MA’LUMOTLARNI QAYTA ISHLASHDA PANDAS KUTUBXONASINING AFZALLIKLARI

Hamroyeva Ozoda

BuxDU Fizika-Matematika va Axborot texnologiyalari fakulteti Kompyuter ilmlari va dasturlash texnologiyasi(turlar bo’yicha) yo’nalishi 3-bosqich talabasi

E-mail: ozodahamroyevaa@gmail.com

Annotatsiya. Ushbu maqolada qanday holatlarda ma’lumotlarni qayta ishlash uchun Pandas kutubxonasidan foydalanish samarali bo’lishi va ma’lumotlar ustida qayta ishlashning amaliy misollari hamda ularning tahlili yoritilgan.

Kalit so’zlar: ma’lumotlarni qayta ishlash, xom ma’lumot, yangi neft, pandas, model, prognoz, transformatsiya, metod, dublikat, vektorlashtirish

Аннотация. В статье рассматриваются ситуации, в которых эффективно использовать библиотеку Pandas для обработки данных, практические примеры обработки данных и их анализ.

Ключевые слова: обработка данных, необработанные данные, новая нефть, pandas, модель, прогноз, преобразование, метод, дублирование, векторизация

Abstract. This article discusses the situations in which it is effective to use the Pandas library for data processing, practical examples of data processing, and their analysis.

Keywords: data processing, raw data, new oil, pandas, model, forecast, transformation, method, duplicate, vectorization

Kirish

Bugungi kunda biz guvohi bo’layotgan Raqamli Texnologiyalarning shiddat sayin rivojlanishi bilan bir qatorda ma’lumotlar hajmi ham tobora ortib bormoqda. Tabiiyki, axborot ko’paygan sayin ularning ichidan sifatli va kerakli ma’lumotlarni ajratib olish masalasi paydo bo’ladi. Bunday sharoitda ma’lumotlarni qayta ishlash (Data Proccessing) bizga samarali yechim bo’lib xizmat qiladi. Sababi qayta ishlanmagan xom ma’lumotlar (Raw Data) deyarli hech qanday qiymatga ega emas. Bu go’yoki “yangi neft” singaridir. Neftdan foydali mahsulotlarni olish uchun, albatta, uning ustida bir qancha qayta ishlash jarayonlari amalga oshiriladi. Xom axborot ham xuddi shunindek foydali mahsulotga aylanishi uchun, uni qayta ishlash zarur. Shunda u biznes, ilmiy tadqiqotlar va shu singari ko’plab sohalarda qaror qabul qilish, innovatsiyalarni rag’batlantirish va yangi imkoniyatlarni yaratish kabi bir qator foydalarni taqdim etadi. Katta hajmdagi ma'lumotlarni samarali tarzda boshqarish uchun bir qator vositalar mavjud, shulardan biri Python dasturlash tilida Pandas kutubxonasi eng ommabop va samarali vositalardan biri hisoblanadi. Pandas kutubxonasi ma'lumotlar bilan ishlashni soddalashtiradi,

ISSN 3030-3907

yuqori samaradorlikka erishishni ta'minlaydi va murakkab ma'lumotlarni tahlil qilishni osonlashtiradi.

