# Title: “NYC Real Estate - Dataset Cleaning”

## Author: “Eanna Lohan”

### Date: “27 October 2020”

# Objective:

# Taking four Real Estate Sales datasets from the New York City Department of Finance Rolling Sales Department, plot the data, combine the dataframes & clean up the column names into a single dataframe ready for further analysis.

# The first step is to load the readxl library to allow us read the excel files and create individual dataframes for each of the four borroughs

library(readxl) # load package to read excel file  
brooklyn <- read\_excel("rollingsales\_brooklyn.xls", skip = 4) # save dataframe of brooklyn borough sales beginning from row 5 of excel file  
  
  
brooklyn <- read\_excel("rollingsales\_brooklyn.xls", skip = 4)  
bronx <- read\_excel("rollingsales\_bronx.xls", skip = 4)  
staten\_island <- read\_excel("rollingsales\_statenisland.xls", skip = 4)  
queens <- read\_excel("rollingsales\_queens.xls", skip = 4) #load 4 dataframes

# Next, we load the tidyverse library and take a glimpse at the brooklyn data

library(tidyverse)# load package

## -- Attaching packages ----------------------------------------------------------------------------------------------------------- tidyverse 1.3.0 --

## <U+2713> ggplot2 3.2.1 <U+2713> purrr 0.3.3  
## <U+2713> tibble 2.1.3 <U+2713> dplyr 0.8.3  
## <U+2713> tidyr 1.0.0 <U+2713> stringr 1.4.0  
## <U+2713> readr 1.3.1 <U+2713> forcats 0.4.0

## -- Conflicts -------------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

glimpse(brooklyn)# provide glimpse of dataframe

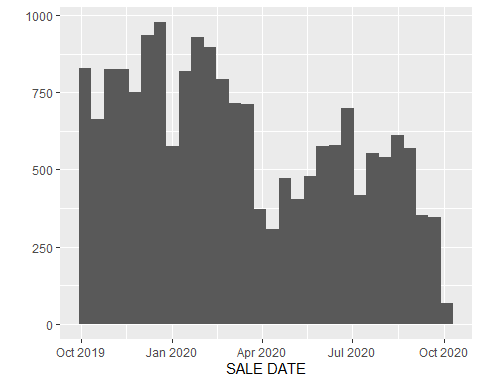
## Observations: 18,598  
## Variables: 21  
## $ BOROUGH <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, …  
## $ NEIGHBORHOOD <chr> "BATH BEACH", "BATH BEACH", "BATH BE…  
## $ `BUILDING CLASS CATEGORY` <chr> "01 ONE FAMILY DWELLINGS", "01 ONE F…  
## $ `TAX CLASS AT PRESENT` <chr> "1", "1", "1", "1", "1", "1", "1", "…  
## $ BLOCK <dbl> 6359, 6360, 6367, 6371, 6371, 6392, …  
## $ LOT <dbl> 70, 48, 24, 19, 60, 65, 15, 29, 56, …  
## $ `EASE-MENT` <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, …  
## $ `BUILDING CLASS AT PRESENT` <chr> "S1", "A5", "A9", "A9", "A9", "A2", …  
## $ ADDRESS <chr> "8684 15TH AVENUE", "14 BAY 10TH STR…  
## $ `APARTMENT NUMBER` <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, …  
## $ `ZIP CODE` <dbl> 11228, 11228, 11214, 11214, 11214, 1…  
## $ `RESIDENTIAL UNITS` <dbl> 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, …  
## $ `COMMERCIAL UNITS` <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, …  
## $ `TOTAL UNITS` <dbl> 2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 2, …  
## $ `LAND SQUARE FEET` <dbl> 1933, 2513, 1571, 2320, 2417, 3222, …  
## $ `GROSS SQUARE FEET` <dbl> 4080, 1428, 1456, 1566, 2106, 1804, …  
## $ `YEAR BUILT` <dbl> 1930, 1930, 1935, 1930, 1930, 1940, …  
## $ `TAX CLASS AT TIME OF SALE` <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …  
## $ `BUILDING CLASS AT TIME OF SALE` <chr> "S1", "A5", "A9", "A9", "A9", "A2", …  
## $ `SALE PRICE` <dbl> 1300000, 849000, 830000, 0, 1188000,…  
## $ `SALE DATE` <dttm> 2020-04-28, 2020-03-18, 2020-06-26,…

mean(brooklyn$`SALE PRICE`) #show mean sales price

## [1] 1059455

# As SALE DATE is stored in a format that represents calendar dates and times, we are able to use a single line of code to display a histogram of property sales by date. The qplot() function used to make the histogram is from the ggplot2 package, which is a core tidyverse package.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 The histogram shows there was a reduced volume of sales in April 2020 which could be as a result of the COVID 19 pandemic lockdown measures

# Bind all dataframes into one using the bind\_rows function, and save the dataframe as “NYC\_property\_sales”. The dataset containing 21 columns similar to the four individual datasets is a positive sign as it indicates they each have the exact same columns.

