Plan rada za projekt: Analiza i međuzavisnost podataka o građanima i automobilima

1. Uvod i definiranje ciljeva projekta

- **Opis projekta:**
  Cilj projekta je analizirati i obraditi podatke iz tri baze podataka u .csv formatu koje sadrže informacije o ljudima, njihovim prihodima i posjedovanju automobila te ih upariti s bazama podataka o prodaji i specifikacijama vozila. Projekt se provodi u programskom jeziku Python.
- **Glavni ciljevi:**
  1. Priprema i sistematizacija podataka iz svih baza za daljnju obradu.
  2. Vizualizacija ključnih parametara radi prepoznavanja zakonitosti i obrazaca u podacima.
  3. Provedba statističke analize za donošenje zaključaka temeljenih na podacima.
  4. Komparacija vizualnih i statističkih rezultata radi zaključivanja o korelacijama i mogućim uzorcima.

2. Faze projekta

2.1. Prikupljanje i priprema podataka

- **Zadatak:**
  Sakupljanje podataka iz tri glavne baze:
  - Baza podataka o ljudima (zanimanje, godišnja primanja, procijenjena razina financijskog statusa, posjedovanje automobila, itd.).
  - Baza podataka o prodaji vozila (spol kupca, novčana primanja, marka, model, cijena, datum prodaje, itd.).
  - Baza tehničkih specifikacija vozila i cijena(marka, model, godina proizvodnje, prijeđeni kilometri, tip goriva, itd.).

- **Aktivnosti:**

1. Čišćenje podataka:
    - Uklanjanje dupliciranih zapisa.
    - Popunjavanje ili uklanjanje nedostajućih vrijednosti.
    - Provjera konzistentnosti podataka (npr. povezivanje marki i modela u svim bazama).
2. Standardizacija podataka:
    - Transformacija formata podataka (npr. datumi, decimalni brojevi, valuta).
    - Normalizacija kategorijskih vrijednosti (npr. "da/ne" za posjedovanje automobila).
3. Povezivanje tablica:
    - Jedinstvena identifikacija za uparivanje podataka među bazama.

2.2. Istraživačka analiza podataka i vizualizacija

- **Zadatak:**
  Istražiti osnovne odnose među podacima putem vizualizacija.
- **Aktivnosti:**
    1. Analiza distribucije pojedinačnih parametara:
        - Histogrami za godišnja primanja, cijene vozila, godišta automobila.
        - Stupčasti grafovi za kategorijske varijable poput zanimanja i posjedovanja vozila.
    2. Analiza korelacija:
        - Scatter plot između cijene vozila i godišnjih primanja.
        - Scatter plot između godišta automobila i cijene.

- Grupni grafovi koji prikazuju usporedbu prema spolu, zanimanju ili drugim kategorijama.

2.3. Statistička analiza

- **Zadatak:**
  Provjeriti hipoteze i kvantificirati odnose između ključnih varijabli putem statističkih testova.
- **Planirana statistička analiza:**
  1. **Korelacijska analiza:**
     - Pearsonov i Spearmanov koeficijent korelacije za kvantitativne varijable (primjerice, cijena vozila i godišnja primanja).
  2. **Testovi razlika:**
     - T-test ili Mann-Whitney U test za usporedbu godišnjih primanja ljudi koji posjeduju automobil i onih koji ne posjeduju.
     - ANOVA test za usporedbu cijena vozila između različitih kategorija zanimanja.
  3. **Regresijska analiza:**
     - Jednostavna ili višestruka linearna regresija za analizu utjecaja više faktora (npr. zanimanja, spola i godišta automobila) na cijenu vozila.
  4. **Distribucijski testovi:**
     - Shapiro-Wilk ili Kolmogorov-Smirnov test za procjenu normalnosti distribucije varijabli poput primanja ili cijene vozila.

2.4. Evaluacija rezultata i zaključci

- **Zadatak:**
  Analizirati rezultate vizualizacije i statističke analize kako bi se donijeli zaključci.
- **Aktivnosti:**
  1. Usporedba rezultata statističkih testova s vizualnim obrascima.

2. Identifikacija ključnih korelacija i uzoraka.
3. Donošenje zaključaka o potvrdama ili promjenama početnih hipoteza.

2.5. Dodatne funkcionalnosti (opcionalno)

- **Izrada interaktivnog *dashboarda*:**
  - Kreiranje web sučelja za prikaz podataka i rezultata.
  - Integracija tablica, grafova i rezultata analiza u interaktivnom obliku.

3. Tehnička implementacija

- **Tehnologije:**
  - Jezik: Python.
  - Biblioteke: pandas (za rad s podacima), matplotlib i seaborn (za vizualizaciju), scipy.stats (za statističku analizu).
- **Struktura koda:**
  - Modul za obradu i čišćenje podataka.
  - Modul za generiranje vizualizacija.
  - Modul za statističku analizu.

4. Očekivani rezultati

- Dobro organizirane baze podataka pogodne za analizu.
- Vizualizacije koje jasno prikazuju osnovne odnose i obrasce među podacima.
- Statistički zaključci koji potvrđuju ili osporavaju hipoteze.
- Dokumentacija svih koraka i rezultata analize.

5. Moguća proširenja

- Razrada interaktivnog sučelja za pregled rezultata (dashboard).
- Primjena strojnog učenja za predikciju (npr. predikcija cijene vozila na temelju podataka).

6. Zaključak

Projekt analize i obrade podataka pružit će vrijedne uvide u odnose između karakteristika ljudi i vozila. Provedeni koraci omogućit će kvalitetnu statističku obradu i vizualizaciju, što će rezultirati preciznijim zaključcima i potencijalnim uvidima u nove zakonitosti među podacima.

Filip Jovanović i Marko Putić