Final_Project_Draft

Eldonna Weber

5/3/2022

Intro

What:

For my final project, I analyzed Zillow home sales data for Mound, MN over the past two years, specially focusing on 3 bedroom, 2+ bathroom houses.

Why:

I have a rental house in Mound, MN that I'm planning to put on the market in June and I'm interested to see if the past home sales data will reflect any interesting trends or help me predict the final sale price of my house.

How:

I started by manually curating a data frame ('moundsales') from Zillow (Data Source).

The data frame includes 150 instances and 12 variables: 'ID', 'Address', 'Beds', 'Baths', 'SQFT', 'Month.Sold', 'Year.Sold', 'Sale.Price', 'Lake.Front', 'Year.Built', 'Lot.Size.Sqft', 'Most.Recent.Tax.Assessment'.

I conducted exploratory data analysis of the data frame, created new data subsets, summarized data via histogram, scatter and box plots, determined probability, determined mean/standard deviation/z-score, and conducted regression.

Body

According to Zillow, the were 150 three bedroom, two-plus bathroom houses sold in Mound, MN between April 2020 and April 2022. Mound is a community on Lake Minnetonka, therefore 32 of the 150 houses are lake front properties. This is an important distinction because these houses sold for significantly more than non-lake front houses, usually in the million dollar range.

My house was built in 1948, has 1859 finished sqft and includes three bedrooms and two baths. The house sits on a 10,000 sqft, non-lake front, lot and was assessed at $245,000 in 2021.

I'd like to understand the relationship between sale price and the most recent tax assessment data and determine the probability of my house selling for over $300,000.

Topics From Class

Topic 1: Data Basics

For this observational study, I created a data frame labeled 'moundsales'. In addition, I created a subset labeled 'nonlf' that only includes non-lake front houses because including lake front houses skews the sale price and tax assessment data.

```
moundsales <- read.table("moundsales.csv", sep = ",", header = TRUE)
```

```
nonlf <- subset(moundsales, moundsales$Lake.Front == 'N')
```

I conducted a number of various data exploration activities in R (e.g., dim, length, names, summary, head, tail, etc.) to anlayze both the 'moundsales' and 'nonlf' data frames. See Appendix for more details.

