Prije djela 2.1.2 nema programskih kodova

NAPOMENA:U PYTHONU UNUTAR LISTI kvant UNOSIMO SVOJE VLASTITE PODATKE

Pogledajmo kako bismo u Pythonu napravili iste tablice i prikazali podatke grafički. Prvo ćemo prevesti tablicu s originalno dobivenim podacima u Python dataframe, kojeg ćemo nazvati "krgrupa".

Možemo koristiti Pandas biblioteku (Python and DataScience). Najprije importamo cijelu biblioteku i zatim kreiramo dataset data.

```
import pandas as pd

data = {
    'Spol': ["Ž", "M", "M", "Ž", "Ž", "M", "M", "Ž", "M", "Ž"],
    'Krvna_grupa': ["A", "0", "B", "0", "0", "A", "AB", "0", "A", "A"]
}

krgrupa = pd.DataFrame(data)
print(krgrupa)

Pohranjujemo ga u dataframe krgrupa i ispisujemo
```

U Pythonu možemo koristiti Pandas za ostvarivanje funkcionalnosti prebrojavanja broja pojavljivanja svake vrijednosti u pojedinim varijablama.

```
import pandas as pd

data = {
    'Spol': ["Ž", "M", "M", "Ž", "Ž", "M", "M", "Ž", "M", "Ž"],
    'Krvna_grupa': ["A", "0", "B", "0", "0", "A", "AB", "0", "A", "A"]
}

krgrupa = pd.DataFrame(data)
```

Koristimo value_counts() za izračun frekvencije pojavljivanja svake vrijednosti u stupcu 'Spol'

```
spol_counts = krgrupa['Spol'].value_counts()

print(spol_counts)
```

Ovaj Python kod koristi funkciju value_counts() za izračun frekvencije pojavljivanja svake vrijednosti u stupcu 'Spol' u DataFrame-u "krgrupa". Nakon izvođenja koda, dobit ćete rezultat koji prikazuje frekvenciju pojave "M" i "Ž".

U Pythonu možemo dobiti frekvenciju pojavljivanja pojedinih krvnih grupa među danim osobama

```
import pandas as pd

data = {
    'Spol': ["Ž", "M", "M", "Ž", "Ž", "M", "M", "Ž", "M", "Ž"],
    'Krvna_grupa': ["A", "0", "B", "0", "0", "A", "AB", "0", "A", "A"]
}
```

```
krgrupa = pd.DataFrame(data)
```

Tablični pregled po krvnoj grupi

```
krvna_grupa_table = pd.crosstab(krgrupa['Krvna_grupa'], krgrupa['Spol'])

print(krvna_grupa_table)
```

```
Spol       M  Ž
Krvna_grupa
0          1  3
A          2  2
AB         1  0
B          1  0
```

Tablice relativnih frekvencija za krvne grupe dobivamo tako da frekvenciju pojedinih krvnih grupa dijelimo s ukupnim brojem dostupnih podataka. Koristit ćemo biblioteku Pandas

```
import pandas as pd

data = {
    'Krvna_grupa': ["0", "A", "AB", "B"],
}

krgrupa = pd.DataFrame(data)

# Izračun relativnih frekvencija za krvne grupe
relk = krgrupa['Krvna_grupa'].value_counts() / len(krgrupa)

print(relk)
```

Ovaj Python kod koristi funkciju value_counts() i dijeli rezultate s ukupnim brojem dostupnih podataka kako bi dobio relativne frekvencije krvnih grupa.

```
0      0.4
A      0.4
AB     0.1
B      0.1
Name: Krvna_grupa, dtype: float64
```

Operacije nad jedinstvenim elementima objekata

U Pythonu se ponovno koristi Pandas i za operacije nad jedinstvenim elementima objekta, što je posebno korisno kada radimo s tablicama; primjerice možemo ispisati tablicu vertikalno

```
import pandas as pd

data = {
    'Krvna_grupa': ["0", "A", "AB", "B"],
}

krgrupa = pd.DataFrame(data)
```

Koristimo funkciju value_counts() za dobivanje frekvencije svake krvne grupe

```
frekvencije = krgrupa['Krvna_grupa'].value_counts().reset_index()
frekvencije.columns = ['Krvna_grupa', 'freq']

print(frekvencije)
```

Ovaj Python kod koristi value_counts() da dobije frekvenciju svake krvne grupe i zatim koristi reset_index() i columns kako bi formatirao tablicu.

```
  Krvna_grupa  freq
0           0     4
1           A     4
2          AB     1
3           B     1
```

Python

Možemo vizualizirati tablicu frekvencija krvnih grupa koristeći biblioteku matplotlib koja služi za grafički prikaz podataka.

```
import pandas as pd
import matplotlib.pyplot as plt

data = {
    'Krvna_grupa': ["0", "A", "AB", "B"],
}

krgrupa = pd.DataFrame(data)
```

Kreiranje grafikona stupčaste frekvencije

```
plt.figure(figsize=(6, 4))  # Postavljanje veličine grafikona
plt.bar(krgrupa['Krvna_grupa'].value_counts().index,
krgrupa['Krvna_grupa'].value_counts(), color='skyblue')
plt.title("Stupčasti prikaz frekvencije krvnih grupa")  # Naslov grafikona
plt.xlabel("Krvna grupa")  # X-osa
plt.ylabel("Broj osoba")  # Y-osa
```

```
plt.xticks(rotation=0)  # Rotacija oznaka na X-osi
plt.tight_layout()

plt.show()  # Prikaz grafikona
```

S vremena na vrijeme, osim tablice relativnih frekvencija, prikazivanje tablice kumulativnih relativnih frekvencija može biti relevantno. Python koristi pandas za izračun kumulativnih relativnih frekvencija i stvara DataFrame s kumulativnim frekvencijama. Nakon izvođenja koda, dobit ćemo ekvivalentnu tablicu kumulativnih relativnih frekvencija.

```
import pandas as pd

data = {
    'Krvna_grupa': ["0", "A", "AB", "B"],
}

krgrupa = pd.DataFrame(data)

## Izračun kumulativnih relativnih frekvencija
relative_frequencies =
krgrupa['Krvna_grupa'].value_counts(normalize=True).sort_index()
cumulative_frequencies = relative_frequencies.cumsum()

# Dodajemo kumulativne relativne frekvencije u novi stupac 'Cumulative_Freq'
cumulative_frequencies_df = pd.DataFrame(cumulative_frequencies, columns=
['Cumulative_Freq'])

print(cumulative_frequencies_df)
```

```
  Cumulative_Freq
0             0.4
A             0.8
AB            0.9
B             1.0
```

```
NA|   Ime  |  Težina |  Visina |   BMI   |
  |--------|---------|---------|---------|
  |   A    |    60   |   165   | 22.038572|
  |   B    |    80   |   180   | 24.69136 |
  |   C    |    85   |   175   | 27.755104|
  |   D    |    84   |   184   | 24.810965|
  |   E    |    62   |   168   | 21.967126|
  |   F    |    65   |   175   | 21.224497|
  |   G    |    86   |   184   | 25.401708|
  |   H    |    86   |   192   | 23.328999|
  |   I    |    95   |   182   | 28.68011 |
  |   J    |    78   |   178   | 24.61810 |
  |   K    |    70   |   175   | 22.85714 |
  |   L    |    68   |   170   | 23.52941 |

1   1   1    3   1   1   1   2   1
```

Naravno, jasno je da se visina 175 javlja najčešće, ali u ovom slučaju to nije toliko zanimljivo. Kada se radi s kvantitativnim podacima, imamo mnogo više informacija koje možemo koristiti prilikom analize podataka. Prije svega, to uključuje mjeru središnje tendencije i mjeru raspršenosti podataka.

