Final Project

Evan Fullwood

2024-12-05

Task 1

First, I am going to read in data about home prices, remove some data, make a simple sales price column, and produce a small snapshot of the data.

```
library(dplyr)
homedata <- readr::read_csv("https://www4.stat.ncsu.edu/online/ST308/Data/ezfullwo_house.csv")
mhomedata <- homedata%>%

  #Remove obs w. Low variable using filter() and ! for FALSE
  filter(LandContour != "Low" )%>%

  #Remove obs before 2007 using filter()
  filter(YrSold > 2006)%>%

  #Make SimpleSp w mutate(), SalesPrice divided by 100000
  mutate(SimpleSP = SalePrice/100000)%>%

  #Remove variables using select(-c())
  select(-c(WoodDeckSF, Condition1))

#create table with kable of first 10 rows and first 6 columns
knitr::kable(mhomedata[1:10, 1:6])
```

SalePrice	FullBath	MoSold	TotRmsAbvGrd	YrSold	OpenPorchSF
208500	2	2	8	2008	61
181500	2	5	6	2007	0
143000	1	10	5	2009	30
307000	2	8	7	2007	57
200000	2	11	7	2009	204
129500	1	2	5	2008	0
144000	1	9	4	2008	0
157000	1	5	5	2008	213
132000	1	7	5	2007	112
149000	1	3	5	2010	0

I am now going to generate some summary statistics about sales price, month sold, and full bath amount. I will include a table that groups this data based on whether or not there is central air.

```r
#Produce summary statistsics about the SalePrice, FullBath, and MoSold variables
homedstats <- mhomedata%>%

  #group by central air first
  group_by(CentralAir)%>%

  summarise(
    #sales price summary
    Spmean = mean(SalePrice),
    Spstdev = sd(SalePrice),
    Spq1 = quantile(SalePrice, .25),
    Spq3 = quantile(SalePrice, .75),

     #FullBath Summary
    Fbmean = mean(FullBath),
    Fbsdev = sd(FullBath),
    Fbq1 = quantile(FullBath, .25),
    Fbq3 = quantile(FullBath, .75),

    #MoSold Summary
    Msmean = mean(MoSold),
    Mstdev = sd(MoSold),
    Msq1 = quantile(MoSold, .25),
    Msq3 = quantile(MoSold, .75),
  )

knitr::kable(homedstats)
```

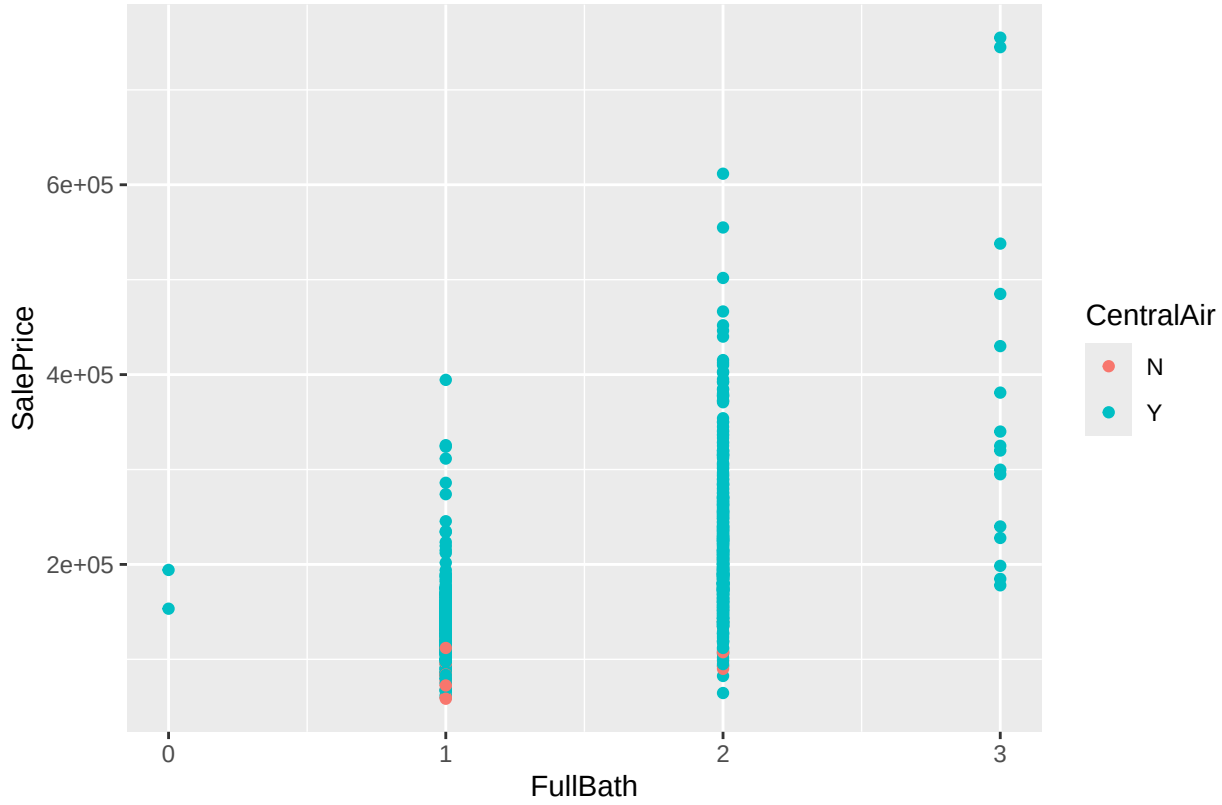
CentralAir	Spmean	Spstdev	Spq1	Spq3	Fbmean	Fbsdev	Fbq1	Fbq3	Msmean	Mstdev	Msq1	Msq3
N	103412.5	31208.32	82750	122750	1.28125	0.4568034	1	2	6.343750	3.096401	4.75	8.25
Y	187875.9	81400.05	135250	216500	1.58728	0.5398992	1	2	6.335589	2.748928	5.00	8.00

