Ames Housing Prices

Statistical Inference Project

Javier Fong - 100437994

Introduction

As for the project for the statistical inference course, I decided to analyze the Ames Housing data set. With this project I expect to get information regarding the housing market in Ames, Iowa, utilizing solely the information at hand and inferring techniques explored in class. I would expect to find differences in house prices based in groups of specific characteristics, and some relationship between characteristics that may not be so obvious.

The data set consist of 1460 observation of house sales in Ames, Iowa. It has 79 variables, but in this project I'll only work with the following:

Categorical variables:

- Street: material of the street. (2 levels)
- HeatingQC: Quality of the heating system. (5 levels)
- KitchenQC: Quality of the kitchen. (5 levels)

Discrete variables:

- YearBuilt: year on which the house was built.
- YearOfSale: year on which the house was sold.
- FullBath: number of full bathrooms.
- HalfBath: number of half bathrooms.

Continuous variables:

- LotArea: Lot area of the house in square feet.
- GarageArea: Area for the garage in square feet.
- SalePrice: Last selling prince is USD.

The data set can be found in this link. (Link to Ames data set)

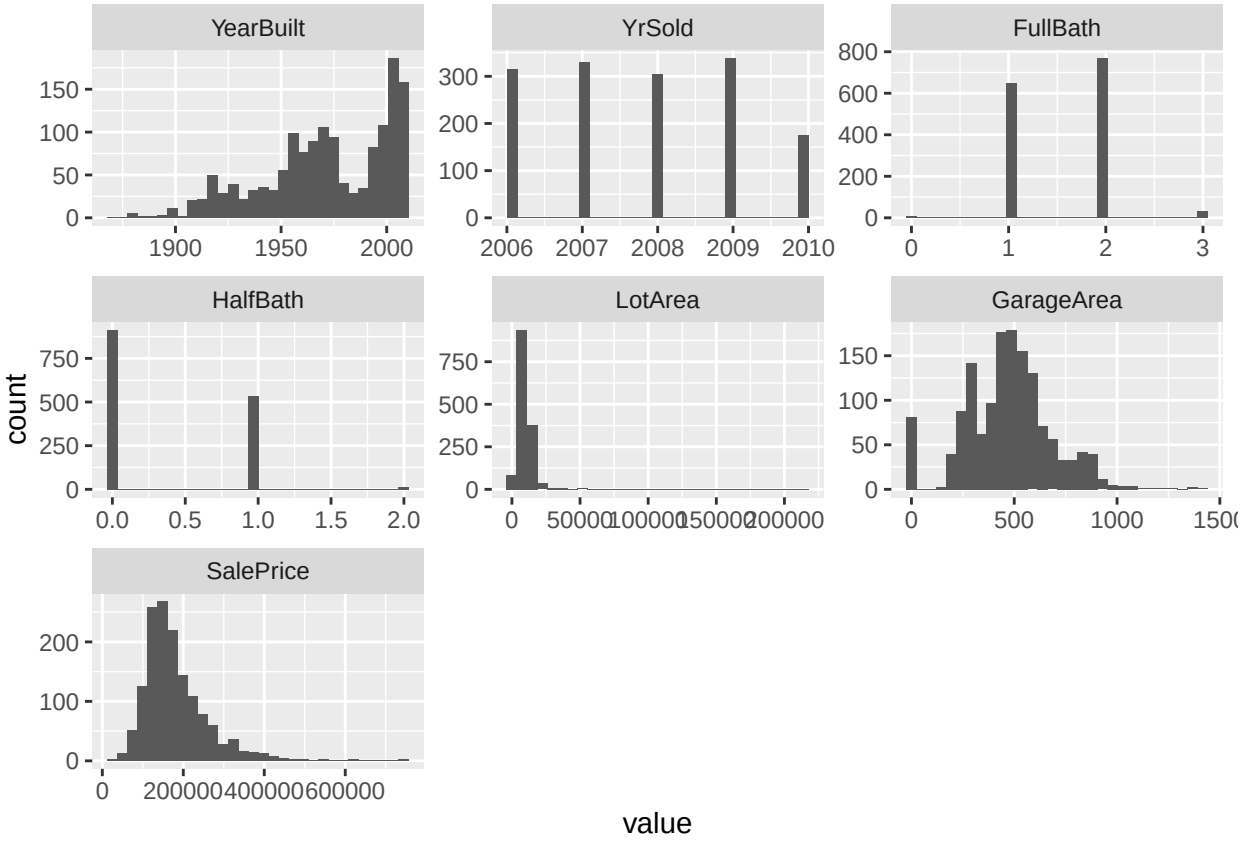
Data Exploration

Quick Summary

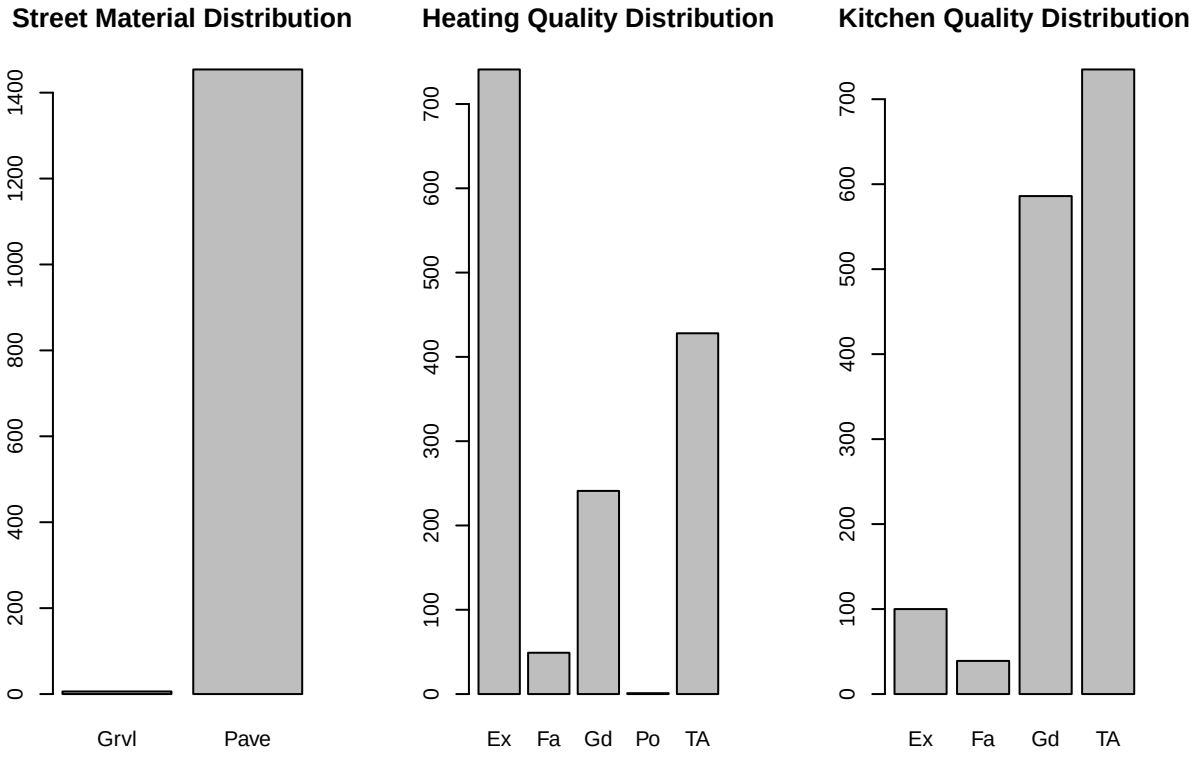
```
##    YearBuilt        YrSold        FullBath        HalfBath
##  Min.   :1872   Min.   :2006   Min.   :0.000   Min.   :0.0000
##  1st Qu.:1954   1st Qu.:2007   1st Qu.:1.000   1st Qu.:0.0000
##  Median :1973   Median :2008   Median :2.000   Median :0.0000
##  Mean   :1971   Mean   :2008   Mean   :1.565   Mean   :0.3829
##  3rd Qu.:2000   3rd Qu.:2009   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :2010   Max.   :2010   Max.   :3.000   Max.   :2.0000
##     LotArea         GarageArea       SalePrice       Street     HeatingQC
##  Min.   :  1300   Min.   :   0.0   Min.   : 34900   Grvl:   6   Ex:741
##  1st Qu.:  7554   1st Qu.: 334.5   1st Qu.:129975   Pave:1454   Fa: 49
```

```
##  Median :  9478   Median : 480.0   Median :163000              Gd:241
##  Mean   : 10517   Mean   : 473.0   Mean   :180921              Po:  1
##  3rd Qu.: 11602   3rd Qu.: 576.0   3rd Qu.:214000              TA:428
##  Max.   :215245   Max.   :1418.0   Max.   :755000
##  KitchenQual
##  Ex:100
##  Fa: 39
##  Gd:586
##  TA:735
##
##
```

