Data Science 04: Náhodné rozdělení dat na tréninkovou a validační množinu (Random Data Splitting)

Import libraries:

In [1]:
```
# Instalace potřebných knihoven
#%pip instal pandas
#%pip install numpy

#%pip install scipy
#%pip install seaborn

#%pip install scikit-learn
#%pip install matplotlib
#%pip install seaborn

# actual installed version of sklearn
#%pip show scikit-learn
```

In [2]:
```
# Import potřebných knihoven
import pandas as pd
import numpy as np

from sklearn.model_selection import train_test_split

import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
```

In [3]:
```
# Soubor je načten a přiřazen do proměnné ‚df‘
path='../../data/01_DataScience/ready_timelaps.csv'
df = pd.read_csv(path)
df.head()
```

Out[3]:

	id	type_brick	type	start_to_verif	verif_to_dest	dest_to_end	total_time
0	1	CORNER	2	6	17	18	41
1	2	HALF	3	3	17	16	36
2	4	BASIC	1	6	14	17	37
3	6	BASIC	1	7	14	14	35
4	9	BASIC	1	8	13	24	45

Náhodné rozdělení dat

In [4]:
```
# Nastavení náhodného semene (random seed) pro reprodukovatelné rozdělení dat
user_seed = 122

# Náhodné rozdělení dat na tréninkovou a validační množinu (80/20)
df_train, df_val = train_test_split(df, test_size=0.2, random_state=user_seed)
```

In [5]:
```
df_train.head()
```

Out[5]:

	id	type_brick	type	start_to_verif	verif_to_dest	dest_to_end	total_time
95	119	BASIC	1	8	16	33	57
12	23	BASIC	1	8	21	18	47
33	47	BASIC	1	8	16	22	46
87	111	BASIC	1	6	16	14	36
28	41	BASIC	1	7	18	16	41

In [6]:
```
df_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 84 entries, 95 to 26
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   id              84 non-null     int64 
 1   type_brick      84 non-null     object
 2   type            84 non-null     int64 
 3   start_to_verif  84 non-null     int64 
 4   verif_to_dest   84 non-null     int64 
 5   dest_to_end     84 non-null     int64 
 6   total_time      84 non-null     int64 
dtypes: int64(6), object(1)
memory usage: 5.2+ KB
```

In [7]:
```
df_val.head()
```

Out[7]:

	id	type_brick	type	start_to_verif	verif_to_dest	dest_to_end	total_time
58	74	BASIC	1	7	15	12	34
22	35	HALF	3	6	21	9	36
3	6	BASIC	1	7	14	14	35
56	72	BASIC	1	9	13	16	38
49	64	BASIC	1	10	10	17	37

In [8]:
```
df_val.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 22 entries, 58 to 82
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   id              22 non-null     int64 
 1   type_brick      22 non-null     object
 2   type            22 non-null     int64 
 3   start_to_verif  22 non-null     int64 
 4   verif_to_dest   22 non-null     int64 
 5   dest_to_end     22 non-null     int64 
 6   total_time      22 non-null     int64 
dtypes: int64(6), object(1)
memory usage: 1.4+ KB
```

Export datové sady (train + valid) do formátu CSV

In [9]:
```
df_train.to_csv('../../data/06_AI/train/train_timelaps.csv', index=False)
df_val.to_csv('../../data/06_AI/val/valid_timelaps.csv', index=False)
```

Autor / Organizace / Datum

Vjačeslav Usmanov, ČVUT v Praze, Fakulta stavební

Přehled změn

Datum (YYYY-MM-DD)	Verze	Autor změny	Popis změny
2026-01-21	1.1	Vjačeslav Usmanov	added DS_04_Data_Splitting.ipynb
2026-02-12	1.2	Vjačeslav Usmanov	changed DS_04_Data_Splitting.ipynb