CS1070: Taming Big Data

Numpy, Pandas and DataFrames

Logistics

- Homework 1 due today

Pandas DataFrame

- 2D labeled data structure with rows and columns
- Each row represents one individual or sample
- Columns can be different types

	NAME	AGE	DESIGNATION
1	a	20	VP
2	b	27	CEO
3	c	35	CFO
4	d	55	VP
5	e	18	VP
6	f	21	CEO
7	g	35	MD

Step 1) Data Exploration

- Before we can get to any actual modeling, we need to make sure we understand the data that we have

- Useful questions to answer in data exploration:
  1. How many columns are there in the data?
  2. How many rows (samples)?
  3. What kind of features are there? Which are ***categorical*** and which are ***numerical***?
  4. What does the data in these features look like? (e.g., what are the statistics of it?)
  5. Is there any missing data?

Boolean Masks

- Also called logical masks
- A way to filter an array for some condition
- “For each value in my array, is that value true/false for my condition?”
- Can then just look at data where the condition is true or just places where it is false
- For example, we might want to look for any IDs that are duplicated. Let’s go back to our notebook to see how we would do that.