Sena YAMAN

yamansena6@gmail.com

Pusula Talent Academy Data Science Intern Case Study

Sağlık Veri Setinde EDA ve Veri Önişleme

1.Amaç

Bu rapor, hasta verilerini içeren veri setini inceleyerek temel istatistikleri, eksik değerleri ve veri kalitesi ile ilgili olası sorunları tespit etmeyi amaçlamaktadır. Analiz sürecinde EDA (Keşifsel Veri Analizi), veri ön işleme adımları ve görselleştirme teknikleri kullanılmıştır.

2.Veri Seti Tanımı

- 2235 satır,13 Sütun
- Kolon isimleri ve 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
```

- Sayısal Değişkenler : 2
- Kategorik Değişkenler : 11

3.Özet İstatistikler

Sayısal değişkenlerin istatistikleri incelenmiştir.

	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

Kategorik değişkenlerin istatistikleri incelenmiştir.

	Cinsiyet	KanGrubu	Uyruk	KronikHastalik
count	2066	1560	2235	1624
unique	2	8	5	220
top	Kadın	0 Rh+	Türkiye	Myastenia gravis
freq	1274	579	2173	38

Genel yorum:

- Bazı kolonlarda eksik veri var (Cinsiyet, KanGrubu, KronikHastalik).
- Uyruk kolonu neredeyse tam dolu ve çoğunluk Türkiye vatandaşlarından oluşuyor.
- KronikHastalik çok çeşitli, en sık hastalık bile nadir (38/1624).
- Cinsiyet dağılımında kadınlar erkeklerden fazla.
- KanGrubu verisi eksikliği göze çarpıyor.

4.Eksik Veri Analizi

Eksik değer bulunan kolonlar tespit edildi.

```
HastaNo              0
Yas                  0
Cinsiyet           104
KanGrubu           365
Uyruk                0
KronikHastalik     345
Bolum                7
Alerji             540
Tanilar             46
TedaviAdi            0
TedaviSuresi         0
UygulamaYerleri    157
UygulamaSuresi       0
```

Veri setindeki eksik değerler, uygun şekilde doldurularak tamamlandı. Kategorik sütunlarda en sık görülen değerler (mode) kullanılırken, bazı bilinmeyen veya eksik bilgiler "Bilinmiyor" veya "KronikHastalik" gibi açıklayıcı sabit değerlerle dolduruldu. Bu sayede analiz ve görselleştirmelerde eksik veri kaynaklı hataların önüne geçildi.

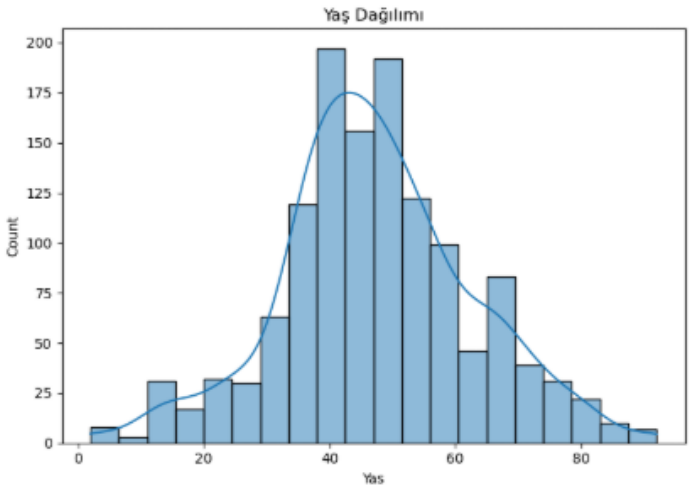
5.Sütunlara Yönelik Metin Standartlaştırma

Metin sütunları, boşluklar temizlenip küçük harfe çevrilerek ve 'nan' değerleri düzeltilerek standartlaştırıldı; böylece veri tutarlılığı ve hatasız analiz sağlandı.

