

Explorační analýza dat (EDA)

Explorační analýza dat (Exploratory Data Analysis - EDA) je přístup k analýze datových sad s cílem shrnout jejich hlavní charakteristiky, často pomocí vizuálních metod. Místo abychom hned aplikovali složité modely, snažíme se nejprve pochopit, co nám data říkají, jakou mají strukturu a kde se nacházejí chyby.

1. Hlavní cíle EDA

- **Odhalení struktury:** Jaké jsou datové typy? Kolik máme záznamů a kolik rysů?
- **Identifikace chyb:** Kde chybí data? Jsou v datech nesmyslné hodnoty (např. záporná cena)?
- **Hledání vztahů:** Které proměnné spolu korelují? Ovlivňuje „věk“ přímo „příjem“?
- **Kontrola předpokladů:** Mají data normální rozdělení (Gaussovo)? Obsahují [odlehle hodnoty](#)?

2. Klíčové techniky a nástroje

A. Deskriptivní statistika (Číselný pohled)

Základní přehled o každém sloupci v datech:

- **Míry centrální tendence:** Průměr (mean), medián, modus.
- **Míry variability:** Rozptyl, směrodatná odchylka, rozsah (min-max).
- **Kvartily:** 25., 50. a 75. percentil (klíčové pro tvorbu boxplotů).

B. Univariační analýza (Jeden sloupec)

Zkoumáme rozdělení jedné proměnné.

- **Histogram:** Ukazuje frekvenci hodnot a tvar rozdělení.
- **Boxplot:** Pomáhá vizualizovat rozptyl a identifikovat outliery.

C. Bivariační a multivariační analýza (Vztahy)

Zkoumáme, jak se proměnné ovlivňují navzájem.

- **Scatter plot (Bodový graf):** Ukazuje vztah mezi dvěma numerickými proměnnými.
- **Korelační matice (Heatmap):** Barevné schéma zobrazující sílu vztahu mezi všemi sloupci. (Hodnoty od -1 do 1).
- **Pairplot:** Matice grafů pro všechny kombinace proměnných v datasetu.

3. Typický postup EDA

Krok	Akce
1. Náhled	<code>`df.head()`, `df.info()`, `df.describe()`</code>
2. Čištění	Identifikace chybějících hodnot a duplicit.
3. Vizualizace	Vykreslení histogramů pro všechny číselné sloupce.
4. Korelace	Zjištění, které rysy (features) jsou nejdůležitější pro cíl.
5. Outliers	Kontrola extrémních hodnot, které by mohly zkreslit model.

4. Nástroje pro EDA v Pythonu

Kromě standardních knihoven existují i nástroje pro „automatické EDA“:

- **Pandas:** Základní manipulace.
- **Matplotlib / Seaborn:** Vizualizace.
- **Pandas Profiling / Sweetviz:** Knihovny, které vygenerují kompletní HTML report o datech jedním příkazem.

```
import seaborn as sns
import matplotlib.pyplot as plt

# Rychlá vizualizace vztahů v datasetu "iris"
df = sns.load_dataset('iris')
sns.pairplot(df, hue='species')
plt.show()

# Korelační heatmapa
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
```

Související články:

- [Základy čištění dat](#)
- [Detekce odlehlých hodnot](#)
- [Škálování dat](#)

Tagy: *ml data_science eda statistics visualization python*

From:

<https://www.serviceit.cz/> - IT ENCYKLOPEDIE

Permanent link:

<https://www.serviceit.cz/doku.php?id=it:ml:eda>

Last update: **2026/01/02 12:48**

