MSDS 6372 Project 1:

Building Regression Models to Predict House Prices

Rajat Chandna

Jonathan Marin

Rene Pineda

Samira Zarandioon

June 16, 2018

# Introduction

# When buying a house, we look at a lot of features, its square feet area, number of bedrooms, bathrooms, frontyards & backyards, location, the look of the kitchen, location, etc. As it happens, the price itself is dependent on many factors. This project tries to understand the underlying factors that go into determined the sale price of houses. The goal of this project is to build a predictive model that can identify and capture the variance in the data as accurately as possible in order to use this data to predict future house prices.

# We are using a dataset of houses in Ames, Iowa. This dataset provides many features of a house that would give a better idea of what and how much do each of these features affect the final sale price. The dataset there are 80 explanatory variables describing every aspect of residential homes in Ames, Iowa such as Street, Neighborhood, LotShape, LandSlope, YearBuilt, FullBath, GarageCars, Fireplaces and Pool Quality, etc. for 2,930 homes.

# We are trying to answer 2 questions:

1. What are the important features and factors that impact house prices, and can we build a model that can accurately predict the price of a house? To answer this, we built many multiple variable regression models, compared their relative performance, and chose the model that did a better job at prediction. We also interpret this model and perform hypothesis testing on it.

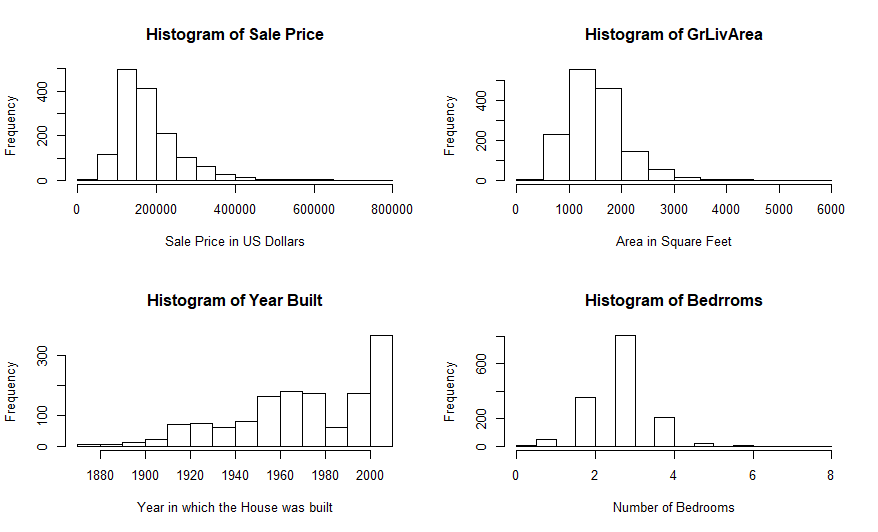
2. If we were to use only two important categorical variables to estimate and predict the price of a house, how can we assess such a model and estimate the influence of each of the two variables in house prices? To answer this, we chose the “Kitchen Quality” and “Neighborhood” variables, and perform two-way ANOVA analysis.

# Data Description

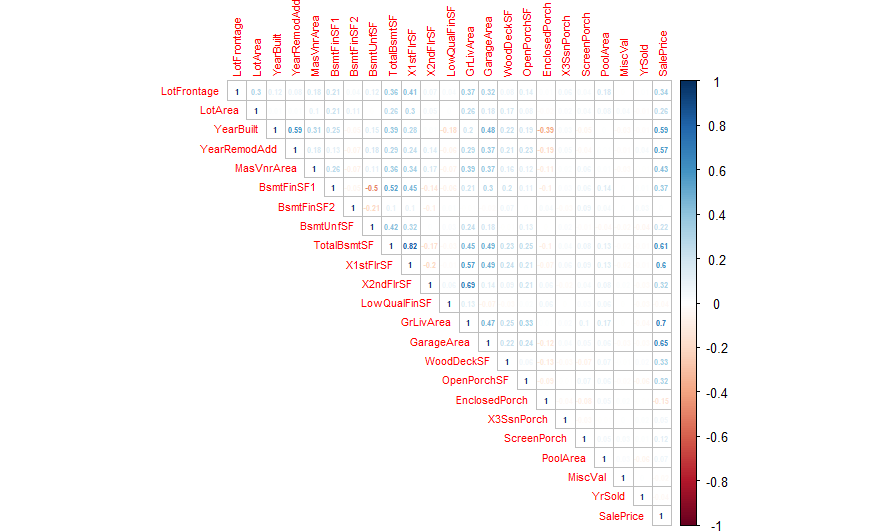
The data set (available at http://www.amstat.org/publications/jse/v19n3/decock/AmesHousing.txt) contains information from the Ames Assessor’s Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010. The dataset has 1460 observations of 80 variables related to the house, its features, and surroundings: 23 nominal, 23 ordinal, 14 discrete, and 20 continuous. A summary table of the variables can be found in the Appendix.

