Theoretical Exercises

Excercise 3.1: (Theoretical) Pseudonymization of data

Consider the following dataset and perform a hash+salt pseudonymization on the value name. For this, use the following hash function:

$$h(s) = \sum_i alph(s_i) \mod 13$$

where $alph(s_i)$ denotes the position in the alphabet of the i^{th} character in the string, e.g. $alph(A) = 1$ and $alph(K) = 11$.

name	height	shoe-size
Franz	165	40
Antje	170	39
Alex	174	42

Excercise 3.2: (Theoretical) Frequencies of data

At the airport check-in desk, the weight of luggage of the passengers is measured. The resulting values are given in the table below.

22 44 11 19 21 17 17 11 11 19 22 17

The weights are categorized in the following categories:

- light: less than 15 kg
- normal: between 15 and 20 kg
- overweight: more than 20 kg

a) Compute the absolute and relative frequencies of all categories, and create a bar-chart and histogram.

b) Create an empirical distribution function and answer the following question: What is the proportion of weights that are less than 18 kg or more than 23 kg?

Practical Exercises

Excercise 3.3: (Practical) Frequencies of data

Consider the Ames housing dataset given in Exercise 2. Load the dataset using a common library, then complete the following exercises:

1. Compute the frequencies for each discrete variable.
2. Plot the frequencies using at least two different methods (e.g. pie and bar chart)
3. Create a histogram to visualize the distribution of the variable ”SalePrice”

Hint: The library Pandas offers many functionalities for plotting and computing the frequencies of values.