

Coursera Capstone Project

The Effects of Housing Features on Price in Minneapolis Neighborhoods, Minnesota USA



December 11, 2020

Coursera

Table of Contents

[Introduction: Business Problem 1](#_Toc58593038)

[Methodology 1](#_Toc58593039)

[Analysis 2](#_Toc58593040)

[Results and Discussion 9](#_Toc58593041)

[Conclusion 10](#_Toc58593042)

## Introduction: Business Problem

I used to work in the mortgage industry processing and closing residential loans. At that time, I did not know some of the factors that influence house prices or how to determine the correlation between the housing features and price. I express particular interest in this topic because of my background. Studying data science with Coursera and IBM, provides me with some knowledge, insight, and curious mind to go little deeper and determine what features buyers are interested in and willing to pay more for them.

Aside from my personal interest in this topic, the target audience I believe the outcome of this project will benefit may be mortgage or lending companies, buyers, individual sellers, real estate brokers, and independent contractors (building, remodeling houses) in the Minneapolis housing market. The results of this project, I hope, will help these interested parties better estimate the house prices with certain features.

In this project, I explore the determinants of house prices in the Minneapolis neighborhood housing market. What are the features in a house that interested parties are looking for and willing to pay more? What features have strong correlation with house price? These questions and more are examined further in this project. Not all features studied in this project have equal or strong correlation to the price. In other words, correlation between housing features varies as noted in the analysis section of this report.

## Methodology

As noted above, the focus of this project is to examine the effect of selected features on residential house price in the Minneapolis neighborhood market. Luckily, the dataset is organized by the city of Minneapolis and freely available at the [Open Minneapolis](https://opendata.minneapolismn.gov/search?tags=assessor) website. On the website, the data is available from 2017 to 2020. I chose 2019 because I thought is most recent and complete. The dataset in csv file is downloaded and has 130,915 rows and 53 columns. The column name “property\_type” has residential, double bungalow, triplex, apartment, commercial, condos, vacant land, cooperative, etc., property types. The focus of this report is on residential property type as such other property types are dropped. After initial data wrangling, the data frame has 75,328 rows and 21 columns remaining.

Out of the 53 columns in the data frame, I choose the following columns for processing:

|  |  |
| --- | --- |
| ZIPCODE | Postal code |
| FORMATTED\_ADDRESS | Full mailing address |
| PARCEL\_AREA\_SQFT | Land area of parcel in square feet |
| X | coordinate of parcel centroid (uses NAD 83 HARN Hennepin County coordinate system) |
| Y | coordinate of parcel centroid (uses NAD 83 HARN Hennepin County coordinate system) |
| PROPERTY\_TYPE | Type of property |
| NEIGHBORHOOD | Name of neighborhood the property is in |
| LANDVALUE | Property's assessed land value as of January 2 of its assessment year |
| BUILDINGVALUE | Property's assessed building value as of January 2 of its assessment year |
| TOTALVALUE | Property's total assessed value as of January 2 of its assessment year |
| BELOWGROUNDAREA | Total square footage below grade |
| ABOVEGROUNDAREA | Total square footage above grade |
| NUM\_STORIES | Number of stories in building |
| GARAGE\_PRESENT | Indicates whether property has a garage |
| PRIMARYHEATING | Primary type of heating in building |
| CONSTRUCTIONTYPE | Style of construction |
| EXTERIORTYPE | Exterior finish of building |
| ROOF | Style of roof |
| FIREPLACES | Number of fireplaces |
| BATHROOMS | Number of bathrooms |
| BEDROOMS | Number of bedrooms |

In the analysis section, I used box plots to determine the relationship between object data types such as primary heating, exterior type, construction type, and roof to provide visual depiction of the relationships. Some integer and float data types are depicted visually for those who may not be familiar with numbers in regression analysis. Foursquare API is used as a requirement for this project to construct a URL, search for venues, request for venues, and explore the geographical location in the city of Minneapolis, Minnesota. Folium, the visualization library, is used to visualize the results. This dataset, [Minneapolis Neighborhoods](https://opendata.minneapolismn.gov/datasets/minneapolis-neighborhoods/geoservice), was downloaded from Open Minneapolis site. Finally, I added the city of Minneapolis [police incidents report of 2019](https://opendata.minneapolismn.gov/datasets/8cd15449ac344aa5a55be7840d67c52d_0) on this report. The neighborhoods that crimes have been committed and the descriptions of the crime are displayed on the map. This is not required but I felt that interested parties may want to know if the neighborhood they are interested in buying a house or investing a large sum of money is the safest environment they can find.

