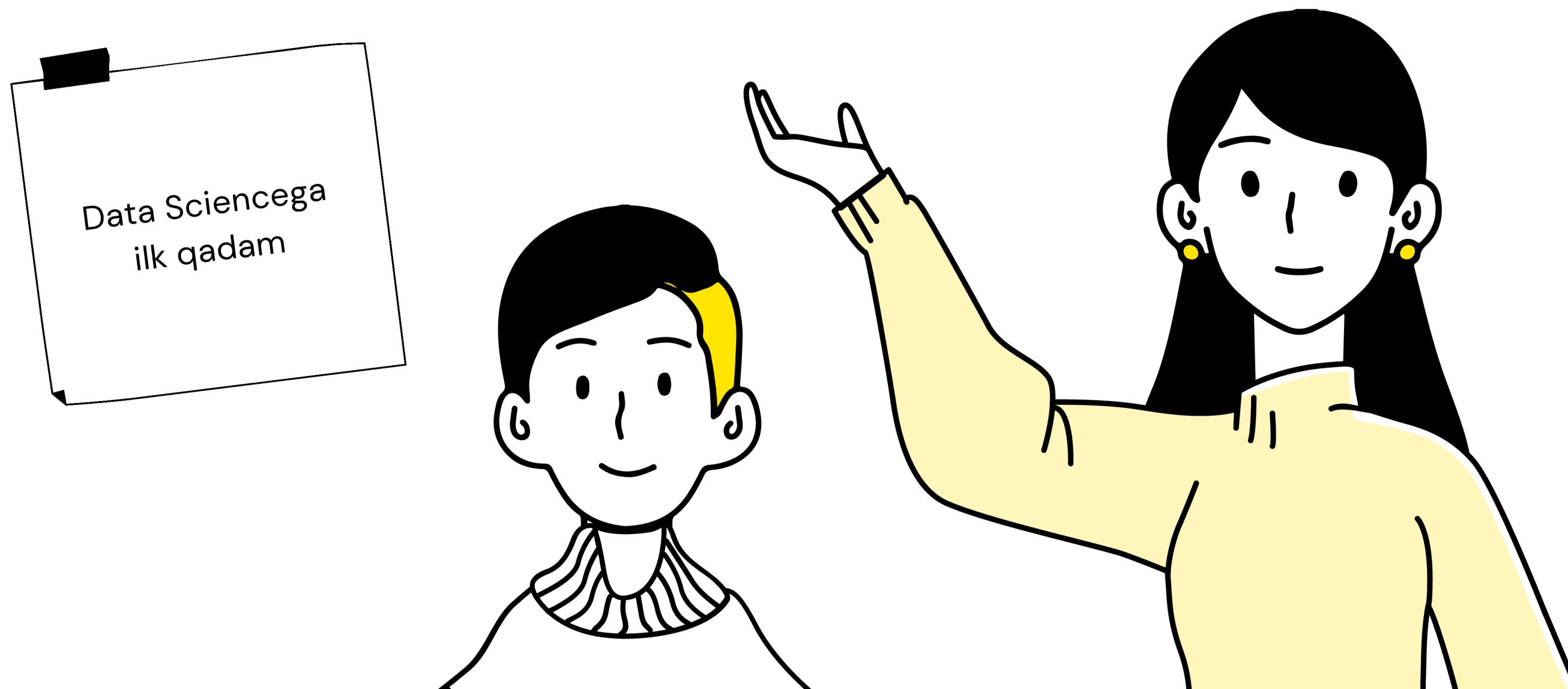


Exploratory Data Analysis (EDA)

