Erin Saunders, Nolan Arendt, Robert Ransom

Syracuse University ischool | MARCH 2023

Ames, IA Housing Market Analysis: 2006-2010

IST687 – Group b, Prof. Toni Hanrahan

**EXECUTIVE SUMMARY**

The Ames, Iowa housing dataset provides valuable information on important house features and the sale price that can help determine the return on investment (ROI) for different housing features in the Ames Housing Market. These features include lot size, year built, neighborhood, exterior quality, basement quality, heating quality, kitchen quality, garage quality, deck square footage, and sale price. Analyzing these features can help inform decisions related to maximizing ROI and guide strategic investments in the housing market. By using this dataset to identify the most impactful features, individuals can make more informed decisions related to buying, selling, or investing in real estate in Ames, Iowa to maximize their ROI.

**METHODOLOGY**

**Source Selection** – Initial discussion among the group led to a research topic revolving around housing markets, purchasing, and recommendations on remodeling or immediate updates that can be made to increase property value for first-time home buyers or investors. Real estate company Zillow provided an initial dataset with sale prices over a period of time but did not address specific variable option listings between observations to identify a correlated change or feature that could address the price variation.

Kaggle.com, known for comprehensive datasets and resources for data science professionals and enthusiasts, was consulted and provided the Ames Housing Dataset, listing home sales between 2006 and 2010; 2930 observations. The dataset broke down house features into 80 variables ranging from lot size and square footage, to siding material and patio construction. Overall, approximately 234,000 individual features could be assessed to identify and model the most impactful variables in the Ames Housing Market. Future investigation should be done with updated sales information to validate developed models and simulations.

Kaggle.com was selected, data set can be found at [Ames Housing Dataset | Kaggle](https://www.kaggle.com/datasets/prevek18/ames-housing-dataset) <<https://www.kaggle.com/datasets/prevek18/ames-housing-dataset>>. Research and presentation of the Ames Dataset was originally conducted by Dean De Cock, Truman State University. Data provided by the Ames, Iowa Assessor’s Office. Data dictionary and abstract found at <https://jse.amstat.org/v19n3/decock/DataDocumentation.txt>; elements are included in this document for the variables selected during investigation.

**Variable Selection and Cleaning** – Electing to focus on internal updates, assuming major renovations such as roof replacement or siding upgrades to address structural integrity or weather proofing issues, the list of variables was pared down to an assumed list of the most impactful internal aspects pertaining to sale price. The following variables were selected for correlation and regression analysis; data definitions and factors are referenced from <https://jse.amstat.org/v19n3/decock/DataDocumentation.txt>:

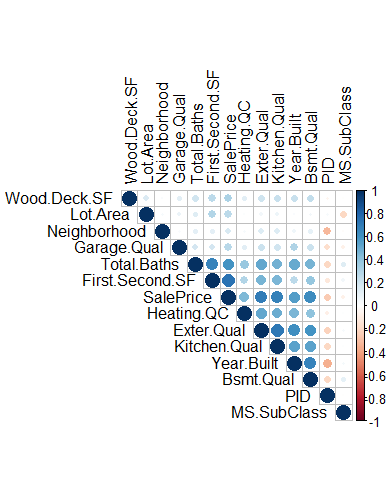
* MS.SubClass: Identifies the type of dwelling involved in the sale,
* Neighborhood: Physical location within Ames city limits,
* Year.Built: Original construction date,
* Exter.Qual: Evaluates the quality of the material on the exterior,
* Bsmt.Qual: Evaluates the height of the basement,
* Heating.QC: Heating quality and condition,
* Kitchen.Qual: Kitchen quality,
* Garage.Qual: Garage quality,
* Lot.Area: Lot size in square feet,
* X1st.Flr.SF: First Floor square feet,
* X2nd.Flr.SF: Second Floor square feet,
* Full.Bath: Full bathrooms above grade,
* Half.Bath: Half bathrooms above grade,
* Wood.Deck.SF: Wood deck area in square feet,
* SalePrice: Sale price $$. Sale price variation will be evaluated based on the selected variables following correlation and linear regression analysis to determine direction and strength (correlation), direction and magnitude (regression) of the identified relationships.

Cleaning of the data focused on three areas: Dummification, Combination, and Categorization. Cleaning was necessary to convert string variables to binary, combine available square footage to calculate the impact of total square footage, and determine the significance of several external attributes.

**Feature Engineering/Dummification** –

* Dummification (use of dummy variables to turn string variables to numeric based on Boolean Logic) addressed Neighborhood, Exter.Qual, Bsmt.Qual, Heating.QC, Kitchen.Qual. Based on the data dictionary, each “quality” variable was categorized on a scale of 1-5; Neighborhood was categorized from 1-28 based on listed name. Conversion from string to numeric was used for calculation of correlation and regression relating to SalePrice. Duplex houses (MS.SubClass = 090) were removed from the dataset to focus on single family homes.
* Combination addressed Bath and First/Second Floor square footage.
  + Half.Bath values were first converted to a decimal representation of the number of half baths. For example, 1 becomes .1, 3 becomes .3. This calculation was necessary to address whole integers representing both Full and Half Baths in the dataset. Without conversion there was no way to differentiate a combined total of 4 bathrooms to mean 2 Full/2 Half, 4 Full, or 8 Half baths. Total.Baths (Full.Bath + Half.Bath) was created in the data frame; Full/Half.Bath were subsequently removed from the data frame to avoid multicollinearity among the variables.
  + X1st.Flr.SF and X2nd.Flr.SF combined into the new variable First.Second.SF; individuals were removed to avoid multicollinearity. The new variable represents the estimated total square footage in the house but does not include additional square footage potentially found with “Low Qual Fin SF,” addressing low quality finished square feet across all floors.
* Categorization was used to group Wood.Deck and First.Second.SF observations for computation of upgrades/modernization impact on SalePrice.

