ДЕРЕКТЕРДІ ИНТЕЛЛЕКТУАЛДЫ ТАЛДАУ

Дәріс 9

Деректерді барлау талдауы: EDA

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

- EDA үдерісін паспорттан жаңа гипотезаларға дейін жүргізу.
- Сиректікті, қатені және статистикалық ауытқыма мәнді ажырату.
- Топтар мен дереккөз шектеулерін ескеріп байланыстарды сипаттау.

Дәріс мазмұны

1. **Сұрақ және паспорт**
2. **Сапа және санаттар**
3. **Үлестірімдер**
4. **Байланыстар және кесінділер**
5. **Қорытындылар және гипотезалар**

EDA: сұрақ, сипаттау, тұрақтылықты тексеру

1

Сұрақты тұжырымдау

Талдау нысанын, белгілерді және бақылау бірлігін нақтылау.

2

Сапаны тексеру

Схеманы, жетіспейтін мәндерді, санаттар мен қайталануларды зерттеу.

3

Үлестірімдерді сипаттау

Ортаны, шашырауды, жиіліктерді есептеп, пішінді көрсету.

4

Байланыстарды зерттеу

Жалпы көріністі мазмұнды кесінділермен салыстыру.

5

Тұрақтылықты тексеру

Бақылауды нақтылап, келесі гипотезаларды тұжырымдау.

- EDA алдыңғы тексерулерге қайта оралуға мүмкіндік береді.
- Ережелер өзгерісі мен себептері журналда тіркеледі.

Белгілер сөздігі: тип мағынамен анықталады

Өріс	Семантикалық тип	Шектеу
Age	Сандық жас	Көлденең кесінді, траектория емес
Position	Номинал санат	Сандық қашықтық жоқ
Stamina	Дағды рейтингі	Физикалық өлшем бірлігі емес
Name	Жазбадағы ат	Бірегей ID болуына кепілдік жоқ
Unnamed: 0	Қызметтік индекс	Белгілерден шығарылады
Penalties	Дағды рейтингі	Соғылған гол саны емес

Wage/Value өлшем бірліктері мен кезеңі белгісіз; мұны айқын көрсетеміз.

Паспорт және алғашқы сандық сипаттама

Python мысалы

```
from pathlib import Path
import numpy as np
import pandas as pd

football = pd.read_csv(Path('data') / 'football.csv')
features = ['Age', 'Acceleration', 'SprintSpeed', 'Stamina']
print(football.shape)
print(football.isna().sum().sum())
print(football[features].describe())
```

Нәтиже: Пішін: (12897, 42); жетіспейтін мәндер саны: 0.

Паспорт

- 12897 жол, 42 баған.
- Жинақтағы CSV ішінде жетіспейтін ұяшықтар жоқ.

Сипаттау

- Белгілердің ортасы, квартильдері және ауқымы.
- Шеткі мән әлі дәлелденген қате емес.

Сапа: мазмұндық қорытындыдан бұрын санау

Деректерді тексереміз

- Жетіспейтін мәндер мен шартты белгілер.
- Таңдалған өрістердегі дәл қайталанулар.
- Санат жиіліктері және жазылу нұсқалары.

Жауаптарды араластырмаймыз

- Аттың қайталануы нысанның қайталануына тең емес.
- Санаттар саны жол санына тең емес.
- Сирек топ міндетті түрде қате емес.

Жиілік шегі — санаттың бар болу заңы емес, көрсету ережесі.

Санаттар: түсінікті қалыпқа келтіру

Python мысалы

```
category_demo = pd.Series(
    [' Alpha ', 'alpha', 'ALPHA', 'Beta', '', None],
    dtype='string'
)
category_key = category_demo.str.strip().str.casefold()
category_key = category_key.replace({'': pd.NA})
print(category_key.value_counts(dropna=False))
assert len(category_key) == len(category_demo)
```

