Tuba ÇALCI – tubacalci@hotmail.com

Data Science Intern Case Study

Project Overview

- **Goal:** Explore the dataset and make it **model-ready**
- **Dataset:** Physical Medicine & Rehabilitation; **2235 rows / 13 columns**

Dataset

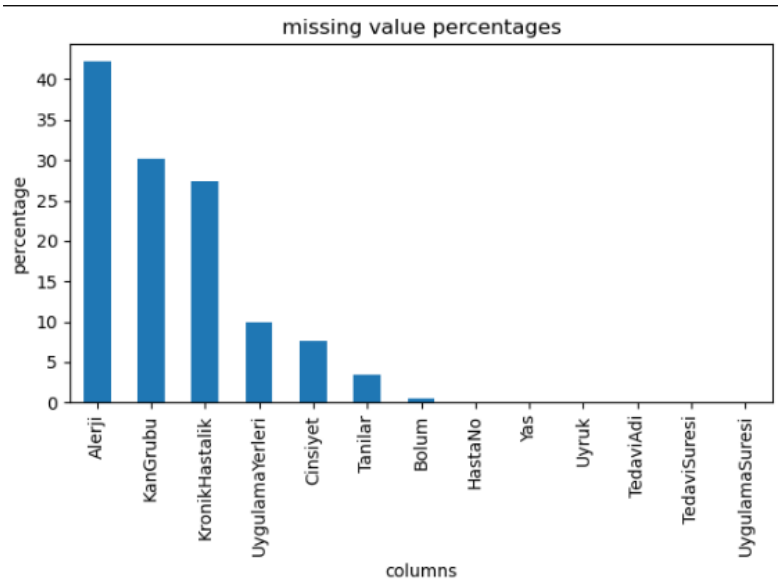
- The dataset contains **2235 rows** and **13 columns**.
- **Target variable:** TedaviSuresi (treatment duration in sessions)
- Key columns: HastaNo (ID), Yas (Age), Cinsiyet (Gender), KanGrubu (Blood Type), Uyruk (Nationality), KronikHastalik, Bolum (Department), Alerji, Tanilar, TedaviAdi, TedaviSuresi (Target), UygulamaYerleri, UygulamaSuresi.

EDA Findings

Missing Values

- Overall missingness: 6/13 columns contain missing values.
- Top missing columns (by %):
    - Alerji → 42.24%
    - KanGrubu → 30.20%
    - KronikHastalik → 27.34%

	missing	missing_%
Alerji	944	42.24
KanGrubu	675	30.20
KronikHastalik	611	27.34
UygulamaYerleri	221	9.89
Cinsiyet	169	7.56
Tanilar	75	3.36
Bolum	11	0.49
HastaNo	0	0.00
Yas	0	0.00
Uyruk	0	0.00
TedaviAdi	0	0.00
TedaviSuresi	0	0.00
UygulamaSuresi	0	0.00



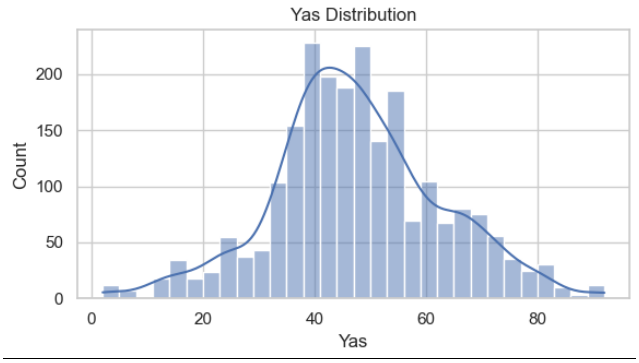
- Columns like HastaNo, Yas, Uyruk, TedaviAdi, TedaviSuresi, UygulamaSuresi are **complete** (no missing data)

Numerical Variables

Note: TedaviSuresi and UygulamaSuresi were originally stored as text with units ("15 Seans", "20 Dakika") and were converted to numeric values using regex before analysis.

