2025

Data Science Intern Case Study Report

AUTHOR
ŞEYMA NUR CİN

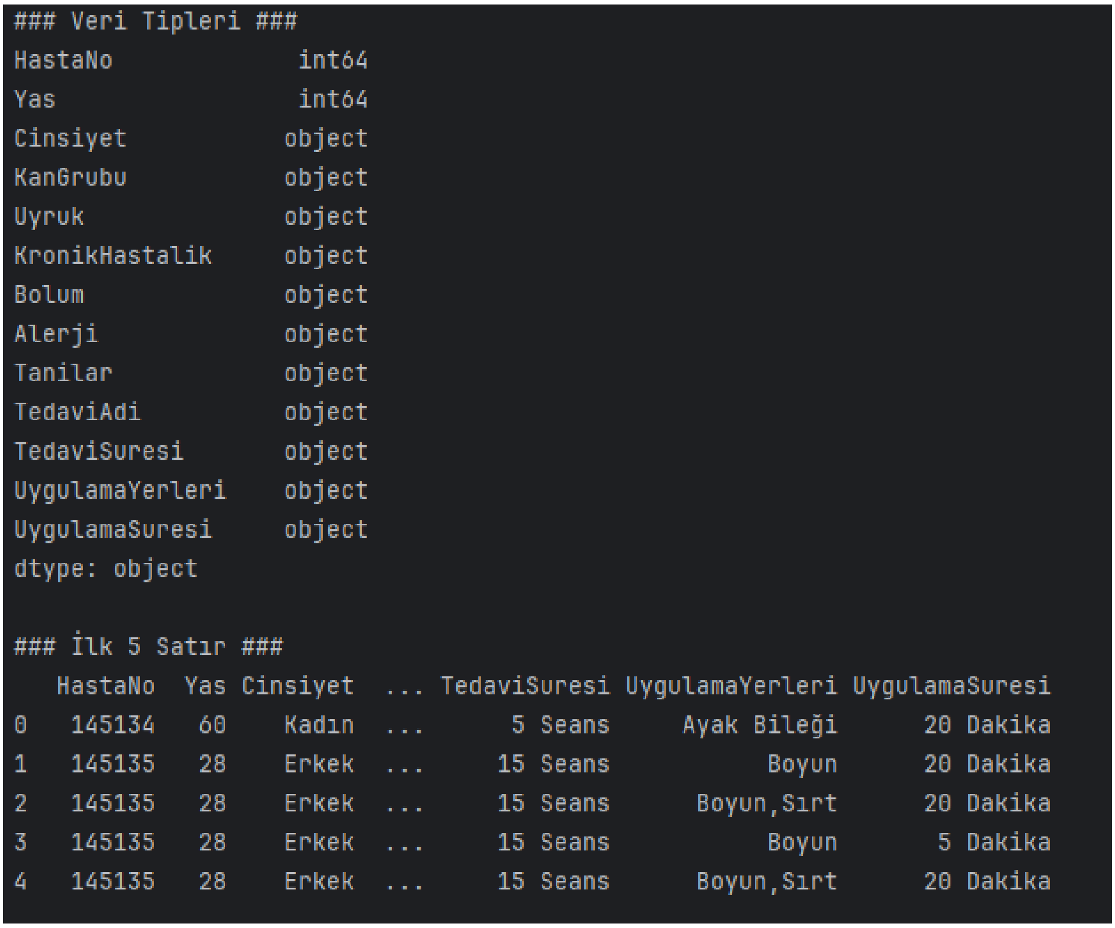
e-mail: cinseymanur@gmail.com

Author:Şeyma Nur Cin

Data Science Intern Case Study Report

Dataset Loading

Load the dataset and check the first few rows and data types:



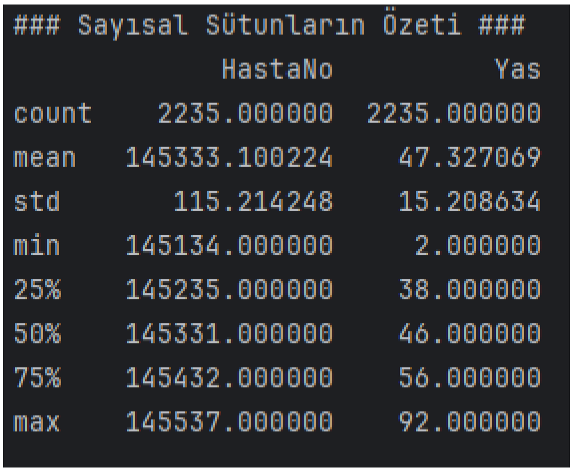
```
### Veri Tipleri ###
HastaNo              int64
Yas                  int64
Cinsiyet            object
KanGrubu            object
Uyruk               object
KronikHastalik      object
Bolum               object
Alerji              object
Tanilar             object
TedaviAdi           object
TedaviSuresi        object
UygulamaYerleri     object
UygulamaSuresi      object
dtype: object

### İlk 5 Satır ###
   HastaNo  Yas  Cinsiyet  ... TedaviSuresi UygulamaYerleri UygulamaSuresi
0   145134   60     Kadın  ...      5 Seans     Ayak Bileği      20 Dakika
1   145135   28     Erkek  ...     15 Seans           Boyun      20 Dakika
2   145135   28     Erkek  ...     15 Seans      Boyun,Sırt      20 Dakika
3   145135   28     Erkek  ...     15 Seans           Boyun       5 Dakika
4   145135   28     Erkek  ...     15 Seans      Boyun,Sırt      20 Dakika
```

Numerical and Categorical Columns Overview

Summary statistics for numerical columns:

print(df.describe())



```
### Sayısal Sütunların Özeti ###
             HastaNo          Yas
count    2235.000000  2235.000000
mean   145333.100224    47.327069
std       115.214248    15.208634
min    145134.000000     2.000000
25%    145235.000000    38.000000
50%    145331.000000    46.000000
75%    145432.000000    56.000000
max    145537.000000    92.000000
```

Author:Şeyma Nur Cin

Categorical Columns Analysis

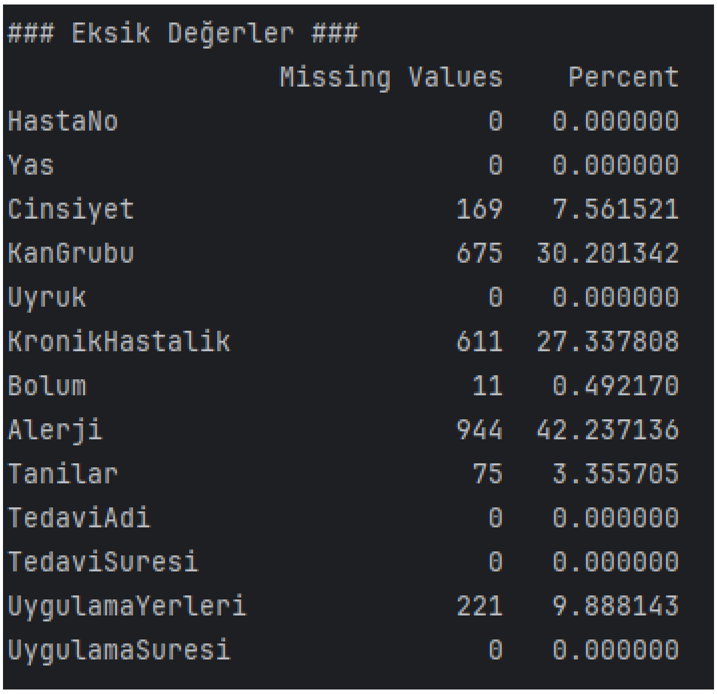
Frequency counts for categorical columns:

for col in df.select_dtypes(include=['object']).columns:

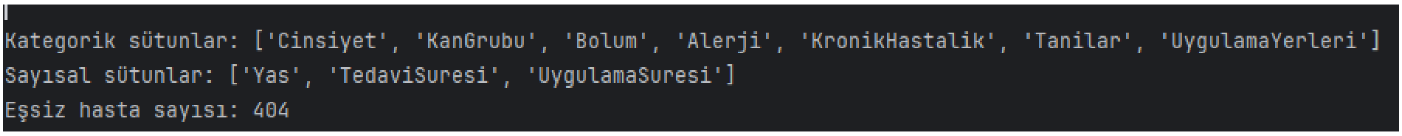
print(df[col].value_counts(dropna=False))

Tables of counts and percentages can be included in the report.

In this step, the categorical columns in the dataset (Gender, Blood Type, Nationality, Chronic Conditions, Department, Allergies, Diagnoses, Treatment Name, Application Sites, Treatment Duration, and Application Duration) were summarized. The code calculated the unique values and their frequencies for each column and identified missing values. Additionally, the Treatment Duration and Application Duration columns were converted to numeric format, and their distributions were visualized using histograms. This allowed a quick overview of general distributions, frequently occurring categories, and missing data in the dataset.



```
### Eksik Değerler ###
                 Missing Values    Percent
HastaNo                       0   0.000000
Yas                           0   0.000000
Cinsiyet                    169   7.561521
KanGrubu                    675  30.201342
Uyruk                         0   0.000000
KronikHastalik              611  27.337808
Bolum                        11   0.492170
Alerji                      944  42.237136
Tanilar                      75   3.355705
TedaviAdi                     0   0.000000
TedaviSuresi                  0   0.000000
UygulamaYerleri             221   9.888143
UygulamaSuresi                0   0.000000
```



```
Kategorik sütunlar: ['Cinsiyet', 'KanGrubu', 'Bolum', 'Alerji', 'KronikHastalik', 'Tanilar', 'UygulamaYerleri']
Sayısal sütunlar: ['Yas', 'TedaviSuresi', 'UygulamaSuresi']
Eşsiz hasta sayısı: 404
```

