Analýza atributů Pokémonů

Julie Vondráčková

29. 6. 2024

1 Úvod

Jako svou statistickou práci jsem si vybrala analyzovat data o Pokémonech z tohoto odkazu:
Kaggle datasheet.
Má analýza zkoumá souvislost mezi typy Pokémonů a jejich výškou, váhou a útokem. Používáme data z datasetu Pokémonů dostupného na Kaggle.
K vizualizaci a výpočtům jsem použila Python 3.12 a tyto knihovny:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import f_oneway, shapiro
```

2 Popisná tabulka dat

Popisná tabulka výšky, váhy a útoku všech 781 Pokémonů v datasetu:

Statistika	Výška (m)	Váha (kg)	Útok
Průměr	1.16	61.38	77.69
Std	1.08	109.35	32.22
Min	0.10	0.10	5.00
25%	0.60	9.00	55.00
50%	1.00	27.30	75.00
75%	1.50	64.80	100.00
Max	14.50	999.90	185.00

Data v tabulce jsem získala pomocí tohoto kódu:

```
relevantni_sloupce = ['type1', 'height_m', 'weight_kg', 'attack']
pokemon_data_cleaned = pokemon_data.dropna(subset=['height_m', '
    weight_kg'])

popisna_statistika = pokemon_data_cleaned[relevantni_sloupce].
    describe()