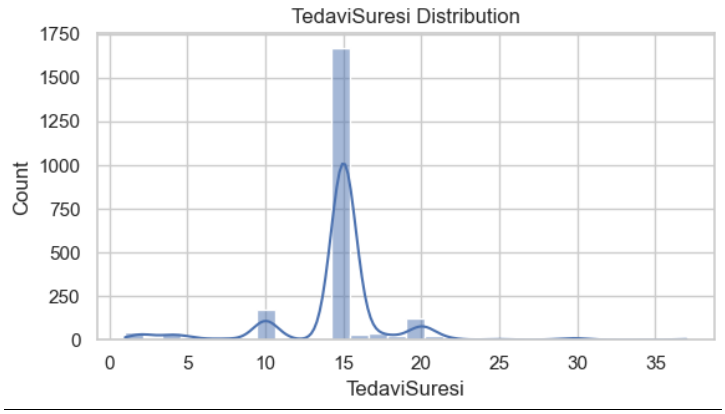
Age (Yas)

- Count: 2235
- Mean / Std: 47.33 / 15.21
- Min / Max: 2 / 92
- Quartiles (25 / 50 / 75): 38 / 46 / 56
- Shape: approximately bell-shaped and centered on middle age.



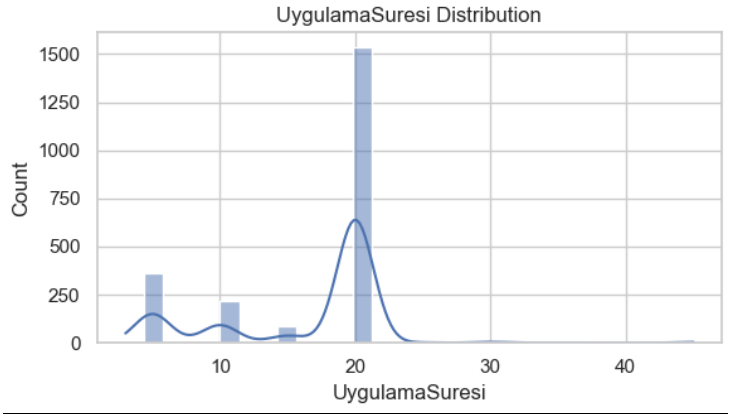
Treatment Duration (TedaviSuresi)

- Count: 2235
- Mean / Std: 14.58 / 3.73
- Min / Max: 1 / 37
- Quartiles (25 / 50 / 75): 15 / 15 / 15
- Insight: extremely strong mode at 15 sessions, indicating standardized treatment protocols.



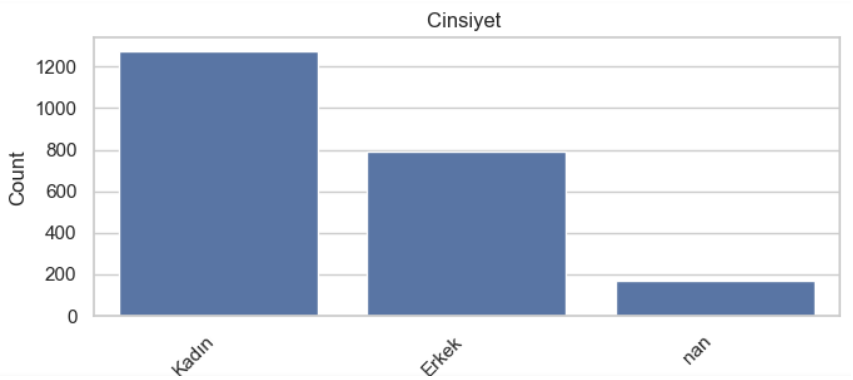
Application Duration (UygulamaSuresi)

- Count: 2235
- Mean / Std: 16.57 / 6.29
- Min / Max: 1 / 45
- Quartiles (25 / 50 / 75): 10 / 20 / 20
- Insight: pronounced peak around 20 with a long left tail toward smaller values and a few larger outliers.