ADABIYOTLAR TAHLILI

Wes McKinneyning “Python for Data Analysis” kitobi Python-da ma'lumotlarni manipulyatsiya qilish, qayta ishlash, tozalash va maydalashning bolt va gaykalari bilan bog'liq. Muallifning maqsadi Python dasturlash tilining qismlari va uning ma'lumotlarga yo'naltirilgan kutubxona ekotizimlari va vositalari bo'yicha qo'llanmani taklif qilishdir, bu dasturchining samarali ma'lumotlar tahlilchisi bo'lishiga yordam beradi. Kitobning sarlavhasida "ma'lumotlar tahlili" bo'lsa-da, asosiy e'tibor ma'lumotlarni tahlil qilish metodologiyasidan farqli o'laroq Python dasturlash, kutubxonalar va vositalarga qaratilgan. Bu o’rganuvchiga ma'lumotlarni tahlil qilish uchun kerak bo'lgan Python dasturlashdir [1]. Stefani Molinning “Hands-On Data Analysis with Pandas” qo’llanmasi Python kutubxonasi pandasdan foydalangan holda ma'lumotlarni tahlil qilish bo'yicha keng qamrovli qo'llanma hisoblanadi. Ushbu kitob nafaqat yangi boshlanuvchilarni asosiy bilimlar bilan ta'minlabgina qolmay, balki tajribali amaliyotchilarga yangi istiqbol va amaliy tushunchalarni taklif etgan. O'quvchilarni pandas kutubxonasining nozik jihatlari bo'ylab mohirona yo'naltirgan, ma'lumotlarni manipulyatsiya qilish, tayyorlash va tadqiq qilishda o'zining jasoratini namoyish etgan. Kitob NumPy, matplotlib, seaborn va sci-kit-learn kabi boshqa muhim Python kutubxonalarini qamrab olish uchun o'z ko'lamini kengaytirib, uni ma'lumotlarni tahlil qilish ishqibozlari uchun keng qamrovli manbaga aylangan [2]. Ильичев В.Ю. hamda Юрик Е.А.ning “Анализ массивов данных с использованием библиотеки pandas для Python” nomli maqolasida ma'lumotlar bazalari bilan ishlash uchun biriktirilgan Pandas kutubxonasi bilan erkin ishlaydigan Python dasturlash tilidan foydalangan holda katta hajmdagi ma'lumotlarni qayta ishlash algoritmini ishlab chiqish tasvirlangan. Unda an'anaviy SQL-ga asoslangan ma'lumotlar bazasini qayta ishlash vositalarida mavjud bo'lmagan Pandas kutubxonasining noyob imkoniyatlarini ko'rsatilgan(faqat relyatsion ma'lumotlar bazalari bilan ishlaydigan tuzilgan so'rovlar tili). Ushbu maqolada qo‘shimcha kutubxonalar ulangan holda ma’lum bir hududdagi quyosh radiatsiyasi va atmosfera oqimlarining xususiyatlari to‘g‘risidagi statistik ma’lumotlar asosida fotoelektr quyosh va shamol stansiyalarining quvvatini hisoblash uchun ushbu algoritmdan yanada foydalanish va rivojlantirish kutilayotganligi keltirilgan [3].

TADQIQOT METODOLOGIYASI

Pandas 2008-yilda Wes McKinney tomonidan ishlab chiqilgan va Python dasturlash tilida ma'lumotlar tahlili uchun qulay va samarali kutubxona sifatida tezda keng tarqaldi. Pandas nomi "panel data" (panel ma'lumotlar) so'zidan kelib

