

Основне академске студије
Информациони инжењеринг

Методе и технике науке о подацима

Обрада и анализа података у језику Python

(материјали за вежбе)

Одабране функције

Библиотека *NumPy*

функција **`numpy.multiply(...)`**

множење низова елемент по елемент

функција **`numpy.around(...)`**

заокруживање вредности

функција **`numpy.linspace(...)`**

формирање низа равномерно распоређених вредности у датом интервалу

функција **`numpy.repeat(...)`**

умножавање елемената низа

Одабране функције

Библиотека *pandas*

функција **`pandas.read_csv(...)`**

учитавање података из датотеке у формату CSV

особина **`pandas.DataFrame.columns`**

издвајање ознака колоне из скупа података

функција **`pandas.cut(...)`**

распоређивање вредности по интервалима

функција **`pandas.Series.astype(...)`**

трансформација серије у задати тип података

функција **`pandas.Series.value_counts(...)`**

пребрајање вредности по различитим елементима серије

функција **`pandas.Series.isin(...)`**

провера заступљености елемената серије међу датим вредностима

Библиотека *pandas*

функција **`pandas.DataFrame.rename(...)`**

промена ознака оса за скуп података

функција **`pandas.Index.to_series(...)`**

трансформација индекса у серију

Одабране функције

Библиотека *Matplotlib*

функција **`matplotlib.pyplot.plot(...)`**

исцртавање графика с линијама и ознакама у две димензије

функција **`matplotlib.pyplot.step(...)`**

исцртавање степеничастог графика

функција **`matplotlib.pyplot.subplot(...)`**

манипулација простором за цртање с координатним системом

функција **`matplotlib.pyplot.xlim(...)`**

манипулација распоном *X* осе

функција **`matplotlib.pyplot.ylim(...)`**

манипулација распоном *Y* осе

класа **`matplotlib.patches.Patch`**

дводимензионални елемент с бојама

Библиотека *Matplotlib*

функција **`matplotlib.colormaps.get_cmap(...)`**

очитавање мапе боја

особина **`matplotlib.colors.ListedColormap.colors`**

издвајање боја из мапе боја

Подаци

скуп података о продаји моторног горива у америчкој савезној држави Ајова (САД)

постоје 1273 појаве и 9 обележја

свака појава обухвата податке за једну календарску годину и један округ обележја за годину, округ, број продајних локација, количину продатог бензина с етанолом у галонима, количину продатог бензина без етанола у галонима...

извор скупа података

скуп података *Iowa Motor Fuel Sales by County and Year*

матична Интернет страна

https://data.iowa.gov/Sales-Distribution/Iowa-Motor-Fuel-Sales-by-County-and-Year/hbwp-wys3/about_data

у оквиру портала отворених података савезне државе Ајова (САД)

подаци доступни у више формата

може се користити формат CSV (CSV датотека доступна кроз опцију *Export*)

скуп података креиран 4. 12. 2014. а најновије повезано ажурирање 4. 4. 2025.

назначено одељење које је објавило скуп података је *Iowa Dept of Revenue*

назначени власник скупа података је *Estelle Montgomery*

лиценца и могућности коришћења скупа података

CC0 1.0 Universal

<https://creativecommons.org/publicdomain/zero/1.0/legalcode>

Задатак 1.

Учитати скуп података из одговарајуће CSV датотеке.

Уклонити из учитаног скупа података све редове у којима су недостајуће вредности присутне.

Проверити садржај учитаног скупа података, као и то какве су колоне садржане.

Из назива колоне уклонити сувишне беле знакове.

Задатак 2.

Унутар учитаног скупа података задржати само обележја о години, округу, броју продајних локација, количини продатог бензина без етанола, количини продатог бензина с етанолом, количини продатог прозирног и обојеног дизела, количини продатог прозирног и обојеног биодизела и количини продатог чистог биодизела.

Преименовати колоне скупа података тако да уместо изворних буду коришћени преведени називи.

Проверити садржај скупа података, као и то какве су колоне садржане.

Задатак 3.

Унутар учитаног скупа података задржати само редове о оним окрузима за које су подаци доступни за сваку од година што се помиње у скупу података.

Проверити садржај скупа података.

Задатак 4.

У скуп података додати ново обележје о бројности продајних локација, а вредности тог обележја извести из одговарајућих вредности постојећег обележја о броју продајних локација.

Вредност *mala* код новог обележја одговара вредностима постојећег обележја које не прелазе медијалну вредност.

Вредност *velika* код новог обележја одговара вредностима постојећег обележја које прелазе медијалну вредност.

Колону за ново обележје позиционирати непосредно иза колоне за постојеће обележје о броју продајних локација.

Проверити садржај скупа података, као и то какве су колоне садржане.

Задатак 5.

У скупу података изменити обележја о продаји горива тако да вредности уместо у галонима буду дате у литрама.

Нове вредности измењених обележја о продаји горива заокружити на нула децимала.

Проверити садржај скупа података.

Задатак 6.

За податке о округу и броју продајних локација спровести прво груписање на основу округа, затим агрегирање бројчаних вредности помоћу функције за аритметичку средину, а онда и сортирање по агрегираним бројчаним вредностима у обрнутом редоследу.

