MA 376 Lesson 2 - Guided Data Exploration (AY23-2)

Getting Started

There are two things we do every time we start working in R Studio:

1. Turn on the Tidyverse Package. Think of it like opening your toolbox so you can use the tools inside of it

```
library(tidyverse)
```

2. Tell R Studio where to find the files you want to use and reads in the house data so we can begin exploring it.

```
setwd("C:/Users/james.starling/OneDrive - West Point/Documents/GitHub/MA256/LSN1_2/")
data <- read_csv("L02_Redfin-Nashville.csv")

# OR reach out to a website where the data is hosted:
data <- read_csv("https://raw.githubusercontent.com/jkstarling/MA256/main/LSN1_2/L02_Redfin-Nashville.cs
```

Calculate Monthly House Payments

Let's create a new variable that estimates our total monthly house payment including mortgage, taxes, insurance, and HOA fees for every home so we can think about what we can afford. To do this, we'll need to make some assumptions:

1. We'll put 20% down so 0.8*PRICE = Loan Amount for our mortgage PMT calculations
2. $5k in closing costs for a 3% fixed APR on a 30 year primary residence
3. We'll use TN average annual effective property tax rate of 0.64% (.0064) of PRICE
4. We'll use TN average annual homeowner's insurance cost of $1,233

Simplify the data

Based on these assumptions, can you identify some limitations we can address during `Step 6 of the Statistical Investigative Process` i.e. `Look Back and Ahead?` For Example, what if the property tax rate or insurance cost in Nashville is higher? Could we find and use the actual tax rate and insurance cost for Nashville homes?

Let's simplify our dataset to only include variables we care about i.e. `address`, `city`, `price`, `home`, `lot size`, `#BD`, `#BA`, and `year built`.

You must save the new dataset somewhere to use it, just like the `read_csv` function.

```
house_data <- data %>% select(ADDRESS, CITY, PRICE,
                              BEDS, BATHS, `SQUARE FEET`,
                              `LOT SIZE`, `YEAR BUILT`)
```

Missing data

The HOA/MONTH column in `house_data` has missing (NA) values. Any time a formula includes a variable with a NA value, it results in an NA value. In this case, NA means there are no HOA fees so to fix this, we need to replace all NA values with $0 so we can use HOA/MONTH in mathematical formulas for total monthly cost. We do this by creating adataframe so we can use the `is.na()` function to replace all NA values with 0:

```
df = data.frame(data$`HOA/MONTH`) #Creates dataframe using HOA/MONTH variable in "data"
df[is.na(df)] = 0 #The is.na() function replaces all "NA" with a numerical 0
```

Updating column names

If you open the dataframe `df`, the column name is not very appealing and it will be difficult to refer to it in formulas. Let's rename it using the `colnames()` function to make our life easier:

```
colnames(df)[1]="HOA.MONTH" #We renamed the 1st column in our df dataframe to "HOA.MONTH"
```

We can now use the `mutate()` command to create a new variable in house_data that is equal to the df dataframe variable HOA.MONTH values. NOTE: If we don't save the dataset with the new variable somewhere, we won't actually be able to use it later, so we overwrote the original dataset with the new one that includes the new variable:

```
house_data = house_data %>% mutate(df)
```

Notice that the column name is copied over exactly as it originally appeared. If the dataframe had multiple columns, they would have all been copied over

Creating new columns

Create new columns that calculate the cost to buy, monthly payments, and rent.

```
MPR = 0.03/12 #Monthly Percentage Rate is the assumed Annual 3% Rate divided by 12 months
house_data = house_data %>% #Again, overwrite original data so it includes new variables
  mutate(CostToBuy=0.2*PRICE+5000, #20% down + estimated $5k in closing costs
         Monthly_PMT=`HOA.MONTH`+(1233/12)+0.0064*PRICE/12+
           ((0.8*PRICE*(MPR*((1+MPR)^360)))/(((1+MPR)^360)-1)),#HOA/mo+Ins/mo+Taxes/mo+Mortgage PMT/mo
         Rent=(BEDS-1)*650) #BEDS-1 because you'll live in one of the rooms
```