NYC\_property\_sales <-bind\_rows(mutate\_all(brooklyn, as.character), mutate\_all(bronx, as.character), mutate\_all(staten\_island, as.character), mutate\_all(queens, as.character))  
  
glimpse(NYC\_property\_sales) #glimpse at combined dataframe

## Observations: 51,714  
## Variables: 21  
## $ BOROUGH <chr> "3", "3", "3", "3", "3", "3", "3", "…  
## $ NEIGHBORHOOD <chr> "BATH BEACH", "BATH BEACH", "BATH BE…  
## $ `BUILDING CLASS CATEGORY` <chr> "01 ONE FAMILY DWELLINGS", "01 ONE F…  
## $ `TAX CLASS AT PRESENT` <chr> "1", "1", "1", "1", "1", "1", "1", "…  
## $ BLOCK <chr> "6359", "6360", "6367", "6371", "637…  
## $ LOT <chr> "70", "48", "24", "19", "60", "65", …  
## $ `EASE-MENT` <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, …  
## $ `BUILDING CLASS AT PRESENT` <chr> "S1", "A5", "A9", "A9", "A9", "A2", …  
## $ ADDRESS <chr> "8684 15TH AVENUE", "14 BAY 10TH STR…  
## $ `APARTMENT NUMBER` <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, …  
## $ `ZIP CODE` <chr> "11228", "11228", "11214", "11214", …  
## $ `RESIDENTIAL UNITS` <chr> "1", "1", "1", "1", "1", "1", "2", "…  
## $ `COMMERCIAL UNITS` <chr> "1", "0", "0", "0", "0", "0", "0", "…  
## $ `TOTAL UNITS` <chr> "2", "1", "1", "1", "1", "1", "2", "…  
## $ `LAND SQUARE FEET` <chr> "1933", "2513", "1571", "2320", "241…  
## $ `GROSS SQUARE FEET` <chr> "4080", "1428", "1456", "1566", "210…  
## $ `YEAR BUILT` <chr> "1930", "1930", "1935", "1930", "193…  
## $ `TAX CLASS AT TIME OF SALE` <chr> "1", "1", "1", "1", "1", "1", "1", "…  
## $ `BUILDING CLASS AT TIME OF SALE` <chr> "S1", "A5", "A9", "A9", "A9", "A2", …  
## $ `SALE PRICE` <chr> "1300000", "849000", "830000", "0", …  
## $ `SALE DATE` <chr> "2020-04-28", "2020-03-18", "2020-06…

# Using the magrittr library, we can tidy up the column names of NYC dataframe, by updating all column names to replace all spaces with underscores, and then update all column names to lower case. This is done using the assignment pipe operator and str\_replace\_all() function.

library(magrittr) # load package

##   
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':  
##   
## set\_names

## The following object is masked from 'package:tidyr':  
##   
## extract

colnames(NYC\_property\_sales) %<>% str\_replace\_all("\\s", "\_") %<>% tolower() # take column names of NYC dataframe, and then update all column names to replace all spaces with underscores, and then update all column names to lower case  
colnames(NYC\_property\_sales)

## [1] "borough" "neighborhood"   
## [3] "building\_class\_category" "tax\_class\_at\_present"   
## [5] "block" "lot"   
## [7] "ease-ment" "building\_class\_at\_present"   
## [9] "address" "apartment\_number"   
## [11] "zip\_code" "residential\_units"   
## [13] "commercial\_units" "total\_units"   
## [15] "land\_square\_feet" "gross\_square\_feet"   
## [17] "year\_built" "tax\_class\_at\_time\_of\_sale"   
## [19] "building\_class\_at\_time\_of\_sale" "sale\_price"   
## [21] "sale\_date"

NYC\_property\_sales %>% glimpse() # Take dataframe, and then, glimpse

## Observations: 51,714  
## Variables: 21  
## $ borough <chr> "3", "3", "3", "3", "3", "3", "3", "3"…  
## $ neighborhood <chr> "BATH BEACH", "BATH BEACH", "BATH BEAC…  
## $ building\_class\_category <chr> "01 ONE FAMILY DWELLINGS", "01 ONE FAM…  
## $ tax\_class\_at\_present <chr> "1", "1", "1", "1", "1", "1", "1", "1"…  
## $ block <chr> "6359", "6360", "6367", "6371", "6371"…  
## $ lot <chr> "70", "48", "24", "19", "60", "65", "1…  
## $ `ease-ment` <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ building\_class\_at\_present <chr> "S1", "A5", "A9", "A9", "A9", "A2", "B…  
## $ address <chr> "8684 15TH AVENUE", "14 BAY 10TH STREE…  
## $ apartment\_number <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ zip\_code <chr> "11228", "11228", "11214", "11214", "1…  
## $ residential\_units <chr> "1", "1", "1", "1", "1", "1", "2", "1"…  
## $ commercial\_units <chr> "1", "0", "0", "0", "0", "0", "0", "0"…  
## $ total\_units <chr> "2", "1", "1", "1", "1", "1", "2", "1"…  
## $ land\_square\_feet <chr> "1933", "2513", "1571", "2320", "2417"…  
## $ gross\_square\_feet <chr> "4080", "1428", "1456", "1566", "2106"…  
## $ year\_built <chr> "1930", "1930", "1935", "1930", "1930"…  
## $ tax\_class\_at\_time\_of\_sale <chr> "1", "1", "1", "1", "1", "1", "1", "1"…  
## $ building\_class\_at\_time\_of\_sale <chr> "S1", "A5", "A9", "A9", "A9", "A2", "A…  
## $ sale\_price <chr> "1300000", "849000", "830000", "0", "1…  
## $ sale\_date <chr> "2020-04-28", "2020-03-18", "2020-06-2…

colnames(NYC\_property\_sales)

## [1] "borough" "neighborhood"   
## [3] "building\_class\_category" "tax\_class\_at\_present"   
## [5] "block" "lot"   
## [7] "ease-ment" "building\_class\_at\_present"   
## [9] "address" "apartment\_number"   
## [11] "zip\_code" "residential\_units"   
## [13] "commercial\_units" "total\_units"   
## [15] "land\_square\_feet" "gross\_square\_feet"   
## [17] "year\_built" "tax\_class\_at\_time\_of\_sale"   
## [19] "building\_class\_at\_time\_of\_sale" "sale\_price"   
## [21] "sale\_date"

# END OF REPORT