Vratit ćemo se BMI podacima kako bismo definirali navedene pojmove.

U Pythonu možemo provesti analizu kvantitativnih podataka koji se nalaze u Excel datoteci nazvanoj "BMI2.xlsx" na listi "Druga". Glavna svrha je izračunati i prikazati različite informacije o tim podacima. Prvo, koristi se biblioteka Pandas za učitavanje podataka iz navedene Excel datoteke. Nakon toga, prvi stupac (Ime) se izbacuje jer nije potreban za analizu.

Nakon obrade podataka, koristi se funkcija print(kvant.dtypes) kako bi se ispisali tipovi podataka za preostale stupce, što pomaže u razumijevanju strukture podataka.

Zatim se analizira stupac "Visina". Funkcija value_counts() koristi se kako bi se izračunala i ispisala frekvencija (broj pojavljivanja) svake jedinstvene vrijednosti u tom stupcu. Ovaj korak pomaže u dobivanju informacija o tome koliko često se pojedine visine pojavljuju u analiziranim podacima.

```
import pandas as pd

# Učitavanje kvantitativnih podataka iz Excel datoteke
kvant = pd.read_excel("BMI2.xlsx", sheet_name="Druga", header=0, nrows=13)

# Izbacivanje prvog stupca (Ime) jer nije potreban
kvant = kvant.iloc[:, 1:]

# Ispis tipova preostalih podataka
print(kvant.dtypes)

# Prikaz tablice frekvencija visine
visina_counts = kvant['Visina'].value_counts()
print(visina_counts)
```

DIO 2.2 NE SADRŽI NIKAKVE PROGRAMSKE DJELOVE

2.2.1

Aritmetička sredina, mod i medijan

U Pythonu za izračun aritmetičke sredine koristimo ugrađenu funkciju mean iz NumPy-a

```
import numpy as np

x = np.array([24, 5, 9, 12, 22, 2, 89, 4, 13])
result = np.mean(x)
print(result)  # Output: 20.0

# Uklanjanje vrijednosti 89 i ponovno računanje aritmetičke sredine
x = x[x != 89]
```

```
result = np.mean(x)
print(result)  # Output: 11.375
```

Medijan

Medijan je statistička mjera središnje tendencije koja predstavlja srednju vrijednost unutar poretka podataka.

U Pythonu koristimo funkciju median() iz biblioteke NumPy za izračun medijana. Prvo, izračunavamo medijan za originalni niz podataka, a zatim izračunavamo medijan za niz x nakon što smo iz njega izbacili vrijednost 89.

```
import numpy as np

# Primjer izračunavanja medijana
# Prvo, koristimo funkciju median() iz biblioteke NumPy za izračun medijana za cijeli niz podataka
data = [24, 5, 9, 12, 22, 2, 89, 4, 13]
median_value = np.median(data)
print(median_value)  # Output: 12.0

# Ako uklonimo vrijednost 89 iz niza 'x' i ponovno izračunamo medijan, dobivamo:
x = [value for value in x if value != 89]
median_x = np.median(x)
print(median_x)  # Output: 10.5
```

Medijan je koristan za mjerenje središnje tendencije, a posebno je koristan kada imate outliers (ekstremne vrijednosti) u vašim podacima, jer neće biti toliko osjetljiv na te ekstremne vrijednosti kao aritmetička sredina. Medijan je vrijednost koja dijeli vaše podatke na dvije jednake polovice, tako da je polovina vaših podataka manja od medijana, a polovina veća.

Mod je statistička mjera koja predstavlja vrijednost koja se najčešće pojavljuje u nizu brojeva. U Pythonu mod možemo izračunati pomoću biblioteke statistics:

```
import statistics

# Stvaranje uzorka brojeva
nuz = [1, 2, 3, 3, 4, 5, 6, 6, 7, 8, 8, 9]

# Izračunavanje moda niza brojeva
mod_values = statistics.multimode(nuz)
print(mod_values)
```

Ovaj kod koristi funkciju multimode() iz statistics biblioteke kako bi pronašao sve vrijednosti koje se pojavljuju najčešće u nizu. Rezultat će sadržavati sve takve mod vrijednosti koje se pojavljuju u nizu nuz

Postoji alternativni način, pomoću numpyja:

```
import numpy as np
```

```
# Stvaranje uzorka brojeva
nuz = np.array([1, 2, 3, 3, 4, 5, 6, 6, 7, 8, 8, 9])

# Pronalazak mod vrijednosti
unique_values, counts = np.unique(nuz, return_counts=True)
max_count = np.max(counts)
mod_values = unique_values[counts == max_count]
print(mod_values)
```

U Pythonu, također možemo korisnički definirati funkciju za mod. To možemo napraviti na slijedeći način:

```
import numpy as np

def mod(x):
    unique, counts = np.unique(x, return_counts=True)
    max_count = np.max(counts)
    mode_values = unique[counts == max_count]
    return mode_values

nuz = np.array([1, 2, 3, 3, 4, 5, 6, 6, 7, 8, 8, 9])
mod_result = mod(nuz)
print(mod_result)
```

Razmatramo pojmove središnje tendencije i raspršenosti podataka u statistici. Središnja tendencija je mjera koja opisuje centralnu ili tipičnu vrijednost u skupu podataka, dok je raspršenost mjera koja opisuje varijaciju podataka i koliko su raznolike vrijednosti.

Razmotrimo primjer s nizovima podataka:

Prvi niz podataka: 8, 9, 9, 10, 10, 10, 10, 11, 11, 12 Drugi niz podataka: 1, 2, 5, 8, 10, 10, 12, 15, 18, 19

Prvi niz podataka je prilično koncentriran oko vrijednosti 10, što ukazuje na malu raspršenost podataka. Drugi niz podataka je znatno raznolikiji, što znači da ima veću raspršenost.

Unatoč razlici u raspršenosti, i aritmetička sredina i medijan oba niza podataka su 10. To nas podsjeća da iako te mjere pružaju informacije o središnjoj tendenciji, ne govore nam ništa o raspršenosti podataka. Da bismo bolje razumjeli podatke, trebamo razmotriti i druge statističke mjere koje opisuju raspršenost, kao što su varijanca i standardna devijacija.

```
import statistics

uski = [8, 9, 9, 10, 10, 10, 10, 11, 11, 12]
siroki = [1, 2, 5, 8, 10, 10, 12, 15, 18, 19]

mean_uski = statistics.mean(uski)
mean_siroki = statistics.mean(siroki)

median_uski = statistics.median(uski)
median_siroki = statistics.median(siroki)
```

```
# Mod u Pythonu ne postoji u Numpy-ju kao u R-u, pa ćemo koristiti svoju funkciju za računanje moda.
def mod(niz):
    brojac = {}
    for value in niz:
        if value in brojac:
            brojac[value] += 1
        else:
            brojac[value] = 1

    max_frekvencija = max(brojac.values())
    moda = [value for value, frekvencija in brojac.items() if frekvencija == max_frekvencija]

    return mod

mod_uski = mod(uski)
mod_siroki = mod(siroki)

print(f"Srednja vrijednost uskog niza: {mean_uski}")
print(f"Srednja vrijednost širokog niza: {mean_siroki}")

print(f"Medijan uskog niza: {median_uski}")
print(f"Medijan širokog niza: {median_siroki}")

print(f"Mod uskog niza: {mod_uski}")
print(f"Mod širokog niza: {mod_siroki}")
```

Razmatramo različite mjere koje nam pomažu da razumijemo raznolikost i središnje tendencije podataka. Jedna od tih mjera je raspon (range), koji se definira kao razlika između maksimalne i minimalne vrijednosti u nizu podataka. Druga važna mjera je varijanca (variance), koja mjeri koliko su pojedine vrijednosti u nizu razdvojene od srednje vrijednosti. Pomoću varijance možemo bolje razumjeti raspršenost podataka.