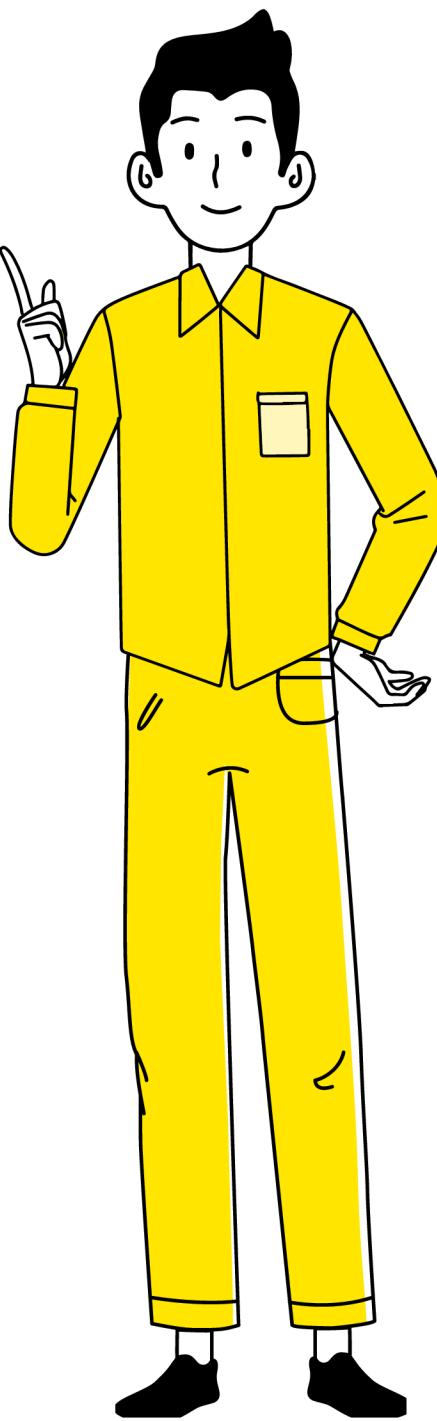


Bexruz Nutfiloyev

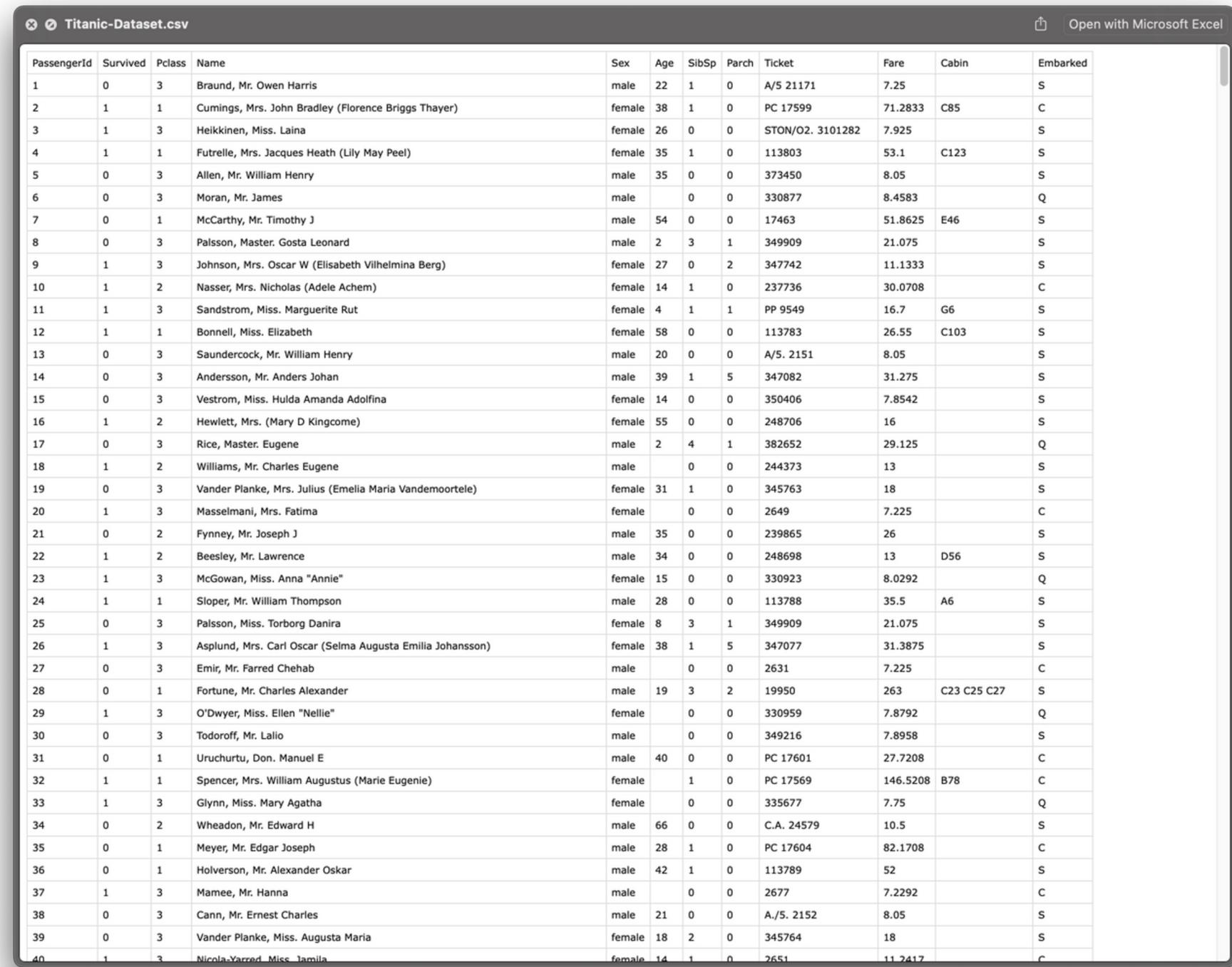
