FA-582 – Assignment 1

Problem 1

Use the datasets provided for Bronx, Brooklyn, Manhattan, Queens, and Staten Island. Do the following:

- Load in and clean the data.
- Conduct exploratory data analysis in order to find out where there are outliers or missing values, decide how you will treat them, make sure the dates are formatted correctly, make sure values you think are numerical are being treated as such, etc.
- Conduct exploratory data analysis to visualize and make comparisons for residential building category classes across boroughs and across time (select the following: 1-, 2-, and 3-family homes, coops, and condos). Use histograms, boxplots, scatterplots or other visual graphs. Provide summary statistics along with your conclusions.

Problem 2

The datasets provided nyt1.csv, nyt2.csv, and nyt3.csv represents three (simulated) days of ads shown and clicks recorded on the New York Times homepage. Each row represents a single user. There are 5 columns: age, gender (0=female, 1=male), number impressions, number clicks, and logged-in. Use R to handle this data. Perform some exploratory data analysis:

- Create a new variable, age_group, that categorizes users as “<20”, “20-29”, “30-39”, “40-49”, “50-59”, “60-69”, and “70+”.
- For each day:
    - Plot the distribution of number of impressions and click-through-rate (CTR = #clicks / #impressions) for these age categories
    - Define a new variable to segment or categorize users based on their click behavior.
    - Explore the data and make visual and quantitative comparisons across user segments/demographics (<20-year-old males versus <20-year-old females or logged-in versus not, for example).
- Extend your analysis across days. Visualize some metrics and distributions over time.