Exploratory Data Analysis

Projektinis darbas

Trukmė: 16 valandų (2022.12.05 - 2022.12.08)
Duomenų šaltinis: atviri juridinių asmenų registro duomenys
Maximalus taškų skaičius: 10 tšk (2.3+5.2+2.5)

Svarbu: kiekvienos dienos pabaigoje įrašykite į grupės susirašinėjimą savo progresą - ties kuriuo uždaviniu sustojote. Notebook .ipynb failas bus pagrindinis atsiskaitymo failas, tad atlikdami užduotis formatuokite notebooką pagal gerąsias README praktikas. Failą pavadinkinkite vardas_pavarde.ipynb ir įdėkite į dėstytojo nurodytą folderį.
Jeigu neaiški užduoties formuluotė, klauskite dėstytojo, kad patikslintų.

Pirma dalis

Dalis verta 2.3 taškų.

1. Iš pridėto *requirements.txt* failo sukurti anacondos aplinką ir žemiau esančias užduotis vykdyti šioje aplinkoje. **[0.25 taško]**
2. Parsisiųsti 2020 ir 2021 įmonių “Juridinių asmenų pateikti finansinės atskaitomybės dokumentai – balanso ataskaitos” ir “Juridinių asmenų pateikti finansinės atskaitomybės dokumentai – pelno (nuostolių) ataskaitos” duomenis. **[0.25 taško]**
3. Susipažinkite su duomenimis **[1.8 taško]**:
    a. Ką reiškia skirtingi duomenų laukai? Iš kokių atributų sudaryti duomenys?
    b. Kiek reikšmių ir požymių turi kiekvienas iš dokumentų?
    c. Kokio duomenų tipo yra kiekvienas iš požymių? Ar duomenų tipai tarp atitinkamų lentelių iš 2020 ir 2021 metų sutampa?
    d. Paaiškinkite kas yra *object* duomenų tipas ir kuo jis skiriasi nuo *string* duomenų tipo.
    e. Kuri lentelė turi daugiausiai nežinomų verčių? Kuris požymis išsiskiria nežinomų verčių skaičiumi?
    f. Ar kažkuri lentelė turi pasikartojančių duomenų?
        i. Kokia stulpelių aibė vienareikšmiškai identifikuoja kiekvienos lentelės eilutę?

Antra dalis

Dalis verta 5.2 taškų.

4. Keletas juridinių asmenų pavadinimų yra tiesiogiai susiję su “Thermo Fisher” bendrove. Raskite pilnus šių bendrovių pavadinimus. **[0.4 taško]**

5. Apsimeskime, kad atributas “form_pav” neegzistuoja. Iš likusių laukelių gaukite visas UAB ir MB bendroves. Palyginkite rekonstrukcijos tikslumą su “form_pav” vertėmis - kiek procentų verčių pavyko atkurti skirtingose duomenyse? **[0.4 taško]**

6. Apskaičiuoti vidutinias reikšmes “pelnas_pries_apmokestinima” ir “nuosavas_kapitalas” laukų kiekvienai juridinio asmens formai (Uždaroji akcinė bendrovė, Mažoji bendrija, etc.) už visą laikotarpį (2020 ir 2021 metai kartu). **[0.4 taško]**

7. Atrinkti unikalias UAB ir MB įmones, kurių “nuosavas_kapitalas” buvo didesnis negu visų UAB ir MB įmonių nuosavo kapitalo mediana už visą laitotarpį (2020 ir 2021 metai kartu). **[0.4 taško]**

