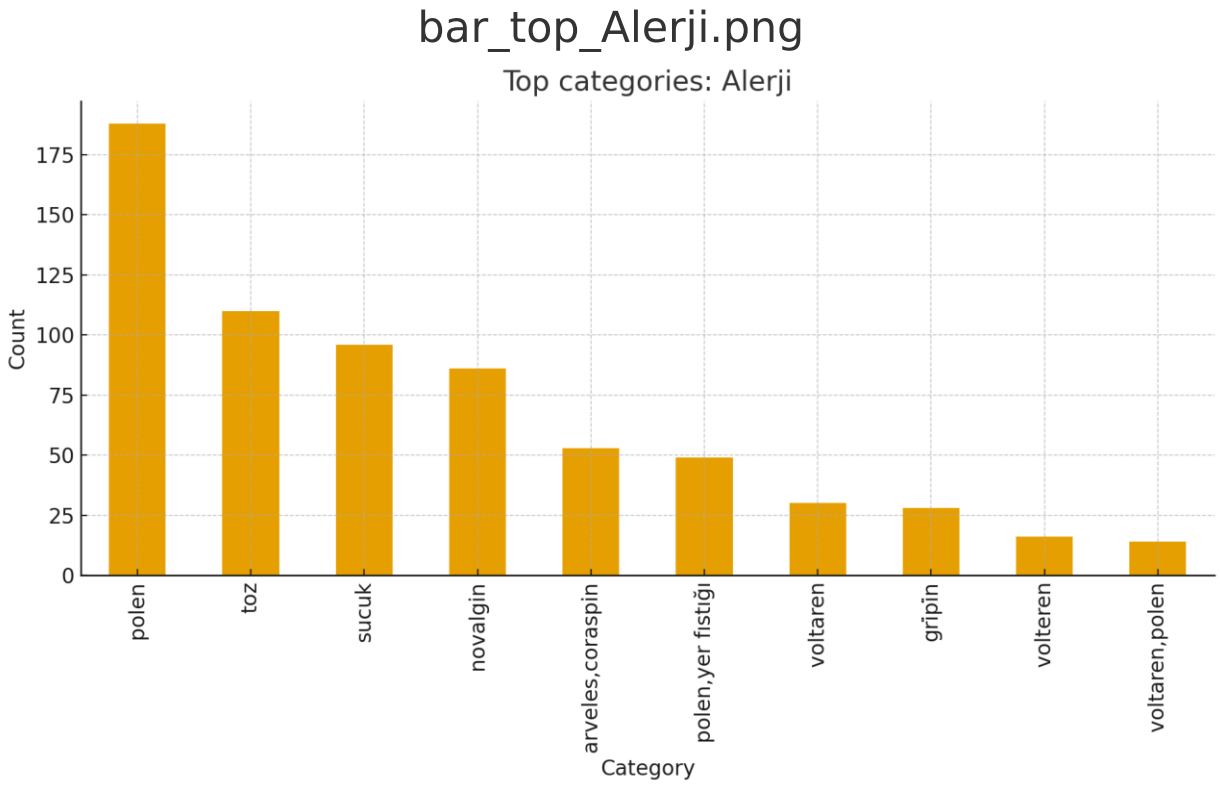
Pusula Data Science Intern Case

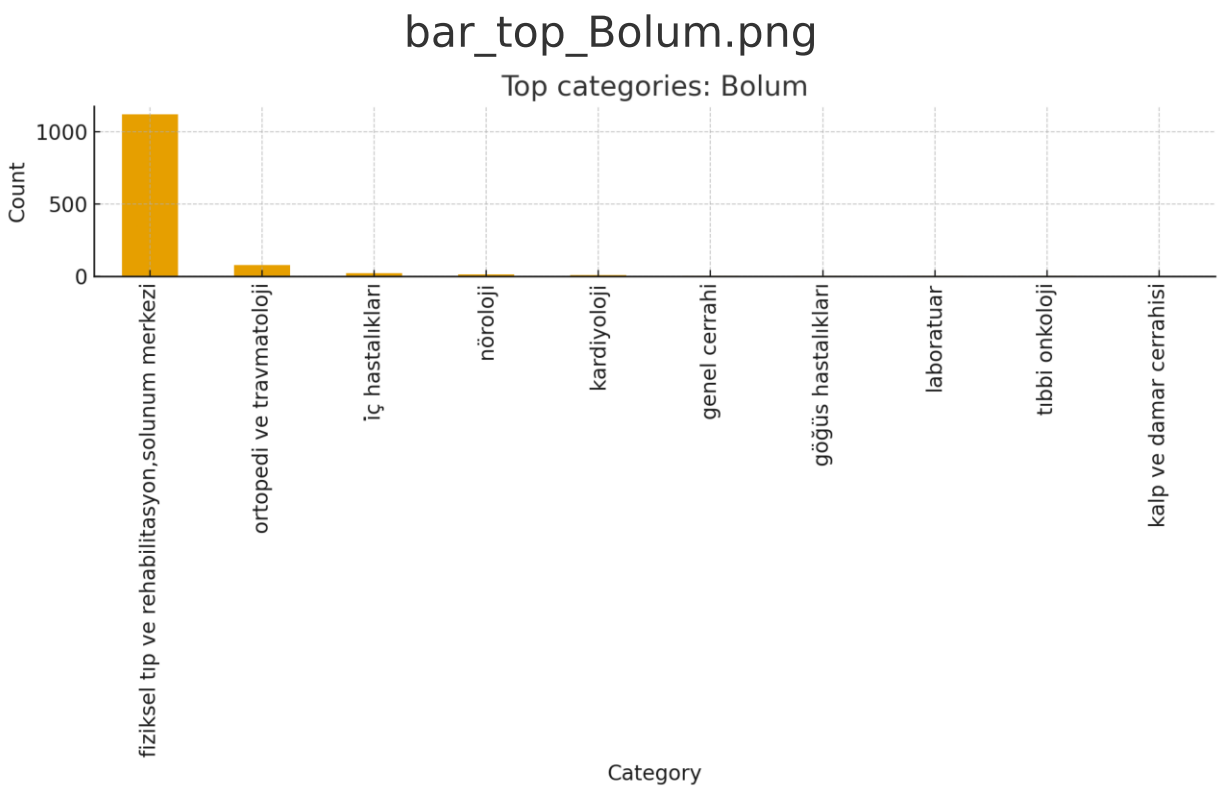
EDA & Preprocessing — Extras

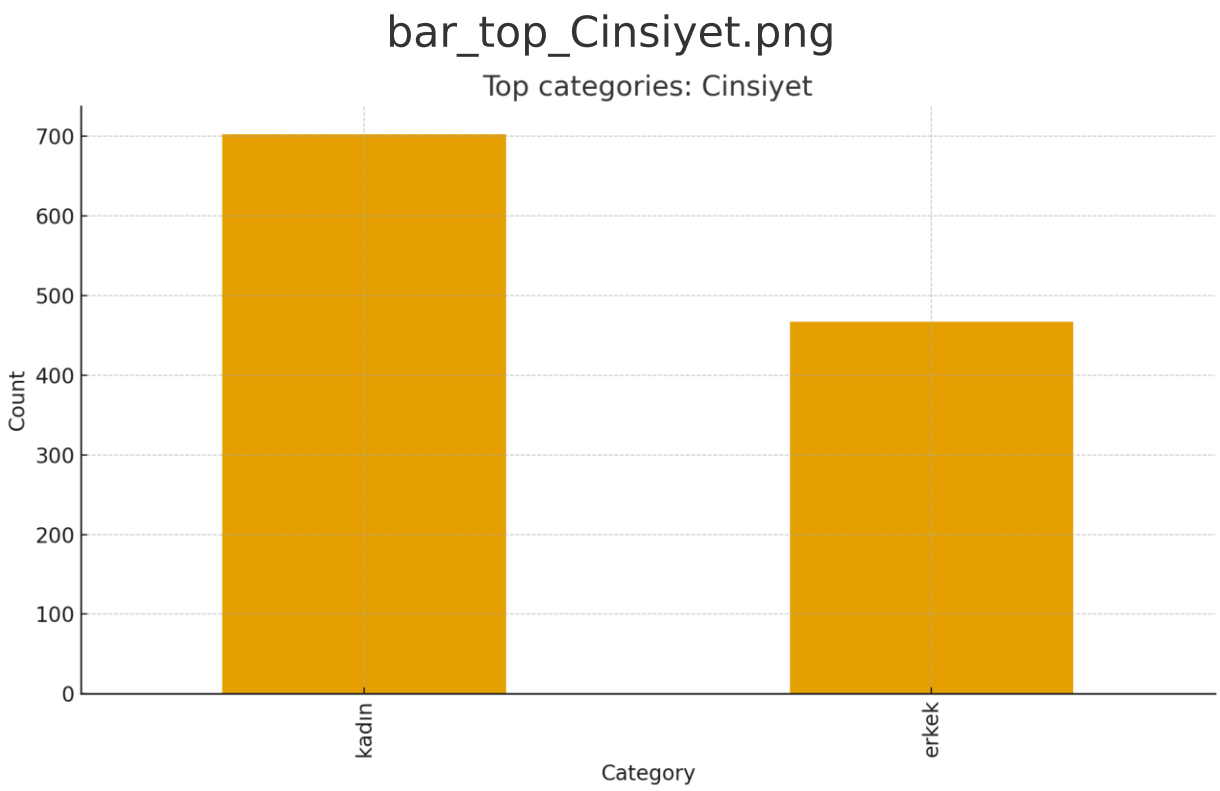
Rows: 1273 | Cols: 13

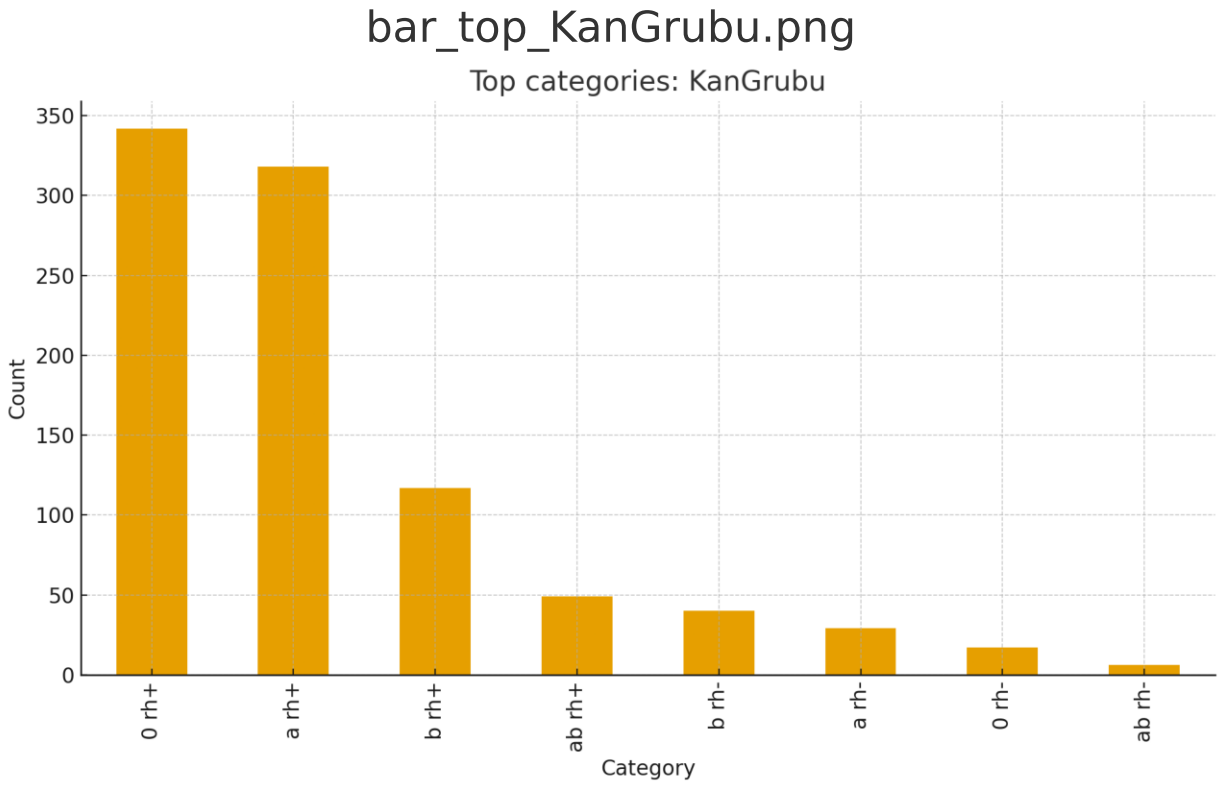
Key EDA Findings:

- Numeric columns: HastaNo, Yas
- Categorical/text columns: Cinsiyet, KanGrubu, Uyruk, KronikHastalik, Bolum, Alerji, Tanilar, TedaviAdi, TedaviSuresi, UygulamaYerleri, UygulamaSuresi
- Missing present in: Cinsiyet, KanGrubu, KronikHastalik, Bolum, Alerji, Tanilar, UygulamaYerleri
- ID `HastaNo` excluded from features to prevent leakage.