ML Engineer at AIIG



EDA nima?

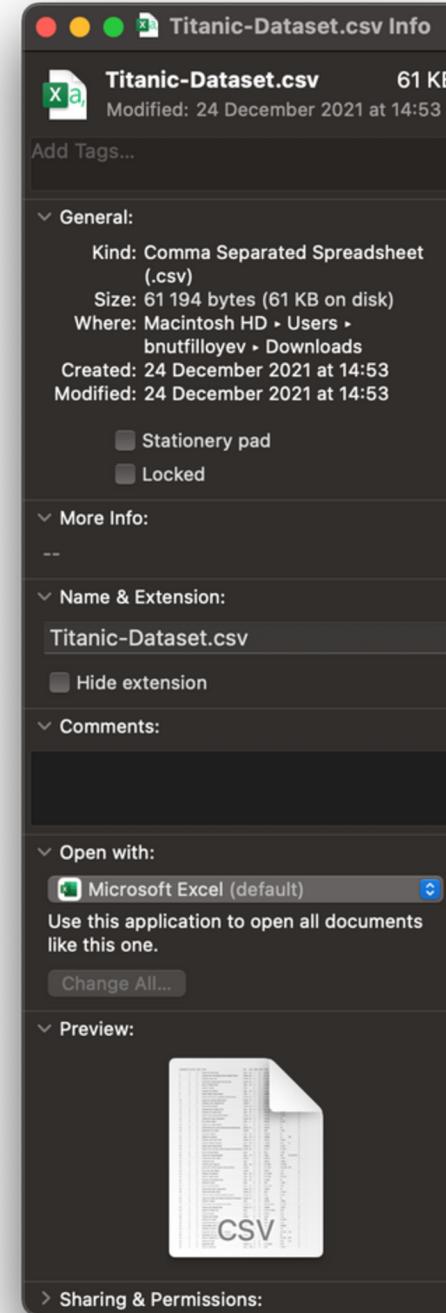


Ma'lumotning shakli va hajmi (rows, columns)