## Analysis

As mentioned above, box plots are used to visualize the relationships between the object data types and the building value or price. I take advantage of one of the benefits of box plots, i.e., comparing the distribution of multiple datasets. Take a look at the box plot below displaying the distribution of multiple heating units that can be installed in houses in Minneapolis.

From the box plots below, it’s clear that the distribution of building value in all the categories under primary heating, exterior type, construction type, and roof are distinct enough to be considered as potential predictors of building value. The tables below each box plot show the average building price for each categorical variable. Interested partied do not have to be a data scientist to compare the visual and numerical representation of the data.

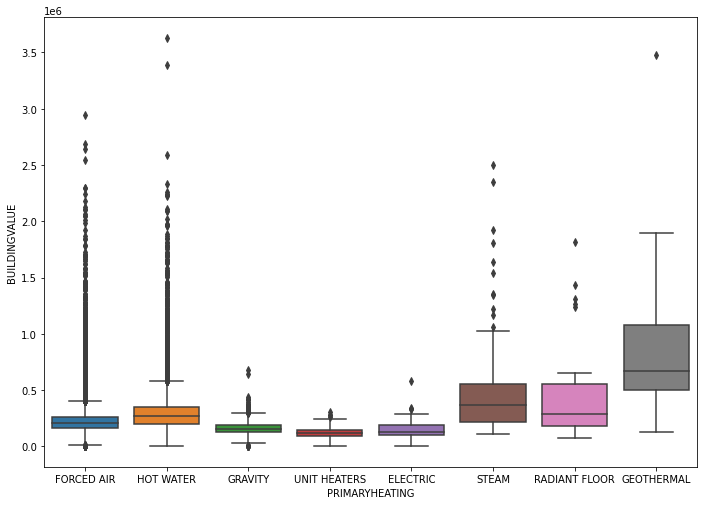
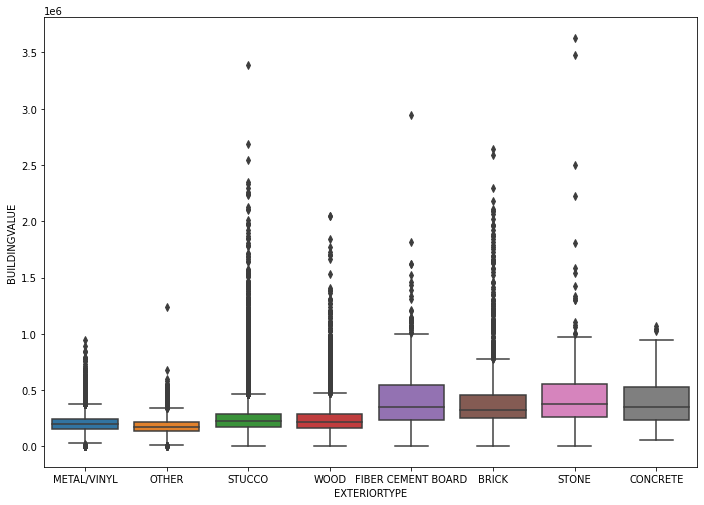


Figure 1: Box Plot displaying the primary heating systems vs. building value

|  |  |  |
| --- | --- | --- |
|  | PRIMARYHEATING | BUILDINGVALUE |
| 0 | ELECTRIC | 148578.625954 |
| 1 | FORCED AIR | 226756.243957 |
| 2 | GEOTHERMAL | 841377.500000 |
| 3 | GRAVITY | 161065.417256 |
| 4 | HOT WATER | 301815.756115 |
| 5 | RADIANT FLOOR | 502555.555556 |
| 6 | STEAM | 483065.034965 |
| 7 | UNIT HEATERS | 121676.555024 |

Table 1: The average price of a house vs. primary heating units

 Figure 2: Box plot showing exterior type vs. average building price

|  |  |  |
| --- | --- | --- |
|  | EXTERIORTYPE | BUILDINGVALUE |
| 0 | BRICK | 406892.531705 |
| 1 | CONCRETE | 420098.305085 |
| 2 | FIBER CEMENT BOARD | 410577.536946 |
| 3 | METAL/VINYL | 205204.886799 |
| 4 | OTHER | 180485.735115 |
| 5 | STONE | 563175.229358 |
| 6 | STUCCO | 251052.487628 |
| 7 | WOOD | 250461.232968 |