**Correlation, Linear Regression, Modeling** – Following cleaning, an initial correlation matrix was created to identify relationships between the variables.



Of the selected variables, the initial correlation matrix assisted in further refining focus on those that had the greatest correlation with SalePrice to answer the remaining business questions. Using the categorized variables, recommendations in response to business questions can be made on which areas to focus and what level improvement should be targeted, as well as initial quality of the purchased home in order to maximize Return on Investment in Ames, IA.

**Linear Regression Analysis**

Text

Description automatically generated with medium confidence

Following the correlation matrix analysis, variables with a low-strength relationship with SalePrice were removed from the data frame. Remaining variables were determined to have the highest potential impact on SalePrice.

**Prediction** – Using the developed linear regression model, categorized standards for initial purchase and final quality categories were identified and applied against the model, predicting the purchase price and potential impact of remodeling on future SalePrice.

Graphical user interface, text

Description automatically generated

**Project Coordination and Planning (Kanban)** – Project coordination was conducted using Microsoft Teams and Planner, laying out tasks in buckets (Backlog, In Progress, Internal/External Blocker, Review/Validation, Complete) with self-nomination capability for product creation.

**RESULTS AND BUSINESS ANSWERS**

**What variables have the greatest impact on SalePrice?**

Timeline

Description automatically generated

The correlation matrix provides an impressive visual representation of the direction and strength of the relationship between multiple variables. Using this matrix, most impactful variables were further focused from the original selection of 16 to the 7 that represent the greatest correlation with SalePrice. Surprisingly, Neighborhood and Lot.Area had a weak correlation with SalePrice, which could be expected in a larger city where certain neighborhoods or larger plots are at a premium. Conversely, Bsmt.Qual (which references basement height, not finish) had one of the strongest correlations. After preliminary research and personal experience, knowing that Ames, IA is a prime location for tornado activity helps to explain that correlation, where smaller basements may not accommodate individuals or families, or incur increased cost to shore up resilience with a relative level of comfort during storms.

**What improvements can be made with the greatest ROI?**

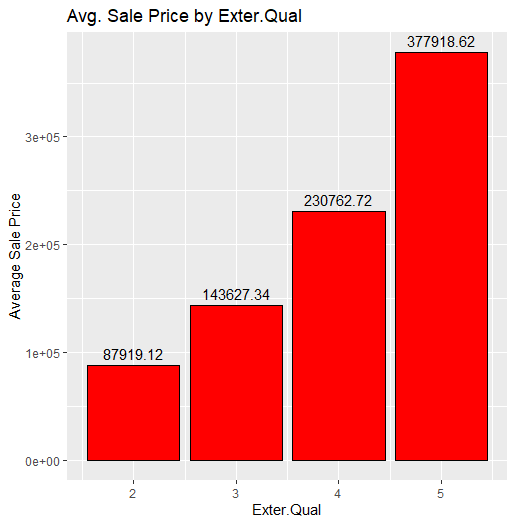
In order to evaluate the impact of different improvements on the ROI, each considered variable quality category was considered against the average SalePrice for listings at that level. For example, Exterior Quality at category 3 (Typical/Average) correlated to an average SalePrice of $143,627. In order to increase the quality, a prospective homebuyer might consider new siding to raise the category to 4 and see potential gain of ~$90K in resale value. R Code example below of grouping and summarization.

Text

Description automatically generated

Each variable is analyzed in this manner providing buyers fields to target for potential improvements, or areas to consider when negotiating purchase price.

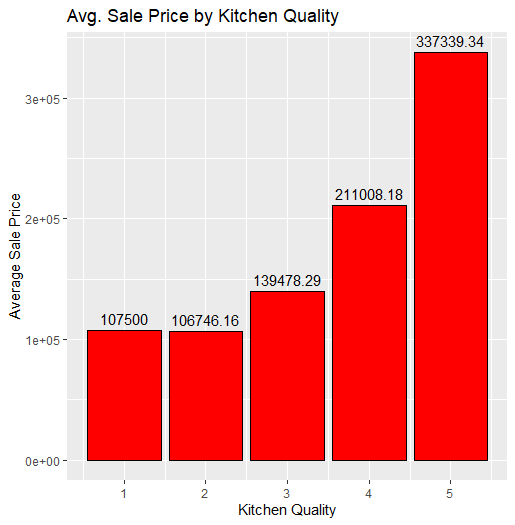
**External Quality**



|  |  |
| --- | --- |
| Exterior Quality | Average Sale Price |
| 2 (Fair) | $87, 919 |
| 3 (Typical) | $143, 627 |
| 4 (Good) | $230, 763 |
| 5 (Excellent) | $377, 919 |

Based strictly on External Quality (excluding current market rates for labor and materials) upgrading from Good to Excellent provides the greatest impact on average sale price (~$147K), while Typical to Good is perhaps more attainable depending on available funds and investment capability (~$90K).

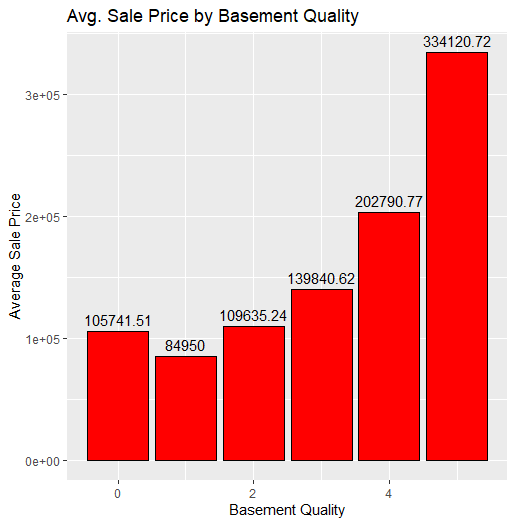
**Kitchen Quality**



|  |  |
| --- | --- |
| Kitchen Quality | Average Sale Price |
| 1 (Poor) | $107, 500 |
| 2 (Fair) | $106, 746 |
| 3 (Typical) | $139, 478 |
| 4 (Good) | $211, 008 |
| 5 (Excellent) | $337, 339 |

Kitchen Quality (excluding current market rates for labor and materials) appears to favor the higher end kitchens, improving average sale price ~$70-200K if initially rated as Typical. Upgrading from Poor to Fair does not appear to impact average sale price. Recommended to purchase at Typical and upgrade to Good quality kitchen for first-time home buyers.