Raspon se računa kao:

range = max(x_1, x_2, ..., x_n) - min(x_1, x_2, ..., x_n)

Za primjer, izračunat ćemo raspon i varijancu za dva različita niza podataka: "uski" i "siroki".

```
import numpy as np

uski = np.array([8, 9, 9, 10, 10, 10, 10, 11, 11, 12])
siroki = np.array([1, 2, 5, 8, 10, 10, 12, 15, 18, 19])

# Računanje raspona
range_uski = np.max(uski) - np.min(uski)
range_siroki = np.max(siroki) - np.min(siroki)

# Računanje varijance
variance_uski = np.var(uski, ddof=1)  # ddof=1 za nepristranu varijancu
variance_siroki = np.var(siroki, ddof=1)
```

```
print("Za uski niz: Raspon =", range_uski, ", Varijanca =", variance_uski)
print("Za siroki niz: Raspon =", range_siroki, ", Varijanca =", variance_siroki)
```

Standardna devijacija

Iako se standardna devijacija može izračunati i iz varijance (uzimanjem korijena), Numpy ima svoju ugrađenu funkciju `std()` koja izračunava standardnu devijaciju. Nizovi podataka o kojima je riječ u ovom odlomku imaju sljedeće standardne devijacije:

```
- std(uski) = 1.154701
- sd(siroki) = 6.218253

# Varijanca
import numpy as np

# Nizovi podataka
uski = [1, 2, 3, 4, 5]
siroki = [10, 20, 30, 40, 50]

# Izračun standardne devijacije
sd_uski = np.std(uski)
sd_siroki = np.std(siroki)

# Ispis rezultata
print(f"Standardna devijacija za uski niz: {sd_uski}")
print(f"Standardna devijacija za siroki niz: {sd_siroki}")
```

NADALJE NEMA PROGRAMSKIH KODOVA SVE DO NASLOVA 2.5

Vratimo se sada na primjer s indeksom tjelesne mase (BMI). Podaci su spremljeni u dataframe "kvant", a imena stupaca (vektora) u dataframe-u su Ime, Težina, Visina i BMI. Opremljeni novim znanjem, sada ćemo lako izračunati neke osnovne kvantitativne mjere tih podataka.

```
import numpy as np
import pandas as pd

# Podaci
data = {
    'Ime': ['Osoba1', 'Osoba2', 'Osoba3', 'Osoba4', 'Osoba5'],
    'Tezina': [70, 80, 75, 95, 60],
    'Visina': [1.75, 1.80, 1.70, 1.90, 1.65],
    'BMI': [22.86, 24.69, 25.95, 26.32, 22.04]
}

# Pretvaranje podataka u DataFrame
df = pd.DataFrame(data)
```

```
# Izračun kvantitativnih mjera
srednja_vrijednost_tezine = np.mean(df['Tezina'])
medijan_tezine = np.median(df['Tezina'])
standardna_devijacija_tezine = np.std(df['Tezina'])
raspon_tezine = np.ptp(df['Tezina'])

# Ispis rezultata
print(f"Srednja vrijednost težine: {srednja_vrijednost_tezine:.5f}")
print(f"Medijan težine: {medijan_tezine:.5f}")
print(f"Standardna devijacija težine: {standardna_devijacija_tezine:.5f}")
print(f"Raspon težine: {raspon_tezine}")
```

U Pythonu koristimo .describe() za prikaz osnovnih statistika za pojedine stupce ili cijele dataframeove.

```
import pandas as pd

# Podaci
data = {
    'Ime': ['Osoba1', 'Osoba2', 'Osoba3', 'Osoba4', 'Osoba5'],
    'Tezina': [60.00, 67.25, 79.00, 76.58, 95.00],
    'Visina': [165.0, 173.8, 176.5, 177.3, 192.0],
    'BMI': [21.22, 22.65, 24.07, 24.24, 28.6]
}

# Pretvaranje podataka u DataFrame
df = pd.DataFrame(data)

# Prikaz osnovnih statistika za stupac "Tezina"
print(df['Tezina'].describe())

# Prikaz osnovnih statistika za cijeli DataFrame "kvant"
print(df.describe())
```

Vizualizacija

Vizualizaciju podataka u Pythonu moguće je ostvariti pomoću matplotliba.

```
import matplotlib.pyplot as plt

# Podaci o težini i imenima osoba
tezina = [60, 67.25, 79, 76.58, 85.25, 95]
ime = ["Osoba1", "Osoba2", "Osoba3", "Osoba4", "Osoba5", "Osoba6"]

# Postavljanje margina za graf
plt.subplots_adjust(bottom=0.15, left=0.15)

# Priprema za prikaz podataka o težini
plt.bar(ime, tezina)
plt.ylim(0, 100)
```

```
plt.title("Podaci o težini")
plt.xlabel("Osobe")
plt.ylabel("Težina (kg)")
plt.xticks(rotation=45, ha="right", fontsize=8)
plt.yticks(fontsize=8)
plt.title("Podaci o težini", fontsize=12)

# Tekstualni opis
plt.text(0.5, -0.25, "Ovdje je prikazan bar grafikon koji prikazuje podatke o težini
različitih osoba.", transform=plt.gca().transAxes)
plt.text(0.5, -0.3, "Imena osoba su prikazana na x-osi, a težina (u kg) na y-osi.",
transform=plt.gca().transAxes)
plt.text(0.5, -0.35, "Grafikon prikazuje raspodjelu težina među osobama.",
transform=plt.gca().transAxes)

# Prikaz grafa
plt.show()
```

```
import matplotlib.pyplot as plt

# Podaci
ime = kvant['Ime']
tezina = kvant['Tezina']

# Postavljanje margina za graf
plt.subplots_adjust(left=0.2, right=0.9, top=0.9, bottom=0.2)

# Priprema za prikaz podataka o težini
plt.barh(ime, tezina, color='darkred')

# Postavljanje osi i naslova
plt.xlim(0, 100)
plt.xlabel('Težina (kg)')
plt.title('Podaci o težini')

# Prikaz grafikona
plt.show()

# Tekstualni opis
print("Ovdje je prikazan horizontalni bar grafikon koji prikazuje podatke o težini
različitih osoba.")
print("Imena osoba su prikazana na y-osi, a težina (u kg) na x-osi.")
print("Grafikon prikazuje raspodjelu težina među osobama.")
```

Histogram je grafički način prikazivanja distribucije podataka. Na x-osi se nalaze različite vrijednosti visine, dok je na y-osi prikazana frekvencija pojavljivanja svake visine. Histogram omogućava vizualno prepoznavanje učestalosti različitih visina u podacima, što je korisno za analizu distribucije podataka.

```
import matplotlib.pyplot as plt

# Postavljanje margina za graf
```

```
plt.subplots_adjust(left=0.15)
plt.figure(figsize=(6, 4))

# Prikaz histograma visine
plt.hist(kvant['Visina'], bins=10, edgecolor='k', color='blue')
plt.title('Histogram visine')
plt.xlabel('Visina')
plt.ylabel('Broj pojavljivanja')

plt.show()
```

