Hoboken Housing Price

Group Member: Abhay Sharma, Raphael Presberg, Shikuan Gao

# Background and Motivation

Hoboken Area is a hot real estate market. There are several reasons foe that. First, in the Hoboken Area there are a lot of international professor and students. After they graduate, students who still want to stay in America need a place to settle down. They will focus their housing location in Hoboken. Second, the office workers in New York City can not afford the high price in Manhattan or Queens, in order to stay in the place near their working area, they will also focus their housing location in Hoboken. Third, according to the statistic, Hoboken market trends indicate an increase of $133,000 (20%) in median home sales over the past year. The average price per square foot for this same period rose to $724, up from $676. From the figure below, we can see that the graph shown exactly we mentioned above.

A close up of text on a white background

Description generated with high confidence

**Figure. Median Sales Price**

Moreover, in our project, we believe that buying a home place is an important life choice. Not only need to meet all the housing condition one’s desire, but also the price should be under one’s financial ability.

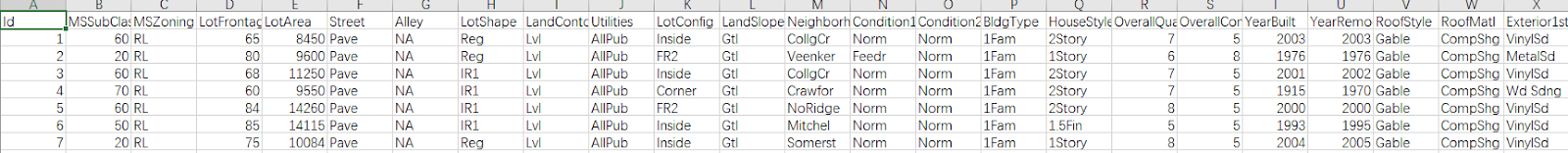
# Objective

As the rising of housing market in Hoboken, in this project report, our goal is to conduct a full factorial design model to analysis the housing price of Hoboken. We try to build a design and to unravel a relation between certain factor of a house and search most related feature to identify the price of a home place. As we mentioned above, it is important to choose a home place, our result can help those normal people in buying a home place, giving an advice for their decision making.

# Data

### Data Description

The raw data of our project is from Kaggle, it is a dataset for Hoboken real estate prices information. By checking the dataset in the below, we can see that there are over 1000 variables and 80 attributes we needed such as street, alley, lot shape and so on. Each attribute influences on the outcome of a single real estate price



**Figure. Dataset**

# Experiment versus Observational

In this section, we would introduce the difference between observational study and experiment design. A designed experiment applies a treatment to individuals and attempts to isolate the effects of the treatment on a response variable. On the order hand, an observational study measures the characteristics of a population by studying individuals in a sample but does not attempt to manipulate or influence the variables of interest. Base on what learned on the difference, we come out the conclusion that our project is more like an observational study than an experiment design. In our project, we measure our dataset as the characteristics of a population, also study each single housing information and looking for the same levels of the factors we design as a sample. During our project, we did not attempt to manipulate or influence the variables of interest.

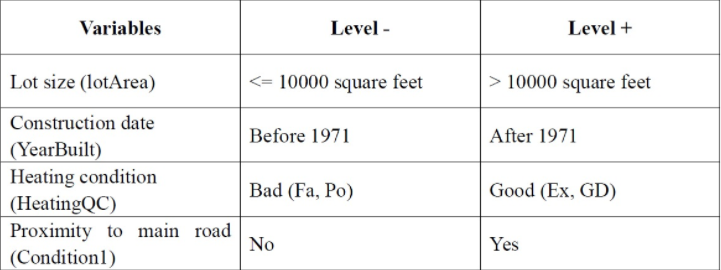
A picture containing object

Description generated with high confidence

**Figure. Experiment versus Observational**

# Factors and Levels

In this section, we would introduce the factors and levels used in our project. First, what we did is create 4 factors analysis creation shown as below table. The factors we choose are lot size, construction date, heating condition and proximity to main road. Each factor has two levels, for example for the lot size, negative level is equal or smaller than 10000 square feet, positive level is larger than 10000 square feet. For the construction date, negative level is construction date before 1971, positive level is construction date after 1971. Last, the response of our design is the house final sales price in dataset. After choosing all the factors, levels and response, we set up arandom factor design of 4 factors on 16 runs. To mention after 2 replications, the total runs will be 32 runs.

