Housing Prices

Justin Valentin, Holden Baranowski, Emmanuel Obeng

2023-12-04

library(readxl)  
library(SKTools)  
library(tidyverse)

library(psych)

library(ggplot2)  
library(dplyr)  
library(stats)  
library(psych)  
library(dplyr)  
  
Housing\_Prices <- read\_xlsx("SOCI510\_Group7\_Housing Prices.xlsx")

# Square Feet by Number of Bedrooms

Housing\_Prices2 <-  
 Housing\_Prices %>%   
 mutate(RoomCount = if\_else(`Bedrooms` < 4, TRUE, FALSE))

# Descriptives

mean(Housing\_Prices2$SquareFeet, na.rm = T)

## [1] 2006.375

median(Housing\_Prices2$SquareFeet, na.rm = T)

## [1] 2007

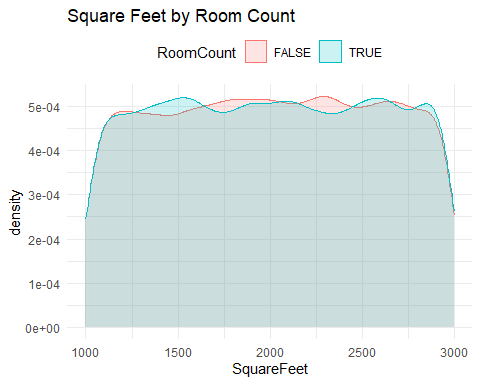
max(Housing\_Prices2$SquareFeet, na.rm = T) - min(Housing\_Prices2$SquareFeet, na.rm = T)

## [1] 1999

sd(Housing\_Prices2$SquareFeet, na.rm = T)

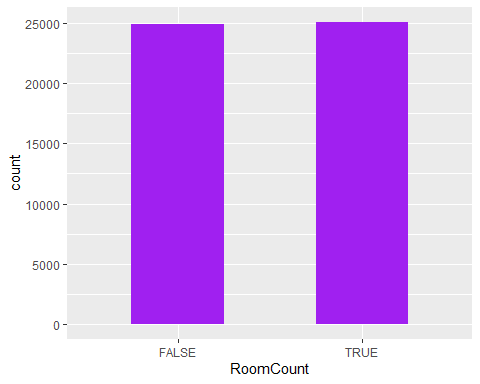
## [1] 575.5132

Housing\_Prices2 %>%  
 ggplot(aes(x = SquareFeet, color = RoomCount, fill = RoomCount)) +  
 geom\_density(alpha = .2) +  
 theme\_minimal() +  
 labs(title = "Square Feet by Room Count") +  
 theme(legend.position = "top")



# Frequencies of RoomCount

Housing\_Prices2 %>%   
 ggplot(., aes(x = RoomCount)) +  
 geom\_bar(stat = "count", width = .5, fill = "purple")



Housing\_Prices2 %>%   
 select(RoomCount) %>%   
 frequencies()

## # A tibble: 2 × 3  
## var value n  
## <chr> <chr> <int>  
## 1 RoomCount FALSE 24903  
## 2 RoomCount TRUE 25097

# Analysis (Independent T-Test)

We will conduct an Independent Samples T-test because there are exactly two distinct groups to compare for this analysis. Houses with less than 4 bedrooms (TRUE) and houses with 4 or more bedrooms (FALSE).

**Step 1: State hypotheses**

H0: μTRUE = μFALSE H1: μTRUE ≠ μFALSE

**Step 2: Determine alpha level**

Alpha = .05 The null hypothesis will be rejected if the p-value is less than .05 which will risk a Type 1 error 5% of the time.

**Step 3: Run test statistic**

t = 0.176 p = 0.860

**Step 4: Make a decision**

Because the p-value is not less than .05, we fail to reject the null hypothesis. If the null hypothesis is true, the probability of these data or data more extreme is .85. There is no sufficient evidence to homes with 4 or more bedrooms have significantly different square footage than homes with less than 4 bedrooms

t.test(Housing\_Prices2$SquareFeet ~ Housing\_Prices2$RoomCount)

##   
## Welch Two Sample t-test  
##   
## data: Housing\_Prices2$SquareFeet by Housing\_Prices2$RoomCount  
## t = 0.17605, df = 49998, p-value = 0.8603  
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0  
## 95 percent confidence interval:  
## -9.18299 10.99539  
## sample estimates:  
## mean in group FALSE mean in group TRUE   
## 2006.830 2005.923

# Price x SquareFeet

# Descriptives

Housing\_Prices2$Price <- as.numeric(Housing\_Prices2$Price)  
mean(Housing\_Prices2$Price, na.rm = T)

## [1] 224827.3

median(Housing\_Prices2$Price, na.rm = T)

## [1] 225052.1

max(Housing\_Prices2$Price, na.rm = T)

## [1] 492195.3

min(Housing\_Prices2$Price, na.rm = T)

## [1] -36588.17

max(Housing\_Prices2$Price, na.rm = T) - min(Housing\_Prices2$Price, na.rm = T)

## [1] 528783.4

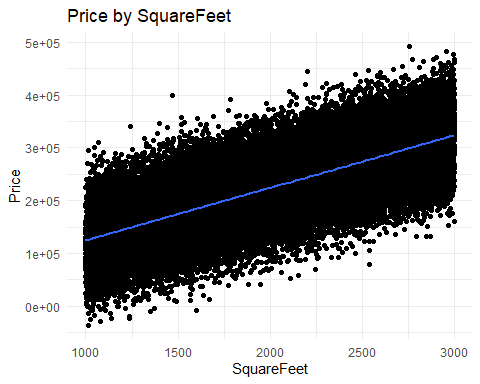
sd(Housing\_Prices2$Price, na.rm = T)

## [1] 76141.84

# Evaluate the Scatter Plot

Per the data, data points showing price based on square feet are roughly condensed around the linear line. Despite a few potential outliers, there is a strong positive or direct linear correlation between the price and Square feet of houses. As square feet increases, price increases.