- Top-20 token binarization used for multi-label list columns.


bar_top_Alerji.png
Top categories: Alerji
Count
0
25
50
75
100
125
150
175
polen
toz
sucuk
novalgin
arveles,coraspin
polen,yer fıstığı
voltaren
gripin
volteren
voltaren,polen
Category


bar_top_Bolum.png
Top categories: Bolum
Count
1000
500
0
fiziksel tıp ve rehabilitasyon,solunum merkezi
ortopedi ve travmatoloji
iç hastalıkları
nöroloji
kardiyoloji
genel cerrahi
göğüs hastalıkları
laboratuar
tıbbi onkoloji
kalp ve damar cerrahisi
Category


bar_top_Cinsiyet.png
Top categories: Cinsiyet
Count
0
100
200
300
400
500
600
700
kadın
erkek
Category


bar_top_KanGrubu.png
Top categories: KanGrubu
Count
350
300
250
200
150
100
50
0
0 rh+
a rh+
b rh+
ab rh+
b rh-
a rh-
0 rh-
ab rh-
Category

Extras Implemented:

- Pipeline-level, modular preprocessing with ColumnTransformer.
- High-cardinality categorical handling via frequency encoding.
- Multi-label splitter (comma-separated) with Top-K binarization per field.
- Exported cleaned dataset, prepared features, target, and fitted pipeline.
- Matplotlib-only charts (one per figure) for numeric distributions and top categories.
- README & requirements for quick reproducibility.
- Ethics/Privacy: patient ID excluded from features; target-only usage for `TedaviSuresi`.