Table: The average house price vs. exterior type of the building

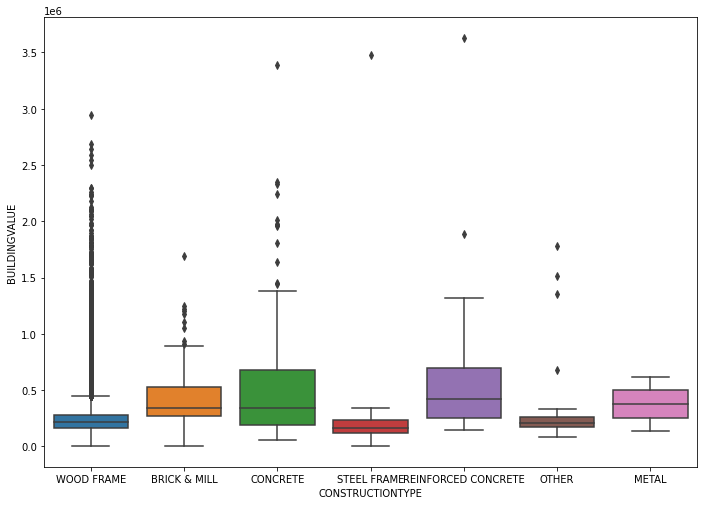


Figure 3: Construction type vs. average building price

|  | **CONSTRUCTIONTYPE** | **BUILDINGVALUE** |
| --- | --- | --- |
| **0** | BRICK & MILL | 444228.421053 |
| **1** | CONCRETE | 561040.000000 |
| **2** | METAL | 376200.000000 |
| **3** | OTHER | 361300.000000 |
| **4** | REINFORCED CONCRETE | 660821.428571 |
| **5** | STEEL FRAME | 367929.411765 |
| **6** | WOOD FRAME | 242848.606534 |

Table 3: Construction type vs. the average building price

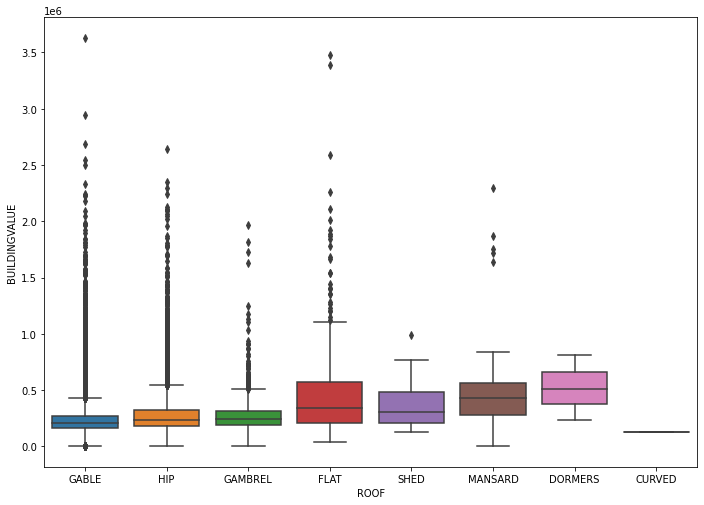


Figure 4: Building roof vs. average price of the building

|  |  |  |
| --- | --- | --- |
|  | ROOF | BUILDINGVALUE |
| 0 | CURVED | 131200.000000 |
| 1 | DORMERS | 517566.666667 |
| 2 | FLAT | 487314.285714 |
| 3 | GABLE | 234420.681748 |
| 4 | GAMBREL | 272766.173362 |
| 5 | HIP | 280136.134454 |
| 6 | MANSARD | 506101.265823 |
| 7 | SHED | 380957.142857 |

