Fachhochschule
Dortmund
University of Applied Sciences and Arts

we
focus
on
students

Data Science

1. Exercise / Practical

Theoretical Exercises

Excercise 1.1: (Theoretical) Distinguishing between data types

Consider the following data set about sales prices in Ames (USA). The presented data is part of a larger dataset [a]:

Lot Area	NBHD	Type	Qual	Cond	Built	1st Flr over Lot Area	1st Flr SF	Mo Sold	SalePrice
11160	NAmes	1Fam	7	5	1968	0,19	2110	4.2010	244000
4920	StoneBr	TwnhsE	8	5	2001	0,27	1338	4.2010	213500
7500	Gilbert	1Fam	7	5	1999	0,14	1028	6.2010	189000
7980	Gilbert	1Fam	6	7	1992	0,15	1187	3.2010	185000
12537	NAmes	1Fam	5	6	1971	0,09	1078	4.2010	149900
1680	BrDale	Twnhs	5	5	1971	0,31	525	3.2010	105500
2280	NPkVill	Twnhs	7	6	1975	0,37	836	6.2010	120000
11520	NridgHt	1Fam	9	5	2005	0,15	1698	6.2010	275000
10171	NridgHt	1Fam	7	5	2004	0,15	1535	3.2010	214000
7132	NridgHt	TwnhsE	8	5	2006	0,19	1370	4.2010	205000
3203	Blmngtn	TwnhsE	7	5	2006	0,36	1145	1.2010	160000
13300	Gilbert	1Fam	7	5	2004	0,06	744	6.2010	184500

a) Discuss, based on the given data, what is basic population, sample, statistical unit, variable and value?

b) Decide for every variable to which data category its values belong to.

Hint: It could be helpful to do some research work.

[a] https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

Excercise 1.2: (Theoretical) Comparison CSV and JSON

Compare the data interchange formats CSV and JSON regarding the following categories:

a) Readability and simpleness for human

b) Support of hierarchical data structures

c) Efficiency in file-size of larger data

Excercise 1.3: (Theoretical) Problems of sales data

Consider the following situation: You are investigating a dataset containing global sales data of a company. The dataset contains information on the goods, prices, dates etc.

a) Discuss at least three potential problems, this dataset could have.

b) How could these problems be addressed to raise the trust in the data? Which dimension of data quality do these problems adress?

Practical Exercises

Excercise 1.4: (Practical) CSV to JSON

You can find a CSV file in the Ilias course room called *AmesHousing.csv*. Your task is to transform the CSV file to the JSON format. For this, create a program which does the following:

a) Loading the data into a proper structure, e.g. list of lists.

b) Transform it into a structure similar to a JSON file, e.g. list of dictionary.

c) Save the transformed data into a JSON file.

Hint: You should parse the file yourself, i.e. you are not allowed to use built-in libraries which parse CSV files.

Excercise 1.5: (Practical) Pandas

Consider the CSV file from the previous exercise. Use a library to load the data and print a summary of it.
Hint: A widely used library for working with data in Python is pandas: `https://pandas.pydata.org`