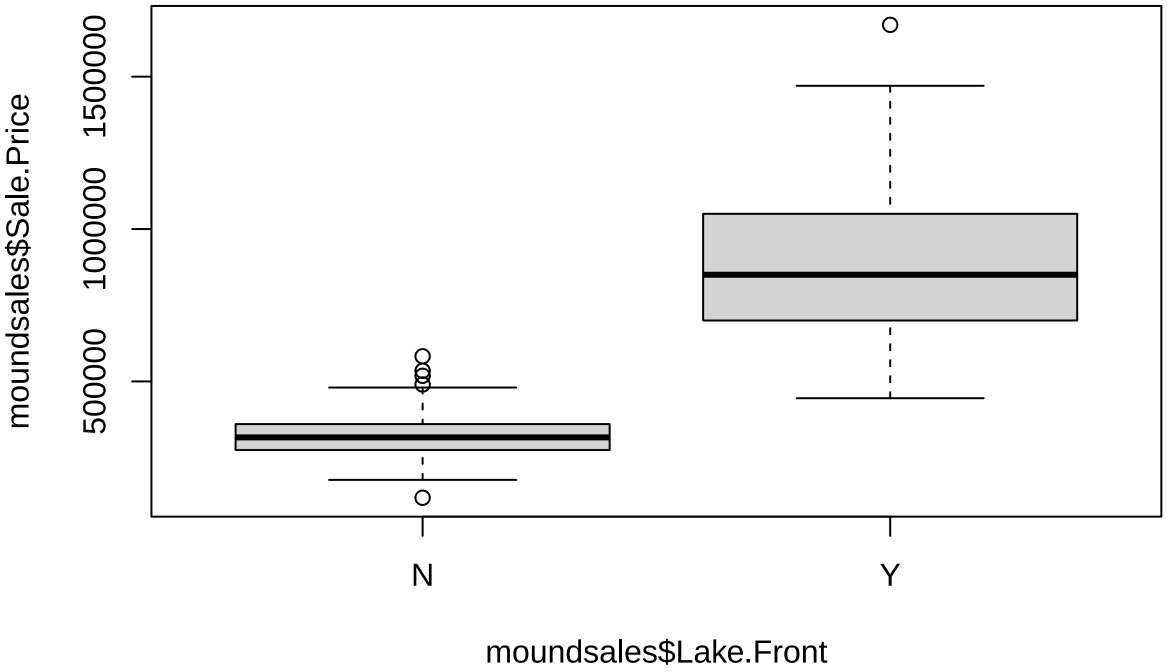
Topic 2: Summarizing Data

The mean sale price for lake front houses was $888,192, while the mean sale price for non-lake front houses is $324,801.

```
lakefront <- subset(moundsales, moundsales$Lake.Front == 'Y')
summary(lakefront)
```

```
##        ID           Address               Beds       Baths            SQFT
##  Min.   :  2.00   Length:32          Min.   :3   Min.   :2.000   Min.   :1302
##  1st Qu.: 38.50   Class :character   1st Qu.:3   1st Qu.:2.000   1st Qu.:1718
##  Median : 88.00   Mode  :character   Median :3   Median :3.000   Median :2239
##  Mean   : 78.38                      Mean   :3   Mean   :3.078   Mean   :2316
##  3rd Qu.:119.75                      3rd Qu.:3   3rd Qu.:4.000   3rd Qu.:2552
##  Max.   :150.00                      Max.   :3   Max.   :5.000   Max.   :4153
##   Month.Sold          Year.Sold      Sale.Price        Lake.Front
##  Length:32          Min.   :2020   Min.   : 445000   Length:32
##  Class :character   1st Qu.:2020   1st Qu.: 700000   Class :character
##  Mode  :character   Median :2020   Median : 850000   Mode  :character
##                     Mean   :2021   Mean   : 888192
##                     3rd Qu.:2021   3rd Qu.:1035000
##                     Max.   :2022   Max.   :1670000
##    Year.Built   Lot.Size.Sqft   Most.Recent.Tax.Assessment
##  Min.   :1910   Min.   : 4356   Min.   : 411000
##  1st Qu.:1946   1st Qu.: 7840   1st Qu.: 637750
##  Median :1972   Median : 9583   Median : 705000
##  Mean   :1968   Mean   :10911   Mean   : 749813
##  3rd Qu.:1990   3rd Qu.:12306   3rd Qu.: 855750
##  Max.   :2020   Max.   :28750   Max.   :1344000
```

```
boxplot(moundsales$Sale.Price ~ moundsales$Lake.Front)
```



Only 46 houses, or 39%, sold for under $300,000, which supports my realator's claim that there has been a shortage of houses under $300,000 for sale in Mound, MN.

```
under300k <- subset(nonlf, nonlf$Sale.Price < 300000)
dim(under300k)
```

```
## [1] 46 12
```

Of the 46 houses that sold for more than $300,000, 16 had a recent tax assessment of $245,000 or below.

```
S300kT245 <- subset(nonlf, nonlf$Sale.Price > 300000 & nonlf$Most.Recent.Tax.Assessment < 246000)
dim(S300kT245)
```

```
## [1] 16 12
```

Topic 3: Probability

The probability of a three bedroom, two-plus bathroom house in Mound, MN selling for $300,000 or more given it is not lake front is 61%.

```
Over299k <- subset(nonlf, nonlf$Sale.Price > 299999)
72/118
```

```
## [1] 0.6101695
```

The probability of a three bedroom, two-plus bathroom house in Mound, MN selling for $1 million or more given it has lake front is 28%.

```
table(moundsales$Lake.Front)
```

```
##
##   N   Y
## 118  32
```

```
milsale <- subset(moundsales, moundsales$Sale.Price > 999999)
View(milsale)
```

```
9/32
```

```
## [1] 0.28125
```