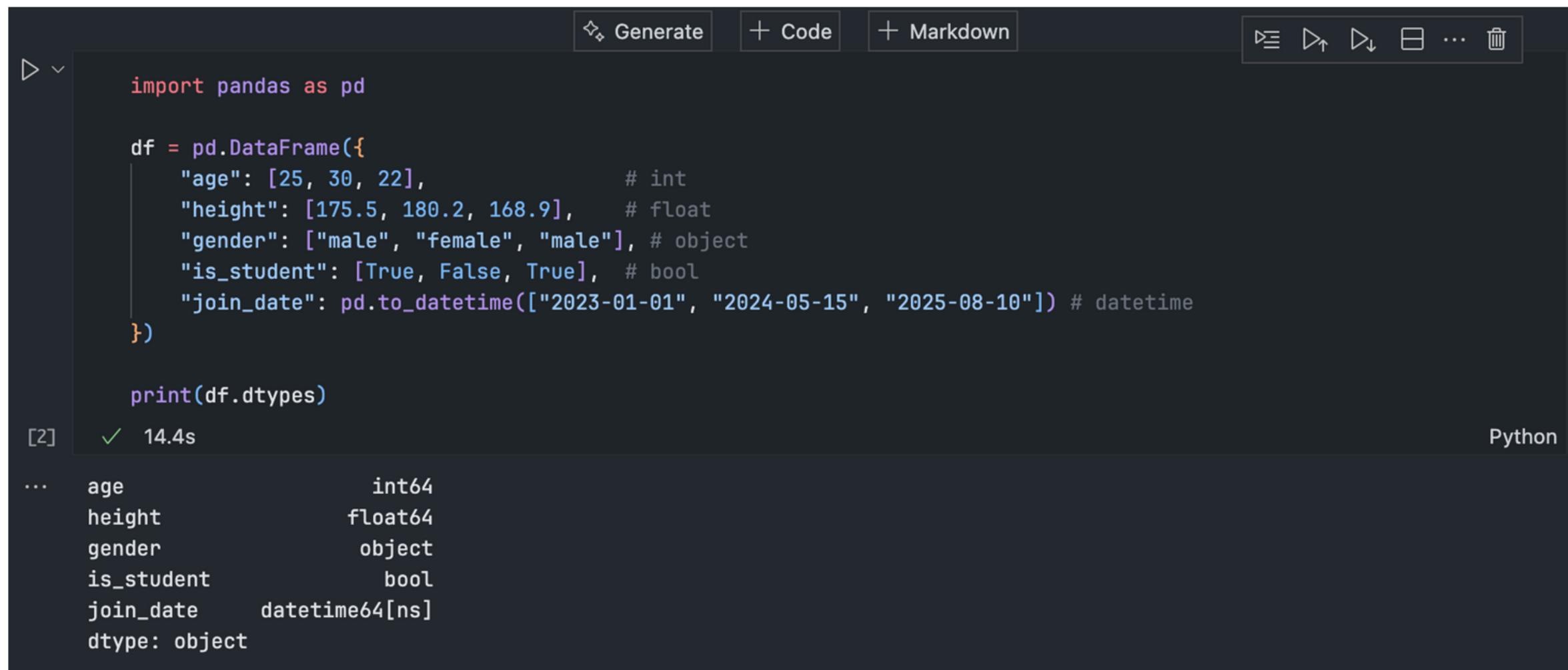


The image shows a screenshot of a CSV file named "Titanic-Dataset.csv" in a Mac OS X window. The window title bar says "Titanic-Dataset.csv". At the top right, there is a button labeled "Open with Microsoft Excel". The main content is a table with 40 rows and 13 columns. The columns are: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked. The first few rows of data are:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saunderscock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16		S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
18	1	2	Williams, Mr. Charles Eugene	male		0	0	244373	13		S
19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	345763	18		S
20	1	3	Masselmani, Mrs. Fatima	female		0	0	2649	7.225		C
21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26		S
22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S
23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0292		Q
24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6	S
25	0	3	Palsson, Miss. Torborg Danira	female	8	3	1	349909	21.075		S
26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson)	female	38	1	5	347077	31.3875		S
27	0	3	Emir, Mr. Farred Chehab	male		0	0	2631	7.225		C
28	0	1	Fortune, Mr. Charles Alexander	male	19	3	2	19950	263	C23 C25 C27	S
29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female		0	0	330959	7.8792		Q
30	0	3	Todoroff, Mr. Lailo	male		0	0	349216	7.8958		S
31	0	1	Uruchurtu, Don. Manuel E	male	40	0	0	PC 17601	27.7208		C
32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female		1	0	PC 17569	146.5208	B78	C
33	1	3	Glynn, Miss. Mary Agatha	female		0	0	335677	7.75		Q
34	0	2	Wheadon, Mr. Edward H	male	66	0	0	C.A. 24579	10.5		S
35	0	1	Meyer, Mr. Edgar Joseph	male	28	1	0	PC 17604	82.1708		C
36	0	1	Holverson, Mr. Alexander Oskar	male	42	1	0	113789	52		S
37	1	3	Mamee, Mr. Hanna	male		0	0	2677	7.2292		C
38	0	3	Cann, Mr. Ernest Charles	male	21	0	0	A./5. 2152	8.05		S
39	0	3	Vander Planke, Miss. Augusta Maria	female	18	2	0	345764	18		S
40	1	3	Nicola-Yarred, Miss. Jamila	female	14	1	0	2651	11.2417		C



