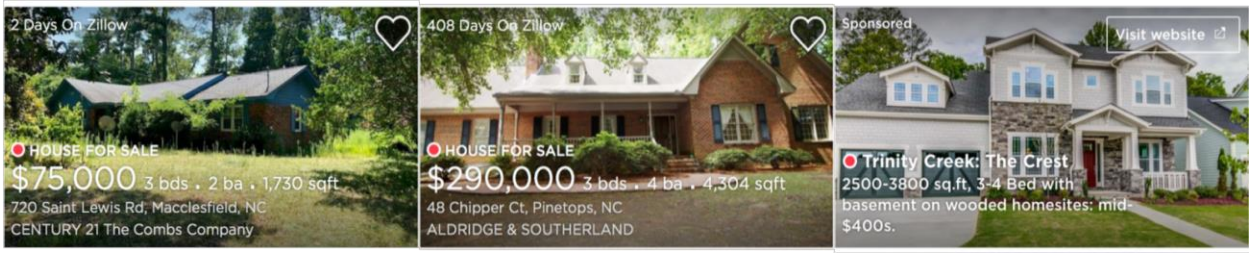
IT137IU: Data Analysis
Lab#3/Assignment#3: Data Wrangling (continue)

Introduction



In this lab you will learn more data wrangling functions. You will be continue working with data on Housing Data - Zillow Research that you downloaded in Lab 2 namely ***Zip_zhvi_4bd.csv*** (Four bedrooms at zip level), and download the file ***Metro_zhvi_4bd.csv*** (Four bedrooms at metro level) that uploaded in Blackboard.

This lab guide follows closely and supplements the material presented in Chapters 3, 7-9, and 14 in the textbook R for Data Science (RDS).

Data Wrangling

Joining tables

Our goal is to merge together the datasets ***Zip_zhvi_4bd.csv*** and ***Metro_zhvi_4bd.csv***. Remember from Lecture 4 that the unique RegionID for a combining data. We have this ID as the single variable RegionID in ***Zip_zhvi_4bd.csv*** and RegionID in ***Metro_zhvi_4bd.csv***.

Exercise 1: [20pts] Let's make the two datasets to become tidy data.

Exercise 2: [20pts] Before merge two datasets together let's find and check if there are the common RegionID or StateName in the two datasets.

Exercise 3: [20pts] Let's merge the two datasets together, which matches pairs of observations whenever their keys or IDs are equal. We match on the variables *RegionID* and *RegionID* and save the merged data set into a new object called *data_metro_zip*

We want to merge ***Metro_zhvi_4bd.csv*** into ***Zip_zhvi_4bd.csv***. The argument by tells R which variables to match rows on, in this case RegionID in ***Metro_zhvi_4bd.csv*** and RegionID in ***Zip_zhvi_4bd.csv***. The number of columns in *data_metro_zip* equals the number of columns in ***Zip_zhvi_4bd.csv*** plus the number of columns in ***Metro_zhvi_4bd.csv*** minus the ID variable you merged on. Check *data_metro_zip* to make sure the merge went as you expected.

Note that if you have two variables with the same name in both files, R will attach a .x to the variable name in ***Metro_zhvi_4bd.csv*** and a.y to the variable name in ***Zip_zhvi_4bd.csv***. For example, if you have a variable named SizeRank in both files, *data_metro_zip* will contain both variables and name it SizeRank.x (the variable in ***Zip_zhvi_4bd.csv***) and SizeRank.y (the variable in ***Metro_zhvi_4bd.csv***).

Exercise 4: **[10pts]** Let's find a solution to avoid having variables with the same names in the two files above.

Missing data

A special value used across all data types is NA. The value NA indicates a missing value (stands for "Not Available"). Properly treating missing values is very important. The first question to ask when they appear is whether they should be missing in the first place. Or did you make a mistake when data wrangling? If so, fix the mistake. If they should be missing, the second question becomes how to treat them. Can they be ignored? Should the records with NAs be removed?

Numerics also use other special values to handle problematic values after division. R spits out -Inf and Inf when dividing a negative and positive value by 0, respectively, and NaN when dividing 0 by 0.

```
-1/0
```

```
## [1] -Inf
```

```
1/0
```

```
## [1] Inf
```

```
0/0
```

```
## [1] NaN
```

You will likely encounter NA, Inf, and NaN values, even in already relatively clean datasets like those produced by the Zillow. The first step you should take is to determine if you have missing values in your dataset.

Exercise 5: **[20pts]** Let's find the number NA's in the *data_metro_zip.*

Exercise 6: **[10pts]** If there are NA value in *data_metro_zip*. Let's replace all NA with 0 in the *data_metro_zip***.**

What to submit:

Your submission should include the following:

1. Lab report answers to the six exercises above and source code.
2. Please create a folder called "yourname_studentID_lab3" that includes all the required files and generate a zip file called "yourname_studentID_lab3.zip".
3. Please submit your work (.zip) to Blackboard.