KORRELYATSIYA VA BOG‘LIQLIK KO‘RSATKICHLARINI OSHIRISH.



Matchonov Shohrux Matkarim o‘g‘li
Jizzax shahridagi Qozon (Vo‘lgabo‘yi) federal universiteti filiali
“Raqamli ta’lim texnologiyalari markazi” boshlig‘i.
shohruhmatchonov@gmail.com

Annotatsiya.

Mazkur maqolada python dasturlash tilining pandas kutubxonasida mavjud imkoniyatlardan foydalanib real ma’lumotlar bazalari taxlili asosida ustunlarning o‘zaro bog‘liqligini topish va bog‘liqlik ko‘rsatkichlarini oshirish masalasi qaraladi.

Kalit so‘zlar: Seriya, DataFrame, Json, csv, mediana, moda, mean, kavariatsiya, korrelyatsiya, regressiya.

Аннотация.

В данной статье рассматривается проблема поиска взаимосвязей столбцов и повышения показателей корреляции на основе анализа реальных баз данных с использованием возможностей библиотеки pandas языка программирования Python.

Ключевые слова: Seriya, DataFrame, Json, CSV, медиана, moda, mean, ковариация, корреляция, регрессия.

Annotation.

This article examines the problem of finding column relationships and increasing correlation indices based on the analysis of real databases using the capabilities of the pandas library of the Python programming language.

Keywords: Series, DataFrame, Json, CSV, median, moda, mean, covariance, correlation, regression.

Dunyo axborot hamjamiyatining rivojlanishi katta massivli axborot resurslarini hosil bo‘lishi va izchillik bilan rivojlanishi ma’lumotlarga ishlov berish mexanizmlarini takomillashtirish muhim masalalardan biri bo‘lib qolmoqda. Chunki ma’lumotlarni tahlillash, turli kesimlarni qamrab olish, chuqur tahlil asosida istiqbolli qaror muqobillarini ishlab chiqish predmet soha bo‘yicha hosil qilinadigan bilimlar (meta ma’lumotlar) sifatiga bog‘liq bo‘ladi. Sifatli shakllantirilgan bilimlar asosida axborot analitiklari faoliyati, tahlillari hamda predmet soha bo‘yicha beradigan xulosalari hatoliklarining kamaytirilishiga erishiladi. Bu masala globallashuv sharoitida shakllanayotgan ijtimoiy-iqtisodiy munosabatlarning to‘g‘ri yo‘lga qo‘yilishi, keng qamrovli tahlillash mexanizmlarini yo‘lga qo‘yishda mashinali o‘qitish usul va vositalarini zamonaviy axborot texnologiyalari asosida akslantirish hamda turli krosplatforma tizimlarida joriylashtirishga bo‘lgan talabni kuchaytirmoqda.[1]

Zamonaviy axborot-kommunikasiya texnologiyalarining hozirgi kundagi rivojlanish tendensiyasi ma’lumotlarga ishlov berishga yangicha qarash, o‘zgacha yondashuvlarni ishlab chiqishni talab etadi. Ma’lumki global axborot makoni katta massivdagi ma’lumotlarni bir vaqtning o‘zida ishlash tamoiliga asoslanadi. Bir vaqtning o‘zida katta xajmli ma’lumotlarni tahlillash (ularga dastlabki ishlov berish, tasniflash, bashoratlash, tahlillash) masalalarini echish tezkor tahlillashni amalga oshirish bilan bog‘liq bir qator muammolarni hal etish jarayoni hisoblanib, bu jarayon bir nechta qism tabaqali yechim talab qiluvchi masalalarni tarkibiga oladi. Bulardan eng muximi va murakkabi - dastlabki ishlov berish jarayoni hisoblanadi. Ma'lumotlarga dastlabki ishlov berishni samarali tashkil etish, ishlov berish mexanizmi unga ishlov berish jarayoni uchun kam vaqt va resurs sarflanishiga olib keladi. Bundan tejalgan mashina resurslari masalaning qolgan qismini echish uchun yo‘naltirilishi mumkin. Bundan dastlabki ishlov berishning o‘zini ham bir qator qism masalalarga ajratib olish maqsadga muvofiq bo‘lib, uni yechish uchun ham turlicha yondashuvlar qo‘llaniladi. Mazkur yondashuvlar kiruvchi ma’lumotlar turiga qarab tanlanishi yoki turli tipli ma’lumotlarga moslashuvchan bo‘lishi maqsadga muvofiq. Aynan shunday yondashuv kiruvchi ma’lumot massivi xajmiga mos ravishda kuchsiz yoki kuchli, murakkab yoki soddaroq mexanizm qo‘llash imkonini beradi. Soddaroq mexanizm mashina resurslariga murojaatni kamaytirsa, murakkab va kuchli mexanizm

resurslardan o‘z o‘rnida va vaqtida foydalanishga asos bo‘ladi. Xuddi shunday bu tasnif, kirayotgan dastlabki ma'lumotlar masalani murakkabligiga qarab o‘zgartirilishi bu moslashuvchan intelektual dastlabki ishlov berish mexanizmni shakllantirish imkoniyatini yaratadi. [2]