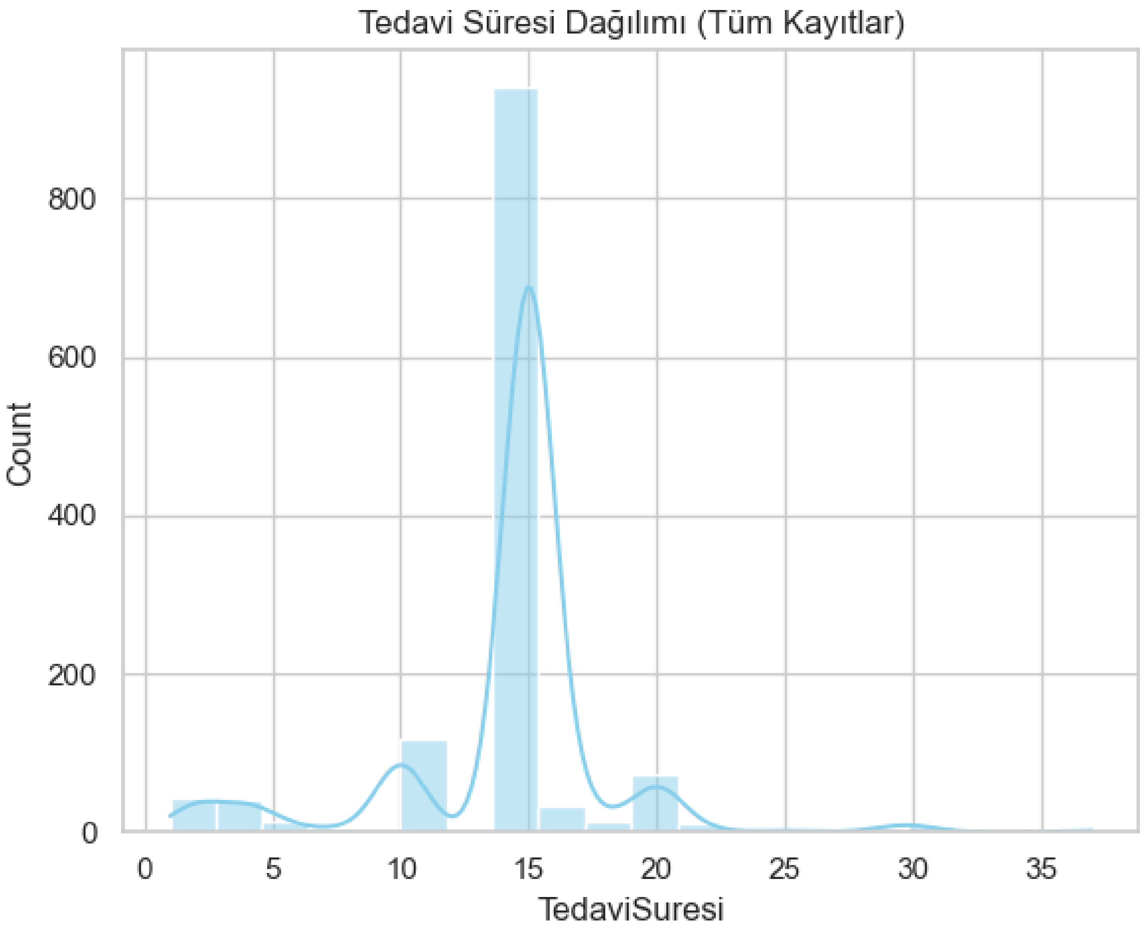
The categorical and numerical columns in the dataset are listed as follows:

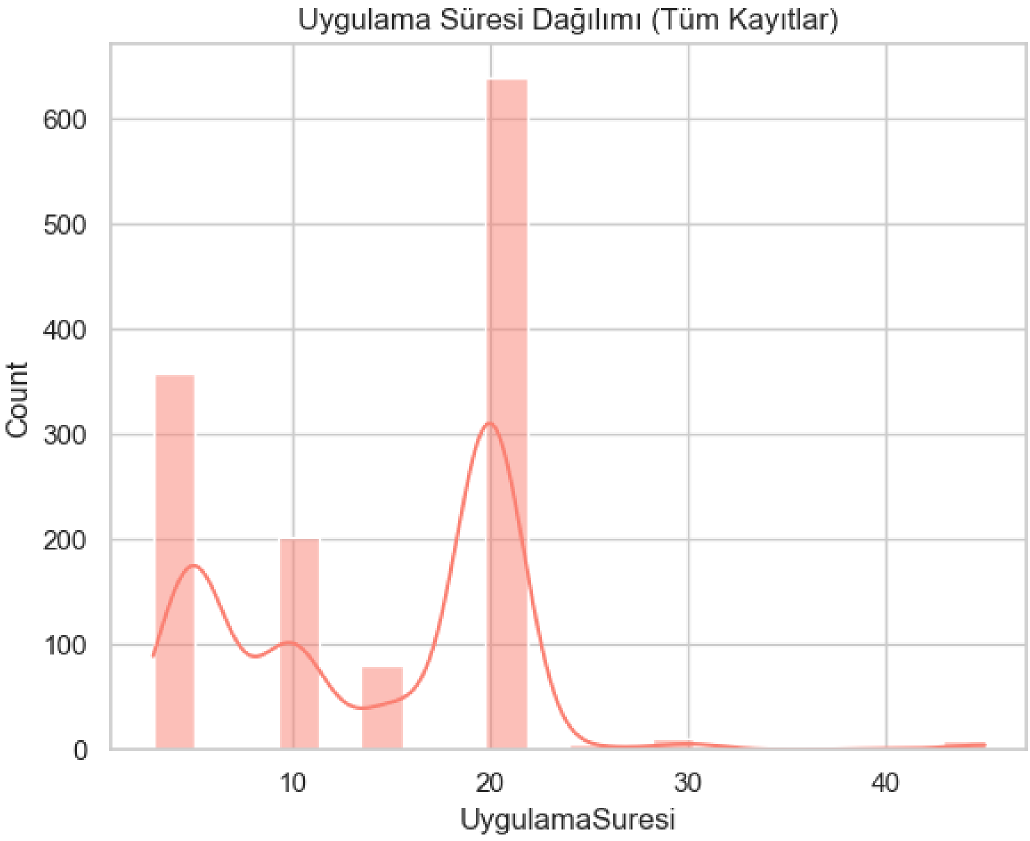
Categorical Columns: 'Gender', 'BloodType', 'Department', 'Allergies', 'ChronicConditions', 'Diagnoses', 'ApplicationSites'

Numerical Columns: 'Age', 'TreatmentDuration', 'ApplicationDuration'

Additionally, the dataset contains 404 unique patient records.

Author:Şeyma Nur Cin





Author:Şeyma Nur Cin

There are histograms showing the distributions of the variables 'ApplicationDuration' and 'TreatmentDuration'.

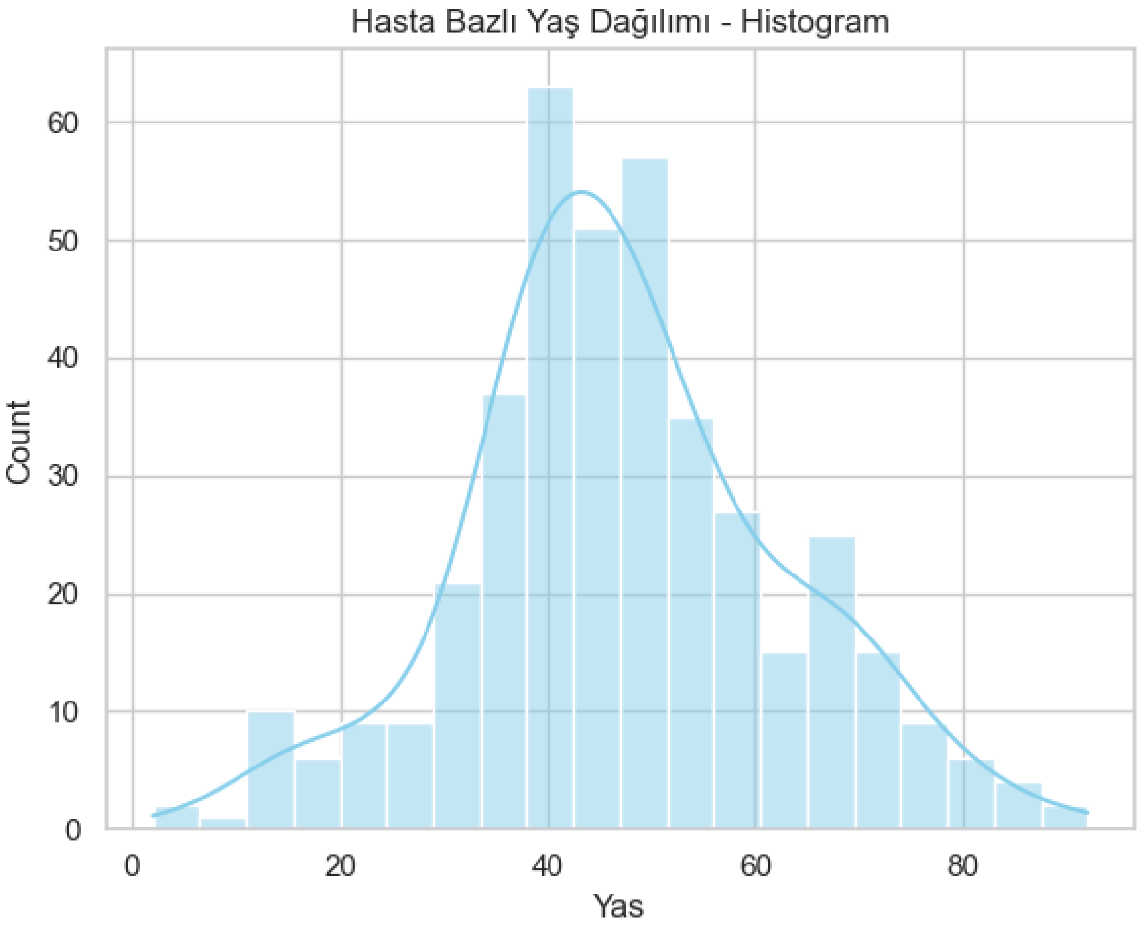
Application Duration Distribution
This graph shows the distribution of application duration within the dataset. It can be observed that application duration is mostly concentrated around 20, with noticeable peaks also around 5 and 12. This distribution indicates that patients have varying application durations.

Treatment Duration Distribution
This graph visualizes the distribution of treatment duration. According to the graph, the most common treatment duration is around 15. Additionally, there are smaller peaks near 12 and 20. This suggests that while most treatments last about 15 sessions, some treatments can be shorter or longer.

These graphs provide important insights into patients' application and treatment durations in the dataset. Analyzing this information can help in better understanding and optimizing healthcare services.

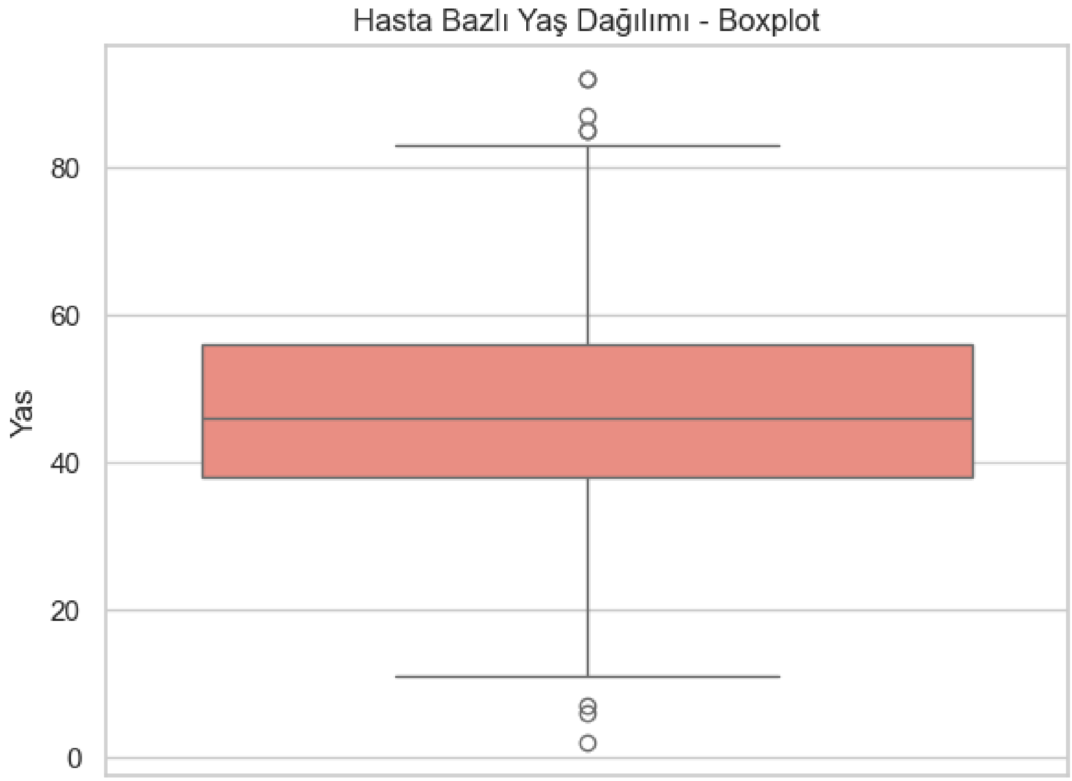
Patient Age Distribution

The patient age distribution is visualized using both a histogram and a boxplot.