Numeric variables



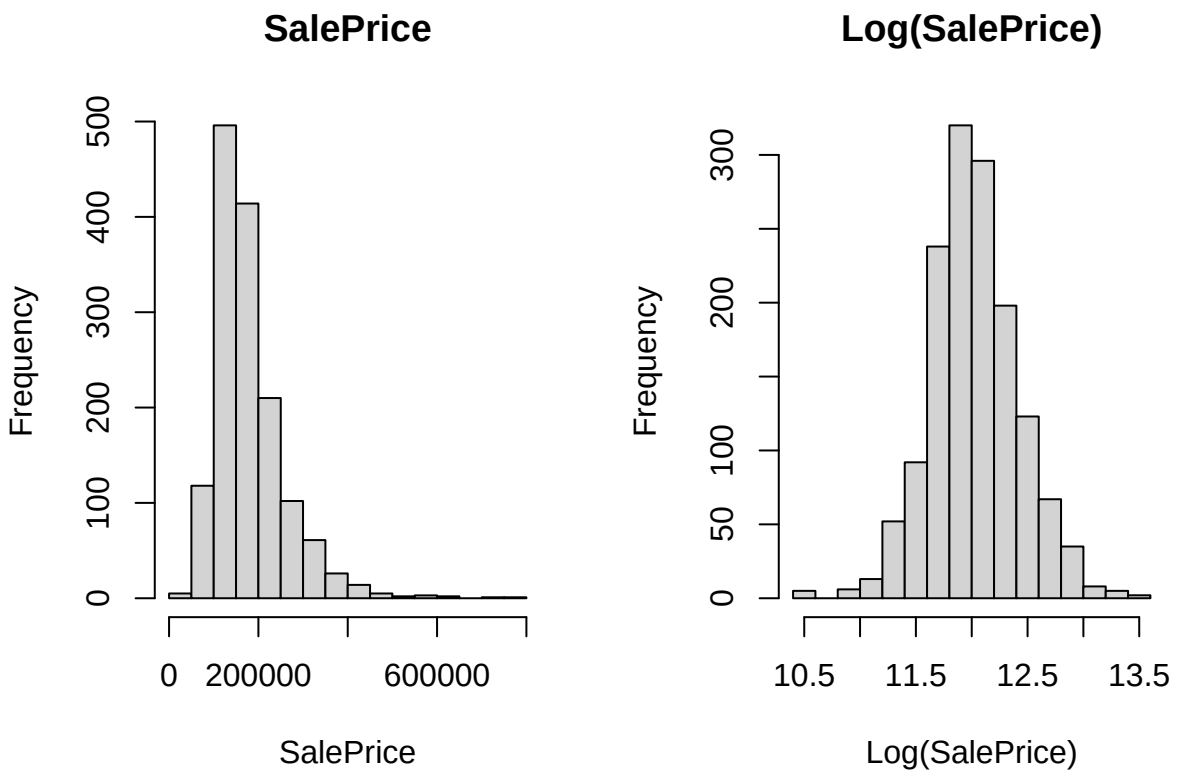
Categorical Variables



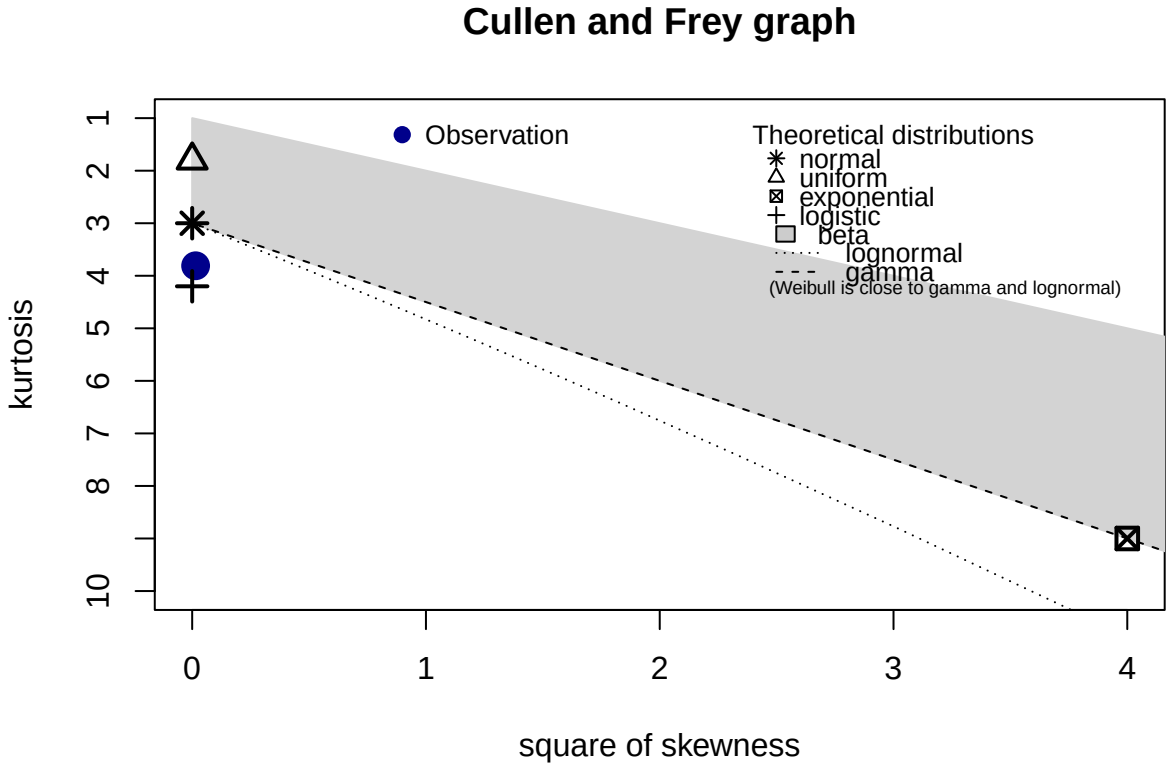
Distribution selection

The continuous variable I will analyze in this project will be **SalesPrice**. This variable describes the selling price of the property.

From the plot below we notice the variable is skew to the left, so I applied a *log*() transformation to improve its symmetry.