Housing\_Prices2 %>%   
 select(SquareFeet, Price) %>%   
 ggplot(., aes(x = SquareFeet, y = Price)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", formula = y ~ x) +  
 theme\_minimal() +  
 labs(title = "Price by SquareFeet")



# Analysis (Pearson’s Correlation)

We will run a Pearson Correlation (r) Test. The objective of this test is to determine whether there is a relationship between Price and Squarefeet of Houses.

**Step 1: State the null hypothesis**

Ho: p = 0 H1: p ≠ 0 where p is the population correlation

**Step 2: Establish the alpha level**

Alpha = 0.05 Reject the null hypothesis if p-value < 0.05 which risks a type 1 error (rejecting the null Alpha hypothesis when we should have failed to reject the null hypothesis) 5% of the time

**Step 3: Run the test statistic and find the p-value**

r= 0.75 t = 254.10 p = < 0.01

**Step 4: Make a decision**

Because the p-value is < 0.05, we reject the null hypothesis that the population correlation is non-significant. If we assume the null hypothesis is true (⍴ = 0), the probability of the observed data (or data even less likely) is 0.01. While the correlation in the sample is strong (0.75), there is sufficient evidence that Price and Squarefeet of houses are linearly related positively in the population. As SquareFeet increases, Price also increase on average.

**Coefficient of Determination:**

r = 0.75 ; R = 0.56 Thus 56% of price increase is explained by an increase in square feet.

Housing\_Prices2$Price <- as.numeric(Housing\_Prices2$Price)  
Housing\_Prices2$SquareFeet <- as.numeric(Housing\_Prices2$SquareFeet)  
  
Housing\_Prices2 %>%   
select(Price, SquareFeet) %>%  
 corr.test() %>%  
 corr\_summary() %>%  
 filter(dv == "SquareFeet")

## # A tibble: 1 × 9  
## iv dv r n t p p.adjust p\_sig p.adjust\_sig  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <lgl> <lgl>   
## 1 Price SquareFeet 0.751 50000 254. 0 0 TRUE TRUE

# SquareFeet x Neighborhood

## Descriptives

Housing\_Prices2 %>%  
 select(Neighborhood) %>%   
 frequencies()

## # A tibble: 3 × 3  
## var value n  
## <chr> <chr> <int>  
## 1 Neighborhood Rural 16676  
## 2 Neighborhood Suburb 16721  
## 3 Neighborhood Urban 16603

# Analysis (ANOVA)

We wanted to see how SquareFeet differs from the three types of Neighborhoods. We will be conducting an ANOVA test for this analysis. There are three different types of neighborhoods we will be comparing different square feet sizes.

**Step 1: State the hypothesis**

Ho: µ(Rural) = µ(Suburb) = µ(Urban) H1: NOT µ(Rural) = µ(Suburb) = µ(Urban)

**Step 2: Establish an alpha**

Alpha = .05 We will reject the null hypothesis if the p-value is less than (<) .05 which risks a type 1 error (rejecting the null hypothesis when we should have failed to reject the null hypothesis) 5% of the time.

**Step 3: Run the Test**

F = 4.6032 p-vale = 0.01002.

**Step 4: We make a decision about the null hypothesis**

Because the p-value is < .05, we reject the null hypothesis that there is a difference in the mean number of losses based on player tier. If we assume the null hypothesis is true, the probability of the observed data is greater than .05.

Housing\_Prices2 <-   
 Housing\_Prices %>%   
 mutate(Neighborhood = as.character(Neighborhood))  
  
Housing\_Prices2 %>%   
 select(Neighborhood) %>%   
 frequencies(perc = T)

## # A tibble: 3 × 5  
## var value n perc valid.perc  
## <chr> <chr> <int> <dbl> <dbl>  
## 1 Neighborhood Rural 16676 0.334 0.334  
## 2 Neighborhood Suburb 16721 0.334 0.334  
## 3 Neighborhood Urban 16603 0.332 0.332

anova(aov(SquareFeet ~ Neighborhood, Housing\_Prices2))

## Analysis of Variance Table  
##   
## Response: SquareFeet  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Neighborhood 2 3.0489e+06 1524441 4.6032 0.01002 \*  
## Residuals 49997 1.6557e+10 331168   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Post-Hoc Test

Two of the Neighborhood types have significantly different numbers compared to the SquareFeet of the houses in their communities. Based on the post-hoc test, Urban has statistically significantly different (higher) SquareFeet from Rural. Urban also has statistically significantly different (higher) SquareFeet when compared to Suburbs.

TukeyHSD(aov(SquareFeet ~ Neighborhood, Housing\_Prices2))

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = SquareFeet ~ Neighborhood, data = Housing\_Prices2)  
##   
## $Neighborhood  
## diff lwr upr p adj  
## Suburb-Rural -1.470005 -16.230564 13.29055 0.9704136  
## Urban-Rural 15.795780 1.009053 30.58251 0.0329177  
## Urban-Suburb 17.265785 2.488988 32.04258 0.0169863