Python for AI

Дәуренов ӘД

Дәріс 10. Python тілінде деректерді зерттеушілік талдау және визуализациялау

Дәрістің мақсаты:

- деректерді зерттеушілік талдаудың мәнін түсіну,
- деректерден заңдылықтарды анықтау және
- оларды көрнекі түрде ұсыну тәсілдерімен танысу.

Деректерді зерттеушілік талдау дегеніміз не?

Күнделікті өмірде көптеген деректер жиналады:

көлік қозғалысының жылдамдығы, аккумулятор заряды, ауа температурасы, студенттердің бағалары немесе кәсіпорын жабдықтарының көрсеткіштері.

Алайда деректердің көп болуы олардың мағынасы бірден түсінікті болады дегенді білдірмейді.

Үлкен кестеге қарап, негізгі үрдісті немесе ерекше жағдайды байқау қиын.

Деректерді зерттеушілік талдау — деректердің құрылымын, негізгі сипаттамаларын, ерекшеліктерін және өзара байланыстарын бастапқы зерттеу үдерісі.
Ағылшын тілінде бұл ұғым **Exploratory Data Analysis**, қысқаша **EDA** деп аталады.

Зерттеушілік талдау мынадай сұрақтарға жауап іздейді:

- Деректерде қандай ақпарат бар?
- Қандай мәндер жиі кездеседі?
- Көрсеткіштер қаншалықты өзгереді?
- Ерекше немесе күмәнді мәндер бар ма?
- Көрсеткіштер арасында байланыс байқала ма?

Мысал: электромобильдің энергия тұтынуы туралы деректер берілген. Зерттеу барысында энергия тұтынуының сапарлар арасында қалай өзгеретінін, қай жағдайда жоғарылайтынын және жылдамдықпен қандай байланысы бар екенін қарастыруға болады.

Зерттеушілік талдаудың жасанды интеллекттегі орны

Жасанды интеллект жүйесін құру үшін алдымен қолданылатын деректерді түсіну қажет.
Егер деректерде қателер, маңызды ақпараттың жетіспеуі немесе біржақты іріктеу болса, модельдің нәтижесі де сенімсіз болуы мүмкін.

Мысалы, электромобильдің жүріс қорын болжайтын модель тек жазғы сапарлардың деректерімен оқытылса, оның қысқы жағдайдағы болжамдары дәл болмауы ықтимал.

Зерттеушілік талдау:

- деректердің есепке сәйкестігін бағалауға;
- жетіспейтін ақпаратты байқауға;
- ықтимал пайдалы көрсеткіштерді анықтауға;
- модель құруға дейін бастапқы болжамдар қалыптастыруға көмектеседі.

Зерттеушілік талдаудың негізгі нәтижесі — деректер туралы түсінік және әрі қарай тексерілетін сұрақтар.

Деректердің түрлері

Дерек түрі	Сипаттамасы	Мысал
Сандық деректер	Санмен өлшенетін немесе саналатын шамалар	Жылдамдық, температура, жолаушылар саны
Санаттық деректер	Нысанның белгілі бір топқа жататынын көрсетеді	Көлік түрі, қозғалтқыш түрі
Реттік деректер	Санаттардың мағыналық реті бар	Төмен, орташа, жоғары
Уақыттық деректер	Оқиғаның уақытын немесе күнін көрсетеді	Өлшеу уақыты, сапар күні

Сандармен жазылған ақпараттың барлығы өлшенетін шама бола бермейді. Мысалы, көліктің сәйкестендіру нөмірінің орташа мәнін есептеу мағыналы нәтиже бермейді. Сондықтан талдаушы бағандағы мәндерді ғана емес, **олардың нені білдіретінін** түсінуі керек.

Деректерді түсіну және олардың сапасын бағалау

Талдаудың басында әрбір жол мен бағанның мағынасы анықталады.
Мысалы, көлік туралы кестедегі бір жол бір сапарды, бір көлікті немесе бір секундтағы өлшеуді білдіруі мүмкін. Бұл айырмашылық нәтижені түсіндіруге тікелей әсер етеді.