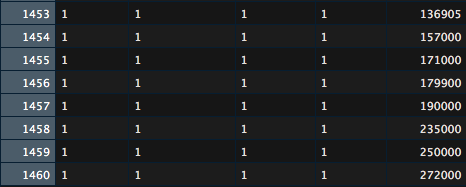


**Figure. Factors Table**

# Factorial Design

### Conducting the Experiment

First, we split to have 2 replications. In order to do that, we split the dataset half and half. Each experiment dataset is 50 percent of the initial dataset. And we use R to manage the factorial design, the negative level is shown as -1 and the positive level is shown as 1. After building the table, we use R code to find all the records answering the requirement. In the process, we met some issues that there are more than one record to answering each factor requirement situation. As the figure shown below, we can see that there are 8 records answering the requirement for each factor to be 1 (positive level).



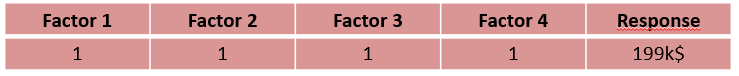
**Figure.**

The outcome has 8 different house price response. In order to solve the problem, we calculate the mean of these house prices. As the equation and table shown below, the mean price of these houses is 199k dollars.

A screenshot of a cell phone

Description generated with very high confidence

**Equation.**



**Table.**

### Response Results

After running 2 replications, as we can see from the figure shown below, for the same combinations, we have different response in the two experiments. The standard order is also changing due to the differences in the responses. For example, in the first replication, the response order is 11, 15, 7 and so on. On the other hand, in the second replication, the response replication is 12, 16, 14 and so on.

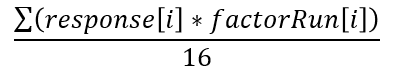
A screenshot of a cell phone

Description generated with high confidence

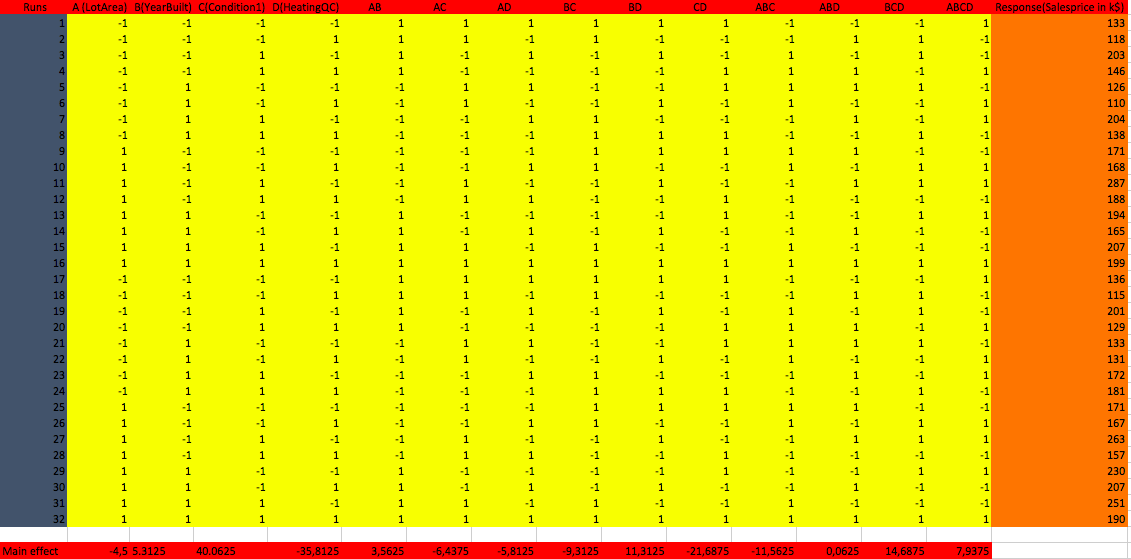
**Figure. Two replication Table.**

### Full Factorial Design

After completing 2 replications and put it all together, we have our full factorial design shown as the table below. This table allows out to study the effect of each factor on the [response variable](https://en.wikipedia.org/wiki/Response_variable), as well as the effects of [interactions](https://en.wikipedia.org/wiki/Interaction_(statistics)) between factors on the response variable. Other than that, we also calculate each factor main factor using the equation show as below:



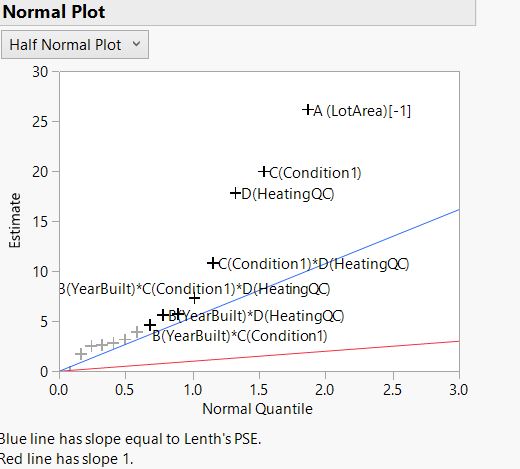
The scores show the importance of each factor will affect the price of a home place, for example, the highest score in the table is 40.0625 by factor heating condition.