ISSN 3030-3907

chiqqan bo'lib, bu kutubxona asosan ko'p o'lchovli ma'lumotlar bilan ishlashni osonlashtiradi. Pandas ma'lumotlarni tozalash, transformatsiya qilish, yig'ish va tahlil qilish kabi vazifalarni osonlashtiruvchi bir qator qulay metodlarga ega. Shuningdek, Pandas, katta hajmdagi ma'lumotlar bilan ishlashda tezlik va samaradorlikni ta'minlashda juda foydalidir. Ma'lumotlarni qayta ishlashni ma'lumotni tozalash, o'zgartirish va transformatsiya qilish, guruhlash va agregatsiya, birlashtirish va qo'shish kabi bir qancha asosiy yo'nalishlarga ajratish mumkin. Ma'lumotlarni tozalash (Data Cleaning) va ma'lumotlarni o'zgartirish (Data Transformation) - bu data engineering va data analytics jarayonlarida eng muhim bosqichlardan biri hisoblanadi. Ushbu jarayonlar yordamida xom (raw) ma'lumotlarni toza, tahlilga tayyor va strukturaviy holatga keltiriladi. Bu jarayon ma'lumotlardagi sifat muammolarini bartaraf etishga qaratilgan bo'lib, Agar ma'lumotlar to'g'ri tozalanmasa va o'zgartirilmasa quyidagi muammolar yuzaga kelishi mumkin: 1) tahlillar noto'g'ri natija beradi, 2) model va prognozlar noto'g'ri bo'ladi, 3) ma'lumotlar bazasi noto'g'ri qarorlar qabul qilishga sabab bo'ladi. Pandas bu jarayonlarni avtomatlashtirish va soddalashtirishda eng kuchli vositalardan biri hisoblanadi. Real hayotda tahlilning 80% vaqti aynan ma'lumotlarni tozalash va o'zgartirishga sarflanadi. Ma'lumotlarni tozalash ma'lumotlar tahlilining asosiy bosqichlaridan biridir. Pandas kutubxonasi bu jarayonni osonlashtiradi va tezlashtiradi. Null (NaN) qiymatlar bilan ishlash: Ma'lumotlar to'plamlarida yo'qolgan qiymatlar (NaN) ko'pincha uchraydi. Pandas *.isnull()*, *.notnull()*, *.dropna()* va *.fillna()* kabi metodlar bilan ma'lumotlardagi NaN qiymatlarni aniqlash va boshqarish imkonini beradi. Bu metodlar yordamida foydalanuvchi yo'qolgan qiymatlarni o'chirish yoki ularni tegishli qiymatlar bilan to'ldirish jarayonlarini amalga oshirishi mumkin. Dublikatlarni aniqlash va olib tashlash uchun Pandas *.duplicated()* va *.drop_duplicates()* metodlarini taqdim etadi. Bu metodlar yordamida foydalanuvchilar ma'lumotlar to'plamidagi takrorlanuvchi qatorlarni aniqlab, ularni olib tashlashlari mumkin, bu esa tahlilning aniqligini oshiradi. Pandas *.loc()* va *.iloc()* metodlari yordamida foydalanuvchilar ma'lum shartlarga asoslangan holda ma'lumotlarni filtrlay oladi. Masalan, ma'lum bir qiymatga ega bo'lgan qatorlarni yoki ustunlarni osongina tanlab olish mumkin. Pandas kutubxonasi yordamida ma'lumotlarni o'zgartirish va transformatsiya qilish jarayonlari juda samarali amalga oshiriladi. Pandas *.assign()*, *.drop()*, va *.rename()* metodlari yordamida ustunlar va qatorlarni qo'shish, o'chirish yoki nomlarini o'zgartirish mumkin. Bu esa ma'lumotlarni tahlil qilish uchun kerakli formatga keltirishda yordam beradi. Pandas yordamida ma'lumotlarni shaklini o'zgartirish uchun *.pivot()* va *.melt()* metodlari mavjud. Bu metodlar ma'lumotlarni tahlil qilishda kerakli strukturalarga keltirishni osonlashtiradi. Pandas ma'lumotlar ustunlari va qatorlarini vektorlashtirish imkoniyatiga ega, ya'ni bir vaqtning o'zida

ISSN 3030-3907

butun ustunni yoki qatorni qayta ishlash mumkin. Bu metod tahlilni tezlashtiradi va kodni oddiylashtiradi. Ma'lumotlarni guruhlash va agregatsiya qilish Pandas kutubxonasining kuchli tomonlaridan biridir. Pandas *.groupby()* metodini taqdim etadi, bu yordamida foydalanuvchilar ma'lumotlarni guruhlash va har bir guruh uchun statistik tahlil (o'rtacha, maksimal, minimal qiymatlar va h.k.) qilish imkoniyatiga ega. Bu jarayon katta hajmdagi ma'lumotlarni boshqarishda juda foydalidir. Pandas kutubxonasi yordamida turli ma'lumotlar to'plamlarini birlashtirish va qo'shish juda oson. *.merge()* va *.join()* metodlari yordamida ikki yoki undan ortiq ma'lumotlar to'plamini birlashtirish mumkin. Bu metodlar yordamida foydalanuvchilar turli manbalardan olingan ma'lumotlarni osongina bog'lab, yagona to'plamga kiritishlari mumkin.