Again, overwrite original data so it includes the new variables (cash flow and ROI):

```
house_data = house_data %>%
  mutate(Cash_Flow = Rent - Monthly_PMT + 1431, #BAH is all profit if rent pays for everything
         Annual_ROI = 12*Cash_Flow/CostToBuy)
#Annual return on investment: Annual profit / Investment cost
```

Since the annual ROI calculation doesn't account for utilities, annual maintenance, vacancy, & turnover expenses, let's only consider homes with an annual ROI $> 20\%$ (0.2)

```
ROI_house_data = house_data %>% filter(Annual_ROI > 0.2)
```

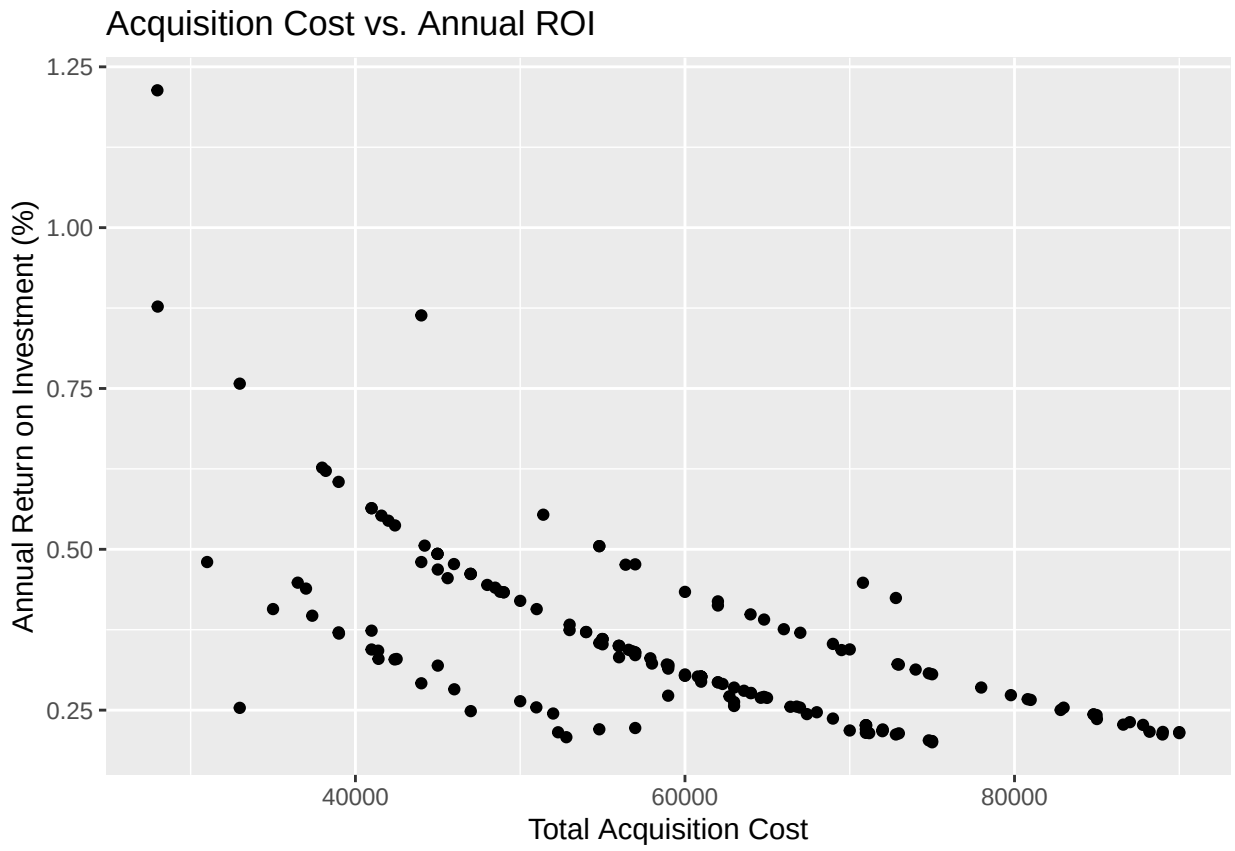
Changing notation NOTE: If your Annual ROI is in scientific notation, you may convert it using the following code to make the output easier to interpret:

```
options(scipen=0)#scipen=0 turns on scientific notation. scipen=999 turns it off
```

