SVEUČILIŠTE U RIJECI

FAKULTET INFORMATIKE I DIGITALNIH TEHNOLOGIJA

Preddiplomski studij informatike

1.Projekt

Uvod u analizu i vizualizaciju podataka

Analiza skupa podataka **Global Air Pollution**

Autor: Zara Čubranić

U Rijeci, travanj 2024.

Sadržaj

[**Uvod 2**](#_dx83miodlmf0)

[**Opis skupa podataka 3**](#_oep8nvsvrc01)

[**NumPy funkcije 4**](#_4t4xg71wcpu)

[**Statističke mjere 8**](#_r1p4jva41bxn)

[**Vrijednosti u korelaciji 10**](#_w6baj6l3b3ax)

[**Kod 14**](#_gatratyzmxcv)

# Uvod

U današnje vrijeme sve veći izazov s kojim se svijet suočava je zagađenje zraka. Visoka razina onečišćenja zraka može imati ozbiljne posljedice po zdravlje ljudi i okoliš, te može pridonijeti klimatskim promjenama. Stoga je važno pratiti kvalitetu zraka i analizirati podatke kako bi se donosile informirane odluke o smanjenju onečišćenja.

U ovom istraživanju koristit ću dataset koji sadrži podatke o kvaliteti zraka u različitim gradovima diljem svijeta. Dataset uključuje informacije o koncentraciji različitih štetnih tvari u zraku, kao što su PM2.5, PM10, dušikov dioksid (NO2), sumporov dioksid (SO2) i ozon (O3). Također, dataset sadrži podatke o meteorološkim uvjetima kao što su temperatura, vlažnost zraka i brzina vjetra.

Ovi podaci mogu nam pružiti uvid u razinu onečišćenja zraka u različitim dijelovima svijeta, te nam pomoći u pronalaženju uzroka i posljedica zagađenja zraka. U nastavku ćemo detaljnije analizirati ovaj dataset kako bismo bolje razumjeli trendove onečišćenja zraka i identificirali moguće mjere za smanjenje negativnih utjecaja na okoliš i zdravlje ljudi.

# Opis skupa podataka

Dataset se sastoji od jedne datoteke CSV formata nazvane global\_air\_pollution\_dataset.csv. Datoteka sadrži detaljne informacije o sljedećim parametrima:

* Country: Naziv države
* City: Naziv grada
* AQI Value: Ukupna vrijednost AQI indeksa za grad
* AQI Category: Ukupna kategorija AQI indeksa za grad (npr. dobra, umjerena, nezdrava)
* CO AQI Value: Vrijednost AQI indeksa za ugljikov monoksid (CO)
* CO AQI Category: Kategorija AQI indeksa za ugljikov monoksid (CO)
* Ozone AQI Value: Vrijednost AQI indeksa za ozon (O3)
* Ozone AQI Category: Kategorija AQI indeksa za ozon (O3)
* NO2 AQI Value: Vrijednost AQI indeksa za dušikov dioksid (NO2)
* NO2 AQI Category: Kategorija AQI indeksa za dušikov dioksid (NO2)
* PM2.5 AQI Value: Vrijednost AQI indeksa za čestice prašine promjera 2.5 mikrometra ili manje (PM2.5)
* PM2.5 AQI Category: Kategorija AQI indeksa za čestice prašine promjera 2.5 mikrometra ili manje (PM2.5)

**Tipovi podataka:**

Dataset primarno sadrži kategorijalne i numeričke podatke.

* **Kategorijalni podaci:** Obuhvaćaju nazive država, gradove i kategorije AQI (npr. dobra, umjerena).
* **Numerički podaci:** Uključuju vrijednosti AQI indeksa za ukupno zagađenje zraka i pojedinačne onečišćujuće tvari (CO, O3, NO2, PM2.5).

**Kategorije podataka:**

Za lakše razumijevanje i analizu, dataset koristi standardni AQI indeks koji kategorizira kvalitetu zraka na sljedeći način:

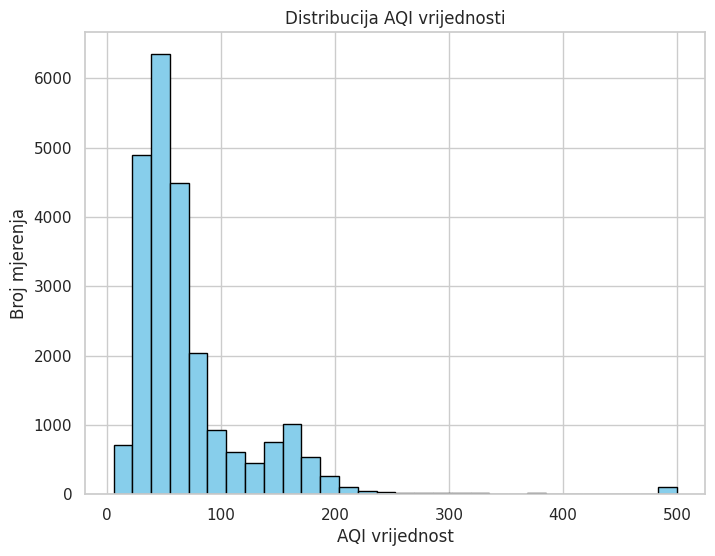
* Dobra
* Umjerena
* Nezdrava za osjetljive grupe
* Nezdrava za sve
* Vrlo nezdrava

Na primjer, prvi red datoteke može sadržavati sljedeće vrijednosti:

* **Country:** United States of America
* **City:** Los Angeles
* **AQI Value:** 101
* **AQI Category:** Moderate
* **CO AQI Value:** 12
* **CO AQI Category:** Good
* **Ozone AQI Value:** 85
* **Ozone AQI Category:** Moderate
* **NO2 AQI Value:** 36
* **NO2 AQI Category:** Good
* **PM2.5 AQI Value:** 42
* **PM2.5 AQI Category:** Moderate

Ove vrijednosti vam govore da je kvaliteta zraka u Los Angelesu umjerena, s umjerenim vrijednostima ozona i čestica prašine PM2.5. Vrijednosti ugljikovog monoksida i dušikovog dioksida su dobre.

Ukupno dataset pokriva 23463 različite lokacije.



# NumPy funkcije

Evo pet funkcija ili operacija iz NumPy biblioteke koje imaju smisla za analizu ovog skupa podataka o zagađenju zraka:

1. **np.mean():**

Ova funkcija može se koristiti za izračun prosječnih vrijednosti AQI (Air Quality Index) ili vrijednosti zagađivača poput CO, O3, NO2 i PM2.5 za određene zemlje, gradove ili regije.

Prosječni AQI za ovaj skup podataka iznosi približno 72.01. Prosječni AQI pruža informaciju o prosječnoj kvaliteti zraka u skupu gradova ili regija obuhvaćenih podacima. Ova vrijednost omogućuje procjenu općenite razine zagađenja zraka u istraživanom području.

# 1. Izračun prosječnog AQI

average\_aqi = np.mean(aqi\_values)

print("Prosječni AQI:", average\_aqi)

Prosječni AQI: 72.01086817542513

1. **np.max():**

Koristeći ovu funkciju, možemo pronaći maksimalne vrijednosti AQI ili vrijednosti zagađivača u skupu podataka. To nam može pomoći identificirati najviše onečišćene gradove ili regije.

Maksimalni AQI u ovom skupu podataka iznosi 500. Maksimalni AQI pokazuje najvišu zabilježenu razinu zagađenja zraka u nekom od gradova ili regija u analiziranom skupu podataka. Vrijednost od 500 označava izuzetno visoku razinu zagađenja, što može ukazivati na ozbiljne probleme s kvalitetom zraka u tom području.

# 2. Pronalaženje maksimalnog AQI

max\_aqi = np.max(aqi\_values)

print("Maksimalni AQI:", max\_aqi)

Maksimalni AQI: 500

1. **np.min():**

Ova funkcija omogućuje pronalaženje minimalnih vrijednosti AQI ili vrijednosti zagađivača. To može biti korisno za identifikaciju područja s najboljom kvalitetom zraka.

Minimalni AQI u skupu podataka iznosi 6. Minimalni AQI pokazuje najnižu zabilježenu razinu zagađenja zraka u analiziranom području. Ova vrijednost može ukazivati na područja s izuzetno dobrim kvalitetom zraka ili na nedostatak podataka za neka područja.

# 3. Pronalaženje minimalnog AQI

min\_aqi = np.min(aqi\_values)

print("Minimalni AQI:", min\_aqi)

Minimalni AQI: 6

1. **np.std():**

Računajući standardnu devijaciju, možemo dobiti informaciju o raspršenosti podataka o zagađenju zraka. Veća vrijednost standardne devijacije može ukazivati na veću varijabilnost kvalitete zraka.

Standardna devijacija AQI iznosi približno 56.05. Standardna devijacija mjeri raspršenost podataka oko prosječne vrijednosti. Veća vrijednost standardne devijacije ukazuje na veću varijabilnost u kvaliteti zraka između različitih gradova ili regija obuhvaćenih podacima.

# 4. Računanje standardne devijacije AQI

std\_dev\_aqi = np.std(aqi\_values)

print("Standardna devijacija AQI:", std\_dev\_aqi)

Standardna devijacija AQI: 56.054025694475605

1. **np.median():**

Medijanu možemo koristiti kao mjeru središnje tendencije za AQI ili vrijednosti zagađivača. To može biti korisno ako postoji nekoliko ekstremnih vrijednosti koje mogu utjecati na prosječnu vrijednost.

Medijan AQI iznosi 55.0. Medijan je središnja vrijednost podataka kada su poredani po veličini. U ovom slučaju, medijan AQI predstavlja središnju razinu zagađenja zraka u skupu podataka. On je manje osjetljiv na ekstremne vrijednosti od prosječne vrijednosti, što ga čini korisnim za procjenu tipične razine zagađenja zraka u analiziranom području.

# 5. Izračunavanje mediane AQI

median\_aqi = np.median(aqi\_values)

print("Medijan AQI:", median\_aqi)

Medijan AQI: 55.0

# Statističke mjere

1. **Izračun kvantila (npr. 25%, 50%, 75%):**

Kvantili predstavljaju vrijednosti koji dijele distribuciju podataka na određene proporcije. Na primjer, 25. percentil je vrijednost ispod koje leži 25% podataka. Ovo može biti korisno za razumijevanje raspodjele AQI vrijednosti u datasetu.

quantiles = np.percentile(aqi\_values, [25, 50, 75])

print("25. percentil:", quantiles[0])

print("50. percentil (medijan):", quantiles[1])

print("75. percentil:", quantiles[2])

25. percentil: 39.0

50. percentil (medijan): 55.0

75. percentil: 79.0

1. **Brojanje vrijednosti koje su veće/manje od određene granice:**

Ova operacija može pružiti informacije o tome koliko često se AQI vrijednosti nalaze izvan određenih granica, što može biti korisno za analizu kvalitete zraka.

# Brojanje vrijednosti izvan određene granice (npr. AQI > 100)

count\_above\_100 = np.sum(aqi\_values > 100)

print("Broj AQI vrijednosti većih od 100:", count\_above\_100)

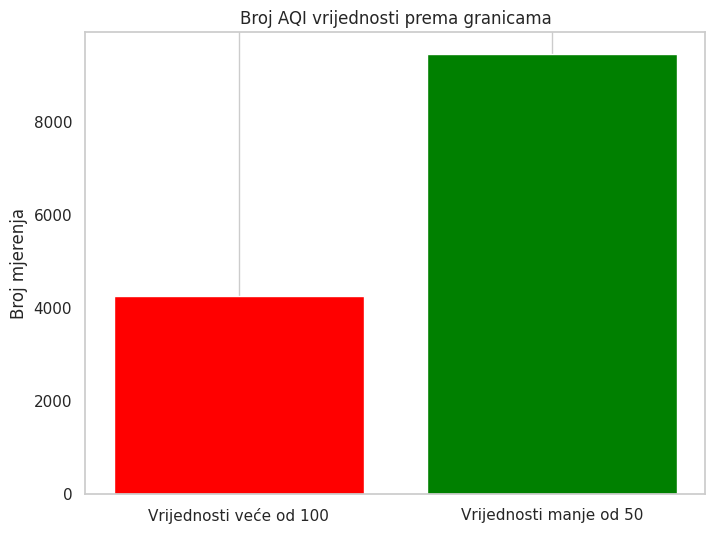
# Brojanje vrijednosti iznad određene granice (npr. AQI < 50)

count\_below\_50 = np.sum(aqi\_values < 50)

print("Broj AQI vrijednosti manjih od 50:", count\_below\_50)

Broj AQI vrijednosti većih od 100: 4244

Broj AQI vrijednosti manjih od 50: 9453



1. **Izračun koeficijenta varijacije:**

Koeficijent varijacije (CV) je mjera relativnog raspona varijabilnosti podataka. To je omjer standardne devijacije i srednje vrijednosti. Visok koeficijent varijacije ukazuje na visoku varijabilnost, dok niska vrijednost ukazuje na malu varijabilnost.

cv\_aqi = (std\_dev\_aqi / average\_aqi) \* 100 # izražen u postotcima

print("Koeficijent varijacije AQI:", cv\_aqi)

Koeficijent varijacije AQI: 77.84106359879291

# Vrijednosti u korelaciji

U analizi kvalitete zraka, istraživanje korelacije između AQI vrijednosti i drugih zagađivača poput CO, NO2 i ozona pruža važne uvide u međusobnu povezanost tih parametara. Koristeći statističke metode poput izračuna korelacijske matrice i vizualizacije toplinskom kartom, možemo identificirati postojanje i jačinu veze između različitih zagađivača zraka i općenite kvalitete zraka.

Korelacija između AQI-a i drugih zagađivača može ukazati na uzročno-posljedične veze te nam pomoći u razumijevanju kako se promjene u koncentracijama određenih zagađivača odražavaju na ukupnu kvalitetu zraka. Na primjer, visoka pozitivna korelacija između AQI-a i CO ili NO2 sugerira da su povećane razine tih zagađivača povezane s lošijom kvalitetom zraka, što može biti korisno za identificiranje izvora zagađenja i razvoj strategija za njegovo smanjenje.

# Učitavanje potrebnih podataka

co\_values = data['CO AQI Value']

no2\_values = data['NO2 AQI Value']

ozone\_values = data['Ozone AQI Value']

# Izračun koeficijenta korelacije

correlation\_aqi\_co = np.corrcoef(aqi\_values, co\_values)[0, 1]

correlation\_aqi\_no2 = np.corrcoef(aqi\_values, no2\_values)[0, 1]

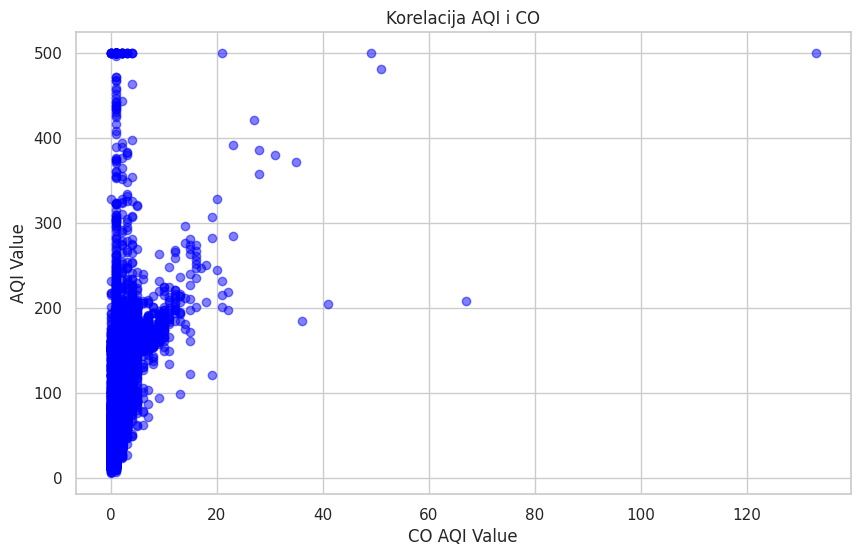
correlation\_aqi\_ozone = np.corrcoef(aqi\_values, ozone\_values)[0, 1]

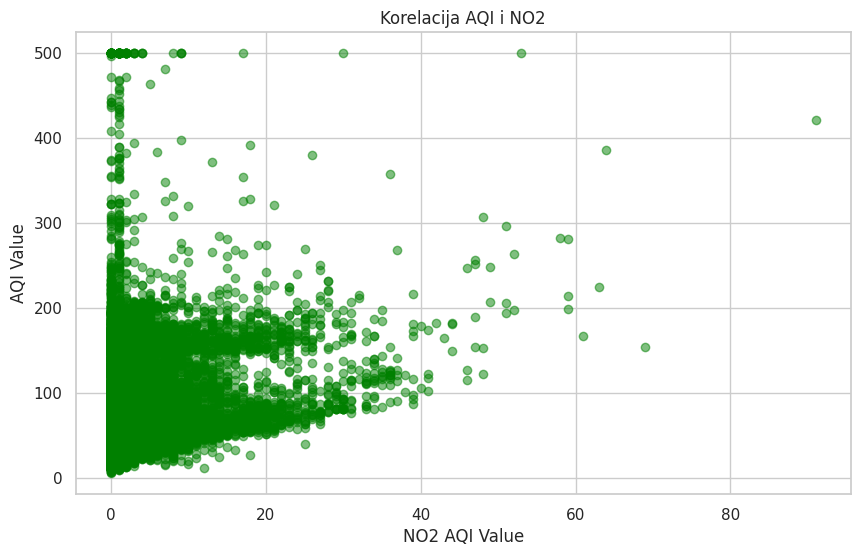
# Ispis rezultata

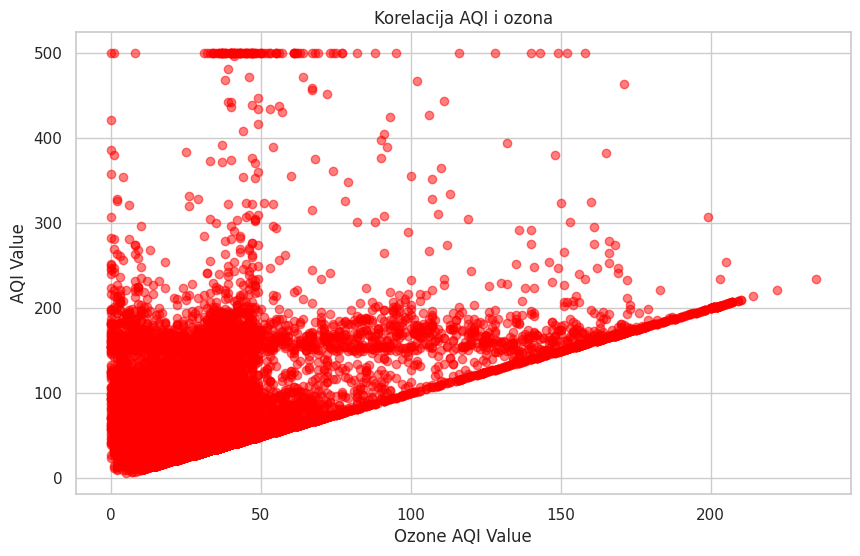
print("Korelacija AQI i CO:", correlation\_aqi\_co)

print("Korelacija AQI i NO2:", correlation\_aqi\_no2)

print("Korelacija AQI i ozona:", correlation\_aqi\_ozone)







Na grafovima koji prikazuju korelaciju između AQI vrijednosti i drugih zagađivača zraka poput CO, NO2 i ozona, svaki točkasti podatak predstavlja jednu točku mjerenja. Osovina x označava vrijednosti određenog zagađivača (npr. CO AQI Value, NO2 AQI Value, Ozone AQI Value), dok osovina y označava AQI vrijednost, koja se koristi kao indikator općenite kvalitete zraka.

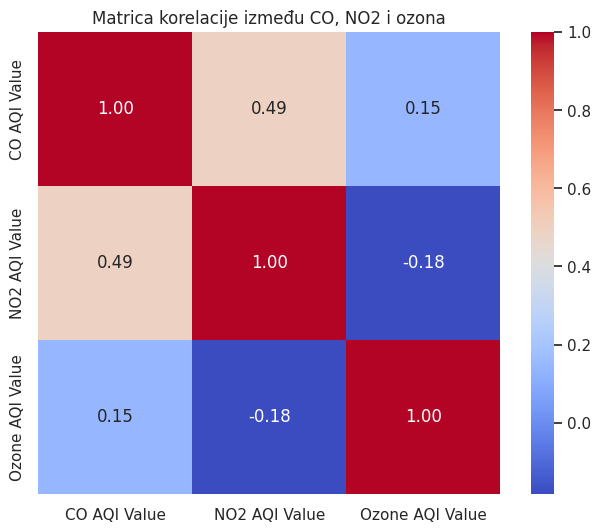
Na svakom grafu, ako postoji jasna korelacija između AQI vrijednosti i određenog zagađivača, primijetit ćemo određeni oblik ili smjer distribucije točaka. Na primjer, pozitivna korelacija između AQI i zagađivača sugerira da su veće vrijednosti zagađivača povezane s većim vrijednostima AQI-a, što ukazuje na lošiju kvalitetu zraka. U slučaju negativne korelacije, više vrijednosti zagađivača povezane su s manjim vrijednostima AQI-a, što sugerira bolju kvalitetu zraka.

## Toplinska karta korelacije zagađivaća

Na prikazanoj toplinskoj karti (heatmap) matrice korelacije između CO, NO2 i ozona, boja svakog kvadratića predstavlja jačinu i smjer korelacije između dva zagađivača.

Vrijednosti na glavnoj dijagonali, gdje su zagađivači upoređivani sa samim sobom, su uvijek 1 jer je korelacija zagađivača sa samim sobom savršeno pozitivna.

Pozitivna korelacija između dva zagađivača očituje se tamnijom bojom, dok negativna korelacija rezultira svjetlijom bojom. Što je boja intenzivnija, to je korelacija jača.



Vrijednosti 1.0 na dijagonalama označavaju savršenu pozitivnu korelaciju između CO, NO2 i ozona sa samim sobom.

Između CO i NO2 imamo umjerenu pozitivnu korelaciju (0.49), što ukazuje da povećanje koncentracije jednog zagađivača obično prati povećanje koncentracije drugog zagađivača.

Između CO i ozona te između NO2 i ozona, imamo slabu ili gotovo zanemarivu korelaciju (-0.18, 0.15), što sugerira da promjene u koncentraciji jednog zagađivača nisu značajno povezane s promjenama u koncentraciji drugog zagađivača.

Ovakva vizualizacija omogućava brzo razumijevanje međusobne povezanosti između različitih zagađivača zraka, što je ključno za identificiranje uzroka i razumijevanje dinamike zagađenja zraka

# Kod

import numpy as np

import pandas as pd

# Učitavanje podataka

data = pd.read\_csv('dataset.csv')

# Izdvajanje AQI vrijednosti

aqi\_values = data['AQI Value']

# 1. Izračun prosječnog AQI

average\_aqi = np.mean(aqi\_values)

print("Prosječni AQI:", average\_aqi)

# 2. Pronalaženje maksimalnog AQI

max\_aqi = np.max(aqi\_values)

print("Maksimalni AQI:", max\_aqi)

# 3. Pronalaženje minimalnog AQI

min\_aqi = np.min(aqi\_values)

print("Minimalni AQI:", min\_aqi)

# 4. Računanje standardne devijacije AQI

std\_dev\_aqi = np.std(aqi\_values)

print("Standardna devijacija AQI:", std\_dev\_aqi)

# 5. Izračunavanje mediane AQI

median\_aqi = np.median(aqi\_values)

print("Medijan AQI:", median\_aqi)

# Statistika

# Izračunavanje kvantila

quantiles = np.percentile(aqi\_values, [25, 50, 75])

print("25. percentil:", quantiles[0])

print("50. percentil (medijan):", quantiles[1])

print("75. percentil:", quantiles[2])

# Brojanje vrijednosti izvan određene granice (npr. AQI > 100)

count\_above\_100 = np.sum(aqi\_values > 100)

print("Broj AQI vrijednosti većih od 100:", count\_above\_100)

# Brojanje vrijednosti iznad određene granice (npr. AQI < 50)

count\_below\_50 = np.sum(aqi\_values < 50)

print("Broj AQI vrijednosti manjih od 50:", count\_below\_50)

# Izračun koeficijenta varijacije

cv\_aqi = (std\_dev\_aqi / average\_aqi) \* 100 # izražen u postotcima

print("Koeficijent varijacije AQI:", cv\_aqi)

# Korelacija

# Učitavanje potrebnih podataka

co\_values = data['CO AQI Value']

no2\_values = data['NO2 AQI Value']

ozone\_values = data['Ozone AQI Value']

# Izračun koeficijenta korelacije

correlation\_aqi\_co = np.corrcoef(aqi\_values, co\_values)[0, 1]

correlation\_aqi\_no2 = np.corrcoef(aqi\_values, no2\_values)[0, 1]

correlation\_aqi\_ozone = np.corrcoef(aqi\_values, ozone\_values)[0, 1]

# Ispis rezultata

print("Korelacija AQI i CO:", correlation\_aqi\_co)

print("Korelacija AQI i NO2:", correlation\_aqi\_no2)

print("Korelacija AQI i ozona:", correlation\_aqi\_ozone)

# Grafovi

import matplotlib.pyplot as plt

# Histogram za distribuciju AQI vrijednosti

plt.figure(figsize=(8, 6))

plt.hist(aqi\_values, bins=30, color='skyblue', edgecolor='black')

plt.title('Distribucija AQI vrijednosti')

plt.xlabel('AQI vrijednost')

plt.ylabel('Broj mjerenja')

plt.grid(True)

plt.show()

# Stupčasti graf za broj AQI vrijednosti većih i manjih od određene granice

plt.figure(figsize=(8, 6))

plt.bar(['Vrijednosti veće od 100', 'Vrijednosti manje od 50'], [count\_above\_100, count\_below\_50], color=['red', 'green'])

plt.title('Broj AQI vrijednosti prema granicama')

plt.ylabel('Broj mjerenja')

plt.grid(axis='y')

plt.show()

# 

# Korelacija

# Definiranje podataka za grafikon

x\_co = co\_values

x\_no2 = no2\_values

x\_ozone = ozone\_values

# Korelacija AQI i CO

plt.figure(figsize=(10, 6))

plt.scatter(x\_co, aqi\_values, color='blue', alpha=0.5)

plt.title('Korelacija AQI i CO')

plt.xlabel('CO AQI Value')

plt.ylabel('AQI Value')

plt.grid(True)

plt.show()

# Korelacija AQI i NO2

plt.figure(figsize=(10, 6))

plt.scatter(x\_no2, aqi\_values, color='green', alpha=0.5)

plt.title('Korelacija AQI i NO2')

plt.xlabel('NO2 AQI Value')

plt.ylabel('AQI Value')

plt.grid(True)

plt.show()

# Korelacija AQI i ozona

plt.figure(figsize=(10, 6))

plt.scatter(x\_ozone, aqi\_values, color='red', alpha=0.5)

plt.title('Korelacija AQI i ozona')

plt.xlabel('Ozone AQI Value')

plt.ylabel('AQI Value')

plt.grid(True)

plt.show()

# Heatamap

import seaborn as sns

correlation\_data = data[['CO AQI Value', 'NO2 AQI Value', 'Ozone AQI Value']]

correlation\_matrix = correlation\_data.corr()

plt.figure(figsize=(8, 6))

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', fmt=".2f", square=True)

plt.title('Matrica korelacije između CO, NO2 i ozona')

plt.show()

Prosječni AQI: 72.01086817542513

Maksimalni AQI: 500

Minimalni AQI: 6

Standardna devijacija AQI: 56.054025694475605

Medijan AQI: 55.0

25. percentil: 39.0

50. percentil (medijan): 55.0

75. percentil: 79.0

Broj AQI vrijednosti većih od 100: 4244

Broj AQI vrijednosti manjih od 50: 9453

Koeficijent varijacije AQI: 77.84106359879291

Korelacija AQI i CO: 0.4306016428516073

Korelacija AQI i NO2: 0.23175791460152015

Korelacija AQI i ozona: 0.4053097578174101