

№ 6 Дәріс. Exploratory Data Analysis (EDA) — Деректерді зерттеушілік талдау.

Дәрістің мақсаты:

- Деректерді зерттеушілік талдаудың мәні мен маңызын түсіндіру
- Негізгі EDA әдістерін қарастыру
- Деректерді визуализациялау және статистикалық талдау тәсілдерін түсіндіру
- EDA нәтижелерін дұрыс интерпретациялауды үйрету

Кілт сөздер:

Exploratory Data Analysis, EDA, histogram, boxplot, scatter plot, correlation, outliers, missing values, pandas, seaborn, data cleaning, feature analysis, statistical distribution.

Кіріспе

Қазіргі заманда деректер кез келген саланың стратегиялық ресурсы болып табылады. Бизнес, ғылым, медицина, қаржы, экономика, білім беру — барлық бағытта шешім қабылдау деректерге сүйенеді. Алайда деректің өзі дайын жауап бермейді. Оны түсініп, құрылымдап, жасырын заңдылықтарды анықтау қажет. Осы кезде Exploratory Data Analysis (EDA) — деректерді зерттеу және түсіндіру әдістері маңызды рөл атқарады.

EDA — бұл деректерді талдау процесінің алғашқы және ең маңызды кезеңі. Ол арқылы деректердің құрылымы, сапасы, ерекшеліктері, жасырын тәуелділіктері анықталады. Бұл кезеңде статистикалық әдістер мен визуализация құралдары пайдаланылады. Нәтижесінде зерттеуші немесе аналитик модель құруға дейін деректердің табиғатын түсінеді, қателер мен ауытқуларды анықтап, гипотезалар құра бастайды.

Тарихи тұрғыдан, EDA ұғымын Джон Тьюки (John Tukey) енгізді. Оның мақсаты деректер арқылы ойлау мәдениетін қалыптастыру болатын. Тьюкидің көзқарасы бойынша, EDA — тек әдістер жиынтығы емес, деректерге сыни көзқараспен қарау, сұрақтар қою және болжам жасау процесі.

Бүгінде EDA машиналық оқыту жүйелерін әзірлеуде, бизнес-аналитикада, ғылыми зерттеулерде, статистикалық бақылауда кең қолданылады. Машиналық оқытудағы көптеген қателіктер дұрыс модель таңдамаудан емес, деректерді терең түсінбеуден туындайды. Сондықтан EDA — Data Science және Machine Learning саласындағы міндетті және жауапты кезең.

1. EDA түсінігі және рөлі

EDA дегеніміз — деректерді визуализация және статистикалық әдістер арқылы зерттеу процесі:

- Өзгермелі айнымалылар арасындағы байланыстарды анықтау;
- Деректердің құрылымы мен үлестірілуін түсіну;
- Шулы деректер мен аномалияларды табу;
- Болжамдарды қалыптастыру;

- Модель үшін маңызды айнымалыларды анықтау.
- Ережесі:* “Data first — model second”

2. EDA кезеңдері

Қадам	Мазмұны
1. Деректерді жүктеу	Dataset-ті оқу (CSV, SQL, API)
2. Жалпы сипаттама	Өлшем, тип, missing values, duplicates
3. Статистикалық талдау	mean, median, std, range, IQR
4. Визуализация	Histogram, boxplot, heatmap
5. Outliers анықтау	Нормадан ауытқыған мәндер
6. Корреляция талдауы	Айнымалылар арасындағы байланыс
7. Нәтижені отчет түрінде беру	Tүсіру, summary report

3. Статистикалық әдістер

Негізгі метрикалар

Метрика	Түсінігі
Mean	Орташа мән
Median	Орталық мән
Mode	Ең жиі кездесетін мән
Standard deviation	Шашыраңқылық
Variance	Дисперсия
Min/Max	Диапазон
Skewness	Таралу асимметриясы
Kurtosis	Құйрық тығыздығы

4. Визуализация әдістері

График	Мақсаты
Histogram	Үлестірілуді көру
Boxplot	Outliers анықтау
Scatterplot	Корреляция қарау
Heatmap	Корреляция матрицасы
Bar chart	Категорияларды салыстыру
Pairplot (Seaborn)	Бірнеше айнымалы қатынасы

5. Outliers және Missing values

Missing values өңдеу тәсілдері

- Delete Rows
- Median/Mean/Mode imputation

- KNN Imputation
- Interpolation

Outliers өңдеу тәсілдері

- Boxplot/IQR Rule
- Z-score
- Winsorization

6. Корреляция және айнымалы байланыстары

Корреляция коэффициенттері:

Түрі	Қолдану аймағы
Pearson	Сандық айнымалылар
Spearman	Реттелген айнымалылар
Kendall	Рангы аз dataset

Корреляция интерпретациясы:

r мәні	Байланыс
0 – 0.3	Әлсіз
0.3 – 0.7	Орташа
0.7 – 1.0	Күшті

7. Python құралдары

Техника	Кітапханалар
Data Loading	pandas
Statistical Summary	numpy, pandas
Visualization	matplotlib, seaborn, plotly
Profiling Tools	pandas-profiling, Sweetviz

8. EDA нәтижесін ұсыну

- Жиынтық кесте (summary table)
- Визуализация (графиктер)
- Қорытынды мен ұсыныстар
- Data Insights (жаңалықтар, тренд)

Қорытынды

Exploratory Data Analysis — деректермен жұмыс жасаудың бастамасы ғана емес, ең маңызды кезеңдерінің бірі. Ол деректердің сапасын бағалауға, ішкі құрылымын түсінуге, маңызды тенденциялар мен заңдылықтарды табуға мүмкіндік береді. EDA-ның мақсаты — тек график салу немесе есеп шығару емес, деректердің шынайы мазмұнын түсіну, дұрыс сұрақтар қою және дұрыс шешімдер қабылдау.

Жақсы орындалған EDA:

- модельдің дәлдігін арттырады;

- кателерді ерте кезеңде анықтайды;
- бизнес және ғылыми гипотезаларды нақтылайды;
- аналитикалық ойлауды дамытады;
- деректерге негізделген шешімдер қабылдауды жеңілдетеді.

Болжам құру, машиналық оқыту алгоритмдерін қолдану немесе статистикалық модельдер жасау кезінде EDA — сенімді негіз қалыптастырады. Сондықтан бұл процесті механикалық емес, шығармашылық және ғылыми ізденіс ретінде қарастыру қажет.

Қорытындылай келе, EDA — сәтті талдау мен модельдеудің кілті. Деректерді түсінбей тұрып дұрыс шешім жасалмайды. Сол себепті әрбір зерттеуші, аналитик немесе Data Scientist EDA әдістерін терең меңгеруі тиіс.

Бақылау сұрақтары

1. EDA дегеніміз не?
2. EDA кезеңдерін атаңыз.
3. Missing values қалай өңделеді?
4. Outliers анықтау әдістерін атаңыз.
5. Корреляция деген не және оның түрлері қандай?
6. Pairplot қай кезде қолданылады?

Пайдаланылған әдебиеттер

1. Tukey, J. *Exploratory Data Analysis*, Addison-Wesley.
2. Wes McKinney — *Python for Data Analysis*
3. Seaborn, Matplotlib ресми құжаттамасы
4. Kaggle Machine Learning EDA notebooks
5. Pandas Documentation, 2024