Table 4: Roof type vs. average building price

|  | **PARCEL\_AREA\_SQFT** | **lat** | **long** | **LANDVALUE** | **BUILDINGVALUE** | **TOTALVALUE** | **BELOWGROUNDAREA** | **ABOVEGROUNDAREA** | **NUM\_STORIES** | **GARAGE\_PRESENT** | **FIREPLACES** | **BATHROOMS** | **BEDROOMS** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **PARCEL\_AREA\_SQFT** | 1.000000 | -0.118563 | -0.037254 | 0.575397 | 0.357893 | 0.482362 | 0.415074 | 0.453493 | 0.132902 | 0.243271 | 0.360146 | 0.353681 | 0.224987 |
| **lat** | -0.118563 | 1.000000 | -0.188511 | -0.286888 | -0.110249 | -0.195394 | -0.164800 | -0.226236 | -0.145498 | -0.023582 | -0.192883 | -0.189273 | -0.147942 |
| **long** | -0.037254 | -0.188511 | 1.000000 | -0.308926 | -0.295164 | -0.327128 | -0.128683 | -0.123491 | -0.073202 | -0.068377 | -0.257766 | -0.143066 | -0.042815 |
| **LANDVALUE** | 0.575397 | -0.286888 | -0.308926 | 1.000000 | 0.673235 | 0.872212 | 0.461243 | 0.620482 | 0.288749 | 0.222736 | 0.500328 | 0.509631 | 0.246548 |
| **BUILDINGVALUE** | 0.357893 | -0.110249 | -0.295164 | 0.673235 | 1.000000 | 0.948879 | 0.558674 | 0.793681 | 0.442547 | 0.300498 | 0.542963 | 0.672218 | 0.416887 |
| **TOTALVALUE** | 0.482362 | -0.195394 | -0.327128 | 0.872212 | 0.948879 | 1.000000 | 0.566451 | 0.789881 | 0.415956 | 0.293857 | 0.572743 | 0.662215 | 0.381013 |
| **BELOWGROUNDAREA** | 0.415074 | -0.164800 | -0.128683 | 0.461243 | 0.558674 | 0.566451 | 1.000000 | 0.594817 | 0.022481 | 0.242454 | 0.501608 | 0.469452 | 0.392869 |
| **ABOVEGROUNDAREA** | 0.453493 | -0.226236 | -0.123491 | 0.620482 | 0.793681 | 0.789881 | 0.594817 | 1.000000 | 0.643532 | 0.256607 | 0.513397 | 0.694657 | 0.592811 |
| **NUM\_STORIES** | 0.132902 | -0.145498 | -0.073202 | 0.288749 | 0.442547 | 0.415956 | 0.022481 | 0.643532 | 1.000000 | 0.118247 | 0.178515 | 0.404506 | 0.416869 |
| **GARAGE\_PRESENT** | 0.243271 | -0.023582 | -0.068377 | 0.222736 | 0.300498 | 0.293857 | 0.242454 | 0.256607 | 0.118247 | 1.000000 | 0.193424 | 0.267859 | 0.170274 |
| **FIREPLACES** | 0.360146 | -0.192883 | -0.257766 | 0.500328 | 0.542963 | 0.572743 | 0.501608 | 0.513397 | 0.178515 | 0.193424 | 1.000000 | 0.445638 | 0.273649 |
| **BATHROOMS** | 0.353681 | -0.189273 | -0.143066 | 0.509631 | 0.672218 | 0.662215 | 0.469452 | 0.694657 | 0.404506 | 0.267859 | 0.445638 | 1.000000 | 0.504669 |
| **BEDROOMS** | 0.224987 | -0.147942 | -0.042815 | 0.246548 | 0.416887 | 0.381013 | 0.392869 | 0.592811 | 0.416869 | 0.170274 | 0.273649 | 0.504669 | 1.000000 |

Table 5: On this table, the correlation between int64 and float64 variable types are calculated using the method corr.

The table above displays the calculated results of correlation between each column heading with variable type integer and float. On this table, viewers can compare the correlation between column headings to the corresponding intersection rows. For example, if you want to find out the correlation between the building value and above the ground area can be found on two different points depending on your reference or starting point (column or row). The correlation between these two headings is 0.793681.

