Probability and Statistics

Practical lectures 1 and 2 - 26/09/2018

About me

- Davide Posillipo, Data Scientist and Statistician
- Project Manager at Nunatac, Milan
- https://www.linkedin.com/in/davide-posillipo/
- https://github.com/DavidePosillipo

Assignments

- Four assignments
- At least one week per assignment
- Mix of coding and statistical analysis

Project

- All the groups will use the same dataset

- Focus on the statistical reasoning more than the coding skills

Outline of the first two lectures

- Setting up the framework
- Intro to Git
- Kaggle and Google Dataset Search
- Intro to Numpy
- Intro to Pandas
- Intro to Data Visualization with Python

Framework

- Python 3
- Anaconda distribution
- Jupyter notebook
- Spyder IDE

Intro to Git

- Version control of the code is a must in software development

- Data science is (also) software production

- Let’s use the official tutorial (https://guides.github.com/activities/hello-world/)

Git: how will we use it?

- Each group should create a **repository**
- Each component of the group should have a **GitHub account**
- Ideally, each component should have his/her **branch**
- One component should coordinate the **merging operation**
- The assignments/final project will be delivered with a **push**

Kaggle and Google Dataset Search

- Kaggle: your home for data science?

- Google Dataset Search: a valuable resource

Python data science stack