Author:Şeyma Nur Cin

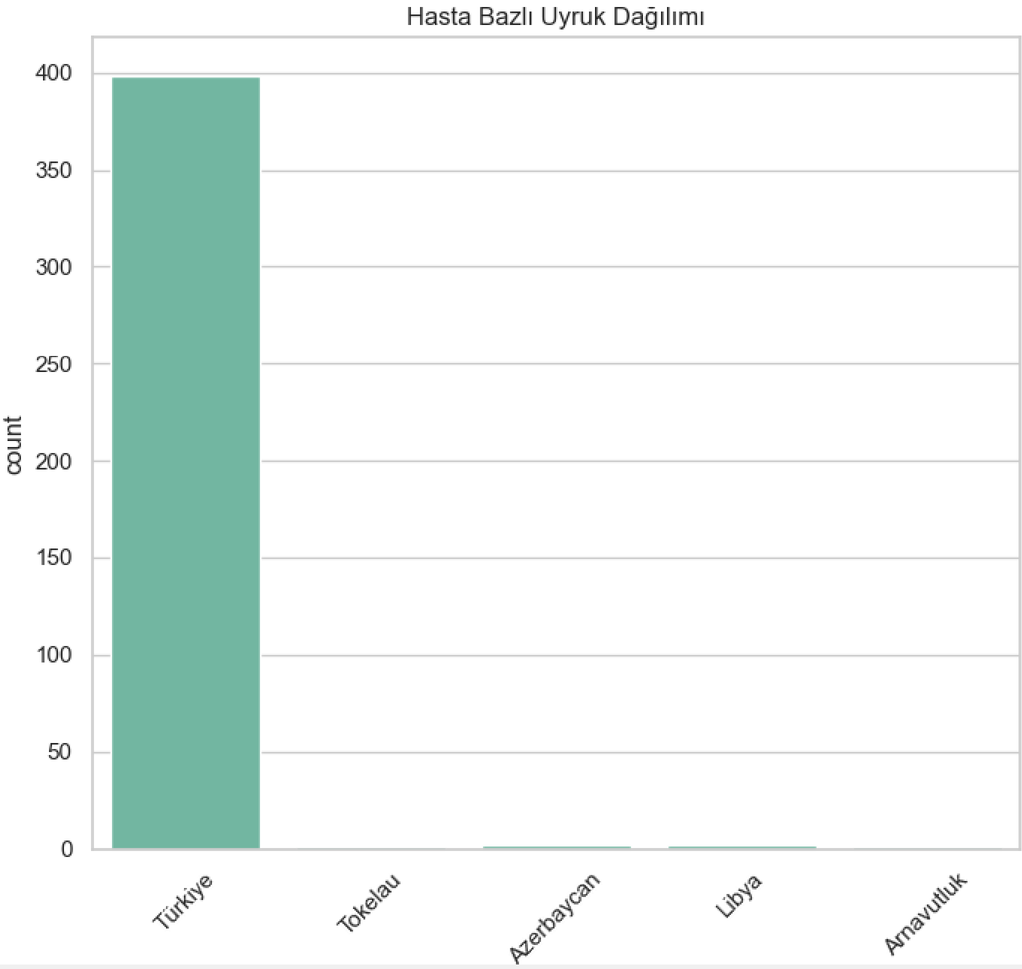
Histogram: The "Patient-Based Age Distribution - Histogram" graph shows the distribution of patients' ages. The distribution generally follows a bell-shaped curve, with a concentration in the 40 to 50 age range, which contains the highest number of patients. As ages increase or decrease from this range, the number of patients decreases. There are no missing values in the age distribution within the dataset.



Boxplot: The "Patient-Based Age Distribution - Boxplot" graph provides summary statistics of the age distribution. The median (Q2) is around 48–50 years. The first quartile (Q1) is approximately 38–40 years, and the third quartile (Q3) is around 55–58 years. This confirms that patients are mostly concentrated in the middle-age group. The boxplot also shows some outliers, including patients over 90 years old and under 10 years old, indicating that there are very few patients in these age groups.

Author:Şeyma Nur Cin

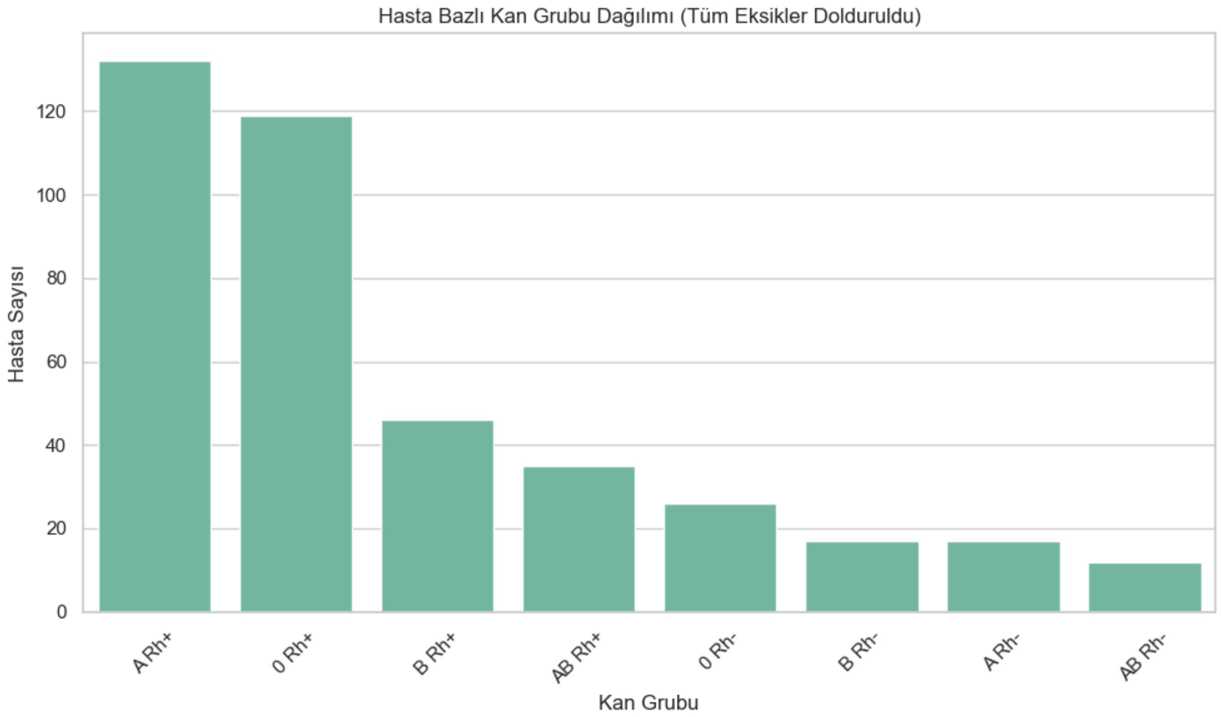
Patient Nationality Distribution



The Patient Nationality Distribution graph shows the distribution of patients in the dataset according to their nationalities. The most notable observation is that patients with Turkish nationality make up the overwhelming majority. Other nationalities (Tokelau, Azerbaijan, Libya, Albania) are represented in much smaller numbers. This indicates that the dataset mainly consists of patients from Turkey.

Author:Şeyma Nur Cin

Patient Blood Type Distribution



The "Patient-Based Blood Type Distribution" graph shows the distribution of patients according to their blood types. According to the graph, the most common blood types are:

A Rh+ (approximately 130 patients)

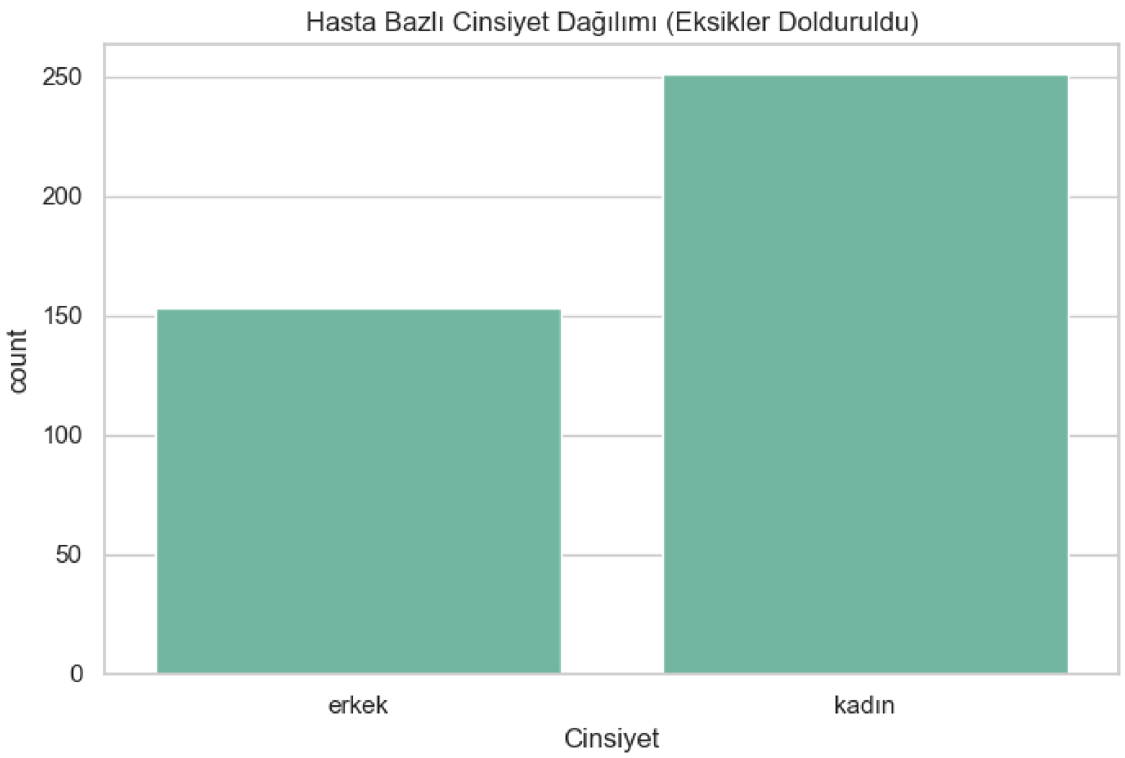
0 Rh+ (approximately 120 patients)

B Rh+ (approximately 45 patients)

Other blood types (AB Rh+, 0 Rh-, B Rh-, A Rh-, AB Rh-) are represented by fewer patients. In this graph, missing blood type data has been imputed. Some missing values were filled using information from the same patient, while the remaining missing values were imputed using the KNN Imputer method. After imputation, these values were displayed using the actual blood type values.

