STAT GU4206/GR5206 Homework 1 [100 pts]
Due 11:59pm Monday, September 23th on Canvas

Your homework should be submitted on Canvas as an R Markdown file. Please submit both the .Rmd and .pdf files (or .html). 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, more practice with filtering and vectorized commands.

Data Description:
The data set `NYChousing.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://www.furmancenter.org`. The dataset contains financial and physical information on the properties including geographic, subsidy, ownership, physical, and financial information. Note that this dataset is a few years old.

Perform the following tasks:

Part 1: Loading, Cleaning the Exploring Data in `R`

i. Load the data into a dataframe called `housing`.

ii. How many rows and columns does the dataframe have?

iii. Run the appropriate function to display the variable names of the dataframe.

iv. Run this command, and explain, in words, what this does:

```
apply(is.na(housing), 2, sum).
```

v. Remove the rows of the dataset for which the variable `Value` is NA.

vi. How many rows did you remove with the previous call? Does this agree with your result from (iv)?

vii. Calculate the third quartile of the property values, i.e., the third quartile Q_3 is the 75th percentile. Use the `quantile()` function to complete this task.

viii. Create a new variable in the dataset called `HighValue` that is equal to "High" if the property's value is greater than Q_3 and is equal to "NotHigh" if the property's value is less than or equal to Q_3.

ix. Display a contingency table that shows the proprtions of `HighValue` split by `Borough`. Note that the `table()` function is the easiest way to tackle this problem but the `table()` function gives raw counts.

x. What is the proportion of properties whose values are in the upper quartile **and** are located in The Bronx? Solve this question in two ways: (1) by using the table from (ix), and (2) by using logical/relational commands and using the function `mean()`.

xi. **Given** a randomly selected property is in The Bronx, what is the probability that its value is in the upper quartile? Solve this question in two ways: (1) by using the table from (ix), and (2) by using logical/relational/filtering commands and using the function `mean()`.

xii. Create a new variable in the dataset called `logValue` that is equal to the logarithm of the property's `Value`. What are the minimum, median, mean, and maximum values of `logValue`?

xiii. 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 `UnitCount`.

xiv. Finally create a new variable in the dataset called `after1950` which equals `TRUE` if the property was built in or after 1950 and `FALSE` otherwise. You'll want to use the `YearBuilt` variable here. This can be done in a single line of code.

Part 2: EDA

2i. Create a multiple boxplot (side-by-side boxplots) comparing property value across the five boroughs. Create a multiple boxplot (side-by-side boxplots) comparing property `logValue` across the five boroughs. Make sure to label the plots appropriately.

2ii. Plot property `logValue` against property `logUnits`. Name the x and y labels of the plot appropriately. `logValue` should be on the y-axis.

2iii. Make the same plot as above, but now include the argument `col = factor(housing$after1950)`. 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$after1950)), fill = unique(factor(housing$after1950)))`.

2iv. 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 1950 (v) for properties built before 1950?

2v. Make a single plot showing property `logValue` against property `logUnits` for Manhattan and Brooklyn. When creating this plot, clearly distinguish the two boroughs.

2vi. 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$Borough[props] == "Manhattan") {
    manhat.props <- c(manhat.props, props)
  }
}

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

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

2vii. For five boroughs, what are the median property values? (Use `Value` here, not `logValue`.)