Topic 4: Normal Distribution

The sale price distribution for the 'nonlf' data set is unimodal and normal with a slight skew to the right. The distribution has the following statistics:

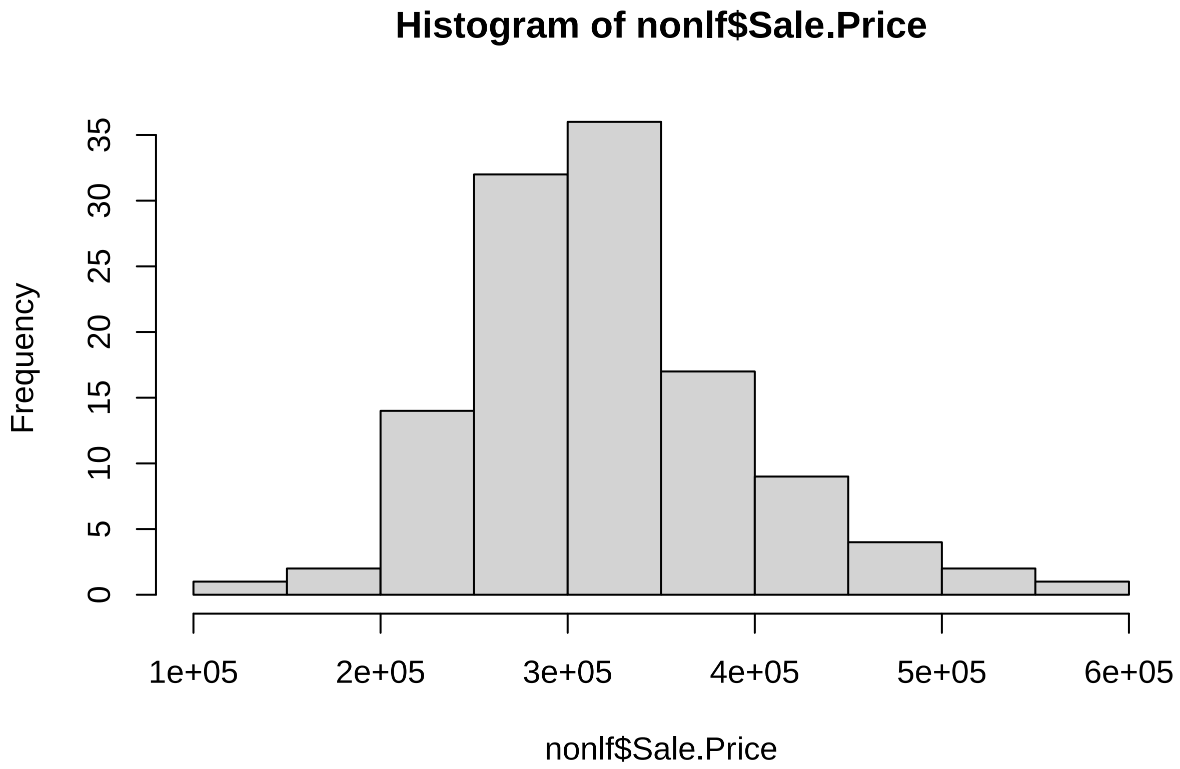
mean = 324801 sd = 74870 x = 300000 Z-score = - .331

