Kuncová Iveta

Projekt E-sport, Statistické porovnání korelací v Pythonu

Otázka: Co z ověřovaných ukazatelů ovlivňuje počet hráčů v zemích?

Podkladová data a soubor python v jupiter notebooku: link

https://drive.google.com/drive/folders/1qHz2vEuzCEYq6PzUA9Y6f4PPyj7lSB5T?usp=sharing

Abstrakt:
Dokument statistickými metodami pomocí pythonu zkoumá čtyři subhypotézy A až D. Výsledným zjištěním je, že mezi zkoumanými proměnnými korelace existují.
V tabulce jsou hypotézy seřazeny od nejsilnější korelace po nejslabší.

Hypotéza:	Proměnná 1	Proměnná 2	Velikost korelace
D	Počet hráčů/mil. obyvatel	Prům. procento obyvatel užívajících internet	0.69
C	Počet hráčů/mil. obyvatel	Prům. výše HDP na obyvatele	0.58
A	Počet hráčů	Prům. počet obyvatel	0.37
B	Počet hráčů	Prům. počet obyv. ve věku 15-19 let	0.31

Legenda:

Porovnávaná období všech proměnných: **1997 - 2021**

Počet hráčů = počet hráčů v rozdělení za jednotlivé země světa

Počet hráčů/mil. obyvatel = (Počet hráčů za jednotlivé země světa * 1 000 000) / průměrný počet obyvatel za jednotlivé země světa

Prům. výše HDP na obyvatele = Průměrné HDP na obyvatele za jednotlivé země světa

Prům. počet obyvatel = Průměrná velikost celé populace za jednotlivé země světa

Prům. počet obyv. ve věku 15-19 let = Průměrná velikost populace ve věku 15 až 29 let za jednotlivé země světa. (Hráči e-sportu se většinou pohybují v rozmezí 16 až 25 let.)

Prům. procento obyvatel užívajících internet = Podíl jedinců v populaci užívajících internet za jednotlivé země světa

1. Hypotézy:

H0 = Korelace mezi … neexistuje:

A0)	Počtem hráčů	a	Prům. počtem obyvatel
B0)	Počtem hráčů	a	Prům. počtem obyv. ve věku 15-19 let
C0)	Počtem hráčů/mil. obyvatel	a	Prům. výší HDP na obyvatele
D0)	Počtem hráčů/mil. obyvatel	a	Prům. procentem obyvatel užívajících internet

H1 = Korelace mezi … existuje:

A1)	Počtem hráčů	a	Prům. počtem obyvatel
B1)	Počtem hráčů	a	Prům. počtem obyv. ve věku 15-19 let
C1)	Počtem hráčů/mil. obyvatel	a	Prům. výší HDP na obyvatele
D1)	Počtem hráčů/mil. obyvatel	a	Prům. procentem obyvatel užívajících internet

2. Ověření předpokladů dat (linearita, normální rozložení):

Pro rozhodnutí, kterou metodu porovnávání korelace použít, je žádoucí znát „předpoklady dat“, tj. zda jsou data lineární a v normálním rozložení.

2. 1. Linearita dat:

Linearita dat je pojem, který se používá k popisu vztahu mezi dvěma proměnnými. Pokud je vztah mezi dvěma proměnnými lineární, pak lze vztah vyjádřit přímkou.

Python: seaborn.scatterplot()

Tato metoda vykreslí rozptylový diagram pro dvě proměnné. Pokud jsou data lineární, budou body na rozptylovém diagramu sledovat klesající nebo stoupající čáru, viz obrázek vpravo:

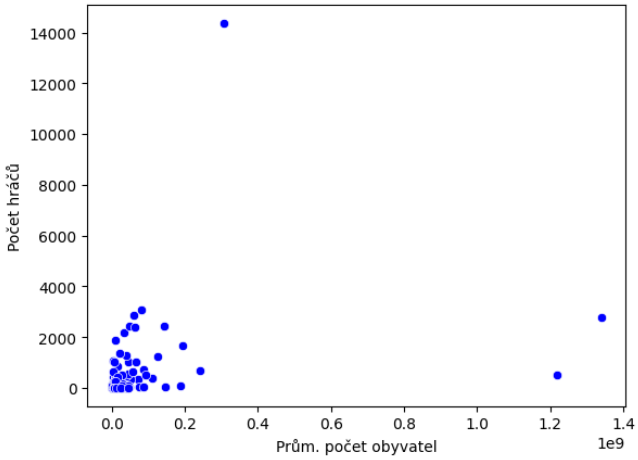
Ukázka lineárního rozložení dat:

Python kód (jupyter notebook):

