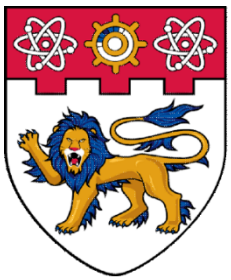


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

# MH3511 Data Analysis with Computer

## Project Work – Group 1

## House Price Prediction


| Name | Matriculation Number |
|---|---|
| CHEW JUN WEI | U2130331A |
| KEVIN YOK | U2130510H |
| LIV TAN KER JIN | U2130630A |
| LOW WAN TING NICOLE | U2131471D |
| NG SI EN | U2130243F |



**Abstract**: For most people, buying a house is deemed to be done at some point in life. It is easy to understand that the sale price is largely determined by basic information such as the number of bedrooms and toilets. However, the relationship between other equally important parameters, such as material of the house exterior, zone of the house, are less clear. Hence, we would like to examine these relationships, accounting for required assumptions, using appropriate tests and regression models. Ultimately, we would like to predict the house price given basic information about the house.

# Table of Contents

## 1. Introduction

In our project, a dataset containing the sale price of properties in Ames, Iowa, along with other characteristics of the properties such as lot area, overall quality, year built, fireplaces, garage area and fence is used. Based on this dataset, we seek to answer the following questions:

1. How are the variables distributed?
2. What justifies the inclusion of the variables in our models?
3. How does the sale price depend on
   - the number of bedrooms and/or toilets?
   - the various sizes of areas in a house – lot, garage?
   - the age of the house?
   - whether the house has been remodelled in the past five years prior to its sale?
   - the zoning, exterior material and overall quality of the house?
4. Is the sale price predictable?

This report delves into data descriptions, analysis, testing, and estimation utilizing R language. For each of our research objectives, we conducted statistical analyses and drew conclusions using the most suitable approach, supplemented with explanations and elaboration.

## 2. Data Description

The dataset, titled “House Price Prediction Cleaned Dataset”, is obtained from Kaggle by Chandramouli Naidu. The original Ames Housing dataset was compiled by Dean De Cock for use in data science education. It is cited as a good alternative for data scientists looking for a modernized and expanded version of the often-cited Boston Housing dataset.

We will work with the “cleaned train.csv” file, which contains 381 columns and 1458 rows. Before proceeding to data analysis, we first performed an intensive preliminary data cleaning which includes:

1. Removing dummy variables
2. Defining a new *Remodelled* dummy variable
   - *Remodelled* = 1 if the house is remodelled within 5 years prior to its sale
   - *Remodelled* = 0 otherwise
3. Defining a new *Toilets* variable which sums the number of toilets in each property using the formula below.
   - *Toilets* = *FullBath* + (*HalfBath*/2)
4. Picking 10 variables to represent all types of data, as shown follow:

| | | |
|---|---|---|
| Numerical | Continuous | 1. Sale price (our dependent variable)<br>2. Lot Area<br>3. Garage Area |
| | Discrete | 4. Number of bedrooms<br>5. Number of toilets<br>6. Age (with respect to year sold) |
| Categorical | Nominal | 7. Exterior material<br>8. Zoning |
| | Ordinal | 9. Overall Quality*<br>10. Remodelled |

*Overall quality is considered as a categorical variable due to the unequal intervals between ratings (i.e. the difference in quality between a house with ratings 1 and 2 respectively is likely smaller than that between 9 and 10. This is due to the more stringent standards at higher levels of quality). With regard to the linear regression model later on, the relationship between sale price and overall quality is unlikely to be linear either. Hence, it is treated as a categorical variable.

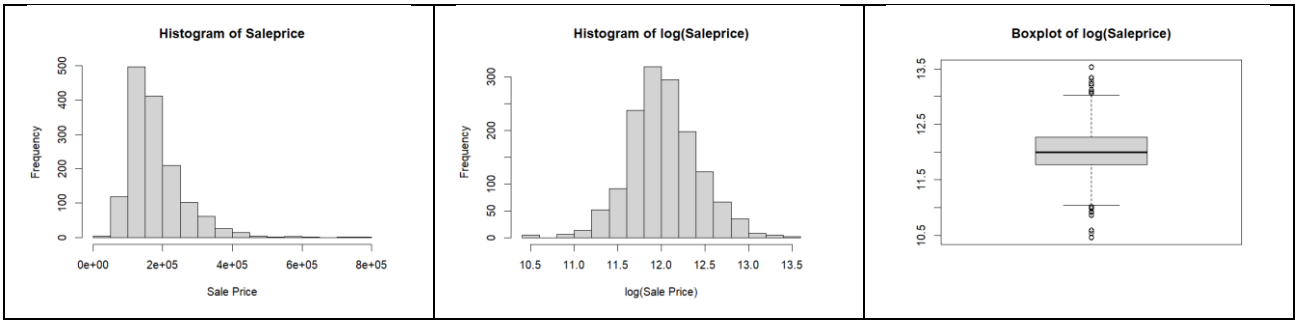
With that, we will be left with a subset containing 10 columns and 1458 rows.

## 3. Summary Statistics

In this section, we investigate the data in more detail. The dependent variable, Sale Price, and the categorical variables are investigated individually to see if they are normally distributed so that the appropriate tests can be applied to justify its inclusion in our regression model.

### 3.1. Summary statistics for Main Variable of Interest

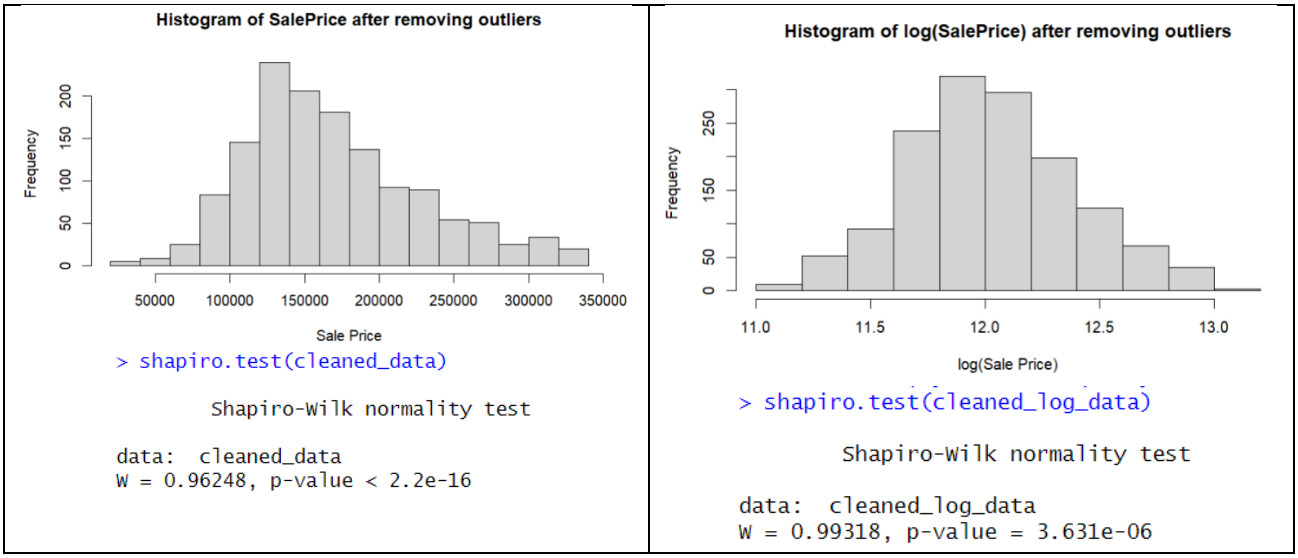
Our main variable of interest is the sale price of the property. The following plots show the overall distribution of the variable sale price.




