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.

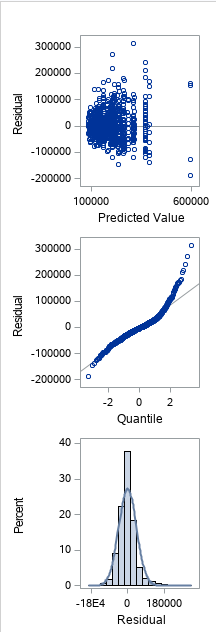
Two-way ANOVA analysis requires observations within different groups and across different groups be independent of one another. But house sale prices within same neighborhood are rarely independent of one another and suffer from spatial correlation. For this analysis, we will assume that house sale prices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

THE 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.***