

STAT GR5206 Homework 1 [100 pts]

Due 8:00pm Monday, October 2nd on Canvas

Your homework should be submitted on Canvas using RMarkdown. Please submit both a knitted .pdf file and a raw .Rmd file. (If you are having trouble knitting to .pdf come to office hours and we'll try to sort it out, but for the homework, knit to .html and then convert to .pdf before handing it in). We will not (and cannot) accept any other formats. Please clearly label the questions in your responses and support your answers by textual explanations and the code you use to produce the result. Note that you cannot answer the questions by observing the data in the "Environment" section of RStudio or in Excel – you must use coded commands.

Goals: data cleaning, EDA, R graphics, practice with filtering and vectorized commands.

The data set `properties.csv` contains property-level data on privately-owned, subsidized rental properties in New York City collected by the Furman Center. The data can be found here: <http://coredata.nyc/>. The dataset contains financial and physical information on the properties including geographic, subsidy, ownership, physical, and financial information.

Perform the following tasks:

Part 1: Loading and Cleaning the Data in R

- i. Load the data into a dataframe called `housing`.
- ii. How many rows and columns does the dataframe have?
- iii. Run this command, and explain, in words, what this does:

```
apply(is.na(housing), 2, sum).
```
- iv. Remove the rows of the dataset for which the variable `assessed_value` equals 0.
- v. How many rows did you remove with the previous call?
- vi. Create a new variable in the dataset called `logValue` that is equal to the logarithm of the property's `assessed_value`. What are the minimum, median, mean, and maximum values of `logValue`?
- vii. Create a new variable in the dataset called `logUnits` that is equal to the logarithm of the number of units in the property. The number of units in each piece of property is stored in the variable `res_units`.
- viii. Finally create a new variable in the dataset called `after2000` which equals `TRUE` if the property was built in or after 2000 and `FALSE` otherwise. You'll want to use the

`year.built` variable here. This can be done in a single line of code.

Part 2: EDA

The column `boro_name` contains the Borough of each property and is one of either **Bronx**, **Manhattan**, **Staten Island**, **Brooklyn**, or **Queens**.

- i. Plot property `logValue` against property `logUnits`. Name the x and y labels of the plot appropriately. `logValue` should be on the y-axis.
- ii. Make the same plot as above, but now include the argument `col = factor(housing$after2000)`. Describe this plot and the covariation between the two variables. What does the coloring in the plot tell us?

Hint: `legend("bottomright", legend = levels(factor(housing$after2000)), fill = unique(factor(housing$after2000)))`.

- iii. The `cor()` function calculates the correlation coefficient between two variables. What is the correlation between property `logValue` and property `logUnits` in (i) the whole data, (ii) just Manhattan (iii) just Brooklyn (iv) for properties built after 2000 (v) for properties built before 2000? You will need to add the argument `use = "pairwise.complete.obs"` to handle NA values.
- iv. Make two plots showing property `logValue` against property `logUnits` for Manhattan and Brooklyn. (If you can fit the information into one plot, clearly distinguishing the two boroughs, that's OK too.)
- v. Consider the following block of code. Give a single line of R code which gives the same final answer as the block of code. There are a few ways to do this.

```
manhat.props <- c()

for (props in 1:nrow(housing)) {
  if (housing$boro_name[props] == "Manhattan") {
    manhat.props <- c(manhat.props, props)
  }
}

med.value <- c()
for (props in manhat.props) {
  med.value <- c(med.value, housing$assessed_value[props])
}
```

```
med.value <- median(med.value, na.rm = TRUE)
```

- vi. Make side-by-side box plots comparing property `logValue` across the five boroughs.
- vii. For five boroughs, what are the median property values? (Use `assessed_value` here, not `logValue`.)