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/raw/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	18	16	40
1	2	HALF	3	4	16	18	38
2	4	BASIC	1	6	14	16	36
3	6	BASIC	1	6	14	16	36
4	9	BASIC	1	8	12	24	44

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
12	24	BASIC	1	10	16	18	44
64	89	END	4	10	14	30	54
39	55	BASIC	1	20	14	8	42
61	78	BASIC	1	7	12	14	33
83	109	BASIC	1	8	14	14	36

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

```
<class 'pandas.core.frame.DataFrame'>
Index: 81 entries, 12 to 26
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   id              81 non-null     int64 
 1   type_brick      81 non-null     object
 2   type            81 non-null     int64 
 3   start_to_verif  81 non-null     int64 
 4   verif_to_dest   81 non-null     int64 
 5   dest_to_end     81 non-null     int64 
 6   total_time      81 non-null     int64 
dtypes: int64(6), object(1)
memory usage: 5.1+ KB
```

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

Out[7]:

	id	type_brick	type	start_to_verif	verif_to_dest	dest_to_end	total_time
67	92	BASIC	1	6	16	15	37
85	111	BASIC	1	8	14	14	36
3	6	BASIC	1	6	14	16	36
34	49	BASIC	1	6	18	24	48
17	29	BASIC	1	6	18	20	44

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

```
<class 'pandas.core.frame.DataFrame'>
Index: 21 entries, 67 to 49
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   id              21 non-null     int64 
 1   type_brick      21 non-null     object
 2   type            21 non-null     int64 
 3   start_to_verif  21 non-null     int64 
 4   verif_to_dest   21 non-null     int64 
 5   dest_to_end     21 non-null     int64 
 6   total_time      21 non-null     int64 
dtypes: int64(6), object(1)
memory usage: 1.3+ KB
```

Export datové sady (train + valid) do formátu CSV

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