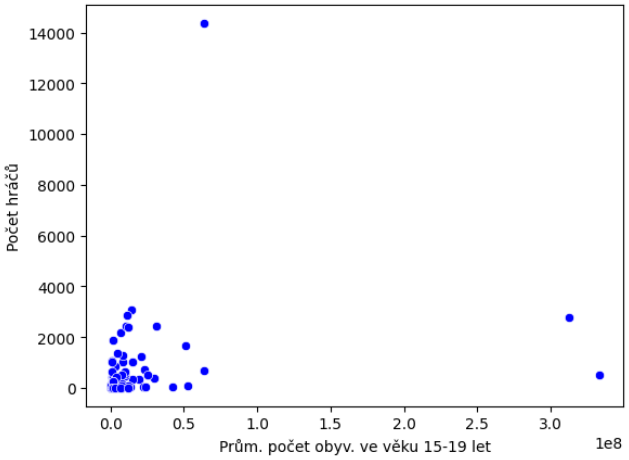
```
import seaborn as sns

sns.scatterplot(data=data, x="Prům. počet obyvatel", y="Počet hráčů", color="blue")

sns.scatterplot(data=data, x="Prům. počet obyv. ve věku 15-19 let", y="Počet hráčů",
color="blue")

sns.scatterplot(data=data, x="Prům. výše HDP na obyvatele", y="Počet hráčů/mil.
obyvatel", color="blue")

sns.scatterplot(data=data, x="Prům. procento obyvatel užívajících internet", y="Počet
hráčů/mil. obyvatel", color="blue")
```

Grafy rozložení dat – test linearity:

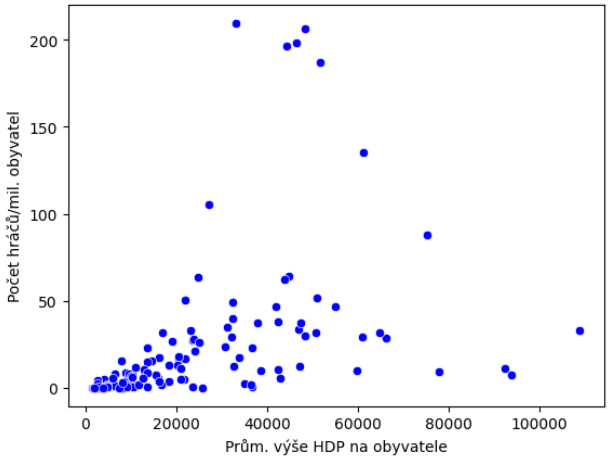
Hypotéza A:



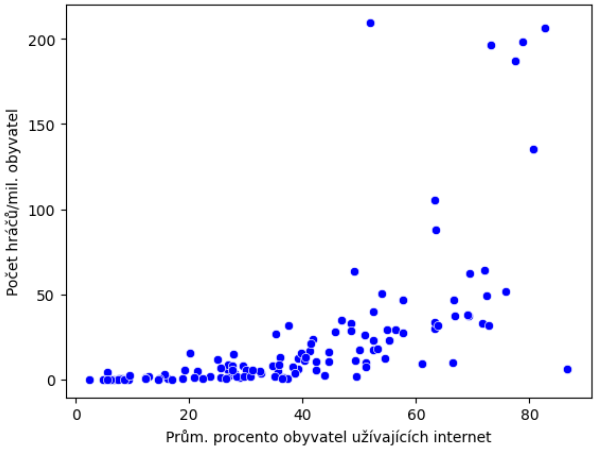
Hypotéza B:



Hypotéza C:



Hypotéza D:



Výsledek (tabulka):

Graf	Linearita
A	ne
B	ne
C	ano
D	ano

Výsledek (shrnutí): ***Ne všechna data jsou jednoznačně lineární.***

2. 2. Normální rozložení:

D'Agostino a Pearson test:

Na základě kombinace zkosení a špičatosti křivky dat testuje to, zda vzorek pochází z normálního rozdělení. Normální rozdělení se vždy testuje pro jeden sloupec tabulky.

P-value: (P-hodnota) vyjadřuje pravděpodobnost, že by se hodnota testového statistického ukazatele, která je alespoň tak velká jako pozorovaná, objevila náhodou, pokud by data pocházela z normální populace.

Pokud je p-value menší než určitý předem stanovený práh (0,05), pak data nepochází z normálního rozdělení.

