Final Project IS605

Sharon Morris

5/06/2017

Kaggle.com Home Price Completion

1. You are to register for Kaggle.com (free) and compete in the House Prices: Advanced Regression Techniques competition.

House Prices: Advanced Regression Techniques competition Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

Load Library

Define Variables 2. Pick one of the quantitative independent variables from the training data set (train.csv) , and define that variable as X. Pick SalePrice as the dependent variable, and define it as Y for the next analysis.

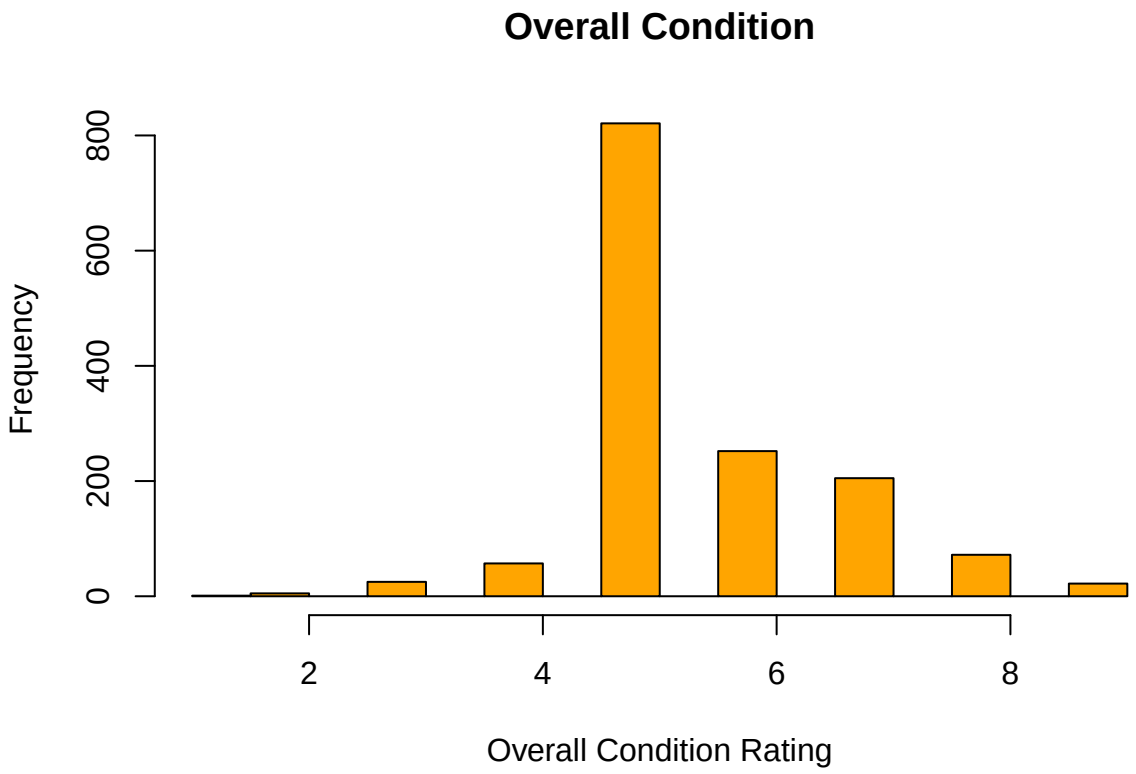
```
train <- read.csv(paste0("https://raw.githubusercontent.com/indianspice/IS605/master/Final%20Project/tra

#Quantitative independent variable OverallCond
X <- train$OverallCond
Y <- train$SalePrice

#Summary of independent and dependent variables
summary(X)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   5.000   5.000   5.575   6.000   9.000
```