# Exploratory Analysis

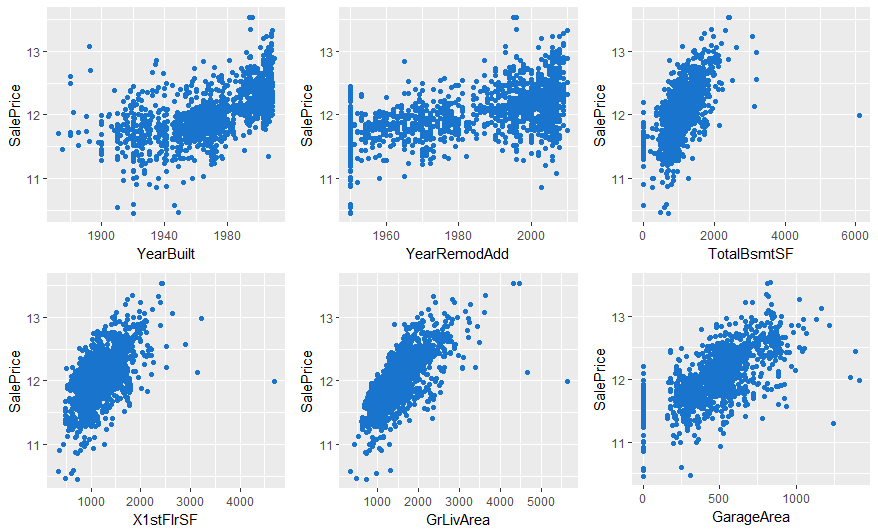
First, we’ll try to understand general features of the houses been sold. Doing some summary statistics and charts, we can draw the following conclusions: the median sale price is $163,000, most of the houses sold have an above the ground area between 1,000 and 2,000 sq ft, a large proportion of houses are fairly recent (built after the year 2000) and 3-bedroom houses are the most popular.



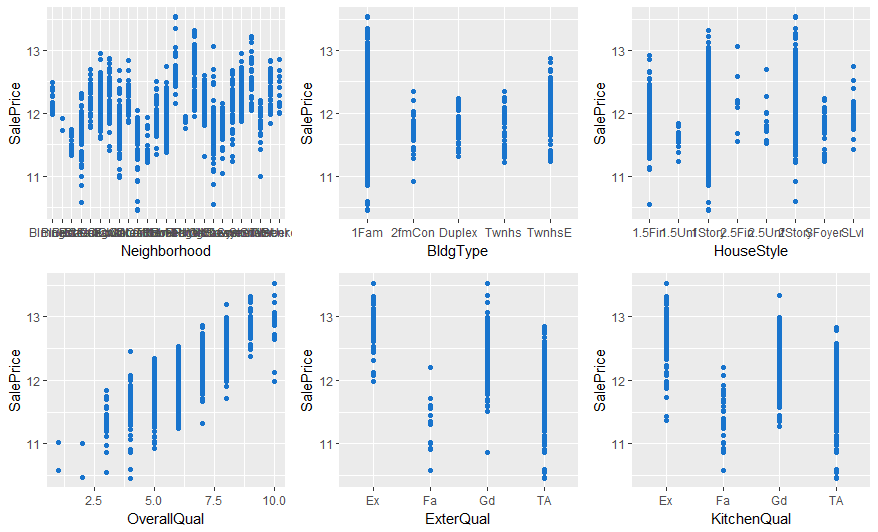
# We also want to understand which numeric variables are strongly correlated with the sale price, and whether multicollinearity may exist. For this, we create a correlation matrix as the one shown below:



# As expected, we find that variables related to the size of the home (TotalBSmtSF, X1stFlrSF, GrLivArea, Garage Area) are strongly, positively correlated with price, and at the same time these variables show moderate to strong correlations with each other. Another important correlation is related to the year that the home was build and remodeled (we also found that 42% of the houses were not remodeled, and that the average number of years before building and adding/remodeling is 13.6)



# We also want to understand which categorical variables can influence Sale Price. We present selected scatterplots below:



# Addressing Objective 1:

## Restatement of Problem

We want to create a model that accurately predicts the sale price of a house in Ames, Iowa, and identify which features of a home are more influential in determining its Sale Price.

