For this assignment, we will be using R Studio and we will learn a minimal set of instructions for visualizing data based on such framework.

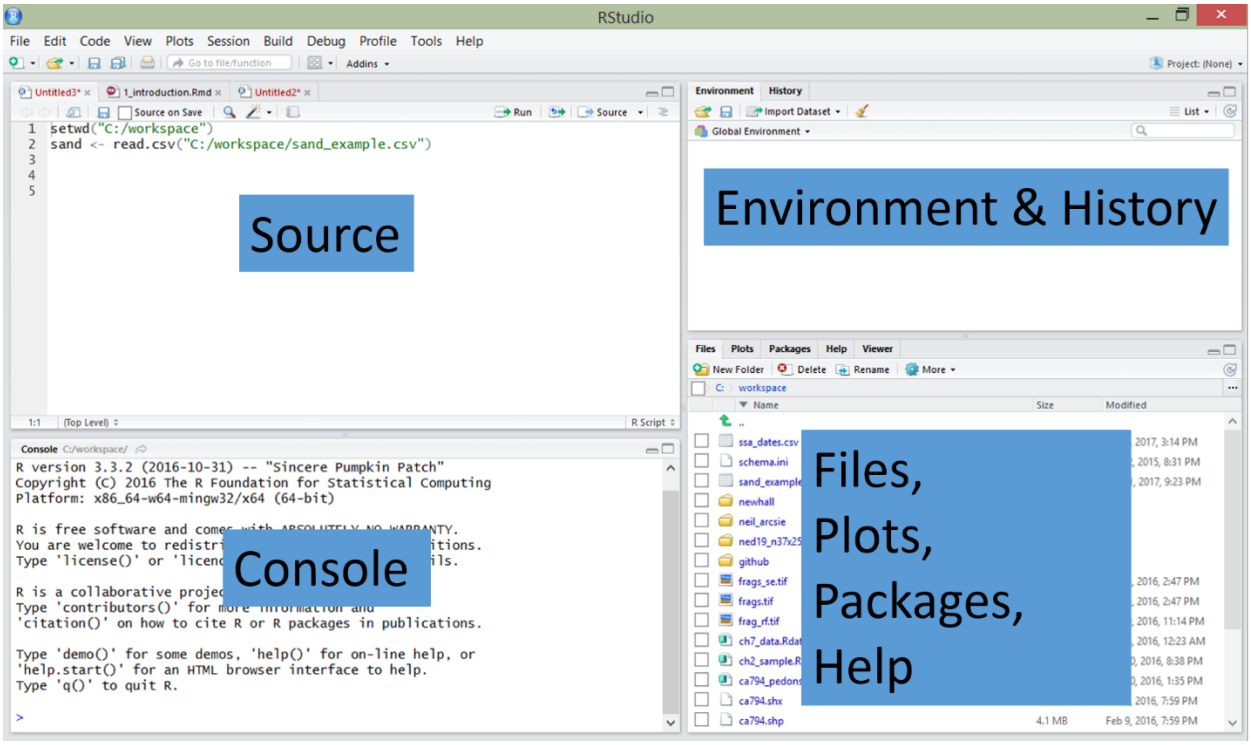
For a more advanced set of features, you can take a look at ggplot2

Setup

R is a free, open-source software and programming language developed in 1995 at the University of Auckland as an environment for statistical computing and graphics.

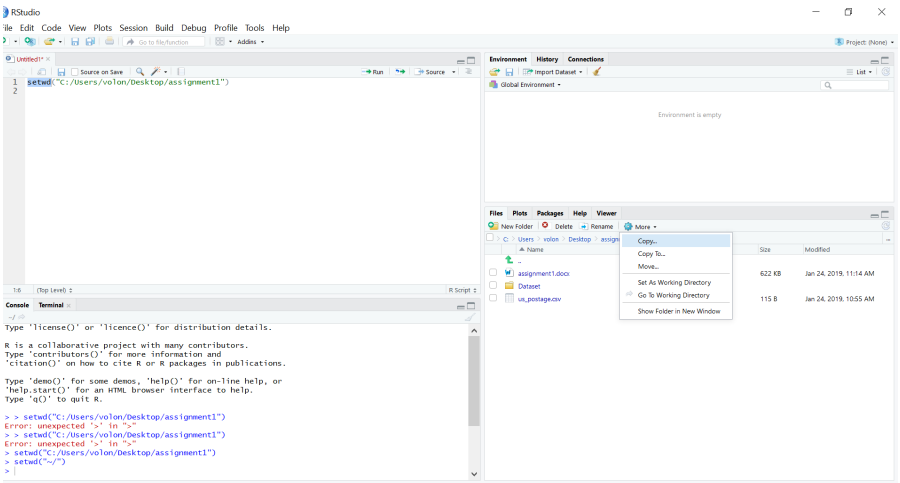
R studio is by far the most used IDE for R. You can either install it as a standalone package or by using Anaconda.