Below is a scatter plot that shows the price based on the amount of bathrooms. Orange dots show homes with no central air.

```r
library(ggplot2)

#create the ggplot arguments and save as object
SPplot <- ggplot(data=mhomedata, aes(x=FullBath, y=SalePrice, color = CentralAir))

#add the scatter plot with geom_point() and the caption with labs(caption =)
SPplot + geom_point() + labs(caption = "No central air is uncommon, and is most often found in homes wi
```



No central air is uncommon, and is most often found in homes with below average sales prices.

Do the variables FullBath and MoSold have an impact on SalePrice?

```
#Run multiple linear regression on specified data
LRmodel <- lm(data = mhomedata, SalePrice ~ FullBath + MoSold)

#summary of LRmodel
summary(LRmodel)
```

```
##
## Call:
## lm(formula = SalePrice ~ FullBath + MoSold, data = mhomedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -155867  -39737   -6900   21435  446637
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51194.2     8997.0   5.690 1.81e-08 ***
## FullBath     85826.4     4516.1  19.004  < 2e-16 ***
## MoSold        -310.0      882.8  -0.351    0.726
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67410 on 768 degrees of freedom
## Multiple R-squared:  0.3208, Adjusted R-squared:  0.3191
```

```
## F-statistic: 181.4 on 2 and 768 DF,  p-value: < 2.2e-16
```

```
#predict the SalePrice using list() with experimental variables
predict(LRmodel, list(FullBath = 2, MoSold = 8))
```

```
##        1
## 220366.8
```

```
predict(LRmodel, list(FullBath = 6, MoSold = 1))
```

```
##        1
## 565842.3
```

After looking at the linear regression, it is likely that FullBath is important to price, but not MoSold, as the p-value is high.

Task 2

Write a function that finds the mean median and IQR of all of the data to analyze averages of all video game scores.

```
#Read in video game data
vgs <- readr::read_csv("https://corgis-edu.github.io/corgis/datasets/csv/video_games/video_games.csv")
```

```
## Rows: 1212 Columns: 36
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): Title, Metadata.Genres, Metadata.Publishers, Release.Console, Rele...
## dbl (25): Features.Max Players, Metrics.Review Score, Metrics.Sales, Metrics...
## lgl  (6): Features.Handheld?, Features.Multiplatform?, Features.Online?, Met...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#use apply(), select data and column, set up function
apply(select(vgs, 'Metrics.Review Score'), MARGIN = 2, FUN = function(x){
  #generate mean, median, IQR
  c(mean=mean(x), median= median(x), IQR=IQR(x))
})
```

```
##        Metrics.Review Score
## mean               68.82838
## median             70.00000
## IQR                19.00000
```

