ДЕРЕКТЕРДІ ИНТЕЛЛЕКТУАЛДЫ ТАЛДАУ

Дәріс 3

Жиіліктер және деректерді топтау

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

- Жиілікті, үлесті және пайызды дұрыс бөліммен есептеу.
- unique, nunique, value_counts және groupby құралдарын ажырату.
- Жазба сандарын, аралық шекараларын және топтық орташа мән мағынасын тексеру.

Дәріс мазмұны

1. **Жиіліктер**
2. **Бірегей мәндер**
3. **Үлестер**
4. **groupby**
5. **Аралықтар**

Жеке жазба мен топ: егжей-тегжей деңгейі әртүрлі

Бақылау

- Бастапқы файлдың бір жолы — ойыншы туралы жазба.
- Белгілер нақты жазбаны сипаттайды.
- Сүзгі бастапқы жолдардың бір бөлігін таңдайды.

Топ

- Кілт ортақ мәні бар жолдарды біріктіреді.
- Агрегат таңдалған жолдарды қысқаша сипаттайды.
- Ықшамдауда жеке айырмалардың бір бөлігі жоғалады.

Топтаудан кейін жолдың мағынасы өзгереді, сондықтан оны қайта көрсету қажет.

Жиілік, үлес және пайыз

$$share_c = n_c / N;\ percent_c = 100 * share_c$$

Абсолюттік қамтылу

- n_c: таңдалған санаттың жол саны.
- Жиіліктер қосындысы есепке алынған жол санына тең.

Салыстырмалы қамтылу

- N сұрақ пен есепке алу ережелері арқылы анықталады.
- normalize=True пайызды емес, үлесті береді.

6 жолдың 3-еуі: үлесі 0.5, пайызы 50%.

value_counts: әр клубтың жазбасы қанша?

Python мысалы

```
import pandas as pd

demo = pd.DataFrame({
    'Club': ['A', 'A', 'B', 'B', 'B', 'C'],
    'Position': ['GK', 'ST', 'GK', 'ST', 'ST', 'CM'],
    'Age': [20, 24, 22, 26, 30, 21],
    'Wage': [10, 30, 20, 40, 60, 15]
})
counts = demo['Club'].value_counts()
print(counts.to_dict())
assert counts.sum() == len(demo)
```

Нәтиже: {'B': 3, 'A': 2, 'C': 1}

Нені санаймыз?

- Бағандағы санаттың әр кездесуін.
- Нәтиже индексінде санаттар бар.

Нені тексереміз?

- Санауыштар қосындысын және барлық тең көшбасшыны.
- Жол бірлігін: толық құрам емес, жазба.

unique, nunique және value_counts жауаптары әртүрлі

Сұрақ	Құрал	Нәтиже
Қандай санаттар кездесті?	unique	Кездесу ретіндегі мәндер
Әртүрлі санат қанша?	nunique	Бір сан
Әр санатта неше жол бар?	value_counts	Санауыштар Series нысаны
Әр топтың көрсеткіші қандай?	groupby + агрегаттау	Series немесе DataFrame
Адам саны қанша?	Сенімді идентификатор	Аты жеткіліксіз

Бөлім мен жетіспейтін мәндер үлесті өзгертеді

Толық файл

- Алым: қажетті санаттың жазбалары.
- Бөлім: файлдың барлық есепке алынған жолы.
- Үлестер қосындысы есепке алу ережелеріне сай болуы тиіс.

Ішкі таңдама

- Алдымен қажетті жиынды таңдаймыз.
- Бөлім: оның есепке алынған жазбаларының саны.
- Жетіспейтін мәндерді айқын қосамыз не шығарамыз.

Бөлімдер әртүрлі әрі айқын берілсе, бір жиілік әртүрлі үлес бере алады.

Жиіліктер Series нысанынан есеп кестесіне

Python мысалы

```
frequency = (
    counts.rename_axis('Club')
    .reset_index(name='n_players')
)
frequency['share'] = frequency['n_players'] / len(demo)
frequency['percent'] = 100 * frequency['share']
print(frequency)
assert frequency.shape == (3, 4)
```

Нәтиже: B: 3 және 50%; A: 2 және 33.33%; C: 1 және 16.67%.

Айқын өрістер

- Club: санат; n_players: жол саны.
- reset_index белгілерді кәдімгі бағанға айналдырады.

Шкалалар

- share бірлік үлесімен өлшенеді.
- percent = 100 * share.

groupby: бөлу, есептеу, жинау

1

Топ кілті

Club немесе Club пен Position үйлесімі.

2

Көрсеткіш

Age, Wage не мағынасына сай таңдалған басқа белгі.

3

Функция

Сұрақ мағынасына сай орташа мән, қосынды, минимум не санауыш.

4

Бақылау

Топ саны, жазба сандары және ұяшықты тәуелсіз тексеру.

- Топтау нысаны дайын жауапты емес, бөлуді анықтайды.
- Өрістерді айқын таңдаймыз: сандық тип қосындыны мағыналы етпейді.