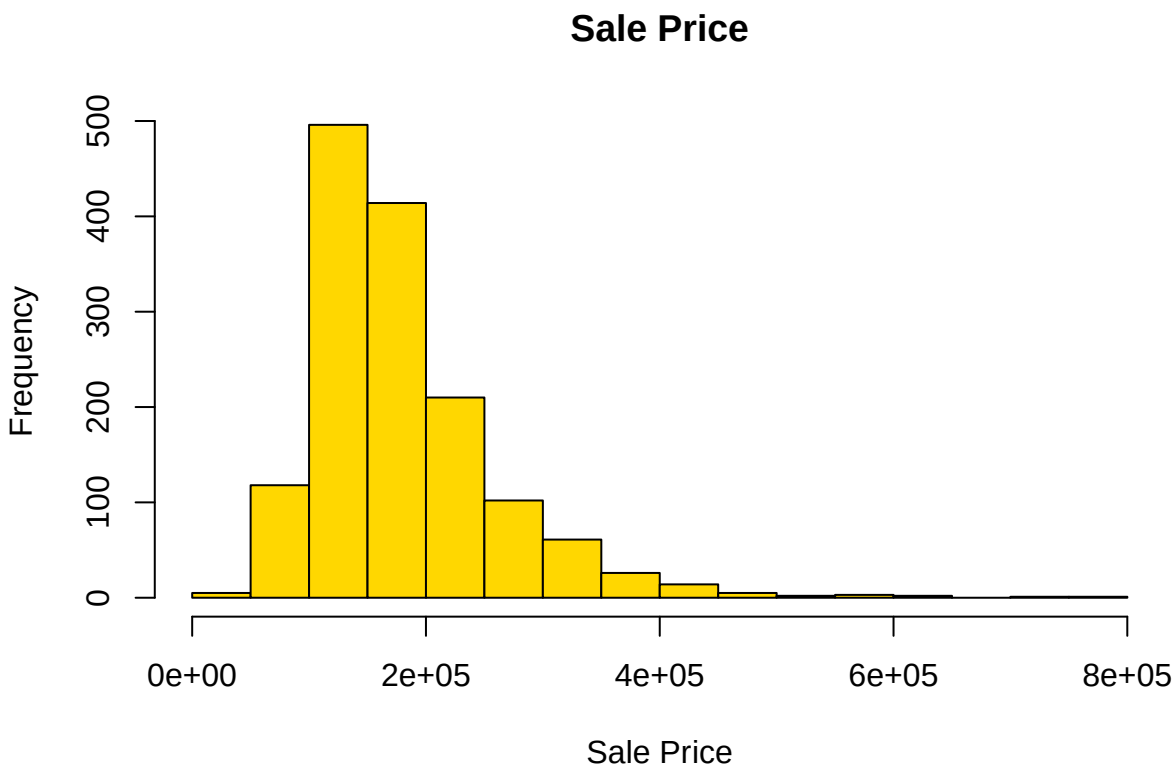
```
summary(Y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  130000  163000  180900  214000  755000
```

```
hist(train$OverallCond,
     main = "Overall Condition",
     xlab = "Overall Condition Rating",
     col = "orange")
```




```
hist(train$SalePrice,
     main = "Sale Price",
     xlab = "Sale Price",
     col = "gold")
```



Probability Calculate as a minimum the below probabilities a through c. Assume the small letter "x" is estimated as the 4th quartile of the X variable, and the small letter "y" is estimated as the 2d quartile of the Y variable. Interpret the meaning of all probabilities. a. $P(X > x|Y > y)$

```
quantile(X)
```

```
##   0%  25%  50%  75% 100%
##    1    5    5    6    9
```

```
quantile(Y)
```

```
##     0%    25%    50%    75%   100%
##  34900 129975 163000 214000 755000
```

The probability of a house at the second quartile has an overall condition rating at the third quartile is 13.5%

```
(sum(Y > 163000 & X > 6) / length(X)) / (sum(Y > 163000) / length(Y))
```

```
## [1] 0.1346154
```

b. $P(X > x, Y > y)$

```
sum(X > 6 & Y > 163000) / length(X)
```

```
## [1] 0.06712329
```

c. $P(X < x|Y > y)$

```
(sum(Y > 163000 & X < 6) / length(X)) / (sum(Y > 163000) / length(Y))
```

```
## [1] 0.7568681
```

Count Table

```
#Sum rows with sales prices > 163000
(YQ2 <- sum(Y > 163000))
```

```
## [1] 728
```

```
#Count of all rows sales prices
(YTotal <- length(Y))
```

```
## [1] 1460
```

```
#Count of all rows of overall condition
(XTotal <- length(X))
```

```
## [1] 1460
```

```
#Rows greater than 163,000 and greater than 6
(Ygt <- sum(Y > 163000 & X > 6))
```

```
## [1] 98
```

```
#Rows greater than 163,000 and less than 6
(Ylt <- sum(Y > 163000 & X < 6))
```

```
## [1] 551
```

```
sum(train[,'OverallCond'] <= 6 & train[,'SalePrice'] > 163000)
```

```
## [1] 630
```

	<=2nd Quartile	**>2nd Quartile**	Total
<= 3rd Quartile	728	630	1,358