*SalePrice* is the property's sale price in United States Dollars (USD). Our primary analysis reveals that the median sale price is $16,300, and the mean sale price is $180,932.90. Notably, the median falls below the mean, indicating a right-skewed distribution. This observation is further supported by the left histogram depicted above. We further conducted the Shapiro-Wilk normality test, confirming the non-normal distribution of our data with a p-value of 2.2 * 10^-6.

Therefore, we implemented a natural logarithm transformation (base $e$) on the variable, which visually exhibited a closer approximation to normal distribution. None of the values are dropped as there is no property having a sale price of $0. The histogram and boxplot of the log-transformed variable are shown above. However, the Shapiro-Wilk normality test shows a p-value of 1.149 * 10^-7 indicating that the data is non-normal.

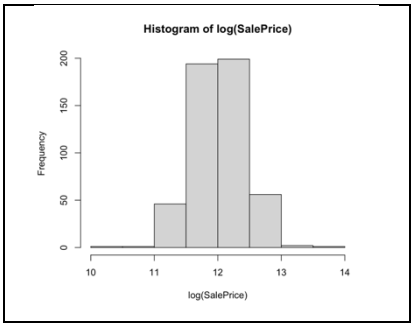
Moving on, we tried removing outliers of both *SalePrice* and log(*SalePrice*) using Interquartile Range (IQR) Method as an attempt to achieve normal distribution. However, we found that after removing outliers, the distribution still doesn't achieve normality, as shown by:

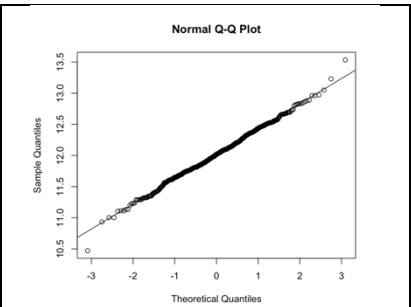



Therefore, we decided to use the original data before removing outliers, with another approach of analysis, i.e. data sampling.

### 3.1.1. Sampling Our Data

We realized that the Shapiro-Wilk normality test has a limitation with sample size, i.e. as the sample size increases, the more likely that the test result is statistically significant. Due to this limitation, we have decided to use a sub-sample with size of 500 to reduce our sample size for testing. The sub-sample is generated randomly. Below is the histogram for the log(*SalePrice*) as well as the qqplot for the sampled data's log(*SalePrice*).




```
> shapiro.test(log(SalePrice))

        Shapiro-Wilk normality test

data:  log(SalePrice)
W = 0.99698, p-value = 0.4837
```

We then used Shapiro-Wilk normality test to test for its normality of this sampled data. The test yielded a p-value of 0.4837, which shows that there is not enough statistical evidence to reject that the data is normal. Since the subsample is generated randomly, we can justify that the overall data is approximately normal as well. We will be using the same method for the subsequent analysis for our data.

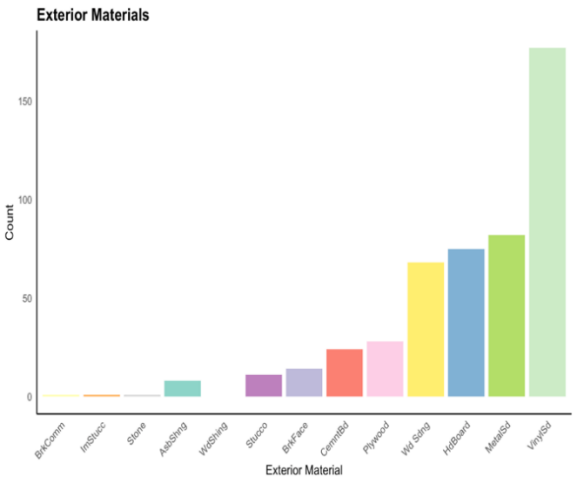
## 3.2. Summary statistics for Categorical Variables

The p-value for the Shapiro Wilk Tests for the different categories with respect to log(*SalePrice*) are tabulated in the following sub-sections for categorical variables. We will use the level of significance of 0.01 for our tests.

### 3.2.1. Exterior Material, *Exterior1st*

| Exterior Material Type | P-value for Shapiro Wilk Test | Normal |
|---|---|---|
| Wood Siding, *Wd Sdng* | 0.6322 | Yes |
| Vinyl Siding, *VinylSd* | 0.7604 | Yes |
| Hard Board, *HdBoard* | 9.97 x 10 -5 | No |
| Metal Siding, *MetalSd* | 0.1832 | Yes |
| Cement Board, *CemntBd* | 0.0666 | Yes |
| Brick Face, *BrkFace* | 0.875 | Yes |
| Stucco, *Stucco* | 0.646 | Yes |
| Plywood, *Plywood* | 0.5007 | Yes |
| Asbestos Shingles, *AsbShng* | 0.3704 | Yes |
| Wood Shingles, *WdShing* | 0.4306 | Yes |

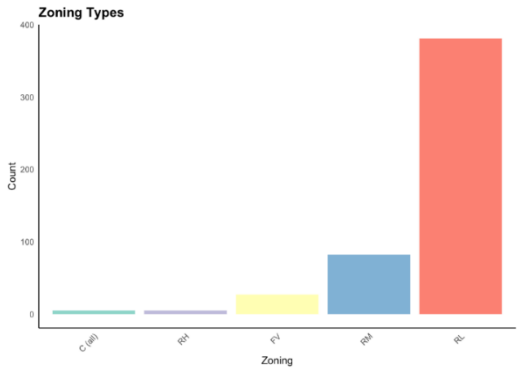
Bar plot Frequency:



