Exploratory Data Analysis in R – ES218

*Manny Gimond - Spring 2022*

Course Description

Exploratory data analysis employs methods such as robust data summaries and data visualization to isolate important patterns and features in the data to shed light on the phenomena being investigated. Students will learn the building blocks of effective graphic design for data exploration and for publication using the *R* programming environment. They will also learn how to manipulate and restructure complex data sets for data analysis. Students will use *R* and *RStudio* to generate dynamic reports that will integrate both analysis and presentation with the goal of generating reproducible research. To this end, students will make use of the GitHub repository along with the versioning application, git, throughout the course.

Course evaluation

Students will be evaluated on their technical as well as theoretical knowledge of data analysis. Evaluations will be based on a **mid-term exam** (contributing **25%** towards the overall grade), **assignments** (contributing **40%** towards the overall grade) and a class project (contributing **35%** towards the overall grade).

Course websites

The lecture notes are hosted on Github: <http://mgimond.github.io/ES218/index.html>  
Assignments and discussion forums will be hosted on **Moodle**.

*Optiona*l course book: Much of the material in this course is inspired from *Visualizing Data* (by William S. Cleveland) and *Exploratory Data Analysis* (by John Tukey). These books are not required for this course.

Topics covered

1. Data analysis environment
   * R and RStudio
   * R packages
2. Working with data in R
   * Data types: numeric, character, factor, logical and dates
   * Data objects: vector, data frame, lists
   * Read and writing data files
3. Manipulating data tables in R using *dplyr* package
   * Subsetting data
   * Filtering by values
   * Summarizing tables by rows and columns
   * Reshaping tables (long vs. wide)
4. Plotting data in R
   * Using the base plotting functions
   * Using the *ggplot* package
5. Univariate analysis
   * Levels and spreads (aka central value and distribution)
   * Quantile plots and Q-Q plots
   * Fits and residuals: an iterative process
   * Re-expressing values (aka transformations)
6. Bivariate analysis
   * Smoothing and banking
   * Robust fitting
   * Time series
   * The 3 R’s of EDA: Robustness, Re-expression, Residuals
   * Traditional regression and robust alternatives
   * P values explained
7. Mapping spatial variables
8. Interacting with relational databases
   * Using SQL
   * Using *dplyr*