Štatistické testovanie hypotéz

URL https://github.com/FIIT-IAU/

Chceme overiť, či má počet valcov motora vplyv na spotrebu.

```
import pandas as pd
import matplotlib
import seaborn as sns
import statsmodels.api as sm
import statsmodels.stats.api as sms
import scipy.stats as stats
from sklearn import preprocessing

cars = pd.read_csv('data/auto-mpg.data',
                   delim_whitespace=True,
                   names = ['mpg', 'cylinders',
'displacement','horsepower',
                            'weight', 'acceleration', 'model_year',
'origin', 'name'],
                   na_values='?')
cars.head()
```

```
C:\Users\matus\AppData\Local\Temp\ipykernel_10072\2311559318.py:1:
FutureWarning: The 'delim_whitespace' keyword in pd.read_csv is
deprecated and will be removed in a future version. Use ``sep='\s+'``
instead
  cars = pd.read_csv('data/auto-mpg.data',
```

```
    mpg  cylinders  displacement  horsepower  weight  acceleration  \
0  18.0          8         307.0       130.0  3504.0          12.0
1  15.0          8         350.0       165.0  3693.0          11.5
2  18.0          8         318.0       150.0  3436.0          11.0
3  16.0          8         304.0       150.0  3433.0          12.0
4  17.0          8         302.0       140.0  3449.0          10.5

   model_year  origin                       name
0          70       1  chevrolet chevelle malibu
1          70       1          buick skylark 320
2          70       1         plymouth satellite
3          70       1              amc rebel sst
4          70       1                ford torino
```

DÚ

a. Overte, či je rozdiel medzi spotrebou 4 a 6-valcových motorov signifikantný.

b. Overte, či je rozdiel medzi spotrebou 4 a 5-valcových motorov signifikantný.

c. Ešte lepší postup je použiť test, ktorý umožňuje otestovať viacero skupín naraz. Zovšeobecnením t-testu pre viacero skupín je **ANOVA (Analysis of variance)**.

Invisible note

Rozdelime

```
mpg4 = cars[cars['cylinders'] == 4]['mpg'].dropna()
mpg5 = cars[cars['cylinders'] == 5]['mpg'].dropna()
mpg6 = cars[cars['cylinders'] == 6]['mpg'].dropna()
```

a)

```
stats.ttest_ind(mpg4, mpg6)

TtestResult(statistic=np.float64(13.718631345338444),
pvalue=np.float64(2.947920641313147e-33), df=np.float64(286.0))
```

t-stat = 13.72 a p-value = 2.95e-33 indikuje signifikantny rozdiel v mpg pretoze t-stat je dost vysoko

```
stats.ttest_ind(mpg4, mpg5)

TtestResult(statistic=np.float64(0.5751716277934285),
pvalue=np.float64(0.5658059728615887), df=np.float64(205.0))
```

t-stat = 0.58 a p-value 0.57 (p > 0.05)

```
stats.f_oneway(mpg4, mpg5, mpg6)

F_onewayResult(statistic=np.float64(93.17980692803185),
pvalue=np.float64(6.208811135418108e-32))
```

ak je p-value mala a f statistic velka, znamena to ze je tam statisticky signifikantny rozdiel medz skupinami

```

```