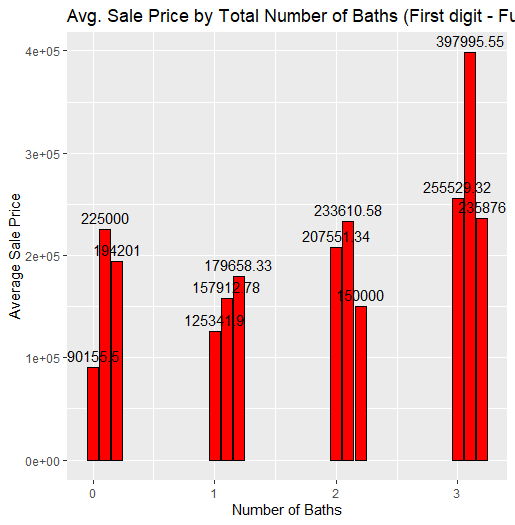
**Basement Quality (Height)**



|  |  |
| --- | --- |
| Basement Quality | Average Sale Price |
| 0 (No bsmt) | $105, 742 |
| 1 (Poor) | $84, 950 |
| 2 (Fair) | $109, 635 |
| 3 (Typical) | $139, 841 |
| 4 (Good) | $202, 791 |
| 5 (Excellent) | $334, 121 |

Basement height is difficult to recommend changes, as any modification in this area would be architectural and could incur additional costs associated with such a large upgrade. Based on the year of the house, modifying foundation type aspects could expose larger structural issues or requirements that may outweigh the value gained.

**Total Baths (Full + Half)**



|  |  |
| --- | --- |
| Total Baths (Full).(Half) | Average Sale Price |
| 0 | $90, 156 |
| 0.1 | $225, 000 |
| 0.2 | $194, 201 |
| 1 | $125, 342 |
| 1.1 | $157, 913 |
| 1.2 | $179, 658 |
| 2 | $207, 551 |
| 2.1 | $233, 611 |
| 2.2 | $150, 000 |
| 3 | $255, 529 |
| 3.1 | $397, 996 |
| 3.2 | $235, 876 |

On average, an additional half bath adds ~$25K if the total number of half baths is below 2. Houses with two half baths with more full baths tended to reduce in price rather than increase with the number of baths. The largest impact seen was from three full, to three full and one half.

<Page Break>

<Space Intentionally Left Blank>

**Neighborhood**

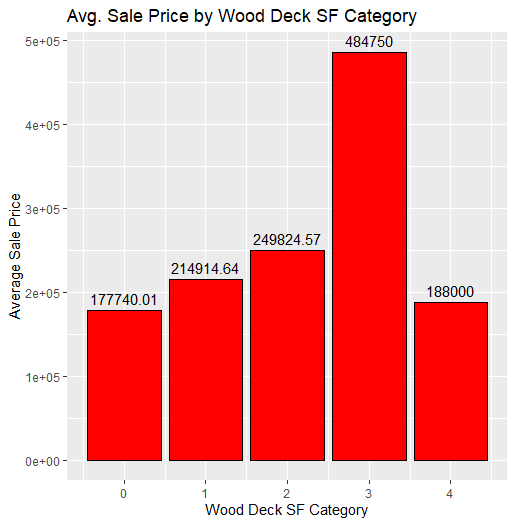
Table

Description automatically generated

|  |  |  |
| --- | --- | --- |
| Nbr # | Neighborhood | Average Sale Price |
| 1 | Blmngtn | $196, 662 |
| 2 | Blueste | $143, 590 |
| 3 | BrDale | $105, 608 |
| 4 | BrkSide | $124, 756 |
| 5 | ClearCr | $208, 662 |
| 6 | CollgCr | $201, 867 |
| 7 | Crawfor | $208, 765 |
| 8 | Edwards | $131, 882 |
| 9 | Gilbert | $190, 647 |
| 10 | Greens | $193, 531 |
| 11 | GrnHill | $280, 000 |
| 12 | IDOTRR | $103, 685 |
| 13 | Landmrk | $137, 000 |
| 14 | MeadowV | $95, 756 |
| 15 | Mitchel | $162, 525 |
| 16 | NAmes | $146, 506 |
| 17 | NoRidge | $330, 319 |
| 18 | NPkVill | $140, 711 |
| 19 | NridgHt | $332, 018 |
| 20 | NWAmes | $189, 601 |
| 21 | OldTown | $124, 318 |
| 22 | SWISU | $133, 590 |
| 23 | Sawyer | $136, 746 |
| 24 | Somerst | $229, 707 |
| 25 | StoneBr | $324, 229 |
| 26 | Timber | $246, 600 |
| 27 | Veenker | $248, 315 |
| 28 | SawyerW | $185, 665 |

Neighborhood is a variable that cannot be upgraded after initial purchase by an individual, unless a high quality home can be physically moved to a new neighborhood which comes at great expense. Average sale prices are provided by neighborhood to assist in selection based on investment capability of the individual. Lower cost neighborhoods may provide larger budgeting for upgrades, however, additional investigation is required on specific upgrade impacts by neighborhood to understand granular impacts of individual options.

