311 EDA

Ana Paula Felix de Queiroz

## 1. Load the Dataset In R

## Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour  
## 1 1 60 RL 65 8450 Pave <NA> Reg Lvl  
## 2 2 20 RL 80 9600 Pave <NA> Reg Lvl  
## 3 3 60 RL 68 11250 Pave <NA> IR1

...

0 320 0 0 <NA> MnPrv Shed  
## MiscVal MoSold YrSold SaleType SaleCondition SalePrice  
## 1 0 2 2008 WD Normal 208500  
## 2 0 5 2007 WD Normal 181500  
## 3 0 9 2008 WD Normal 223500  
## 4 0 2 2006 WD Abnorml 140000  
## 5 0 12 2008 WD Normal 250000  
## 6 700 10 2009 WD Normal 143000

Inspect the contents of the dataframe:

## 'data.frame': 1460 obs. of 81 variables:  
## $ Id : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ MSSubClass : int 60 20 60 70 60 50 20 60 50 190 ...  
## $ MSZoning : chr "RL" "RL" "RL" "RL" ...  
## $ LotFrontage : int 65 80 68 60 84 85 75 NA 51 50 ...  
## $ LotArea : int 8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 .........  
## $ SaleType : chr "WD" "WD" "WD" "WD" ...  
## $ SaleCondition: chr "Normal" "Normal" "Normal" "Abnorml" ...  
## $ SalePrice : int 208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...

## [1] 1460 81

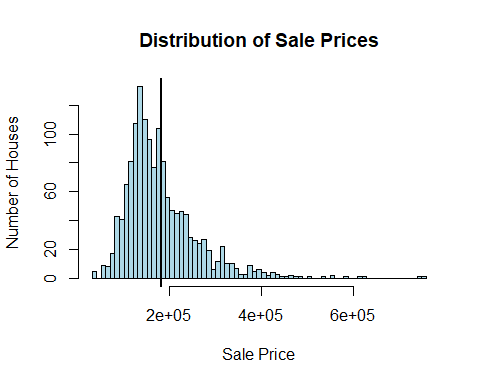
## 

## 2. Explore Data Distributions

Write code to produce histograms showing the distributions of SalePrice, TotRmsAbvGrd, and OverallCond.

### Sale Price

In the cell below, produce a histogram for SalePrice.



Now, print out the mean, median, and standard deviation:

## [1] "Mean: 180921.195890411"

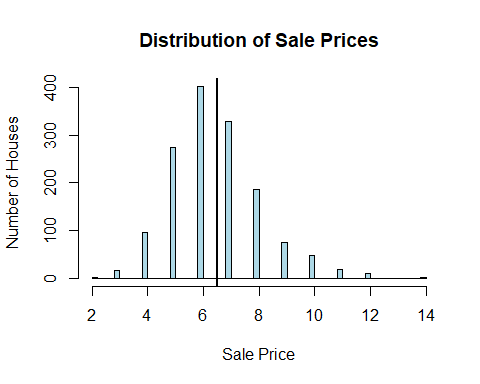
## [1] "Median: 163000"

## [1] "Standard Deviation: 79442.5028828866"

### 

### Total Rooms Above Grade

In the cell below, produce a histogram for TotRmsAbvGrd.



Now, print out the mean, median, and standard deviation:

## [1] "Mean: 6.51780821917808"

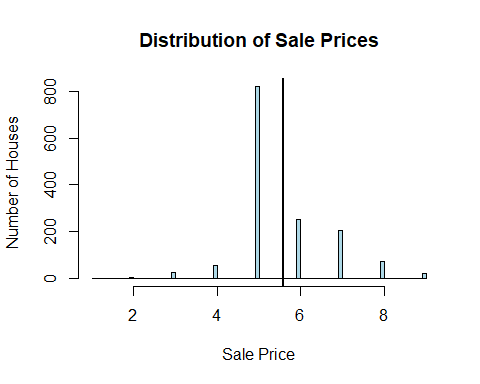
## [1] "Median: 6"

## [1] "Standard Deviation: 1.62539329058405"

### 

### Overall Condition

In the cell below, produce a histogram for OverallCond.



Now, print out the mean, median, and standard deviation:

## [1] "Mean: 5.57534246575342"

## [1] "Median: 5"

## [1] "Standard Deviation: 1.11279933671273"

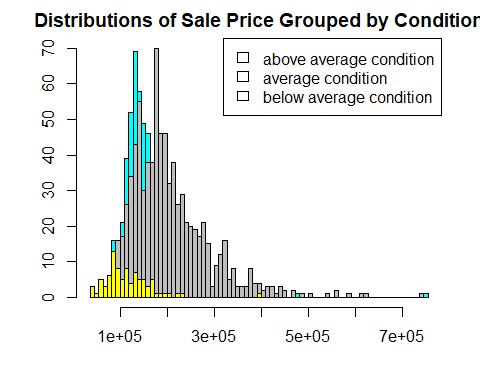
## 3. Explore Differences between Subsets

As you might have noted in the previous step, the overall condition of the house seems like we should treat it as more of a categorical variable, rather than a numeric variable.

One useful way to explore a categorical variable is to create subsets of the full dataset based on that categorical variable, then plot their distributions based on some other variable. Since this dataset is traditionally used for predicting the sale price of a house, let’s use SalePrice as that other variable.

In the cell below, create three variables, each of which represents a record-wise subset of df (meaning, it has the same columns as df, but only some of the rows).

* below\_average\_condition: home sales where the overall condition was less than 5
* average\_condition: home sales where the overall condition was exactly 5
* above\_average\_condition: home sales where the overall condition was greater than 5



## 

## 4. Explore Correlations

To understand more about what features of these homes lead to higher sale prices, let’s look at some correlations. We’ll return to using the full df, rather than the subsets.

In the cell below, print out both the name of the column and the Pearson correlation for the column that is ***most positively correlated*** with SalePrice (other than SalePrice, which is perfectly correlated with itself).

We’ll only check the correlations with some kind of numeric data type.

You can import additional libraries, although it is possible to do this just using pandas.

## Most Positively Correlated Column: OverallQual

## Maximum Correlation Value: 0.7909816

## Most Negativelly Correlated Column: KitchenAbvGr

## Minimun Correlation Value: -0.1359074

Once you have your answer, edit the code below so that it produces a scatter plot of the relevant columns.

A picture containing text, screenshot, diagram

Description automatically generated

## 5. Engineer and Explore a New Feature

Here the code is written for you, all you need to do is interpret it.

We note that the data spans across several years of sales:

##   
## 2006 2007 2008 2009 2010   
## 314 329 304 338 175

Maybe we can learn something interesting from the age of the home when it was sold. This uses information from the YrBuilt and YrSold columns, but represents a truly distinct feature.