Негізгі сұрақтар:

- Деректер қайдан алынған?
- Қандай уақыт аралығын қамтиды?
- Бір жол нені сипаттайды?
- Өлшем бірліктері қандай?
- Қандай жағдайлар деректерге кірмеген?

Өлшем бірліктеріне ерекше назар аудару қажет. Бір бағанда қашықтық километрмен, ал басқа дереккөзде метрмен көрсетілсе, оларды тікелей салыстыру қате қорытындыға әкеледі.
Деректер сапасының жиі кездесетін мәселелеріне бос мәндер, қайталанатын жазбалар, қате өлшемдер және әртүрлі жазылған санаттар жатады.

Тазалау кезінде қателер түзетіледі, ал зерттеушілік талдау кезінде деректердің мағынасы мен ерекшеліктері қарастырылады.

Іс жүзінде бұл үдерістер бір-бірімен байланысты: талдау барысында жаңа мәселе анықталса, деректерді қайта тексеру қажет болуы мүмкін.

Деректердің негізгі статистикалық сипаттамалары

Деректерді қысқаша сипаттайтын төрт көрсеткіш

Орташа мән

Барлық мәндердің қосындысы / мәндер саны

10 12 13 15 100

(10 + 12 + 13 + 15 + 100) / 5 = 30 мин

Шеткі мәндерге сезімтал

Медиана

Реттелген деректердің ортасындағы мән

10 12 **13** 15 100

Медиана = 13 мин

Жұп санда: ортадағы екі мәннің орташа шамасы

Ең кіші және ең үлкен мәндер

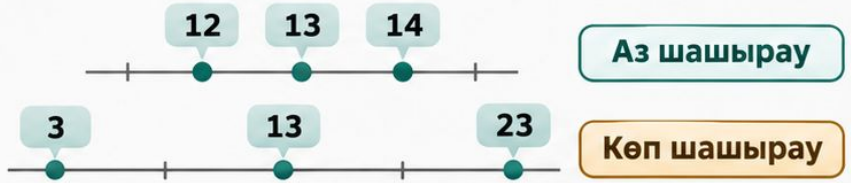
Деректердің шекараларын көрсетеді



Ауқым = 100 – 10 = 90 мин

Шашырау

Мәндердің бір-бірінен айырмашылығы



Екі қатардың да орташа мәні = 13

Көрсеткіштер: ауқым, дисперсия, стандартты ауытқу

Деректердің таралуы

Деректердің таралуы қандай мәндердің жиі, ал қандай мәндердің сирек кездесетінін көрсетеді.

Мысалы, қаладағы көліктердің жылдамдығын зерттегенде:

- көліктердің көбі белгілі бір жылдамдық аралығында жүруі;
- кептеліс кезінде төмен жылдамдықтардың жиі кездесуі;
- жоғары жылдамдықтардың сирек тіркелуі мүмкін.

Таралу шамамен симметриялы, бір жаққа созылған немесе бірнеше айқын шоғырдан тұруы мүмкін.

Егер сапар ұзақтығында екі шоғыр байқалса, деректерде қысқа қалалық сапарлар мен ұзақ қалааралық сапарлар аралас болуы ықтимал. Мұндай жағдайда оларды бөлек зерттеу пайдалы.

Ерекше мәндер және аномалиялар

Ерекше мән — басқа бақылаулардан айтарлықтай ерекшеленетін мән.
Мысалы, температура өлшемдерінің басым бөлігі 20–40 °C аралығында болып, бір өлшем 150 °C көрсетсе, бұл мәнді тексеру қажет.
Оның бірнеше себебі болуы мүмкін:

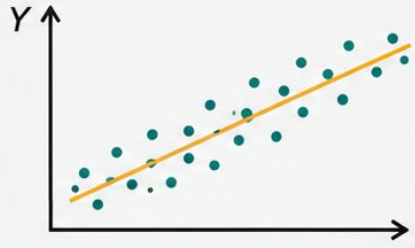
- датчиктің немесе жазбаның қатесі;
- өлшем бірлігінің дұрыс көрсетілмеуі;
- сирек кездесетін нақты оқиға;
- жабдық жұмысының өзгеруі.