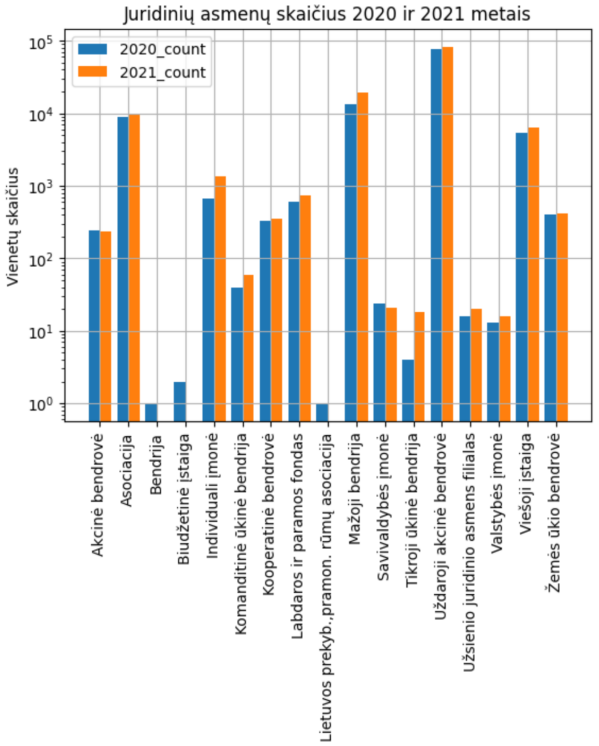
8. Įvertinkite kurios įmonės išsiskiria ypač didele “nuosavas_kapitalas” verte (patenka į 99th percentilę). Patikrinkite įmonių patenkančių į TOP-10 vertes rekvizitai.vz.lt svetainėje. Ranka įrašomi duomenys sukuria klaidos galimybę. Ar yra bendrovė, kurios “nuosavas_kapitalas” jums kelia įtarimų? **[0.4 taško]**

9. Vizualiai atvaizduoti juridinių asmenų formos pasiskirstymą per visą laikotarpį (2020 ir 2021 metais kartu). Vizualizacijos turi sutapti su pavaizduota dešinėje. **[0.4 taško]**

    a. naudojantis *matplotlib* funkcionalumu.

    b. naudojantis pandas funkcionalumu.

    c. naudojantis *seaborn* funkcionalumu.



10. Kurį mėnesį buvo įregistruota daugiausiai juridinių asmenų 2020 ir 2021 metais? Atvaizduokite stulpelinę diagramą. **[0.4 taško]**

11. Vizualiai atvaizduoti juridinių asmenų statuso pasiskirstymą kiekvienais metais. **[0.4 taško]**

12. Kiekvienai įmonei apskaičiuoti tendenciją. Jei “grynasis_pelnas” + “nuosavas_kapitalas” 2021 metais buvo didesnis už 2020 metais deklaruotą, tai tendencija “teigiama”, jei ne - “neigiama”. Atvaizduoti tendencijos duomenis stulpelinėje diagramoje. **[0.4 taško]**

13. Surasti UAB ir MB įmones, kurios 2020 vykdė veiklą, o 2021 - bankrutavo. Bankrutavusių įmonių statusai: “Bankrutavęs”, “Bankrutuojantis”, “Išregistruotas”, “Inicijuojamas likvidavimas”, “Likviduojamas”, “Likviduojamas dėl bankroto”. Toliau veiklą vykdanti įmonė pažymėta "Teisinis stat neįregistruotas" statusu. Kokia bankrutavusių įmonių “nuosavas_kapitalas”suma 2021 metais? Kokia bankrutavusių įmonių “grynasis_pelnas” suma 2021 metais? Palyginti šias sumas su veikiančių įmonių atitinkamomis sumomis. **[0.4 taško]**

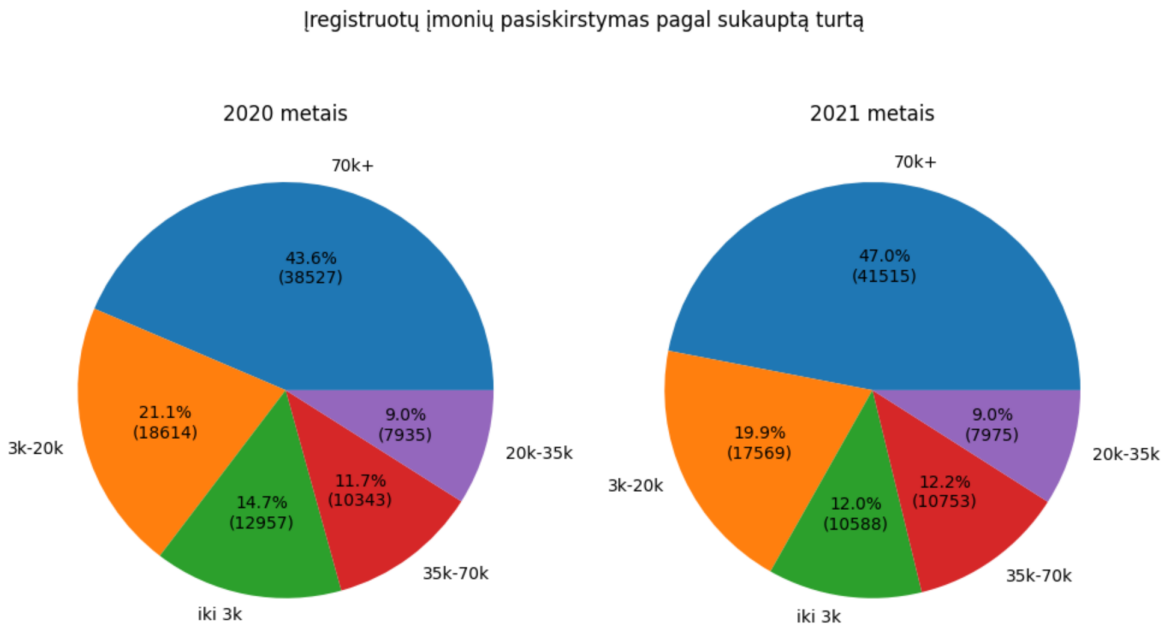
14. Sukurti naują stulpelį “turtas”, kuris skaičiuojamas sudedant “trumpalaikis_turtas” ir “ilgalaikis_turtas” **[0.4 taško]**.