6.Görselleştirme

6.1.Sayısal Değişkenlerin Görselleştirilmesi

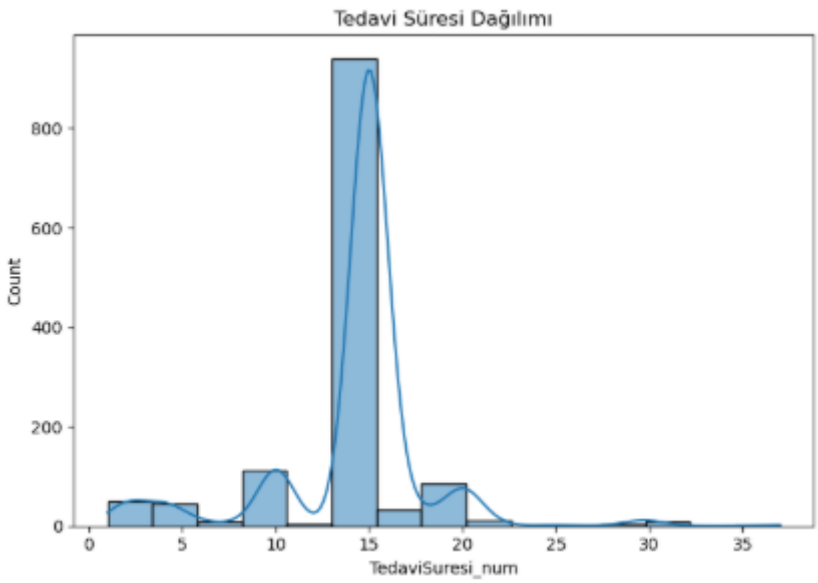
Veri Setindeki Yaş Dağılımı



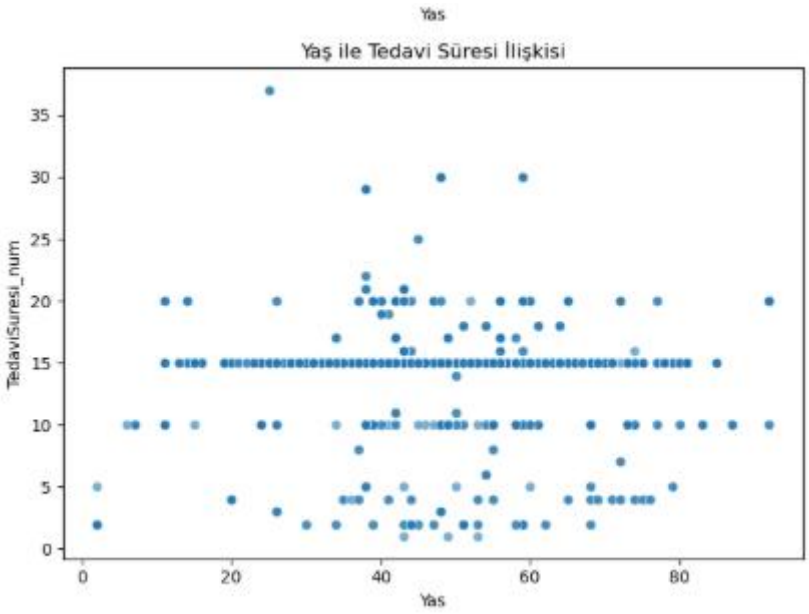
- Yaşlar yaklaşık olarak **normal dağılıma** benziyor; ortalama civarında (40–50 yaş arası) yoğunluk daha yüksek.
- En yüksek sayıda birey **40–50 yaş aralığında** bulunuyor.
- 20 yaşın altı ve 70 yaşın üstü kişiler nispeten az.
- Dağılım simetrik sayılabilir ancak sağa doğru hafif bir uzama (55–80 yaş arası) gözleniyor.

Tedavi Süresi Dağılımı

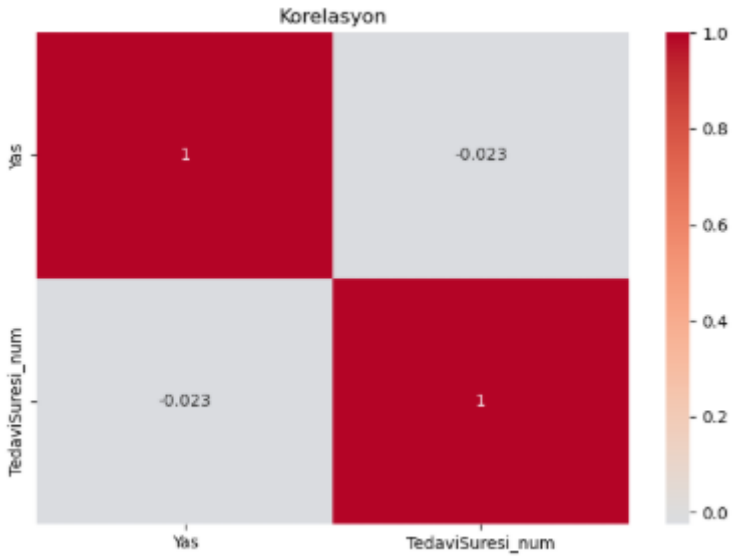
- Tedavi süreleri büyük ölçüde **10–15 gün aralığında yoğunlaşmaktadır**.