Plotting ROI

One would think that the more money I risk, the greater my compensation/ROI should be. Let's compare CostToBuy with Annual ROI to see if that's true. Since both are quantitative, a scatter plot is ideal:

```
ROI_house_data%>%
  ggplot(aes(x = CostToBuy,y=Annual_ROI))+geom_point()+
  labs(x="Total Acquisition Cost",y="Annual Return on Investment (%)",
       title = "Acquisition Cost vs. Annual ROI")
```



How do you interpret this plot? It seems that cheaper homes have a higher ROI, but why do we think they are cheaper? In stats, our primary goal is to try to understand the variability in data. Let's do that now:

(a brief side story...) In all investments, one can apply a concept called risk/reward ratio. Essentially, the more likely you are to lose your investment, the greater the reward/return that investment demands to convince people to invest in it. That's why startup company stock prices are so cheap and have high potential

profit (because most businesses fail in the first few years), and why relatively safe stocks like Walmart, Coca Cola, & Disney are relatively expensive but don't have much/if any potential for growth in stock price.

In our case, the high ROI homes could be in high crime areas, could be run down and/or be in a horrible school district. Perhaps they sit vacant for extended periods of time because no one wants to live there, and then once you do find someone, they break their lease early because they realize how bad it is to live there. They could be farther from hotspots like work, colleges, downtown, malls, restaurants, hospitals, etc.