Ерекше мәнді автоматты түрде қате деп есептеуге болмайды.
Өнеркәсіпте мұндай мән ақаудың алғашқы белгісі болуы мүмкін. Көлік жүйесінде ол күрт тежеуді, қатты қызуды немесе қалыптан тыс энергия тұтынуды көрсетуі ықтимал.
Талдаушы ерекше мәннің пайда болу себебін зерттеп, оны сақтау немесе түзету туралы негізделген шешім қабылдайды.

Көрсеткіштер арасындағы байланыс

Екі көрсеткіш бірге қалай өзгереді?

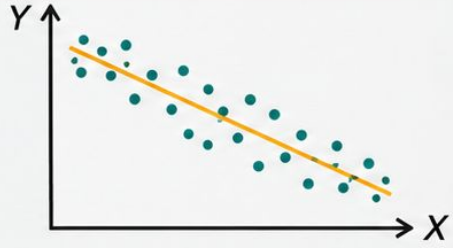
Оң байланыс



Бір көрсеткіш өссе, екіншісі де өсуге бейім.

Оқу уақыты ↗ — емтихан нәтижесі ↗

Теріс байланыс



Бір көрсеткіш өссе, екіншісі төмендеуге бейім.

Жылдамдық ↗ — сапар уақыты ↘

(Бірдей қашықтықта)

Айқын сызықтық байланыс жоқ



Жүйелі сызықтық өзгеріс байқалмайды.

Пирсон корреляциясы: $-1 \leq r \leq +1$

−1	0	+1
Күшті теріс	Сызықтық байланыс әлсіз	Күшті оң

r нөлге жақын болса да, сызықтық емес байланыс болуы мүмкін.

Корреляция себеп-салдар байланысын дәлелдемейді.

Нәтижеге басқа факторлар да әсер етуі мүмкін.

Деректерді визуализациялау

Деректерді визуализациялау — ақпаратты графиктер мен диаграммалар арқылы көрсету.

Кесте нақты мәндерді оқуға ыңғайлы болса, график жалпы көріністі байқауға көмектеседі.

Визуализация арқылы:

- топтарды салыстыруға;
- уақыт бойынша өзгерісті бақылауға;
- таралуды көруге;
- байланысты зерттеуге;
- ерекше мәндерді байқауға болады.

График түрі қойылған сұраққа сәйкес таңдалуы керек.

Зерттеу сұрағы	Қолайлы график	Мысал
Көрсеткіш уақыт бойынша қалай өзгереді?	Сызықтық график	Сапар кезіндегі аккумулятор заряды
Қай топта көрсеткіш жоғары?	Бағандық диаграмма	Көлік түрлері бойынша орташа тұтыну
Қандай мәндер жиі кездеседі?	Гистограмма	Сапар ұзақтығының таралуы
Екі көрсеткіш арасында байланыс бар ма?	Нүктелік диаграмма	Жылдамдық пен энергия тұтынуы
Топтардың таралуы қалай ерекшеленеді?	Қораптық диаграмма	Әртүрлі бағыттардағы сапар ұзақтығы

Деректерді визуализациялау түрлері

Мақсатқа сай графикті таңдаңыз

Сызықтық график



Уақыт бойынша өзгерісті көрсетеді.

Мысал: аккумулятор заряды

Бағандық диаграмма



Жеке санаттарды салыстырады.

Мысал: орташа тәуліктік жүріс

Гистограмма



Сандық деректердің таралуын көрсетеді.

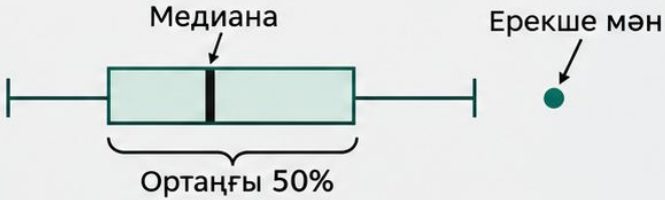
Нүктелік диаграмма



Екі көрсеткіштің байланысын көрсетеді.

Әр нүкте — бір бақылау

Қораптық диаграмма



Медиананы және шашырауды көрсетеді.

Ерекше мән міндетті тұрде қате дерек емес.

