

Python Programming and Machine Learning for Economists – Exercises

Michael E. Rose, PhD

Course at IfW Kiel (Advanced Studies Programme), August 2022


Work only in the GitHub repo that you shared with me!

For each of the exercises below, create *one script* (no jupyter Notebook) that contains both code and if applicable answers as comments in order.

Save the script in the main folder, properly named in correspondence to the name of the exercise. Avoid colons, dots and blanks in the filename. All scripts need to adhere to PEP8 and must be readable to someone that knows Python. Above all, they must execute without error on my machine.

Apart from the script, there should be one folder named "output" to store output such as figures and tables.

1 Optional Exercises for Pandas and Plotting

 Sketch for today: [French Taunting](#)

1. Tips

- Load seaborn's tips dataset using `seaborn.load_dataset("tips")` into a `pandas.DataFrame()`.¹
- Convert the short weekday names to their long version (e.g., "Thursday" instead of "Thur") using `.replace()`.
- Create a scatterplot of "tip" vs. "total_bill" colored by "day" and facets (either by row or by column) by variable "sex". Label the axis so that the unit becomes apparent. Save the figure as `./output/tips.pdf`

2. Occupations

- Import the pipe-separated dataset from <https://raw.githubusercontent.com/justmarkham/DAT8/master/data/u.user> into a `pandas.DataFrame()`. The data is on occupations and demographic information.
- Print the last 10 entries and then the first 25 entries.
- What is the type of each column?
- Count how often each occupation is present! Store the information in a new object.
- How many different occupations are there in the dataset? What is the most frequent occupation? (👉 Try to use a programmatic solution for these questions using the new object!)
- Sort the new object by index. Then create a figure and an axis. Plot a histogram for occupations on that axis (👉 Do not use `.hist()`). Add an appropriate label to the x-axis. How does the figure look like if you don't sort by index beforehand? Save the figure as `./output/occupations.pdf`!

3. Iris


- Read the Iris dataset from <https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data> as `pandas.DataFrame()`. The data is on measures of flowers and the values are on "sepal_length (in cm)", "sepal_width (in cm)", "petal_length (in cm)", "petal_width (in cm)" and "class". Since the data doesn't provide the column names, add the column names after reading in, or alternatively provide while reading in.
- Set the values of the rows 10 to 29 of the column 'petal_length (in cm)' to missing.
- Replace missing values with 1.0.
- Save the data as comma-separated file named `./output/iris.csv` without index.
- Visualize the distribution of all of the continuous variables by "class" with a catplot of your choice. Optionally, try to tilt/rotate the labels using `.set_xticklabels()`, which accepts a `rotation` parameter). Save the figure as `./output/iris.pdf`.

4. Memory


¹If this fails, check the corresponding slide in unit "Plotting w/ pandas (matplotlib), and w/ seaborn" for a workaround!

- (a) Load the comma-separated data from <https://query.data.world/s/wsjbxdqhw6z6izgdxijv5p21fqh7gx> into a `pandas.DataFrame()` (large file!).
- (b) Inspect the object using `.info()` and afterwards `.info(memory_usage="deep")`. What is the difference between the two calls? How much space does the DataFrame require in memory?
- (c) Create a copy of the object with only columns of type object by using `.select_dtypes(include=['object'])`.
- (d) Look at the summary of this object new (using `.describe()`). Which columns have very few unique values compared to the number of observations?
- (e) Does it make sense to convert a column of type object to type category if more than 50% of the observations contain unique values? Why/Why not?
- (f) Convert all columns of type object of the original dataset to type category where you deem this appropriate.
- (g) What is the final size in memory?
- (h) Could above routine have sped up somewhere? (🔗 Look at the documentation for `.read_csv()`).
- (i) Subset the `pandas.DataFrame()` to all the numeric columns only. Then store the file twice in folder `./output/`, namely as CSV file using `.to_csv()` and secondly as feather file using `.to_feather()` (Do not attempt to check in & push these files, as this they would be blocked by GitHub!). By how much do the file sizes differ approximately, and why is that? (🔗 Check out the [corresponding documentation](#))

2 Exercises for Unsupervised Machine Learning

 Sketch for today: [Village Witch](#)


1. Feature Engineering

- (a) Load the Breast Cancer dataset using `sklearn.datasets.load_breast_cancer()` as per usual.
- (b) Extract polynomial features (without bias!) and interactions up to a degree of 2 using `PolynomialFeatures()`. How many features do you end up with?
- (c) Create a `pandas.DataFrame()` using the polynomials. Use the originally provided feature names to generate names for the polynomials ( `.get_feature_names()` accepts a parameter) and use them as column names. Also add the dependent variable to the object and name the column "y". Finally save it as comma-separated textfile named `./output/polynomials.csv`.

2. Principal Component Analysis

- (a) Read the textfile `./data/olympics.csv` (in your git repository) into a `pandas.DataFrame()` using the first column as index. The data lists the individual performances of 33 male athletes during the [Decathlon of the 1988 Olympic summer games](#) (100m sprint, running long, (broad) jump, shot put, high jump, 400m run, 110m hurdles, discus throw, pole vault, javelin throw, 1500m run). Print summary statistics for each of the variables and decide (and act accordingly): Does it make sense to drop variable "score" before proceeding?
- (b) Scale the data such that all variables have unit variance. Which `pandas.DataFrame()` method can you use to assert that all variables have unit variance?
- (c) Fit a plain vanilla PCA model. Store the components in a `pandas.DataFrame()` to display the loadings of each variable. Which variables load most prominently on the first component? Which ones on the second? Which ones on the third? How would you thus interpret those components?
- (d) How many components do you need to explain at least 90% of the data?

3. Clustering

- (a) Load the iris dataset using `sklearn.datasets.load_iris()`. The data is on classifying flowers.
- (b) Scale the data such that each variable has unit variance.
- (c) Assume there are three clusters. Fit a K-Means model, an Agglomerative Model and a DBSCAN model (with `min_sample` equal to 2 and `epsilon` equal to 1) with Euclidean distance. Store only the cluster assignments in a new `pandas.DataFrame()`.
- (d) Compute the silhouette scores using `sklearn.metrics.silhouette_score()` for each cluster algorithm from c). Why do you have to treat noise assignments from DBSCAN differently? Which model has the highest Silhouette score?
- (e) Add variables "sepal width" and "petal length" including the corresponding column names to the `pandas.DataFrame()` that contains the cluster assignments. (Beware of the dimensionality!)
- (f) Rename noise assignments to "Noise".
- (g) Plot a three-fold scatter plot using "sepal width" as x-variable and "petal length" as y-variable, with dots colored by the cluster assignment and facets by cluster algorithm. ( Melt the `pandas.DataFrame()` with above variables as ID variables.) Save the plot as `./output/cluster_petal.pdf`. Does the noise assignment make sense intuitively?