Part 2 — Workshop 2

TECH2: Introduction to Programming, Data, and Information Technology

Richard Foltyn
Norwegian School of Economics (NHH)

October 11, 2024

See GitHub repository for notebooks and data:

`https://github.com/richardfoltyn/TECH2-H24`

1 Exercise: Data cleaning

Before doing actual data analysis, we usually first need to clean the data. This might involve steps such as dealing with missing values and encoding categorical variables as integers.

Load the Titanic data set in `titanic.csv` and perform the following tasks:

1. Report the number of observations with missing `Age`, for example using `isna()`.
2. Compute the average age in the data set. Use the following approaches and compare your results:
    1. Use the `mean()` method.
    2. Convert the `Age` column to a NumPy array using `to_numpy()`. Experiment with NumPy's `np.mean()` and `np.nanmean()` to see if you obtain the same results.
3. Replace the all missing ages with the mean age you computed above, rounded to the nearest integer. Note that in "real" applications, replacing missing values with sample means is usually not a good idea.
4. Convert this updated `Age` column to integer type using `astype()`.
5. Generate a new column `Female` which takes on the value one if `Sex` is equal to `"female"` and zero otherwise. This is called an *indicator* or *dummy* variable, and is preferrable to storing such categorical data as strings. Delete the original column `Sex`.
6. Save your cleaned data set as `titanic-clean.csv` using `to_csv()` with `,` as the field separator. Tell `to_csv()` to *not* write the `DataFrame` index to the CSV file as it's not needed in this example.

2 Exercise: Working with strings

Most of the data we deal with contain strings, i.e., text data (names, addresses, etc.). Often, such data is not in the format needed for analysis, and we have to perform additional string manipulation to extract the exact data we need. This can be achieved using the pandas string methods.

To illustrate, we use the Titanic data set for this exercise.

1. Load the Titanic data and restrict the sample to men. (This simplifies the task. Women in this data set have much more complicated names as they contain both their husband's and their maiden name)
2. Print the first five observations of the `Name` column. As you can see, the data is stored in the format *"Last name, Title First name"* where title is something like Mr., Rev., etc.
3. Split the `Name` column by `,` to extract the last name and the remainder as separate columns. You can achieve this using the `partition()` string method.

4. Split the remainder (containing the title and first name) using the space character `" "` as separator to obtain individual columns for the title and the first name.
5. Store the three data series in the original `DataFrame` (using the column names `FirstName`, `LastName` and `Title`) and delete the `Name` column which is no longer needed.
6. Finally, extract the ship deck from the values in `Cabin`. The ship deck is the first character in the string stored in `Cabin` (A, B, C, ...). You extract the first character using the `get()` string method. Store the result in the column `Deck`.

Hint: Pandas's string methods can be accessed using the `.str` attribute. For example, to partition values in the column `Name`, you need to use

```
df['Name'].str.partition()
```