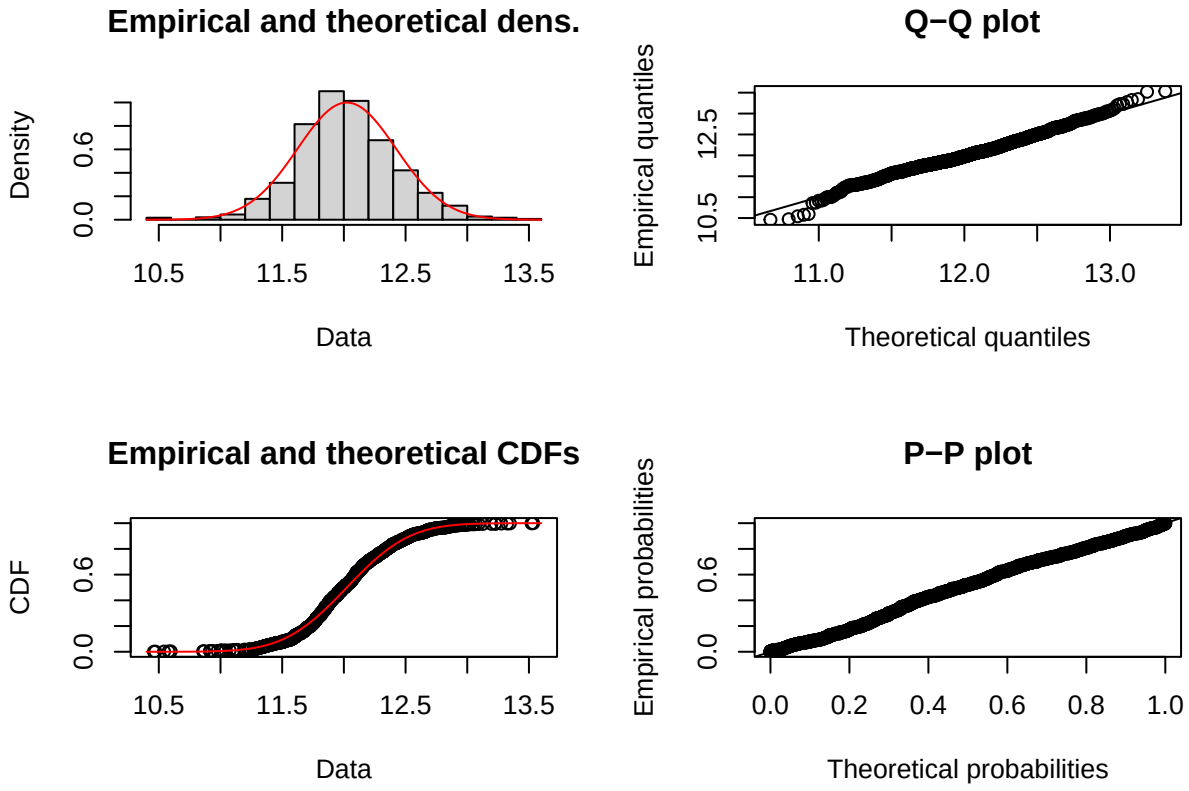
Now, using the function *descdist()* we find the distribution closest related to our sample on the *Cullen and Frey* graph. From this graph we gather that a **normal** distribution seems to closely describe our sample.



Normal Distribution

Using the function *fitdist* we fitted a normal distribution to our sample (transformed with a log). Using the charts below, we can see that the distribution is a pretty good fit to our sample. It only seems to have some issues at the extremes.

```
fit.normal = fitdist(x, "norm", method = "mle")
plot(fit.normal)
```



Maximum Likelihood

Using the MLE method (as a parameter in the fitdist function) we obtain the following parameter estimation for the mean and the standard deviation:

Table 1: Parameter Estimation with Maximum Likelihood

	Estimate	Standard Dev.
mean	12.024051	0.0104506
sd	0.399315	0.0073894

With this values we know that our fitted distribution looks like:

$$X \sim N(\mu = 12.0241, \sigma^2 = 0.1595)$$

One-sample Inference

Estimators for the mean

Estimator $\hat{\mu}_1 = \overline{x}$ (Geometric Sample Mean)

$$\hat{\mu}_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} xi = 12.0240509$$

Given that,

$$\hat{\mu}_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} xi = \frac{1}{n} n\mu = \mu$$

we can say that $\hat{\mu}_1$ is an unbiased estimator of the population mean (μ).