### 3.2.2. Zoning, *MSZoning*

| Zone Type | P-value for Shapiro Wilk Test | Normal |
|---|---|---|
| Residential Low Density, *RL* | 0.3604 | Yes |
| Residential Medium Density, *RM* | 0.204 | Yes |
| Residential High Density, *RH* | 0.8543 | Yes |
| Floating Village Residential, *FV* | 0.4452 | Yes |
| Commercial, *C* | 0.9902 | Yes |

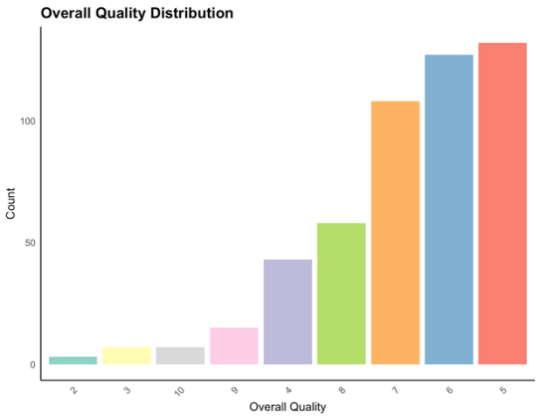
Bar plot Frequency:




### 3.2.3. Overall Quality, *OverallQual*

| Overall Quality | P-Value for Shapiro-Wilk Test | Normal |
|---|---|---|
| 1 | Too little data points | - |
| 2 | 2.2 x 10 -16 | No |
| 3 | 0.7967 | Yes |
| 4 | 0.04865 | Yes |
| 5 | 3.107 x 10 -5 | No |
| 6 | 0.01469 | Yes |
| 7 | 0.001485 | No |
| 8 | 0.3407 | Yes |
| 9 | 0.04178 | Yes |
| 10 | 0.6415 | Yes |

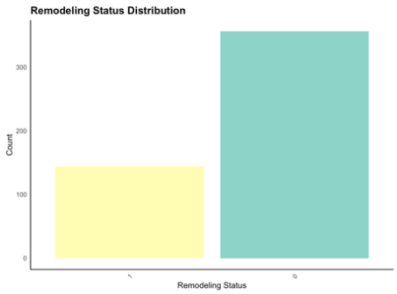
Bar plot Frequency:




### 3.2.4. Remodelling, *Remodelled*

| Remodeled | P-Value for Shapiro-Wilk Test | Normal |
|---|---|---|
| Yes, *1* | 0.974 | Yes |
| No, *0* | 0.03641 | Yes |

Bar plot Frequency:

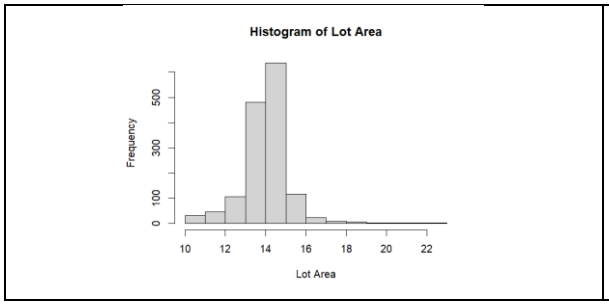


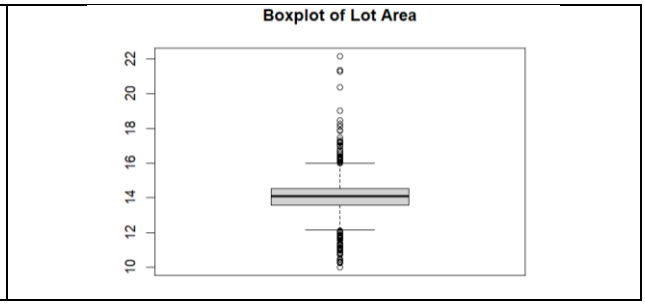

With this, we conclude that *Exterior1st* and *OverallQual* are non-normal variables and we will use non-parametric tests to justify the inclusion of these variables in our model. For *MSZoning* as well as *Remodelled*, we conclude that they are normally distributed, thus we will use t-test as well as ANOVA test in our analysis to justify their inclusions.

## 3.3. Summary statistics for Numerical Variables

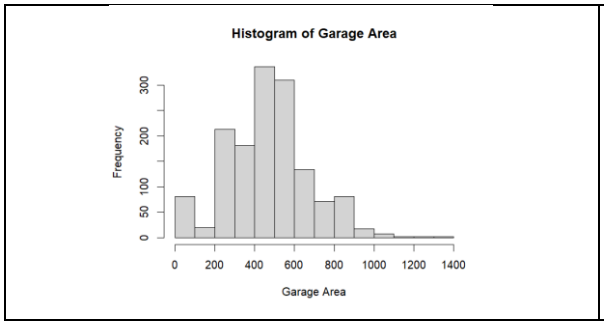
The histogram and boxplot from the variables are tabulated in the following sub-sections.

