Rolling House Sales Analysis - Manhattan, NY
Team: Sam Coyne / Ariel Francisco / Rob Rivens
SMU Data Science / Section 402

Assignment:

Using the Rolling Data Sales website, examine the Manhattan, NY, housing sales data set, obtained from - http://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page

Goal: Create an RStudio project for the analysis of this data set.

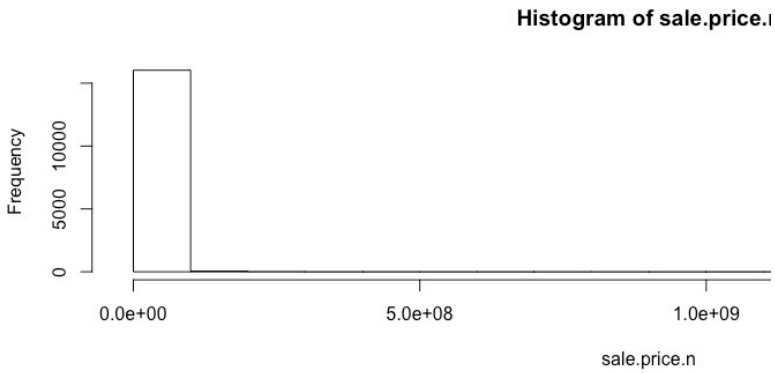
Our README file is posted in the project root directory; explains the premise of our assignment

Our Data Directory contains files we used to load in and clean up the data:

- Includes the original Excel file
- Includes the .csv file we referenced in RStudio
- Data Cleasing Techniques:
  - Used function (T/F) to count 0's True = $0 sale price
    - `manhattan$SALE.PRICE.N <- as.numeric(gsub("[^[:digit:]]","", manhattan$SALE.PRICE))`
    - `count(is.na(manhattan$SALE.PRICE.N))`
  - Changed headers to lower case
    - `names(manhattan) <- tolower(names(manhattan))`
  - Removed leading digits
    - `manhattan$gross.sqft <- as.numeric(gsub("[^[:digit:]]","", manhattan$gross.square.feet))`
    - `manhattan$land.sqft <- as.numeric(gsub("[^[:digit:]]","", manhattan$land.square.feet))`
    - Changed this from character to numeric
      - `manhattan$year.built <- as.numeric(as.character(manhattan$year.built)`

)

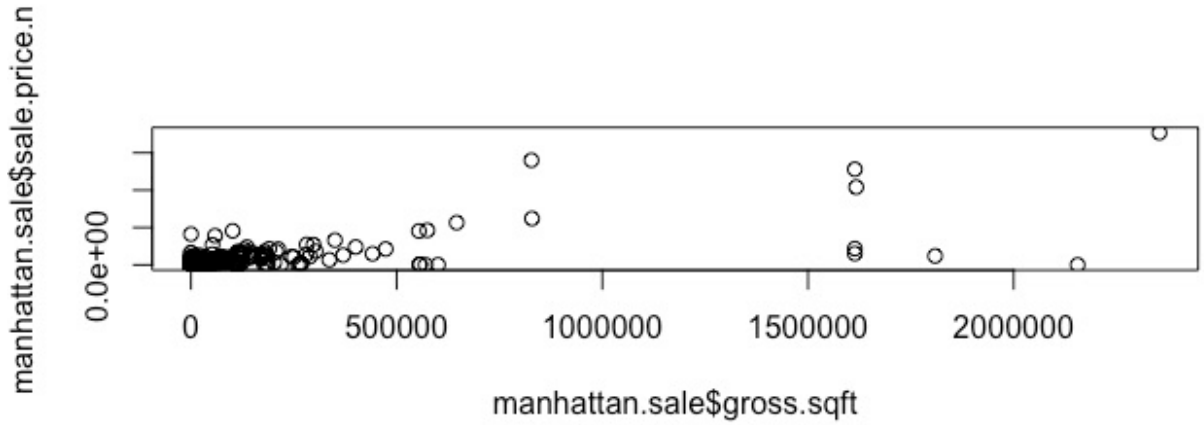
- Performed more exploratory data analysis with histogram
  - attach(manhattan)
  - hist(sale.price.n)
  - detach(manhattan)