count = popisna_statistika.loc['count']
mean = popisna_statistika.loc['mean']
std = popisna_statistika.loc['std']
min_val = popisna_statistika.loc['min']
q25 = popisna_statistika.loc['25%']
q50 = popisna_statistika.loc['50%']
q75 = popisna_statistika.loc['75%']
max_val = popisna_statistika.loc['max']
```

3 Vizualizace

3.1 Rozdělení jednotlivých vlastností

Nejprve jsem si vykreslila histogramy výšky (v metrech), váhy (v kilogramech) a útoku.

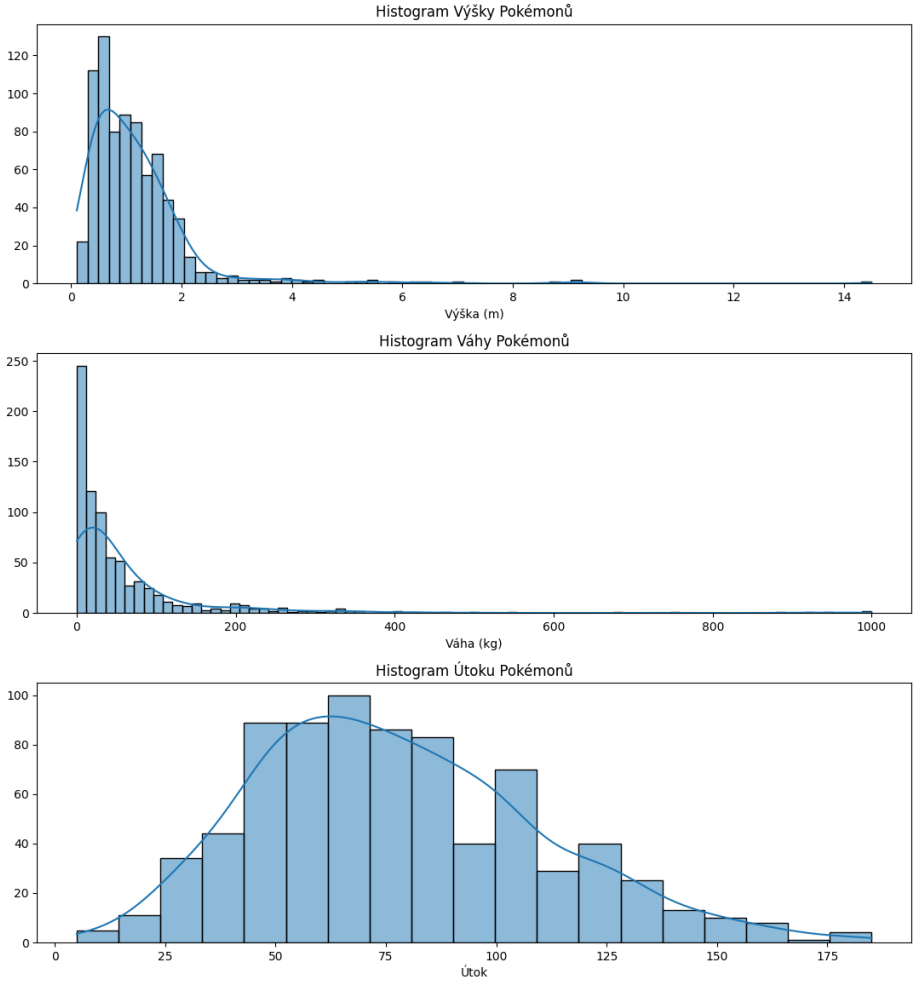

Figure 1: Histogramy výšky, váhy a útoku Pokémonů

Popis: Na histogramu výšky je patrné, že většina Pokémonů má výšku mezi 0.5 a 2 metry. Váha Pokémonů je velmi rozmanitá, většina z nich váží do 100 kg, ale jsou zde i extrémní případy vážící až 999.9 kg. Histogram útoku ukazuje, že útoky jsou poměrně rovnoměrně rozloženy, s průměrem kolem 77.

3.2 Distribuce jednotlivých vlastností podle typu

Vykreslila jsem boxploty pro výšku, váhu a útok podle typu Pokémona.

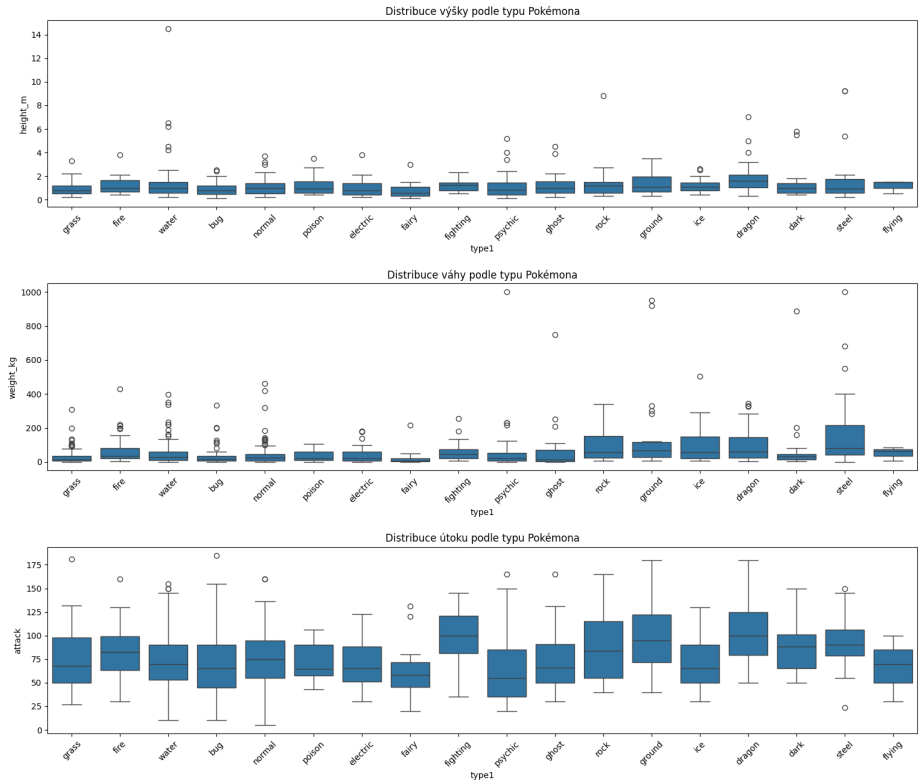

Figure 2: Distribuce výšky, váhy a útoku podle typu Pokémona

Popis: Boxploty ukazují rozložení výšky, váhy a útoku Pokémonů podle jejich typu. Je vidět, že některé typy (např. Dragon, Steel) mají větší rozptyl v hodnotách váhy a útoku. Výška Pokémonů je relativně konzistentní napříč různými typy.

3.3 Scatter Ploty

Scatter ploty byly vytvořeny pro vizualizaci vztahů mezi výškou a váhou, výškou a útokem a váhou a útokem, barevně odlišeny podle typu Pokémona.

Figure 3: Scatter ploty zobrazující vztahy mezi výškou, váhou a útokem Pokémonů

Popis: Scatter ploty ukazují, že existuje pozitivní korelace mezi výškou a váhou Pokémonů, což je očekávané. Vztah mezi výškou a útokem a mezi váhou a útokem není tak výrazný, ale stále lze vidět, že těžší a vyšší Pokémoni mají tendenci mít vyšší útok.

4 Statistická analýza

4.1 ANOVA (ANalysis Of VAriance) Testy

ANOVA testy byly provedeny pro určení, zda existují statisticky významné rozdíly ve výšce, váze a útoku mezi různými typy Pokémonů. Výsledky jsou následující:

Atribut	F-Statistika	p-hodnota
Výška	2.50	7.07e-04
Váha	5.20	5.94e-11
Útok	5.49	9.76e-12

Popis: P-hodnoty ukazují, že existují významné rozdíly ve výšce, váze a útoku mezi různými typy Pokémonů. To znamená, že typ Pokémona má vliv na jeho fyzické a bojové vlastnosti. Konkrétně, typy jako Dragon a Steel mají výrazné rozdíly ve srovnání s ostatními typy.

4.2 Test normality

Pro testování normality distribuce výšky, váhy a útoku byl použit Shapiro-Wilk test. Výsledky jsou následující:

Atribut	W-statistika	p-hodnota
Výška	0.76	1.25e-31
Váha	0.29	0.00e+00
Útok	0.97	1.43e-13

Popis: Výsledky Shapiro-Wilk testu naznačují, že distribuce výšky, váhy a útoku Pokémonů se významně liší od normální distribuce. To je patrné zejména u váhy, kde je p-hodnota extrémně nízká, což potvrzuje, že váhy Pokémonů nejsou normálně rozloženy.

Shapiro–Wilkův test