Клубтар бойынша орташа жас, көлем және қосынды

Python мысалы

```
grouped = demo.groupby('Club', sort=True)
print(grouped['Age'].mean().to_dict())
print(grouped.size().to_dict())
print(grouped['Wage'].sum().to_dict())
```

Нәтиже: Age: A=22, B=26, C=21. Көлемдері: 2, 3, 1. Wage: 40, 120, 15.

Бір топтау

- Әр Club кілті жолдар жиынын анықтайды.
- Таңдалған агрегатқа қарай нәтиже өзгереді.

Үш сұрақ

- Орташа жас; жазба саны; Wage қосындысы.
- Шағын клубты қолмен тексереміз.

size және count әрдайым сәйкес келмейді

Әдіс	Нені санайды	Бір Wage белгісіз болса
size	Топтың барлық жолын	Жазба санауышта қалады
Wage бойынша count	Толтырылған Wage мәндерін	Санауыш азаяды
Club бойынша nunique	Әртүрлі Club мәндерін	Жазба саны болып саналмайды
Wage бойынша mean	Белгілі мәндердің орташа мәнін	Қолданылатын жиын өзгереді

Топтың жазба санын және көрсеткішке қатысатын мәндер санын көрсетіңіз.

Сандық аралықтар: ені және шекаралары

Python мысалы

```
auto = demo['Age'].value_counts(bins=3, sort=False)
print(auto)
age_group = pd.cut(
    demo['Age'], bins=[17, 21, 25, 30],
    labels=['18-21', '22-25', '26-30'],
    include_lowest=True
)
print(age_group.value_counts(sort=False).to_dict())
assert age_group.notna().all()
```

Нәтиже: {'18-21': 2, '22-25': 2, '26-30': 2}

Автоматты аралықтар

- bins ауқымның тең енін береді.
- Аралықтардағы жазба саны әртүрлі болуы мүмкін.

Жарияланған шекаралар

- cut салыстырылатын ауқымдарды бекітеді.
- Шекараға кіру мен шектен шығуды тексереміз.

Жалпы орташа мәнге топ көлемдері қажет

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

Дұрыс салмақ

- n_g: топтық орташа мәндегі бақылау саны.
- Үлкен топ көбірек бастапқы мән қосады.

Басқа сұрақ

- Топтардың қарапайым орташа мәні оларға тең салмақ береді.
- Көлемдер әртүрлі болса, ол әдетте өзгеше болады.

(2*22 + 3*26 + 1*21) / 6 = 23.8333; (22+26+21)/3 = 23.

football топтары: көлемі және түсіндіру шектері

Көлем

- Екі жазбаның орташа мәні бір ойыншыға сезімтал.
- Агрегаттың жанына жазба санын көрсетеміз.

Қамту

- Клубтың қамтылуы файлға қатысты.
- Тарихи кесінді қазіргі құрамға тең емес.

Шкалалар

- Wage/Value: дереккөз бірліктері.
- Penalties: гол емес, рейтинг.

Шағын топтың жоғары орташа мәні оның құрамын және жазба санын тексеруді қажет етеді.

Топтық талдаудың жиі қателері

Санауыш

- nunique жол санына тең емес.
- Минимум мен максимумның барлық тең мәнін сақтаймыз.

Үлес

- Үлес пен пайыз әртүрлі шкаланы қолданады.
- Бөлімді ыңғайлылық емес, сұрақ анықтайды.

Ықшамдау

- bins тең жазба санын емес, енді білдіреді.
- Орташа мәндерді негізделген салмақпен біріктіреміз.

Қорытындыдан бұрын кесте жолының бірлігін, бөлімді және бақылау қорытындысын атаймыз.

Топтық есептің бес негізі

1. Жиілік жазбаларды санайды; үлес оларды жарияланған жиынға бөледі.
2. unique, nunique және value_counts әртүрлі жауап түрін береді.
3. groupby бөлуді анықтайды, агрегат топ көрсеткішін қалыптастырады.
4. size жолдарды, count толтырылған мәндерді санайды.
5. Аралықтар мен топтық орташа мәндер шекаралар мен салмақтарды бақылауды қажет етеді.

Бөлімдер мен топтарды тексеріңіз

1. Жиілік үлес пен пайыздан қалай ерекшеленеді?
2. count қашан size мәнінен кіші болады?
3. Name бойынша unique неліктен адам санын анықтамайды?
4. bins=4 неге кепілдік береді және неге кепілдік бермейді?
5. Топтық орташа мәндердің орташа мәні жалпы мәннен неге өзгеше болуы мүмкін?

Өзіндік жұмыс

demo ішіне бір жетіспейтін Wage қосыңыз. Клубтар бойынша size, count және mean салыстырыңыз; содан кейін дұрыс салмақпен жалпы орташа мәнді қалпына келтіріңіз.

Жиіліктер мен топтау жөніндегі дереккөздер

Құжаттама

1. pandas: топтау

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

2. pandas: value_counts

https://pandas.pydata.org/docs/reference/api/pandas.Series.value_counts.html

3. pandas: cut аралықтары

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

Дәрістің негізі

- Берілген 3-дәрістің презентациясы
- Берілген 3-зертханалық жұмыстың тапсырмалары
- Берілген 3-зертханалық жұмыстың жауаптары бар нұсқасы