# Projekt: Odhaľovanie a ošetrovanie odľahlých hodnôt (Heart Disease) – verzia pre 3. ročník

Meno študenta: Samuel Weis  
Trieda: 3. ročník  
Dátum: 16.9 2025

## 🎯 Cieľ projektu

Na dátach o srdcových ochoreniach ukážeme, ako nájsť a ošetriť odľahlé hodnoty (outliers) v dvoch číselných stĺpcoch:  
• trestbps – pokojový systolický tlak  
• chol – cholesterol  
  
Použijeme tri jednoduché metódy: Z‑score, Percentily a IQR. Porovnáme grafy pred/po a napíšeme krátky záver.

## 🗂 Dataset

Súbor: heart.csv (verzia UCI / Kaggle Heart Disease)  
Poznámka: Do priečinka s kódom si ulož „heart.csv“. Nepoužívaj dlhé absolútne cesty typu C:\Users\...

## ⚙️ Príprava prostredia

Použijeme pandas, numpy, matplotlib, seaborn.

# (ak treba nainštalovať v termináli)  
# pip install pandas numpy matplotlib seaborn  
  
import numpy as np  
import pandas as pd  
import matplotlib.pyplot as plt  
import seaborn as sns  
  
sns.set\_style('whitegrid')

## 🪜 Postup krok za krokom (skopíruj do kódu)

1. KROK 1 – Načítaj CSV a vyber stĺpce:

# Súbor 'heart.csv' ulož do rovnakého priečinka ako tento .py/.ipynb  
df = pd.read\_csv('heart.csv') # napr. z Kaggle: age, sex, trestbps, chol, ...  
print(df.columns)  
  
cols = ['trestbps', 'chol']  
new\_df = df[cols].copy()  
  
print(new\_df.describe())

1. KROK 2 – Rýchle grafy (boxplot + histogram):

fig, axes = plt.subplots(2, 2, figsize=(12, 8))  
sns.boxplot(data=new\_df, y='trestbps', ax=axes[0,0], color='#0180CC')  
sns.histplot(data=new\_df, x='trestbps', kde=True, ax=axes[0,1], color='#0180CC')  
sns.boxplot(data=new\_df, y='chol', ax=axes[1,0], color='#0180CC')  
sns.histplot(data=new\_df, x='chol', kde=True, ax=axes[1,1], color='#0180CC')  
plt.tight\_layout()  
plt.show()

1. KROK 3 – Pomocné funkcie (nájdi a „capni“ odľahlé hodnoty):

# 3 metódy prahov  
def zscore\_bounds(s):  
 m, sd = s.mean(), s.std()  
 return (m - 3\*sd), (m + 3\*sd)  
  
def percentile\_bounds(s):  
 return s.quantile(0.01), s.quantile(0.99)  
  
def iqr\_bounds(s):  
 q1, q3 = s.quantile(0.25), s.quantile(0.75)  
 iqr = q3 - q1  
 return (q1 - 1.5\*iqr), (q3 + 1.5\*iqr)  
  
# univerzálne „capovanie“ (ohraničenie do intervalu)  
def cap\_series(s, lower, upper):  
 return s.clip(lower=lower, upper=upper)  
  
# aplikácia na jeden stĺpec a jednu metódu  
def apply\_capping(df\_col, method\_func, method\_name):  
 lo, hi = method\_func(df\_col)  
 capped = cap\_series(df\_col, lo, hi)  
 return capped.rename(df\_col.name + f'\_capped\_{method\_name}')

1. KROK 4 – Aplikuj metódy na oba stĺpce:

results = new\_df.copy()  
for col in ['trestbps','chol']:  
 results[col + '\_capped\_zscore'] = apply\_capping(new\_df[col], zscore\_bounds, 'zscore')  
 results[col + '\_capped\_percentile'] = apply\_capping(new\_df[col], percentile\_bounds, 'percentile')  
 results[col + '\_capped\_iqr'] = apply\_capping(new\_df[col], iqr\_bounds, 'iqr')  
  
print(results.filter(like='trestbps').describe())  
print(results.filter(like='chol').describe())

1. KROK 5 – Vizualizácie pred vs. po:

def kde\_compare(col):  
 plt.figure(figsize=(10,5))  
 sns.kdeplot(results[col], label='Original', linewidth=2)  
 sns.kdeplot(results[col + '\_capped\_zscore'], label='Z-score', linewidth=2)  
 sns.kdeplot(results[col + '\_capped\_percentile'], label='Percentile', linewidth=2)  
 sns.kdeplot(results[col + '\_capped\_iqr'], label='IQR', linewidth=2)  
 plt.title(f'{col}: hustota pred/po capovaní')  
 plt.legend()  
 plt.show()  
  
def box\_compare(col):  
 melted = results[[  
 col,  
 col + '\_capped\_zscore',  
 col + '\_capped\_percentile',  
 col + '\_capped\_iqr'  
 ]].melt(var\_name='Method', value\_name=col)  
 plt.figure(figsize=(12,5))  
 sns.boxplot(data=melted, x='Method', y=col)  
 plt.title(f'{col}: boxplot porovnanie')  
 plt.xticks(rotation=15)  
 plt.show()  
  
kde\_compare('trestbps')  
box\_compare('trestbps')  
kde\_compare('chol')  
box\_compare('chol')

1. KROK 6 – Krátky záver do Wordu (3–5 viet):

• Koľko a aké veľké odľahlé hodnoty sme videli v trestbps a chol?  
• Ktorá metóda capovania pôsobila najrozumnejšie (podľa grafov)?  
• Ako sa zmenila medián/rozptyl po capovaní?  
• Ako by si tieto upravené dáta použil ďalej (napr. jednoduchý model)?

## 📝 Čo odovzdať

1) Word dokument: Úvod (1 odstavec o datasete), Postup (stručne), Grafy (printscreen), Záver (3–5 viet).  
2) .py alebo .ipynb súbor s kódom.

## 💡 Tipy a časté chyby

• Cestu k súboru píš relatívne: 'heart.csv'. Absolútne cesty typu C:\Users\... často zlyhávajú.  
• Ak neexistujú stĺpce 'trestbps' alebo 'chol', skontroluj názvy (print(df.columns)).  
• Ak grafy nevidia, skontroluj, či kód dobehol bez chyby.