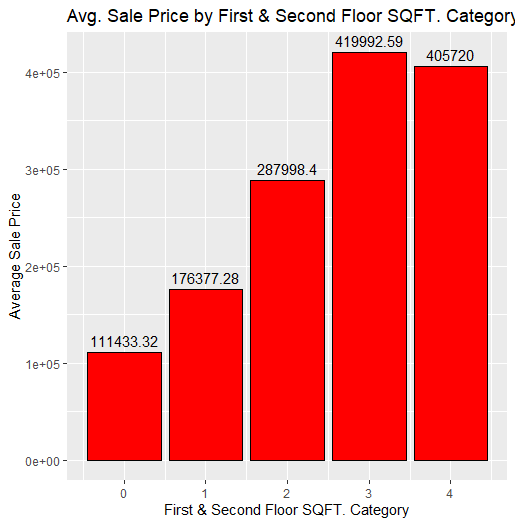
**Wood Deck Square Footage** **(Categorized)**



|  |  |
| --- | --- |
| Wood Deck SF | Average Sale Price |
| 0 (No Deck) | $177, 744 |
| 1 (250 sq.ft) | $214, 915 |
| 2 (500 sq.ft) | $252, 268 |
| 3 (750 sq.ft) | $484, 750 |
| 4 (1000 sq.ft) | $188, 000 |

Outdoor gathering space between the 500 and 750 square foot area have the greatest impact when growing from category 2 to 3. Given the area required to provide those additional spaces, it stands to reason that a house on a larger plot in a less urbanized neighborhood would benefit the most from installation or improvement of an outdoor space.

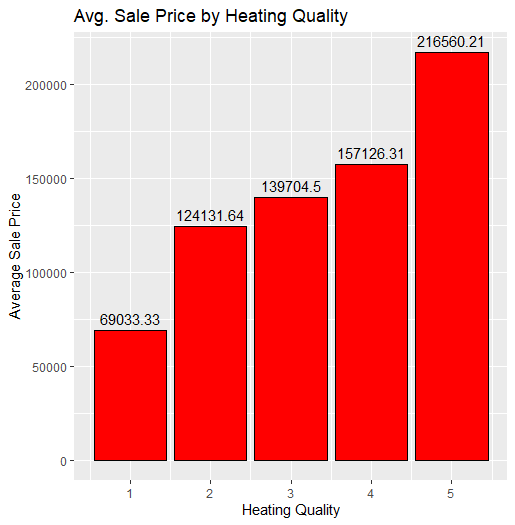
**First/Second Floor Square Footage (Categorized)**



|  |  |
| --- | --- |
| 1st/2nd Sq Ft | Average Sale Price |
| 0 (No 2nd Floor) | $111, 433 |
| 1 (1000-1999 sf) | $176, 377 |
| 2 (2000-2999 sf) | $287, 998 |
| 3 (3000-3999 sf) | $419, 993 |
| 4 (>= 4000 sf) | $750, 000 |

Square footage is a difficult area to adjust or upgrade as it typically involves expansion on the outside structure of the home. Additional foundation modifications required to expand square footage, or conversion of existing non-living space (garage or basement), must be considered and balanced against the desire for additional living area when traded for functional space.

**Heating Quality**



|  |  |
| --- | --- |
| Heating Quality | Average Sale Price |
| 1 (Poor) | $69, 033 |
| 2 (Fair) | $124, 132 |
| 3 (Typical) | $139, 705 |
| 4 (Good) | $157, 126 |
| 5 (Excellent) | $216, 642 |

Of the selected variables, Heating Quality had the lowest impact on Sale Price. The potential gain from installing or upgrading a higher quality system does not necessarily outweigh the additional cost of a home with an adequate system already installed. Upgrading from a category 1 to category 2 has relatively the same financial impact as upgrading from category 4 to category 5 (~$55K).

**How does square footage impact SalePrice?**

Chart, scatter chart

Description automatically generatedInitial plot of Sq Ft vs. Price

Chart, scatter chart

Description automatically generatedRemove outliers (SF > 4500)

Square footage, as expected, has a linear relationship with Sale Price; as Square footage increases, Sale Price increases at a similar rate. In the two scatter plots, the primary concentration of sales occurred for houses below 2500 square feet, between $100K and $300K. Square footage of the home is fixed (outside of remodeling and expansion); additional value in the property must be developed with internal and external updates to categorically raise the worth of individual components.

**What external modifications can be made with the greatest ROI?**

Upgrading external quality, installing new paneling, or replacing weather beaten siding provides the greatest increase in SalePrice. Exterior Quality is categorized from Poor to Excellent; upgrading from Good to Excellent provides the greatest impact on average sale price (~$147K), while Typical to Good is perhaps more attainable depending on available funds and investment capability (~$90K).

**What internal modifications can be made with the greatest ROI?**

Depending on the initial quality of internal systems, Kitchen Quality has the greatest impact on Sale Price regardless of the category increase over Basement Quality or Heating Quality. On average, Kitchen improvements returned ~$40-50K increase in Sale Price per increase in Quality category.

**Of the selected variables, which has the least impact on SalePrice?**

Per the developed linear regression model, Heating Quality has the least impact as categories increase. With a coefficient value of only $1,774, Heating Quality does not provide adequate return on investment for middle-tier categories. Only when increasing from Poor to Fair, or from Good to Excellent, are the returns realized for a potentially large investment.

**For various available budgets, what recommendations could be made for current house configuration and potential improvements for greatest ROI?**

Graphical user interface, text

Description automatically generated

Prediction testing consisted of price modeling of an ideal home purchase for a first-time buyer with an initial budget of $350,000. Focusing on Wood Deck category, Exterior Quality, Heating Quality, Square Footage, Basement Quality and Kitchen Quality; predictive analysis provides a selection of variables that can be changed in the model to determine future returns. Expected ROI for the initial purchase price of ~$350K, with the ability to change all selected features including Square Footage, predicts a return value of $156, 597, approximately 150% growth in Sale Price.

