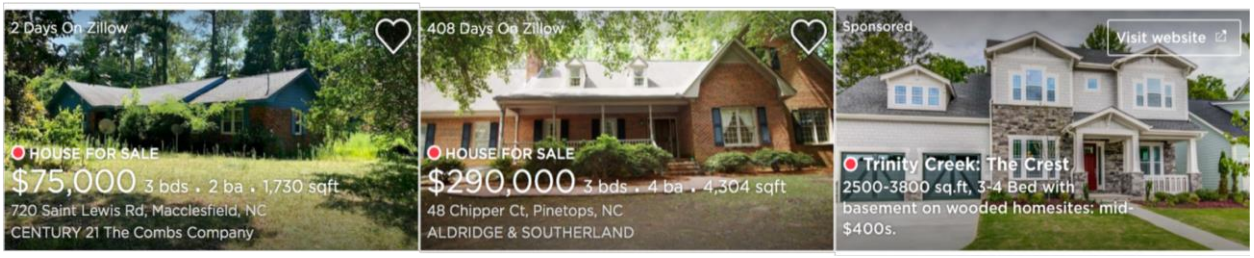
IT137IU: Data Analysis
Lab#4/Assignment#4: Exploration Data Analysis And Summary Statistic

Introduction



The goal of this lab is to acquire skills in running descriptive statistics. In this lab 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 in Lab 3. The objectives of the lab are as follows:

1. Learn how to use various R functions to summarize characteristics
2. Learn how to make presentation-ready tables of descriptive statistics

This lab guide follows closely and supplements the material presented in Chapters 1, 3, 5 and 22 in the textbook R for Data Science (RDS).

1A. Summarizing a single variable

In Lab 3, our goal is to merge together the datasets ***Zip_zhvi_4bd.csv*** and ***Metro_zhvi_4bd.csv*** to become a new larger dataset namely *data_metro_zip*. Moreover, recall from Lecture 3 our two important data types: categorical and numeric. Let's first summarize a numeric variable - house price - using some basic descriptive statistics.

Numeric variables

We can use the function `summarize()` to calculate mean house price. The first argument inside `summarize()` is the data object *data_metro_zip* and the second argument is the function calculating the specific summary statistic, in this case `mean()`, which unsurprisingly calculates the mean of the variable you indicate in between the parentheses.

We get the value ***NA***, which as we learned in Lab 3 represents a missing value. If a variable has missing values, functions like `mean()` will return an ***NA***.

```
mean(data_metro_zip$Price)
```

```
[1] NA
```

Exercise 1: [10pts] Let's use function `summarize()` to find out how many missing values in variable ***Price?***

Exercise 2: [10pts] Let's calculate the mean of variable ***Price*** with these missing values? [**Hint**] How can you deal with these missing values?

Exercise 3: [10pts] Does the average house price differ by RegionType? [**Hint**] Please see a sample result of this question as shown in picture below (your solution might be different because it depends on how you deal with missing values)

```
# A tibble: 3 x 2
  RegionType `mean(Price, na.rm = TRUE)`
  <chr>                            <dbl>
1 country                        281178.
2 msa                            237053.
3 zip                            327720.
```

We can calculate more than one summary statistic within `summarize()`. For example, to get the mean, median, standard deviation of ***Price***, and give column labels for the variables in the resulting summary table.

Exercise 4: **[10pts]** Let's find out the median, standard deviation of variable ***Price***.

Categorical variables

Let's next summarize a single categorical variable ***RegionType*** indicates whether a house belongs to which region levels zip/country/msa. The variable has three categories: zip, country and msa.

Exercise 5: [10pts] Let's find out the percent of region types. [**Hint**] you'll need to combine the functions `group_by()`, `summarize()` and `mutate()` using %>%. Below is a sample result of this question.

```
# A tibble: 3 x 3
  RegionType       n      freq
  <chr>        <int>     <dbl>
1 country        284 0.0000620
2 msa         242536 0.0529
3 zip        4341224 0.947
```

1B. Summarizing two variables

The functions we've gone through so far describe one variable. It is often the case that we are interested in understanding whether two variables are associated with one another.

Let's go through the ways we can describe the association between: (1) two categorical variables; (2) one categorical variable and one numeric variable.

Two categorical variables

To summarize the relationship between two categorical variables, you'll need to find the proportion of observations for each combination, also known as a cross tabulation. Let's create a cross tabulation of the categorical variables ***RegionType*** and ***StateName***.

Exercise 6: **[20pts]** Let's find proportion of observations for ***RegionType*** and ***StateName*** in the *data_metro_zip.* [**Hint**] We do this by using both variables in the `group_by()` command. Below is a sample result of this question.

```
# A tibble: 102 x 4
# Groups:   RegionType [3]
   RegionType StateName     n    freq
   <chr>      <chr>     <int>   <dbl>
 1 country    ""          284 1
 2 msa        "AK"       1136 0.00468
 3 msa        "AL"       5396 0.0222
 4 msa        "AR"       4828 0.0199
 5 msa        "AZ"       2840 0.0117
 6 msa        "CA"       9372 0.0386
 7 msa        "CO"       4828 0.0199
 8 msa        "CT"       1420 0.00585
 9 msa        "DE"        568 0.00234
10 msa        "FL"       8236 0.0340
# i 92 more rows
# i Use `print(n = ...)` to see more rows
```

One categorical, one numeric

A typical way of summarizing the relationship between a categorical variable and a numeric variable is to take the mean of the numeric variable for each level of the categorical variable.

Exercise 7: [20pts] Let's find the mean of house prices for variable ***StateName*** in the *data_metro_zip***.**

2. Tables for presentation

The output from the descriptive statistics we've ran so far is not presentation ready. For example, taking a screenshot of the following results table produces unnecessary information that is confusing and messy.

```
# A tibble: 3 x 4
  RegionType `mean(Price, na.rm = TRUE)`  Median      SD
  <chr>                                  <dbl>   <dbl>   <dbl>
1 country                              281178. 267196.  74297.
2 msa                                  237053. 198843. 144920.
3 zip                                  327720. 248201. 308172.
```

Furthermore, you would like to show a table, say, in your final project that does not require you to take a screenshot, but instead can be produced via code, that way it can be fixed if there is an issue, and is reproducible.

One way of producing presentation tables in R is through the `flextable` package. First, you will need to save the tibble or data frame of results into an object. For example, let's save the above results into an object named ***regiontype.summary***

You then input the object into the function `flextable()`. Save it into an object called *my_table*

Exercise 8: [10pts] Let's create the *my_table* as described above, and save it to a .png file. Below is a sample result

RegionType	mean(Price, na.rm = TRUE)	Median	Standard Deviation
country	281,178.3	267,196.2	74,296.57
msa	237,053.2	198,843.0	144,919.66
zip	327,719.9	248,200.8	308,172.04

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_lab4" that includes all the required files and generate a zip file called "yourname_studentID_lab4.zip".
3. Please submit your work (.zip) to Blackboard.