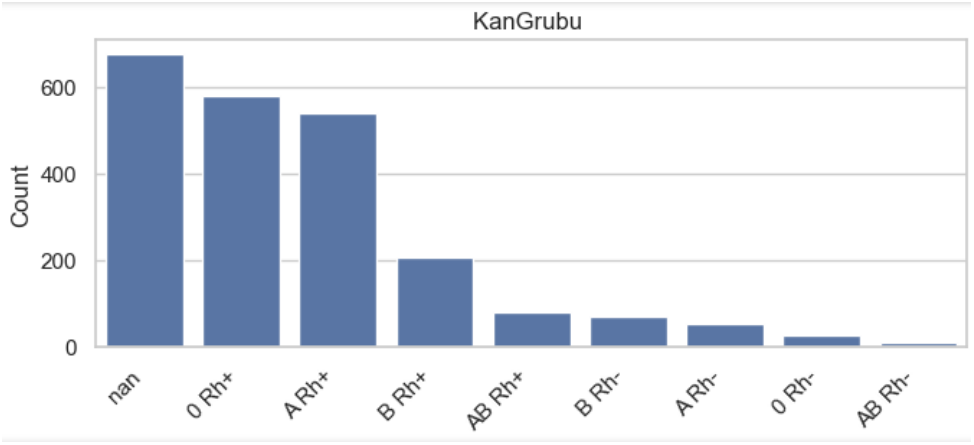


Categorical Variables

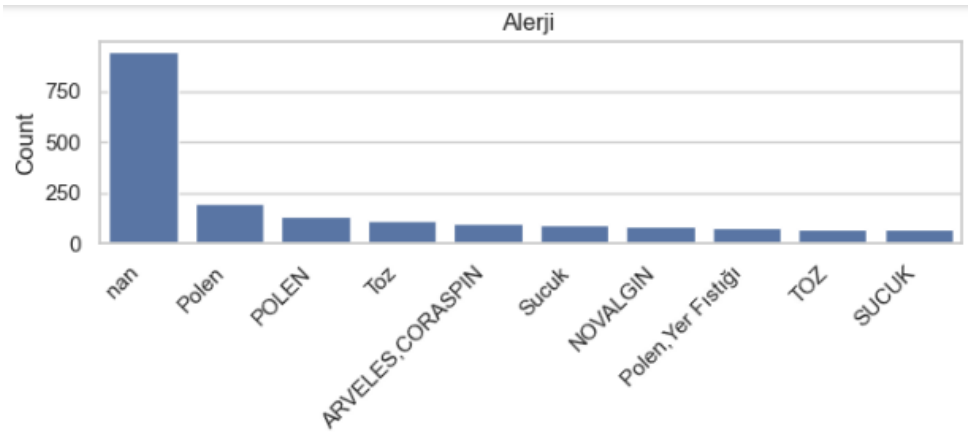
- **Gender (Cinsiyet)**
  - Female: **1274 (57%)**
  - Male: **792 (35%)**
  - Missing: **169 (7.6%)**



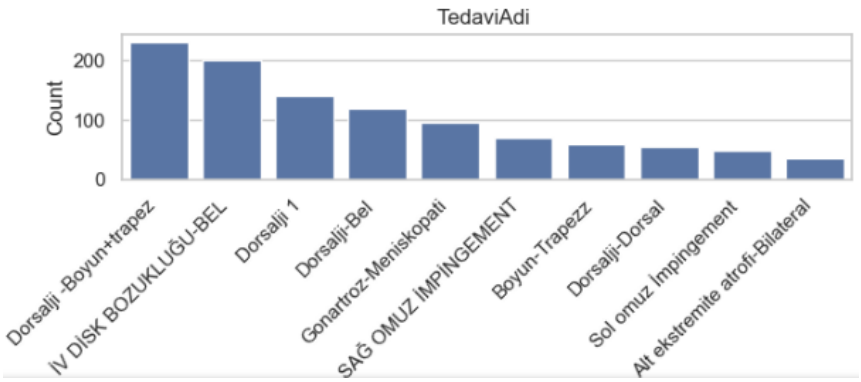
- **Blood Type (KanGrubu)**
    - Top groups: 0 Rh+ (579), A Rh+ (540), B Rh+ (206)
    - Rare groups: AB Rh- (8), 0 Rh- (26)
    - Missing: **675 (30.2%)**