    a. Suskirstyti įmones į 5 grupes pagal lauką “turtas”.

    b. Palyginti kaip šios grupės keičiasi 2020 ir 2021 metais. Palyginimą atvaizduoti skrituline diagrama ir gauti tokį patį kaip apačioje pavaizduotas paveikslas (*skaitinės reikšmės gali skirtis*):

        i. naudojantis *matplotlib* funkcionalumu.

        ii. naudojantis pandas funkcionalumu.



15. Rasti TOP-100 pelningiausių įmonių pagal “grynasis pelnas” ir TOP-100 įmonių kurių “nuosavas kapitalas” didžiausias. Ar yra sutampančių įmonių šiuose sąrašuose?

Skaičiuojant TOP įmonės įtraukti viso laikotarpio įmones (naudoti 2020 ir 2021 metų duomenis kartu). **[0.4 taško]**

16. Ar TOP įmonės pagal “grynasis pelnas” iš 15 užduoties skiriasi pagal metus, t.y. 2020 ir 2021 metais? Kurios įmonės iškrenta iš TOP sąrašo 2021 metais? **[0.4 taško]**

Trečia dalis

Dalis verta 2 taškų.

Jūs norite įsigyti labai mažą arba mažą įmonę.

1. Jums prieinamą finansinę informaciją papildykite (https://atvira.sodra.lt/imones/rinkiniai/index.html)
    a. darbuotojų skaičiumi (2020, 2021 metais)
    b. atlyginimų istoriją.
    c. veiklos sritimi.
2. Galutinis duomenų masyvas turi atrodyti kaip **3_uzduotis_data_sample.csv. [0.5 taško]**
    a. Svarbu atkreipti dėmesį į finasinių duomenų ir sodros duomenų datas ir galimus prasilenkimus. Rekomeduojam vidutinį darbuotojų skaičių imti finansinių metų paskutinį mėnesį.

Užduotys

1. Išsiaiškinkite micro ir mažos įmonės apibrėžimą.
2. Remdamiesi apibrėžimu išfiltruokite micro ir mažas įmones. Įmonės turi turėti statusą MB arba UAB.
3. Indentifikuokite labiausiai augusias įmones. **[1 taškas]**
    a. Pastaba: Nepamirškite atsižvelgti į išimtis (padidinti apyvartą nuo 10 eurų iki 10000 yra daug lengviau nei nuo 10000 iki 1000000). Augimą gali indentifikuoti daugiau nei viena dimensija, pvz. apyvarta, darbuotojų skaičius, darbuotojų atlyginimų didėjimas, maržos didėjimas ir kt.
4. Sukurkite kriterijus, ir atrinkite top 10, **jūsų** manymu geriausių įmonių. **[0.5 taško]**
5. Kokia suma įvertintumėt savo atrinktas įmones?

Pabaiga

Išeksportuokite naudotą anaconda aplinką į naujai sukurtą requirements.txt failą. Notebook failą pavadinkinkite vardas_pavarde.ipynb ir įdėkite į dėstytojo nurodytą folderį.