**R CODE**

# Import libraries

library**(**tidyverse**)**

library**(**ggplot2**)**

library**(**RCurl**)**

library**(**jsonlite**)**

library**(**ggmap**)**

library**(**maps**)**

library**(**mapproj**)**

library**(**imputeTS**)**

library**(**RColorBrewer**)**

library**(**stringr**)**

library**(**e1071**)**

library**(**rpart**)**

library**(**rpart.plot**)**

library**(**kernlab**)**

library**(**caret**)**

library**(**corrplot**)**

# install.packages('corrplot')

# Read in Ames Data

df **<-** read.csv**(**file **=** 'C:/Users/rjran/OneDrive/Documents/MADS/IST 687 - Intro to Data Science/Assignments/Final Project/AmesHousing.csv'**)**

# Visualize first 5 rows of dataset

head**(**df, 5**)**

# Look at structure of dataset

str**(**df**)**

# Glimpse (tidyverse)

glimpse**(**df**)**

# Subsetting selected columns

df **<-** subset**(**df, select **=** c**(**PID, MS.SubClass, Neighborhood, Year.Built, Exter.Qual, Bsmt.Qual, Heating.QC, Kitchen.Qual, Garage.Qual, Wood.Deck.SF, Lot.Area, X1st.Flr.SF, X2nd.Flr.SF, Full.Bath, Half.Bath, SalePrice**))**

