ДЕРЕКТЕРДІ ИНТЕЛЛЕКТУАЛДЫ ТАЛДАУ

Дәріс 4

Агрегаттау және жиынтық кестелер

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

- Қосынды, орташа мән, медиана және санауышты міндет мағынасына сай таңдау.
- Атаулы агрегаттар мен екі кілт бойынша жиынтық кестелер құру.
- Бос үйлесімдерді, салмақтарды және қорытынды мәндерді тексеру.

Дәріс мазмұны

1. **Агрегат таңдау**
2. **agg**
3. **Жол деңгейі**
4. **pivot_table**
5. **Қорытындылар**

Әртүрлі сұраққа әртүрлі агрегат қажет

Сұрақ	Агрегат	Нені ескеру керек
Жалпы көлем қандай?	sum	Жазба санына тәуелді
Орташа деңгей қандай?	mean	Үлкен мәндерге сезімтал
Мәндердің ортасы қайда?	median	Бүкіл шашырауды көрсетпейді
Шектері қандай?	min, max	Бір жазбамен анықталуы мүмкін
Бақылау қанша?	size / count	Жолдар не бос емес мәндер

Орташа мән: деңгей және сезімталдық

$$\text{mean} = \text{sum}(x_i) / n$$

Мазмұны

- Барлық мәнді және олардың шамасын қолданады.
- Бастапқы көрсеткіштің өлшем бірлігін сақтайды.

Шектеуі

- Үлкен мән mean нәтижесін едәуір ығыстыра алады.
- Үлестірімсіз орташа мән әдеттегі жазбаны толық сипаттамайды.

Wage: 10, 10, 100. Қосынды = 120; орташа мән = 40.

Медиана: орталық орын мәндердің теңдігін білдірмейді

Қалай есептеледі?

- Алдымен мәндер реттеледі.
- Тақ n: орталық мән.
- Жұп n: екі орталық мәннің орташа мәні.

Қалай түсіндіріледі?

- Жеке шеткі мәндерге азырақ сезімтал.
- mean = median барлық бақылау тең дегенді білдірмейді.
- Салыстыруды n, min және max арқылы толықтырамыз.

Орталық орын, шашырау және топ көлемі деректердің әртүрлі қырын сипаттайды.

Бір есептегі атаулы агрегаттар

Python мысалы

```
import pandas as pd

demo = pd.DataFrame({
    'Club': ['A', 'A', 'A', 'B', 'B'],
    'Position': ['GK', 'ST', 'ST', 'GK', 'CM'],
    'Wage': [10, 10, 100, 20, 40],
    'SprintSpeed': [40, 70, 90, 50, 80]
})
summary = demo.groupby('Club', as_index=False).agg(
    n=('Wage', 'size'), wage_mean=('Wage', 'mean'),
    wage_median=('Wage', 'median'), wage_sum=('Wage', 'sum')
)
print(summary)
```

Есептің талаптары

- Бір жол бір клубқа сәйкес келеді.
- Өріс атаулары айқын беріледі.

Тексеру

- A: n=3; mean=40; median=10; sum=120.
- B: n=2; mean=30; median=30; sum=60.

Орташа мән мен медиана теңдігі нені дәлелдемейді?

Мәндер	mean	median	Қорытынды
30	30	30	Бір жазба: сәйкестік міндетті
20, 40	30	30	Статистикалар тең болса да, мәндер әртүрлі
10, 10, 100	40	10	Үлкен мән орташа мәнді ығыстырды

Екі статистиканың сәйкестігі үлестірімді анықтамайды; әр таңдалған топтың көлемі мен ауқымы бөлек тексеріледі.

Есептеулерде жол деңгейі сақталуы тиіс

1

Бастапқы жол

Ойыншы сипаттамалары туралы жазба.

2

Бір кілт

Club: қамтылған әр клубқа бір жол.

3

Екі кілт

Club және Position: байқалған әр жұпқа бір жол.

4

Жиынтық пішін

Сол жұптар кестенің екі осіне орналастырылған.

- as_index=False экспортқа ыңғайлы жазық өрістер береді.
- Әртүрлі егжей-тегжей деңгейінің қайталанған қорытындыларын қосуға болмайды.

pivot_table: екі кілт және ұяшықтағы көрсеткіш

pivot_table

- index пен columns екі осьтің кілтін береді.
- values көрсеткішті таңдайды; aggfunc есептеуді анықтайды.
- Қайталанатын жұптар агрегатталады.

pivot

- Пішінді агрегаттаусыз өзгертеді.
- Кілттер жұбы бір мәнді анықтауы тиіс.
- Қайталанатын жұптар дерек мағынасына сай шешімді қажет етеді.

pivot пен pivot_table таңдауы жұптардың бірегейлігіне және агрегаттау сұрағына тәуелді.

Клуб пен позиция бойынша жазба санының жиынтығы

Python мысалы

```
counts = demo.pivot_table(
    index='Club', columns='Position', values='Wage',
    aggfunc='size', fill_value=0, observed=True
)
print(counts)
assert counts.to_numpy().sum() == len(demo)
```

Нәтиже: A: CM=0, GK=1, ST=2. B: CM=1, GK=1, ST=0.