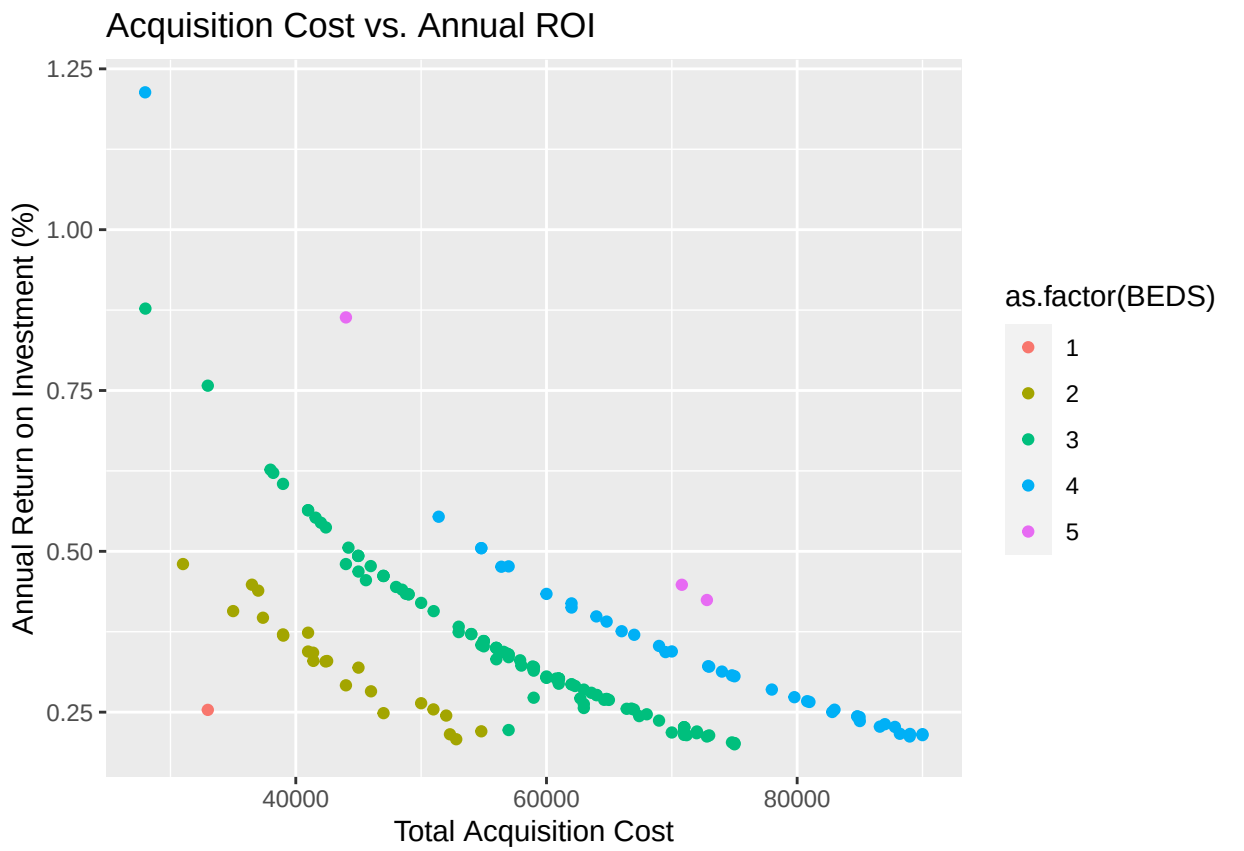
Perhaps those properties demand a higher ROI because the type of folks who would typically live there are more likely to not pay rent, damage your home, need to be evicted, etc. Even though the home is cheaper to acquire, you may be more likely to lose everything, so sellers can't charge much to purchase the home which makes ROI seem awesome.

With that said, we don't know until we put eyes on. It could just be old/outdated, and falling apart in need of a big expensive renovation. In real life, we would look at crime maps, school districts, drive around the neighborhood, see the house in person, etc. before completely writing off a property with such a phenomenal annual ROI. This due diligence is what helps us understand if a home is a high-risk investment or a unicorn opportunity.

A better, more informative, plot.

Let's dive into this data more deeply by adding a 3rd dimension (color) which reflects the number of bedrooms:

```
ROI_house_data %>%
  ggplot(aes(x = CostToBuy,y=Annual_ROI, color=as.factor(BEDS))) +
  geom_point() +
  labs(x="Total Acquisition Cost",y="Annual Return on Investment (%)",
       title = "Acquisition Cost vs. Annual ROI")
```




```
  # geom_smooth(method = "lm")
```

We see an almost perfectly linear trend within categories based on the number of bedrooms, perhaps because the maximum rental income is capped based on the number of bedrooms so the more expensive a 3BR home is, the less ROI.