Python: scipy.normaltest()

```
# Normal distribution test:
# Normal distribution is always tested for one column of a table.
# If the p-value is less than a certain pre-set threshold (0.05), then the data do not come from a normal distribution.
from scipy import stats

print(stats.normaltest(data["Počet hráčů"]))
print(stats.normaltest(data["Počet hráčů/mil. obyvatel"]))
print(stats.normaltest(data["Prům. výše HDP na obyvatele"]))
print(stats.normaltest(data["Prům. počet obyvatel"]))
print(stats.normaltest(data["Prům. počet obyv. ve věku 15-19 let"]))
print(stats.normaltest(data["Prům. procento obyvatel užívajících internet"]))
```

✓ 0.0s

```
NormaltestResult(statistic=225.09172698161382, pvalue=1.3241967721243712e-49)
NormaltestResult(statistic=107.58764136839392, pvalue=4.3415071076559595e-24)
NormaltestResult(statistic=38.56032073757092, pvalue=4.233823426609993e-09)
NormaltestResult(statistic=202.5601007432329, pvalue=1.0342677828014017e-44)
NormaltestResult(statistic=203.0852987523566, pvalue=7.954038670598638e-45)
NormaltestResult(statistic=9.833171528727362, pvalue=0.0073240943909007915)
```

Výsledek (tabulka):

	P-value (vědecký zápis, e-forma) **P-value („nevědecký" zápis)**
Počet hráčů	1,3241967721243712e-49 **0,000000000000000000000000000013241967721243712**
Počet hráčů/mil. obyvatel	4,3415071076559595e-24 **0,000000000000000004341507107655959 5**
Prům. výše HDP na obyvatele	4,233823426609993e-09 **0,000000000042338234266099 93**
Prům. počet obyvatel	1,0342677828014017e-44 **0,000000000000000000000000000000010342677828014017**
Prům. počet obyv. ve věku 15-19 let	7,954038670598638e-45 **0,000000000000000000000000000000079540386705986 38**
Prům. procento obyvatel užívajících internet	- **0,0073240943909007915**

Výsledek (shrnutí): ***Data nemají normální rozdělení****, protože všechny p-hodnoty (nevědecký zápis) jsou menší než 0,05.*

3. Metoda porovnání korelace:

Kendall's Tau (Kendall Rank Correlation Coefficient):

Kendall's Tau (Kendallovo Tau) je neparametrická metrika korelace, která měří směrovou korelaci mezi dvěma posloupnostmi. Nevyžaduje lineární data v normálním rozložení, se kterými tento dokument pracuje.

Python: scipy.kendalltau()
Výsledné ukazatele:
„Statistic" je míra korelace mezi dvěma porovnávanými soubory dat. Používá se k testování hypotézy korelace H0 a H1. Čím je číslo blíže 1 nebo -1, tím je korelace silnější.
Může nabývat hodnot od -1 do 1:

-1 znamená negativní korelaci
0 znamená žádnou korelaci
1 znamená pozitivní korelaci

„P-value" je statistický nástroj, který se používá k testování hypotéz. V tomto případě je nulová hypotéza, že neexistuje žádná asociace mezi dvěma soubory dat (tj. tau = 0). Pokud je p-hodnota menší než určitý předem stanovený práh (často 0,05), pak nulovou hypotézu H0 zamítáme a přijímáme alternativní hypotézu H1, že existuje nějaká asociace mezi dvěma soubory dat. Pokud je p-hodnota větší než tento práh, pak nemůžeme nulovou hypotézu zamítnout.

```
# Correlation analysis:
# statistic - the magnitude of the correlation
# p-value - if it is less than 0.05, it is likely that there exists a real correlation, not just a coincidence
from scipy import stats

print(stats.kendalltau(data["Prům. počet obyv. ve věku 15-19 let"], data["Počet hráčů"]))
print(stats.kendalltau(data["Prům. procento obyvatel užívajících internet"], data["Počet hráčů"]))
print(stats.kendalltau(data["Prům. výše HDP na obyvatele"], data["Počet hráčů/mil. obyvatel"]))
print(stats.kendalltau(data["Prům. počet obyvatel"], data["Počet hráčů/mil. obyvatel"]))
```

✓ 0.0s

```
SignificanceResult(statistic=0.30977218500546366, pvalue=3.4318329221465256e-07)
SignificanceResult(statistic=0.3800806122554475, pvalue=3.971690688428846e-10)
SignificanceResult(statistic=0.5793548387096774, pvalue=9.537725301391349e-22)
SignificanceResult(statistic=-0.14296774193548387, pvalue=0.018054226919800007)
```

Výsledek (tabulky):

Korelace pro hypotézy:	Počet hráčů	
	Statistics	**Pvalue (E-forma)**
A: Prům. počet obyvatel	0.3702218659989922	1.1089236960257302e-09
B: Prům. počet obyv. ve věku 15-19 let	0.30977218500546366	3.4318329221465256e-07
	Počet hráčů/mil. obyvatel	
	Statistics	**Pvalue (E-forma)**
C: Prům. výše HDP na obyvatele	0.5793548387096774	9.537725301391349e-22