Не өзгерді

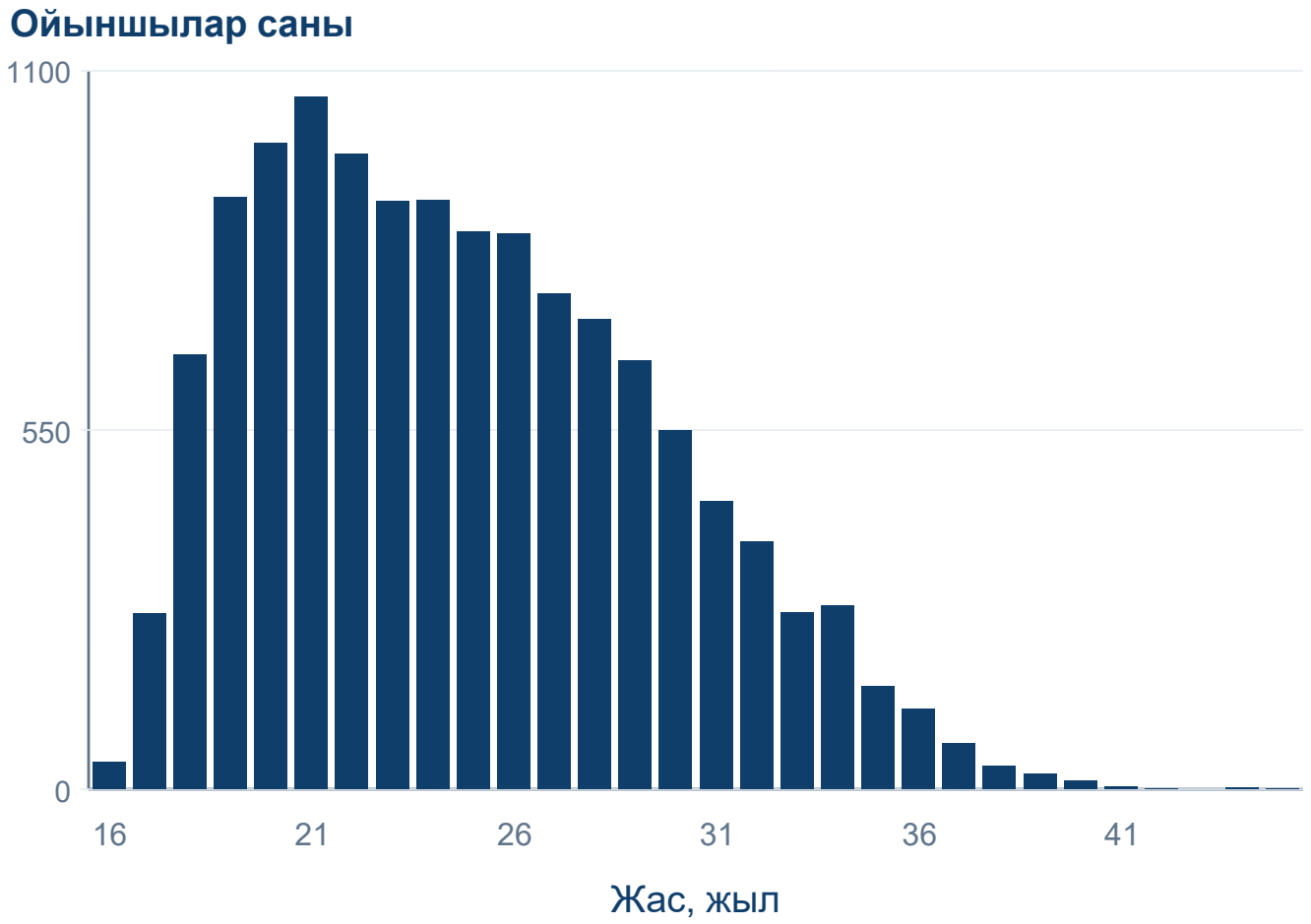
- alpha: 3 жазба; beta: 1 жазба.
- Бос жол мен None екі жетіспейтін мән береді.

Не сақталады

- Жол саны өзгермейді.
- Бастапқы жазылуы тексеруге қолжетімді.

Расталған сәйкестіктер сөздігі кездейсоқ алмастырулардан қауіпсіз.

Жас: үлестірім және таңдама құрамы



Үлестірімге қойылатын сұрақтар

- Орта қайда және шашырау қаншалықты кең?
- Басқа bins кезінде сипаттама сақтала ма?
- Жас құрамы позициялар арасында қалай ерекшеленеді?
- Қай қорытындыға іріктеу тәсілі туралы мәлімет қажет?

Әртүрлі ойыншылардың үлестірімі бір ойыншының траекториясы емес және құрамды тексеруді талап етеді.

IQR: ерекшелік белгісі, түпкілікті үкім емес

IQR = Q3 - Q1; шекаралар: Q1 - 1.5*IQR, Q3 + 1.5*IQR

Статистикалық ереже

- Ағымдағы жиын квартильдеріне тәуелді.
- Көпшілікке қатысты ерекше мәндерді бөледі.

Пәндік шешім

- Белгінің мағынасы мен мәнмәтінді талап етеді.
- Сирек нақты мән сақталады.

[1, 2, 3, 4, 100] үшін: Q1=2; Q3=4; шектері [-1, 7].

Белгі бөлек сақталады; жоюға қосымша негіз қажет.