Author:Şeyma Nur Cin

Patient Gender Distribution



The "Patient-Based Gender Distribution" graph visualizes the distribution of patients by gender. The graph shows that females are more numerous than males. The number of female patients is approximately 250, while the number of male patients is around 150. There were missing values in the gender column, which were filled using the most common value, and then the categorical data was processed and encoded.

Chronic Disease Processing and Patient-Based Analysis

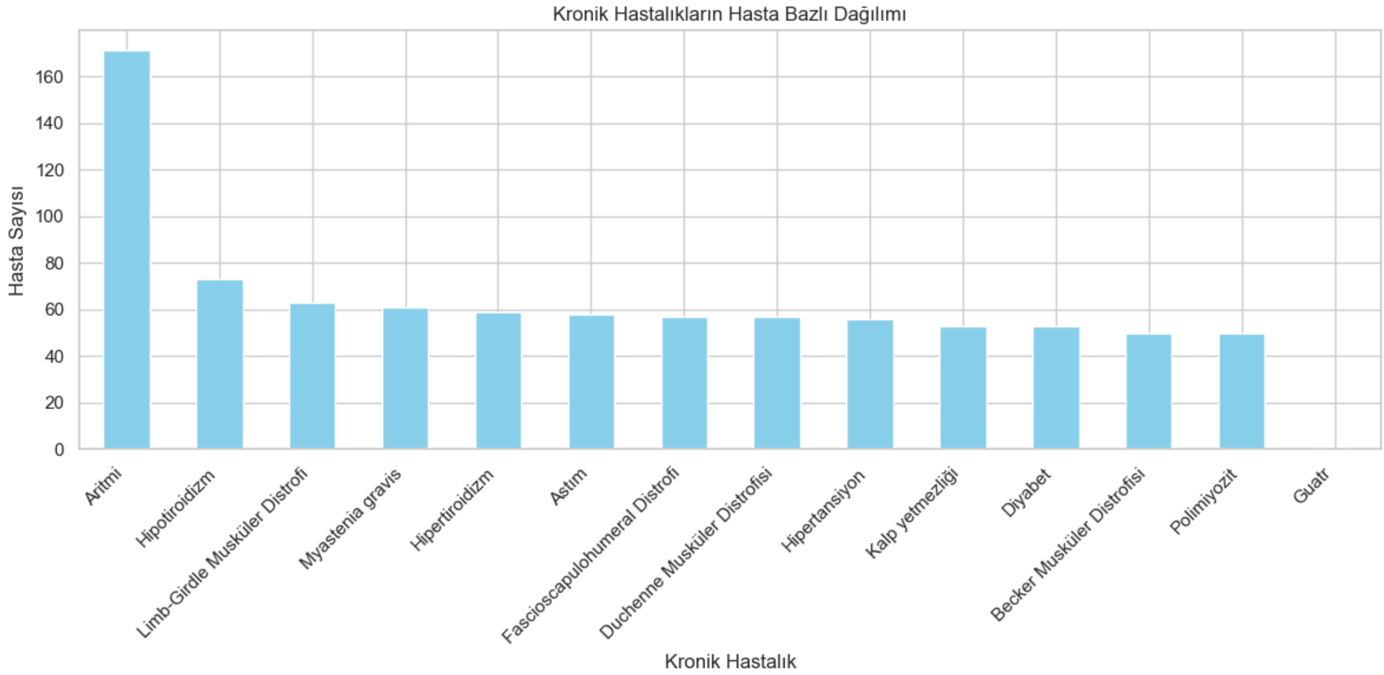
In this step, the dataset's "Chronic Diseases" (KronikHastalik) column was processed and analyzed at the patient level. The main steps were:

1. **Data Cleaning:**
   - Standardized disease names by correcting common typos (e.g., 'hiportiroidizm' → 'Hipotiroidizm').
   - Filled missing values using the most frequent chronic disease observed in the same department (Bolum). If no common value existed, a placeholder 'Unknown' was used.
2. **Data Transformation:**
   - Converted the string of diseases into a list of individual diseases for each record.
   - Applied one-hot encoding to create binary columns for each unique chronic disease.
3. **Patient-Based Aggregation:**

- For each patient (HastaNo), unique diseases were counted only once, ensuring that repeated entries for the same patient did not inflate counts.
- Computed the number of patients affected by each chronic disease.

4. **Visualization:**
   - Created a bar plot showing the patient-based distribution of chronic diseases.
   - Exploded the disease lists so that each chronic disease appeared on a separate row, allowing visualization of age distribution by disease.
   - Used a boxplot to show the distribution of patient ages across different chronic diseases.

Purpose: These steps allow for a clean, standardized view of chronic disease prevalence and its relationship with patient age, providing insights into which diseases are most common and which age groups are affected.



Distribution of Chronic Diseases by Patient Count

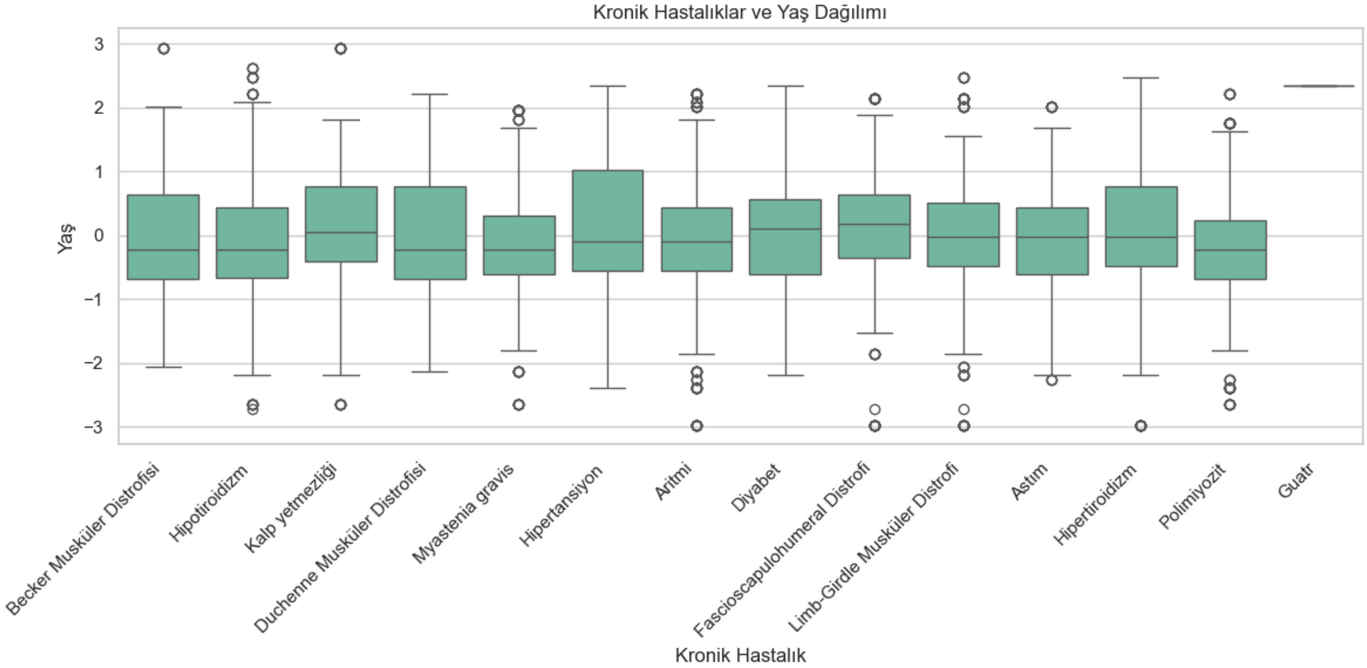
The "Patient-Based Distribution of Chronic Diseases" bar chart shows the number of patients for each chronic disease. According to this bar chart:

- The most common chronic disease is **Arrhythmia**, observed in approximately 170 patients.
- The second most frequent is **Hypothyroidism**, seen in around 75 patients.
- Following these, **Limb-Girdle Muscular Dystrophy, Myasthenia Gravis, Hyperthyroidism, Asthma, Fascioscapulohumeral Dystrophy, Hypertension, Heart Failure, Diabetes,** and **Becker Muscular Dystrophy** appear in roughly 50–60 patients each, with similar counts.

Poliomyelitis and **Goiter** are observed in fewer patients.

Author:Şeyma Nur Cin

Analysis of Age Distribution for Chronic Diseases



This chart compares the age distributions of patients with different chronic diseases in the dataset. The X-axis represents various chronic diseases, while the Y-axis shows standardized age values. The boxplot drawn for each chronic disease provides information about the average age, age range, and potential outliers for individuals with that disease.

Key observations from the chart are:

- **Overall Age Concentration:** In most chronic disease groups, the age distribution shows a similar trend. The median age (the dark line within the box) is generally concentrated between -0.5 and +1 in standardized age values, indicating that middle-aged individuals dominate the dataset.
- **Outliers:** Some diseases, particularly Becker Muscular Dystrophy, Hypothyroidism, Diabetes, Hypertension, Polymyositis, and Goiter, show noticeable outliers (points indicated by circles) outside the general age distribution. This indicates that there are also younger or older individuals within the population affected by these diseases. For example, Becker Muscular Dystrophy shows relatively young outliers, whereas Hyperthyroidism may have older outliers.
- **Age Effect by Disease:** For some diseases, age distributions are similar, while for others (e.g., Arrhythmia and Hypertension), there is a wider age range. This suggests that these diseases may affect a broader age group.
- **Muscular Dystrophies:** The age distributions for muscular dystrophy-related diseases such as Becker Muscular Dystrophy, Duchenne Muscular Dystrophy, and Fascioscapulohumeral Dystrophy are similar. This may imply that these types of diseases tend to affect comparable age groups.

Author:Şeyma Nur Cin

Department-Based Patient Distribution and Treatment Duration Analysis

In this phase, a data processing and analysis pipeline was applied to the Department column in the dataset. The procedure included the following steps:

Filling Missing Values:

The fill_department_pipeline function was used to fill missing Department entries based on the patient's TedaviAdi (Treatment Name). The main portion of the treatment name (e.g., the part before the "+" sign) was considered, and the departments of other patients with the same treatment were examined. The most frequently occurring department was assigned to the missing value. If no suitable department was found, the value “Unknown” was used.