Приказати груписане, агрегиране и сортиране податке за првих пет округа с највећим просечним бројем продајних локација.

Задатак 7.

Формирати подскуп података који се односи само на 2019. годину.

Утврдити која је највећа количина продатог бензина с етанолом унутар подскупа података и приказати је.

Утврдити која је највећа количина продатог бензина без етанола унутар подскупа података и приказати је.

...

Задатак 7.

...

На основу подскупа података нацртати графикон расејања тако да X оса одговара количини продатог бензина с етанолом а Y оса количини продатог бензина без етанола.

Подесити величину тачке на графикону тако да одговара броју продајних локација.

Подесити и X и Y осу на графикону тако да прикладно покривају распоне приказаних вредности.

Подесити наслов графикона и називе оса.

Задатак 8.

Посматрати подскуп података који се односи само на 2019. годину.

Над подскупом података израчунати и приказати дескриптивне статистике за број продајних локација.

На основу подскупа података одредити колико округа се по свом броју продајних локација уклапа у следеће опсеге: $[0, 50)$, $[50, 100)$, $[100, 150)$ и $[150, 200)$.

...

Задатак 8.

...

На основу подскупа података нацртати тортни графикон који одговара расподели округâ по датим опсезима броја продајних локација.

Подесити ознаке поред исечака на графикону тако да буду приказани одговарајући опсези броја продајних локација.

Подесити ознаке на самим исечцима тако да буду приказани одговарајући удели у процентима с прецизношћу од две децимале.

Исечке одмакнути једне од других.

Подесити боје исечака користећи неку од уграђених мапа боја.

Подесити наслов графикона.

Задатак 9.

Посматрати подскуп података који се односи само на 2019. годину.

Утврдити која је највећа количина продатог прозирног и обојеног биодизела унутар подскупа података и приказати је.

Утврдити која је највећа количина продатог чистог биодизела унутар подскупа података и приказати је.

...

Задатак 9.

...

На основу подскупа података нацртати два хистограма за количину продатог биодизела, при чему се један хистограм односи на податке за прозирни и обојени биодизел а други на податке за чисти биодизел.

За оба хистограма користити идентичне границе интервала: оквирна ширина појединачног интервала је 20000000 а први интервал креће од 0.

Хистограме треба приказати тако да буду један до другог.

Подесити за оба хистограма Y осу тако да прикладно покрива распон одговарајућих вредности.

Подесити наслов графикана и називе оса за оба хистограма.

Задатак 10.

Формирати подскуп података који се односи само на 2019. годину и округе чији назив почиње великим латиничним словом С.

Проверити називе округа унутар подскупа података.

Над подскупом података израчунати и приказати дескриптивне статистике за количину продатог прозирног и обојеног дизела.

...

Задатак 10.

...

На основу подскупа података нацртати стубичасти графикон који приказује количину продатог прозирног и обојеног дизела по појединачним окрузима.

Стубове на графикону обојити тако да је могуће разликовати округе по бројности продајних локација.

Подесити на графикону текстуалне ознаке за округе испод стубова и позиционирати те ознаке тако да буду читљиве.

Подесити Y осу на графикону тако да прикладно покрива распон одговарајућих вредности.

Подесити наслов графикона и називе оса.

Поставити одговарајућу легенду.

Задатак 11.

За податке о години, количини продатог бензина без етанола и количини продатог бензина с етанолом спровести груписање на основу године а потом и агрегирање количина продатог горива помоћу функције за сабирање.

Приказати груписане и агрегиране податке.

...

Задатак 11.

...

На основу груписаних и агрегираних података нацртати линијски графикон где X оса одговара години а Y оса количини продатог горива тако да подаци за бензин без етанола и бензин с етанолом буду приказани засебним линијама.

За линије на графикону користити различите боје и различите врсте ознаке, а такође подесити прозирност линија на графикону.

Подесити Y осу тако да почетна вредност буде 0.

Подесити наслов графикона и називе оса.

Поставити одговарајућу легенду.

...

Задатак 11.

...

На основу груписаних и агрегираних података нацртати степеничасти графикон где X оса одговара години а Y оса количини продатог горива тако да подаци за бензин без етанола и бензин с етанолом буду приказани засебним линијама.

За степеничасте линије на графикону користити различите боје и различите врсте линије, а такође подесити прозирност степеничастих линија на графикону.

Подесити Y осу тако да почетна вредност буде 0.

Подесити наслов графикона и називе оса.

Поставити одговарајућу легенду.

Основна литература

NumPy. NumPy. Internet: <https://numpy.org/>

NumPy. NumPy documentation — NumPy v2.3 Manual. Internet: <https://numpy.org/doc/stable/index.html>

pandas. pandas - Python data analysis library. Internet: <https://pandas.pydata.org/>

pandas. pandas documentation — pandas 2.3.0 documentation. Internet: <https://pandas.pydata.org/docs/>

Matplotlib. Matplotlib — Visualization with Python. Internet: <https://matplotlib.org/>

Matplotlib. Using Matplotlib — Matplotlib 3.10.3 documentation. Internet: <https://matplotlib.org/stable/users/index>