- En yüksek sıklık yaklaşık **13–15 gün aralığında** gözlenmektedir.
- **5 günün altında** ve **20 günün üzerinde** tedavi süreleri nispeten az sayıda bulunmaktadır.
- Dağılım genel olarak **sağa çarpık** olup, uzun süreli tedaviler (20 gün ve üzeri) sağ kuyruğu uzatmaktadır.



Yaş ile Tedavi Süresi İlişkisi



- Yaş ile tedavi süresi arasında **belirgin bir doğrusal ilişki görülmemektedir**.
- Tedavi süresi çoğunlukla **10–15 gün aralığında** yoğunlaşmış olup, bu durum farklı yaş grupları için de benzerdir.
- **Genç yaşlarda (0–20 arası)** ve **ileri yaşlarda (70 ve üzeri)** tedavi süresi dağılımı daha seyrektir.
- Bazı bireylerde tedavi süresi **20 günün üzerine çıkmakta**, ancak bu durum yaşa bağlı sistematik bir artış göstermemektedir.

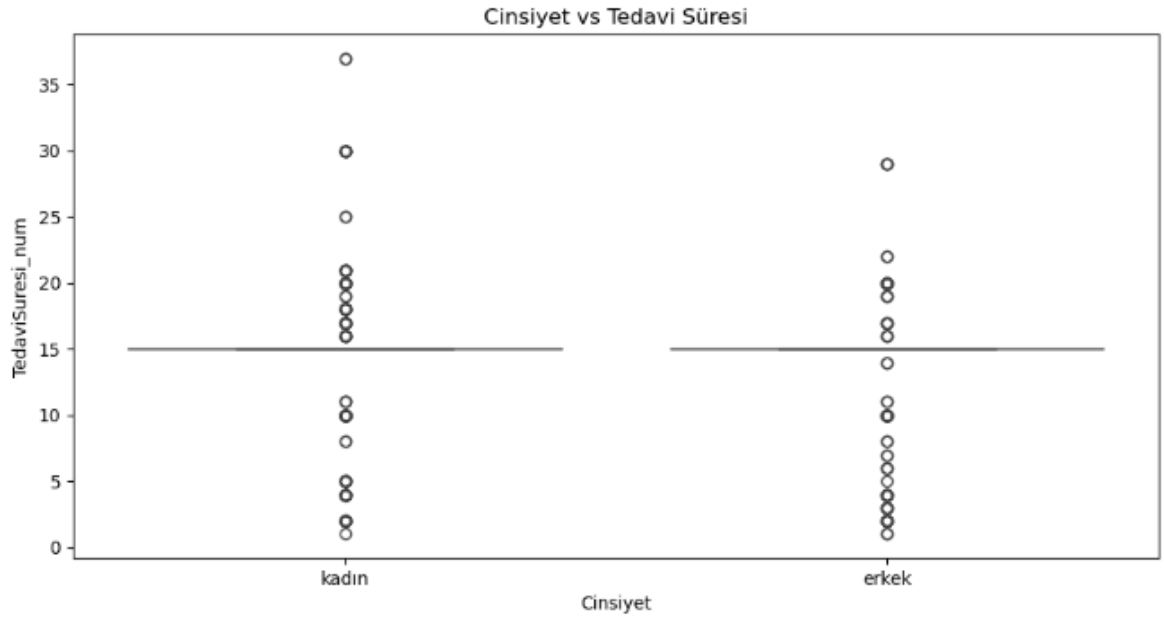


- Yaş ile tedavi süresi arasındaki korelasyon katsayısı **-0.023** olup, bu değer **neredeyse sıfıra yakın** çıkmıştır.
- Bu sonuç, **yaş ile tedavi süresi arasında anlamlı bir ilişki olmadığını** göstermektedir.

- Korelasyon değerinin negatif olması, yaş arttıkça tedavi süresinin çok küçük bir oranda azalabileceğini ima etse de bu ilişki **istatistiksel olarak önemsizdir**.
- Dolayısıyla yaş değişkeni, tedavi süresini tahmin etmek için **güçlü bir belirleyici değildir**.

