Project 1 guidelines

Important Notes

- This document is meant to answer some of your questions regarding what to do and what to deliver in project 1. It is not meant to replace the project statement, nor to contain complete information regarding the project tasks. **READ** the project statement **FIRST** and **THEN** this document to find answers to your questions.
- We won't tell you exactly what to do: if we did, it wouldn't be Exploratory Data Analysis.
- You don't have to submit the entire data set, but you must indicate in the report which data sources you used and submit small samples so that we can test your code. The samples should be representative of the data set, that means if you collect data for 7 days, you must submit 7 samples (1 small sample per day). Do not exceed 1-2 MB.
- The most important part of your work is going to be the observations and conclusions you derive from your analysis, together with the plots and data that support your claims. That being said, try to use the basic statistics covered in the book and handouts in creative ways: there's a lot you can do with those simple tools.
- Each problem must be zipped or tarred in a separate archive. You will submit 5 archives in total, 1 per problem. Each archive will contain the scripts and the report corresponding to the same problem. For naming files and archives, check the bullet point below, highlighted in yellow.
- Please, comment generously your scripts: it will improve their readability and quality.

- Write professional reports that include plots and interesting observations.
- For naming files, follow the conventions described in the project statement. However, prefix your ubit name to every file you submit. E.g. if the file is called EDAnyt.R, and your ubit name is luigidit, call the file luigidit_EDAnyt.R.
- For naming archives, just call them prob1, prob2, etc. It is going to make life easier for you and for us as well.
- The project deadline is strict: do not ask for extensions because you are just going to make a bad impression in front of the professor and the TAs. Work harder, instead.

Problem 1: Data acquisition

- **Main Task**: Collect data and import it in R
- **What to do**: write R script to collect data of interest (suggested Twitter, but can be something else) and to save it to disk. Collect the data for at least one week (you're going to use it in problem 4 and 5). The data will be formatted as JSON objects (most likely), so you need to write a second script to import the JSON objects, previously saved to disk by the first script, into R structures for analysis. Packages are available to help you with both tasks, but you need to read the package documentation, understand what each package does and which API it uses, and write down all these observations and explanations in the report.
- **Deliverables**: R script for data mining, R script for importing JSON, small data samples, report that mentions data source, data structure, period of time data has been collected, and any comments regarding imported packages (as described above).

Problem 2: Simple EDA

- **Main Task**: Performing EDA on a simple data set
- **What to do**: In few words, problems 1, 2, 3, and 4 on page 38. Answering those questions will help you write the code and the report. The book has part of the solution but many things are missing. Write one script for problem 1 and 2, another for problem 3, and a report containing plots for BOTH scripts and observations/analysis of the results.
- **Deliverables**: 2 R scripts and the report.

Problem 3: Case study

- **Main Task**: performing EDA on a real world data set
- **What to do**: Very similar to problem 2 but with a different data set. Like problem 2, you need to write two R scripts, one for the analysis on a single data set (Brooklyn) and the other for extended analysis on all data sets (Manhattan, Queens, etc.). There is another difference: you must answer questions 1-6 page 48-49. Answering those questions will help you write the code and the report.
- **Deliverables**: 2 R scripts and the report.

Problem 4: Data product

- **Main Task**: Combine different data sets and perform statistical analysis for business recommendation
- **What to do**: collect data from Twitter for a week (must choose right keywords to select relevant tweets). Write 1 R scripts to analyze the data and make recommendation on how to improve RealDirect profit. This problem is more open ended than the others, so try to use all your creativity and smarts.
- **Deliverables**: R script, small samples, and report.

Problem 5: Stream processing

- **Main Task**: Create UI to access statistical results
- **What to do**: Collect Twitter data for a week regarding elections (or any other topic you like) and perform statistical analysis day by day and for the entire week. Create a UI with RShiny to display trends and summarization. Collect plots and observations in the report.
- **Deliverables**: R script, small samples, and report.