Posebno je praktično to što možemo ručno podesiti razne parametre našeg grafa, poput veličine, naziva osi, boje i sl.

```
import matplotlib.pyplot as plt
import numpy as np

# Podaci
visina = kvant['Visina']  # Ovdje ubacimo naše podatke

# Postavljanje stila grafikona
plt.style.use('ggplot')

# Postavljanje parametara grafikona
plt.figure(figsize=(6, 6))  # Veličina grafikona
plt.hist(visina, bins=len(visina), color='aquamarine', edgecolor='black')
plt.title('Visina ispitanika', fontsize=14)
plt.xlabel('Visina', fontsize=12)
plt.ylabel('Frekvencija pojavljivanja', fontsize=12)

# Prilagodba veličine oznaka na osima
plt.xticks(fontsize=10)
plt.yticks(fontsize=10)

plt.show()
```

Python koristi numpy.histogram za izračunavanje histograma i pristupa relevantnim podacima poput granica (breaks), brojača (counts), gustoće (density), srednjih točaka (mids), i drugih sličnih podataka.

```
import numpy as np
import matplotlib.pyplot as plt

# Podaci o visini
visina = kvant['Visina']  # Zamijenite ovo sa stvarnim podacima

# Izračunavanje histograma
hist, bin_edges = np.histogram(visina, bins='auto', density=False)

# Prikazivanje histograma
plt.hist(visina, bins=bin_edges, color='aquamarine', edgecolor='black')
plt.title('Visina ispitanika', fontsize=14)
plt.xlabel('Visina', fontsize=12)
```

```
plt.ylabel('Frekvencija pojavljivanja', fontsize=12)

# Pristup izračunatim podacima
print("breaks:", bin_edges)
print("counts:", hist)
print("density:", hist / len(visina))
print("mids:", (bin_edges[1:] + bin_edges[:-1]) / 2)
print("xname:", "visina")
print("equidist:", True)
print("class_attr:", "histogram")

plt.show()
```

```
import numpy as np
import matplotlib.pyplot as plt

# Podaci o visini
visina = kvant['Visina']  # Zamijenimo ovo sa stvarnim podacima

# Postavljanje stila grafikona
plt.style.use('ggplot')

# Postavljanje parametara grafikona
plt.figure(figsize=(6, 6))  # Veličina grafikona
plt.hist(
    visina,
    bins=len(visina),
    color='darkmagenta',
    edgecolor='black',
    density=True  # Omogućavanje relativne frekvencije
)
plt.title('Visina ispitanika', fontsize=14)
plt.xlabel('Visina', fontsize=12)
plt.ylabel('Relativna frekvencija', fontsize=12)

# Prilagodba veličine oznaka na osama
plt.xticks(fontsize=10)
plt.yticks(fontsize=10)

plt.show()
```

Možemo učiniti i da naši stupci imaju različite širine

```
import numpy as np
import matplotlib.pyplot as plt

# Podaci o visini
visina = kvant['Visina']  # Zamijenimo ovo sa stvarnim podacima

# Postavljanje stila grafikona
plt.style.use('ggplot')
```

```
# Postavljanje parametara grafikona
plt.figure(figsize=(6, 6))  # Veličina grafikona
bin_edges = [165, 168, 172, 178, 180, 182, 192]
plt.hist(
    visina,
    bins=bin_edges,
    color='darkmagenta',
    edgecolor='black',
    density=True  # Omogućavanje relativne frekvencije
)
plt.title('Visina ispitanika - različite širine', fontsize=14)
plt.xlabel('Visina', fontsize=12)
plt.ylabel('Relativna frekvencija', fontsize=12)
plt.ylim(0, 0.06)

# Prilagodba veličine oznaka na osima
plt.xticks(fontsize=10)
plt.yticks(fontsize=10)

plt.show()
```

Završit ćemo ovaj kratki pregled grafičkog prikaza podataka korištenjem jednog tipa grafa, takozvanog boxplota (funkcija u Pythonu je ggplot`). Boxplot se na engleskom naziva "box and whiskers diagram", a može se prevesti i kao "pravokutni dijagram". On nam je koristan u kontekstu analize grupiranja podataka, njihovog rasporeda i identifikacije ekstremnih vrijednosti (outliera).

```
import matplotlib.pyplot as plt

# Podaci o visini
visina = kvant['Visina']  # Zamijenimo ovo sa stvarnim podacima

# Postavljanje stila grafikona
plt.style.use('ggplot')

# Postavljanje parametara grafikona
plt.figure(figsize=(6, 6))  # Veličina grafikona
plt.boxplot(
    visina,
    vert=False,  # Horizontalni prikaz
    widths=0.6,  # Širina "pravokutnih" dijelova
    patch_artist=True,  # Omogućava stilizaciju "pravokutnih" dijelova
    flierprops=dict(marker='o', markersize=5, markerfacecolor='red',
markeredgecolor='red')  # Stilizacija outlier-a
)
plt.title('Boxplot visine osoba', fontsize=14)
plt.xlabel('Visina', fontsize=12)

# Prilagodba veličine oznaka na osama
plt.yticks([])  # Uklanjanje oznaka s y-osi

plt.show()
```

Možemo prikazati i horizontalni boxplot

```
import matplotlib.pyplot as plt

# Podaci o indeksu tjelesne mase (BMI)
bmi = kvant['BMI']  # Zamijenite ovo sa stvarnim podacima

# Postavljanje stila grafikona
plt.style.use('ggplot')

# Postavljanje parametara grafikona
plt.figure(figsize=(6, 3))  # Veličina grafikona (prilagodite prema potrebi)
plt.boxplot(
    bmi,
    vert=False,  # Horizontalni prikaz
    widths=0.6,  # Širina "pravokutnih" dijelova
    patch_artist=True,  # Omogućava stilizaciju "pravokutnih" dijelova
)
plt.title('Boxplot indeksa tjelesne mase', fontsize=14)
plt.xlabel('BMI', fontsize=12)

# Prilagodba veličine oznaka na osima
plt.yticks([])  # Uklanjanje oznaka s y-osi

plt.show()
```

Na jednom grafu možemo pokazati podatke za sve 3 liste

```
import matplotlib.pyplot as plt

# Podaci o BMI
bmi_data = kvant[:, [1, 2, 3]]  # Zamijenite ovo sa stvarnim podacima

# Postavite odgovarajuće margine za grafikon
plt.figure(figsize=(6, 3))  # Veličina grafikona
plt.style.use('ggplot')

# Nacrtajte boxplot za odabrane kolone
plt.boxplot(bmi_data, vert=False, widths=0.6, patch_artist=True)
plt.title('Boxplot svih BMI podataka', fontsize=14)
plt.xlabel('BMI', fontsize=12)
plt.yticks([])  # Uklonite oznake s y-ose

plt.show()
```

Možemo, radi veće preglednosti prikazati i sva 3 boxplota

```
import matplotlib.pyplot as plt

# Podaci za boxplotove
visina_data = kvant['Visina']  # Zamijenite ovo sa stvarnim podacima
tezina_data = kvant['Tezina']  # Zamijenite ovo sa stvarnim podacima
bmi_data = kvant['BMI']  # Zamijenite ovo sa stvarnim podacima
```

```
# Postavimo raspored za prikaz boxplotova
plt.figure(figsize=(12, 4))  # Veličina grafikona
plt.style.use('ggplot')

plt.subplot(131)
plt.boxplot(visina_data, vert=False, widths=0.6, patch_artist=True)
plt.title('Visina')

plt.subplot(132)
plt.boxplot(tezina_data, vert=False, widths=0.6, patch_artist=True)
plt.title('Težina')

plt.subplot(133)
plt.boxplot(bmi_data, vert=False, widths=0.6, patch_artist=True)
plt.title('BMI')

plt.tight_layout()  # Organizira panel za prikaz(layout grafova)
plt.show()
```

###Koeficijent korelacije Za izračun koeficijenta korelacije, numpy ima ugrađenu funkciju corrcoef

```
import numpy as np

# Primjer korištenja funkcije corrcoef() za izračunavanje koeficijenta korelacije
xv = np.array([-2, 0, 2, 4, 4])
yv = np.array([2, 2, 3, 4, 4])

correlation_coefficient = np.corrcoef(xv, yv)[0, 1]
print("Rezultat:", correlation_coefficient)


# Učitavanje podataka za težinu, visinu i indeks tjelesne mase
tezina = np.array([60, 80, 85, 84, 62, 65, 86, 86, 95, 78, 70, 68])
visina = np.array([165, 180, 175, 184, 168, 175, 184, 192, 182, 178, 175, 170])
bmi = np.array([22.03857, 24.69136, 27.75510, 24.81096, 21.96712, 21.22449, 25.40170,
23.32899, 28.68011, 24.61810, 22.85714, 23.52941])

# Izračunajte korelaciju između visine i BMI
correlation_height_bmi = np.corrcoef(visina, bmi)[0, 1]
print("Korelacija između visine i BMI:", correlation_height_bmi)

# Izračunajte korelaciju između visine i težine
correlation_height_weight = np.corrcoef(visina, tezina)[0, 1]
print("Korelacija između visine i težine:", correlation_height_weight)
```

Scatterplot

```
import matplotlib.pyplot as plt

# Prikaz zavisnosti između visine i BMI korištenjem scatterplot-a
```

```
plt.figure(figsize=(6, 4))  # Veličina grafikona
plt.style.use('ggplot')

plt.scatter(kvant['Visina'], kvant['BMI'], marker='o', s=40)
plt.title("Visina i BMI")
plt.xlabel("Visina")
plt.ylabel("BMI")

plt.show()
```

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

# Možemo iskoristiti pairplot za prikaz grafova zavisnosti između kvantitativnih
nizova. Njega uvozimo iz posebne seaborn biblioteke.
# Na primjer, za prikaz zavisnosti visine, težine i BMI

data = kvant[['Visina', 'Tezina', 'BMI']]
sns.set(style="ticks")
sns.pairplot(data, markers='o', plot_kws={'s': 40})
plt.suptitle("Zavisnost između visine, težine i BMI", y=1.02)

plt.show()
```

Da bismo u Pythonu učitali Excel datoteku, koristimo Pandas paket

```
import pandas as pd

# Učitavanje podataka iz Excel datoteke "Skok-u-dalj.xlsx" u DataFrame "dalj"
dalj = pd.read_excel("Skok-u-dalj.xlsx", sheet_name="Skok-u-dalj", header=None)

# Možemo i pridružiti imena stupcima
dalj.columns = ["daljina", "ime", "prezime", "rodjenje", "klub", "dat_skok"]
```

U ovom kontekstu, imamo jedan numerički vektor koji sadrži ukupno 30 podataka. Stupci "rodjenje" i "dat_skok" predstavljaju datume, što znači da se mogu urediti i sortirati. Na primjer, možemo sortirati stupac "rodjenje" kako bismo dobili podatke u određenom redoslijedu, kao što je prikazano u sljedećem Python kodu:

```
# Sortiranje stupca "rodjenje" u DataFrame-u "dalj"
sorted_dalj = dalj.sort_values(by="rodjenje")

import pandas as pd

# Prikaz sažetka podataka za DataFrame "dalj"
summary_dalj = dalj.describe(include="all")

# Ispis tablice srednjim formatiranjem
print(summary_dalj.to_markdown())
```

Ovaj Python kod koristi funkciju describe() za prikaz sažetka podataka u DataFrame dalj, a zatim koristi to_markdown() metodu kako bi ispisao tablicu sa srednjim formatiranjem.

Nadalje, dobivene podatke, možemo prikazati boxplotom.

```
import matplotlib.pyplot as plt

# Postavljanje margina za grafikon
plt.figure(figsize=(5, 5))
plt.margins(0.2)

# Nacrtamo boxplot za stupac "daljina" u DataFrame "dalj"
plt.boxplot(dalj["daljina"], vert=False)
plt.title("Boxplot preskočene daljine")

plt.show()
```

Možemo nacrtati histogram za stupac "daljina u našem DataFrame-u"

```
import matplotlib.pyplot as plt

# Postavljanje margina za grafikon
plt.figure(figsize=(4, 4))
plt.subplots_adjust(left=0.2, right=0.9, top=0.9, bottom=0.2)

# Nacrtajte histogram za stupac "daljina" u DataFrame "dalj"
plt.hist(dalj["daljina"], bins=10, color="darkmagenta")
plt.title("Histogram duljine skoka")
plt.xlabel("Daljina")
plt.ylabel("Frekvencija")

plt.show()
```

```
import pandas as pd

# Grupiranje duljine skoka prema klubovima i prikaz sažetka za svaku grupu
grouped = dalj.groupby("klub")
summary = grouped["daljina"].describe()
print(summary)
```

Možemo i izdovjiti natjecatelje iz pojedinih gradova i nacrtati boxplot za svaku skupinu

```
import matplotlib.pyplot as plt

# Izdvajamo natjecatelje iz Zagreba
dalj_zg = dalj[dalj["klub"] == "Zagreb"]

# Izdvajamo natjecatelje iz Rijeke
dalj_ri = dalj[dalj["klub"] == "Rijeka"]

# Izdvajamo natjecatelje iz Varaždina
dalj_vz = dalj[dalj["klub"] == "Varaždin"]

# Postavljanje margina za grafikone
plt.subplots(figsize=(12, 4))
```

```
plt.subplots_adjust(wspace=0.4)

# Crtamo boxplot za natjecatelje iz Zagreba
plt.subplot(131)
plt.boxplot(dalj_zg["daljina"], vert=False)
plt.title("Zagreb")
plt.xlabel("Daljina")

# Crtamo boxplot za natjecatelje iz Rijeke
plt.subplot(132)
plt.boxplot(dalj_ri["daljina"], vert=False)
plt.title("Rijeka")
plt.xlabel("Daljina")

# Crtamo boxplot za natjecatelje iz Varaždina
plt.subplot(133)
plt.boxplot(dalj_vz["daljina"], vert=False)
plt.title("Varaždin")
plt.xlabel("Daljina")

plt.show()
```

Isto tako, možemo prikazati i broj natjecatelja iz pojedinog grada

```
import Numpy
# Broj natjecatelja iz Zagreba, Rijeke i Varaždina
broj_natjecatelja_zg = dalj_zg.shape[0]
broj_natjecatelja_ri = dalj_ri.shape[0]
broj_natjecatelja_vz = dalj_vz.shape[0]

print(broj_natjecatelja_zg)
print(broj_natjecatelja_ri)
print(broj_natjecatelja_vz)
```

Možemo, zbog lakše manipulacije datumima, u naš dataframe dodati još 1 stupac koji ima samo numeričku godinu rođenja.

```
import pandas as pd

# Pretvorite stupac "rodjenje" u tip datuma (date)
dalj['rodjenje'] = pd.to_datetime(dalj['rodjenje'])

# Dodajte novi stupac "godina" s godinama rođenja
dalj['godina'] = dalj['rodjenje'].dt.year
```

Možemo i grupirati duljinu skoka po godinama rođenja i prikazati sažetak za svaku pojedinu grupu

```
import pandas as pd

# Grupiranje duljine skoka po godinama rođenja i prikaz sažetka za svaku grupu
grouped = dalj.groupby('godina')['daljina']
summary = grouped.describe()
```

```
print(summary)
```

Naravno, te je podatke moguće vizualizirati

```
import matplotlib.pyplot as plt

plt.figure(figsize=(12, 4))

dalj_2004 = dalj[dalj['godina'] == 2004]
dalj_2005 = dalj[dalj['godina'] == 2005]

plt.subplot(1, 2, 1)
plt.hist(dalj_2004['daljina'], bins=10, color='skyblue')
plt.title('Godište 2004')
plt.xlabel('Daljina')
plt.ylabel('Frekvencija')

plt.subplot(1, 2, 2)
plt.hist(dalj_2005['daljina'], bins=10, color='salmon')
plt.title('Godište 2005')
plt.xlabel('Daljina')
plt.ylabel('Frekvencija')

plt.tight_layout()
plt.show()
```

U Pythonu, umjesto Tidyverse-a iz R-a, možemo koristiti Pandas, popularnu biblioteku za analizu i obradu podataka. Za čitanje podataka i izračunavanje sažetka za stupac "daljina" iz data frame-a možete koristiti sljedeći ekvivalentni Python kôd: import pandas as pd (ukoliko nije instaliran, instaliramo ga u našem terminalu pomoću pip install Pandas)

```
# Pretpostavljamo da postoji DataFrame "dalj" s odgovarajućim podacima

# Izračunajte sažetak za stupac "daljina"
summary = dalj["daljina"].describe()

# Ispis rezultata
print(summary)
```

Pandas nudi describe() funkciju koja daje sažetak za numeričke stupce u DataFrame-u, uključujući minimum, maksimum, srednju vrijednost, kvartile itd. Ovaj kôd će vam dati ekvivalentne rezultate kao i R kôd s Tidyverse-om

Kod Pythona čitanje i parsiranje HTML web stranica te dohvaćanje tablica može se postići korištenjem biblioteke BeautifulSoup i requests za dohvaćanje HTML sadržaja sa web stranice.

```
import requests
from bs4 import BeautifulSoup

# Adresa web stranice
```

```
wikistr = "https://en.wikipedia.org/wiki/"
listurban = "List_of_urban_areas_in_the_European_Union"
url = wikistr + listurban

# Dohvaćanje HTML sa web stranice
response = requests.get(url)
html_content = response.text

# Parsiranje HTML-a koristeći BeautifulSoup
soup = BeautifulSoup(html_content, 'html.parser')

# Dohvaćanje tablica
tables = soup.find_all('table', {'class': 'wikitable'})

# Pretvaranje tablica u DataFrame (ovdje pretpostavljamo da koristimo pandas)
import pandas as pd
dfs = [pd.read_html(str(table)) for table in tables]
```

```
# Iz naše html stranice možemo odabrati određene sadržaje i pohraniti ih u novi
DataFrame
import pandas as pd

# Odabir određenih stupaca iz DataFrame-a 'eu' i pretvaranje u novi DataFrame
eu_cities = eu.iloc[:, [1, 3, 4, 8]]

# Ispis strukture novog DataFrame-a
print(eu_cities.info())
```

```
# Možemo zamjeniti imena stupaca u eu_cities DataFrame-u
eu_cities.columns = ["Grad", "Drzava", "BrojSt", "Gustoca"]

# Nadalje, možemo konvertirati BrojSt i Gustoca u numeričke vektore
eu_cities["BrojSt"] = eu_cities["BrojSt"].str.replace(",", "").astype(float)
eu_cities["Gustoca"] = eu_cities["Gustoca"].str.replace(",", "").astype(float)

# Ispis strukture DataFrame-a s ograničenim brojem redova za prikaz
print(eu_cities.head(3))

# Ispis strukture eu_cities DataFrame-a
eu_cities.info()

# Možemo vidjeti gdje se Zagreb nalazi u listi gradova
zagreb_info = eu_cities[eu_cities["Grad"] == "Zagreb"]
print(zagreb_info)
```

```
# Sortiranje eu_cities po Gustoca u opadajućem redoslijedu
eu_sorted_gust = eu_cities.sort_values(by='Gustoca', ascending=False)

# Ispis prvih 10 gradova s najvećom gustoćom naseljenosti
print(eu_sorted_gust.head(10))

        Grad    Drzava   BrojSt Gustoca
```

```
67      Genoa     Italy   540000  695021
63     Vienna   Austria  1890000   56148
6     Athens    Greece  3450000  529046
57     Bilbao     Spain   775000  525051
62    Málaga     Spain   686000  509469
61 Santa Cruz     Spain   506000  465238
45 Sofia Pernik Bulgaria  947000  45713
48    Madrid     Spain  6211000  455122
22  Bucharest   Romania 1862000  45225
44 Barcelona     Spain 4800000  4477
```

Možemo sortirati gradove prema gustoći naseljenosti u opadajućem redoslijedu

```
import pandas as pd

# Sortiranje eu_cities prema gustoći naseljenosti u opadajućem redoslijedu
eu_sort_gust = eu_cities.sort_values(by='Gustoca', ascending=False)

# Ispis prvih 10 gradova s najvećom gustoćom naseljenosti
print(eu_sort_gust.head(10))
```

```
# Koristit ćemo matplotlib za kreiranje histograma s našim podacima
import matplotlib.pyplot as plt
import numpy as np

# Podaci za histogram (proučavanje BrojSt, podijeljeno s 1.000.000)
data = eu_cities['BrojSt'] / 1000000

# Postavljanje margina za grafikon
plt.subplots_adjust(left=0.1, right=0.9, top=0.9, bottom=0.1)

# Nacrtaj histogram
plt.hist(data, bins=11, range=(0.5, 11.5), color='aquamarine4')

# Postavke osi i naslova
plt.xlabel("Veličina urbanog područja (mil.)")
plt.ylabel("Broj gradova")
plt.title("Veličina EU gradova")

# Postavljanje oznaka na x-osi
plt.xticks(np.arange(0.5, 12, 1))

# Prikaži grafikon
plt.show()
```

```
import matplotlib.pyplot as plt
import numpy as np

# Postavljanje margina za grafikon
plt.subplots_adjust(left=0.15, right=0.95, top=0.9, bottom=0.2)

# Niz podataka (prilagođeni brojevi)
```

```
data = np.array(eu_cities['BrojSt']) / 1000000

# Nacrtaj histogram relativnih frekvencija
plt.hist(data, bins=11, density=True, color='darkmagenta')

# Postavljanje naslova i oznaka osi
plt.title("Veličina EU gradova")
plt.xlabel("Veličina urbanog područja (mil.)")
plt.ylabel("Relativna frekvencija")

# Postavljanje granica x-osi
plt.xlim(0.5, 11.5)

# Dodavanje oznaka na x-osi
plt.xticks(np.arange(0.5, 12, 1), rotation=0)

# Dodavanje "rug" (ticks) na x-osi
plt.scatter(data, np.zeros_like(data), marker='|', color='black', s=30)

# Prikaz grafa
plt.show()
```

Ovaj kôd crta histogram relativnih frekvencija na temelju podataka iz stupca eu_cities[broj_st], pri čemu se koristi density= TRUE. Oznake i stilovi grafikona također su postavljeni, a na x-osi se dodaju oznake s funkcijom xtics. Također je dodan "rug" na x-osi pomoću funkcije scatter.

Boxplot je koristan za prikazivanje raspodjele podataka i identificiranje potencijalnih outlier-a. U ovom slučaju, prikazat ćemo boxplot za veličinu EU gradova

```
import matplotlib.pyplot as plt

# Postavljanje margina za grafikon
plt.subplots_adjust(left=0.2)

# Crtanje horizontalnog boxplot-a za veličinu EU gradova
plt.boxplot(eu_cities["BrojSt"] / 1000000, vert=False)

# Postavljanje oznaka
plt.title("Veličina EU gradova")
plt.xlabel("Veličina urbanog područja (mil.)")
plt.ylabel("Broj gradova")

# Prikaz grafikona
plt.show()
```

Python

Python koristi pandas za ostvarivanje funkcionalnosti selektiranja podataka iz nekog dataframe-a. Python kod će grupirati podatke prema stupcu 'Drzava' i primijeniti funkciju sum na stupcu 'BrojSt' kako bi izračunao ukupnu sumu za svaku državu.

Rezultat će biti DataFrame s dvjema kolonama, 'Drzava' i 'BrojSt', koji sadrži agregirane vrijednosti za svaku državu.

```
import pandas as pd

# Pretvaranje DataFrame-a eu_cities u Pandas DataFrame
eu_cities_df = pd.DataFrame(eu_cities)

# Agregacija podataka prema stupcu 'BrojSt' grupiranom po stupcu 'Drzava' i
primjenjujući funkciju sum
result = eu_cities_df.groupby('Drzava')['BrojSt'].sum().reset_index()

# Ispis rezultata
print(result)
```

U Pythonu možemo prilično jednostavno grupirati i sortirati podatke

```
import pandas as pd

# Grupiranje i agregacija
df1 = eu_cities.loc[1:20].groupby('Drzava')['BrojSt'].sum().reset_index()
df2 = eu_cities.loc[1:20].groupby('Drzava')['BrojSt'].count().reset_index()

# Spajanje data frame-ova
result_df = pd.merge(df1, df2, on='Drzava')

# Preimenovanje stupaca
result_df.columns = ['Drzava', 'Stanovnici', 'Broj_gradova']

# Sortiranje prema Broj_gradova u opadajućem redoslijedu
result_df = result_df.sort_values(by='Broj_gradova', ascending=False)

print(result_df)
```

Python

U Pythonu, postoje 2 načina da dohvatimo prvih 20 gradova po naseljenosti

```
import pandas as pd

# Učitavanje podataka iz DataFrame-a eu_cities
eu_cities = pd.DataFrame(eu_cities)  # Ovdje dodamo prave podatke

# Filtriranje prvih 20 redova
eu_cities = eu_cities.iloc[:20]

# Korištenje operatera za obradu podataka
result = (
    eu_cities
    .groupby('Drzava')
    .agg(Broj_gradova=pd.NamedAgg(column='BrojSt', aggfunc='count'),
         Stanovnici=pd.NamedAgg(column='BrojSt', aggfunc='sum'))
```

```
    .reset_index()
    .sort_values(by='Broj_gradova', ascending=False)
)

# Ispis rezultata
print(result)
```

```
import pandas as pd

# Učitavanje podataka iz DataFrame-a eu_cities
eu_cities = pd.DataFrame(eu_cities)

# Filtriranje prvih 20 redova
eu_cities = eu_cities.iloc[:20]

# Korištenje data.table za obradu podataka
import data_table as dt
dtdf = dt.Frame(eu_cities)
result = (
    dtdf[:, dt.sum(dt.f.BrojSt), dt.count(), dt.by(dt.f.Drzava)]
    .sort(-dt.f.Broj_gradova)
)

# Ispis rezultata
print(result)
```

U ovome dijelu skripte namjerno ostavljam i R i Python kod kako bi kontekst bio jasniji

R

Program "R" dolazi s raznovrsnim paketima koji su vrlo korisni za različite potrebe u analizi podataka. Popis dostupnih paketa može se pronaći korištenjem funkcije data(package = "ime_paketa"), koja prikazuje informacije o paketima i omogućuje pristup dokumentaciji i funkcionalnostima svakog od njih. Tako se, na primjer, u R-ovom paketu "datasets" nalazi skup podataka "faithful", koji sadrži podatke o erupcijama gejzira "Old Faithful" koji se nalazi u američkom Nacionalnom parku Yellowstone. Taj gejzir je poznat po relativno pravilnim erupcijama, gdje se događaju svakih 40 minuta do dva sata. U skupu podataka "faithful" nalaze se dvije kolone s 272 redaka. U prvoj koloni su navedeni vremena trajanja erupcija, a u drugoj koloni vremena čekanja između erupcija (u minutama). S obzirom na broj redaka, nećemo ispisati sve elemente tog skupa podataka, već ćemo odmah pogledati njihovu strukturu, sažetak podataka i boxplotove.

```
library("datasets")

# Ispis strukture podataka 'faithful'
str(faithful)
# 'data.frame': 272 obs. of 2 variables:
#  $ eruptions: num 3.6 1.8 3.33 2.28 4.53 ...
#  $ waiting  : num 79 54 74 62 85 55 88 85 51 85 ...
```

```r
# Sažetak podataka
summary(faithful)
#   eruptions        waiting
# Min.   :1.600   Min.   :43.0
# 1st Qu.:2.163   1st Qu.:58.0
# Median :4.000   Median :76.0
# Mean   :3.488   Mean   :70.9
# 3rd Qu.:4.454   3rd Qu.:82.0
# Max.   :5.100   Max.   :96.0

# Grafovi boxplot
par(mfrow = c(1, 2), mar = c(2, 2, 2, 2))
boxplot(faithful$eruptions, main = "Trajanje erupcije")
boxplot(faithful$waiting, main = "Vrijeme između")
```

U Pythonu ne postoji ekvivalentna funkcija koja prikazuje popis dostupnih paketa zajedno s informacijama o njima na način sličan funkciji data(package = "ime_paketa") u R-u.

U Pythonu, možete dobiti popis instaliranih paketa koristeći naredbu pip list ili pip freeze u naredbenom retku, ovisno o verziji Pythona koju koristite. Međutim, to će samo izlistati nazive paketa bez dodatnih informacija.

Da biste dobili više informacija o određenom paketu, možete koristiti pip show ime_paketa, gdje zamjenjujete "ime_paketa" stvarnim nazivom paketa. Ova naredba će prikazati informacije o verziji, autorima, opisu i drugim metapodacima za taj paket.

U Pythonu također postoji funkcija help(), koja vam omogućuje da dobijete dokumentaciju i informacije o modulima i funkcijama. Na primjer, ako želite dobiti informacije o paketu numpy, možete koristiti help(numpy). Ova funkcija će prikazati dokumentaciju za paket, modul ili funkciju koje specificirate kao argument.

R

Ovaj kod stvara dva jednodimenzionalna grafa (stripchart) za prikazivanje podataka iz skupa podataka "faithful". Prvi graf prikazuje trajanje erupcija, dok drugi graf prikazuje vrijeme između erupcija. Opcija jitter = TRUE dodaje malo nasumičnog pomaka podacima radi bolje preglednosti kad postoje višestruki zapisi s istim vrijednostima. Funkcija main koristi se za postavljanje naslova na grafove.

```r
par(mfrow = c(2, 1), mar = c(2, 2, 2, 2))
stripchart(faithful$eruptions, jitter = TRUE, main = "Stripchart trajanja erupcija")
stripchart(faithful$waiting, jitter = TRUE, main = "Stripchart vremena između
erupcija")
```

Kako Python nema sadržan paket za gejzir, možemo uvesti dataframe sa tim podacima

```python
import matplotlib.pyplot as plt
import pandas as pd

# Pretvaranje skupa podataka 'faithful' u DataFrame (pod pretpostavkom da je skup
podataka već dostupan)
```

```
faithful_df = pd.DataFrame({'eruptions': faithful.eruptions, 'waiting':
faithful.waiting})

# Postavljanje grafova u dva reda
plt.figure(figsize=(6, 8))

# Stripchart za trajanje erupcija
plt.subplot(2, 1, 1)
plt.stripplot(data=faithful_df, x='eruptions', jitter=True)
plt.title("Stripchart trajanja erupcija")

# Stripchart za vrijeme između erupcija
plt.subplot(2, 1, 2)
plt.stripplot(data=faithful_df, x='waiting', jitter=True)
plt.title("Stripchart vremena između erupcija")

plt.tight_layout()
plt.show()
```

```
# Postavljanje margina za grafikon
par(mar = c(4, 4, 2, 2))

# Izrada scatter-plot grafa
plot(
  faithful$waiting, faithful$eruptions,
  pch = 16, cex = 0.8,
  xlab = "Vremena između erupcija (min)",
  ylab = "Vrijeme erupcije (min)",
  main = "Scatter-plot grafikon za gejzir 'Old Faithful'"
)
```

```
import matplotlib.pyplot as plt

# Postavljanje margina za grafikon
plt.subplots_adjust(left=0.1, right=0.9, top=0.9, bottom=0.1)

# Izrada scatter-plot grafa
plt.scatter(faithful['waiting'], faithful['eruptions'], marker='o', s=40)
plt.xlabel("Vremena između erupcija (min)")
plt.ylabel("Vrijeme erupcije (min)")
plt.title("Scatter-plot grafikon za gejzir 'Old Faithful'")

# Prikaži grafikon
plt.show()
```

```
Poželjno je i sortirati podatke po veličini kako bi prikaz bio jasniji
# Postavljanje margina za grafikon
par(mar = c(2, 4, 1, 1))

# Ispis vremena erupcija u sortiranom redoslijedu
plot(
  sort(faithful$eruptions),
```

```
  pch = 1,
  cex = 0.7,
  cex.axis = 0.8,
  ylab = "Vrijeme erupcije (min)",
  cex.lab = 0.8
)
Možemo dodati i rug
# Postavljanje margina za grafikon
par(mar = c(2, 4, 1, 1))

# Ispis vremena erupcija u sortiranom redoslijedu
plot(
  sort(faithful$eruptions),
  pch = 1,
  cex = 0.7,
  cex.axis = 0.8,
  ylab = "Vrijeme erupcije (min)",
  cex.lab = 0.8
)

# Dodavanje "rug" (ticks) na y-osi
rug(faithful$eruptions, side = 2)

# Postavljanje margina za grafikon
par(mar = c(2, 4, 1, 1))

# Ispis vremena između erupcija u sortiranom redoslijedu
plot(
  sort(faithful$waiting),
  pch = 1,
  cex = 0.7,
  cex.axis = 0.8,
  ylab = "Vrijeme između erupcija (min)",
  cex.lab = 0.8
)

# Dodavanje "rug" (ticks) na y-osi
rug(faithful$waiting, side = 2)

# Postavljanje margina za grafikon
par(mar = c(4, 4, 2, 2))

# Nacrtaj histogram vremena između erupcija
hist(
  faithful$eruptions,
  xlab = "Vrijeme između erupcija (min)",
  main = "Histogram frekvencija vremena između erupcija",
  cex.axis = 0.7,
  cex.main = 0.8,
  cex.lab = 0.8,
  ylab = "Frekvencija",
```

```
  col = "aquamarine4"
)
```

```
import matplotlib.pyplot as plt
import numpy as np

# Postavljanje margina za grafikone
plt.figure(figsize=(8, 10))
plt.subplots_adjust(hspace=0.5)

# Ispis vremena erupcija u sortiranom redoslijedu
plt.subplot(3, 1, 1)
sorted_eruptions = np.sort(faithful['eruptions'])
plt.plot(sorted_eruptions, 'o', markersize=5)
plt.title('Vrijeme erupcija (min)')
plt.xlabel('Redoslijed')
plt.ylabel('Vrijeme erupcije (min)')

# Dodavanje "rug" (ticks) na y-osi
plt.subplot(3, 1, 2)
plt.plot(sorted_eruptions, np.zeros_like(sorted_eruptions), '|', markersize=20)
plt.title('Vrijeme erupcija s "rug" (ticks)')
plt.xlabel('Vrijeme erupcije (min)')

# Ispis vremena između erupcija u sortiranom redoslijedu
plt.subplot(3, 1, 3)
sorted_waiting = np.sort(faithful['waiting'])
plt.plot(sorted_waiting, 'o', markersize=5)
plt.title('Vrijeme između erupcija (min)')
plt.xlabel('Redoslijed')
plt.ylabel('Vrijeme između erupcija (min)')

plt.show()

# Nacrtaj histogram vremena između erupcija
plt.figure(figsize=(8, 5))
plt.hist(sorted_eruptions, bins=15, color='aquamarine', edgecolor='black')
plt.title('Histogram frekvencija vremena između erupcija')
plt.xlabel('Vrijeme između erupcija (min)')
plt.ylabel('Frekvencija')

plt.show()
```

```
# Postavljanje margina za grafikon
par(mar = c(5, 4, 2, 2))

# Nacrtaj histogram vremena između erupcija
hist(
  faithful$eruptions,
  main = "Histogram relativnih frekvencija",
  breaks = 20,
  probability = TRUE,
```

```
  cex.axis = 0.7,
  cex.main = 0.8,
  xlab = "Vrijeme između erupcija (min)",
  cex.lab = 0.8,
  ylab = "Frekvencija",
  col = "aquamarine4"
)

# Dodaj krivulju glatke gustoće
points(
  density(faithful$eruptions, bw = 0.1),
  type = 'l',
  col = 'red',
  lwd = 2
)
```

```
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import gaussian_kde

# Postavljanje margina za grafikon
plt.figure(figsize=(8, 6))
plt.subplots_adjust(left=0.1, right=0.9, top=0.9, bottom=0.1)

# Nacrtaj histogram vremena između erupcija
plt.hist(
    faithful['eruptions'],
    bins=20,
    density=True,
    color='aquamarine4',
    edgecolor='black'
)
plt.title('Histogram relativnih frekvencija')
plt.xlabel('Vrijeme između erupcija (min)')
plt.ylabel('Frekvencija')

# Izračunaj i nacrtaj glatku krivulju gustoće
density = gaussian_kde(faithful['eruptions'])
x = np.linspace(faithful['eruptions'].min(), faithful['eruptions'].max(), 1000)
plt.plot(x, density(x), 'r', linewidth=2)

plt.show()
```

Završit ćemo ovu malu analizu podataka o gejziru "Old Faithful" analizom korelacije između vremena trajanja erupcija i vremena između njih. Prikazani dijagram raspršenja sugerira da postoji visoka pozitivna korelacija između tih podataka. To isto potvrđuje i izračun korelacije:

```
cor(faithful$waiting, faithful$eruptions)
# [1] 0.9008112
```

Python ekvivalent:

```python
import pandas as pd
import numpy as np

# Učitaj podatke
data = pd.read_csv("putanja/do/podataka.csv")  # Zamijenite "putanja/do/podataka.csv" sa stvarnom putanjom do datoteke

# Izračunaj korelaciju
correlation = np.corrcoef(data['waiting'], data['eruptions'])[0, 1]
print(f"Korelacija: {correlation}")
```