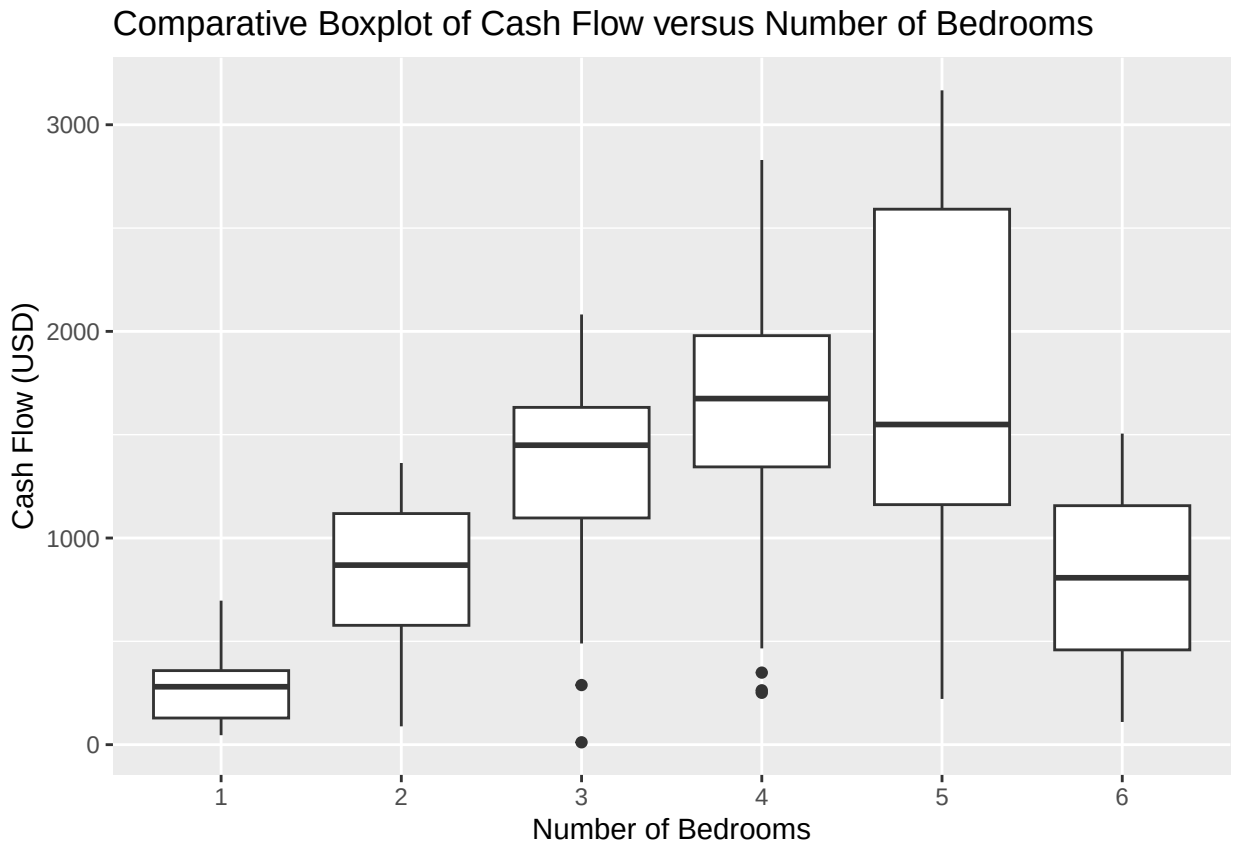
Note: to overlay the plot with the best fitting least squares regression lines, but we must add + after labs(...) and remove the #hashtag on the next line.

An investment strategy (using cashflow)

My investment strategy is to buy the home with the largest # of BRs and the least acquisition cost without sacrificing quality. There are other concerns as well, #like proximity to universities, size of bedrooms, number of bathrooms, etc. but the main consideration is potential income vs expense.

This is great analysis, but ROI is not everything. Perhaps we care more about how much cash is going into our pocket each month rather than %ROI. After all, they say cash is king, and accruing cash faster will allow me to buy more homes sooner. Let's compare the number of bedrooms to the amount of Cashflow for all homes with a positive ROI.

```
CF_house_data=house_data %>%
  filter(Annual_ROI>0)
CF_house_data %>%
  ggplot(aes(x = as.factor(BEDS), y = Cash_Flow)) + geom_boxplot() +
  labs(title  = "Comparative Boxplot of Cash Flow versus Number of Bedrooms",
       x = "Number of Bedrooms", y="Cash Flow (USD)")
```

