IT137IU: Data Analysis
Lab#5/Assignment#5: Exploration Data Analysis And Summary Statistic (continue)

Introduction

2 Days On Zillow
HOUSE FOR SALE
$75,000 3 bds . 2 ba . 1,730 sqft
720 Saint Lewis Rd, Macclesfield, NC
CENTURY 21 The Combs Company
408 Days On Zillow
HOUSE FOR SALE
$290,000 3 bds . 4 ba . 4,304 sqft
48 Chipper Ct, Pinetops, NC
ALDRIDGE & SOUTHERLAND
Sponsored
Visit website
Trinity Creek: The Crest
2500-3800 sq.ft, 3-4 Bed with basement on wooded homesites: mid-$400s.

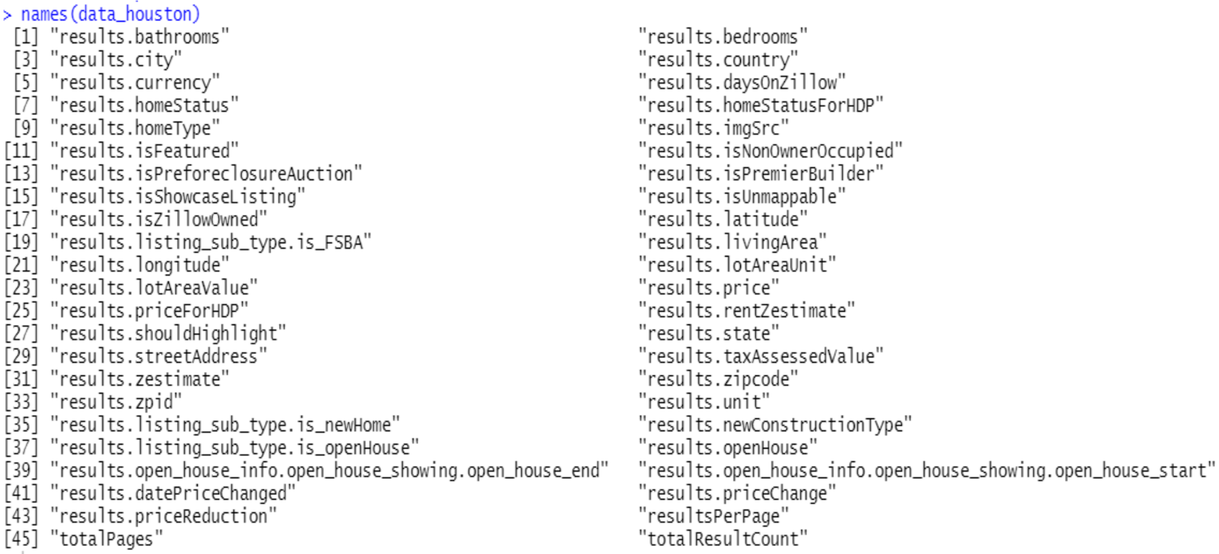
In this lab you will acquire the skills needed to process and present data in R. The objectives of the guide are as follows:

1. Learn basic data wrangling operations.
2. Learn how to use various R functions to summarize characteristics.

I uploaded a file on Blackboard containing housing price data at city levels taken from Zillow Group. Download the file ***Zillow-Houston-TX.csv*** and save it into the same folder where your Lab 5 file resides.

Data Wrangling

To see the names of variables in the dataset `data`, use the `names()` command.



```
> names(data_houston)
 [1] "results.bathrooms"                                       "results.bedrooms"
 [3] "results.city"                                            "results.country"
 [5] "results.currency"                                        "results.daysOnZillow"
 [7] "results.homeStatus"                                      "results.homeStatusForHDP"
 [9] "results.homeType"                                        "results.imgSrc"
[11] "results.isFeatured"                                      "results.isNonOwnerOccupied"
[13] "results.isPreforeclosureAuction"                         "results.isPremierBuilder"
[15] "results.isShowcaseListing"                               "results.isUnmappable"
[17] "results.isZillowOwned"                                   "results.latitude"
[19] "results.listing_sub_type.is_FSBA"                        "results.livingArea"
[21] "results.longitude"                                       "results.lotAreaUnit"
[23] "results.lotAreaValue"                                    "results.price"
[25] "results.priceForHDP"                                     "results.rentZestimate"
[27] "results.shouldHighlight"                                 "results.state"
[29] "results.streetAddress"                                   "results.taxAssessedValue"
[31] "results.zestimate"                                       "results.zipcode"
[33] "results.zpid"                                            "results.unit"
[35] "results.listing_sub_type.is_newHome"                     "results.newConstructionType"
[37] "results.listing_sub_type.is_openHouse"                   "results.openHouse"
[39] "results.open_house_info.open_house_showing.open_house_end" "results.open_house_info.open_house_showing.open_house_start"
[41] "results.datePriceChanged"                                "results.priceChange"
[43] "results.priceReduction"                                  "resultsPerPage"
[45] "totalPages"                                              "totalResultCount"
```