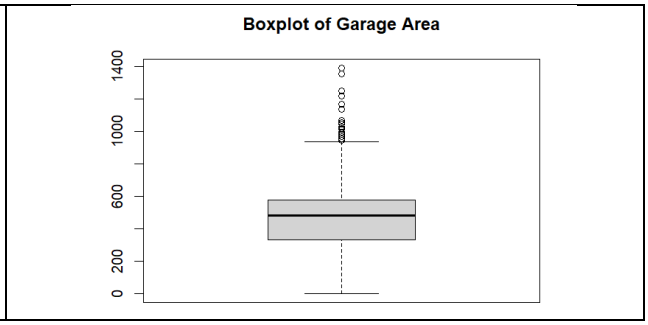
### 3.3.1. Lot Area, *LotArea*



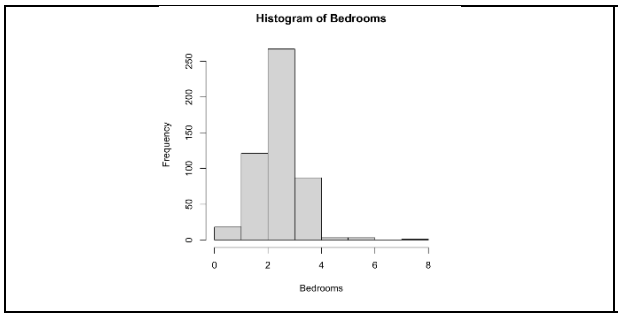




### 3.3.2. Garage Area, *GarageArea*



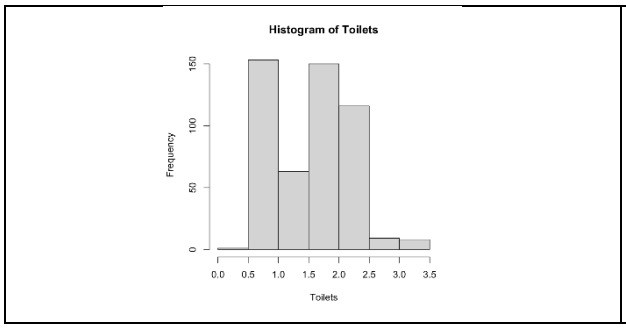



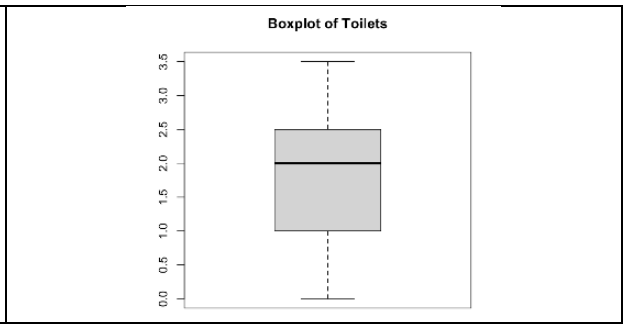

### 3.3.3. Number of bedrooms, *Bedroom*



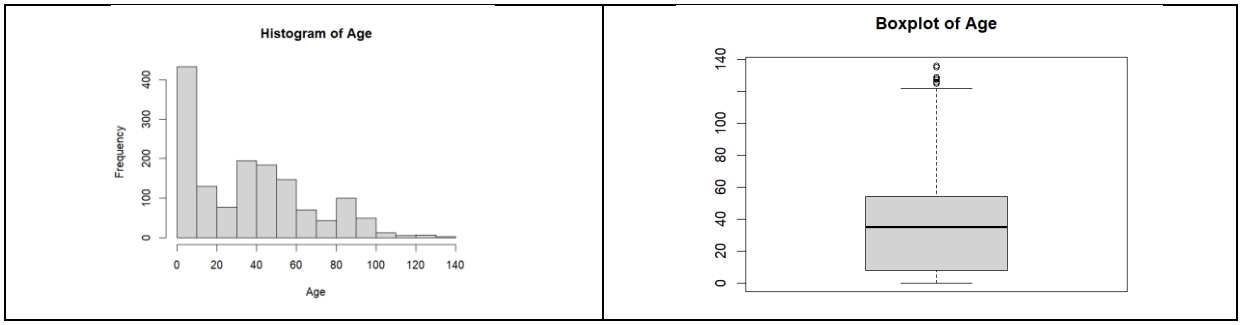


### 3.3.4. Number of toilets, *Toilets*



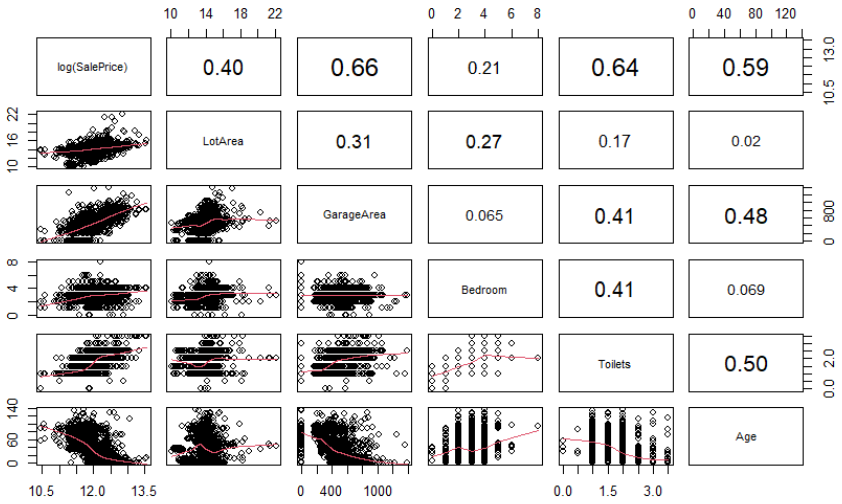
### 3.3.5. Age, *Age*




## 4. Statistical Analysis

Now, we will conduct statistical analysis for our numerical values to see if they are justified to be included in our regression models.

### 4.1. Correlation between log(*SalePrice*) and other variables




Scatter plots and correlation coefficients are valuable tools for examining potential linear relationships between a property's sale price and its characteristics. Characteristics with a correlation score exceeding 0.8 are deemed strong, while those falling between 0.4 and 0.8 are considered moderate, and those below 0.4 are regarded as weak. When we focus on the first row, in general the correlation for all our variables with respect to log(*SalePrice*) are considered not weak and hence we decide to include it into our regression model. We have decided to use bedrooms also even though it is weak, but it is not negligible.

Among the characteristics of properties, there are a few interesting observations from this tabulation:

- *Age* and *Toilets* are positively correlated (ρ = 0.50)
- *GarageArea* and *Age* are positively correlated (ρ = 0.48)
- *GarageArea* and *Toilets* are positively correlated (ρ = 0.41)
- *Bedroom* and *Toilets* are positively correlated (ρ = 0.41)

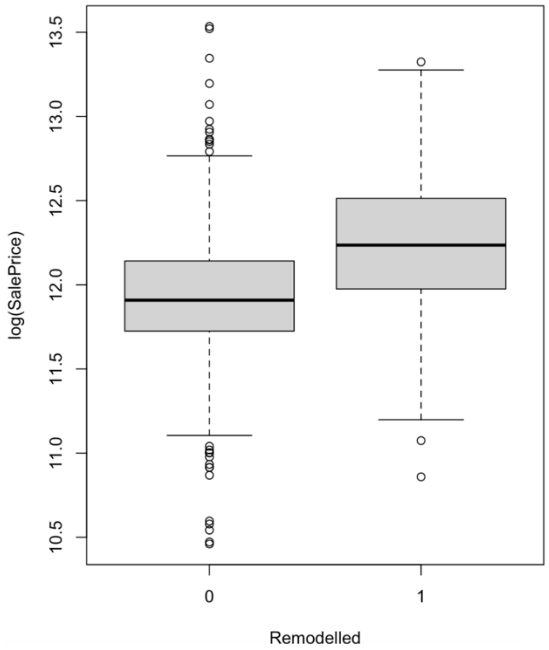
We shall perform some statistical tests to confirm some of our observations in the next section.

## 4.2. Statistical Tests

### 4.2.1. Relationship between log(*SalePrice*) and *Remodelled*

The dataset contains information of the year a house is remodelled. In order to investigate the effect remodelling has on sale price, we divided the dataset into two samples, based on whether the house has been remodelled within 5 years before its sale and whether it was not.