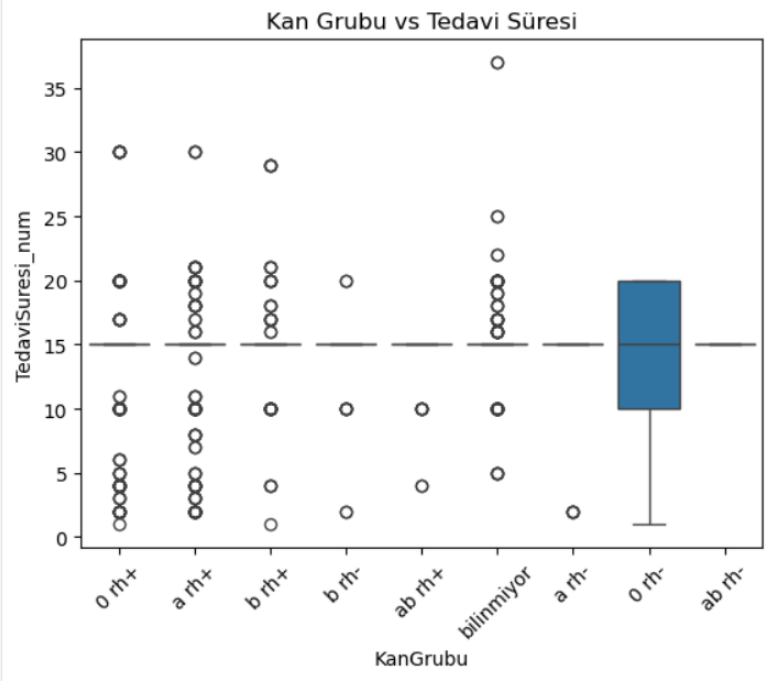
6.2.Kategorik Değişkenlerin Görselleştirilmesi

Cinsiyet ve Tedavi Süresi



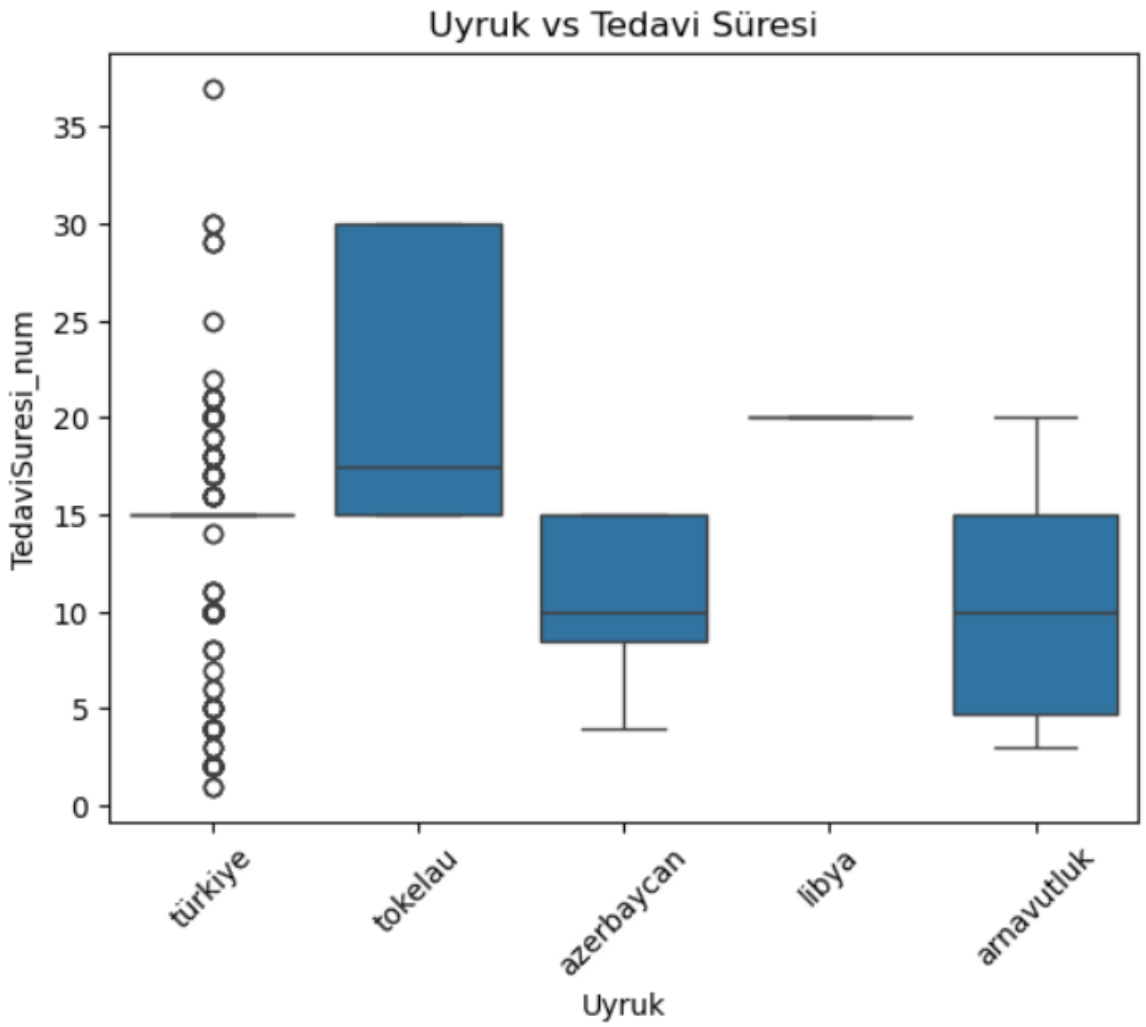
- Kadınlar grubunda tedavi süreleri genellikle 0 ile 37 arasında değişmektedir. Noktaların yoğunlaştığı aralık yaklaşık olarak 15 civarındadır.