The columns that will interest us the most in this dataset are:

- `price`: the price of a house.
- `lotAreaValue`: the surface of a house.
- `homeType`: type of a house.

In particular, we will look at summary statistics for `price` and `lotAreaValue`, either for the whole data set or independently for each type of house. Notice that both of these variables are numeric. They are vectors of numbers, each representing an observation.

Exercise 1: **[10pts]** Let's remove all the prefix "results." in variables contain them and deal with the missing values in each variable as described above.

Measures of central tendency & Measures of dispersion

Measures of central tendency map a vector of observations onto a single number that represents, roughly put, "the center". Since what counts as a "center" is ambiguous, there are several measures of central tendencies. Different measures of central tendencies can be more or less adequate for one purpose or another. The type of variable (nominal, ordinal or metric, for instance) will also influence the choice of measure. We will visit three prominent measures of central tendency here: *(arithmetic) mean*, *median* and *mode*.

Exercise 2: **[10pts]** Let's find out the mean, median, and mode of variable `price` for different types of houses.

Measures of dispersion indicate how much the observations are spread out around a measure of central tendency, let's say, "a center". Intuitively put, they provide a measure for how diverse, variable, clustered, concentrated or smeared out the data observations are. In the following, we will cover three common notions: *variance*, *standard deviation* and *quantiles*.

Exercise 3: **[20pts]** Let's find out the variance, standard deviation and interquartile range (IQR) of variable `price` for different types of houses.

Hint: Please refer IQR from the Wikipedia Interquartile range.

Note that the IQR is the difference between the 75th and 25th percentiles. It is a measure of spread, and more generally, an indicator of inequality. Another measure of spread or inequality is the 90/10 ratio. To calculate this ratio, we'll first need to calculate the 90th and 10th percentiles using the quantile() command, where we indicate the percentile using the argument p =. We can do all of this inside summarize()

Exercise 4: **[20pts]** Let's find out the ratio 90/10 of variable `price` for different types of houses.

Covariance and correlation

The covariance between $\vec{x}$ and $\vec{y}$ measures, intuitively put, the degree to which changes in one vector correspond with changes in the other. Formally, covariance is defined as follows (notice that we use n-1 in the denominator to obtain an unbiased estimator if the means are unknown):

$$\mathrm{Cov}(\vec{x},\vec{y}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu_{\vec{x}})\,(y_i - \mu_{\vec{y}})$$

Exercise 5: **[20pts]** Let's find out the covariance between two vectors (variables) `price` and `lotAreaValue`.

Note: There are different unit measurements in variable `lotAreaValue` (sqft, acres). Therefore, you should change them to the same unit measurement.

Covariance is a very useful notion to show how two variables, well, co-vary. But the problem with this notion of covariance is that it is not invariant under linear transformation.

To compensate for this problem, we can look at **Bravais-Pearson correlation**, which is covariance standardized by standard deviations:

$$r_{\vec{x}\vec{y}} = \frac{\mathrm{Cov}(\vec{x},\vec{y})}{\mathrm{SD}(\vec{x})\,\mathrm{SD}(\vec{y})}$$

Exercise 6: **[20pts]** Let's find out the correlation between two vectors (variables) `price` and `lotAreaValue`. What's the meaning of this correlation value?

What to submit:

Your submission should include the following:

1. Lab report answers to all exercises above and source code.
2. Please create a folder called "yourname_studentID_lab5" that includes all the required files and generate a zip file called "yourname_studentID_lab5.zip".
3. Please submit your work (.zip) to Blackboard.