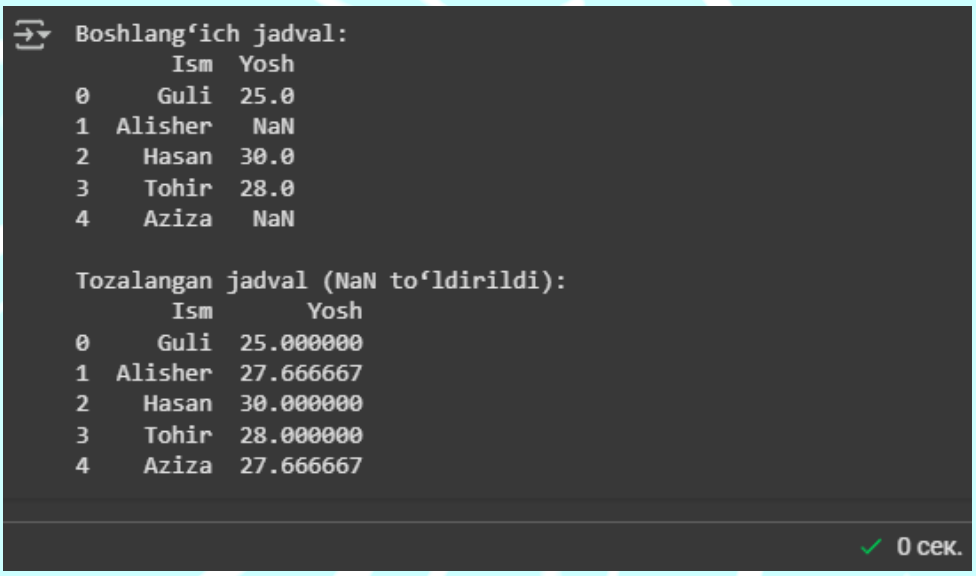
MUHOKAMA VA NATIJALAR

Quyida amaliy misol sifatida NaN qiymatlarni aniqlash va ularni o’rtacha qiymat bilan to’ldirishni ko’ramiz. Bunda biz pandas va numpy kutubxonalaridan foydalanishizmiz kerak bo’ladi. Avval bizga ma’lumot kerak bo’adi, uni *.csv* va *.json* kabi fayllardan ham olishimiz mumkin yoki lug’at(*dictionary)* shaklida o’zimiz yozsakham bo’adi. Biz ham shu yo’ldan foydalandik. So’ngra u ma’lumotni *DataFrame*ga o’tkazamiz. Bu orqali kelayotgan ma’lumotlar Pandas o’zi ishlay oladigan jadvalga aylanadi. Va bu jadvalni ko’rish uchun *console*ga chiqarib ko’ramiz. Keyingi bosqichda ustunda mavjud bo’lgan NaN qiymatlarni o’rnini o’rtacha qiymat bilan to’ldirib jadvalga o’zgartirish kiritadi. Bu uchun biz *.fillna()* metodidan foydalanamiz. O’zgargan jadvalni ko’rish uchun uni yana consolega chiqaramiz.

```
import pandas as pd
import numpy as np
data = {
 'Ism': [Guli, Alisher, Hasan, Tohir, Aziza],
   'Yosh': [25, np.nan, 30, 28, np.nan]
}
df = pd.DataFrame(data)
print("Boshlang‘ich jadval:")
print(df)
df['Yosh'] = df['Yosh'].fillna(df['Yosh'].mean())
print("\nTozalangan jadval (NaN to‘ldirildi):")
print(df)
```

Natija:

ISSN 3030-3907



1-rasm. NaN qiymatlarni aniqlash va ularni o’rtacha qiymat bilan to’ldirish.