Жұмыс мысалы: шеткі мәннің әсері

Python мысалы

```
toy = pd.Series([1., 2., 3., 4., 100.])
q1, q3 = toy.quantile([0.25, 0.75])
iqr = q3 - q1
lower, upper = q1 - 1.5 * iqr, q3 + 1.5 * iqr
flag = ~toy.between(lower, upper)
print(lower, upper)
print(toy.mean(), toy.median())
print(toy.loc[flag].tolist())
```

Нәтиже: -1.0 7.0; 22.0 3.0; [100.0]

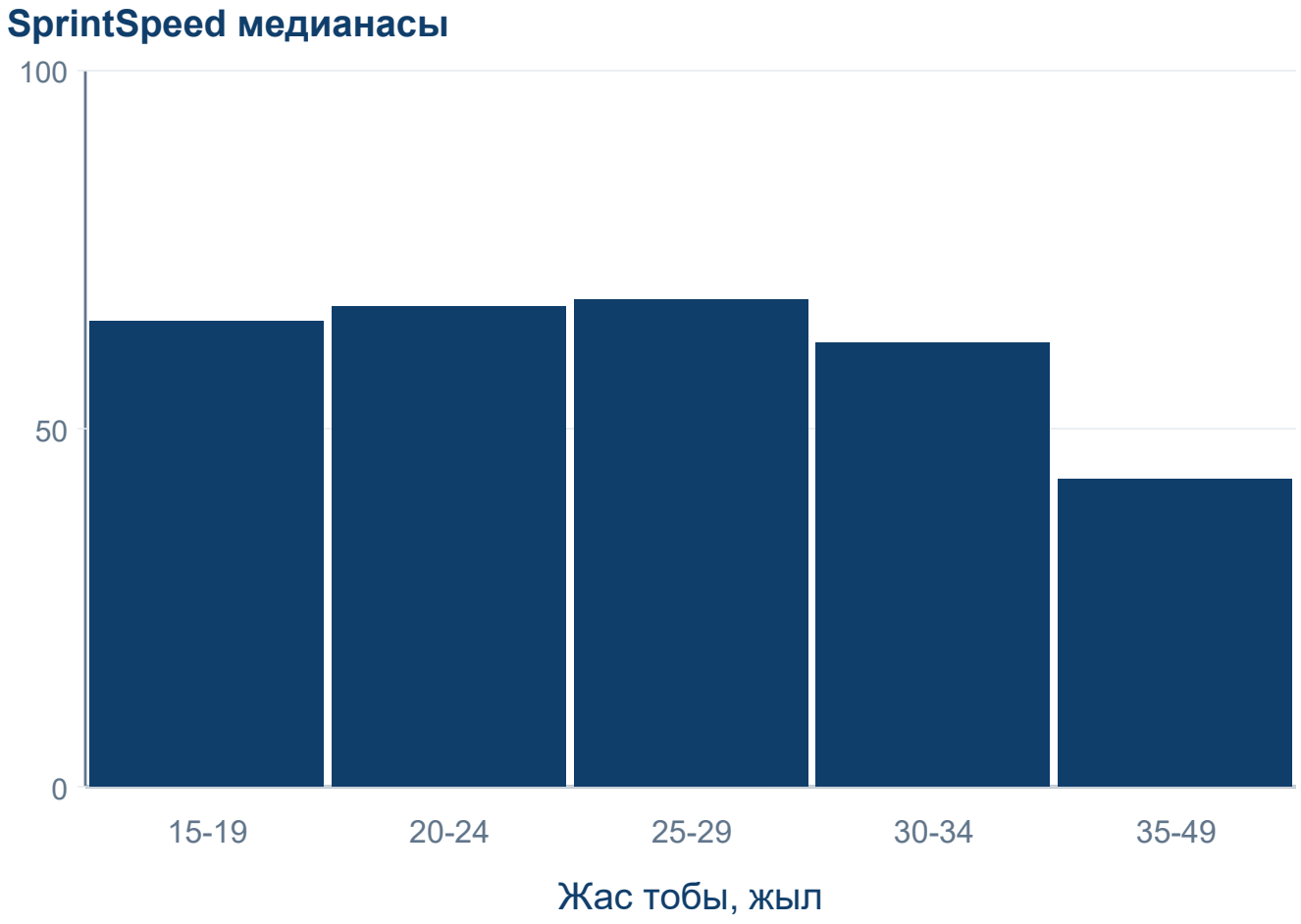
Есептеу

- Шектер: -1 және 7; үміткер: 100.
- Орташа мән 22, медиана 3.