One-Hot Encoding:

The filled Department column was transformed into a numerical format using the one_hot_department function via one-hot encoding. Each department was represented as a separate binary column.

Patient-Based Department Distribution:

For each patient, the visited departments were listed uniquely, and the number of distinct patients per department was calculated. This allowed visualization of the patient distribution across departments.

Average Treatment Duration by Department:

The average TedaviSuresi (Treatment Duration) was computed for each department and displayed using a bar chart, showing which departments generally have longer or shorter treatments.

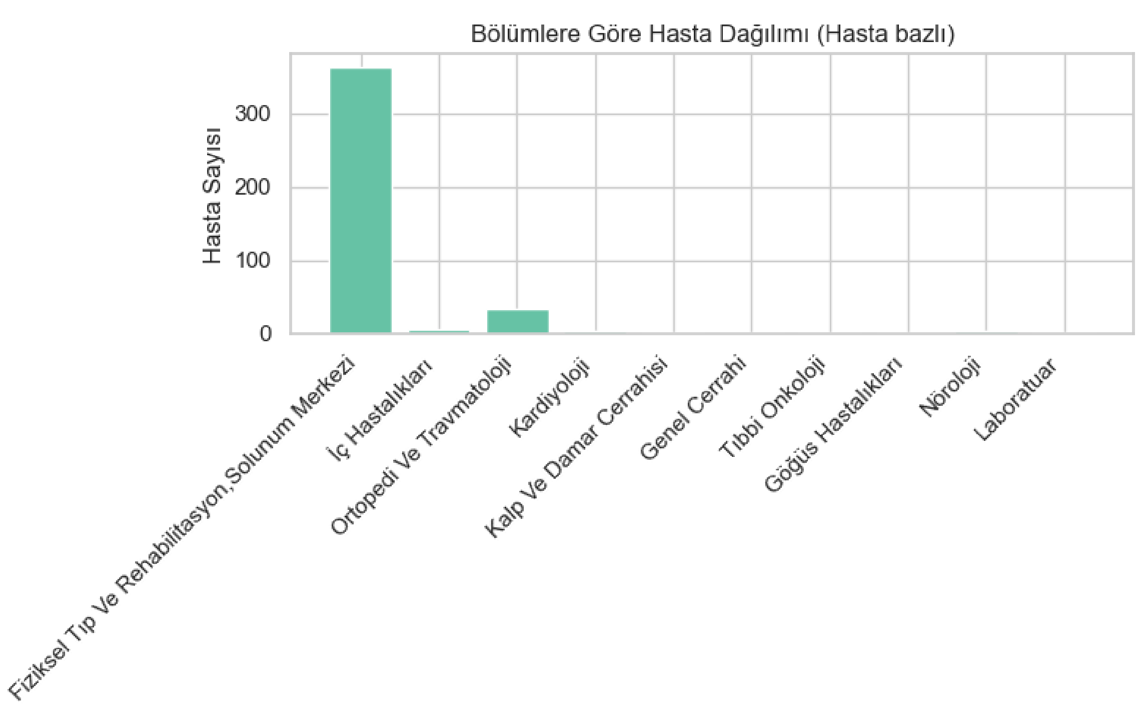
Treatment Duration Distribution by Department:

Boxplots were used to illustrate the distribution of treatment durations for each department. This provided insights into central tendencies, variability, and potential outliers in treatment durations.

Overall, this pipeline facilitated the cleaning and encoding of categorical department data while providing comprehensive visualizations of patient distribution and treatment durations

Author:Şeyma Nur Cin

Patient Distribution by Department



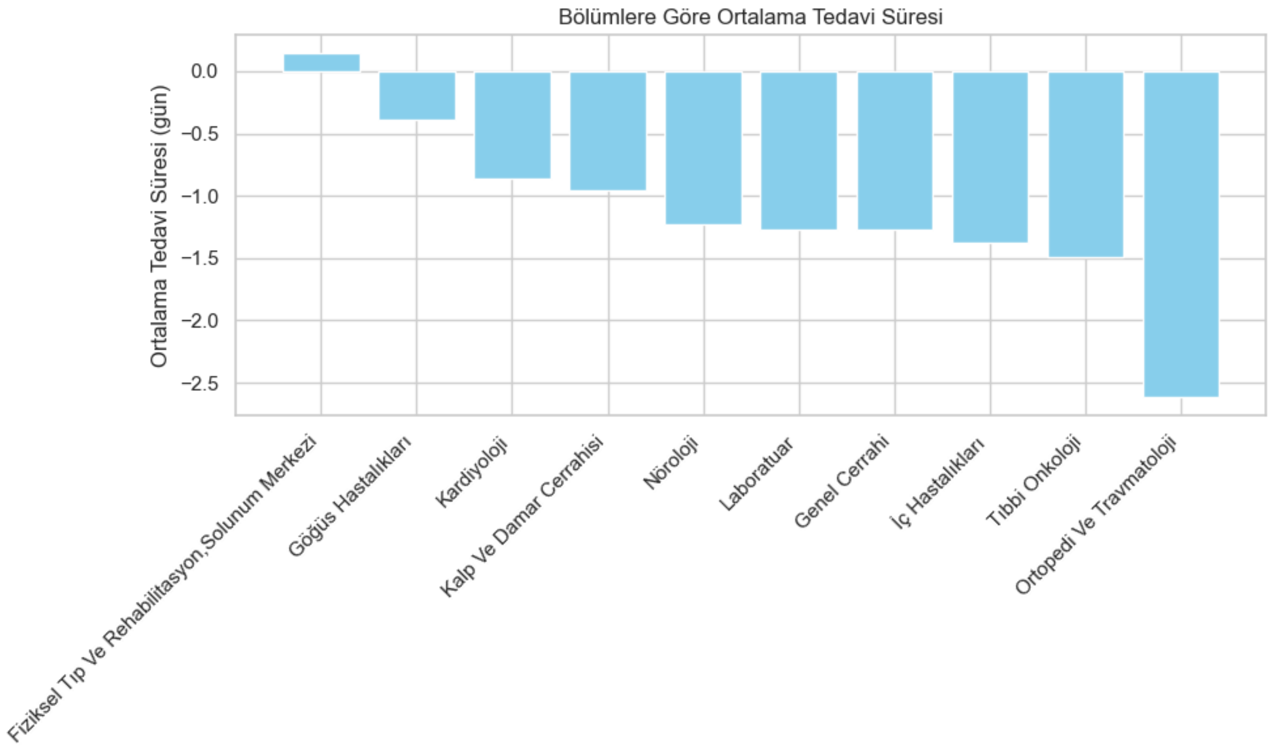
The **Patient Distribution by Department (Patient-Based)** chart shows the number of patients visiting each department. According to this bar chart:

- The **Physical Medicine and Rehabilitation, Respiratory Center** department stands out as the busiest, with approximately 350 patients.
- **Internal Medicine** and **Orthopedics and Traumatology** departments serve more patients compared to others.
- Departments such as **Cardiology, Cardiovascular Surgery, General Surgery, Medical Oncology, Chest Diseases, Neurology, and Laboratory** have relatively few patients.

This indicates that the hospital primarily focuses on **Physical Medicine and Rehabilitation** services.

Author:Şeyma Nur Cin

Treatment Duration Analysis by Department



Two charts were used to analyze treatment durations: **Treatment Duration Distribution (Boxplot)** and **Average Treatment Duration (Bar Chart)**. Both charts present the treatment durations using standardized values.

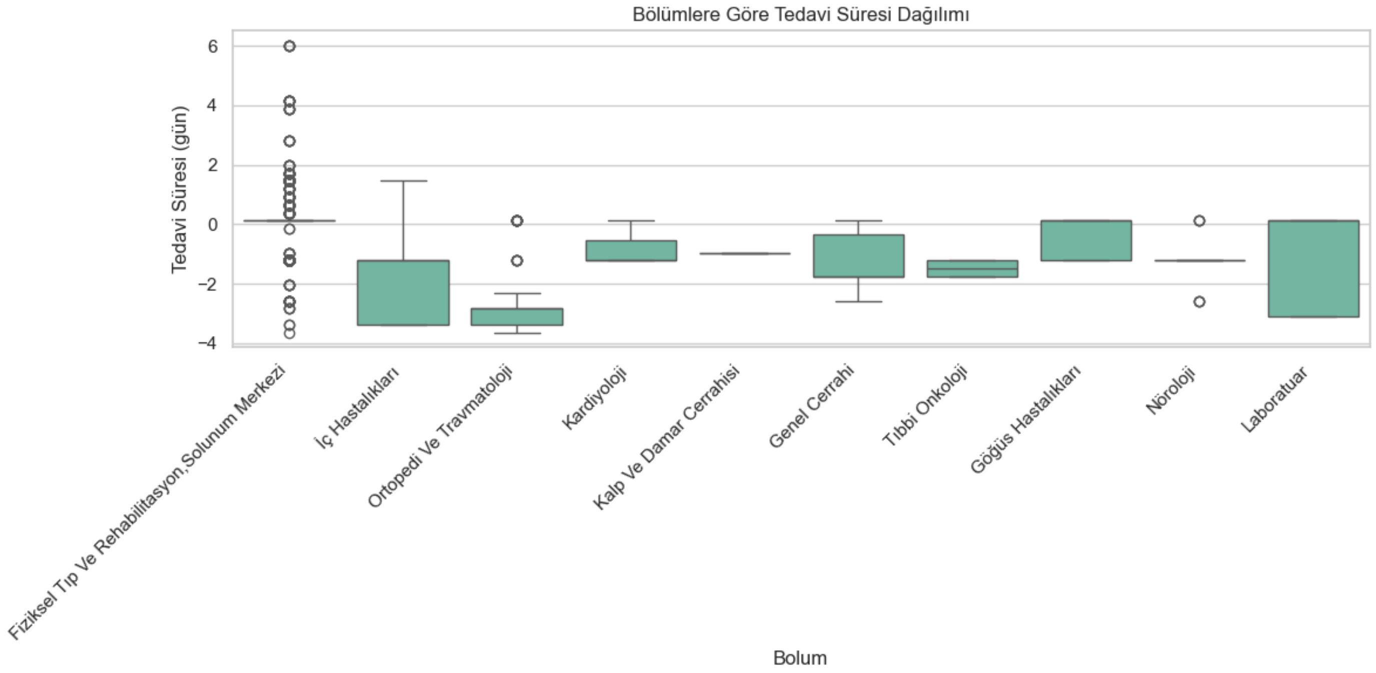
Average Treatment Duration

The **Average Treatment Duration by Department** chart shows the mean treatment duration for each department. Key observations include:

- The **Orthopedics and Traumatology** department has the longest average treatment duration.
- Departments such as **Medical Oncology, Internal Medicine, and General Surgery** also show relatively long average treatment durations.
- The **Physical Medicine and Rehabilitation, Respiratory Center** department has the shortest average treatment duration.

Author:Şeyma Nur Cin

Treatment Duration Distribution



The **Treatment Duration Distribution by Department** chart (boxplot) provides a more detailed view of the variation in treatment durations across departments:

- In the **Physical Medicine and Rehabilitation, Respiratory Center** department, although the average treatment duration is short, the distribution is wide. This suggests that some patients are discharged quickly while others receive longer treatments. The presence of numerous outliers (points extending above and below the boxes) illustrates this variation.
- In the **Orthopedics and Traumatology** department, the boxplot also shows high variability in treatment durations, in addition to a longer average duration.
- Other departments generally display a more homogeneous distribution of treatment durations.

These analyses provide valuable insights into how treatment processes vary across hospital departments. While the **Physical Medicine and Rehabilitation** department has the highest patient volume, its treatment durations are relatively short but highly variable. Conversely, departments like **Orthopedics and Traumatology** experience longer and more variable treatment durations. This information can inform resource planning and service improvements.

Author:Şeyma Nur Cin

Processing Allergy Data and Patient-Based Analysis

In this phase, a data processing and analysis pipeline was applied to the **Allergy** column in the dataset. The steps were performed as follows:

Normalization:
The normalize_allergies function was used to standardize allergy names. For example, "volteren" and "Voltaren" were converted to a single form ("voltaren"). This ensured that minor spelling differences or similar names were consolidated under a single category.

Filling Missing Values:

The fill_allergies function was used to fill empty or missing allergy data with the most frequent value. This step improved data completeness and analysis accuracy.

Conversion to Set:

The to_set function converted each allergy entry into a set of unique elements. This ensured that the same allergy for a patient was counted only once.

Pipeline Application:

The three steps above were combined into a **scikit-learn Pipeline** and applied to the entire dataset. As a result, the column df_processed['Alerji_Set'] was obtained.

One-Hot Encoding:

Using MultiLabelBinarizer, the allergy sets were transformed into a **one-hot encoded** numerical format, with each allergy represented as a separate column. This provides a suitable data format for machine learning and statistical analyses.

Patient-Based Allergy Counting:

The number of unique patients for each allergy type was calculated. This allowed identification of how many different patients have each allergy. Entries labeled as "Unknown" were excluded from the analysis.

Visualization:

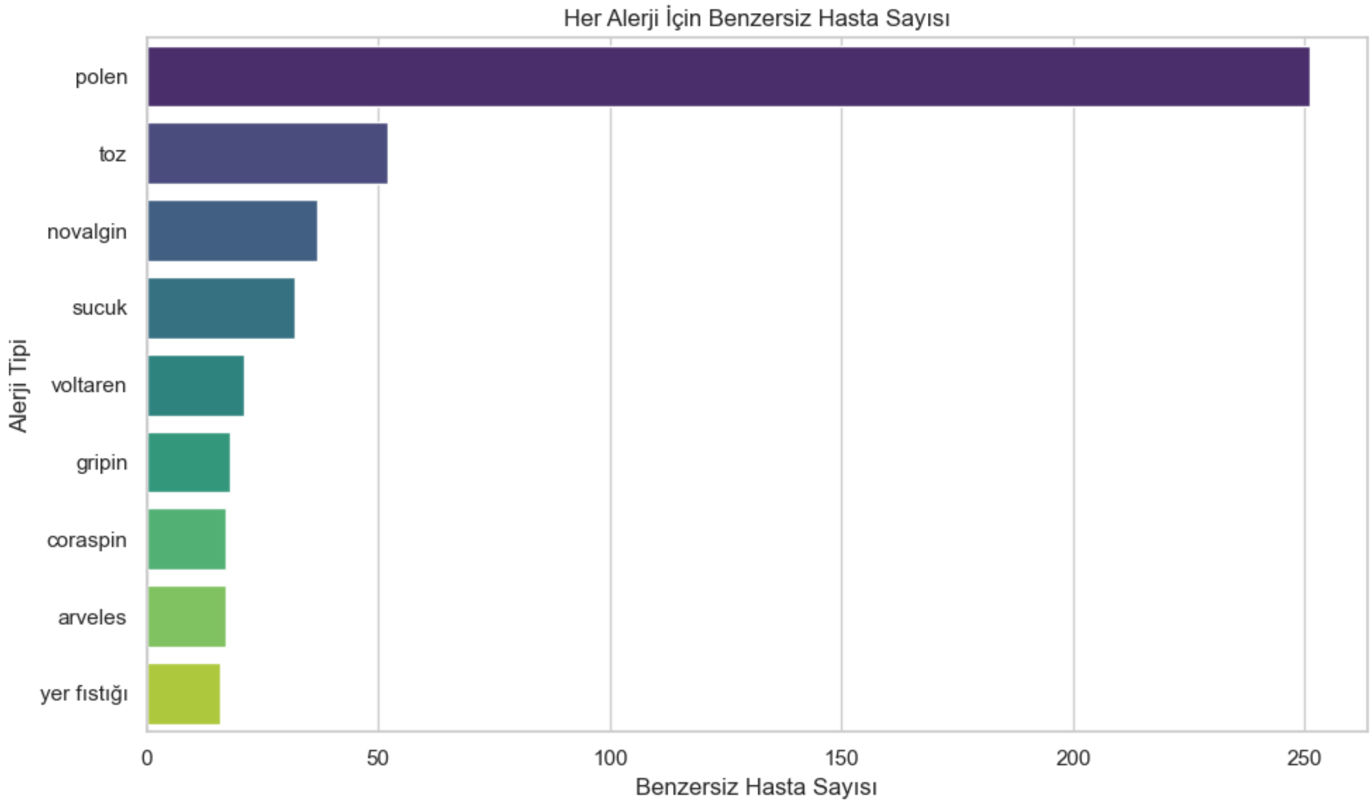
The patient-based allergy distribution was visualized using a bar plot. The chart shows the number of unique patients for each allergy type, allowing a quick overview of the prevalence of allergies in the dataset.

Example Patient Verification:

The allergy data for a specific patient ID was queried, and that patient's allergy set was verified. This step was used to confirm the correctness of the data processing pipeline.

Author:Şeyma Nur Cin

Analysis of Unique Patient Counts by Allergy Type



This horizontal bar chart shows the number of unique patients for each allergy type in the dataset. The chart compares the patient counts across different allergy categories.

Most Common Allergy Types:

- **Pollen allergy** is by far the most common, affecting approximately 250 patients. This indicates that a significant portion of patients in the dataset exhibit allergic reactions to pollen.
- **Dust allergy** follows with around 50 patients.
- **Novalgin** and **Sucuk** allergies also have a notable presence, each affecting roughly 40 patients.

Less Common Allergy Types:

- **Voltaren, Gripin, Coraspin, Arveles,** and **Peanut** allergies are less frequent, with patient counts ranging from 5 to 20.

Comments and Insights:

This analysis provides valuable information about the allergy profile of patients in the dataset. Sensitivities to allergens such as pollen, dust, Novalgin, and Sucuk are more prevalent in this patient group.

Author:Şeyma Nur Cin

Processing and Analysis of Treatment Name Data

In this stage, a data cleaning, normalization, and analysis pipeline was applied to the **TedaviAdi** (Treatment Name) and **TedaviSuresi** (Treatment Duration) columns of the dataset. The steps were carried out as follows:

Text Cleaning and Normalization:

- The temizle_ve_normalize_et function was used to standardize treatment names.
- Upper and lower case differences were corrected, and Turkish characters were normalized.
- Unnecessary words, symbols, and numbers were removed; "+" signs were replaced with commas, and multiple spaces were reduced to a single space.
- Common spelling errors and abbreviations were corrected using a predefined manual mapping dictionary (manual_mapping).

As a result, each treatment name was converted into a consistent and analysis-ready format.

Fuzzy Grouping (Merging Similar Treatments):

- The fuzzy_group function used fuzzy matching to merge similar treatment names into a single category.
- A similarity threshold of 86% was set, and treatments with high similarity scores were assigned to the same category.

Patient-Based Imputation:

- For each patient, missing treatment names were forward- and backward-filled (ffill and bfill).
- This ensured that missing treatment information for a patient was filled using other available records of the same patient.

Conversion of Treatment Duration to Numeric Format:

- The sureyi_sayiya_cevir function converted the **TedaviSuresi** column from string format to numeric format.
- This made treatment durations suitable for statistical analysis and visualization.

Frequency Analysis and Top 10 Treatments:

- The 10 most frequent treatments were identified from the cleaned data.
- Average treatment durations were calculated only for these top 10 treatments.

Visualization:

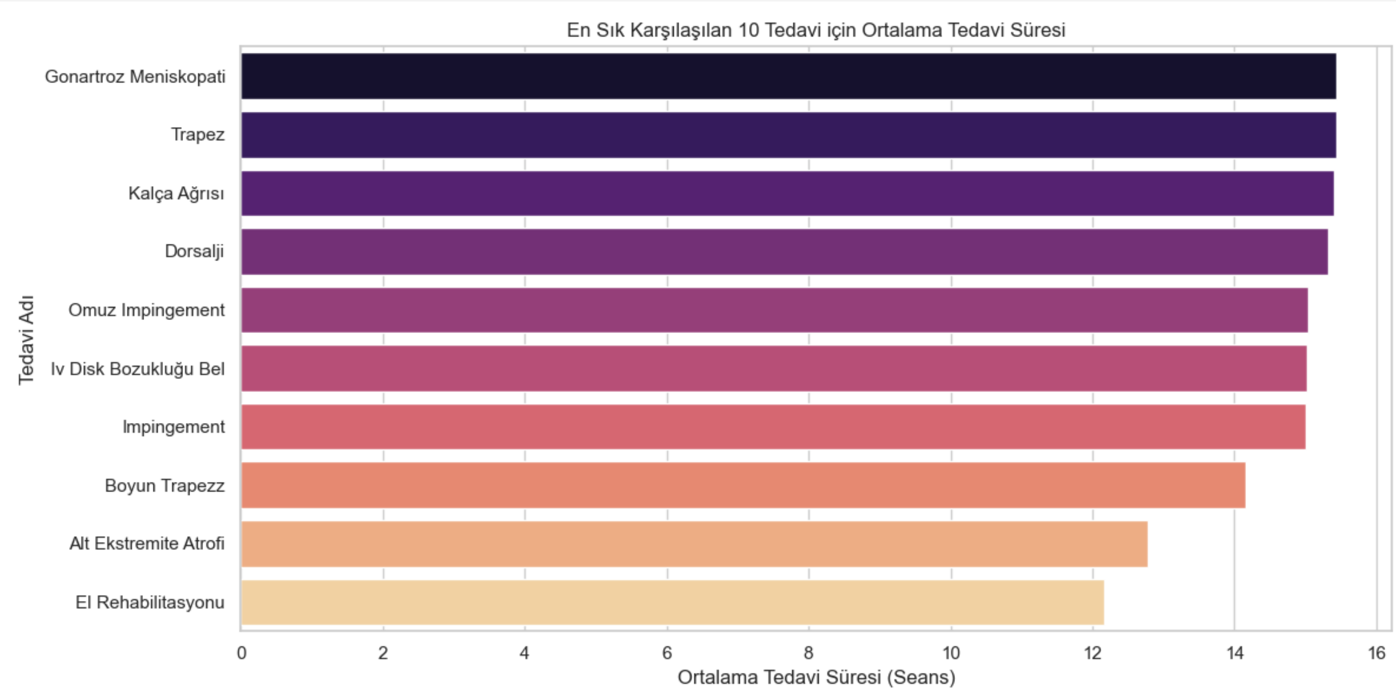
- Using the seaborn library, a horizontal bar chart was created to visualize the average treatment durations for the top 10 treatments.
- The chart allows easy comparison of the average treatment duration for each treatment.

Author:Şeyma Nur Cin

Conclusion:

This pipeline cleaned, normalized, and grouped similar treatment names. Treatment durations were converted to numeric format, making the data ready for analysis and visualization. As a result, comprehensive insights were obtained regarding treatment duration distributions per patient and the most frequently administered treatments.

Analysis of Average Treatment Duration for the Top 10 Most Frequent Treatments



This horizontal bar chart shows the average treatment durations (sessions) for the 10 most frequently administered treatment types in the dataset. In the chart, treatment names are displayed on the Y-axis, while average treatment durations are on the X-axis. Although the color gradient does not indicate ranking or intensity based on duration, it clearly visualizes the average number of sessions for each treatment.

Ranking by Treatment Duration:

According to the chart, the average treatment durations for the top 10 most frequent treatments are as follows:

- Treatments for **Gonarthrosis Meniscopathy, Trapezius, Hip Pain, Dorsalgia, Shoulder Impingement, and Lumbar IV Disc Disorders** have the longest average durations, around 15–16 sessions, placing them at the top.
- **Impingement** treatment follows with an average of approximately 15 sessions.
- **Neck Trapezius** treatment has an average of around 14 sessions.
- **Lower Extremity Atrophy** treatment averages about 12–13 sessions.
- **Hand Rehabilitation** has the shortest average duration in this group, approximately 12 sessions.

Observations and Insights:

This analysis demonstrates that certain treatment types require longer intervention periods. Specifically:

- **Joint-related issues** (Gonarthrosis Meniscopathy, Hip Pain, Shoulder Impingement, Impingement) and **spinal problems** (Dorsalgia, Lumbar IV Disc Disorders) require more sessions.
- **Musculoskeletal and neuromuscular treatments** (Trapezius, Neck Trapezius, Hand Rehabilitation, Lower Extremity Atrophy) also involve extended treatment durations.

These findings support several conclusions:

- **Treatment Planning:** Clinics and rehabilitation centers can plan patient schedules and allocate resources more effectively by considering these average durations.
- **Healthcare Burden:** Longer treatments mean more time and resources for patients. Conditions requiring longer interventions may reflect a higher overall burden on the healthcare system.
- **Rehabilitation Processes:** The relatively shorter average durations for rehabilitation-focused treatments (Lower Extremity Atrophy, Hand Rehabilitation) may indicate faster recovery potential or more efficient treatment protocols in these areas.

Processing and Patient-Based Analysis of Application Locations Data

In this stage, the **UygulamaYerleri (Application Locations)** column in the dataset was processed through data cleaning, missing value imputation, and one-hot encoding. The steps were carried out as follows:

Patient- and Treatment-Based Missing Value Imputation:

- Missing **UygulamaYerleri** values were filled on a per-patient and per-treatment basis.
- Filled values were based on other records of the same patient with the same treatment.
- If no valid value was available for a given patient and treatment, missing entries were assigned as **"Unknown"**.

Conversion to List and Cleaning:

- Data in the **UygulamaYerleri** column were converted into a list format, splitting entries by commas.
- Empty or unnecessary elements were removed, retaining only valid application locations.

One-Hot Encoding:

- Using **MultiLabelBinarizer**, each application location was transformed into a separate numerical column.
- This produced binary (0/1) indicators for each location, making the data suitable for machine learning and statistical analysis.

Merging and Duplicate Removal:

- One-hot encoded columns were merged back into the original **df_processed** dataframe.
- Duplicate rows were removed based on **HastaNo (Patient ID)**, **TedaviAdi_clean (Treatment Name)**, and **UygulamaSuresi (Application Duration)** columns.

Sample Patient Check:

- For a specific patient (**HastaNo = 145141**), application locations and durations were checked to verify the accuracy of the data processing steps.

This pipeline ensures that the **UygulamaYerleri** data are cleaned, missing values are filled on a patient-specific basis, and the data are transformed into a numerical format suitable for analysis and visualization.

Processing and Analysis of Diagnoses Data

In this stage, the **Tanilar (Diagnoses)** column in the dataset was processed and analyzed using a dedicated pipeline. The operations were performed as follows:

Data Preparation

Data Type Conversion and Standardization:

The **HastaNo** column was converted to string type, and **Tanilar** values were converted to lowercase, stripped of leading/trailing spaces, ensuring consistency across the dataset.

Extra Words Removal:

A predefined list of irrelevant words and phrases (**ek_kelimeler**) such as anatomical regions or generic terms was removed from diagnosis entries to focus on clinically relevant terms.

Diagnoses Cleaning Pipeline

Text Cleaning:

The temizle method removed punctuation, extra spaces, and irrelevant keywords.

Fuzzy Grouping of Similar Diagnoses:

Using **fuzzy matching** (rapidfuzz), similar diagnosis terms were grouped under a single standardized category. A similarity threshold of 85% was used to combine closely matching diagnoses.

Patient-Based Standardization:

Each patient's diagnoses were mapped to the cleaned and standardized groups. Missing or empty diagnosis entries for a patient were filled with the most common diagnosis for the corresponding treatment or assigned **"bilinmiyor" (unknown)** if no reference existed.

Frequency Analysis and Visualization

Unique Patient Diagnoses:

For each patient, a set of unique diagnoses was created, ensuring no duplicate diagnoses were counted multiple times.

Top 10 Diagnoses:

The 10 most frequent diagnoses across all patients were identified and visualized using a horizontal bar plot, showing the number of patients affected by each diagnosis.

Patient-Based Data Checks

- Example patients were examined (e.g., **HastaNo = 145141**) to verify that cleaned and standardized diagnoses were correctly mapped.

Additional Exploratory Analyses

- **Age vs. Treatment Duration:**

A scatter plot was generated to explore the relationship between patient age and treatment duration, with patients color-coded by gender.

- **Gender-Based Treatment Duration Distribution:**

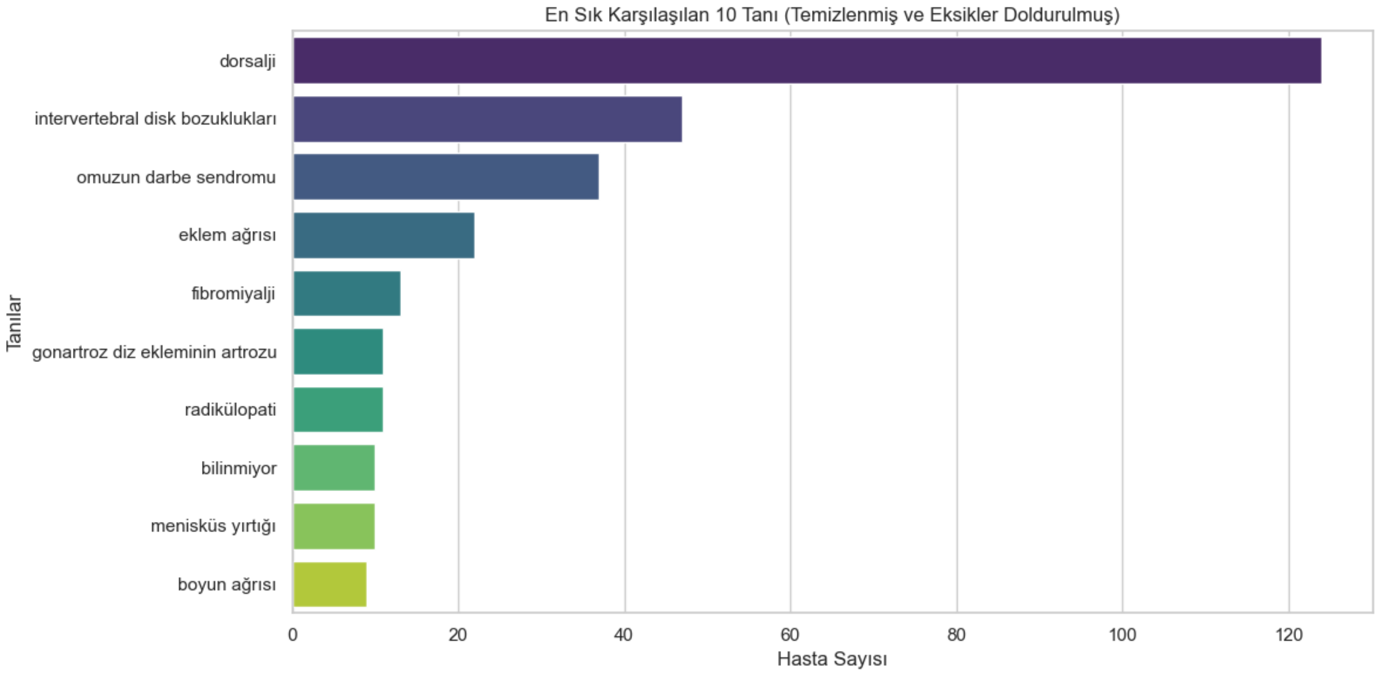
  A box plot was created to examine differences in treatment duration across genders.

Summary:

This pipeline effectively cleaned, normalized, and standardized diagnoses data, accounted for missing values, and grouped similar diagnoses using fuzzy matching. The resulting dataset supports patient-level diagnosis analysis, frequency counts, and exploratory visualizations linking demographic features (age, gender) to treatment duration.

Author:Şeyma Nur Cin

Top 10 Most Common Diagnoses (Cleaned and Missing Values Filled)



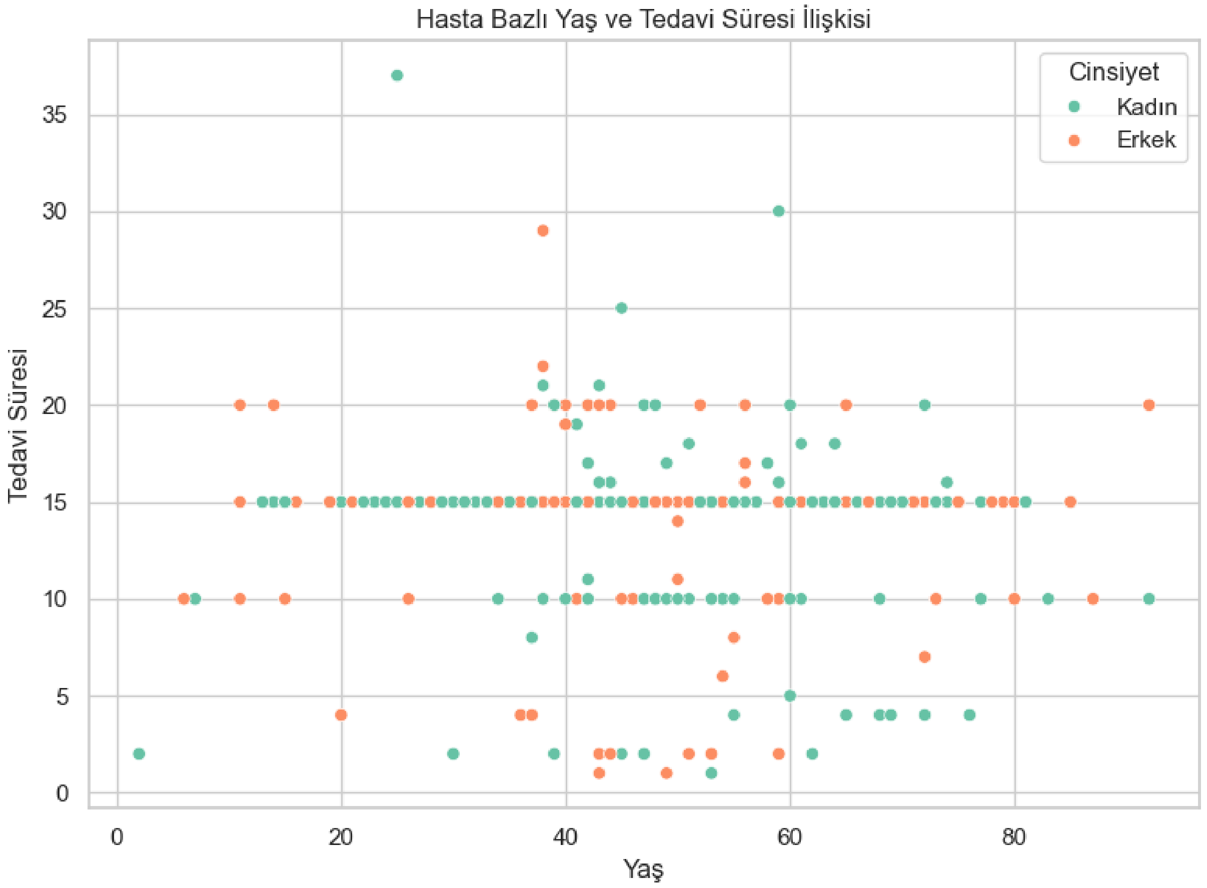
This horizontal bar chart displays the **10 most frequently diagnosed conditions** in the dataset, along with the number of patients for each diagnosis. The data has been cleaned and missing values have been filled.

- **Dorsalji** is the most common diagnosis, affecting approximately 120 patients.
- This is followed by **Intervertebral Disc Disorders**, with around 50 patients.
- **Shoulder Syndrome** ranks third, with about 40 patients.
- **Joint Pain, Fibromyalgia, Gonarthrosis (Knee Osteoarthritis), Radiculopathy, Unknown, Meniscus Tear,** and **Neck Pain** have lower patient counts.

These findings indicate that the majority of patients in the dataset experience **musculoskeletal and spinal issues**.

Author:Şeyma Nur Cin

Patient-Based Age and Treatment Duration Relationship



This scatter plot visualizes the relationship between **patients' ages** and **treatment session durations**, with points colored by gender.

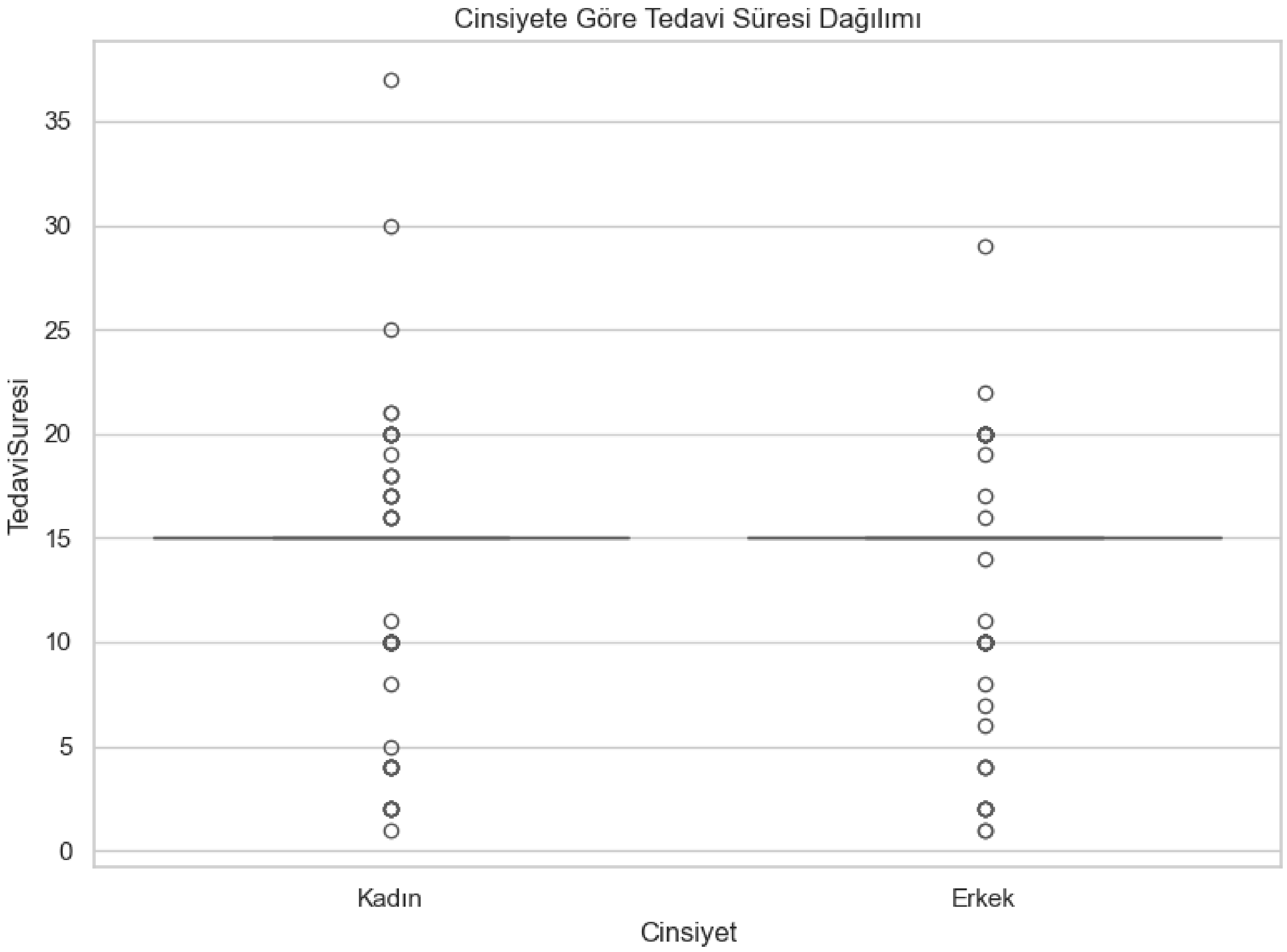
Observations:

- Overall, there is **no clear linear relationship** between age and treatment session duration. In other words, as a patient's age increases, there is **no systematic trend** of treatment duration increasing or decreasing.
- Treatment session durations are generally concentrated between **5 and 20 sessions**. Some outlier cases exist, particularly among younger patients (approximately **20–40 years old**), who have longer treatment durations (**30 sessions or more**).
- Patients of both genders appear across similar age groups and treatment durations. No significant separation or trend is observed based on gender.

Interpretation:
This visualization suggests that **age alone is not a determining factor** for treatment session length. Instead, factors such as **the type of condition, treatment method, and individual recovery process** likely play a more important role.

Author:Şeyma Nur Cin

Treatment Duration Distribution by Gender (Session-Based)



This boxplot compares the distribution of treatment session durations between **female and male patients**.

Observations:

- The distribution of treatment durations appears **similar for both genders**. The **median treatment duration** is approximately **15 sessions** for each group.
- Both groups show a **similar number of outliers** within comparable ranges. This suggests that there is **no significant difference in treatment durations between genders**.

Interpretation:
This analysis indicates that **treatment session durations are likely influenced more by factors such as the type of treatment or the patient's overall health** rather than by gender.

Author:Şeyma Nur Cin

Report Summary

This report encompasses a detailed analysis conducted based on the provided data. The analyses focus on patient demographics, disease distributions, treatment durations, and department-based statistics.

Key Findings and Analyses:

Patient Demographics:

- **Age Distribution:** There are no missing values in the age data. Patient-level analyses indicate that ages are generally concentrated in the 40s, though the dataset spans a wide age range.
- **Gender Distribution:** Some missing values existed in the gender column, which were imputed with the most common value. The gender data was then converted to a numerical format for analysis. The resulting visualizations show that the number of female and male patients is roughly balanced.
- **Nationality Distribution:** No missing values were found in nationality. Patient-level charts reveal that the vast majority of patients are Turkish, with very few from other nationalities.
- **Blood Type Distribution:** Missing blood type values were filled using other records of the same patient or with the KNN Imputer method. Analysis shows that the most common blood types are A Rh+ and O Rh+. The charts clearly reflect the numerical dominance of these groups.

Disease and Allergy Distributions:

- **Chronic Diseases:** Patient distribution by chronic disease shows that Arrhythmia has the highest patient count, while other chronic diseases have fewer patients.
- **Diagnoses:** Among the top 10 most frequent diagnoses, dorsalgia and intervertebral disc disorders are the most common.
- **Allergies:** Analysis of allergy types shows that pollen allergy is the most prevalent, followed by notable occurrences of dust and novalgin allergies.

Department and Treatment Duration Analysis:

- **Patient Distribution by Department:** The Physical Medicine and Rehabilitation and Respiratory Center departments serve noticeably more patients than other departments.
- **Average Treatment Duration:** Department-level analysis reveals that Orthopedics and Traumatology have the longest average treatment durations, whereas Physical Medicine and Rehabilitation and the Respiratory Center have the shortest.
- **Treatment Duration Distribution (Boxplot):** This analysis provides detailed insight into the variability and outliers of treatment durations. Physical Medicine and Rehabilitation and the Respiratory Center show a wide range of treatment durations with several outliers.
- **Most Frequent Treatments:** The top 10 most frequently applied treatments were examined, showing that Gonarthrosis Meniscopathy, Trapezius, and Hip Pain treatments have similar and the longest average session durations. This session-based analysis offers insights into treatment protocol lengths.

- **Application Duration:** Analysis of application durations indicates clustering around specific session counts, with a notable concentration around 20 sessions.

These analyses provide a foundational understanding of the hospital’s patient profile, the most common conditions, departmental workload, and treatment efficiency. The findings can support strategic decisions aimed at optimizing services and improving patient care quality.