By visual inspection, the boxplot shows a difference in distributions for the two groups.




To determine whether there is a difference in the sale price between houses that were remodelled and houses that were not, we carried out a two-sample t-test approach. As mentioned above, the use of t-test as the categorical data for remodelled is justified by the normal distribution regularity condition being satisfied. Before carrying out the test, we will first determine if both samples have an equal variance.

```
> var.test(housing_data$logprice ~ housing_data$Remodelled)

        F test to compare two variances

data:  housing_data$logprice by housing_data$Remodelled
F = 0.90368, num df = 1033, denom df = 423, p-value = 0.2076
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7678403 1.0577794
sample estimates:
ratio of variances
         0.9036816
```

Based on the variance test, the F-statistic is 0.904 and the p-value is 0.2076 > 0.05 with a confidence interval of [0.77, 1.06]. At the 95% significance level, we do not reject the null hypothesis in favour of the alternative hypothesis. There is sufficient evidence to conclude that both samples have an equal variance.

Next, we conduct the two-sample t-test with the following hypothesis:

$$H_0 : \mu_{remodelled} = \mu_{not\ remodelled}$$
$$H_1 : \mu_{remodelled} \neq \mu_{not\ remodelled}$$

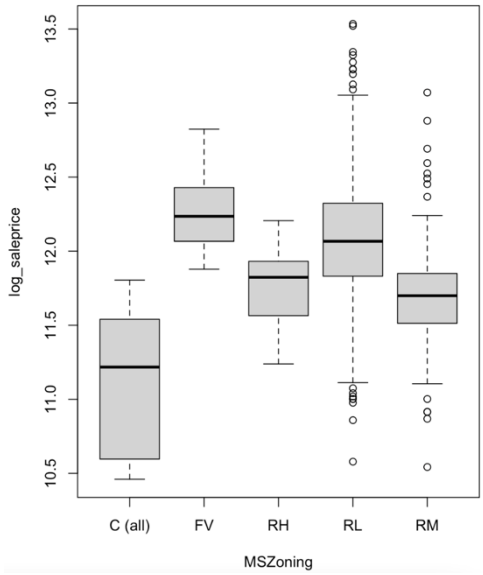
Where μ is the mean of log(*SalePrice*)

```
> t.test(remodelled_sp, remodelled_no_sp, var.equal = T)

        Two Sample t-test

data:  remodelled_sp and remodelled_no_sp
t = 14.897, df = 1456, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2778462 0.3621157
sample estimates:
mean of x mean of y
 12.25094  11.93096
```

From the summary, the t-statistic is 14.897 and the p-value is 2.2e-16 < 0.05 with a confidence interval of [0.278, 0.362]. At the 95% significance level, we reject the null hypothesis in favour of the alternative hypothesis. There is sufficient evidence to conclude that whether a house has been remodelled within 5 years of its sale has an effect on its sale price.

### 4.2.2. Relationship between log(*SalePrice*) and *MSZoning*

*MSZoning* is a categorical variable and each house is categorised into one of the five labels: C (all), FV, RH, RL, RM. The following plot illustrates the distributions of log(*SalePrice*) in the different zones.




By visual inspection, it appears that the different zones have different distributions. To determine if a particular zone influences log(*SalePrice*), an analysis of variance (ANOVA) test will be conducted with the following hypothesis.

$$H_0: \mu_C = \mu_{FV} = \mu_{RH} = \mu_{RL} = \mu_{RM}$$

$$H_1: At\ least\ one\ of\ \mu_i\ is\ not\ equal$$

```
> summary(aov(housing_data$logprice ~ factor(housing_data$MSZoning)))
                               Df Sum Sq Mean Sq F value Pr(>F)
factor(housing_data$MSZoning)    4  40.94  10.234   77.51 <2e-16 ***
Residuals                     1453 191.85   0.132
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the statistical tests carried out in 3.2.2, *MSZoning* follows a normal distribution under 1% level of significance. Therefore, the usage of the ANOVA test is appropriate here. The test resulted in a p-value of 2e-16 < 0.05. At the 95% significance level, we reject the null hypothesis in favour of the alternative hypothesis. There is sufficient evidence to conclude that log(*SalePrice*) differs with respect to *MSZoning*.

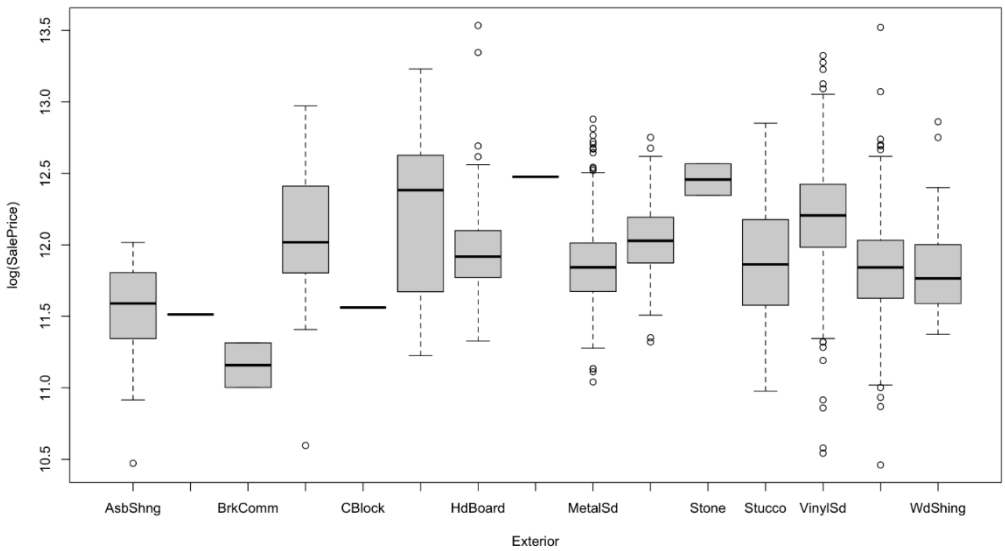
As the ANOVA test only tells us that at least one of the zones has a different average log(*SalePrice*), we can use a pairwise t-test to check which zones are different from one another.