The overall strategy will be: 1. Create different models using widely adopted regression techniques, 2. Check the model’s assumptions and perform influential point analysis, 3. Assess and compare each model’s predicting power, 4. Select the model and features that have the better performance based on predetermined criteria.

## Model Selection

We created 5 models:

1. An initial model that includes all variables in the dataset

2. A LASSO Model

3. Models using automated variable selection techniques (forward, backwards and stepwise), using the p-value as criteria for including and removing variables.

## Checking Assumptions

|  |  |  |  |
| --- | --- | --- | --- |
| All Variables | Forward Selection | Backward elimination | Stepwise Selection |
| Residual Plots (Sample only, have to substitute with the actual plots) | | | |
|  |  |  |  |
| The residual plot of saleprice to predicted values resembles to dense cloud around 0. There is some suspicion of non-constant variance. | | | |
| Histogram of residuals with the normal distribution superimposed. | | | |
|  |  |  |  |
| Text here | | | |
| Q-Q Plot of the Residuals | | | |
|  |  |  |  |
| Text Here | | | |
| Influential point analysis: Leverage and Cook’s D | | | |
|  |  |  |  |
|  |  |  |  |
| Text here | | | |
| **Conclusion:** based on the graphs, we do not see strong evidence against normality and equality of variance, although we notice the presence of some outliers that will require further study, both in the Q-Q plots, leverage and the Cook’s D plots. Independence is questionable as houses can be sold in the same period of time and neighborhood, belong to the same owner, etc. but we will assume it. | | | |

## Comparing Competing Models

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictive Models** | **Adjusted R2** | **AIC** | **BIC** |
| All features included |  |  |  |
| LASSO |  |  |  |
| Forward |  |  |  |
| Backward |  |  |  |
| Stepwise |  |  |  |
| Custom model |  |  |  |

## Parameter Interpretation

The results of the selected linear regression model are shown below:

| **Parameter** | **Estimate** | **95% Confidence Limits** | | **Interpretation** |
| --- | --- | --- | --- | --- |
| **Intercept** |  |  |  |  |
| **logGrLivArea** |  |  |  |  |
| **Neighborhood BrkSide** |  |  |  |  |
| **Neighborhood Edwards** |  |  |  |  |
| **Neighborhood NAmes** |  |  |  |  |

## Conclusion

Insert text here

# Objecive 2: Two-way ANOVA

## Route of the Analysis

**The Motivation and Design.**

Different people weigh different features quite differently while deciding upon a house purchase and what price they should pay taking into account presence or absence of their highly desired features. Some people want houses in quite neighborhoods, while others want near busy downtowns. Some weigh higher high build qualities more and yet another look for greater Living areas, while deciding on fair price to pay for the houses.

Motivated by my wife’s interest in cooking and kitchen quality, I would like access the influence on kitchen quality on sale prices of houses in different neighborhoods in the city of AMES. **Since it is common knowledge that house sale prices differ from neighborhood to another, we could have chosen neighborhood as a block but since this analysis is in observational setting, we would also like to examine influence of neighborhood on house sale prices. Hence, we are treating both Kitchen Quality and Neighborhood are our independent variables and House Sale Price as dependent variable**.

The dataset for the analysis is “CleanedDataForANOVA.csv”, which is a dataset obtained after initial cleaning on Kaggle House Sale Prices data set.

**The DataSet.**

The description and levels of independent variables is as:

KitchenQual: Kitchen quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Neighborhood: Physical locations within Ames city limits

Blmngtn Bloomington Heights

Blueste Bluestem

BrDale Briardale

BrkSide Brookside

ClearCr Clear Creek

CollgCr College Creek

Crawfor Crawford

Edwards Edwards

Gilbert Gilbert

IDOTRR Iowa DOT and Rail Road

MeadowV Meadow Village

Mitchel Mitchell

Names North Ames

NoRidge Northridge

NPkVill Northpark Villa

NridgHt Northridge Heights

NWAmes Northwest Ames

OldTown Old Town

SWISU South & West of Iowa State University

Sawyer Sawyer

SawyerW Sawyer West

Somerst Somerset

StoneBr Stone Brook

Timber Timberland

Veenker Veenker

**The Questions of Interest.**

The main goal of the analysis is to:

1. Determine the influence of 4 different levels of kitchen quality on house sale price. Is the difference between prices paid for houses with better quality kitchens differ significantly than prices paid for houses with not so great quality kitchens, after taking into account of the neighborhoods?
2. Determine the influence of Neighborhood on the house sale prices in city AMES?
3. Determine whether the influence of 4 different levels of kitchen quality on house sale price vary from neighborhood to neighborhood?