Графикті дұрыс оқу және құрастыру

Графиктің көрнекі болуы жеткіліксіз: ол ақпаратты дұрыс жеткізуі керек.
Негізгі талаптар:

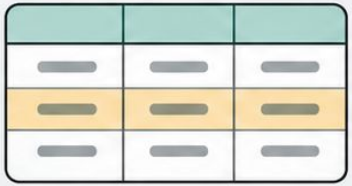
- **Түсінікті тақырып:** графикте не көрсетілгені анық болуы керек.
- **Осьтердің атауы:** әр осьтің мағынасы жазылады.
- **Өлшем бірліктері:** минут, километр, пайыз сияқты бірліктер көрсетіледі.
- **Орынды масштаб:** айырмашылықтар жасанды түрде үлкейтілмеуі керек.
- **Оқылатын безендіру:** артық түстер мен сәндік элементтер ақпаратты қабылдауға кедергі келтірмеуі керек.

Мысалы, бағандық диаграмманың тік осін нөлден бастамау шағын айырмашылықты өте үлкен етіп көрсетуі мүмкін.
Түс те мағынаға сәйкес қолданылғаны дұрыс. Қауіпті температураны қызыл түспен белгілеу түсінуді жеңілдетеді. Бірақ ақпаратты тек түспен жеткізбей, атаулар мен белгілерді де қолдану қажет.

Python құралдарының рөлі

Деректерді талдау және визуализациялау — бір ортада

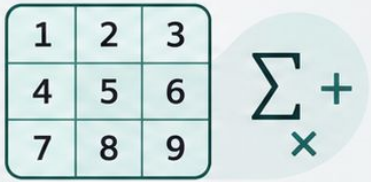
Pandas



Кестелік деректер

Оқу • іріктеу
Топтау • жинақтау

NumPy



Сандық массивтер

Массивтермен жұмыс
Сандық есептеулер

Matplotlib



Графиктер мен диаграммалар

Сызықтық график • гистограмма
Бағандық • нүктелік диаграмма

Seaborn



Статистикалық визуализация

Таралуларды көрсету
Топтарды салыстыру

Құралдар Есептейді және бейнелейді

Талдаушы Зерттеу сұрағын қояды → графикті таңдайды → нәтижені түсіндіреді

Графикті түсіндіру және қорытынды жасау — талдаушының міндеті.

Нәтижені түсіндіру

Талдау нәтижесі есептелген сандар мен салынған графиктердің тізімімен шектелмейді. Олардың мағынасын түсіндіру қажет.
Мысалы:
Бақылау: қысқы сапарларда энергия тұтынуы жоғары болған.
Ықтимал түсіндірме: төмен температура, салонды жылыту және жол жағдайлары әсер етуі мүмкін.
Қосымша тексеру: сапар ұзақтығы, жылдамдық және жүктеме бойынша ұқсас сапарларды салыстыру қажет.

Бұл мысалда байқалған нәтиже мен оның ықтимал себебі бөлек көрсетілген.
Қорытынды жасағанда деректердің көлемін, қайдан алынғанын және қандай жағдайларды қамтитынын ескеру қажет. Бір көліктің бірнеше сапары бойынша алынған нәтижені барлық электромобильдерге бірден таратуға болмайды.

Бақылау сұрақтары

1. Деректерді зерттеушілік талдау қандай мақсатта жүргізіледі?
2. Жасанды интеллект моделін құру алдында деректерді зерттеу не үшін қажет?
3. Орташа мән мен медиананың айырмашылығы қандай?
4. Орташа мәндері бірдей деректердің таралуы әртүрлі болуы мүмкін бе?
5. Ерекше мәндерді неге бірден жоюға болмайды?
6. Корреляция неліктен себеп-салдар байланысын дәлелдемейді?
7. Бағандық диаграмма мен гистограмманың айырмашылығы қандай?
8. Екі сандық көрсеткіштің байланысын қандай диаграмма арқылы зерттеуге болады?
9. Графиктегі масштаб нәтижені қабылдауға қалай әсер етеді?
10. Талдау қорытындысын жасағанда қандай шектеулерді ескеру қажет?