- Erkekler grubunda da tedavi süreleri benzer bir aralıkta (0 ile 30 civarı) yoğunlaşmaktadır ve ortalama tedavi süresi de yaklaşık 15'tir.

Kan Grubu ve Tedavi Süresi



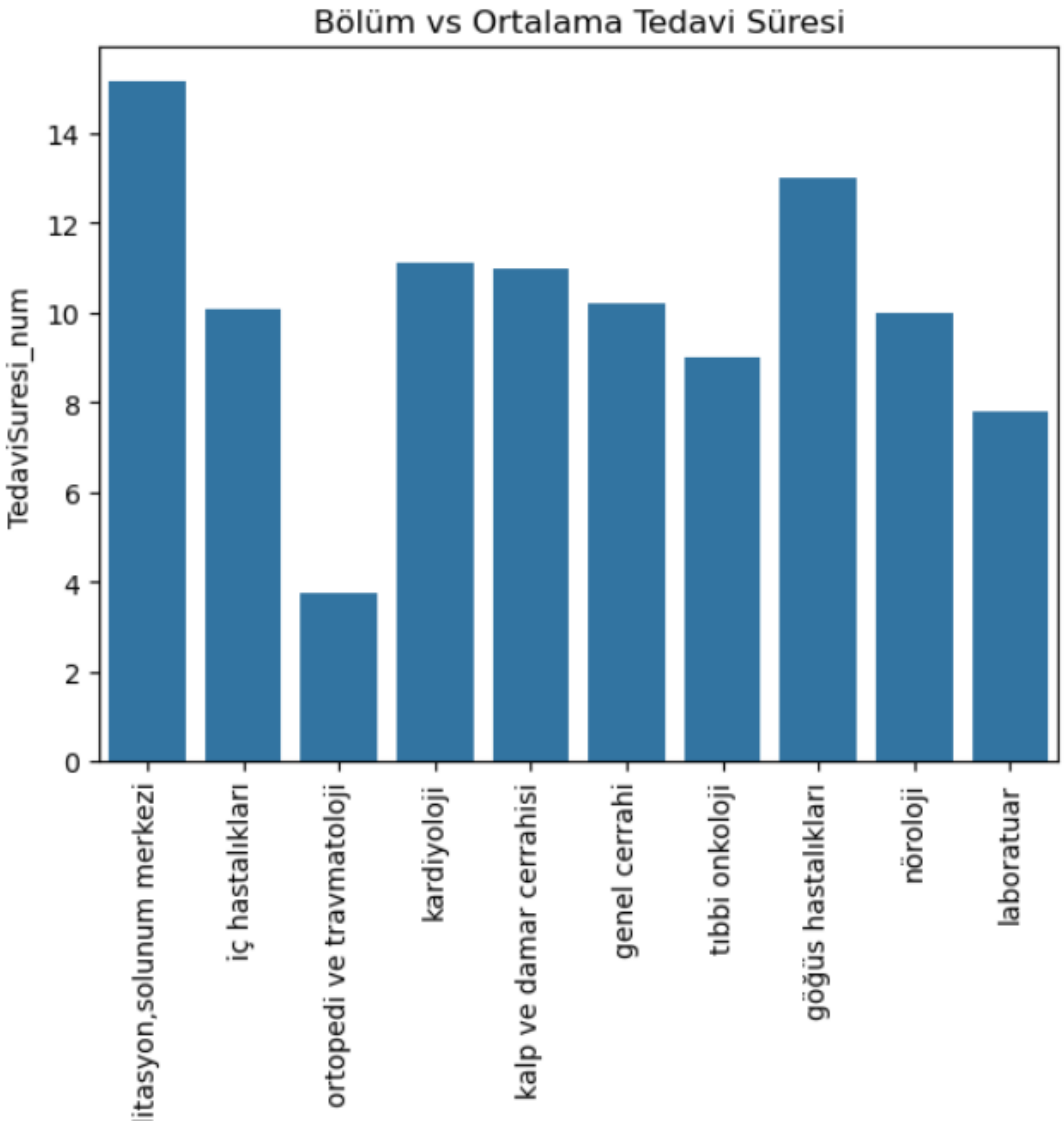
- **0 Rh+:** Bu kan grubundaki hastaların tedavi süreleri genellikle 5 civarında yoğunlaşmaktadır.