### Table. Full Factorial Design

# Half Normal Plot

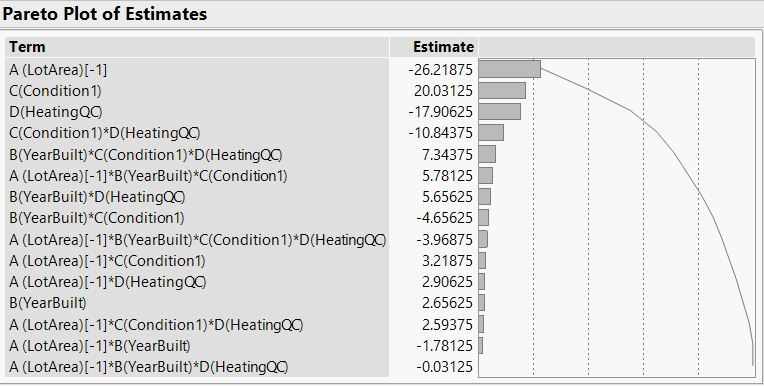
The half normal probability plot is a graphical tool that uses these ordered estimated effects to help assess which factors are important and which factors are not important.



The half normal plot was plotted and from that we see that the factors ‘Lot Area’, ‘House close to the main road’, ‘Heating Condition’ and interaction effect of ‘Lot Area’ and ‘Heating Condition’ are away from the blue line. Hence, we can conclude that they are not normally distributed and therefore significant. Hence, from Half normal plot the significant effects we obtained are: A, C, D, CD, BCD. Other effects like the interaction of ‘Year Built’ and ‘Heating condition’ are close to the blue line, hence we can conclude that they are normally distributed and therefore are not significant.

# Pareto Plot

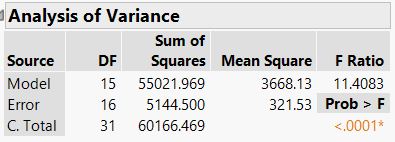
A Pareto chart is a bar graph. The lengths of the bars represent frequency or cost (time or money) & are arranged with longest bars on the left and the shortest to the right. In this way the chart visually depicts which situations are more significant.

****

The pareto plot shows us the frequency of each effect. The bars represent the individual values of the effects in descending order, the bars are segregated in descending order as we go downwards.

# Analysis of Variance (ANOVA)

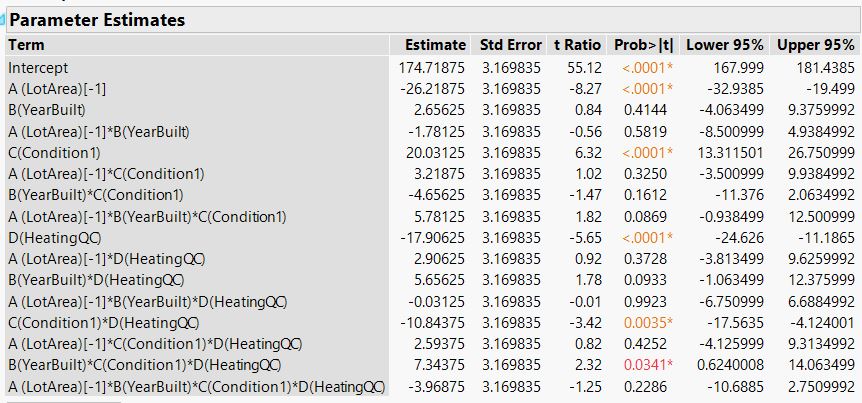
The F-test analysis is the basis for model evaluation of both single factor and multi-factor experiments. This analysis is commonly output as an ANOVA table by statistical analysis software, as illustrated by the table below.



The most important output of the table is the F-ratio (11.4083). The F-ratio is equivalent to the Mean Square (variation) between the groups of 3668.13 divided by the Mean Square error within the groups of 321.53.

The Model F-ratio of 11.4083 implies the model is significant. The p-value ('Probability of exceeding the observed F-ratio assuming no significant differences among the means') of less than 0.0001 indicates that there is only 1 % probability that a Model F-ratio this large could occur due to noise (random chance).

The following graph shows the ‘Simultaneous Pairwise Difference’ Confidence Intervals of all the effects.



# Regression Model

After the analysis of variance, we studied the Regression model. The regression equation was calculated by taking into account the significant effects that we obtained by Half Normal Plot.