- Kept only actual sales
  - manhattan.sale <- manhattan[manhattan$sale.price.n!=0,]
  - plot sales >0 on raw data and on log scale-

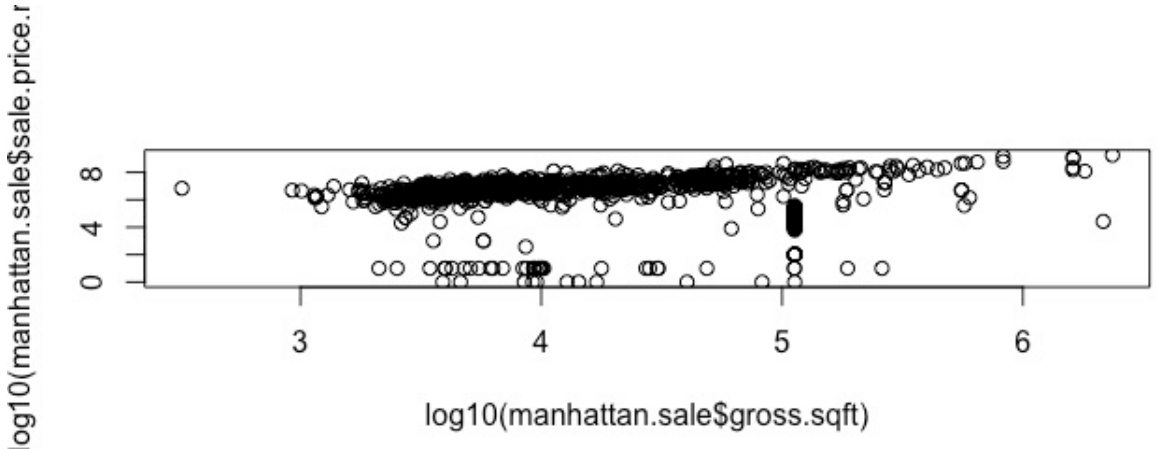
-

  plot(manhattan.sale$gross.sqft,manhattan.sale$sale.price.n)



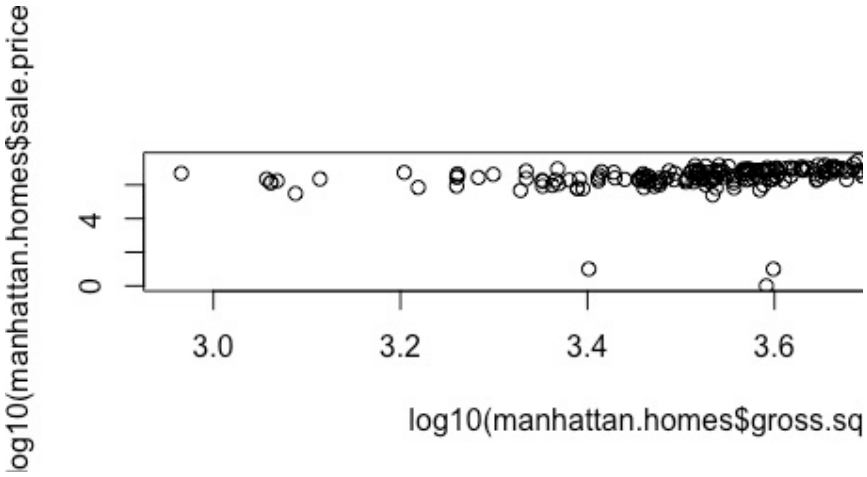
-

plot(log10(manhattan.sale$gross.sqft),log10(manhattan.sale$sale.price.n))

