ДЕРЕКТЕРДІ ИНТЕЛЛЕКТУАЛДЫ ТАЛДАУ

Дәріс 1

Деректерді интеллектуалды талдауға кіріспе. Series және DataFrame

Оспан Ә.Ғ., PhD, доцент-зерттеуші

ҚазҰУ, Жасанды интеллект және Big Data кафедрасы

2026-2027 оқу жылы

Дәрістің мақсаты мен мазмұны

Мақсат және оқу нәтижелері

- Бақылауды, белгіні және орынды аналитикалық сұрақты анықтау.
- Series пен DataFrame құрып, индекс белгісі мен орынды ажырату.
- CSV оқып, жиынның схемасын, көлемін және шектеулерін тексеру.

Дәріс мазмұны

1. **Деректерді интеллектуалды талдау міндеті**
2. **Белгілер**
3. **Series**
4. **DataFrame**
5. **Оқу және тексеру**

Сұрақты тексерілетін есептеуге айналдыру

1

Сұрақ

Көрсеткішті, салыстырылатын топтарды және қорытындының шегін атау.

2

Бақылау

Бір жол мен әрбір белгінің нені білдіретінін анықтау.

3

Есептеу

Деректерді тексеріп, мағынасы тұрғысынан орынды амалды таңдау.

4

Қорытынды

Санды, шектеулерді және нәтижені қайталау тәсілін түсіндіру.

- Кесте пішіні жолдың мағынасын анықтамайды: оны дереккөз белгілейді.
- Қайталанатын нәтиже деректерді, кодты және түсіндірмені қамтиды.

Сақтау типі және белгінің мағынасы

Өріс	Мағынасы	Орынды есептеу
Age	Жас	Орташа мән, медиана, ауқым
Position	Позиция санаты	Жиіліктер және топтарды салыстыру
SprintSpeed	Сипаттама рейтингі	Рейтингтерді салыстыру
Name	Дереккөздегі аты	Жазбаларды көрсету; бірегейлігіне кепілдік жоқ
Unnamed: 0	Бұрынғы индекс	Техникалық салыстырып тексеру; аналитикалық белгі емес

Алдымен мағына мен өлшем бірліктерін анықтап, содан кейін тип пен амалды таңдаймыз.

Series: мәндер индекс белгілерімен байланысады

Нысанның бөліктері

- Мәндер: бір айнымалының деректері.
- Индекс: элементтермен байланысқан белгілер.
- name және dtype: атау мен сақтау тәсілі.

Қол жеткізудің екі тәсілі

- .loc индекс белгісін таңдайды.
- .iloc нөлден саналатын орынды таңдайды.
- Белгілердің бірегейлігі бөлек тексеріледі.

Series іріктеу мен арифметика кезінде мәннің өз белгісімен байланысын сақтайды.

Series: құрудан тексерілетін санға дейін

Python мысалы

```
import pandas as pd

scores = pd.Series(
    [72, 80, 68], index=['p1', 'p2', 'p3'], name='rating'
)
print(scores.loc['p2'])
print(scores.iloc[0])
print(round(scores.mean(), 2))
```

Нәтиже: 80; 72; 73.33

Деректерді алу

- p2 — индекс белгісі.
- 0 — бірінші орын.

Есептеу

- Орташа мән: мәндер қосындысы / 3.
- Нәтиженің көрсетілуін дөңгелектейміз.

Series қосындысында белгілер сәйкестендіріледі

Белгі	Бірінші Series	Екінші Series	Қосынды
a	10	2: екінші орында	12
b	20	1: бірінші орында	21
c	Мән жоқ	5 қоссақ	Толтыру ережесінсіз жетіспейтін мән

Сәйкестендіру орындарды емес, бірдей белгілерді байланыстырады; арифметика алдында дереккөздер индекстерінің айырмасын тексереміз.

DataFrame: ортақ жол осі, әртүрлі белгілер

Кестенің ұйымдасуы

- Жол бақылауды білдіреді.
- Баған белгіні білдіреді.
- Ұяшықта белгінің мәні сақталады.

Құру және бақылау

- Тізімдер сөздігі бағандарды анықтайды.
- Индекс пен columns осьтерді атайды.
- shape пен dtypes құрылымды тексереді.

Есептеудің алдында екі осьтің мағынасын және нәтиженің күтілетін пішінін атаңыз.

Бір бағаң: Series пе, DataFrame бе?

Python мысалы

```
players = pd.DataFrame({
    'Name': ['Ada', 'Boris', 'Chen'],
    'Age': [20, 24, 22],
    'SprintSpeed': [70, 80, 75]
}, index=['p1', 'p2', 'p3'])
print(players.shape)
print(players['Age'].shape)
print(players[['Age']].shape)
```

Нәтиже: (3, 3); (3,); (3, 1)

Бір кілт

- players['Age'] Series қайтарады.
- Нәтижеде бір ось бар.

Кілттер тізімі

- players[['Age']] DataFrame қайтарады.
- Нәтижеде екі ось сақталады.