Ma'lumot turlari va ularning ma'nosi



The screenshot shows a Jupyter Notebook cell with the following code:

```
import pandas as pd

df = pd.DataFrame({
    "age": [25, 30, 22],           # int
    "height": [175.5, 180.2, 168.9], # float
    "gender": ["male", "female", "male"], # object
    "is_student": [True, False, True], # bool
    "join_date": pd.to_datetime(["2023-01-01", "2024-05-15", "2025-08-10"]) # datetime
})

print(df.dtypes)
```

The cell output is:

[2] ✓ 14.4s

```
... age          int64
height        float64
gender        object
is_student     bool
join_date    datetime64[ns]
dtype: object
```

The cell has a Python label in the bottom right corner.

Statik xulosalar

```
import pandas as pd

# Misol uchun Titanic dataset
df = pd.read_csv('Titanic-Dataset.csv')

# Faqat raqamli ustunlar uchun qisqa statistika
stats = df.describe() # count, mean, std, min, 25%, 50%, 75%, max

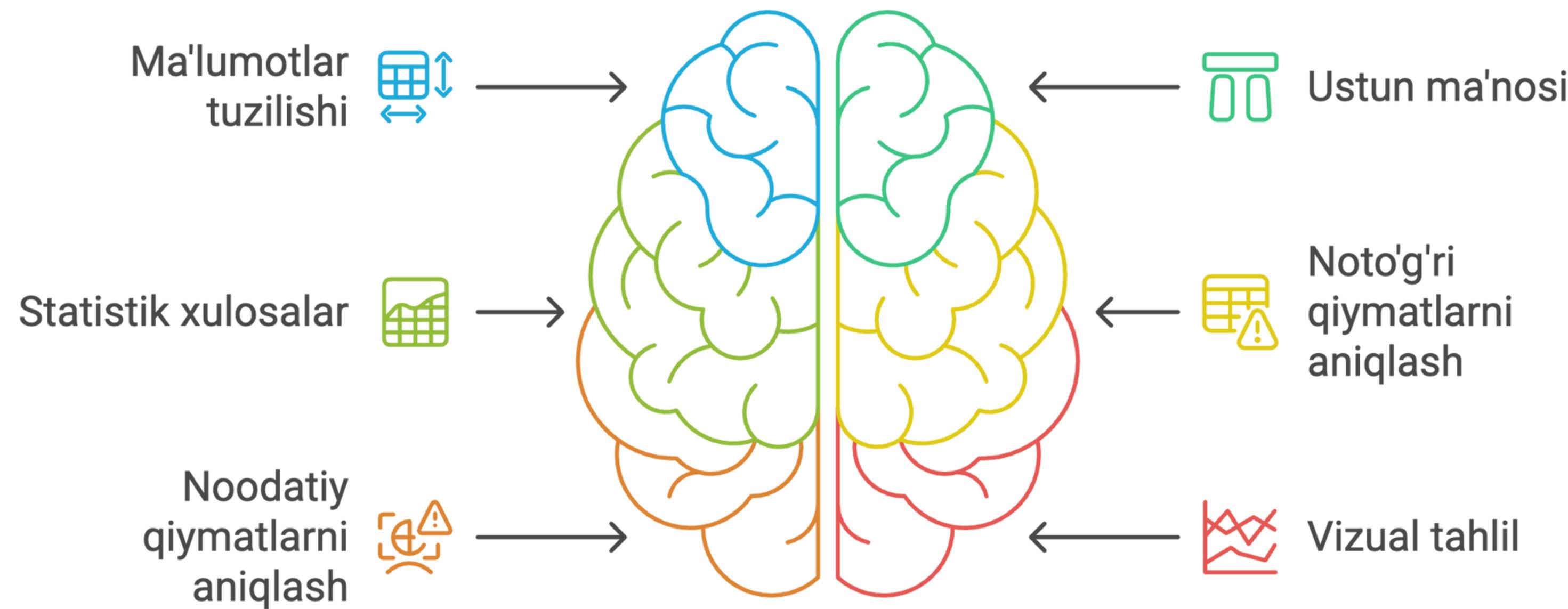
print("📊 Asosiy statistik ma'lumotlar:\n", stats)

# Qo'shimcha – faqat kerakli ustunlar bo'yicha alohida ko'rsatkichlar
print("\nMean qiymatlar:", df.mean(numeric_only=True))
print("\nMedian qiymatlar:", df.median(numeric_only=True))
print("\nMinimum qiymatlar:", df.min(numeric_only=True))
print("\nMaximum qiymatlar:", df.max(numeric_only=True))
print("\nStandart og'ish (std):", df.std(numeric_only=True))

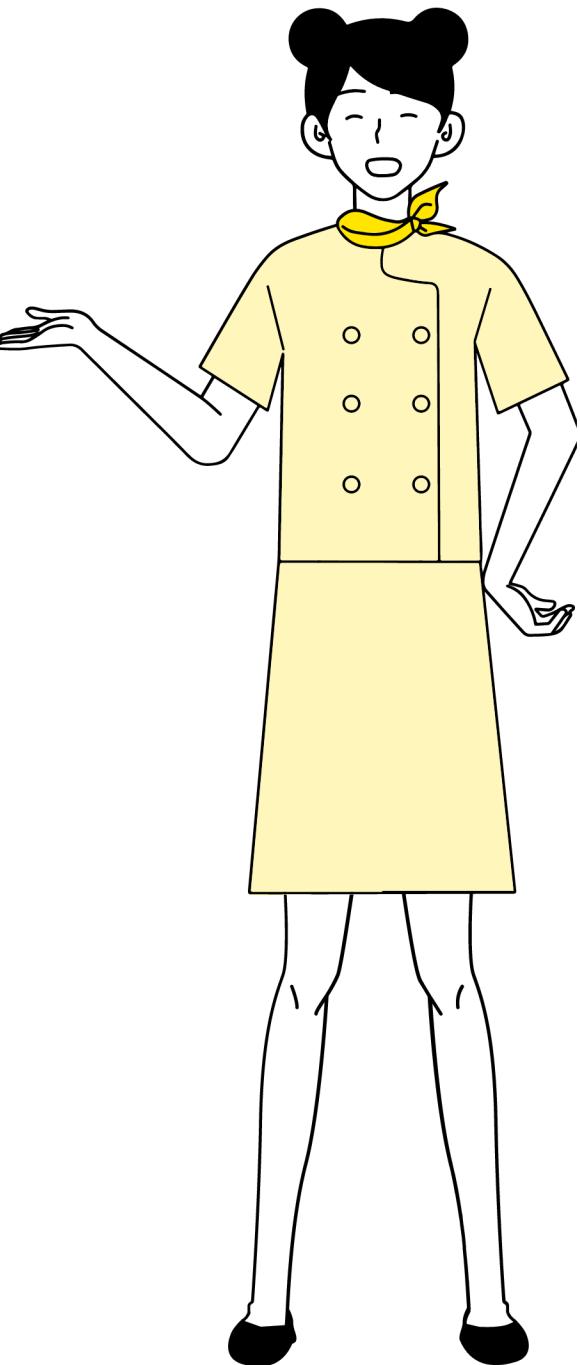
"""
Izoh
• describe() – bitta buyruq bilan eng koʻp ishlataladigan statistikani chiqaradi.
• mean() – oʻrtacha qiymat
• median() – median qiymat (ortadagi)
• min() / max() – eng kichik va eng katta qiymat
• std() – standart ogʼish (ma'lumotlar tarqalishi)
• numeric_only=True – faqat raqamli ustunlarni hisoblaydi, matn ustunlar xatolik bermasligi uchun.
"""

0.0s Python
```

Ma'lumotlarni tahlil qilishning asosiy jihatlari



Qayerda ishlatiladi?





