15.071: The Analytics Edge **Fall 2022**

Homework 0: Introduction to R

Out: September 07; Due: September 11, 11:59 pm.
Please post the assignment in pdf format with file name "Lastname_15071_HW0.pdf".

Problem 1: Setting up `R` and RStudio

For this course, we will be using the `R` programming language and the RStudio graphical user interface. Please follow the instructions below to install and set up these two pieces of software. Note: the versions listed here are the latest versions of 'R' as of mid August. If you have trouble with installing 'R', please first ask a classmate/Google, and if that doesn't help then email TAs.

a. Install `R` on your computer using the following steps—depending on your operating system.

   (i) On Windows: visit **https://cran.r-project.org/bin/windows/base/** and select "R 4.2.1 for Windows" at the top. Open the downloaded installer and use all default options.

   (ii) If you have a Mac, visit **https://cran.r-project.org/bin/macosx/** and select the relevant **.pkg** link to download the correct version of `R` for your computer. Most machines will be running a recent version of Mac and should download R-4.2.1. Open the downloaded installer and use all default options.

b. Install RStudio on your computer using the following steps—depending on your operating system.

   (i) If you have a Windows computer, visit **https://www.rstudio.com/products/rstudio/download/** and download "RStudio-2022.07.1-554.exe" for Windows 10/11 from the "All Installers" section. Run the installer and use all default options to install RStudio on your computer.

   (ii) If you have a Mac, visit **https://www.rstudio.com/products/rstudio/download/** and download "RStudio-2022.07.1-554.dmg" for macOS 10.15+ from the "All Installers" section. Open the downloaded **.dmg** file, which should open a folder with icons labeled "Applications" and "RStudio". Drag the "RStudio" icon into the "Applications" folder.

c. Open RStudio on your computer (from the start menu in Windows or from the Applications directory on a Mac). On a Mac, you may get a warning asking you if you want to open RStudio for the first time, since it was downloaded from the Internet; in this case click "Open".

d. Run the following commands in the "Console" section of RStudio—in the bottom-left part. These commands install packages that you will need in this course. You can copy-paste the commands one-by-one into the Console area and hit Enter. Note: You need to be connected to the Internet.

```
install.packages("dplyr")
install.packages("tidyr")
install.packages("ggplot2")
install.packages("ROCR")
install.packages("rpart")
install.packages("rpart.plot")
install.packages("caret")
install.packages("randomForest")
install.packages("tidyverse")
install.packages("tm")
install.packages("SnowballC")
```

e. Finally, run the following command. It calls a flower database named `iris` and provides summary statistics. Please report the output as part of your submission—either by copying and pasting it, or by taking a screenshot.

```
summary(iris)
```

Problem 2: Predicting Wine Quality using Linear Regression

You will construct a linear regression model to predict the quality of Bordeaux wines. The data are provided in the file `wine.csv`. The dataset comprises 46 observations (rows), each of which corresponds to a vintage year of Bordeaux wines auctioned in 2015; and it comprises 9 columns, described in Table 1.

Table 1: Columns in the dataset `wine.csv`.

Column	Type	Description
`Year`	Descriptor	Vintage year of the wines.
`LogAuctionIndex`	Dependent variable	Numerical value that captures an index related to the weighted average of prices of various Bordeaux wines by vintage year, and sold at auction in 2015. The index is actually the logarithm of this weighted average.
`WinterRain`	Independent variable	Amount of rain during the winter season (mm).
`HarvestRain`	Independent variable	Amount of rain during the harvest season (mm).
`GrowTemp`	Independent variable	Average temperature in the growing season (celsius).
`HarvestTemp`	Independent variable	Average temperature in the harvest season (celsius).
`Age`	Independent variable	Age of wine (years since vintage).
`FrancePop`	Independent variable	Population of France (millions of people).
`USAlcConsump`	Independent variable	US Alcohol Consumption (per capita, in liters of equivalent 100% alcohol).

a) Navigate to the directory containing `wine.csv` (either with the `setwd()` function or with the RStudio interface). Run the following commands to read the data into `R`, to construct a linear regression model that predicts `LogAuctionIndex` based on all variables and all observations, and to display the results:

```
wine <- read.csv("wine.csv")
mod <- lm(LogAuctionIndex ~ . -Year, data=wine)
summary(mod)
```

What is the R^2 of your model?