# Removing duplex

df **<-** df**[**df**$**MS.SubClass **!=** 090,**]**

# Dummifying Neighborhood column

df**$**Neighborhood **<-** ifelse**(**df**$**Neighborhood **==** "Blmngtn", 1,

ifelse**(**df**$**Neighborhood **==** "Blueste", 2,

ifelse**(**df**$**Neighborhood **==** "BrDale", 3,

ifelse**(**df**$**Neighborhood **==** "BrkSide", 4,

ifelse**(**df**$**Neighborhood **==** "ClearCr", 5,

ifelse**(**df**$**Neighborhood **==** "CollgCr", 6,

ifelse**(**df**$**Neighborhood **==** "Crawfor", 7,

ifelse**(**df**$**Neighborhood **==** "Edwards", 8,

ifelse**(**df**$**Neighborhood **==** "Gilbert", 9,

ifelse**(**df**$**Neighborhood **==** "Greens", 10,

ifelse**(**df**$**Neighborhood **==** "GrnHill", 11,

ifelse**(**df**$**Neighborhood **==** "IDOTRR", 12,

ifelse**(**df**$**Neighborhood **==** "Landmrk", 13,

ifelse**(**df**$**Neighborhood **==** "MeadowV", 14,

ifelse**(**df**$**Neighborhood **==** "Mitchel", 15,

ifelse**(**df**$**Neighborhood **==** "NAmes", 16,

ifelse**(**df**$**Neighborhood **==** "NoRidge", 17,

ifelse**(**df**$**Neighborhood **==** "NPkVill", 18,

ifelse**(**df**$**Neighborhood **==** "NridgHt", 19,

ifelse**(**df**$**Neighborhood **==** "NWAmes", 20,

ifelse**(**df**$**Neighborhood **==** "OldTown", 21,

ifelse**(**df**$**Neighborhood **==** "SWISU", 22,

ifelse**(**df**$**Neighborhood **==** "Sawyer", 23,

ifelse**(**df**$**Neighborhood **==** "Somerst", 24,

ifelse**(**df**$**Neighborhood **==** "StoneBr", 25,

ifelse**(**df**$**Neighborhood **==** "Timber", 26,

ifelse**(**df**$**Neighborhood **==** "Veenker", 27,

ifelse**(**df**$**Neighborhood **==** "SawyerW", 28,

0**))))))))))))))))))))))))))))**

# Dummifying exter.qual

df**$**Exter.Qual **<-** ifelse**(**df**$**Exter.Qual **==** "Ex", 5,

ifelse**(**df**$**Exter.Qual **==** "Gd", 4,

ifelse**(**df**$**Exter.Qual **==** "TA", 3,

ifelse**(**df**$**Exter.Qual **==** "Fa", 2,

ifelse**(**df**$**Exter.Qual **==** "Po", 1,

0**)))))**

# Removing any nulls in exter.qual

df**$**Exter.Qual**[**is.na**(**df**$**Exter.Qual**)]** **<-** 0

# Dummifying bsmt.qual

df**$**Bsmt.Qual **<-** ifelse**(**df**$**Bsmt.Qual **==** "Ex", 5,

ifelse**(**df**$**Bsmt.Qual **==** "Gd", 4,

ifelse**(**df**$**Bsmt.Qual **==** "TA", 3,

ifelse**(**df**$**Bsmt.Qual **==** "Fa", 2,

ifelse**(**df**$**Bsmt.Qual **==** "Po", 1,

0**)))))**

# Removing any nulls in bsmt.qual

df**$**Bsmt.Qual**[**is.na**(**df**$**Bsmt.Qual**)]** **<-** 0

# Dummifying heating.qc

df**$**Heating.QC **<-** ifelse**(**df**$**Heating.QC **==** "Ex", 5,

ifelse**(**df**$**Heating.QC **==** "Gd", 4,

ifelse**(**df**$**Heating.QC **==** "TA", 3,

ifelse**(**df**$**Heating.QC **==** "Fa", 2,

ifelse**(**df**$**Heating.QC **==** "Po", 1,

0**)))))**

# Removing any nulls in heating.qc

df**$**Heating.QC**[**is.na**(**df**$**Bsmt.Qual**)]** **<-** 0

# Dummifying kitchen.qual

df**$**Kitchen.Qual **<-** ifelse**(**df**$**Kitchen.Qual **==** "Ex", 5,

ifelse**(**df**$**Kitchen.Qual **==** "Gd", 4,

ifelse**(**df**$**Kitchen.Qual **==** "TA", 3,

ifelse**(**df**$**Kitchen.Qual **==** "Fa", 2,

ifelse**(**df**$**Kitchen.Qual **==** "Po", 1,

0**)))))**

# Removing any nulls in kitchen.qual

df**$**Kitchen.Qual**[**is.na**(**df**$**Kitchen.Qual**)]** **<-** 0

# Dummifying garage.qual

df**$**Garage.Qual **<-** ifelse**(**df**$**Garage.Qual **==** "Ex", 5,

ifelse**(**df**$**Garage.Qual **==** "Gd", 4,

ifelse**(**df**$**Garage.Qual **==** "TA", 3,

ifelse**(**df**$**Garage.Qual **==** "Fa", 2,

ifelse**(**df**$**Garage.Qual **==** "Po", 1,

0**)))))**

# Removing any nulls in garage.qual

df**$**Garage.Qual**[**is.na**(**df**$**Garage.Qual**)]** **<-** 0

# Combine full bath and half bath. To combine, all half baths will be equal to 0.1.

df**$**Half.Bath **<-** df**$**Half.Bath **\*** 0.1

# Create new column, total\_baths. 2.1 will mean two full, one half bathroom. 3.4 means three full, four half bathrooms.

df**$**Total.Baths **<-** df**$**Full.Bath **+** df**$**Half.Bath

# Create new column to look at first and second floor square footage.

df**$**First.Second.SF **<-** df**$**X1st.Flr.SF **+** df**$**X2nd.Flr.SF

# Remove Full.Bath and Half.Bath now that we used the two columns to create Total\_Baths. Also remove X1st/X2nd Flr SF now that we used the two columns to create First\_Second\_SF

df **<-** subset**(**df, select **=** **-**c**(**Full.Bath, Half.Bath, X1st.Flr.SF, X2nd.Flr.SF**))**

# Organization of columns

df **<-** subset**(**df, select **=** c**(**PID, MS.SubClass, Neighborhood, Year.Built, Exter.Qual, Bsmt.Qual, Heating.QC, Kitchen.Qual, Garage.Qual, Wood.Deck.SF, Lot.Area, First.Second.SF, Total.Baths, SalePrice**))**

head**(**df, 5**)**

# BUSINESS QUESTION: WHAT VARIABLES HAVE THE GREATEST IMPACT ON SALE PRICE?

# Create correlation matrix to identify most correlated features with SalePrice, both negatively and positively correlated

res **<-** cor**(**df**)**

corrplot**(**res, type **=** "upper", order **=** "hclust", tl.col **=** "black", tl.sort **=** 45**)**

correlation\_matrix **<-** cor**(**df**)**

correlation\_matrix

# BUSINESS QUESTION: WHAT IMPROVEMENTS CAN BE MADE WITH THE GREATEST ROI?

# Quality of exterior grade by sale price

exter.qual.df **<-** df %>% group\_by**(**Exter.Qual**)** %>%

summarise**(**mean\_salary**=**mean**(**SalePrice**)**,

.groups **=** 'drop'**)**

# Almost 100K jump from 3-4, <100k jump from 4-5

# Maybe a condo because you have professionals assist, or condos will be more expensive to new homebuyers because of upkeep