**THE Assumption for the analysis.**

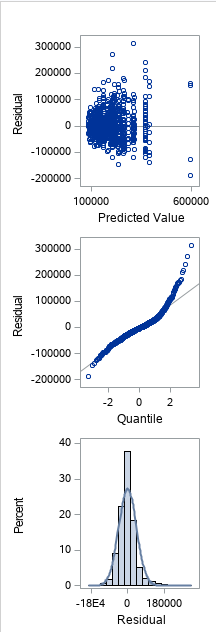
Two-way ANOVA analysis requires observations within different groups and across different groups be independent of one another. But house sale prices within same neighborhood are rarely independent of one another and suffer from spatial correlation. For this analysis, we will assume that house sale prices are independent of each other and proceed with analysis.

## Main Analysis

***Inputting the data and plotting the means plot.***



***Above plot depicting mean sale prices based upon levels of Kitchen quality across different neighborhoods alludes to a non-additive model as difference in mean sale prices among Kitchen Quality appear to be different across different neighborhoods. Also, spread of the error bars hint at higher spreads at higher mean values, thus towards non-constant variance. Residual plot could clarify this trend further.***



***Above residual plot reveals a funnel shaped pattern for distribution of residuals across fitted values. Also, a pattern of deviation towards left side could be seen from Normal QQ Plot. A log transformation would help in rectifying such situation.***

***Also, since houses in given data set are concentrated under 4000 sq ft of GrLiving Area, we would remove any observations that have GrLivingArea > 4000 since such houses are not representative of houses in the dataset while having high influence on sale prices.***



***After log transformation on sales prices and some outliers removal, plotting the mean log sale prices based upon levels of Kitchen quality across different neighborhoods.***



***As seen from plot above, that after log transformation on sales price, the variance is nearly equal across different levels of Kitchen Quality and Neighborhood. Also, plot alludes towards interaction between Kitchen Quality and Neighborhood and hence towards non-additive model in which average log sales prices among levels of Kitchen Quality (from Excellent, Good, Average to Fair) differ with the different Neighborhoods. Hence, running the saturated, non-additive model first.***

| **Class Level Information** | | |
| --- | --- | --- |
| **Class** | **Levels** | **Values** |
| **KitchenQual** | 4 | Ex Fa Gd TA |
| **Neighborhood** | 25 | Blmngtn Blueste BrDale BrkSide ClearCr CollgCr Crawfor Edwards Gilbert IDOTRR MeadowV Mitchel NAmes NPkVill NWAmes NoRidge NridgHt OldTown SWISU Sawyer SawyerW Somerst StoneBr Timber Veenker |

|  |  |
| --- | --- |
| **Number of Observations Read** | 1456 |
| **Number of Observations Used** | 1456 |

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 70 | 155.8014039 | 2.2257343 | 42.54 | **<.0001** |
| **Error** | 1385 | 72.4578965 | 0.0523162 |  |  |
| **Corrected Total** | 1455 | 228.2593005 |  |  |  |

| **R-Square** | **Coeff Var** | **Root MSE** | **LogSalePrice Mean** |
| --- | --- | --- | --- |
| 0.682563 | 1.902582 | 0.228727 | 12.02194 |

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **KitchenQual** | 3 | 7.89484489 | 2.63161496 | 50.30 | **<.0001** |
| **Neighborhood** | 24 | 23.12545364 | 0.96356057 | 18.42 | **<.0001** |
| **KitchenQu\*Neighborho** | 43 | 5.22502408 | 0.12151219 | 2.32 | **<.0001** |

***As we see from above tables that:***

1. ***Overall Model with Kitchen Quality and Neighborhood explaining/predicting log house sales prices is significant at alpha = 0.05, with p-value < 0.0001.***
2. ***Variables Kitchen Quality and Neighborhood combined explain about 68.2% variability in sale prices. This practical effect size is quite significant.***
3. ***As seen from Type 3 SS table, the effects of Kitchen Quality are significant even after accounting for effects of Neighborhood and interaction. Also, effects of Neighborhood are significant even after accounting for effects of Kitchen Quality and interaction. And finally, so does the effects of interaction are significant after accounting for effects of Neighborhood and Kitchen Quality.***

***Now since interaction is significant, the interpretation of the main effects would not be correct in isolation to interaction. We will perform multiple comparisons in presence of interaction once we find out interaction is important. We will first validate model assumptions.***