Shapiro–Wilkův test je statistický test normality, který byl představen Samuel Shapirovem a Martinem Wilkem v roce 1965. Jeho cílem je zjistit, zda daný soubor dat pochází z normálního rozdělení.

Definice

Testová statistika W je definována jako:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

kde:

- $x_{(i)}$ jsou uspořádané vzestupně (od nejmenšího k největšímu) hodnoty ve vzorku,
- a_i jsou konstanty závislé na očekávaných hodnotách a kovariancích uspořádaných statistických hodnot normálního rozdělení,
- x_i jsou původní hodnoty ve vzorku,
- $\bar{x}$ je průměr hodnot ve vzorku,
- n je počet pozorování ve vzorku.

Interpretace výsledků

Hodnota testové statistiky W se pohybuje mezi 0 a 1. Pokud je hodnota W nízká, znamená to, že data pravděpodobně nepocházejí z normálního rozdělení. Kritické hodnoty pro různé hladiny významnosti jsou tabulkově dostupné, a podle těchto hodnot se určuje, zda nulovou hypotézu (data pocházejí z normálního rozdělení) zamítneme nebo ne.

Použití

Shapiro–Wilkův test se často používá v oblastech, kde je třeba ověřit předpoklad normality před provedením dalších statistických analýz, jako jsou regresní analýzy nebo analýzy rozptylu (ANOVA).

5 Závěr

Na základě provedené analýzy mohu potvrdit, že existují statisticky významné rozdíly ve výšce, váze a útoku mezi různými typy Pokémonů. Tyto rozdíly byly identifikovány pomocí ANOVA testů a potvrzeny Shapiro-Wilk testy normality. Vizualizace v podobě histogramů, boxplotů a scatter plotů poskytují jasný pohled na rozložení a vztahy mezi těmito atributy.

6 Použitý kód

Níže uvádím použitý kód v jazyce Python pro načtení dat, výpočet popisné statistiky, vytvoření vizualizací a provedení statistických testů.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import f_oneway, shapiro

# Načtení datasetu
cesta_k_souboru = "path_to_your_pokemon.csv"
pokemon_data = pd.read_csv(cesta_k_souboru)

# Příprava dat
relevantni_sloupce = ['type1', 'height_m', 'weight_kg', 'attack']
pokemon_data_cleaned = pokemon_data.dropna(subset=['height_m', '
    weight_kg'])

# Popisná statistika
popisna_statistika = pokemon_data_cleaned[relevantni_sloupce].
    describe()

# Box Ploty
plt.figure(figsize=(18, 12))

# Výška podle typu
plt.subplot(3, 1, 1)
sns.boxplot(x='type1', y='height_m', data=pokemon_data_cleaned)
plt.title('Distribuce výšky podle typu Pokémona')
plt.xticks(rotation=45)

# Váha podle typu
plt.subplot(3, 1, 2)
sns.boxplot(x='type1', y='weight_kg', data=pokemon_data_cleaned)
plt.title('Distribuce váhy podle typu Pokémona')
plt.xticks(rotation=45)

# Útok podle typu
plt.subplot(3, 1, 3)
sns.boxplot(x='type1', y='attack', data=pokemon_data_cleaned)
plt.title('Distribuce útoku podle typu Pokémona')
plt.xticks(rotation=45)

plt.tight_layout()
plt.show()

# Scatter Plots with a single legend
fig, axes = plt.subplots(3, 1, figsize=(18, 12))

# Výška vs. Váha
scatter1 = sns.scatterplot(x='height_m', y='weight_kg', hue='type1'
