



—
BC COMS 2710:
Computational Text Analysis

Lecture 4 – Pandas Visualizations



- Tutorial 1.1 & 1.2
 - *Should* be submitted already

- Tutorial 1.3
 - Released today
 - *Should* be submitted tomorrow (Friday 05/07)

- Homework 01
 - Due Monday (05/10)

- Reading 01
 - Due Sunday night (05/09)
 - Overleaf template is online schedule
 - Make your own copy of the template on overleaf



- Weekly
 - Mondays: 1:00 – 3:30pm
 - Thursdays: 5:00 – 6:00pm

- Gauri:
 - Fridays 1:00pm – 2:00pm
 - *Will add more times*



Office hours:

- Disable waiting room
- Breakout rooms for 1-on-1 discussions
- Main room for broad discussions
- Breakout rooms for groups

Autograder

- Some hidden tests on Gradescope



Pandas



- **Aurélien Geron** wrote an excellent notebook going through pandas:
 - https://nbviewer.jupyter.org/github/ageron/handson-ml2/blob/master/tools_pandas.ipynb
- BabyPandas online textbook:
 - https://eldridgejm.github.io/dive_into_data_science/02-data_sets/dataframes.html
- Cultural Analytics Textbook:
 - Linked on course website



- Counting values
- Query
- Grouping
- Functions
- Merge
- Times



—

Matplotlib Visualizations

—



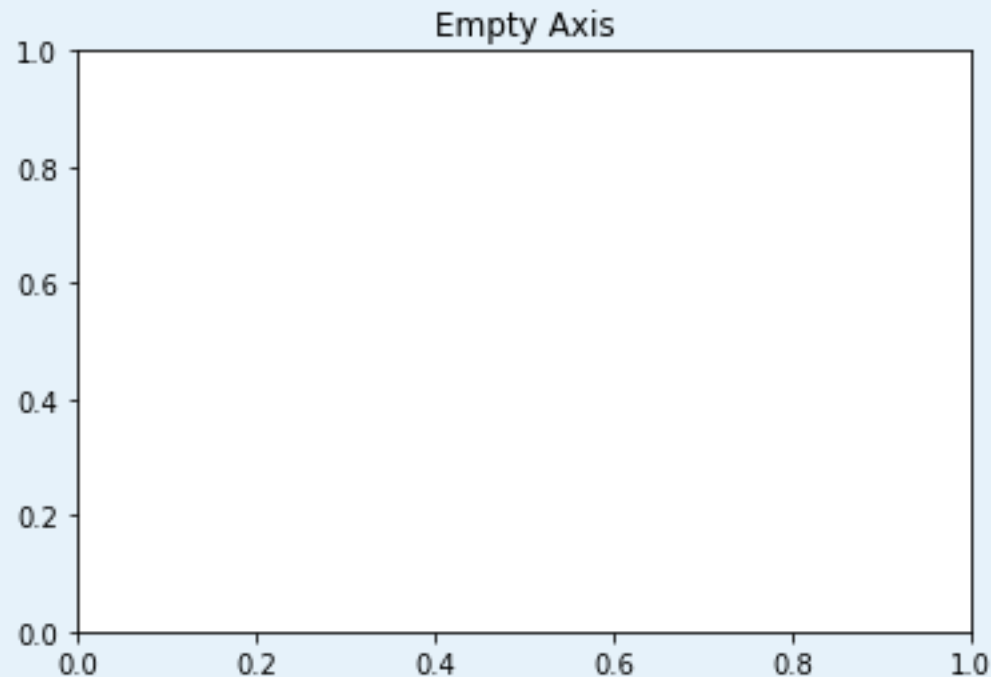
- Categorical
 - Differences are not quantifiable
 - Set of discrete values
- Numerical:
 - Meaningful differences
 - Examples:
 - Time
 - Temperature



- Python's primary plotting package
- Widely used for data visualization
- Easy to use for simple visualizations, but allows for fine-grained control for experienced users •
- We will look only into `matplotlib.pyplot`
 - Highest-level module
 - Create figures, add elements such as lines and text

Slides from [Jorge Mendez](#)

- Figure: the whole figure.
 - Contains Axes, artists (titles, legends).
 - Should have at least one Axes





- Axes: a plot.
 - There may be more than one per *Figure*.
 - Has a title, an x-label, and a y- label
 - Artists (titles, legends).



- Pandas integrates Matplotlib plotting functionality
- `df.plot()` - plots DataFrame
- `df.plot(kind = "...")`:
 - 'bar' or 'barh' for bar plots
 - 'hist' for histogram
 - 'box' for boxplot
 - 'kde' or 'density' for density plots
 - 'area' for area plots
 - 'scatter' for scatter plots
 - 'hexbin' for hexagonal bin plots
 - 'pie' for pie plots