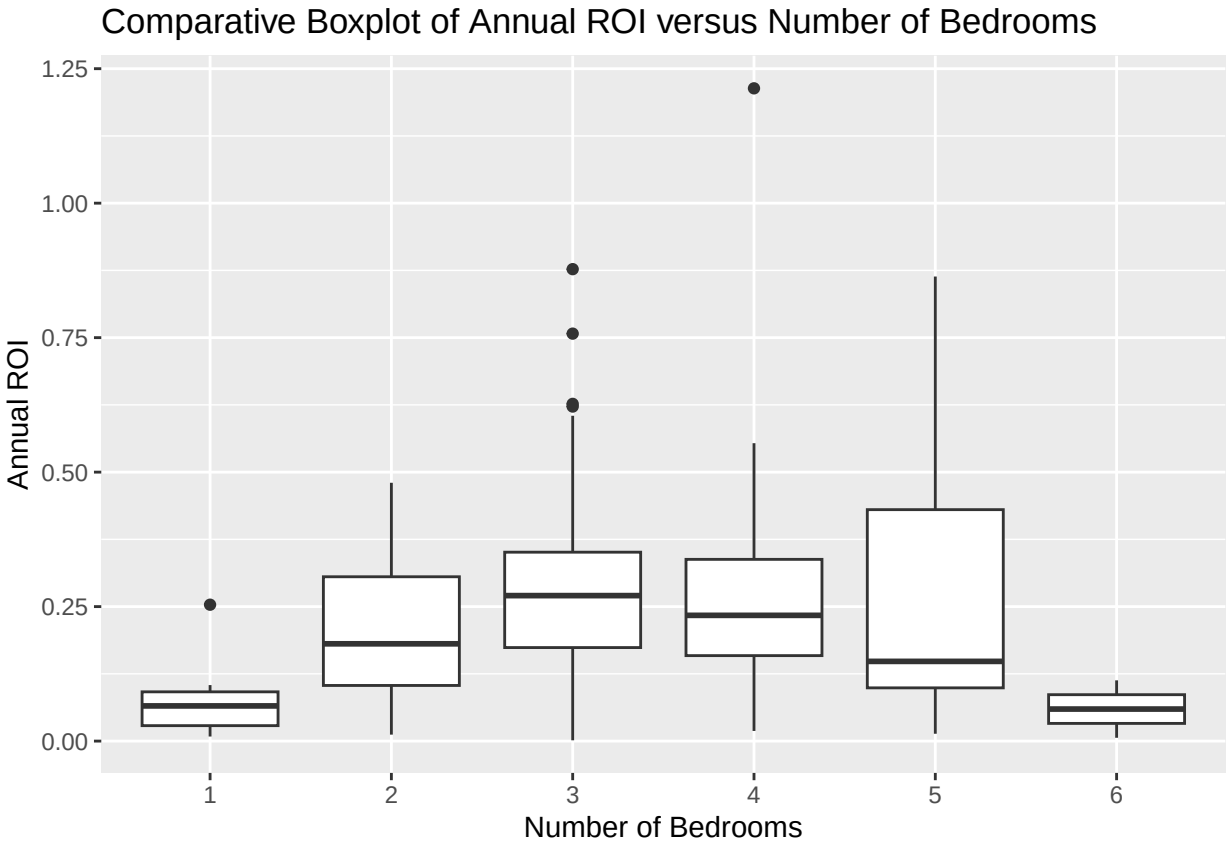


When comparing the 1 & 2 BR boxplots, the median cash flow almost triples when you spend a little extra for a 2BR. When comparing the 2 & 3 BR boxplots, your median cash flow almost doubles when you spend a little extra for a 3BR. However spending extra for a 4BR does not seem to do much in regard to median cash flow (maybe $200 increase). Spending extra for a 5BR home seems to reduce the median cash flow, but it is not uncommon to cash flow upwards of $2500/mo which is significantly better than what 4BR homes are offering. 6BR homes don't seem worth it at all. Bigger homes should cost more when considering materials cost to build them, but the 6BR home has cash flows comparable to that of a 2BR home.