Which video game publishers receive the highest ratings, lowest? What are the mean, median, and IQR of the three highest and three lowest publishers?

```r
#Split the data into dataframes based on publisher
publishersplt <- split(vgs, vgs$Metadata.Publishers)

#lapply must be used since the data is formatted as a list
pbscrdata <- lapply(publishersplt, FUN = function(x){

  #extract scores from data and save as metrics
  metrics <- x$'Metrics.Review Score'
  #Combine and compute desired metrics
 c(mean = mean(metrics), median = median(metrics), IQR = IQR(metrics))
})

pbscrdata
```

```
## $`2K`
##     mean   median      IQR
## 73.14286 75.00000 12.00000
##
## $Activision
##     mean   median      IQR
## 67.83908 70.00000 19.00000
##
## $`Activision,Konami`
##   mean median    IQR
##     89     89      0
##
## $`Activision,Sony`
##   mean median    IQR
##     68     68      0
##
## $Atari
##   mean median    IQR
##  60.00  57.00  21.25
##
## $`Atari,Namco`
##   mean median    IQR
##   68.5   68.5    5.5
##
## $Capcom
##     mean   median      IQR
## 73.91429 74.00000 11.50000
##
## $`Capcom,Nintendo`
##   mean median    IQR
##     75     75      0
##
## $`Capcom,Rockstar`
##   mean median    IQR
##     88     88      0
##
## $Disney
##   mean median    IQR
##  67.50  67.50  15.25
```

```
##
## $EA
##     mean   median      IQR
## 70.17576 71.00000 14.00000
##
## $`EA,Namco`
##   mean median    IQR
##   67.5   67.5    6.5
##
## $`EA,Sony`
##   mean median    IQR
##     58     58      0
##
## $Eidos
##     mean   median      IQR
## 64.68182 67.50000 17.25000
##
## $Konami
##     mean   median      IQR
## 66.08511 69.00000 21.00000
##
## $Microsoft
##     mean   median      IQR
## 75.19048 79.00000 24.00000
##
## $`Microsoft,SquareEnix`
##   mean median    IQR
##     64     64      0
##
## $Midway
##     mean   median      IQR
## 59.43478 62.00000 24.50000
##
## $Namco
##     mean   median      IQR
## 66.78788 67.00000 18.00000
##
## $`Namco,Sony`
##   mean median    IQR
##     83     83      5
##
## $`Namco,Ubisoft`
##   mean median    IQR
##     85     85      0
##
## $Nintendo
##     mean   median      IQR
## 75.72941 77.00000 15.00000
##
## $`Nintendo,Sega`
##   mean median    IQR
##   68.5   68.5    1.5
##
## $`Nintendo,SquareEnix`
```

```
##   mean median    IQR
##     65     65      0
##
## $Rockstar
##     mean   median      IQR
## 78.83333 81.00000 11.00000
##
## $Sega
##   mean median    IQR
##  65.32  67.00  19.50
##
## $Sony
##   mean median    IQR
##   75.8   79.0   15.0
##
## $`Sony,Ubisoft`
##   mean median    IQR
##     71     71      6
##
## $SquareEnix
##     mean   median      IQR
## 75.80645 75.00000 13.00000
##
## $THQ
##     mean   median      IQR
## 66.85246 68.00000 15.00000
##
## $Ubisoft
##     mean   median      IQR
## 66.82609 64.50000 17.25000
```

```
#I'm attempting to turn the list back into a data frame using enframe(),
#just so I can sort the data based on mean, and find the highest score.
 dataframestats <- tibble::enframe(pbscrdata, value='mean')%>%
   #pbscrdata has 3 values: mean, median, and IQR. unnest_wider will generate a column for each.
   tidyr::unnest_wider(col = 'mean')%>%
   #arrange descending based on mean
   arrange(desc(mean))

dataframestats
```

```
## # A tibble: 31 x 4
##    name               mean median   IQR
##    <chr>             <dbl>  <dbl> <dbl>
##  1 Activision,Konami  89       89     0
##  2 Capcom,Rockstar    88       88     0
##  3 Namco,Ubisoft      85       85     0
##  4 Namco,Sony         83       83     5
##  5 Rockstar           78.8     81    11
##  6 SquareEnix         75.8     75    13
##  7 Sony               75.8     79    15
##  8 Nintendo           75.7     77    15
##  9 Microsoft          75.2     79    24
## 10 Capcom,Nintendo    75       75     0
```

```
## # i 21 more rows
```

It appears that Konami has the highest critic score, although it seems like they only have 1 game, along with Capcom Rockstar and Namco Ubisoft.

```
BADstats <- tibble::enframe(pbscrdata, value='mean')%>%
   tidyr::unnest_wider(col = 'mean')%>%
   #arrange ascending
   arrange(mean)

BADstats
```

```
## # A tibble: 31 x 4
##    name                  mean median   IQR
##    <chr>                <dbl>  <dbl> <dbl>
##  1 EA,Sony               58     58     0
##  2 Midway                59.4   62    24.5
##  3 Atari                 60     57    21.2
##  4 Microsoft,SquareEnix  64     64     0
##  5 Eidos                 64.7   67.5  17.2
##  6 Nintendo,SquareEnix   65     65     0
##  7 Sega                  65.3   67    19.5
##  8 Konami                66.1   69    21
##  9 Namco                 66.8   67    18
## 10 Ubisoft               66.8   64.5  17.2
## # i 21 more rows
```

EA(Sony) and Midway are the worst.