exter.qual.df

# Quality of kitchen grade by sale price

kitchen.qual.df **<-** df %>% group\_by**(**Kitchen.Qual**)** %>%

summarise**(**mean\_salary**=**mean**(**SalePrice**)**,

.groups **=** 'drop'**)**

# 1 AND 2 NO DIFFERENT

# Investment wise or first time home buyer, buy a 3 and improve to 4

kitchen.qual.df

# Quality of basement grade by sale price

bsmt.qual.df **<-** df %>% group\_by**(**Bsmt.Qual**)** %>%

summarise**(**mean\_salary**=**mean**(**SalePrice**)**,

.groups **=** 'drop'**)**

# IT IS BETTER TO HAVE NO BASEMENT THAN A BAD BASEMENT

# 3-5 jump. I would spend extra money to get a 3 basement grade then improve

bsmt.qual.df

total.baths.df **<-** df %>% group\_by**(**Total.Baths**)** %>%

summarise**(**mean\_salary**=**mean**(**SalePrice**)**,

.groups **=** 'drop'**)**

# Half.Baths add about 25k

# Adding a second half bath only worked once, stick with 1

total.baths.df

# BUSINESS QUESTION: HOW DOES SQUARE FOOTAGE IMPACT SALES PRICE?

# Strongest correlated feature: First.Second.SF

# Outliers identified, will query and remove

df %>%

ggplot**(**aes**(**SalePrice, First.Second.SF**))** **+**

geom\_point**()** **+** geom\_smooth**(**method **=** "lm"**)** **+**

labs**(**title **=** "Linear Model",

subtitle **=** "Price per Square Foot",

x **=** "Sales Price",

y **=** "1st + 2nd Floor Square Footage"**)**

# Removal of outliers in First.Second.SF

df **<-** df**[**df**$**First.Second.SF **<** 4500,**]**

# Strongest correlated feature: First.Second.SF with outliers removed

# Distinct correlation, as sales price increases, First.Second.SF also increases

df %>%

ggplot**(**aes**(**SalePrice, First.Second.SF**))** **+**

geom\_point**()** **+** geom\_smooth**(**method **=** "lm"**)** **+**

labs**(**title **=** "Linear Model",

subtitle **=** "Price per Square Foot",

x **=** "Sales Price",

y **=** "1st + 2nd Floor Square Footage"**)**

# BUSINESS QUESTION: WHAT EXTERNAL MODIFICATIONS CAN BE MADE WITH THE GREATEST ROI?

# Neighborhood, Exter.Qual, Wood.Deck.SF

# Neighborhood by Sale Price

neighborhood.df **<-** df %>% group\_by**(**Exter.Qual, Neighborhood**)** %>%

summarise**(**Average\_SalePrice**=**mean**(**SalePrice**)**,

.groups **=** 'drop'**)**

# If I were looking to find the cheapest possible neighborhood, I would select neighborhood 14.

print**(**neighborhood.df, n **=** 30**)**

#Neighborhood by Average Sale Price

neighborhood.avg **<-** df %>% group\_by**(**Neighborhood**)** %>%

summarise**(**Average\_SalePrice **=** mean**(**SalePrice**)**,

.groups **=** 'drop'**)**

#Print results

print**(**neighborhood.avg, n **=** 28**)**

# Quality of exterior grade by sale price

# Almost 100K jump from 3-4, <100k jump from 4-5

# Maybe a condo because you have professionals assist, or condos will be more expensive to new homebuyers because of upkeep

# 2 --> 3 $55k

# 3 --> 4 $87k

# 4 --> 5 $147k

exter.qual.df

exter.qual.ggplot **<-** df %>%

group\_by**(**Exter.Qual**)** %>%

summarise**(**mean\_price **=** mean**(**SalePrice**))**

# Create ggplot object with means plotted by letter grade and data labels

ggplot**(**exter.qual.ggplot, aes**(**x **=** Exter.Qual, y **=** mean\_price**))** **+**

geom\_bar**(**stat **=** "identity", fill **=** "red", color **=** "black"**)** **+**

geom\_text**(**aes**(**label **=** round**(**mean\_price, 2**))**, vjust **=** **-**0.5**)** **+**

labs**(**title **=** "Avg. Sale Price by Exter.Qual", x **=** "Exter.Qual", y **=** "Average Sale Price"**)**

min**(**df**$**Wood.Deck.SF**)**

max**(**df**$**Wood.Deck.SF**)**

# Dummifying wood deck into categories to compare

df**$**Wood.Deck.SF.Category **<-** ifelse**(**df**$**Wood.Deck.SF **>=** 1000, 4,

ifelse**(**df**$**Wood.Deck.SF **>=** 750, 3,

ifelse**(**df**$**Wood.Deck.SF **>=** 500, 2,

ifelse**(**df**$**Wood.Deck.SF **>=** 250, 1,

0**))))**

# Create visual to view wood deck sf in categories by sale price

wood.deck.df **<-** df %>% group\_by**(**Wood.Deck.SF.Category**)** %>%

summarise**(**Average\_SalePrice**=**mean**(**SalePrice**)**,

.groups **=** 'drop'**)**

# Improvement from a 2-3 in terms of deck square footage is an increase in over $200k

wood.deck.df

wood.deck.ggplot **<-** df %>%

group\_by**(**Wood.Deck.SF.Category**)** %>%

summarise**(**mean\_price **=** mean**(**SalePrice**))**

# Create ggplot object with means plotted by letter grade and data labels

ggplot**(**wood.deck.ggplot, aes**(**x **=** Wood.Deck.SF.Category, y **=** mean\_price**))** **+**

geom\_bar**(**stat **=** "identity", fill **=** "red", color **=** "black"**)** **+**

geom\_text**(**aes**(**label **=** round**(**mean\_price, 2**))**, vjust **=** **-**0.5**)** **+**

labs**(**title **=** "Avg. Sale Price by Wood Deck SF Category", x **=** "Wood Deck SF Category", y **=** "Average Sale Price"**)**

# Dummifying First.Second.SF into categories to compare

df**$**First.Second.SF.Categories **<-** ifelse**(**df**$**First.Second.SF **>=** 4000, 4,

ifelse**(**df**$**First.Second.SF **>=** 3000, 3,

ifelse**(**df**$**First.Second.SF **>=** 2000, 2,

ifelse**(**df**$**First.Second.SF **>=** 1000, 1,

0**))))**

first.second.sf.categories **<-** df %>% group\_by**(**First.Second.SF.Categories**)** %>%

summarise**(**Average\_SalePrice**=**mean**(**SalePrice**)**,

.groups **=** 'drop'**)**

first.second.sf.categories

first.second.sf.ggplot **<-** df %>%

group\_by**(**First.Second.SF.Categories**)** %>%

summarise**(**mean\_price **=** mean**(**SalePrice**))**

# Create ggplot object with means plotted by letter grade and data labels

ggplot**(**first.second.sf.ggplot, aes**(**x **=** First.Second.SF.Categories, y **=** mean\_price**))** **+**

geom\_bar**(**stat **=** "identity", fill **=** "red", color **=** "black"**)** **+**

geom\_text**(**aes**(**label **=** round**(**mean\_price, 2**))**, vjust **=** **-**0.5**)** **+**

labs**(**title **=** "Avg. Sale Price by First & Second Floor SQFT. Category", x **=** "First & Second Floor SQFT. Category", y **=** "Average Sale Price"**)**

# BUSINESS QUESTION: WHAT INTERNAL MODIFICATIONS CAN BE MADE WITH THE GREATEST ROI?

# 1 -> 2 -754

# 2 -> 3 32k

# 3 -> 4 71k

# 4 -> 5 126k

kitchen.qual.df

kitchen.qual.ggplot **<-** df %>%

group\_by**(**Kitchen.Qual**)** %>%

summarise**(**mean\_price **=** mean**(**SalePrice**))**

# Create ggplot object with means plotted by letter grade and data labels

ggplot**(**kitchen.qual.ggplot, aes**(**x **=** Kitchen.Qual, y **=** mean\_price**))** **+**

geom\_bar**(**stat **=** "identity", fill **=** "red", color **=** "black"**)** **+**

geom\_text**(**aes**(**label **=** round**(**mean\_price, 2**))**, vjust **=** **-**0.5**)** **+**

labs**(**title **=** "Avg. Sale Price by Kitchen Quality", x **=** "Kitchen Quality", y **=** "Average Sale Price"**)**