A different investment strategy (using ROI)

The jump from 3 to 4 BR and potential from 4 to 5 BR homes is most interesting/significant so let's now compare ROI vs #BR's using CF data in boxplot form:

```
CF_house_data %>%
  ggplot(aes(x = as.factor(BEDS), y = Annual_ROI)) + geom_boxplot()+
  labs(title  = "Comparative Boxplot of Annual ROI versus Number of Bedrooms",
       x = "Number of Bedrooms", y="Annual ROI")
```



Notice your median and 75% quartile annual ROI are worse for a 4BR compared to a 3BR which means the additional cash flow and ROI from a 4BR home does not justify its cost. You should purchase a 3BR home unless you find a phenomenal deal on a 4BR or 5BR.

A personal story - Additional BR ("I" = a former mathlete instructor)

In April 2019, I bought a 3BR/4BA home valued at $244k and invested $44k over the next month to update the home and convert 2 large rooms into 2 more bedrooms by adding closets. 2 months after completing the renovations, the home appraised for $345k and rented for $3k/mo, cash flowing $1700/mo profit. Thanks to 2.5 years' appreciation, today it's worth $475k.

Could we apply this same strategy by buying a 2, 2.5, or 3 BA home, and paying for an additional BR to be built?

What would be the approximate value gain assuming the bedroom is 10 ft x 15 ft? We can figure this out by comparing the average price of 2BR 1350-1550 sqft homes, 3BR 1500-1700 sqft homes, and 4BR 1650-1850 sqft homes to estimate the value gain of a bedroom addition project.

```
Addition_data=house_data%>%
  filter(BEDS==2 & `SQUARE FEET`>=1350 & `SQUARE FEET` <= 1550 |
           BEDS==3 & `SQUARE FEET`>=1500 & `SQUARE FEET` <= 1700 |
           BEDS==4 & `SQUARE FEET`>1650 & `SQUARE FEET`<1850)
```

We use & (ampersand) to filter on multiple conditions simultaneously (e.g. "AND"). We use | (vertical line above the enter key) to filter "OR" conditions

Now let's compare the average price for each BR/BA combination within the SQFT ranges depending on size to determine the value of a 1 BR addition: NOTE that the order of the group_by function will simplify or complicate analysis so when you want to analyze the benefit of a 1x bedroom addition, having BEDS 2nd in the group_by() function makes it easier to compare.

```
Addition_data%>%
  group_by(BATHS, BEDS)%>% #calculates mean(PRICE) for every combination of BEDS & BATHS
  summarise(Average_Price=mean(PRICE))
```

```
## # A tibble: 11 x 3
## # Groups:   BATHS [5]
##    BATHS  BEDS Average_Price
##    <dbl> <dbl>         <dbl>
##  1   1       2       285000
##  2   1       3       169900
##  3   2       2       428000
##  4   2       3       278938
##  5   2       4       306963.
##  6   2.5     2       213250
##  7   2.5     3       308888.
##  8   2.5     4       294900
##  9   3       3       356750
## 10   3       4       428000
## 11   4       4       547000
```

Based on this information, there are only 3x scenarios that could generate profit: - Converting a 2BA home from 3BR to 4BR has a potential average gain of $28,025. - Converting a 2.5BA home from 2BR to 3BR has a potential average gain of $95,638. - Converting a 3BA home from 3BR to 4BR has a potential average gain of $71,250.

When I had a bathroom addition completed, it cost me $15,000 so all 3 of these scenarios seem profitable, but if I could only pick one, the $95,638 seems best.

Another personal story - Additional BA (where "I" is the same former mathlete instructor)

I once owned a 4BR/2.5BA home where the bathroom situation was garbage. There was an owner's suite on the 3rd floor with it's own bathroom, but 2 other BR's on the 3rd floor and 1 more bedroom in the basement all shared the other full BA that was in the hall on the 3rd floor. I spent $15k to install a 3rd full bathroom in the basement to make the basement suite like an additional owner's suite. My property has gained $80k in value almost immediately and allowed me to profit an extra $600/mo in rent.