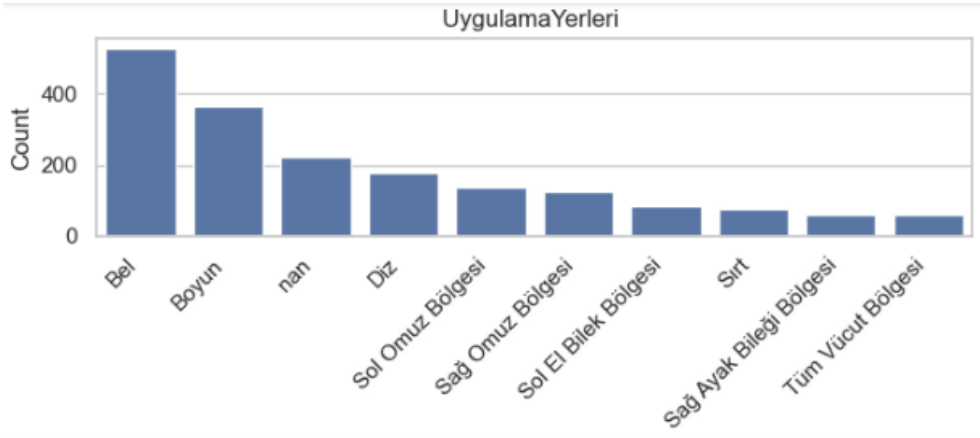
- **Nationality (Uyruk)**
    - Dominated by **Türkiye (2173, ~97%)**
    - Minor: Tokelau (27), Albania (13), Azerbaijan (12), Libya (10)
- **Chronic Conditions (KronikHastalik)**
    - Missing: **611 (27.3%)**
    - Frequent: Myasthenia Gravis (38), Arrhythmia (36), Facioscapulohumeral Dystrophy (36), Asthma/Endocrine disorders (30–34 cases each)
- **Department (Bolum)**
    - Mostly **Physical Medicine & Rehabilitation (2045, ~91%)**
    - Others: Orthopedics (88), Internal Medicine (32), Neurology (17)
    - Missing: 11
- **Allergies (Alerji)**
    - Missing: **944 (42.2%)**
    - Frequent entries: Pollen (198 + 134), Dust (119 + 74), “Sucuk” (91 + 73), ARVELES/CORASPIN (102), Novalgin (90)



- **Diagnoses (Tanilar)**
    - Top: “Dorsalji, Lumbo-sacral region” (149), “Shoulder trauma syndrome” (128), “Intervertebral disk disorders” (116)
    - Missing: 75 (~3.4%)
- **Treatment Name (TedaviAdi)**
    - Top: Dorsalji-Boyun+Trapez (231), Lumbar disk disorder (200), Dorsalji-1 (140), Dorsalji-Bel (120)
    - Other frequent: Gonarthrosis-Meniscus (95), Shoulder Impingement (70), etc.

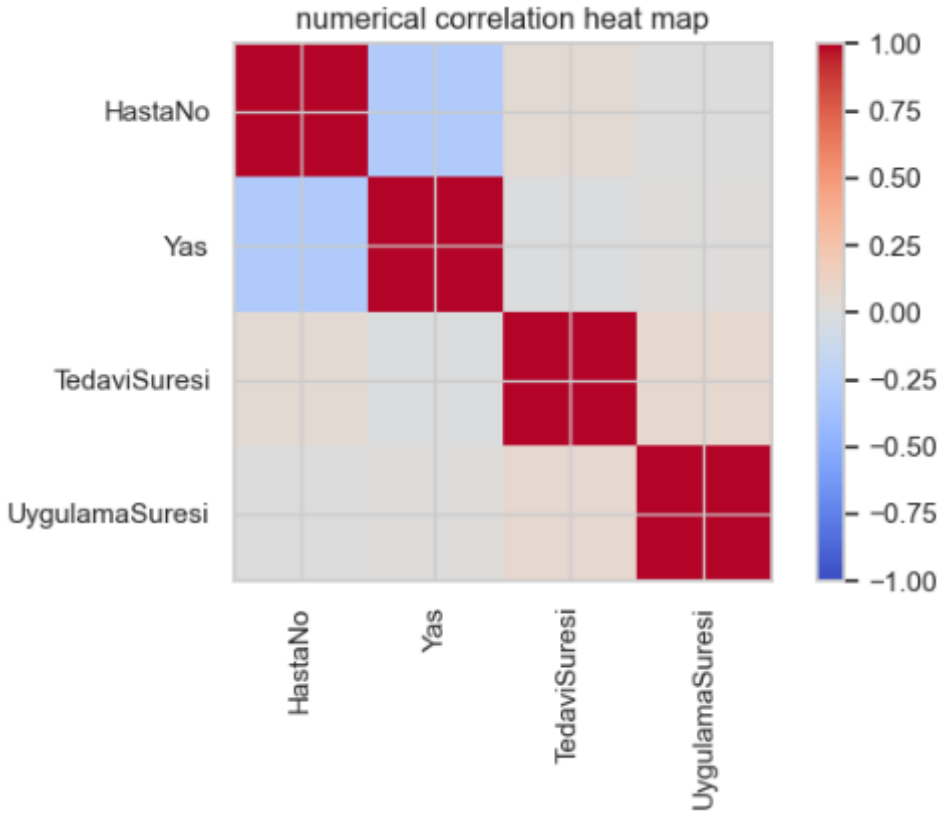


- **Application Site (UygulamaYerleri)**
    - Top: Lumbar (Bel) (528), Neck (Boyun) (363), Knee (177), Shoulder (137–127)
    - Missing: **221 (9.9%)**

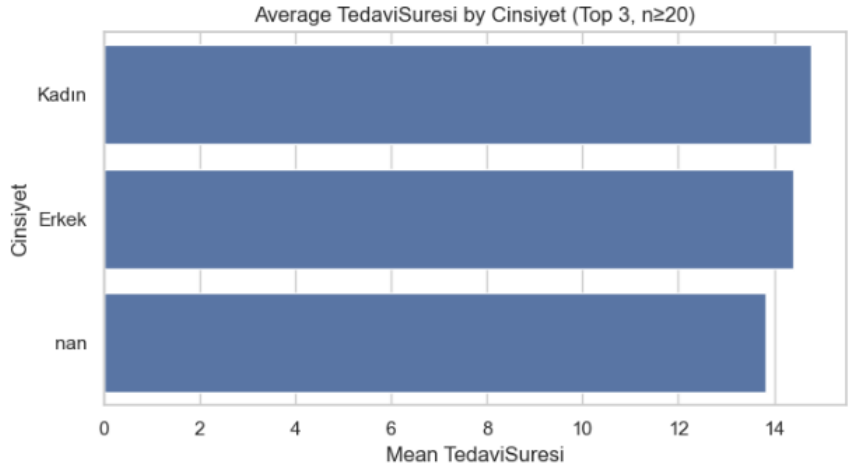


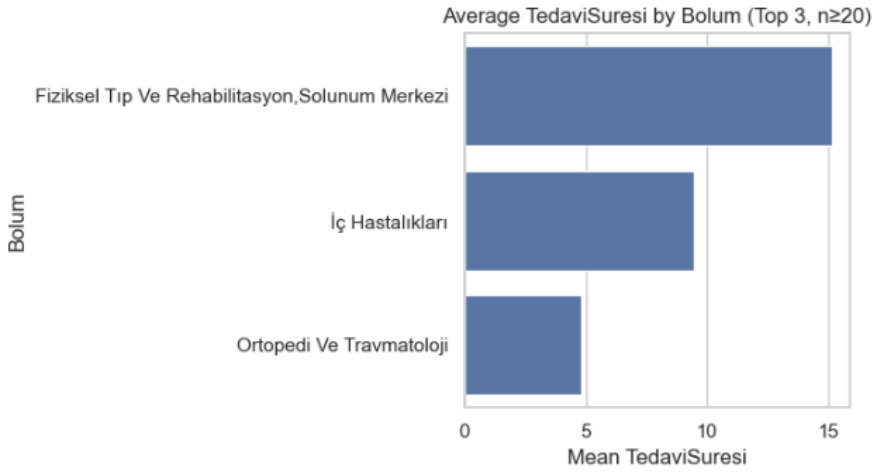
Relationships with Target (TedaviSuresi)