Estimator $\hat{\mu}_2 = Me(x)$ (Sample Median)

$$\hat{\mu}_2 = Me(X) = 12.0015055$$

By definition, we know that

$$F(\hat{\mu}_2) = \frac{1}{2} = \int_{-\infty}^{\hat{\mu}_2} f(x)dx$$

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\hat{\mu}_2} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{2}$$

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx + \frac{1}{\sigma\sqrt{2\pi}}$$

$$\int_{\mu}^{\hat{\mu}_2} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{2}$$

But,

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{0} e^{\frac{-Z^2}{2}} dz = \frac{1}{2}$$

Which means,

$$\frac{1}{2} + \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{\hat{\mu}_2} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{2}$$

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{\hat{\mu}_2} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = 0 \implies \mu = \hat{\mu}_2$$

With this we conclude that $\hat{\mu}_2$ is also an unbiased estimator of the mean.

Estimator Comparison

If we examine the variance of both estimators we get that:

$$Var(\hat{\mu}_1) = \frac{\sigma^2}{n}$$

$$Var(\hat{\mu}_2) = \frac{\pi}{2}\frac{\sigma^2}{n}$$

Which means,

$$Var(\hat{\mu}_1) \leq Var(\hat{\mu}_2)$$

So, even though both estimators are unbiased, the geometric mean is a more precise estimator of the population mean.

We can see this in our sample because, as we estimate the variance as well:

$$\hat{Var}(\hat{\mu}_1) = \frac{S'^2}{n} = 0.0001093 \leq 0.0001717 = \frac{\pi}{2}\frac{S'^2}{n} = \hat{Var}(\hat{\mu}_2)$$

Estimators Error

Given that both estimators are unbiased, we used CV to calculate the error of the estimators. With the following results.

$$CV(\hat{\mu}_1) = \frac{Var(\hat{\mu}_1)}{\hat{\mu}_1} = 0.0008694$$

$$CV(\hat{\mu}_2) = \frac{Var(\hat{\mu}_2)}{\hat{\mu}_2} = 0.0010917$$

95% Confidence Interval

Using the T-statistic, we calculated the confidence interval for both estimator as:

The error of $\hat{\mu}_1$ is less that $\hat{\mu}_2$

$$P(12.0035442 \leq \hat{\mu}_1 \leq 12.0445576) = 0.95$$

$$P(12.0008328 \leq \hat{\mu}_2 \leq 12.0021781) = 0.95$$

Proportion in Population

Now we'll examine the variable *HalfBath*, that describe the amount half bathrooms in the property (bathrooms without shower). But we'll reduce the levels to just 2:

$$hasHalfBathroom = \begin{cases} 0, HalfBathroom = 0 \\ 1, HalfBathrom > 0 \end{cases}$$

This is pretty close to the real variable given that only 12 observations have more than 1 & half bathrooms. For this analysis we assume this variable can be regarded as random.

The proportion of observation that belong to this group is the mean value of the observations, given that we estimate that each observation behaves as a *Bernulli* r.v.

$$\hat{p} = \overline{X} = 0.3746575$$

$$Var(\overline{X}) = \frac{\hat{p}(1-\hat{p})}{n} = 0.0001605$$

95% Confidence Interval

Using the T-statistic, giving that we do not know the real variance of the population, we get the following confidence intervals at 95%:

$$CI_{0.95}(p) = [\hat{p} - t_{1459:0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + t_{1459:0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}] = [0.3498086, 0.3995065]$$

Inference with more than one sample

In this section we'll create groups using the variable *KitchenQual* that describes the current quality of the kitchen at the property. This variable can take 4 different values, Ex = Excellent, Fa = Fair, Gd = Good and TA = Typical/Average. Now we calculate the mean *SalePrice* for each of this groups and the cv for this estimators.

Table 2: Estimate mean Sale Price value by Kitchen Quality

Kitchen Quality	mean	cv
Ex	12.63361	0.3053004
Fa	11.50458	0.5159091
Gd	12.22234	0.0984061
TA	11.81059	0.0885820

There seems to be a relationship between the quality of the kitchen and the sale price of the house, because the order from highest to lowest mean Sale Price is Excellent, Good, Typical and Fair.