```
> pairwise.t.test(housing_data$logprice, housing_data$MSZoning, p.adjust.method = 'none')

        Pairwise comparisons using t tests with pooled SD

data:  housing_data$logprice and housing_data$MSZoning

   C (all) FV      RH      RL
FV < 2e-16 -       -       -
RH 1.7e-05 1.1e-06 -       -
RL < 2e-16 0.00054 0.00025 -
RM 1.1e-06 < 2e-16 0.54523 < 2e-16

P value adjustment method: none
```

The pairwise t-test shows that the mean sale price of all zones are different (all p-values are < 0.05) except for RH and RM, with a p-value of 0.545.

### 4.2.3. Relationship between log(*SalePrice*) and *Exterior*

Exterior is a categorical variable and each house is categorised according to its exterior material: VinylSd, MetalSd, Wd Sdng, HdBoard, BrkFace, WdShing, CemntBd, Plywood, AsbShng, Stucco, BrkComm,AsphShn, Stone, ImStucc, CBlock. The following plot illustrates the distributions of log(*SalePrice*) for the different exterior house materials.




Since the results in 3.2.1 shows that Exterior follows a non-normal distribution, we have to employ a non-parametric test. To determine if exterior material affects log(*SalePrice*), we will test the null hypothesis that the distributions of log(*SalePrice*) across all materials are the same. We will do so by using the Kruskal-Wallies test, a non-parametric method for analysing data given that the distributions are not normal (as shown in 3.2.1).

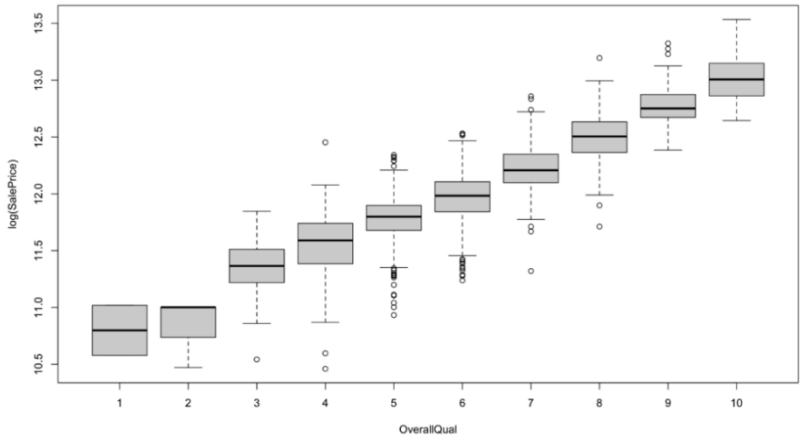
$H_0$: All distributions of log(*SalePrice*) across all exterior materials are identical;
$H_1$: The distributions of log(*SalePrice*) across all exterior materials are not all identical (i.e. at least one of the populations tends to yield larger observations than one of the other populations)

Both the manual and direct method for testing yield the same result, where the p-value < 0.01 and the null hypothesis is rejected with 99% confidence.

```
> s_sq;T_stat;pvalue
[1] 177265.4
[1] 303.4604
[1] 0
> kruskal.test(comp$price,comp$exterior) #rejected, not the same

        Kruskal-Wallis rank sum test

data:  comp$price and comp$exterior
Kruskal-Wallis chi-squared = 303.46, df = 14, p-value < 2.2e-16
```

### 4.2.4. Relationship between log(*SalePrice*) and *OverallQual*

*OverallQual* is a categorical variable and is also proven to be non-normal through the Shapiro-Wilk test in 3.2.3. Each house is categorised according to its overall quality at 10 levels, ranging from 1 to 10, 1 being very bad and 10 being excellent. The following plot illustrates the distributions of log(*SalePrice*) for the different overall qualities.




To determine if *OverallQual* affects log(*SalePrice*), we will test the null hypothesis that the distributions of log(*SalePrice*) across all qualities are the same. We will do so using the Kruskal-Wallies test, a non-parametric method for analysing data given that the distributions are not normal (as shown in 3.2.3).

$H_0$: All distributions of log(*SalePrice*) across all overall quality ratings are identical;
$H_1$: The distributions of log(*SalePrice*) across all overall quality ratings are not all identical

Both the manual and direct method for testing yield the same result, where the p-value < 0.01 and the null hypothesis is rejected with 99% confidence.

```
> s_sq;T_stat;pvalue
[1] 177265.4
[1] 968.5993
[1] 0
> kruskal.test(comp2$price,comp2$qual) #rejected, not the same