# 0 -> 1 -21k

# 1 -> 2 25k

# 2 -> 3 30k

# 3 -> 4 63k

# 4 -> 5 134k

bsmt.qual.df

bsmt.qual.ggplot **<-** df %>%

group\_by**(**Bsmt.Qual**)** %>%

summarise**(**mean\_price **=** mean**(**SalePrice**))**

# Create ggplot object with means plotted by letter grade and data labels

ggplot**(**bsmt.qual.ggplot, aes**(**x **=** Bsmt.Qual, y **=** mean\_price**))** **+**

geom\_bar**(**stat **=** "identity", fill **=** "red", color **=** "black"**)** **+**

geom\_text**(**aes**(**label **=** round**(**mean\_price, 2**))**, vjust **=** **-**0.5**)** **+**

labs**(**title **=** "Avg. Sale Price by Basement Quality", x **=** "Basement Quality", y **=** "Average Sale Price"**)**

heating.qc.df **<-** df %>% group\_by**(**Heating.QC**)** %>%

summarise**(**Average\_SalePrice**=**mean**(**SalePrice**)**,

.groups **=** 'drop'**)**

heating.qc.df

heating.qc.ggplot **<-** df %>%

group\_by**(**Heating.QC**)** %>%

summarise**(**mean\_price **=** mean**(**SalePrice**))**

# Create ggplot object with means plotted by letter grade and data labels

ggplot**(**heating.qc.ggplot, aes**(**x **=** Heating.QC, y **=** mean\_price**))** **+**

geom\_bar**(**stat **=** "identity", fill **=** "red", color **=** "black"**)** **+**

geom\_text**(**aes**(**label **=** round**(**mean\_price, 2**))**, vjust **=** **-**0.5**)** **+**

labs**(**title **=** "Avg. Sale Price by Heating Quality", x **=** "Heating Quality", y **=** "Average Sale Price"**)**

# 1 -> 2 55k !!!! BIG ROI, LARGEST JUMP OUTSIDE 4 -> 5

# 2 -> 3 15k

# 3 -> 4 18k

# 4 -> 5 59k

heating.qc.df

# Adding a second half bath decreased cost every single time except for homes with only one full

# KEY POINT: IF YOU HAVE MORE THAN ONE FULL BATHROOM, DO NOT ADD MORE THAN ONE FULL BATH.

total.baths.df

total.baths.ggplot **<-** df %>%

group\_by**(**Total.Baths**)** %>%

summarise**(**mean\_price **=** mean**(**SalePrice**))**

# Create ggplot object with means plotted by letter grade and data labels

# DONT BUILD A SECOND HALF BATH UNLESS ONLY 1 FULL

ggplot**(**total.baths.ggplot, aes**(**x **=** Total.Baths, y **=** mean\_price**))** **+**

geom\_bar**(**stat **=** "identity", fill **=** "red", color **=** "black"**)** **+**

geom\_text**(**aes**(**label **=** round**(**mean\_price, 2**))**, vjust **=** **-**0.5**)** **+**

labs**(**title **=** "Avg. Sale Price by Total Number of Baths (First digit - Full Bathrooms, Second digit - Half baths)", x **=** "Number of Baths", y **=** "Average Sale Price"**)**

# BUSINESS QUESTION: OF THE SELECTED VARIABLES WHICH HAS THE LEAST IMPACT ON SALES PRICE?

# When looking at the correlation matrix, the MS.SubClass has the least impact on sales price. MS.SubClass in the correlation matrix has a correlation of -0.07.

SalePrice\_Corr **<-** cor**(**df**[-**1**]**, df**$**SalePrice**)**

SalePrice\_Corr

# BUSINESS QUESTION: FOR VARIOUS AVAILABLE BUDGETS, WHAT RECOMMENDATIONS COULD BE MADE FOR CURRENT HOUSE CONFIGURATION AND POTENTIAL IMPROVEMENTS FOR GREATEST ROI?

# Exterior: For the greatest ROI potential

# Wood.Deck.SF.Category --> 2, 2-3 232k increase

# Exterior.Qual --> 4, 4-5 147k increase

# Heating.QC --> 1, 1-2 55k

# First.Second.SF.Category --> 3, 3-4 330k increase

# Bsmt.Qual --> 4, 4-5 134k increase

# Kitchen.Qual --> 4-5, 126k increase

model\_one **<-** lm**(**SalePrice **~** Wood.Deck.SF.Category **+** Exter.Qual **+** Heating.QC **+** First.Second.SF.Categories **+** Bsmt.Qual **+** Kitchen.Qual, data **=** df**)**

# View summary of the model

summary**(**model\_one**)**

model\_one

# Use input data to test against the model

perfect\_roi\_home **=** data.frame**(**Wood.Deck.SF.Category **=** 2,

Exter.Qual **=** 4,

Heating.QC **=** 1,

First.Second.SF.Categories **=** 3,

Bsmt.Qual **=** 4,

Kitchen.Qual **=** 4**)**

perfect\_roi\_home\_finished **=** data.frame**(**Wood.Deck.SF.Category **=** 3,

Exter.Qual **=** 5,

Heating.QC **=** 2,

First.Second.SF.Categories **=** 4,

Bsmt.Qual **=** 5,

Kitchen.Qual **=** 5**)**

# Utilize predict function to call the model you have created, a test dataframe (test), and type="response"

# Predicted price to get the "perfect ROI potential home"

predict**(**model\_one, perfect\_roi\_home, type**=**"response"**)**

# Predicted price once the "perfect ROI potential home" is completed

predict**(**model\_one, perfect\_roi\_home\_finished, type**=**"response"**)**

# Investment profit (ROI) of $156597, now to account for labor and materials.

print**(**517200**-**360603**)**