    , data=pokemon_data_cleaned, palette='tab10', ax=axes[0])
axes[0].set_title('Výška vs. Váha podle typu Pokémona')
axes[0].set_xlabel('Výška (m)')
axes[0].set_ylabel('Váha (kg)')
handles, labels = scatter1.get_legend_handles_labels()
```

```
scatter1.legend_.remove()

# Výška vs. Útok
scatter2 = sns.scatterplot(x='height_m', y='attack', hue='type1',
    data=pokemon_data_cleaned, palette='tab10', legend=False, ax=
    axes[1])
axes[1].set_title('Výška vs. Útok podle typu Pokémona')
axes[1].set_xlabel('Výška (m)')
axes[1].set_ylabel('Útok')

# Váha vs. Útok
scatter3 = sns.scatterplot(x='weight_kg', y='attack', hue='type1',
    data=pokemon_data_cleaned, palette='tab10', legend=False, ax=
    axes[2])
axes[2].set_title('Váha vs. Útok podle typu Pokémona')
axes[2].set_xlabel('Váha (kg)')
axes[2].set_ylabel('Útok')

# Add a single legend to the upper right corner of the figure
fig.legend(handles, labels, loc='upper right', bbox_to_anchor=(1,
    1), ncol=1)

plt.tight_layout()
plt.show()

# Histogramy
plt.figure(figsize=(18, 12))

# Histogram Výšky
plt.subplot(3, 1, 1)
sns.histplot(pokemon_data_cleaned['height_m'], kde=True)
plt.title('Histogram Výšky Pokémonů')
plt.xlabel('Výška (m)')
plt.ylabel('Frekvence')

# Histogram Váhy
plt.subplot(3, 1, 2)
sns.histplot(pokemon_data_cleaned['weight_kg'], kde=True)
plt.title('Histogram Váhy Pokémonů')
plt.xlabel('Váha (kg)')
plt.ylabel('Frekvence')

# Histogram Útoku
plt.subplot(3, 1, 3)
sns.histplot(pokemon_data_cleaned['attack'], kde=True)
plt.title('Histogram Útoku Pokémonů')
plt.xlabel('Útok')
plt.ylabel('Frekvence')

plt.tight_layout()
plt.show()

# ANOVA Testy
anova_vyska = f_oneway(*[group['height_m'].dropna() for name, group
     in pokemon_data_cleaned.groupby('type1')])
anova_vaha = f_oneway(*[group['weight_kg'].dropna() for name, group
     in pokemon_data_cleaned.groupby('type1')])
```

```
anova_utok = f_oneway(*[group['attack'].dropna() for name, group in
     pokemon_data_cleaned.groupby('type1')])

anova_vysledky = {
    'Atribut': ['Výška', 'Váha', 'Útok'],
    'F-Statistika': [anova_vyska.statistic, anova_vaha.statistic,
        anova_utok.statistic],
    'p-hodnota': [anova_vyska.pvalue, anova_vaha.pvalue, anova_utok
        .pvalue]
}

anova_vysledky_df = pd.DataFrame(anova_vysledky)
print("ANOVA Výsledky:")
print(anova_vysledky_df)

# Test normality (Shapiro-Wilk test)
shapiro_vyska = shapiro(pokemon_data_cleaned['height_m'])
shapiro_vaha = shapiro(pokemon_data_cleaned['weight_kg'])
shapiro_utok = shapiro(pokemon_data_cleaned['attack'])

normalita_vysledky = {
    'Atribut': ['Výška', 'Váha', 'Útok'],
    'W-statistika': [shapiro_vyska.statistic, shapiro_vaha.
        statistic, shapiro_utok.statistic],
    'p-hodnota': [shapiro_vyska.pvalue, shapiro_vaha.pvalue,
        shapiro_utok.pvalue]
}

normalita_vysledky_df = pd.DataFrame(normalita_vysledky)
print("Výsledky testu normality (Shapiro-Wilk):")
print(normalita_vysledky_df)
```