There is a 63% probability that my house will sell for over $300k.

```
format(nonlf$Sale.Price, scientific = FALSE)
```

```
##   [1] "341000" "346000" "330000" "315000" "322000" "318000" "360000" "427900"
##   [9] "315000" "535000" "299251" "400000" "273000" "280000" "350000" "118400"
##  [17] "290000" "302000" "335000" "302000" "490000" "300000" "386000" "310000"
##  [25] "250000" "430000" "370000" "387500" "365000" "295000" "385000" "320000"
##  [33] "292000" "452061" "425000" "395000" "287000" "582500" "425000" "330000"
##  [41] "232500" "375000" "310000" "300000" "480000" "305000" "518500" "350000"
##  [49] "342500" "355000" "333000" "354000" "320000" "420000" "396000" "340000"
##  [57] "339900" "290000" "271900" "320000" "312999" "215000" "281500" "355000"
##  [65] "370000" "325000" "435000" "305000" "360000" "289500" "449051" "280000"
##  [73] "325000" "240000" "235000" "400000" "390000" "300000" "250000" "465000"
##  [81] "237000" "279000" "177000" "265000" "271000" "350000" "270000" "320678"
##  [89] "270000" "272500" "275000" "268000" "339000" "200000" "262000" "281685"
##  [97] "315000" "350000" "278900" "275000" "435000" "225000" "319000" "236400"
## [105] "335000" "245000" "265000" "430000" "270000" "226600" "295000" "240000"
## [113] "275000" "289900" "228000" "330000" "329900" "222000"
```

```
hist(nonlf$Sale.Price)
```



```
sd(nonlf$Sale.Price)
```

```
## [1] 74870.19
```

Given:

mean = 324801 sd = 74870 x = 300000

```
(300000 - 324801)/74870
```

```
## [1] -0.3312542
```