Yana bir namunaviy misol sifatida *.loc()* metodini ishlatib ko’ramiz. Dastlab DataFrame yaratamiz. Keyin esa consoleda butun jadvalni ko'rsatamiz. So’ngra biror qatordagi barcha ma'lumotlarni chiqarib ko’amiz (masalan 2-qator - Hasan). Keyingi bosqichda belgilangan qatorlar va ustunlarni chiqaramiz(1 va 2-qatordagi 'Ism' va 'Yosh'). Shu tariqa faqat bir ustun: 'Shahar' ustunini chiqaramiz. Va so’ngra shart asosida tanlaymiz: faqat yoshlari 25 dan katta bo'lganlar.

```
import pandas as pd
data = {
   'Ism': ['Ali', 'Vali', 'Hasan', 'Husan'],
   'Yosh': [25, 30, 28, 22],
   'Shahar': ['Toshkent', 'Samarqand', 'Buxoro', 'Andijon']
}
df = pd.DataFrame(data)
print(" Butun jadval:")
print(df)
print("\n .loc[2] — 2-chi indeksdagi qator:")
print(df.loc[2])
print("\n .loc[1:2, ['Ism', 'Yosh']] — 1–2-qator, faqat 'Ism' va 'Yosh':")
print(df.loc[1:2, ['Ism', 'Yosh']])
print("\n .loc[:, 'Shahar'] — faqat 'Shahar' ustuni:")
print(df.loc[:, 'Shahar'])
print("\n .loc[df['Yosh'] > 25] — 'Yosh' > 25 bo'lganlar:")
print(df.loc[df['Yosh'] > 25])
```

Natija:

ISSN 3030-3907

```
Butun jadval:
     Ism  Yosh     Shahar
0    Ali    25   Toshkent
1   Vali    30  Samarqand
2  Hasan    28     Buxoro
3  Husan    22    Andijon

 .loc[2] – 2-chi indeksdagi qator:
Ism        Hasan
Yosh          28
Shahar    Buxoro
Name: 2, dtype: object

 .loc[1:2, ['Ism', 'Yosh']] – 1-2-qator, faqat 'Ism' va 'Yosh':
     Ism  Yosh
1   Vali    30
2  Hasan    28

 .loc[:, 'Shahar'] – faqat 'Shahar' ustuni:
0     Toshkent
1    Samarqand
2       Buxoro
3      Andijon
Name: Shahar, dtype: object

 .loc[df['Yosh'] > 25] – 'Yosh' > 25 bo'lganlar:
     Ism  Yosh     Shahar
1   Vali    30  Samarqand
2  Hasan    28     Buxoro
```

2-rasm. Shartga asoslangan holda ma’lumotlarni filtrlash.

XULOSA VA TAKLIFLAR

Pandas kutubxonasi ma’lumotlarni tahlil qilish va qayta ishlash jarayonlarini soddalashtiruvchi kuchli vositadir. U orqali jadvallar bilan ishlash, ustunlar kesish, filtrlar qo‘llash, statistik tahlil va grafiklar qurish oson va tez amalga oshiriladi. Ayniqsa, katta hajmdagi ma’lumotlarni avtomatlashtirilgan tarzda tozalash va tuzish imkoniyati uni amaliy loyihalarda ajralmas vositaga aylantiradi. Shu o’rinda quyidagi takliflarni keltirishimiz mumkin: birinchidan ma’lumotlar bilan ishlaydigan dasturchilar va talabalar Pandas kutubxonasini chuqur o‘rganishlari tavsiya etiladi. Ikkinchidan oliy ta’limda ma’lumotlarni qayta ishlash bo‘yicha amaliy mashg‘ulotlarga Pandas asosida yondashish samaradorlikni oshiradi. Uchinchidan Pandas kutubxonasini boshqa vositalar (NumPy, Matplotlib) bilan birgalikda qo‘llash orqali kompleks tahliliy tizimlar yaratish mumkin. To’rtinchidan soha mutaxassislari real loyiha misollarida Pandasdan foydalanishni ko‘proq amaliyotga joriy etishlari lozim.

FOYDALANILGAN ADABIYOTLAR RO‘YXATI

[1]. Wes McKinney (2013) “Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython”

[2]. Stefanie Molin (2019) “Hands-On Data Analysis with Pandas: Efficiently perform data collection, wrangling, analysis, and visualization using Python”

ISSN 3030-3907

[3]. Ильичев В.Ю., Юрик Е.А. (2020) “Анализ массивов данных с использованием библиотеки pandas для Python”

ISSN 3030-3907