Asosiy qism: Biz ko‘pincha ikki yoki undan oqrtiq o‘zgaruvchilar o‘rtasidagi munosabatni o‘rnatishimiz yoki munosabat borligini aniqlashimiz zarurati paydo bo‘ladi. Bunday hollarda biz korrelyatsiya va regressiya masalalarini qo‘llashimiz zarurati paydo bo‘ladi.

Korrelyatsiya – ikki yoki undan ortiq o‘zgaruvchilar o‘rtasidagi munosabatni belgilab beradi.

Regressiya – munosabatlarning mohiyatini ya’ni munosabatga asoslanib bashorat qilish imkoniyatini beradi.

Ushbu maqolada biz regressiya masalasini ko‘rib o‘tamiz. Buning uchun bizga pythondagi pandas kutubxonasi kerak bo‘ladi. Biz Jizzax viloyati “toqimachilik maxsulotlarini ishlab chiqarish” va “kiyim ishlab chiqarish” ustunlari o‘rtasidagi bog‘liqlikni o‘rganamiz. Ushbu ma’lumotlar2016 va 2030-yillar ko‘rsatkichlarini o‘z ichiga oladi. Ma’lumot juda katta bo‘lganligi sababli biz uning “Sanoat” yo‘nalishidagi “Ishlab chiqaradigan sanoat” bo‘limi ma’lumotlaridan foydalandik va biz ma’lumotlarini “Data.csv” fayliga import qilib oldik. Biz pandas kutubxonasidagi *read_csv* yoki *read_json* funksiyasidan foydalanib ma’lumotlarni o‘rganishimiz mumkin. [3]

```
Z=pd.read_csv("Data.csv")
print(Z.to_string())
```

	yil	toqimachilik maxsulotlarini ishlab chiqarish	kiyim ishlab chiqarish
0	2016	168	42
1	2017	304	57
2	2018	468	88
3	2019	390	120

1-rasm. Ma’lumotlarning qisqa korinishi.

Pandas kutubxonasidagi ma’lumotlar to‘plamlari odatda DataFrame deb ataladigan ko‘p o‘lchovli jadvallarda taxlil qilinadi. Series bitta ustunni, DataFrame esa butun jadvalli taxlil qiladi.

```
Z=pandas.read_csv("Data.csv")
```

Ma’lumotlarimizni Json yoki CSV fayldan yuklab olganimizdan keyin dastlabki ishlov berish bosqichiga o‘tamiz. Dastlabki ishlov berish bosqichlari quyidagilar:

- Bo‘sh qiymatlarni qatori bilan o‘chirish:

```
Z.dropna(inplace=True)
```

- O‘rtacha qiymat bilan almashtish:

```
y=Z["ustun nomi"].mean()
```

$$y = (\sum_{i=0}^{n} x_i) / n$$

- Mediana bilan almashtirish:

```
y=Z["ustun nomi"].median()
```

Raqamlarni eng kichikdan kattagacha tartiblab, agar raqamlar soni toq bo'lsa, o'rta qiymat mediana hisoblanadi, aks xolda mediana bizda mavjud bo'lgan ikkita o'rta qiymatning o'rtacha qiymati hisoblanadi.

- Mode(chastota) bilan almashtirish:

```
y=Z["ustun nomi"].mode()[0]
Z["ustun nomi "].fillna(y, inplace=True)
```

Eng ko‘p uchraydigan qiymat bilan almashtirish.

- Almashtirishdan keyin dublikatlarni tozalash bosqichiga o‘tamiz, ya’ni bir xil qatorlarni olib tashlash bosqichi:

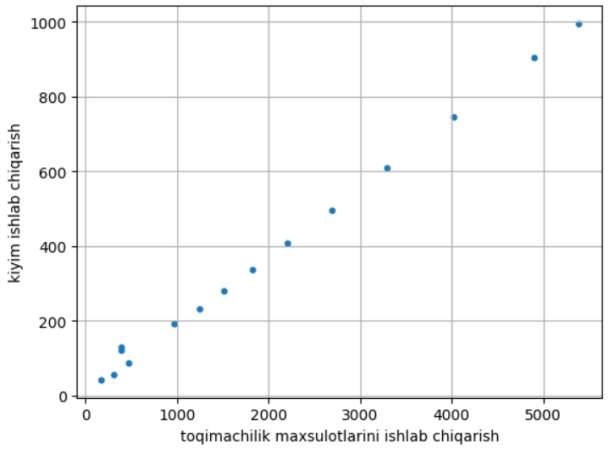
```
Z.drop_duplicates(inplace=True)
```

Yuqoridagi bosqichlarni amalga oshirmasdan turib korrelyatsiya ko‘rsatkichini olib bo‘lmaydi, ya’ni ustunlar orasidagi bog‘liqlikni topib bo‘lmaydi. Ikki o‘zgaruvchi o‘rtasidagi