Answer: z-score = - .331

```
1 - pnorm(-.331)
```

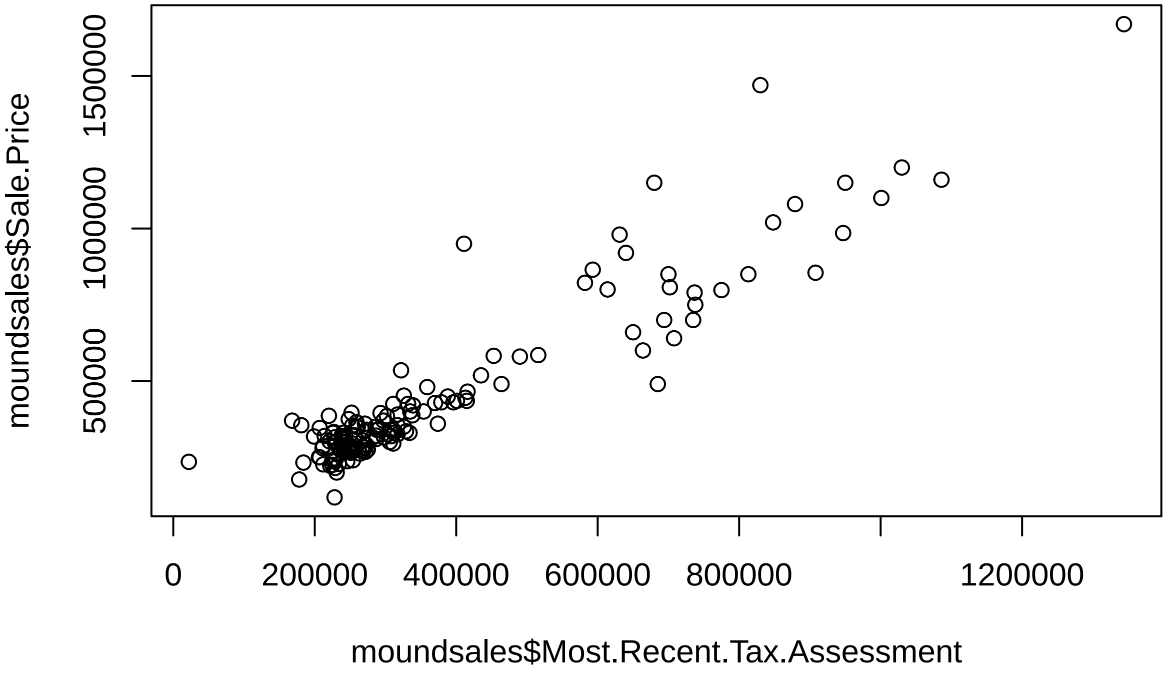
```
## [1] 0.6296778
```

Answer: 63%

Topic 5: Regression

According to the regression data below, there is a very high correlation between the 'Sale.Price' and 'Most.Recent.Tax.Assessment' data. The intercept value indicates that one could expect a house to sell for a premium of $20,934 over the most recent tax assessment for the house.

```
plot(moundsales$Sale.Price ~ moundsales$Most.Recent.Tax.Assessment)
```



```
lm(moundsales$Sale.Price ~ moundsales$Most.Recent.Tax.Assessment)
```

```
##
## Call:
## lm(formula = moundsales$Sale.Price ~ moundsales$Most.Recent.Tax.Assessment)
##
## Coefficients:
##                           (Intercept)  moundsales$Most.Recent.Tax.Assessment
##                             20934.465                                  1.131
```

```
summary(lm(moundsales$Sale.Price ~ moundsales$Most.Recent.Tax.Assessment))
```

```
##
## Call:
## lm(formula = moundsales$Sale.Price ~ moundsales$Most.Recent.Tax.Assessment)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -305751  -44941  -16431   30781  510236
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           2.093e+04  1.560e+04   1.342    0.182
## moundsales$Most.Recent.Tax.Assessment 1.131e+00  3.580e-02  31.596   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97490 on 148 degrees of freedom
## Multiple R-squared:  0.8709, Adjusted R-squared:   0.87
## F-statistic: 998.3 on 1 and 148 DF,  p-value: < 2.2e-16
```

Conclusion

In conclusion, I determined that there is a very high correlation between sale price and the most recent tax assessment for 3 bedroom, 2+ bath houses sold in Mound, MN in the past two years. During this time, 72 non-lake front houses were sold for $300,000 or more and there is a 63% probability that my house will sell for $300,000 or more. As a side note, I also learned that if my property were to have lake front, there would be a 28% probability of selling my property for $1 million or more!

APPENDIX

'moundsales' Data Set

```
moundsales <- read.table("moundsales.csv", sep = ",", header = TRUE)
```

```
dim(moundsales)
```

```
## [1] 150  12
```

```
length(dim(moundsales))
```

```
## [1] 2
```

```
names(moundsales)
```

```
##  [1] "ID"                         "Address"
##  [3] "Beds"                       "Baths"
##  [5] "SQFT"                       "Month.Sold"
##  [7] "Year.Sold"                  "Sale.Price"
##  [9] "Lake.Front"                 "Year.Built"
## [11] "Lot.Size.Sqft"              "Most.Recent.Tax.Assessment"
```

```
head(moundsales)
```

```
##   ID            Address Beds Baths SQFT Month.Sold Year.Sold Sale.Price
## 1  1 4674 Cumberland Rd    3   2.0 2369        Apr      2022     341000
## 2  2  6641 Halstead Ave    3   3.0 2092        Apr      2022     980000
## 3  3   4767 Richmond Rd    3   2.0 1500        Apr      2022     346000
## 4  4 4515 Manchester Rd    3   2.0 1931        Mar      2022     330000
## 5  5  6117 Beachwood Rd    3   2.5 2400        Mar      2022     315000
## 6  6     4812 Lanark Rd    3   2.0 2010        Mar      2022     322000
##   Lake.Front Year.Built Lot.Size.Sqft Most.Recent.Tax.Assessment
## 1          N       1984          6534                     290000
## 2          Y       1980         10018                     631000
## 3          N       1981          5662                     207000
## 4          N       1920          8276                     240000
## 5          N       1970         12632                     282000
## 6          N       1987          9583                     256000
```

```
tail(moundsales)
```

```
##      ID           Address Beds Baths SQFT Month.Sold Year.Sold Sale.Price
## 145 145 5736 Lynwood Blvd    3     2 1939        Apr      2020     289900
## 146 146   4714 Hanover Rd    3     2 1630        Apr      2020     228000
## 147 147    4959 Leslie Rd    3     3 2729        Apr      2020     330000
## 148 148    5447 Breezy Rd    3     2 2047        Apr      2020     329900
## 149 149     2740 Grove Ln    3     2 1426        Apr      2020     222000
## 150 150   3201 Charles Ln    3     4 2475        Apr      2020    1100000
##     Lake.Front Year.Built Lot.Size.Sqft Most.Recent.Tax.Assessment
## 145          N       1910         16553                     273000
## 146          N       1972          6534                     234000
## 147          N       1965         17860                     334000
## 148          N       1946         10018                     312000
## 149          N       1984         10055                     222000
## 150          Y       1988         13068                    1001000
```

```
summary(moundsales$Sale.Price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  118400  287625  337000  444991  451309 1670000
```

```
summary(moundsales$Most.Recent.Tax.Assessment)
```