Екі ось

- Жолдар: Club; бағандар: Position.
- Ұяшық: бастапқы жазбалар саны.

Бақылау

- Нөл жұп жазбаларының жоқтығын білдіреді.
- Санауыштар қосындысы бастапқы жол санына тең.

Бос ұяшық: нөл ме, бағалаудың жоқтығы ма?

Көрсеткіш	Үйлесім жолдары жоқ	Көрсету тәсілі
Жазба саны	Белгілі нөлдік санауыш	Нөл орынды
Орташа SprintSpeed	Бағалау анықталмаған	Жетіспейтін мәнді сақтау
Wage қосындысы	Деректер толықтығын білу қажет	Нөл тек жарияланған ереже бойынша
Белгісіз өлшем	Мән бақыланбаған	Өлшенген нөлмен алмастырмау

margins: орташа мәннің жалпы қорытындысы

Python мысалы

```
mean_total = demo.pivot_table(
    index='Club', columns='Position', values='SprintSpeed',
    aggfunc='mean', margins=True, margins_name='Total',
    observed=True
)
print(mean_total.loc['Total', 'Total'])
print(demo['SprintSpeed'].mean())
```

Нәтиже: 66.0; 66.0

Орташа рейтинг

- Қорытынды бастапқы бақылаулар бойынша есептеледі.
- Топ көлемдері автоматты ескеріледі.

Қайта санау

- Total ішінде агрегатталған көлем бар.
- Қорытынды жолды қайта қоспаймыз.

Ойыншылар бойынша және клубтар бойынша орташа мән

mean_players = sum(n_g * mean_g) / sum(n_g)

Бақылаулар бойынша

- Клуб салмағы белгілі мәндер санына тең.
- Позицияның бастапқы жолдарының тікелей mean нәтижесіне тең.

Клубтар бойынша

- Барлық клубтық орташа мәннің салмағы бірдей.
- Басқа аналитикалық сұраққа жауап береді.

Салмақтары 3 және 2: (3*66.6667 + 2*65)/5 = 66.

Тарихи football.csv клубтары бойынша есеп

Python мысалы

```
from pathlib import Path
import numpy as np

football = pd.read_csv(Path('data') / 'football.csv')
club_report = football.groupby('Club', as_index=False).agg(
    n=('Wage', 'size'), wage_mean=('Wage', 'mean'),
    wage_median=('Wage', 'median')
)
club_report['same'] = np.isclose(
    club_report['wage_mean'], club_report['wage_median'],
    rtol=1e-10, atol=1e-8
)
print(club_report.head(3))
```

Бірнеше көрсеткіш

- Бір Club кілті, үш түсінікті агрегат.
- Түсіндіру үшін жазба саны қажет.

Теңдік

- Рұқсат шегі айқын берілген.
- mean/median сәйкестігі Wage бірдей дегенді білдірмейді.

Жиынтық кестені тексеру: құрылым, қосынды, мағына

Құрылым

- Жарияланған әр топқа бір жол.
- Кілттер мен өріс атаулары бірмәнді.

Сандар

- Санауыштар бастапқы файлмен сәйкес.
- Ұяшық тікелей сүзгілеуге сәйкес.

Түсіндіру

- Нөлдер мен жетіспейтін мәндер мағынасы бөлек.
- Салмақтар мен қорытынды жолдар қайталап санауға әкелмейді.

Әдемі көрсету құрылымды, бастапқы жолдарды және көрсеткіш мағынасын тексеруді алмастырмайды.

Агрегаттаудың бес ережесі

1. Агрегат дерек типі бойынша ғана емес, сұрақ бойынша таңдалады.
2. Атаулы agg есеп схемасын айқын етеді.
3. pivot_table қайталанатын кілт үйлесімдерін агрегаттайды.
4. Бос санауышты нөлмен көрсетуге болады; бос орташа мән анықталмаған.
5. Жалпы орташа мәнге дұрыс салмақ қажет, ал қорытындылар қайта саналмайды.

Жиынтық кесте мен қорытындыларын тексеріңіз

1. Wage қосындысы мен орташа мәні неліктен әртүрлі реттеу беруі мүмкін?
2. mean және median теңдігі нені дәлелдемейді?
3. pivot орнына pivot_table қашан қажет?
4. Бос санауышқа нөл неге орынды, ал орташа мәнге неге қауіпті?
5. Топтық орташа мәндерден жалпы орташа мәнді қалай қалпына келтіруге болады?

Өзіндік жұмыс

demo ішінде Wage қосындыларын және орташа SprintSpeed мәндерін pivot_table және unstack бар groupby арқылы алыңыз. Ұяшықтарды салыстырып, бос үйлесімдер ережелерін түсіндіріңіз.

Агрегаттар мен жиынтық кестелер дереккөздері

Құжаттама

1. pandas: топтар бойынша агрегаттау

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

2. pandas: пішінді өзгерту және жиынтық кестелер

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

3. pandas: pivot_table

https://pandas.pydata.org/docs/reference/api/pandas.pivot_table.html

Дәрістің негізі

- Берілген 4-дәрістің топтауды жалғастыру жөніндегі презентациясы
- Берілген 4-зертханалық жұмыстың тапсырмалары
- Берілген 4-зертханалық жұмыстың жауаптары бар нұсқасы