- **A Rh+:** A Rh+ grubunda, tedavi süreleri hem 5 civarında bir yığılma göstermekte hem de 20'li değerlere kadar çıkabilmektedir.
- **B Rh+:** B Rh+ grubunda da benzer şekilde, 5 civarında bir yoğunluk ve 20'li değerlere kadar uzanan tedavi süreleri görülmektedir.
- **B Rh-:** Bu grupta tedavi süreleri daha dağınık bir dağılım göstermekle birlikte, bazı hastalar 20'li değerlere ulaşmaktadır.
- **AB Rh+:** Bu grupta da tedavi süreleri 5 civarında ve daha üst değerlerde çeşitlilik göstermektedir.
- **Bilinmiyor:** Kan grubu bilinmeyen hastalarda tedavi süreleri en geniş aralıkta dağılmakta, 5'ten başlayıp 35'e kadar çıkabilmektedir.
- **A Rh-:** Bu grupta veri noktası az olsa da, tedavi süreleri 15 civarında görünmektedir.
- **0 Rh-:** Bu kan grubunda, **tedavi süreleri en belirgin şekilde 10 ile 20 aralığında yoğunlaşmıştır.** Grafikteki kutu grafiği (box plot) de bu durumu desteklemektedir.
- **AB Rh-:** Bu grupta tedavi süresi verisi sınırlıdır ve 15 civarında görünmektedir.

Uyruk ve Tedavi Süresi



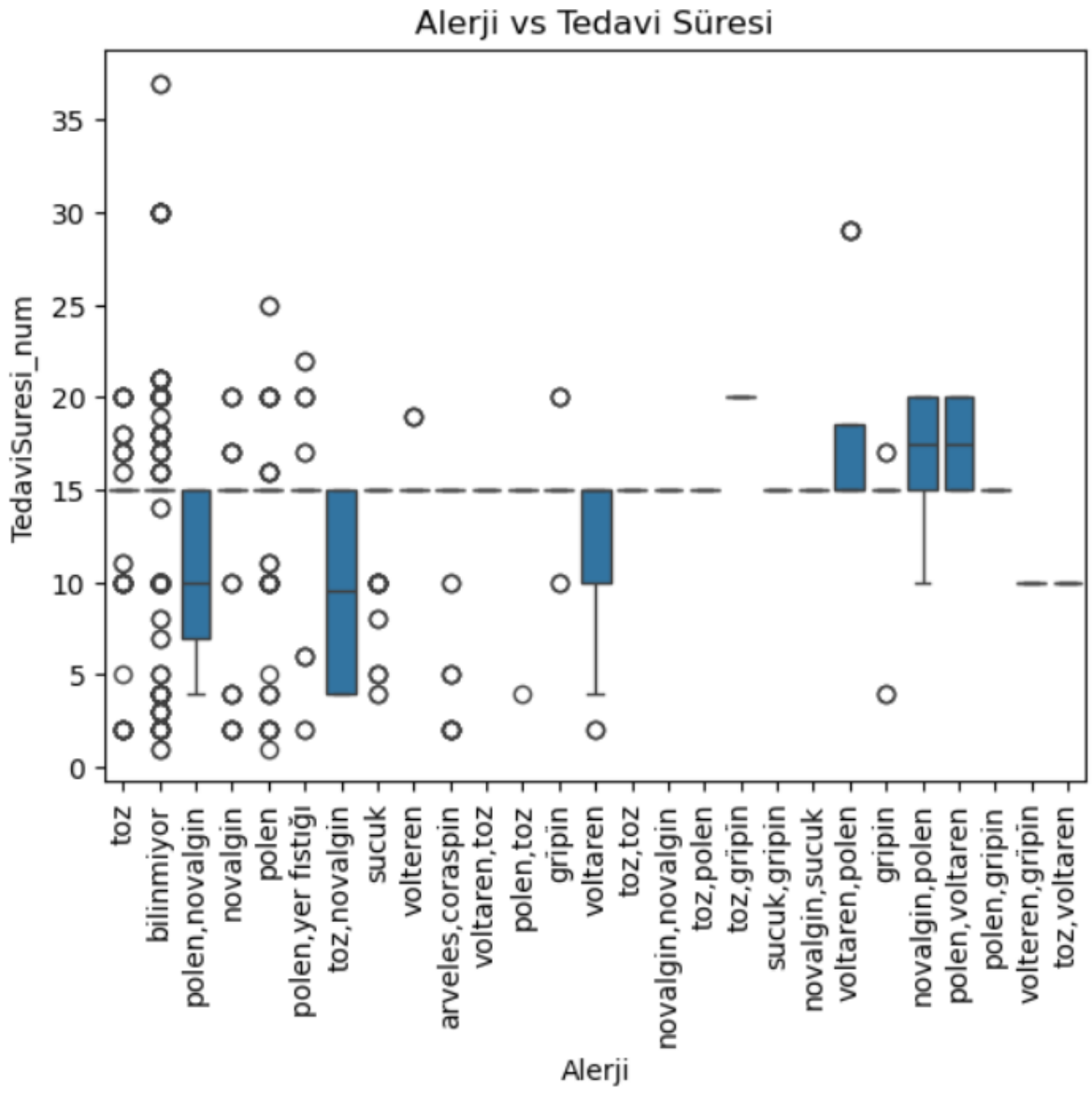
- **Türkiye** uyruklu bireylerde tedavi süreleri en **değişken** ve en **uzun olabilen** gruptur.
- **Azerbaycan** ve **Arnavutluk** uyruklu bireylerde tedavi süreleri daha **kısadır** ve birbirine yakındır.
- **Tokelau** uyruklu bireylerde tedavi süreleri orta düzeydedir ve daha az değişkendir.
