

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/clean_timelaps.csv'
        df = pd.read_csv(path)
        df.head()
```

```
Out[3]:
```

	id	x	y	z	time	delay	type_delay	total_time
0	1	220	95	0	44	0	0	44
1	2	220	252	0	35	0	0	35
2	3	220	440	0	36	109	1	145
3	4	220	690	0	36	0	0	36
4	5	220	940	0	34	0	0	34

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 (60/40)
        df_train, df_val = train_test_split(df, test_size=0.4, random_state=user_seed)
```

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

Out[5]:

	id	x	y	z	time	delay	type_delay	total_time
145	150	1315	220	1000	29	0	0	29
70	75	220	1190	500	33	0	0	33
231	239	220	940	2000	35	6	3	41
191	199	1315	220	1500	36	0	0	36
46	51	3690	220	250	50	0	0	50

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

```
<class 'pandas.core.frame.DataFrame'>
Index: 161 entries, 145 to 187
Data columns (total 8 columns):
#   Column      Non-Null Count  Dtype
---  -
0   id           161 non-null    int64
1   x            161 non-null    int64
2   y            161 non-null    int64
3   z            161 non-null    int64
4   time         161 non-null    int64
5   delay        161 non-null    int64
6   type_delay   161 non-null    int64
7   total_time   161 non-null    int64
dtypes: int64(8)
memory usage: 11.3 KB
```

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

Out[7]:

	id	x	y	z	time	delay	type_delay	total_time
12	13	220	2940	0	32	0	0	32
72	77	220	1690	500	33	23	2	56
212	220	2190	220	1750	35	0	0	35
100	105	252	220	750	53	0	0	53
40	45	2190	220	250	45	0	0	45

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

```
<class 'pandas.core.frame.DataFrame'>
Index: 108 entries, 12 to 3
Data columns (total 8 columns):
#   Column      Non-Null Count  Dtype
---  -
0   id           108 non-null    int64
1   x            108 non-null    int64
2   y            108 non-null    int64
3   z            108 non-null    int64
4   time         108 non-null    int64
5   delay        108 non-null    int64
6   type_delay   108 non-null    int64
7   total_time   108 non-null    int64
dtypes: int64(8)
memory usage: 7.6 KB
```

Export datové sady (train + valid) do formátu CSV

```
In [9]: df_train.to_csv('../data/01_DataScience/final_timelaps.csv', index=False)

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