korrelyatsiya mavjudligini aniqlashning eng oson usullaridan biri ularni taqqoslash sxemasida ko‘rib chiqishdir. [4]

```
Z=pandas.read_csv("Data.csv")
x=Z["toqimachilik maxsulotlarini ishlab chiqarish"]
y=Z['kiyim ishlab chiqarish']
pandas.DataFrame(numpy.array([x,y]).T).plot.scatter(0, 1, s=12, grid=True)
plt.xlabel('toqimachilik maxsulotlarini ishlab chiqarish')
plt.ylabel('kiyim ishlab chiqarish')
plt.show()
```

Ushbu misol quyidagi grafikni xosil qiladi.



2-rasm. Ikkita ustunning o‘zaro bog‘liqlik grafigi.

Ikki o‘zgaruvchi o‘rtasidagi munosabatlarning kuchini aniqlashning bir usuli bu ularning kavariatsiyasini topishdir. Kavariatsiya ikki o‘zgaruvchining birgalikda o‘zgarishini o‘lchaydi. Agar bizda X va Y ustun raqamlari bo‘lsa ularning o‘rtacha qiymatdan og‘ishlari quyidagi formula orqali hisoblanadi

$$dX_i = xi - \bar{x}$$
$$dY_i = yi - \bar{y}$$

Bu yerda x_i – X ustunining i elementi, y_i-Y ustunining i elementi, $\bar{x}$ va $\bar{y}$ lar X va Y ustunlarining o‘rtacha qiymatlari. Agar X va Y birgalikda o‘zgarishga moyil bo‘lsa u xolda ularning o‘rtacha qiymatdan og‘ishlari bir xil bo‘ladi. Kovariatsiya quyidagi formula orqali xisoblanadi.

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n} dx_i\, dy_i$$

Pythonda biz cov funksiyasidan foydalanishimiz mumkin

```
Natija=Z["toqimachilik maxsulotlarini ishlab chiqarish"].corr(Z["kiyim ishlab chiqarish"])
```

Kovariatsiyadan chiqgan natijani tushintirish qiyin. Ko‘pincha Kovariatsiya indeksini xususiy xollar uchun natija sifatida qaraladi. Kovariatsiyani mos ravishda standart og‘ishlarga bo‘lish orqali natijani -1 va 1 oraliqdagi raqamga aylantirishimiz mumkin. Bu natija Person korrelyatsiyasi deyiladi va formula quyidagicha bo‘ladi. [5]

$$r = \frac{1}{n}\sum_{i=1}^{n} \frac{dx_i}{\delta_x}\frac{dy_i}{\delta_y}$$

Bu yerda δ_y va δ_x mos ravishda X va Y ning standart og‘ishlari. δ_x va δ_y lar quyidagi formulalar orqali xisoblangan:

$$\delta_x = \sqrt{\frac{\sum_{i-1}^{n}(xi-\bar{x})^2}{n-1}} \qquad \delta_y = \sqrt{\frac{\sum_{i-1}^{n}(yi-\bar{y})^2}{n-1}}$$

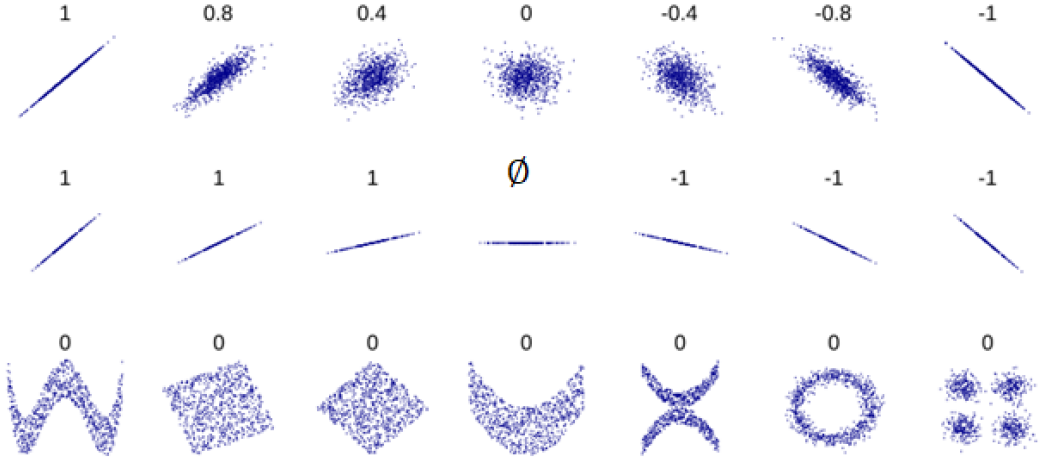
X va Y o‘zgaruvchilar uchun standart og‘ish constant(o‘zgarmas) bo‘lgani uchun korrelyatsiya formulasini quyidagicha soddalashtirish mumkin.