***As seen from residual plots above, the residuals are fairly randomly distributed across fitted values. The spread appears to be small as compared to other levels at value 13(log sale price) but that’s just because there are only two values at this value. The QQ Plot and Histogram reveal fairly normally distributed residuals at different levels. So, our model assumptions are met.***

## Appendix

Code for Analysis 1

Code for Analysis 2

Dataset description

|  |  |  |  |
| --- | --- | --- | --- |
| **Discrete Variables** | **Nominal Variables** | **Ordinal** | **Continuous** |
| Order: Observation number | MS SubClass: Identifies the type of dwelling involved in the sale. | Lot Shape: General shape of property | Lot Frontage: Linear feet of street connected to property |
| Year Built : Original construction date | PID: Parcel identification number - can be used with city web site for parcel review. | Utilities: Type of utilities available | Lot Area: Lot size in square feet |
| Year Remod/Add : Remodel date (same as construction date if no remodeling or additions) | Street: Type of road access to property | Land Slope: Slope of property | Mas Vnr Area: Masonry veneer area in square feet |
| Bsmt Full Bath : Basement full bathrooms | Alley: Type of alley access to property | Overall Qual: Rates the overall material and finish of the house | BsmtFin SF 1: Type 1 finished square feet |
| Bsmt Half Bath : Basement half bathrooms | Land Contour: Flatness of the property | Overall Cond: Rates the overall condition of the house | BsmtFin SF 2: Type 2 finished square feet |
| Full Bath : Full bathrooms above grade | Lot Config: Lot configuration | Exter Qual: Evaluates the quality of the material on the exterior | Bsmt Unf SF: Unfinished square feet of basement area |
| Half Bath : Half baths above grade | Neighborhood: Physical locations within Ames city limits (map available) | Exter Cond: Evaluates the present condition of the material on the exterior | Total Bsmt SF: Total square feet of basement area |
| Bedroom : Bedrooms above grade (does NOT include asement bedrooms) | MS Zoning: Identifies the general zoning classification of the sale. | Bsmt Qual: Evaluates the height of the basement | 1st Flr SF: First Floor square feet |
| Kitchen : Kitchens above grade | Condition 1: Proximity to various conditions | Bsmt Cond: Evaluates the general condition of the basement | 2nd Flr SF : Second floor square feet |
| TotRmsAbvGrd: Total rooms above grade (does not include bathrooms) | Condition 2: Proximity to various conditions (if more than one is present) | Bsmt Exposure : Refers to walkout or garden level walls | Low Qual Fin SF: Low quality finished square feet (all floors) |
| Fireplaces : Number of fireplaces | Bldg Type: Type of dwelling | BsmtFin Type 1 : Rating of basement finished area | Gr Liv Area: Above grade (ground) living area square feet |
| Garage Yr Blt : Year garage was built | House Style: Style of dwelling | BsmtFinType 2 : Rating of basement finished area (if multiple types) | Garage Area: Size of garage in square feet |
| Garage Cars : Size of garage in car capacity | Roof Style: Type of roof | HeatingQC: Heating quality and condition | Wood Deck SF: Wood deck area in square feet |
| Mo Sold : Month Sold (MM) | Roof Matl: Roof material | Electrical: Electrical system | Open Porch SF: Open porch area in square feet |
| Yr Sold : Year Sold (YYYY) | Exterior 1: Exterior covering on house | KitchenQual: Kitchen quality |  |
|  | Exterior 2: Exterior covering on house (if more than one material) | Functional: Home functionality (Assume typical unless deductions are warranted) | Enclosed Porch: Enclosed porch area in square feet |
|  | Mas Vnr Type: Masonry veneer type | FireplaceQu: Fireplace quality | 3-Ssn Porch: Three season porch area in square feet |
|  | Foundation: Type of foundation | Garage Finish : Interior finish of the garage | Screen Porch: Screen porch area in square feet |
|  | Heating : Type of heating | Garage Qual: Garage quality | Pool Area: Pool area in square feet |
|  | Central Air: Central air conditioning | Garage Cond: Garage condition | Misc Val: $Value of miscellaneous feature |
|  | Garage Type: Garage location | Paved Drive: Paved driveway | SalePrice: Sale price $$ |
|  | Misc Feature: Miscellaneous feature not covered in other categories | Pool QC: Pool quality |  |
|  | Sale Type: Type of sale | Fence: Fence quality |  |
|  | Sale Condition: Condition of sale |  |  |