        Kruskal-Wallis rank sum test

data:  comp2$price and comp2$qual
Kruskal-Wallis chi-squared = 968.6, df = 9, p-value < 2.2e-16
```

### 4.3. Multiple Linear Regression

In this section, our goal is to construct a multiple linear regression model for log(*SalePrice*). To assess the adequacy of our model, we'll partition the data into training and testing sets. We will employ the train-test-split method, dividing the data into an 80-20 ratio for training and testing, respectively.

The result on the train data is shown in the R output below. The fitted model is:

$$\begin{aligned}\log(SalePrice) = {} & \beta_0 + \beta_1 LotArea + \beta_2 GarageArea + \beta_3 BedroomAbvGr + \beta_4 Toilets + \beta_5 OverallQual_2 + \\ & \beta_6 OverallQual_3 + \beta_7 OverallQual_4 + \beta_8 OverallQual_5 + \beta_9 OverallQual_6 + \beta_{10} OverallQual_7 + \\ & \beta_{11} OverallQual_8 + \beta_{12} OverallQual_9 + \beta_{13} OverallQual_{10} + \beta_{14} Age + \beta_{15} Remodelled + \\ & \beta_{16} MSZoning_{FV} + \beta_{17} MSZoning_{RH} + \beta_{18} MSZoning_{RL} + \beta_{19} MSZoning_{RM} + \beta_{20} Exterior1st_{AsphShn} + \\ & \beta_{21} Exterior1st_{BrkComm} + \beta_{22} Exterior1st_{BrkFace} + \beta_{23} Exterior1st_{CBlock} + \beta_{24} Exterior1st_{CementBd} + \\ & \beta_{25} Exterior1st_{HdBoard} + \beta_{26} Exterior1st_{ImStucc} + \beta_{27} Exterior1st_{MetalSd} + \beta_{28} Exterior1st_{Plywood} + \\ & \beta_{29} Exterior1st_{Stone} + \beta_{30} Exterior1st_{Stucco} + \beta_{31} Exterior1st_{VinylSd} + \beta_{32} Exterior1st_{WdSdng} + \\ & \beta_{33} Exterior1st_{WdShing}\end{aligned}$$

**Assumptions for Ordinary Least Square Linear Regression**

Before we proceed with the ordinary least square (OLS) linear regression, we need to ensure the regularity assumptions, known as the Gauss-Markov Assumptions are satisfied (Wooldridge, 2009). This justifies the use of OLS estimates as the Best Unbiased Linear Estimator (BLUE) (Stock & Watson, 2015).

- Zero conditional mean assumption. The conditional distribution of error term, $u_i$ given our regressors has a mean of zero, i.e. $E(u_i \mid X_i) = 0$. Suppose if the zero conditional mean assumption is not satisfied, we will need to introduce instrument variables to correct the omitted variable bias (a form on endogeneity issue in OLS).
- Independent and identical data distribution. This assumption holds when our dataset is drawn from the same population in Ames, Iowa (identical) and is done by random selection (independent). This assumption is crucial as it justifies the usage of statistical concepts such as Central Limit Theorem and Law of Large Numbers, used throughout the OLS regression model.
- Linear in Parameters. The model in the population is written as shown above, where $\beta_0, \beta_1, \ldots, \beta_{33}$ are linear in parameters.
- No perfect collinearity. In the sample (and therefore in the population), none of the independent variables is constant, and there are no exact linear relationships among the independent variables.

### 4.3.1. Results

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.008e+01  1.555e-01  64.829  < 2e-16 ***
LotArea                     4.270e-06  5.236e-07   8.155 9.16e-16 ***
GarageArea                  3.602e-04  3.150e-05  11.432  < 2e-16 ***
BedroomAbvGr                3.234e-02  8.005e-03   4.040 5.70e-05 ***
Toilets                     9.752e-02  1.280e-02   7.617 5.46e-14 ***
factor(OverallQual)2        1.764e-01  1.835e-01   0.961 0.336671
factor(OverallQual)3        5.126e-01  1.365e-01   3.755 0.000182 ***
factor(OverallQual)4        5.917e-01  1.287e-01   4.596 4.78e-06 ***
factor(OverallQual)5        7.306e-01  1.283e-01   5.696 1.56e-08 ***
factor(OverallQual)6        8.631e-01  1.286e-01   6.713 3.01e-11 ***
factor(OverallQual)7        9.899e-01  1.293e-01   7.655 4.11e-14 ***
factor(OverallQual)8        1.162e+00  1.305e-01   8.906  < 2e-16 ***
factor(OverallQual)9        1.420e+00  1.338e-01  10.611  < 2e-16 ***
factor(OverallQual)10       1.458e+00  1.392e-01  10.470  < 2e-16 ***
Age                        -6.433e-04  3.132e-04  -2.054 0.040194 *
factor(Remodelled)1         1.544e-02  1.441e-02   1.072 0.284070
factor(MSZoning)FV          4.874e-01  7.779e-02   6.266 5.25e-10 ***
factor(MSZoning)RH          4.500e-01  8.826e-02   5.099 4.00e-07 ***
factor(MSZoning)RL          5.660e-01  7.278e-02   7.777 1.66e-14 ***
factor(MSZoning)RM          4.279e-01  7.337e-02   5.832 7.13e-09 ***
factor(Exterior1st)AsphShn -2.001e-01  1.850e-01  -1.082 0.279500
factor(Exterior1st)BrkComm -4.523e-01  1.369e-01  -3.305 0.000981 ***
factor(Exterior1st)BrkFace  1.458e-01  5.523e-02   2.641 0.008387 **
factor(Exterior1st)CBlock   7.524e-02  1.847e-01   0.407 0.683817
factor(Exterior1st)CemntBd  1.226e-01  5.378e-02   2.279 0.022874 *
factor(Exterior1st)HdBoard  4.479e-02  4.884e-02   0.917 0.359238
factor(Exterior1st)ImStucc  9.127e-02  1.854e-01   0.492 0.622603
factor(Exterior1st)MetalSd  6.220e-02  4.807e-02   1.294 0.195930
factor(Exterior1st)Plywood  7.554e-02  5.079e-02   1.487 0.137215
factor(Exterior1st)Stone    3.391e-01  1.350e-01   2.511 0.012178 *
factor(Exterior1st)Stucco  -1.167e-02  5.993e-02  -0.195 0.845601
factor(Exterior1st)VinylSd  7.254e-02  4.876e-02   1.488 0.137147
factor(Exterior1st)Wd Sdng  5.515e-02  4.801e-02   1.149 0.250863
factor(Exterior1st)WdShing -4.898e-03  5.993e-02  -0.082 0.934865
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

NB: $OverallQual_1$, $MSZoning_C$ and $Exterior1st_{AsbShng}$ are omitted from the regression model to prevent the case of perfect multicollinearity.

### 4.3.2. Summary and Findings

Surprisingly, we see that remodelling is insignificant, with a rather small impact on log(*SalePrice*) compared to overall quality. This is in contrary to our previous test in 4.2.1 which shows that there is a statistical difference in the mean of log(*SalePrice*) for remodelled vs not remodelled, showing that remodelled has an effect on log(*SalePrice*). This difference in results might be because of the partialling-out effect of the other variables in a multiple regression model. It seems that what is of importance to buyers is not about whether the house is remodelled, but the overall quality of the house at purchase (i.e. a house of good quality without remodelling could sell for nearly as well as a remodelled house of the same quality).

Another point of interest is that the ascending trend of coefficients for the overall quality dummies from 2 to 10 conforms to the ordinal nature of the *OverallQual* variable (i.e. as quality moves further from 0 (the base), it has an increasingly positive effect on log(*SalePrice*). For overall quality, it is mostly significant except for the value ‘2’ as there are too little data points in ‘2’ so there is higher standard deviation and hence causing it to be insignificant.

We can also observe that houses in residential zones sell better than that in other zones (with Commercial as the base – see appendix A for legend). Particularly, residential areas with low density (RL) have the highest positive impact on log(*SalePrice*).

Finally, we apply the model to the test data to generate the predicted values. We use the predicted log(*SalePrice*) to get the predicted sale price (using exponential transformation). Comparing the predicted sale price to the actual sale price, we calculated the absolute difference (i.e. square root of the mean squared error). Our predicted sale price using the model above deviates from the actual sale price by ±$34072, which is approximately ±20.9% of the median sale price ($163,000). This corresponds to our adjusted R-squared value as shown above (80.01%), which means that our model explains approximately 80% of the variation in sale price for this dataset.

## 5. Conclusion and Discussion

In this study, we investigated numerous factors influencing the sale price of houses in Ames, Iowa. Through statistical analyses and modelling in R, we addressed several research objectives aimed at understanding the dynamics of property pricing.

Firstly, we examined the distribution of our variables, providing insights into their distribution. This initial exploration allowed us to identify the appropriate approach for our statistical tests. Since the dataset consisted of normal and non-normally distributed variables, we applied both parametric and non-parametric methods for our analyses.

Next, we justified the inclusion of our numerical variables in our model by considering that they showed a positive and significant correlation with the independent variable. These numerical variables include number of bedrooms and toilets, sizes of lot and garage, and age of the house. For our categorical variables, we justified using statistical tests like two sample t-test, ANOVA as well as non-parametric tests. These variables include whether the house underwent remodelling within five years preceding its sale as well as the zoning, exterior material and overall quality of the house, for which the tests show significant difference.

Although the variable remodelled appears to be insignificant, we believe that its inclusion is necessary due to the significant results of the two-sample t-tests, showing that there is a difference in mean value for ‘remodelled’ and ‘not remodelled’.

Lastly, we found out that house price is predictable with our regression model, up to 80% accuracy. While the report's findings are intriguing, it's crucial to note that they are based on data from just one season available online. Moreover, as data capture techniques evolve, more sophisticated indices may be considered. A more extensive and in-depth analysis of house sale price data, employing advanced analytical methods, would be required to make more robust predictions.

# 6. Appendix

Listing of code and output from R.

Variance and two sample t-tests on *Remodelled* (4.2.1)