$$r = \frac{cov(X,Y)}{\delta_x \, \delta_y}$$

Aynan python dasturlash tilidagi pandas kutubxonasi Person korrelyatsiyasidan foydalanadi.

Korrelyatsiya koeffitsienti yuqorida aytib o‘tilganidek -1 va 1 oralig‘ida chiqadi va koeffitsienti 1 ga yaqinlashgani sari bog‘liqlik ortib boradi. Agar korrelyatsiya -1 ga yaqinlashsa teskati munosabat(teskari bog‘liqlik) xosil bo‘ladi. Koeffitsienti 0 ga yaqinlashgani sayin o‘zgaruvchilar o‘rtasidagi munosabatlar yo‘qolib boradi. Korrelyatsiya faqat chiziqli munosabatlarni o‘lchaydi.

Quyidagi grafikda chiziqli bo‘lmagan bazi munosabatlar borligini ko‘rishingiz mumkin.



3-rasm. Chiziqli bo‘lmagan bazi munosabatlar uchun korrelyatsiya koeffitsienti

Etibor bersangiz markazda ∅ ya’ni korrelyatsiya aniqlanmagan, chunki standart og‘ish $\delta_y = 0$ va kovariatsiyani 0 ga bo‘lganimiz uchun natija mantiqsiz chiqadi. Olingan natija yani boznng natija nomli o‘zgaruvchimiz 0.9985827486608085 (99.8 %) ko‘rsatkichni berdi. Bu natija ikkala ustun o‘rtasida ijobiy karrelyatsiya aniqlanganidan dalolat beradi ya’ni Jizzax viloyatida kiyim ishlab chiqarish 99.8% toqimachilik maxsulotlarini ishlab chiqarish sanoatiga bog‘liqligini ko‘rsatadi.

Xulosa.

Xulosa qilib shuni aytish mumkinki korrelyatsiya faqat chiziqli munosabatlarni o‘lchashi mumkinligi va chiziqli bo‘lmagan munosabatlar ko‘pligini maqolani o‘qish davomida ko‘rishingiz mumkin. Bizning kiruvchi ma’lumotlarimiz 99.8 % bir biriga bog‘liqligini ko‘rishingiz mumkin bu esa chiziqli munosabatga mos kelganligini ko‘rsatadi. Bundan keyingi tadqiqotlarimizda ko‘proq chiziqli bo‘lmagan munosabatlarni taxlil qilish, mashinali o‘qitishning turli xil algoritmlari, bashoratlashning xar xil usullarini o‘rganishni o‘z oldimizga maqsad qilib qo‘ydik.

Foydalanilgan adabiyotlar.

1. Озод Жўраевич Бабомурадов; Шохрух Маткарим ўғли Матчонов. BI-ТИЗИМЛАР ВА УЛАРНИ АМАЛИЙ МАСАЛАЛАРНИ ЕЧИШДА ҚЎЛЛАШ. RAE 2022, 1, 212-234.

2. Бабамурадов Озод Жўраевич; Асатов Тимур Нурмухаматович; Матчонов Шохрух Маткарим ўғли. МАЪЛУМОТЛАРНИ ТАҲЛИЛЛАШ МАСАЛАСИДА ДАСТЛАБКИ ИШЛОВ БЕРИШНИНГ НАЗАРИЙ АСОСЛАРИ. RAQAMLI TEXNOLOGIYALARNI O‘QUV JARAYONIGA JORIY ETISHNING ZAMONAVIY HOLATI VA ISTIQBOLLARI. Jizzax 2023.

3. Shahrukh Matkarim Matchonov Ogli; Ildarova Sarvinoz Arslan’s Qizi; Kurbanova Lola Ulughbek Qizi. BI-SYSTEMS AND THEIR USE IN SOLVING PRACTICAL PROBLEMS. Euro. Sch. J. 2024, 5, 1-4.

4. Patrick Schober, Christa Boer. Correlation Coefficients: Appropriate Use and Interpretation February 2018Anesthesia & Analgesia 126(5):1