**Linear Relationships**

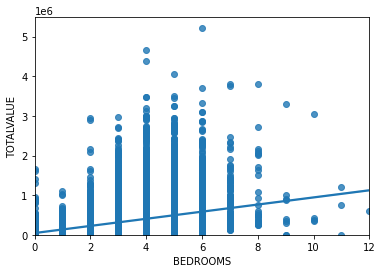
****

Figure 5: Correlation between number of bedrooms and total building value

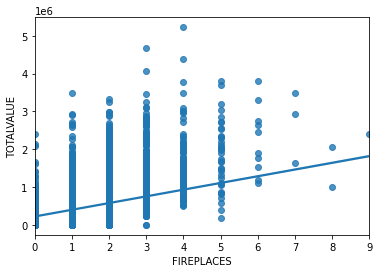


Figure 6: Correlation between number of fireplaces and total building value

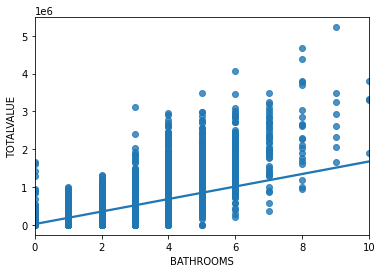
****

Figure 7: Correlation between number of bathrooms and total building value

## Results and Discussion

The analysis of the house data, using the object variables, reveals noticeable increase in average building value for some variables and lower average price for others. For example, the primary heating has average building value of $841,377.50 for a house with Geothermal heating unit and $121676.55 for a house fitted with Unit heaters at the lower end. These clearly shows that the determinants chosen in this study affect the house prices in the Minneapolis neighborhoods area. Interested parties are therefore presented with clear choice what material to use if they want to increase the value of the house. It may be cost of variables that affect the value of a house or other factor. Whatever the reason, there are clear variation in values of house fitted with different units.

While searching and replacing missing values, I noticed that the most frequent categorical variables in their respective categories are not those that add highest average value to the house. In other words, the materials most used by builders ranked at the very bottom with respect to increase value of the house. For example, forced air ranked 5th out of 7, wood frame 6th out of 6, stucco 5th out of 7, and gable ranked 6th out of 7 categories. For better visualization and comprehension of these observations, please refer to the categorical figures and their respective tables above. Grouping the construction type and roof type, the combination of the reinforced concrete and flat roof type is the most expensive ($946,416.70).

Applying the Pearson Correlation Coefficient and P-value, I found the correlation between the number of bedrooms, bathrooms, and fireplace to be statistically significant though, the linear relationships in all three of them are moderately strong. These results are consistent with table 5 above and the three figures below it.

|  |  |  |
| --- | --- | --- |
|  | **Pearson Correlation Coefficient** | **P-value** |
| Bedrooms | 0.3810128178868299 | 0.0 |
| Bathrooms | 0.6622152395520273 | 0.0 |
| Fireplaces | 0.5727427014314311 | 0.0 |

Foursquare API is used to request interesting venues in Minneapolis neighborhoods. Some venues are found and listed on the map. When I run k-means to cluster the neighborhood, I am surprise to find only 1 cluster in Minneapolis neighborhood. I started with k=5 to k=2 getting error in all of them. When I set k=1, it works just fine with one dot on the map. As I noted above, police incident report of 2019 is added where descriptions of crimes and the neighborhoods those crimes were committed are noted on the map. The crime data is not a requirement for this project but I thought the target audiences may be interest to know the categories, frequency, and the neighborhoods the crimes occurred.

In model evaluation, the results are both visualized and measured quantitatively to determine the accuracy of the model. In this study, I used the R^2 to determine how accurate the models are. R square or the coefficient of determination, is a measure to show how close the data is to the fitted regression line. The value of the R^2 is the percentage of variation of the response variable (y) that is explained by a linear model. For linear regression, ridge regression, and second-degree polynomial I used the total value (which include the land and building values) to calculate the R^2 values. In second-degree polynomial, both the training data and testing data are used to calculate R^2 values and the results are 0.7847696530237769 and 0.7713001435645377 respectively. I can now say that ~78.48% of the

Variation of total value is explained by this polynomial fit. The ridge regression on trained data has R^2 of 0.710036153823197 and simple linear regression is 0.6239123580483767. Please refer to the code on Jupyter Notebook for complete results.

## Conclusion

In conclusion, the determinants studied in this project have some correlation to the residential properties at different levels. In the categorical variables section, for example, I found noticeable increase in average house values with certain categorical variables. For example, houses with reinforced concrete have higher average value than other categorical variables in the same category. Finally, I observed that the most commonly used variables under the primary heating, construction type, exterior type, and the roof are the least valuable in terms of the average house values. This brings us to the question that may be explored further in future researches. And that is, why are the variables that least increase the average value of a house the most commonly used by builders or owners?