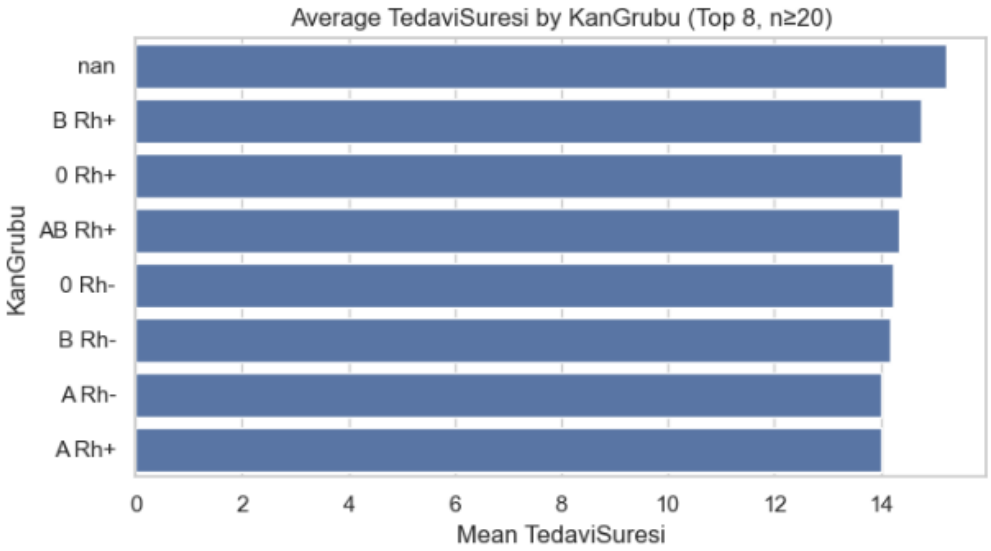
- **Numerical correlations**
  - Correlation with Yas: very weak
  - Correlation with UygulamaSuresi: very weak
  - No strong linear relationships detected in numerical variables.



- **Categorical comparisons**
    - Average TedaviSuresi by **Gender, Department, Blood Type** shows only slight variations.
    - No categorical group stands out with a significantly longer or shorter treatment duration.







Key Insights (EDA)

- **Standardization**:
    - TedaviSuresi heavily concentrated at 15 sessions → reflects a fixed rehabilitation protocol.
    - UygulamaSuresi clustered around 20 minutes, also standardized.
- **Age**:
    - Centered around middle age (mean ≈ 47), most patients between 30–60.
- **Missing data**:
    - High in Alerji (42%), KanGrubu (30%), KronikHastalik (27%).
    - These variables need imputation or careful handling before modeling.
- **Numerical correlations**:
    - Very weak between TedaviSuresi and other numeric variables (Yas, UygulamaSuresi).
    - No strong linear relationship observed.
- **Categorical variables**:
    - Gender, Department, and Blood Type show only slight variations in average treatment duration.
    - No categorical group strongly determines treatment length.

Data Pre-Processing

Data splitting:

- Rows with missing TedaviSuresi were removed.
- The dataset was split into training (80%) and testing (20%) subsets.

Column types:

- Numerical features (Yas, UygulamaSuresi, etc.) and categorical features (Cinsiyet, KanGrubu, Bolum, etc.) were identified separately.
- Categorical variables were cast to object type for consistency.

Missing value handling:

- Numerical variables were imputed using the median.
- Categorical variables were imputed using the mode (most frequent value), and "Unknown" was assigned when no mode was available.

Rare category handling:

- Categories with very low frequency were grouped into an "Other" category.

Encoding:

- All categorical variables were transformed using One-Hot Encoding.
- Training and testing sets were aligned to have exactly the same columns.

Scaling:

- Numerical variables were standardized using StandardScaler (mean = 0, std = 1).

Outcome:

- Final training set: 1,788 rows, 88 features
- Final testing set: 447 rows, 88 features
- No missing values remained; datasets are fully consistent and model-ready.

Validation:

- A baseline RandomForest model achieved an R^2 score of ≈ 0.89, confirming that preprocessing was successful.