Түсіндіру

- Шеткі мән орташа мәнді едәуір ығыстырады.
- Бұдан мән қате деген шешім шықпайды.

Жас пен жылдамдық: байланысқа мәнмәтін қажет



Байланысты абайлап оқимыз

- Жұптар бір тарихи көлденең кесіндіден алынған.
- Жас топтарының позициялық құрамы әртүрлі болуы мүмкін.
- Корреляция жеке өзгерісті емес, байланысты сипаттайды.
- Түсіндірмені тексеруге салыстырмалы кесінділер не жаңа деректер қажет.

Жалпы көрініс сипаттаудан топ құрамын тексеруге ауысыңыз.

Корреляциялар және позициялар бойынша қорытынды

Python мысалы

```
correlation = football[features].corr(method='pearson')
print(correlation)
top = football['Position'].value_counts().head(4).index
subset = football.loc[football['Position'].isin(top)]
groups = subset.groupby('Position').agg(
    n=('Stamina', 'size'),
    mean_age=('Age', 'mean'),
    median_stamina=('Stamina', 'median')
)
print(groups)
```

Сандық байланыс

- Пирсон сызықтық бірлескен өзгерісті сипаттайды.
- Белгілер жиыны мағынасы бойынша таңдалады.

Топтық мәнмәтін

- Қорытынды n және жас құрамын көрсетеді.
- Жалпы байланыс топ ішінде салыстыруды талап етеді.

Араластыру: жалпы жиын және жеке топтар

Жалпы байланыс

- Әртүрлі позициялар мен құрамдарды біріктіреді.
- Топтар арасындағы айырманы көрсетуі мүмкін.
- Себепті анықтамайды.

Кесінділерді тексеру

- Байланыс бағыты мен шамасын салыстырамыз.
- Көлемдер мен енгізу ережелерін хабарлаймыз.
- Балама түсіндірмелерді тұжырымдаймыз.

Кесінді бақылауды нақтылайды, бірақ өздігінен себептік механизмді дәлелдемейді.

Тексерілетін бақылауды қалай рәсімдеу керек

1

Бақылау

Нақты шаманы, топты немесе үлестірім пішінін көрсету.

2

Дәлел

Кестені, суретті, көлемді және есептеу ережесін келтіру.

3

Шектеу

Бұл деректер мен рәсім нені анықтамайтынын атау.

4

Келесі сұрақ

Кесінді, жаңа деректер немесе жеке тексеру жоспарын ұсыну.

- Тұрақтылықты тексеру түсінікті сипатталуы тиіс.
- Тек күтуді қолдайтын нұсқаларды таңдамаймыз.

EDA шегі: гипотеза және растау

Барлау

- Құрылымды байқап, гипотезалар жасайды.
- Кең іздеуге және қайта оралуға жол береді.
- Шектеулер мен белгісіздікті көрсетеді.

Растау

- Алдын ала анықталған сұрақты талап етеді.
- Алғышарттар мен көптік іздеуді ескереді.
- Тәуелсіз деректерді қажет етуі мүмкін.

Іздеуден кейін табылған айырманы алғышарттарды жеке тексермей дәлелденген әсер деп атамаңыз.

Дәрістің негізгі нәтижелері

1. EDA сұрақты, сапаны, үлестірімдер мен кесінділерді байланыстырады.
2. Белгі типі тек dtype арқылы емес, мағынамен анықталады.
3. IQR ерекшелікті белгілейді; қате туралы шешімге мәнмәтін қажет.
4. Жалпы байланыс топтардың әртекті құрамын көрсетуі мүмкін.
5. Бақылау дәлелмен, шектеумен және келесі сұрақпен бірге беріледі.

Бақылау сұрақтары және өзіндік жұмыс

1. EDA нәтижесі расталған гипотезадан қалай ерекшеленеді?
2. Бірегей дерлік баған модельге неге әрдайым пайдалы емес?
3. Пәндік қате IQR үміткерінен қалай ерекшеленеді?
4. Позиция ішіндегі байланыс жалпы байланыстан неге ерекшеленуі мүмкін?
5. Бақылаудың аналитикалық шешімге тұрақтылығын қалай тексеруге болады?

Өзіндік жұмыс

football бойынша бір бақылауды таңдаңыз. Екі түсіндірме ұсынып, әрқайсысына оларды ажыратуға көмектесетін кесіндіні немесе жаңа деректерді атаңыз.

Дереккөздер және құжаттама

Құжаттама

1. pandas: сипаттамалық есептеулер

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

2. pandas: топтау

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

3. NIST: EDA деген не

https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

Дәрістің негізі

- Берілген 9-дәрістің презентациясы
- Берілген 9-зертханалық жұмыс