- **Libya** uyruklu bireyler hakkında yeterli veri olmamakla birlikte, mevcut veri ortalamanın üzerindedir.

Bölüm ve Ortalama Tedavi Süresi



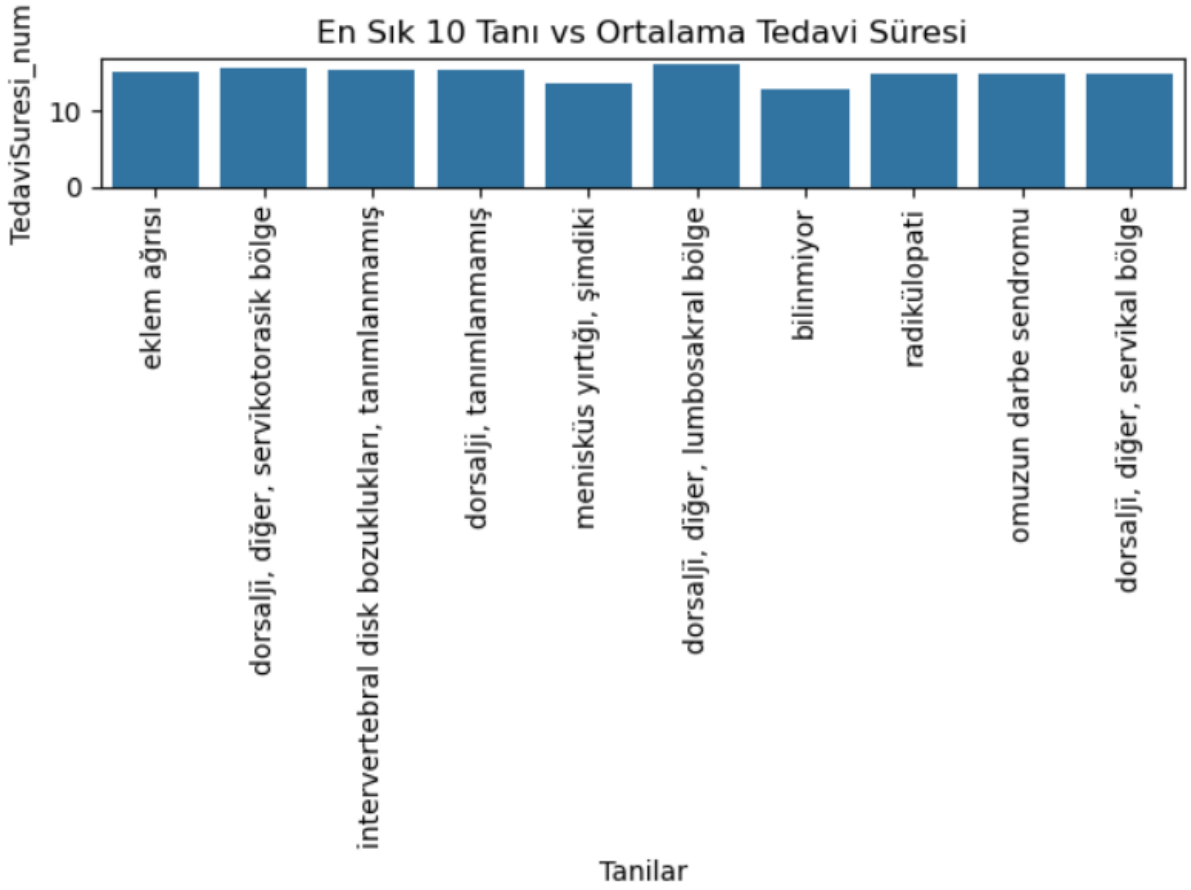
- **Fiziksel Tıp ve Rehabilitasyon, Solunum Merkezi:** Bu bölüm, en uzun ortalama tedavi **süresine** sahiptir. Ortalama süre yaklaşık 14.5 gündür.
- **Göğüs Hastalıkları:** Bu bölüm, yaklaşık 12.5 günlük ortalama tedavi süresi ile ikinci en uzun tedavi süresine sahiptir.
- **Kardiyoloji ve Kalp Damar Cerrahisi:** Bu iki bölümün ortalama tedavi süreleri birbirine oldukça yakındır ve yaklaşık 11.5 gündür.
- **Genel Cerrahi:** Ortalama tedavi süresi yaklaşık 10.5 gündür.
- **Nöroloji:** Ortalama tedavi süresi yaklaşık 10 gündür.
- **Tıbbi Onkoloji:** Bu bölümün ortalama tedavi süresi yaklaşık 9 gündür.
- **Laboratuvar:** Ortalama tedavi süresi yaklaşık 8 gündür.
- **İç Hastalıkları:** Bu bölümün ortalama tedavi süresi yaklaşık 10 gündür.
- **Ortopedi ve Travmatoloji:** Bu bölüm, **en kısa ortalama tedavi süresine** sahiptir. Ortalama süre yaklaşık 4 gündür.

Alerji ve Tedavi Süresi



Genel olarak, alerjen türlerine göre tedavi sürelerinde önemli farklılıklar görülmemektedir. Ancak, **"toz" alerjeninde gözlemlenen yüksek değişkenlik ve belirgin aykırı değerler**, bu durumda tedavi