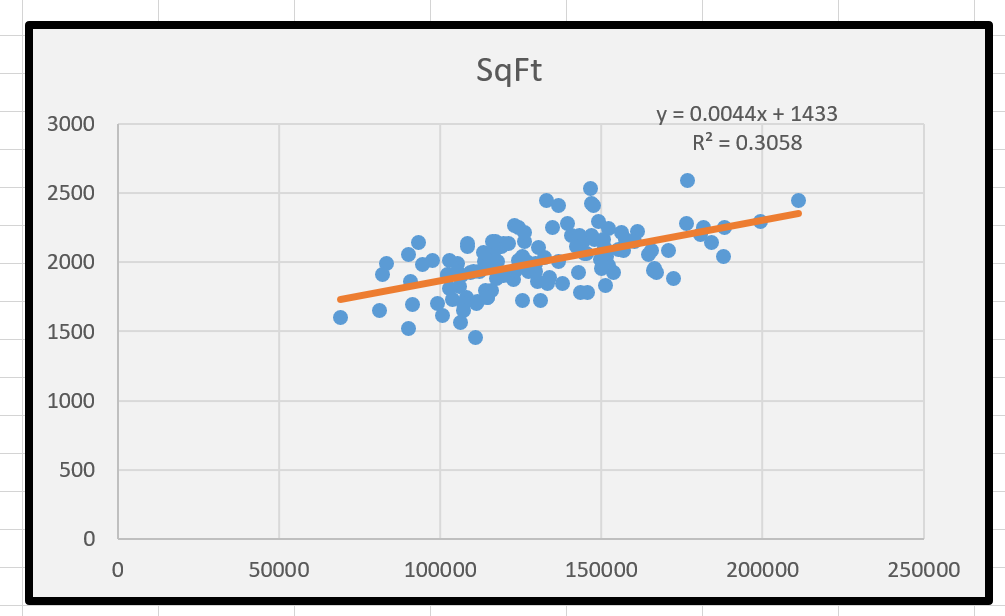
**Square feet:** 1000 square feet increase holding bedrooms, bathrooms and offers unchanged would translate to ~$62K increase in the home price. The positive sign is expected and in line with team’s expectations. In general, larger homes cost more (as per the data set provided, this might not hold true if we had looked at another dataset with Manhattan data)

**Bedrooms:** One-bedroom increase holding all other variables constant will increase the home price by ~$9K. The positive sign is expected and in line with team’s expectations. More bedrooms cost more in general.

**Bathrooms:** An additional bathroom will increase the price by ~$13K. The positive sign is expected and in line with team’s expectations. Bathrooms are more expensive to renovate or add to the existing house structure, therefore the price of the house increases.

**Intercept:** If a home has one less offer, it would imply the house price is expected to be lower than a similar house by approximately ~14K. The sign is negative and non-intuitive to the team.

There aren’t any houses that are 0 square feet with 0 bedrooms and bathrooms and 0 offers. Because there are no houses for $0, the intercept is far outside of our data, and counts as extrapolation (something we shouldn’t do, theoretically). We can’t do a scatterplot of multiple x’s, but when we look at a scatterplot of square footage and price, we see that the point farthest left is not close to intercepting the x axis. The cheapest house in our dataset is $69,100. The negative value in our equation makes the regression equation work, but can’t be taken literally here.



* **What does the R-squared mean?**

The adjusted R squared for the model is 68.8% meaning 68.8% of variation in home prices can be explained by changes in square footage, number of bedrooms and bathrooms, and number of offers for a particular home.

# Two-way sensitivity

* **Create a spreadsheet prediction of the model. Perform a two-way sensitivity analysis and use conditional formatting to highlight the results.**

The model equation and the coefficients are the following:

**Equation:** price = -17347.38 + 61.84\*sqft + 9319.75\*bedrooms + 12646.35\*bathrooms - 13601.01\*offers



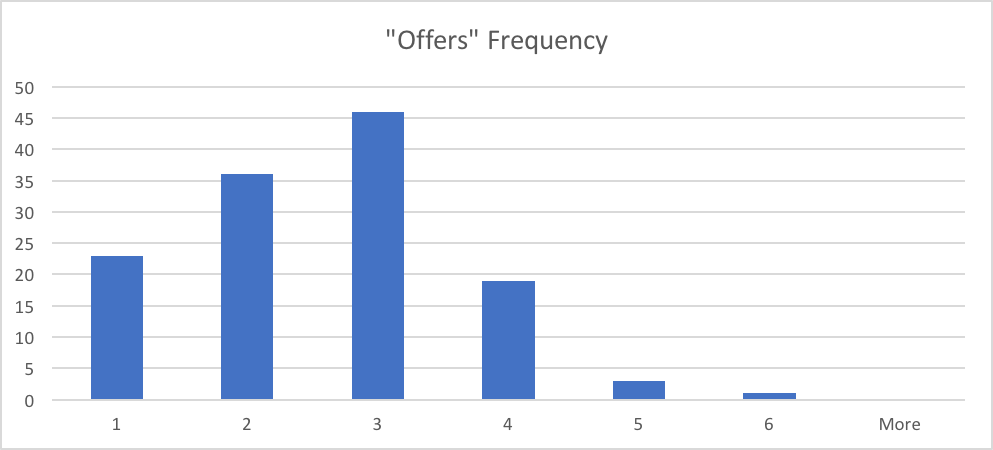
**Interpretation:** According to the two-way sensitivity analysis, the cheapest house has the fewest bedrooms and the least square feet. The most expensive house has the most bedrooms and the most square feet. This is intuitive.

# Conclusion

* **What would explain non-intuitive results in your regression using the data which you were provided?**

First, the frequency distribution chart below shows the Offers data is skewed to the right, hence not normal which doesn’t meet one of the assumptions of ordinary least squares, therefore contributing to a “wrong” sign of the coefficient.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Bins* | *Frequency* | *Cumulative %* | *Bins* | *Frequency* | *Cumulative %* |
| 1 | 23 | 17.97% | 3 | 46 | 35.94% |
| 2 | 36 | 46.09% | 2 | 36 | 64.06% |
| 3 | 46 | 82.03% | 1 | 23 | 82.03% |
| 4 | 19 | 96.88% | 4 | 19 | 96.88% |
| 5 | 3 | 99.22% | 5 | 3 | 99.22% |
| 6 | 1 | 100.00% | 6 | 1 | 100.00% |
| More | 0 | 100.00% | More | 0 | 100.00% |



Second, the dataset used to estimate the regression doesn’t support the claim that the price is larger with an increasing number of offers. To demonstrate it, a weighted average price for each number of offers was calculated as shown in the table below. As the number of Offers increases, the weighted average price drops from $145K for houses with one offer to $90K for houses with six offers. (Having more observations where count of records is 19, 3 and 1 would help to provide a better conclusion).



Once the weighted average price is plotted against the number of Offers and a trendline is fitted, the R-squared is 89%. The trend in the historical data clearly demonstrates as the number of offers increases the price declines for the given dataset.

To further explain non-intuitive result a few more regressions were estimated. Using stepwise regression and recoding Brick and Neighborhood using dummy variables (0 and 1) ultimately resulted in the regression below.



Taking into account Brick and Neighborhood variables did improve the model fit and the coefficient of Offers changed from (-$13.6K) to (-$7.6K) but remained negative. The adjusted R squared improved from 68.8% to 85.5% once brick and neighborhood variables were considered. However, it was still not enough to explain why the Offers coefficient is negative.

* **What additional data would assist you in explaining the non-intuitive results?**

We understand that there are many times when you can’t get access to all of the data that you’d like. That said, there is some data that would help greatly to clarify the non-intuitive results.

|  |  |  |
| --- | --- | --- |
| Question | Data | How the data could help |
| What is the economic environment for the house price dataset? | Dates of house sales in our dataset  General economic trends for the geographic area | If the data were collected during a recession or other buyer’s market, that might explain the lower prices. |
| Is our dataset a random sample of houses in this town? | More house sales, including for smaller and larger houses | The current data set is very compact and includes only 128 data points (it seems to be missing both small and large houses). It’s understandable that we are working with a sample, but if the sample isn’t representative of the whole population then we may get skewed results. |
| What was the posted sale price of each house? | It seems we have closing prices, but we’d benefit from seeing what the houses were advertised for sale as. | If a house were advertised at lower than the norm for a house of that size in that neighborhood, that could drive more offers. |
| How long was the house on the market? | Number of days the house was on the market | Houses that are on the market a long time have the opportunity to get more offers and also may be overpriced. |

# Glossary

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