```
> boxplot(housing_data$logprice ~ housing_data$Remodelled, xlab='Remodelled', ylab = 'log(SalePrice)')
>
> # get log salprice from houses that are remodelled and not remodelled
> remodelled_sp <- housing_data[housing_data$Remodelled == 1,]$logprice
> remodelled_no_sp <- housing_data[housing_data$Remodelled == 0,]$logprice
>
> # test for equal variance
> var.test(housing_data$logprice ~ housing_data$Remodelled)

        F test to compare two variances

data:  housing_data$logprice by housing_data$Remodelled
F = 0.90368, num df = 1033, denom df = 423, p-value = 0.2076
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7678403 1.0577794
sample estimates:
ratio of variances
         0.9036816

> # 2 sample t-test
> t.test(remodelled_sp, remodelled_no_sp, var.equal = T)

        Two Sample t-test

data:  remodelled_sp and remodelled_no_sp
t = 14.897, df = 1456, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2778462 0.3621157
sample estimates:
mean of x mean of y
 12.25094  11.93096
```

ANOVA test for *MSZoning* (4.2.2)

```
> # anova test for MSZoning
> boxplot(housing_data$logprice ~ housing_data$MSZoning, xlab='MSZoning', ylab='log_saleprice')
>
> summary(aov(housing_data$logprice ~ factor(housing_data$MSZoning)))
                             Df Sum Sq Mean Sq F value Pr(>F)
factor(housing_data$MSZoning)    4  40.94  10.234   77.51 <2e-16 ***
Residuals                     1453 191.85   0.132
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> pairwise.t.test(housing_data$logprice, housing_data$MSZoning, p.adjust.method = 'none')

        Pairwise comparisons using t tests with pooled SD

data:  housing_data$logprice and housing_data$MSZoning

   C (all) FV      RH      RL
FV < 2e-16 -       -       -
RH 1.7e-05 1.1e-06 -       -
RL < 2e-16 0.00054 0.00025 -
RM 1.1e-06 < 2e-16 0.54523 < 2e-16

P value adjustment method: none
```

Krusal-Wallies Test for log(*SalePrice*) on *Exterior* (4.2.3)

```
> #for exterior
>
> exterior<-df$Exterior1st
> price<-log(df$SalePrice)
> rank<-rank(price, ties.method="average")
>
> comp<-data.frame(exterior,price,rank)
>
> ranksumi<-aggregate(comp$rank,list(comp$exterior),FUN=sum)
> ni<-aggregate(comp$rank, list(comp$exterior),FUN=length)
> N<-length(exterior)
>
> s_sq<-(sum(rank^2)-N*(N+1)^2/4)/(N-1)
> T_stat<-(sum(ranksumi[,2]^2/ni[,2])-(N*(N+1)^2/4))/s_sq
> pvalue<-1-pchisq(T_stat,14)
>
> s_sq;T_stat;pvalue
[1] 177751.9
[1] 304.0446
[1] 0
>
> kruskal.test(comp$price,comp$exterior) #rejected, not the same

        Kruskal-Wallis rank sum test

data:  comp$price and comp$exterior
Kruskal-Wallis chi-squared = 304.04, df = 14, p-value < 2.2e-16
```

Krusal-Wallies Test for log(*SalePrice*) on *OverallQual* (4.2.4)

```
> #for overall quality
>
> qual<-df$OverallQual
> comp2<-data.frame(qual,price,rank)
> ranksumi<-aggregate(comp2$rank,list(comp2$qual),FUN=sum)
> ni<-aggregate(comp2$rank, list(comp2$qual),FUN=length)
> N<-length(qual)
>
> s_sq<-(sum(rank^2)-N*(N+1)^2/4)/(N-1)
> T_stat<-(sum(ranksumi[,2]^2/ni[,2])-(N*(N+1)^2/4))/s_sq
> pvalue<-1-pchisq(T_stat,9)
>
> s_sq;T_stat;pvalue
[1] 177265.4
[1] 968.5993
[1] 0
>
> kruskal.test(comp2$price,comp2$qual) #rejected, not the same

        Kruskal-Wallis rank sum test

data:  comp2$price and comp2$qual
Kruskal-Wallis chi-squared = 968.6, df = 9, p-value < 2.2e-16
```

Code for boxplots (4.2.3 and 4.2.4)

```
> boxplot(comp$price~comp$exterior, xlab="Exterior",ylab="log(SalePrice)")
> boxplot(comp2$price~comp2$qual, xlab="OverallQual",ylab="log(SalePrice)")
```

Code for multiple linear regression (4.3.1)

```
> set.seed(1) # Set the seed for reproducibility
> sample_size <- floor(0.8 * nrow(data))  # 80% train, 20% test
> train_indices <- sample(seq_len(nrow(data)), size = sample_size)
>
> train_data <- data[train_indices, ]
> test_data <- data[-train_indices, ]
>
> model <- lm(log_SalePrice~LotArea+GarageArea+BedroomAbvGr+
+               Toilets+factor(OverallQual)+Age+factor(Remodelled)+factor(MSZoning)+factor(Exterior
1st),data=train_data)
>
> summary(model)
> mse <- mean((exppredicted_values - actual_values)^2)
> sqrt(mse)
[1] 34072.34
>
> median(data$SalePrice)
[1] 163000
```

## 7. References


Anna Montoya, DataCanary. (2016). House Prices - Advanced Regression Techniques. *Kaggle*. https://kaggle.com/competitions/house-prices-advanced-regression-techniques

Stock, J. H., & Watson, M. W. (2015). 6.5 The Least Squares Assumptions in Multiple Regression. In *Introduction to Econometrics* (3rd ed., Ser. Global Edition, pp. 245–247). essay, Pearson Education Limited.

Wooldridge, J. M. (2009). 3.3 The Expected Value of the OLS Estimators. In *Introductory Econometrics* (4th ed., pp. 84–102). essay, South-Western Cengage Learning.