Korelace pro hypotézy:	Počet hráčů	
	Statistics	**Pvalue (E-forma)**
D: Prům. procento obyvatel užívajících internet	0.6908387096774193	3.1152834988283753e-30

Hypotézy:

	Existuje korelace?	Velikost korelace	Pravděpodobnost, že je korelace statisticky náhodná
		Statistics	P-value s hodnotami v normálním „nevědeckém“ tvaru
A	Ano, slabá, pozitivní	0.3702218659989922	0.00000000001108923696025730 2
B	Ano, slabá, pozitivní	0.30977218500546366	0.0000034318329221465256
C	Ano, střední, pozitivní	0.5793548387096774	0.000000000000000000000000009537725301391349
D	**Ano, silná, pozitivní**	**0.6908387096774193**	0.000000000000000000000000000031152834988283753

Výsledek:

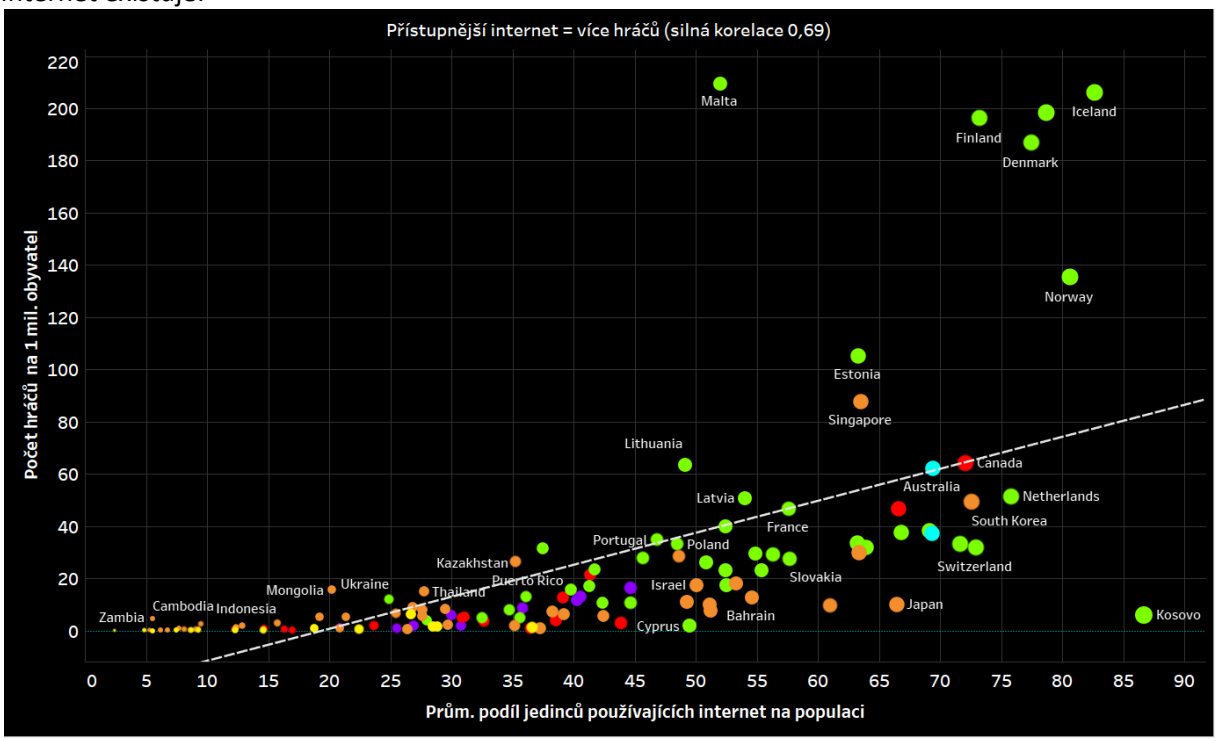
Všechny p-value jsou výrazně menší než 0,05 a tedy korelace mezi uvedenými hodnotami existují, a proto zamítáme Hypotézu H0 a potvrzujeme Hypotézu H1.

Nejsilnější korelace jsou mezi:

Počtem hráčů/1mil. obyvatel a
1) **Podílem jedinců v populaci užívajících internet (0,69, hypotéza D1)**
2) **Průměrným HDP na obyvatele (0,58, hypotéza C1).**

Nejsilnější korelace:

Hypotéza D1: Korelace mezi Počtem hráčů/mil. obyvatel a Prům. procentem obyvatel užívajících internet existuje:



Hypotéza C1: Korelace mezi Počtem hráčů/mil. obyvatel a Prům. výší HDP na obyvatele existuje: