Osnove nauke o podacima objašnjenje projekta

Faza 1 - izbor podataka

Dataset sadrži podatke iz Svetske banke, sa sledećim kolonama:

1. **Country**: Naziv zemlje (tip: tekstualni podaci)
2. **Year**: Godina za koju su podaci prikupljeni (tip: ceo broj)
3. **GDP (USD)**: Bruto domaći proizvod u američkim dolarima (tip: decimalni broj)
4. **Population**: Populacija zemlje (tip: decimalni broj)
5. **Life Expectancy**: Očekivani životni vek u godinama (tip: decimalni broj)
6. **Unemployment Rate (%)**: Stopa nezaposlenosti u procentima (tip: decimalni broj)
7. **CO2 Emissions (metric tons per capita)**: Emisija CO2 po glavi stanovnika u metričkim tonama (tip: decimalni broj)
8. **Access to Electricity (%)**: Procenat populacije koja ima pristup električnoj energiji (tip: decimalni broj)

Dataset je preuzet sa Kaggle-a.

Faza 2 - vizuelizacija podataka

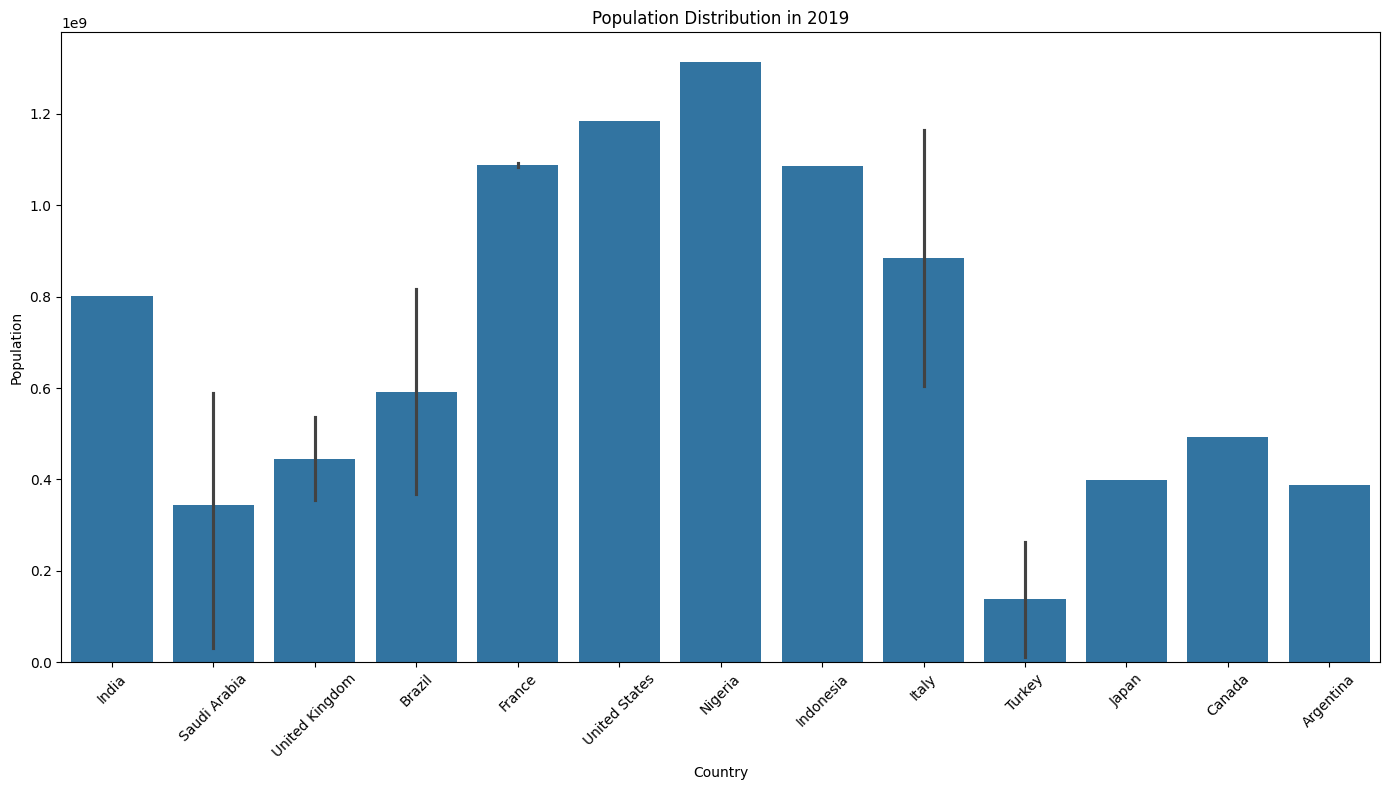
*Prva vizuelizacija*:

Ova vizualizacija prikazuje distribuciju populacije po državama za poslednju dostupnu godinu (2019).

Najveća populacija: Vizuelno je jasno da je Indija vodeća sa populacijom blizu 1,2 milijarde.

Značajne države: Sjedinjene Američke Države i Nigerija takođe imaju visoku populaciju, sa vrednostima preko milijardu.

Niža populacija: Zemlje poput Saudijske Arabije i Turske imaju nižu populaciju.



*Druga vizualizacija:*

Ova vizualizacija istražuje odnos između očekivanog životnog veka i bruto domaćeg proizvoda (BDP) izraženog u američkim dolarima (na logaritamskoj skali).

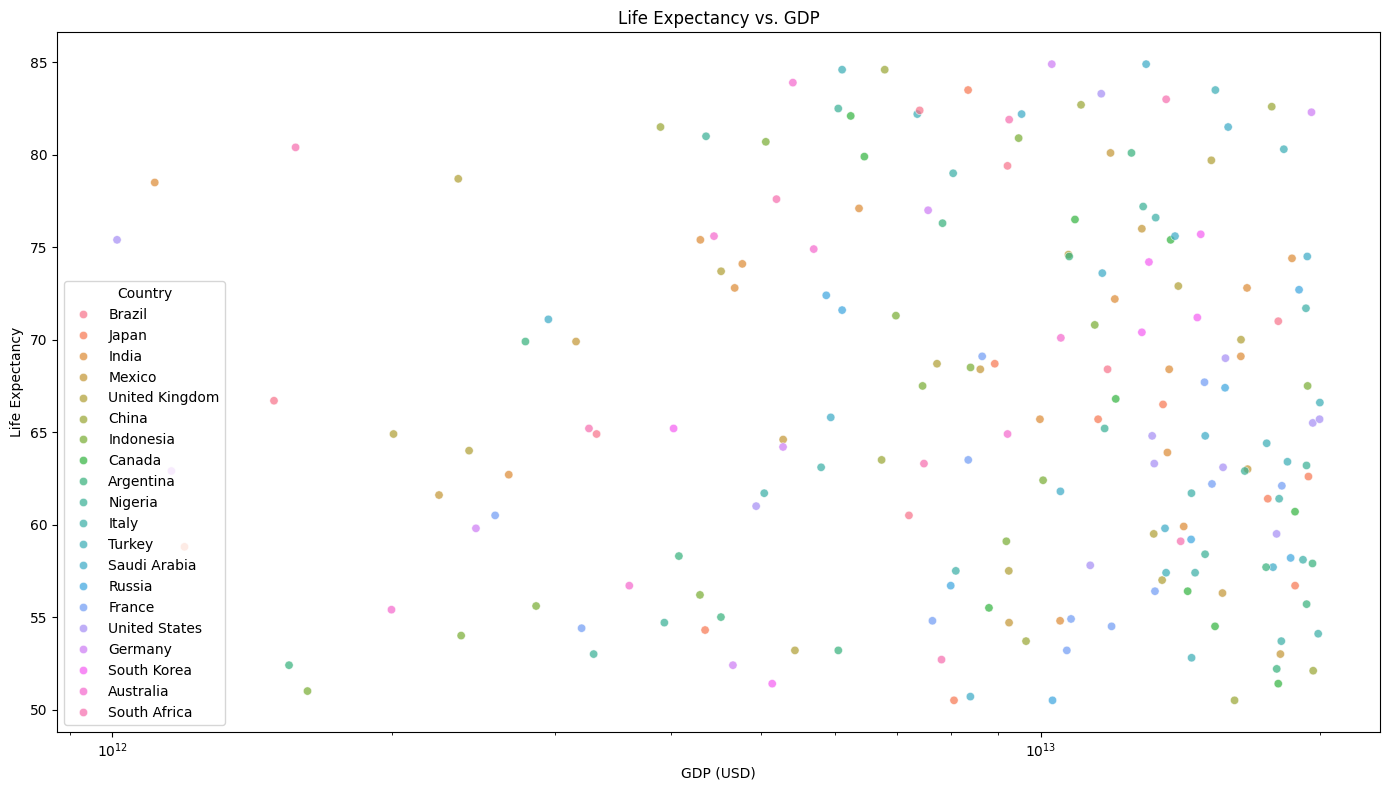
Pozitivan trend: Generalno se primećuje pozitivan trend – zemlje sa višim BDP-om imaju tendenciju da imaju viši očekivani životni vek. Ovo ukazuje na značaj ekonomskog razvoja u poboljšanju kvaliteta života.

Izdvajanje zemalja:

Japan se ističe visokim očekivanim životnim vekom (preko 80 godina) i relativno visokim BDP-om.

Zemlje poput Nigerije imaju niži očekivani životni vek i manji BDP.

Raspon očekivanog životnog veka varira od oko 50 godina u manje razvijenim zemljama do preko 80 godina u razvijenim zemljama, dok BDP pokazuje veliki raspon, od niskih do vrlo visokih vrednosti.



*Treća vizualizacija:*

Ova vizualizacija prikazuje stopu nezaposlenosti u različitim zemljama tokom poslednjih godina, korišćenjem heatmap prikaza.

Trendovi nezaposlenosti:

Argentina, Rusija i Sjedinjene Američke Države beleže visoke stope nezaposlenosti u određenim godinama (preko 20%).

Nigerija i Saudijska Arabija imaju uglavnom niže stope nezaposlenosti.

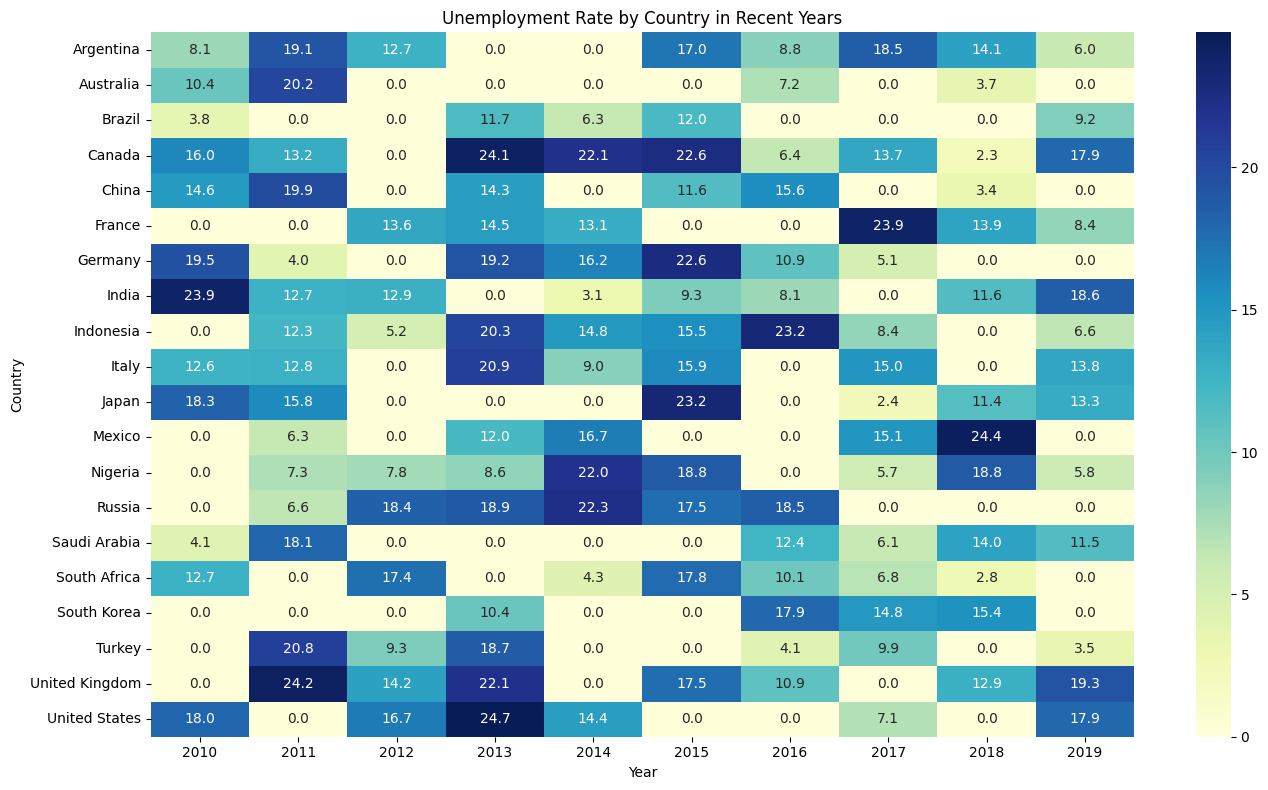
Kanada pokazuje značajne poraste stope nezaposlenosti u pojedinim godinama, poput 2013. godine.

Zemlje sa niskom nezaposlenošću:

Japan i Južna Koreja generalno imaju niže stope nezaposlenosti u poređenju sa drugim zemljama.

Vremenske promene:

U većini zemalja vidljive su značajne promene stope nezaposlenosti između 2010. i 2019. godine.

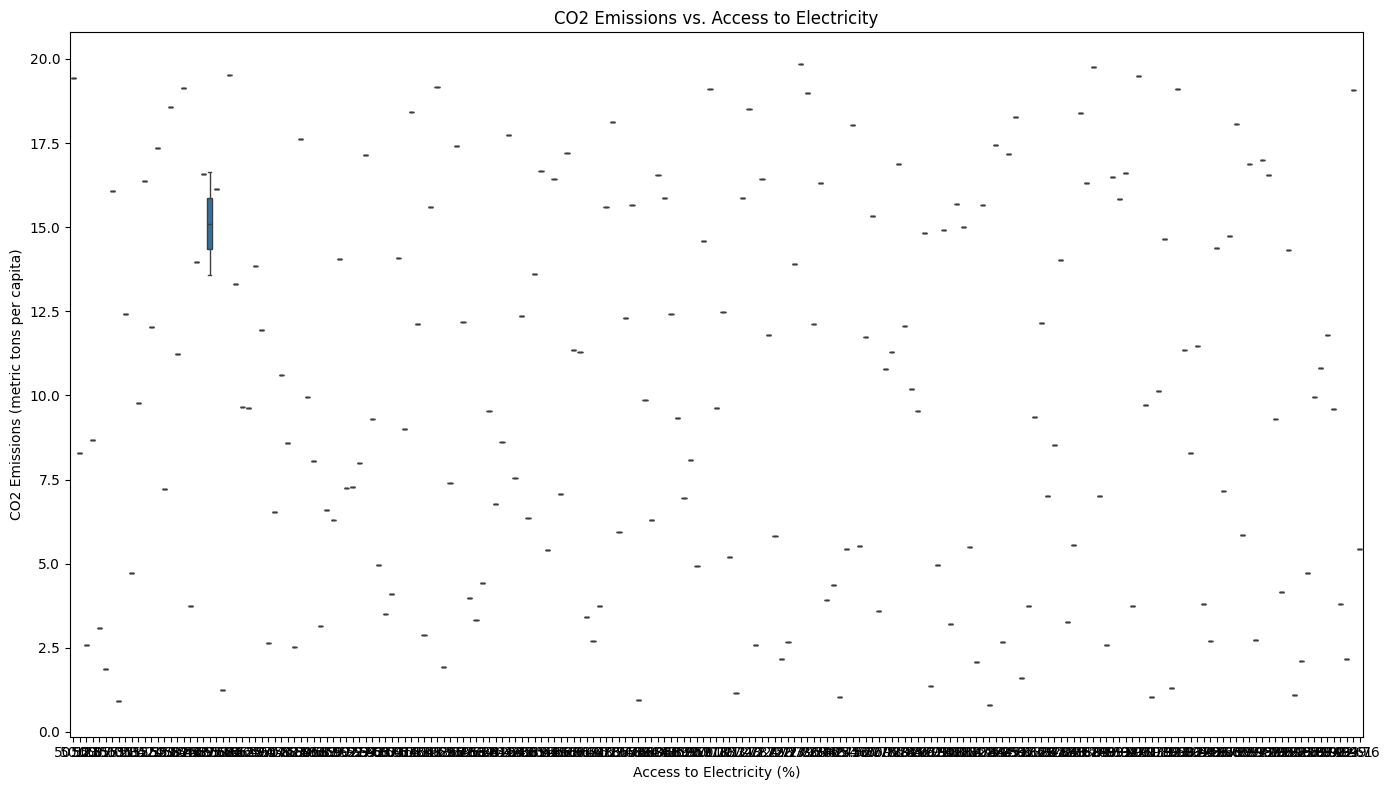


*Četvrta vizualizacija*:

Ova vizualizacija istražuje odnos između emisije CO2 (u metričkim tonama po glavi stanovnika) i procenta populacije sa pristupom električnoj energiji.

Raspon emisija CO2:

Emisije CO2 po glavi stanovnika značajno variraju, krećući se od niskih (ispod 5 metričkih tona) do visokih vrednosti (preko 15 metričkih tona).



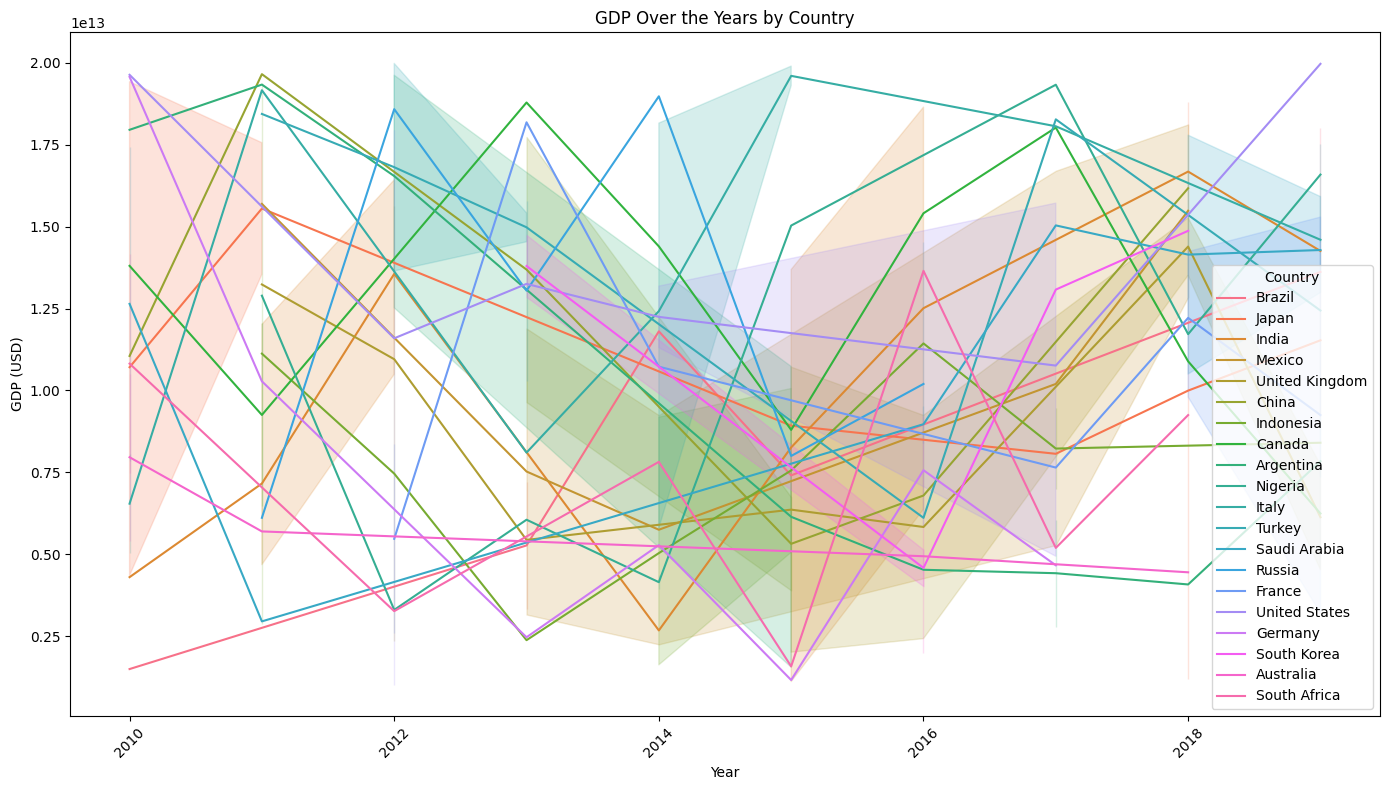
*Peta vizualizacija:*

Ova vizualizacija prikazuje promene BDP-a različitih zemalja tokom godina.

Rast BDP-a: Zemlje poput Sjedinjenih Američkih Država i Kine pokazuju stabilan rast BDP-a tokom analiziranog perioda. Indija takođe beleži pozitivan trend.

Nekoliko zemalja, poput Nemačke i Francuske, pokazuje stabilne trendove sa manjim varijacijama.

Nizak BDP: Zemlje poput Nigerije i Argentine imaju relativno nizak BDP u poređenju sa razvijenim zemljama.



Faza 3 - Priprema podataka i istraživačka analiza

U data set-u nije bilo nedostajućih vrednosti, ali je svakako odrađena provera i popunjavanje.

Takođe, data set nije imao outlier-a, i svi podaci su unutar predviđenih granica.

#### Ključni odnosi između varijabli kada je u pitanju korelacija:

#### **GDP (USD):**

Ima blago negativnu korelaciju sa pristupom električnoj energiji (Access to Electricity (%)): **-0.15**.

Niska korelacija sa ostalim promenljivima, što sugeriše da BDP nije značajno povezan sa populacijom, očekivanim životnim vekom ili stopom nezaposlenosti.

1. **Population (Populacija)**:

Ima malu pozitivnu korelaciju sa pristupom električnoj energiji: **0.08**.

Korelacija sa ostalim varijablama je vrlo mala, bliska 0.

1. **Life Expectancy (Očekivani životni vek)**:

Ima vrlo slabu negativnu korelaciju sa GDP (USD) (-0.06).

Korelacija sa svim ostalim varijablama je niska.

1. **Unemployment Rate (%) (Stopa nezaposlenosti)**:

Ima vrlo malu pozitivnu korelaciju sa emisijom CO2 (**0.04**).

Korelacija sa ostalim varijablama je zanemarljiva.

1. **CO2 Emissions (metric tons per capita)**:

Korelacija sa svim ostalim varijablama je slaba ili bliska 0.

1. **Access to Electricity (%)**:

Blago negativna korelacija sa GDP (USD) (**-0.15**).

Slaba pozitivna korelacija sa populacijom (**0.08**).

Nema značajnih korelacija između varijabli u ovom dataset-u.

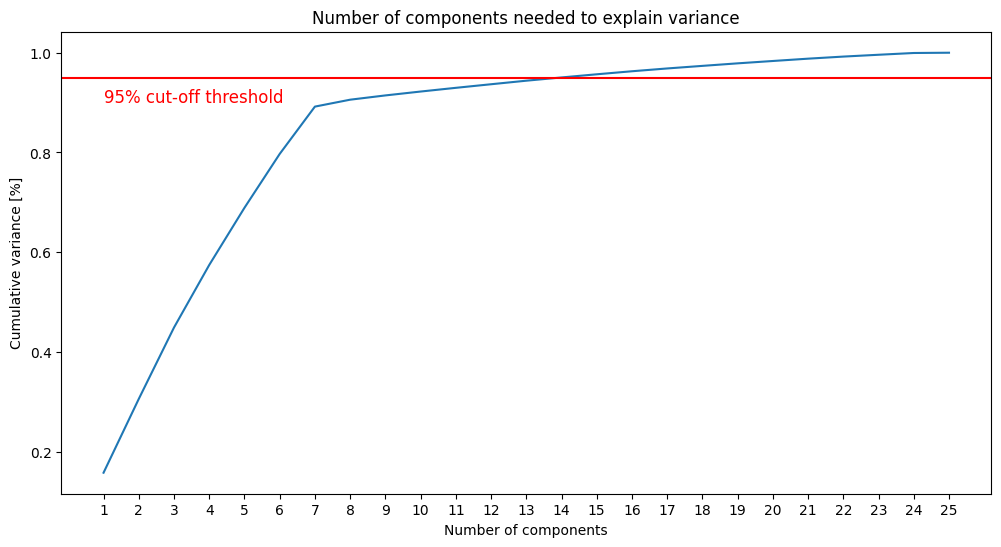
Korelacija između GDP (USD) i pristupa električnoj energiji je najjača negativna vrednost (**-0.15**), ali je još uvek vrlo slaba.

Varijable su uglavnom nezavisne, što sugeriše da su međusobne veze male ili zanemarljive.

Deskriptivna statistika:

Pipeline transformiše podatke tako da na kraju imaju oblik (200, n), gde je n broj kolona nakon transformacije.

PCA je zadržao manji broj dimenzija od originalnog broja, dok je zadržano 95% varijanse podataka.



t-SNE je algoritam za vizualizaciju koji smanjuje dimenzionalnost na 2 ili 3 dimenzije radi otkrivanja klastera. Scatterplot prikazuje klastere u dataset-u u 2D prostoru.

**Pipeline-ovi** standardizuju i transformišu podatke za analizu.

**PCA** smanjuje dimenzionalnost uz minimalan gubitak informacije.

**t-SNE** vizualizuje podatke u 2D prostoru, otkrivajući klastere i strukturu podataka.

Faza 4 - Statistička analiza i testiranje hipoteza

**Shapiro-Wilk test** je korišćen za proveru normalnosti distribucije podataka u dve numeričke kolone:

* Rezultat za prvu kolonu ('Year'): P-vrednost = 0.0000 → Podaci nisu normalno distribuirani.
* Rezultat za drugu kolonu ('GDP (USD)'): P-vrednost = 0.0000 → Podaci nisu normalno distribuirani.

Ovi rezultati ukazuju da se za dalju analizu moraju koristiti neparametarski testovi.

**Mann-Whitney U-test** je neparametarski test korišćen za poređenje medijana dve grupe podataka:

* Statistički značajna razlika između kolona 'Year' i 'GDP (USD)' sa P-vrednošću = 0.0000000000.

Izračunate su srednja vrednost i standardna devijacija za obe grupe:

* **Grupa 1 ('Year')**:
  + Srednja vrednost: 2014.5
  + Standardna devijacija: 2.88
* **Grupa 2 ('GDP (USD)')**:
  + Srednja vrednost: 10.57e+12
  + Standardna devijacija: 5.55e+12

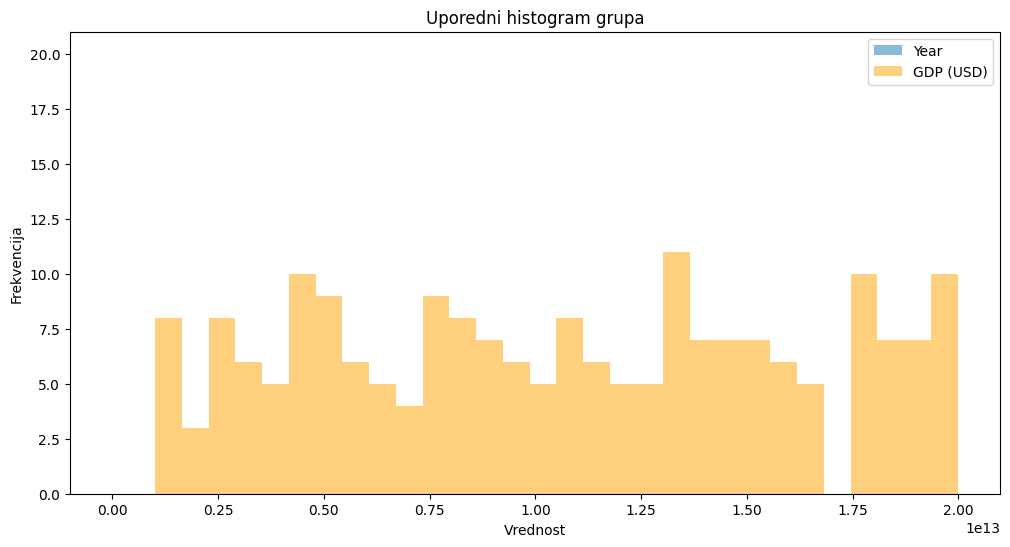
**Cohen-ov d** je korišćen za procenu veličine efekta između dve grupe:

* Cohen-ov d: -2.6942
* Interpretacija: Veličina efekta je velika.

**Za prvu grupu ('Year') izračunat je 95% interval poverenja**:

* Interval poverenja: (2014.10, 2014.90)

Ovaj interval pokazuje granice u kojima se sa 95% sigurnosti nalazi prava srednja vrednost.



Faza 5 - Prediktivno modeliranje i mašinsko učenje

Za klasifikaciju i regresiju pripremljeni su ulazni i izlazni podaci:

* Ulazne kolone ('X'):GDP (USD), Population, Unemployment Rate (%), CO2 Emissions (metric tons per capita), Access to Electricity (%).
* Izlazna kolona ('y'):Binarna varijabla označava da li je životni vek iznad medijane.

Podaci su standardizovani i podeljeni na trening i test skupove (70% trening, 30% test).

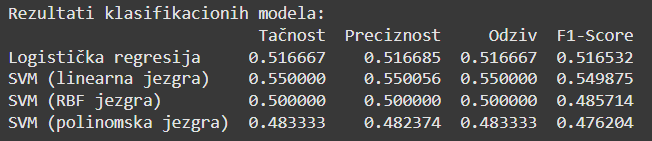
Logistička regresija je evaluirana korišćenjem K-Fold Cross-Validation sa 5 preklopa:

* Prosečna tačnost: ~0.52

Evaluirano je više klasifikacionih modela:

* **Logistička regresija**
* **SVM sa linearnim jezgrima**
* **SVM sa RBF jezgrima**
* **SVM sa polinomskim jezgrima**

Za svaki model izračunati su sledeći metrički pokazatelji: **Tačnost**, **Preciznost**, **Odziv**, **F1-Score**.



Primena linearne regresije na iste podatke za predikciju kontinuiranih vrednosti:

* Mean Squared Error (MSE): ~0.26

Konfuzione matrice prikazuju performanse klasifikacionih modela na test skupu:

* **Logistička regresija:**
  + Vizualizacija matrice pokazuje tačne i pogrešne klasifikacije.
* **SVM (linearna, RBF i polinomska jezgra):**
  + Svaka matrica prikazuje uspešnost modela u predikciji klasa "Nizak" i "Visok" životni vek.
  + **SVM (linearna jezgra)** ima najbolje ukupne performanse sa najvećim brojem tačno klasifikovanih instanci (33/60).

Faza 6 - Napredno mašinsko učenje

**K-Means** algoritam je korišćen za klasterovanje podataka nakon redukcije dimenzionalnosti pomoću PCA (2 komponente):

* **Broj klastera:** 2
* Vizualizacija prikazuje klastere u PCA prostoru.
* Konfuziona matrica pokazuje usklađenost klastera sa binarnim ciljem (nizak vs visok životni vek).

**Stablo odlučivanja** je implementirano i evaluirano na test skupu:

* **Rezultat:**
  + Tačnost: ~0.6
  + Preciznost: ~0.6
  + Odziv: ~0.6
  + F1-Score: ~0.6
* **Vizualizacija:** Konfuziona matrica prikazuje performanse modela u predikciji klasa.

**Neuronska mreža** sa 2 sakrivena sloja je implementirana za binarnu klasifikaciju:

* Tačnost: ~0.87
* Grafikon greške tokom treninga i validacije prikazuje stabilnost modela kroz epohe.
* Scatterplot prikazuje stvarne naspram predikovanih vrednosti.

**Poboljšana neuronska mreža** sa regularizacijom (L2) i dropout slojevima je implementirana radi boljeg generalizovanja:

* Tačnost: ~0.55
* Grafikon greške tokom iteracija prikazuje smanjenje pretreniranja zahvaljujući regularizaciji.
* Konfuziona matrica prikazuje preciznost klasifikacije.