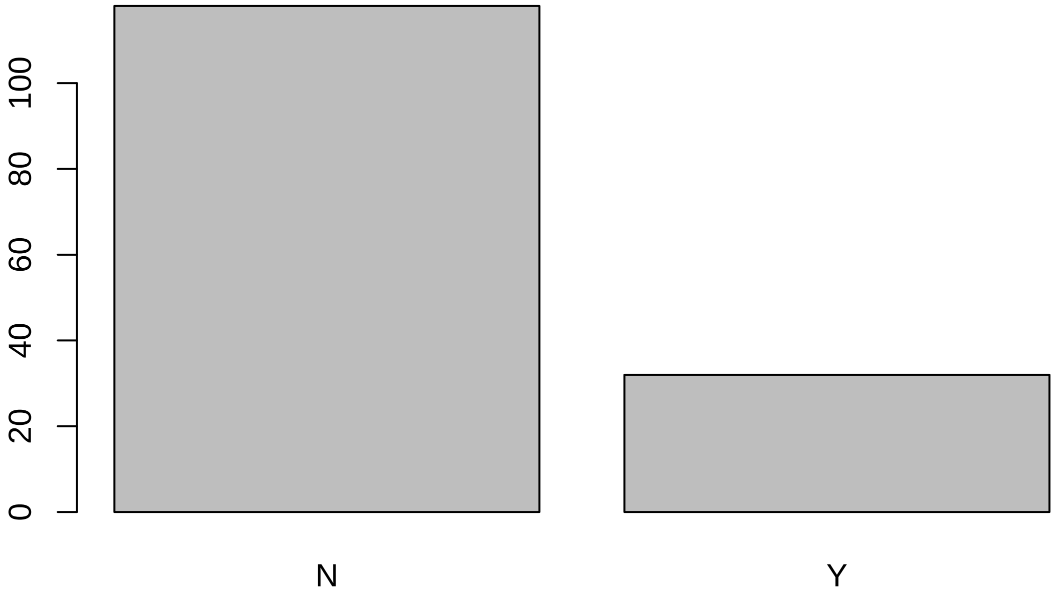
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22000  241250  278500  374900  399750 1344000
```

```
table(moundsales$Lake.Front)
```

```
##
##   N   Y
## 118  32
```

```
milsale <- subset(moundsales, moundsales$Sale.Price > 999999)
View(milsale)
```

```
lakefront <- table(moundsales$Lake.Front)
barplot(lakefront)
```



### ‘nonlf’ Data Set

```
nonlf <- subset(moundsales, moundsales$Lake.Front == 'N')
```

```
summary(nonlf$Sale.Price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  118400  275000  316500  324801  358750  582500
```

```
summary(nonlf$Most.Recent.Tax.Assessment)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   22000  237250  259000  273229  307750  464000
```

```
summary(nonlf$SQFT)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1117    1542    1820    1876    2154    3566
```

```
summary(nonlf$Lot.Size.Sqft)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    3484    6534    9583   10029   11761   23522
```

```r
summary(nonlf$Year.Built)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1910    1954    1974    1969    1987    2020
```

```r
under300k <- subset(nonlf, nonlf$Sale.Price < 300000)
dim(under300k)
```

```
## [1] 46 12
```