Now, using the *hasHalfBath* variable, we estimate the proportion by *KitchenQuality*

Table 3: Proportion of hasHalfBath by KitchenQuality

Kitchen Quality	Est. Proportion	MSE
Ex	0.5500000	0.0045000
Fa	0.1794872	0.0210388
Gd	0.4607509	0.0009202
TA	0.2925170	0.0009626

It is interesting to notice that the group with the highest MSE is the one with the smaller sample size, FA with only 35 observations. In contrast with the *Gd* group, which has a size of 471 observation and the smallest MSE.

Now, we'll compare the mean Sale Price of the two largest groups by Kitchen Quality, Typical against Good.

Table 4: Mean Sale Price by Kitchen Quality group

Kitchen Quality	n	Mean
Gd	586	12.22234
TA	735	11.81059

The pooled Variance of both groups, based in the sample variance if each one is:

$$S^2 = \frac{(n_1 - 1){S'}_1^2 + (n_2 - 1){S'}_2^2}{n_1 + n_2 - 2} = 0.0275554$$

Using this estimator, and the T-statistic, we can calculate a confidence interval of the difference between both means. Which look like this,

$$CI_{0.95}(\mu_1 - \mu_2) = \overline{X_1} - \overline{X_2} \pm t_{n_1+n_2-2;0.025} S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = [0.4087511, 0.4147386]$$

This tells us that the difference is between 0.4087511 and 0.4147386. Given that this range is positive, we can say that the mean sale price of properties with good kitchens is higher than the sale price of typical kitchens.

We can compare this two means by a hypothesis testing of equality of means. For that we calculate a test statistic that reads as

$$T = \frac{\overline{X_1} - \overline{X_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where we use the same S estimator as in the confidence interval.

The idea is to compare this test statistc against the T-statistic $t_{n_1+n_2-2;\alpha}$ and there we have 3 hypothesis:

a. $C_a = (T > t_{n_1+n_2-2;\alpha})$
b. $C_b = (T < -t_{n_1+n_2-2;\alpha})$
c. $C_a = (|T| > t_{n_1+n_2-2;\alpha/2})$

Based on this comparisons, we can know which group mean is larger. In our case the values are as follows:

$$T = 44.7884437$$
$$t_{n_1+n_2-2;\alpha} = 1.6460097$$
$$t_{n_1+n_2-2;\alpha/2} = 1.9617641$$

This comparison confirms our finding in the confidence intervals, that the mean sale price for good quality kitchens is higher than the mean sale price with average kitchens.

Now, we'll do a similar analysis, but instead of mean Sale Price by group, we'll analyze the proportion of half bathrooms based on the kitchen quality.

We use a similar procedure as the means difference of means, and got the following interval for the difference in proportions

$$CI_{0.95}(p_1 - p_2) = \hat{p_1} - \hat{p_2} \pm t_{n_1+n_2-2;0.025} S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = [0.1600449, 0.1764228]$$

This positive difference tells us that there is a higher probability for a house with good kitchen to have a half bathroom, than a house with an average kitchen to have a half bathroom.

We can test via equiality of proportions which $\hat{p}$ is higher with the following results,

$$T = 8.6195019$$
$$t_{n_1+n_2-2;\alpha} = 1.6480145$$
$$t_{n_1+n_2-2;\alpha/2} = 1.9648876$$

Here we confirm our finding that housed with good kitchens have a higher proportion of half bathroom occurrence than houses with average kitchens.

Conclusions

1. We can describe the sale price of houses at Ames, Iowa very accurately as normal distribution with parameters

$$log(X) \sim N(\mu = 12.0241, \sigma^2 = 0.1595)$$

2. Even though the geometric mean and the median are both unbiased predictors for the mean of a Normal distribution, the geometric mean is better given its smaller variance.
3. Groups with higher size have smaller variance.
4. There seems to be a relationship between KitchenQuality and SalePrice. Houses with better kitchens have a higher selling price.
5. Also houses with "good" kitchens tend to be more likely to have half bathrooms.