Amaliy ish № 3

Ma'lumotlarni tozalash

Ishning maqsadi:

Pandas kutubxonasidan foydalanib, ma'lumotlarni xatolar, yetishmayotgan qiymatlar va dublikatlardan tozalashni o'rganish..

Nazariy qism:

1. Nima uchun ma'lumotlarni tozalash kerak?

   Ma'lumotlarda ko'pincha yetishmayotgan qiymatlar, dublikatlar yoki noto'g'ri formatlar kabi xatolar mavjud. Ushbu muammolar tahlil natijalarini buzishi mumkin.

   Ma'lumotlarni tozalash - bu ma'lumotlarni tahlil qilish uchun tayyorlash jarayoni:

   - Yo'qolgan qiymatlarni olib tashlash yoki to'ldirish.
   - Dublikatlarni yo'q qilish.
   - Ma'lumotlarni mos formatga aylantirish.
2. Yo'qolgan qiymatlar qanday hal qilinadi?

   Yo'qotilgan qiymatlar (NaN) quyidagilardek qilish mumkin:

   - Qiymatlari yetishmayotgan satrlar yoki ustunlarni olib tashlash.
   - Ularni o'rtacha, median, rejim yoki boshqa qiymat bilan to'ldirish.
3. Dublikatlar nima?

   Dublikatlar - ma'lumotlarning takroriy satrlari. Ortiqcha bo'lmaslik uchun ularni olib tashlash muhimdir.
4. Ma'lumotlarni tozalash uchun Pandas usullari:
   - `.drop()` — qiymatlari yetishmagan satr/ustunlarni o‘chiradi.
   - `.fill()` — yetishmayotgan qiymatlarni to‘ldiradi.
   - `.drop_duplicates()` — dublikatlarni olib tashlaydi.

Amaliy qism:

1. **Kerakli vositalarni o'rnatish:**
   - Python (3.8+ versiyasi) va Pandas kutubxonasi o‘rnatilganligiga ishonch hosil qiling. Agar Pandas o'rnatilmagan bo'lsa, buyruqni bajaring:

```
pip install pandas
```

2. Ma'lumotlarni tayyorlash:
    - oldingi amaliy ishingizda yuklab olgan Titanic ma'lumotlar to'plamidan foydalaning (`train.csv`).
3. Ma'lumotlar bilan ishlash:
    - Yangi Python faylini yarating (masalan, `amaliy_2.py`) va quyidagi amallarni bajaring:

```
import pandas as pd

# Ma'lumotlarni yuklash
data = pd.read_csv('train.csv')

#1. Qiymatlari yetishmayotgan qatorlarni olib tashlash
cleaned_data = data.dropna()
print ("Yo'qolgan qiymatlari bo'lgan qatorlarni olib
tashlangandan keyin ma'lumotlar:")
print(cleaned_data.head())

# 2. Yetishmayotgan qiymatlarni median bilan to'ldirish
data['Age'].fillna(data['Age'].median(), inplace=True)
print ("'Age' ustunidagi yetishmayotgan \n qiymatlarni
to'ldirgandan keyingi ma'lumotlar:")
print(data.head())

№3. Dublikatlarni olib tashlang
data.drop_duplicates(inplace=True)
print("\nDublikatlar olib tashlanganidan keyin ma'lumotlar:")
print(data.head())
```

4. **Dasturni ishga tushirish:**

    Faylni saqlang va uni terminal orqali ishga tushiring:

    ```
    python amaliy_2.py
    ```

    Natijada ko’rinadi:

- Yo'qolgan qiymatlari bo'lgan qatorlarni olib tashlashdan keyin ma'lumotlar.
- Ustundagi etishmayotgan qiymatlarni to'ldirgandan keyin ma'lumotlar `Age`.
- Dublikatlarni olib tashlaganingizdan keyin ma'lumotlar.

Qo'shimcha materiallar:

- Ma'lumotlarni tozalash bo'yicha qo'llanma: Pandas yordamida ma'lumotlarni tozalash .
- Rasmiy Pandas hujjatlari: Pandas hujjatlari .

Amaliy vazifa:

Vazifa: № 2

Titanic ma'lumotlar to'plamidan foydalanib (`train.csv`), quyidagi amallarni bajaring:

1. Yo'qolgan qiymatlarni olib tashlashdan oldin va keyin qatorlar sonini hisoblang.
2. Ustundagi etishmayotgan qiymatlarni to'ldiring `Embarked` eng tez-tez uchraydigan qiymat (rejim).
3. Ma'lumotlarda dublikatlar mavjudligini tekshiring va agar shunday bo'lsa, ularni olib tashlang.
4. Tozalangan ma'lumotlarning dastlabki 10 qatorini chop eting.