Bank

Kreditga layoqatli guruhlarni aniqlash



Tibbiyot

Qon tahlilidagi anomaliyalarni aniqlash



E-commerce

Eng ko'p sotiladigan mahsulotlarni aniqlash



Sport

Eng samarador o'yinchilarni aniqlash

Xulosa

- **EDA** – ma'lumotlarni tahlil qilishning birinchi va eng muhim bosqichi bo'lib, u datasetni chuqur tushunishga yordam beradi.

Xulosa

- **EDA** – ma'lumotlarni tahlil qilishning birinchi va eng muhim bosqichi bo'lib, u datasetni chuqur tushunishga yordam beradi.
- Ma'lumotning **shakli (rows, columns)**, ustunlarning turlari va ma'nolari aniqlanadi.

Xulosa

- **EDA** – ma'lumotlarni tahlil qilishning birinchi va eng muhim bosqichi bo'lib, u datasetni chuqur tushunishga yordam beradi.
- Ma'lumotning **shakli (rows, columns)**, ustunlarning turlari va ma'nolari aniqlanadi.
- **Qisqa statistik xulosalar** (mean, median, min, max, std) yordamida umumiyl tendensiya va qiymatlar diapazoni ko'rildi.

Xulosa

- **EDA** – ma'lumotlarni tahlil qilishning birinchi va eng muhim bosqichi bo'lib, u datasetni chuqur tushunishga yordam beradi.
- Ma'lumotning **shakli (rows, columns)**, ustunlarning turlari va ma'nolari aniqlanadi.
- **Qisqa statistik xulosalar** (mean, median, min, max, std) yordamida umumiyligi tendensiya va qiymatlar diapazoni ko'rildi.
- **Yetishmayotgan qiymatlar** aniqlanadi va ularni to'ldirish yoki olib tashlash strategiyasi belgilanadi.

Xulosa

- **EDA** – ma'lumotlarni tahlil qilishning birinchi va eng muhim bosqichi bo'lib, u datasetni chuqur tushunishga yordam beradi.
- Ma'lumotning **shakli (rows, columns)**, ustunlarning turlari va ma'nolari aniqlanadi.
- **Qisqa statistik xulosalar** (mean, median, min, max, std) yordamida umumiyligi tendensiya va qiymatlar diapazoni ko'rildi.
- **Yetishmayotgan qiymatlar** aniqlanadi va ularni to'ldirish yoki olib tashlash strategiyasi belgilanadi.
- **Vizualizatsiya** (histogram, boxplot, scatterplot, heatmap) orqali ustunlar o'rtasidagi bog'liqliklar va taqsimotlar tushuniladi.

Xulosa

- **EDA** – ma'lumotlarni tahlil qilishning birinchi va eng muhim bosqichi bo'lib, u datasetni chuqur tushunishga yordam beradi.
- Ma'lumotning **shakli (rows, columns)**, ustunlarning turlari va ma'nolari aniqlanadi.
- **Qisqa statistik xulosalar** (mean, median, min, max, std) yordamida umumiyligi tendensiya va qiymatlar diapazoni ko'rildi.
- **Yetishmayotgan qiymatlar** aniqlanadi va ularni to'ldirish yoki olib tashlash strategiyasi belgilanadi.
- **Vizualizatsiya** (histogram, boxplot, scatterplot, heatmap) orqali ustunlar o'rtasidagi bog'liqliklar va taqsimotlar tushuniladi.
- EDA natijasi keyingi bosqichlarda (**tozalash, transformatsiya, model qurish**) qaysi yo'nalishda ishlash kerakligini aniq belgilab beradi.

Savollar?