### Final Equation

Y= 174.7- 26.2 (A)+ 20.03 (C)- 17.9 (D)- 10.8 (CD)+ 7.3 (BCD)

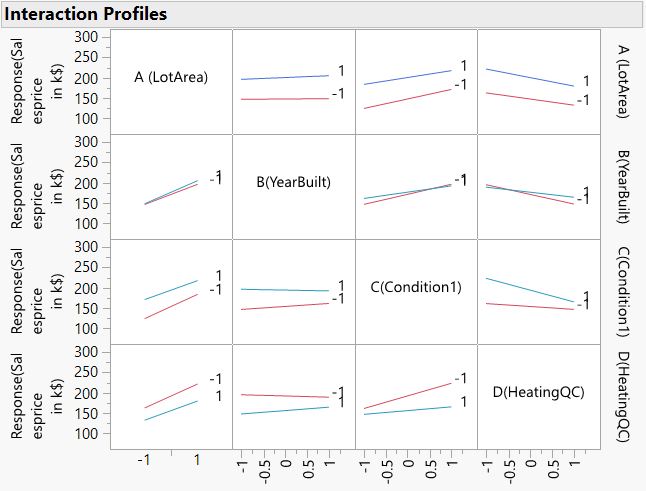
### Final Equation in terms of actual Factors:

Y= 174.7- 26.2 (Lot Area) + 20.03 (House close to main road)- 17.9 (Heating Condition)- 10.8 (House close to main road) \* (Heating Condition) + 7.3 (Year Built) \* (House close to main road) \* (Heating Condition)

From regression equation we can predict the price of the house that will be sold in the future according to the given effects.

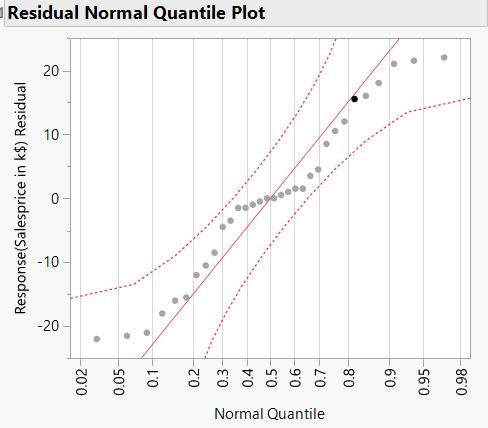
# Interaction Plots

An interaction plot displays the mean of the response variable for each level of a factor with the level of a second factor held constant. Interaction plots are useful for judging the presence of interactions – when the response at one factor level depends upon the level(s) of other factors. Parallel lines in an interaction plot indicate no interaction; non-parallel lines indicate an interaction between the factors. The greater the departure of the lines from the parallel state, the higher the degree of interaction.



# Analysis of Residuals

The Normal Plot of the Residuals was plotted, and the output showed that almost all residuals would pass the normality test, which gives us enough evidence that the residuals were normally distributed.

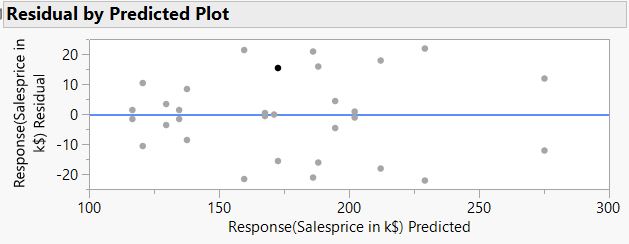


The Residuals vs Predicted values graph was also plotted.

Residual= Observed Value- Predicted Value.

In the plot below, where the prediction made by the model is on the x-axis, and the accuracy of the prediction is on the y-axis. The distance from the line at 0 is how bad the prediction was for that value.

According to our plot, we wouldn’t say that the graph is ideal but still it indicates that the residuals and the fitted values are uncorrelated, as they should be in a homoscedastic linear model with normally distributed errors.



**Conclusion**

From the Analysis of Variance table, we found that the value of the F- Statistic to be 11.4083 and its P-Value to be 0.0001. The large model F-Value of 11.4083 and its small P-Value of 0.0001 implies that the model is significant. There is only a 0.01 % chance that the model F-Value so large could have occurred due to noise.

From the analysis of the ‘Half Normal Plot’ we can say that we have five significant effects which are A (Lot Area), C (House close to main road), D (Heating Condition), CD (House close to main road) \* (Heating Condition) and BCD (Year Built) \* (House close to main road) \* (Heating Condition).

The top 3 significant affects are the size of the house, convenience to the main road and the heating condition of the house. Surprisingly in Hoboken, looking for a satisfied lot size house cost a less than other housing conditions. If one have limited financial ability and willing to buy a house in Hoboken. He can still be looking for a brand-new building but lies on somewhere that far away from city center.