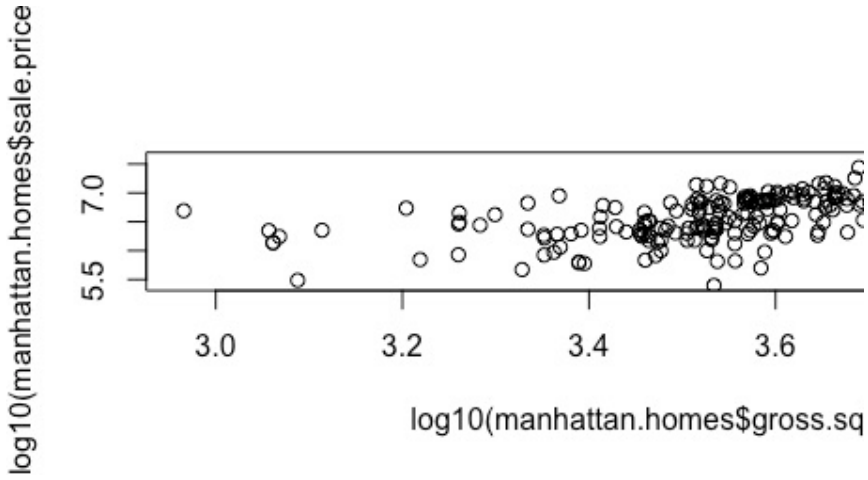


- Focused on 1-, 2-, and 3-family homes #grepl returns TRUE if string contains "FAMILY"
- manhattan.homes <- manhattan.sale[which(grepl("FAMILY",manhattan.sale$building.class.category)),]
- plot "family homes"

- plot(log10(manhattan.homes$gross.sqft),log10(manhattan.homes$sale.price.n))

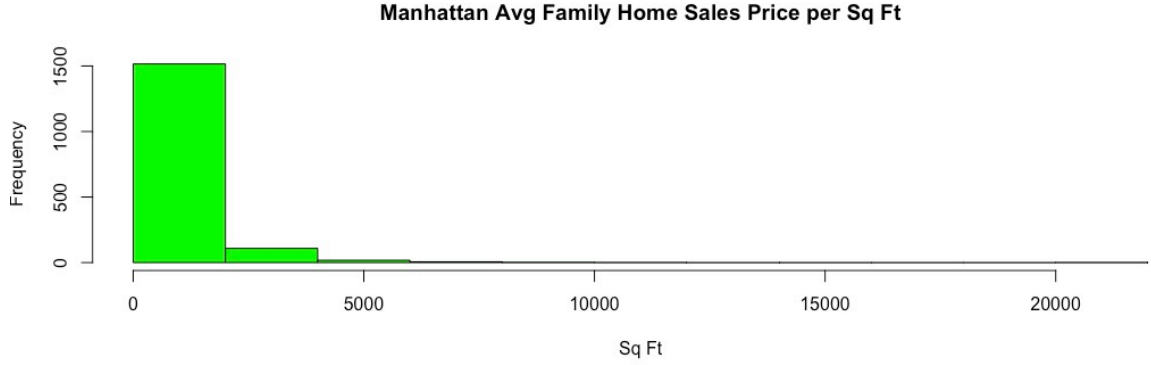


- #summary of "family homes"
  - summary(manhattan.homes[which(manhattan.homes$sale.price.n<100000),])
- #remove outliers that seem like they weren't actual sales
  - manhattan.homes$outliers <-(log10(manhattan.homes$sale.price.n) <=5)
- manhattan.homes <-manhattan.homes[which(manhattan.homes$outliers==0),]
  - plot(log10(manhattan.homes$gross.sqft),log10(manhattan.homes$sale.price.n))



-

- SUMMARY
  - The data doesn't reveal much on the natural scale, but the log10 scale reveals a linear trend for gross sq. ft vs. sales price
  - The correlation is positive with a few deviations from the expected pattern, but a positive correlation seems to exist.
  - HISTOGRAM OF SALE PRICE PER SQ FACETIME:
    - x <- manhattan$SALE.PRICE/ manhattan$GROSS.SQUARE.FEET
    - hist(x,main="Manhattan Avg Family Home Sales Price per Sq Ft",xlab = "Sq Ft")



-Analysis directory contains files for exploratory data analysis on the clean data