# Data Analytics – Exercises

## (Week 03)

In these exercises, you will continue preparing the apartment data. In detail, you will learn how to combine the data with other data and work with pivot tables. In the data analytics process model, these exercises cover part of the step "Preparing & storing" data (see figure 1). Results of the exercises must be uploaded as separate files (no .zip files) by each student on Moodle. Details on how to submit the results can be found in the tasks below.

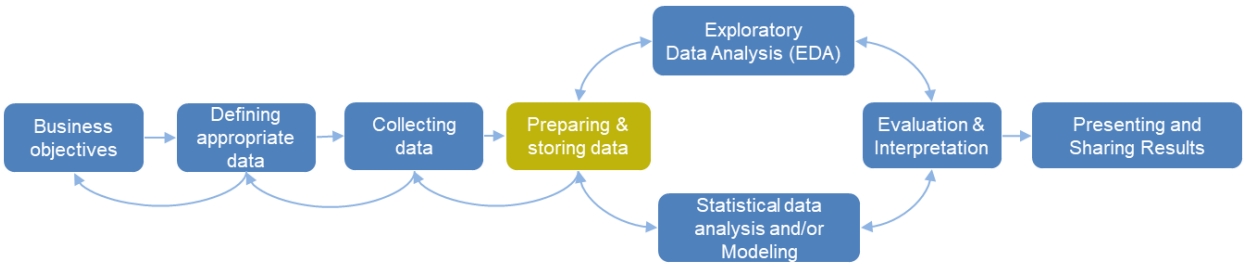


**Figure 1:** Data analytics process model (see slides of week 01)

## Task 1

In these exercises, you will learn to create new variables. Therefore, we will use the prepared rental apartment data. The tasks are:

1. Run the Jupyter notebook 'apartments_data_preparation_zuerich.ipynb' step by step and try to understand what the Python code does.
2. In the Jupyter notebook, go to the section 'Create additional variables from the apartment's descriptions' and look at the example under 'Create new binary (0/1) variable 'luxurious''.
3. Based on this example, create the following binary (0/1) variables '**furnished**', '**balcony**', and '**central**'. You are free to create additional variables. Note that there are only german words in the apartment descriptions which must be used to define the pattern for each for these variables. You can use https://www.deepl.com for translation.
4. In the Jupyter notebook, go to the example: 'Create categorical variable based on apartment area'.
5. Based on this example, create a new categorical variable based on the variable 'price_per_m2'. The variable should contain three levels 'low', 'medium', 'high'.

   **To be submitted on Moodle:** Jupyter notebook as html-file 'apartments_data_preparation_zuerich.html' with the additional variables described above.

# Task 2

In these exercises, you will learn to prepare data on the municipality level and combine these data with the apartment data. The tasks are:

a) Run the Jupyter notebook 'combining_and_organizing_data.ipynb' step by step and try to understand, what the code does.
b) Open the file 'municipality_data.xlsx'. It contains municipality-level data which are merged with the apartment data in the Jupyter notebook.

**<<note that the following parts c) … h) are MS Excel work only, no Python required>>**

c) Now, we would like to add one additional municipality-level variable to 'municipality_data.xlsx'. There is a file with mean taxable income per capita of Swiss municipalities available which was downloaded from the Bundesamt for Statistik: 'fiscal_data.xlsx'. The data includes mean taxable income per capita of Swiss municipalities.
d) Note that the 'mean_taxable_income' per municipality and person is needed (not the total income per municipality).
e) Use the Excel - function '=sverweis()' (german) or '=vlookup()' (english) in MS Excel and the municipality id provided in both files (named 'bfs_number' and 'Regions-ID', respectively) to join the '**mean_taxable_income**' as additional column to 'municipality_data.xlsx'.
   ➔ **Hint (1):** Use ChatGPT or web tutorials to find explanations and examples for Excel functions 'sverweis()' (geman) or 'vlookup()' (english).
   ➔ **Hint (2)**: The Regions-ID in the Excel File with the taxable income is in non-numeric and must be brought to a numeric form in Excel before it can be used.

**<<End of MS Excel work>>**

f) In the Jupyter notebook, look at the example in the section 'Join municipality data to rental apartment data using .merge()'.
g) Based on this example, merge '**mean_taxable_income**' as additional variable to the apartment data.

**To be submitted on Moodle:** see Task 3

# Task 3

In these exercises, you will learn to work with pivot tables in Python. The tasks are:

a) Use the Jupyter notebook 'combining_and_organizing_data.ipynb' as prepared in task 2.
b) Go to the section 'Pivoting data using .pivot_table()' and look at the example.
c) Add the variable `price_per_m2` to the existing pivot table.
d) In the pivot table, compare the `price` with the `price_per_m2`.

e) Is there a relationship between the variable area and the price_per_m2? If so, how can this be explained? Include a short explanation in the Jupyter notebook.
f) Create a new pivot table and use `aggfunc='count'` as the aggregation function. How is the number of apartments distributed in the pivot table? Include a short description in the Jupyter notebook cell.
g) Create a new pivot table with the mean values of `rooms`, `area`, `price`, and `price_per_m2` per municipality. Sort the pivot table by price and area with the most expensive municipalities on top.

**To be submitted on Moodle:** Jupyter notebook as html-file 'combining_and_organizing_data.html' with the additional variable mean_taxable_income from task 2 and the pivot tables from task 3.