Once you open R studio, first thing you have to do is to setup a working directory



Description: a screen capture of the interface of R.

1) Create a new directory "*Assignment_1*" copying inside the entire content of the assignment
2) In R studio navigate to the directory "Assignment_1" by using the File tab in the Files, Plots, Packages,… window.
3) Now we need to set this as our working directory. We can either:
    a. Run the command setwd(".") in the source panel.
    b. Using the interface by clicking on "More"->"Set as working directory"



Tip: In the source panel, if you set the cwd() command, it should output the Current Working Directory (CWD)

Viz idioms

In this assignment, we will study three different datasets and for each of them, we will visualize different characteristics.

We will learn to create three types of graphs

Scatterplots

Scatter plots can be used in many different ways for either show correlation or as a baseline for creating line charts.

COMMAND - plot(1:length(attributeToVisualize), attributeToVisualize)

Sometimes additionally to the scatterplot it is convenient to extract a fitted line to better show trends (especially when considering time-varying data)

COMMAND - scatter.smooth(x=1:length(attributeToVisualize), y= attributeToVisualize)

Histogram

A histogram is a plot that lets you discover, and show, the distribution (shape) of a set of data.

COMMAND – histo(attributeToVisualize)

A histogram shows a discrete distribution of values along an ax. Sometimes we may want to study the same data assuming they are representing a continuous phenomenon. To do so we should compute a density function out of these values.

COMMAND – densityToVisualize <- density(attributeToVisualize)

plot(densityToVisualize)

Notice that the density values cannot be computed if the input array presents some Nan Values. To clean up your input data in R you can use the command

COMMAND – newValues <- attributeToVisualize[!is.na(attributeToVisualize)]

Bubble Chart

A bubble chart is a powerful visualization idiom that combines three attributes in the same graph. A bubble chart is created in R by means of the command symbols (type ?symbols for help in R studio).

COMMAND – symbols(attributeToVisualize 1, attributeToVisualize2, attributeToVisualize3)

REMEMBER – the third value, is supposed to be a vector of radii

Studying the dataset

All dataset we will be using in this assignment are stored in csv files. A csv file can be easily imported with

COMMAND – newDataset <- read.csv(file="filename.csv", header=TRUE, sep=",")

As usual you can use ?read.csv for help.

After a dataset is loaded into a variable you can access to its attributes by typing newDataset$attributeName

After loading a dataset remember to study its content before deciding a valuable viz idiom. You can do this by either:

- Double click on the variable name in the Environment view
- Run the command View(newDataset)

Assignment

Unemployment dataset.

What information is this dataset showing? (Based on what you can deduct from the dataset itself)

What is the number and type [categorical,ordinal,quantitative] of attributes is shows?

Viz idiom choice – we would like to show the number of people unemployed and how this number changed over time.

Birth rate dataset.

What information is this dataset showing? (Based on what you can deduct from the dataset itself)

What is the number and type [categorical,ordinal,quantitative] of attributes is shows?

Viz idiom choice – we would like to show the distribution of newborns over the year of 2008

Crime rate dataset

What information is this dataset showing? (Based on what you can deduct from the dataset itself)

What is the number and type [categorical,ordinal,quantitative] of attributes is shows?

Viz idiom choice – we would like to show, in a single graph, the relationships between the murder and burglary rate. We would like to show this in relation to the population of each place.

Submission

Create a folder in Github called submission and upload there this word file with your answers, a .r file with the code you wrote to generate the visualizations and one image for each graph you created.