Data Valorization:
First Delivery in Kaggle

Lionel Fillatre

fillatre@unice.fr

Important dates

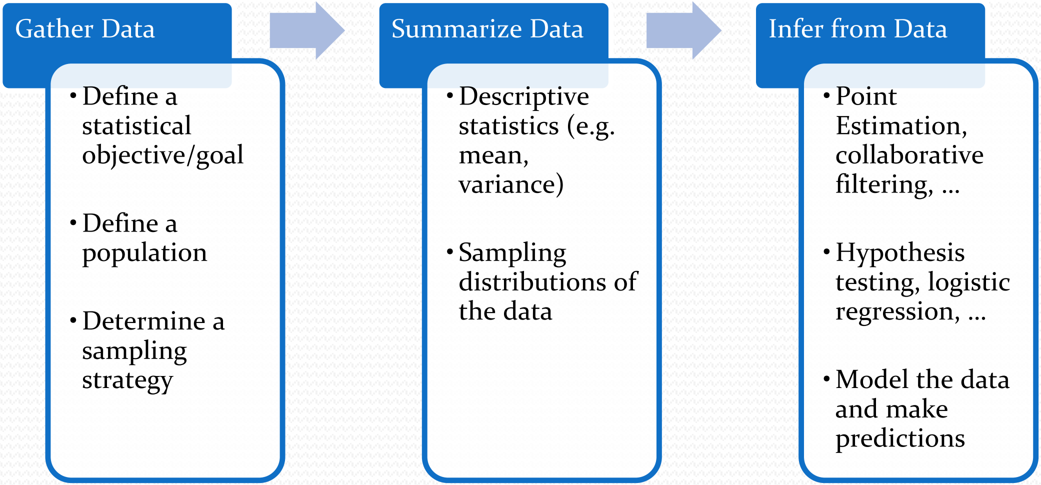
- Kaggle: three deadlines at 21h00 (-1 per day late)

1. Introduction and Data Description 06/02/2019
2. Data Gathering and Sampling 13/02/2019 => Kaggle team registration
3. Data Visualization 27/02/2019 => 1st written exam (no official lecture)
4. Shiny Application 06/03/2019
5. Point Estimation 13/03/2019
6. Logistic Regression 20/03/2019
7. Hypothesis Testing 27/03/2019 => 1st Kaggle delivery
8. Naïve Bayes Test 03/04/2019
9. Correspondence Analysis 10/04/2019
10. Recommendation System 24/04/2019 => 2nd written exam (no official lecture)
11. Reinforcement Learning 15/05/2019 => 2nd Kaggle delivery

Kaggle Challenge

- You have to define the question you aim to answer (classification, dimension reduction, regression, etc.)
- You have to provide numerical and theoretical justification of your analysis
- I encourage you to reuse an existing (or several) Kaggle Kernel
- You can write, run, and view best practice code and visualizations of this dataset on Kaggle Kernels.
- **You must exploit the theoretical tools and practical methods presented in this course!**

You must follow this analysis scheme



Gather Data

- Define a statistical objective/goal
- Define a population
- Determine a sampling strategy

Summarize Data

- Descriptive statistics (e.g. mean, variance)
- Sampling distributions of the data

Infer from Data

- Point Estimation, collaborative filtering, …
- Hypothesis testing, logistic regression, …
- Model the data and make predictions

1st Delivery

- Content: business goal, technical goal, data description, statistical analysis of data, data preprocessing in R to extract relevant information, brief description of the future analysis which will be detailed in the 2nd delivery, etc.

- You have to decide the balance between computer science and applied mathematics

- The 2nd delivery will focus on advanced analytics (Logistic regression, Bayes classification, Regression, Boosting, Random forest, neural networks, Deep learning, etc.). This advanced method can be seen or not in the lecture. It is your choice.

- All your assertions must be based on data analysis.

Talk

- Recorded talk (10 minutes in video) of the chosen challenge

- For each talk, all team members must talk and you can use a slideshow.

- For each video, record the video and post it somewhere (YouTube, etc.). Send an email with the URL of the video to your labs professor.

Assesment Criteria

- Eight criteria
  1. Quality of the talk
  2. Clarity of the R notebook
  3. Clarity of the business goal
  4. Relevance of the technical goal with respect to the business goal
  5. Depth of the statistical analysis
  6. Justification and complexity of the preprocessing steps
  7. Quality of the data-based reasoning
  8. Motivation for future advanced analytics

- Grade
  - Each criterion is graded from 0 to 4 (very poor, poor, fair, good, excellent)
  - The initial grading scale is 0-32
  - But the final grading scale is 0-20 (by applying the Rule of Three).