Could we apply this same strategy by buying a 2, 3, or 4 BR home, and paying for an additional full bathroom to be built? Using the same code as before, but having BATHS 2nd in the group_by command makes it very easy to analyze:

```
Addition_data%>%
  group_by(BEDS, BATHS)%>% #calculates mean(PRICE) for every combination of BEDS & BATHS
  summarise(Average_Price=mean(PRICE))
```

```
## # A tibble: 11 x 3
## # Groups:   BEDS [3]
##     BEDS BATHS Average_Price
##    <dbl> <dbl>         <dbl>
```

```
##  1     2   1         285000
##  2     2   2         428000
##  3     2   2.5       213250
##  4     3   1         169900
##  5     3   2         278938
##  6     3   2.5       308888.
##  7     3   3         356750
##  8     4   2         306963.
##  9     4   2.5       294900
## 10     4   3         428000
## 11     4   4         547000
```

Based on this information, there are only 5x scenarios that could generate profit: 1. Converting a 2BR home from 1BA to 2BA has a potential average gain of $143,000. 2. Converting a 3BR home from 1BA to 3BA has a potential average gain of $109,038. 3. Converting a 3BR home from 2BA to 3BA has a potential average gain of $77,812. 4. Converting a 4BR home from 2BA to 3BA has a potential average gain of $121,037. 5. Converting a 4BR home from 3BA to 4BA has a potential average gain of $119,000.

You can conduct a similar analysis to consider the addition of 1BD AND 1BA: - Converting a 3BD/1BA to a 4BD/2BA has a potential average gain of $137,063 - Converting a 3BD/2BA to a 4BD/3BA has a potential average gain of $149,062 - Converting a 3BD/3BA to a 4BD/4BA has a potential average gain of $190,250

Considering everything at once, since 3BR homes had the best ROI and CF for the price, AND have the most average profit potential given 1BD/1BAaddition, buying a 3/3 and upgrading it to a 4/4 seems to be the optimal investment strategy in the short term and the long term. If I'm willing to learn and do some of the work myself I could make almost $200k profit while cash flowing, $1500/mo in rent.

Yet another personal story - lot/home size

Does lot size or home size influence the price more? We'll do a linear model including these multiple quantitative variables, and will compare their slopes to determine which has a larger affect. The R code for linear models comes from the MA256 course guide:

```
Price_Model = house_data %>% #Finds the best fitting least squares regression line
  lm(PRICE ~ `SQUARE FEET` + `LOT SIZE`, data = ., na.action = na.exclude)
summary(Price_Model) #Gives the equation for the line & allows us to compare variable coefficients for
```

```
##
## Call:
## lm(formula = PRICE ~ `SQUARE FEET` + `LOT SIZE`, data = ., na.action = na.exclude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -461422 -117616   -9040   93192  905075
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.623e+05  2.683e+04  -6.049  6.1e-09 ***
## `SQUARE FEET`  2.636e+02  1.055e+01  24.985  < 2e-16 ***
## `LOT SIZE`     8.251e-01  2.570e-01   3.211  0.00152 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 197900 on 223 degrees of freedom
```

```
##   (124 observations deleted due to missingness)
## Multiple R-squared:  0.7876, Adjusted R-squared:  0.7857
## F-statistic: 413.4 on 2 and 223 DF,  p-value: < 2.2e-16
```

Since SQFT has a coefficient of $263.63, that's like saying 1 sqft change in size results in a $263.63 change in price. Compared to lot size, which only has an $0.83 cents change in price per 1 sqft change in lot size, home size is more influential. With that said, there is some error in our logic since home size also includes some elements of lot size which means we might be double counting.