Кестеге алғашқы шолу: әртүрлі сұрақтар

Құрал	Қандай сұраққа жауап береді	Шектеуі
head / tail	Шеткі жазбалар қалай көрінеді?	Бұл тек үзінді
shape / columns	Көлемі мен өрістер тізімі қандай?	Деректердің шынайылығын тексермейді
info	Типтері мен толтырылуы қандай?	Шкалалардың мағынасын анықтамайды
describe	Негізгі статистикалар қандай?	Өрістерді мағынасына сай таңдауды қажет етеді
isna	Қызметтік жетіспейтін мәндер қайда?	Дереккөздің барлық қатесін танымайды

CSV оқып, талаптарға сәйкестігін тексереміз

Python мысалы

```
from pathlib import Path

football = pd.read_csv(Path('data') / 'football.csv')
print(football.shape)
print(football[['Name', 'Age', 'Position']].head(3))
football.info()
print(football.isna().sum().sum())
```

Нәтиже: Пішіні: (12897, 42). Жетіспейтін ұяшықтар: 0.

Орналасуы

- Жинақтың түбірінен іске қосамыз.
- Path тасымалданатын салыстырмалы жолды береді.

Файлға қойылатын талаптар

- Көлемі мен міндетті өрістерді тексереміз.
- Бүкіл жиындағы жетіспейтін ұяшықтарды санаймыз.

football.csv паспорты: жорамал қосуға болмайды

Уақыт және қамту

- Сипаттамалардың тарихи кесіндісі.
- Өзектілігі мен толық қамтылуына кепілдік жоқ.

Шкалалар

- Wage/Value: бірліктері мен кезеңі расталмаған.
- SprintSpeed пен Penalties — рейтингтер.

Сәйкестендіру

- Name қайталануы мүмкін.
- Unnamed: 0 аналитикалық белгі емес.

Паспорт қорытындыны деректердің шығу тегі мен шкалалары туралы нақты белгілі мәліметпен шектейді.

Нәтижені қайталауға қалай мүмкіндік береміз?

1

Кірісті бекіту

Жолды, файл нұсқасын, өрістер мағынасын және шектеулерін белгілеу.

2

Қадамдарды жазу

Импорттарды, жүктеуді, түрлендірулерді және нәтиже формуласын жазу.

3

Тексеру

Пішінді, бақылау мысалының мәндерін және көрсеткіш шегін тексеру.

4

Бөлек сақтау

Туынды файлдар work_results ішінде; бастапқы файл сақталады.

- Кәдімгі кестені экспорттау: index=False.
- Таза қайта іске қосу жасырын тәуелділіктің жоқтығын растайды.

Деректерді сипаттау және себептерді түсіндіру

Сипаттау

- Таңдалған файлда не байқалатынын хабарлайды.
- Топтарды, көрсеткішті және жазба санын атайды.
- Шкала мен қамту шектеулерін сақтайды.

Себептік қорытынды

- Зерттеудің бөлек дизайнын қажет етеді.
- Орташа мәндердің жалғыз айырмасынан туындамайды.
- Оқиғаның қисынды көрінуімен расталмайды.

Бақыланатын топтар арасындағы айырма өздігінен сол айырманың себебін анықтамайды.

Шындыққа ұқсас жауап беретін қателер

Құрылым

- Индекс белгісін орынмен шатастыру.
- DataFrame орнына Series алынғанын байқамау.

Мағына

- Идентификатордың орташа мәнін есептеу.
- Рейтингке физикалық бірлік телу.

Дәлел

- Файлды тек head бойынша бағалау.
- Топтардың сипаттамасын себеп ретінде ұсыну.

Кодтың орындалуын алынған сан мен оның атауының мағынасынан бөлек тексереміз.

Дәрістен кейін нені орындай алу қажет?

1. Талдау сұрақтан және бір жолдың анықтамасынан басталады.
2. Series мәндерді белгілермен байланыстырады; арифметика индекстерді сәйкестендіреді.
3. DataFrame әртүрлі типтегі белгілерді бақылаулардың ортақ осінде сақтайды.
4. Бір кілт пен кілттер тізімін таңдау әртүрлі өлшемділік береді.
5. CSV бүкіл файл бойынша тексеріледі; шкалалар мағынасы паспортқа жазылады.

Деректер құрылымын түсінуді тексеріңіз

1. Неліктен жол саны адам санына әрдайым тең емес?
2. Series индекс белгісі элементтің орнынан несімен ерекшеленеді?
3. df["Age"] және df[["Age"]] кезінде пішін қалай өзгереді?
4. Неліктен Unnamed: 0 қосындысы пайдалы көрсеткіш емес?
5. Бұл файл Penalties туралы қандай тұжырым жасауға мүмкіндік береді?

Өзіндік жұмыс

Белгілері қайта реттелген және ішінара әртүрлі екі Series құрыңыз. Қосындыны болжаңыз, DataFrame жинап, оның пішіні мен жетіспейтін мәндерін түсіндіріңіз.

Дереккөздер мен құжаттама

Құжаттама

1. pandas: Series және DataFrame

https://pandas.pydata.org/docs/user_guide/dsintro.html

2. pandas: алғашқы қадамдар

https://pandas.pydata.org/docs/user_guide/10min.html

3. pandas: оқу және жазу

https://pandas.pydata.org/docs/user_guide/io.html

Дәрістің негізі

- Берілген 1-дәрістің презентациясы
- pandas бойынша берілген оқу құралы
- Берілген 1-зертханалық жұмыстың тапсырмалары