sürelerinin daha geniş bir yelpazede olabileceğini düşündürmektedir. Ayrıca, bazı alerjen kombinasyonlarında (örneğin, "polen, yer fıstığı" veya "arveles, coraspin") tedavi sürelerinin daha kısa olduğu söylenebilir.

En Sık 10 Tanı ve Ortalam Tedavi Süresi



Grafiğe göre, **eklem ağrısı, dorsalji ile ilişkili çeşitli bölgelerdeki ağrılar ve intervertebral disk bozuklukları gibi tanılar, ortalama olarak daha uzun tedavi sürelerine** sahip görünmektedir. Diğer tanılar ise daha kısa ve birbirine yakın ortalama tedavi sürelerine sahiptir. Ortopedi ve travmatoloji ile ilişkili tanılarda tedavi sürelerinin daha uzun olması beklenebilirken, bu grafikte bu tür tanılar da orta ve uzun tedavi süreleri aralığında yer almaktadır.

7.Veri Ön İşleme

Kategorik Değişkenlerin Dönüşümü:

- "Cinsiyet", "KanGrubu", "Uyruk", "Bolum" ve "TedaviSuresi" gibi kategorik değişkenler, makine öğrenmesi modellerinde kullanılabilmesi için **One-Hot Encoding** yöntemiyle sayısal değerlere dönüştürülmüştür. drop_first=True parametresiyle kategorilerin birinin çıkarılmasıyla çoklu doğrusallık (multicollinearity) önlenmiştir.
- "Alerji" ve "UygulamaYerleri" değişkenleri, **Label Encoding** ile sayısal değerlere dönüştürülmüştür.

Frekans Kodlama:

- "KronikHastalik", "Tanilar" ve "TedaviAdi" gibi değişkenler, her bir kategorinin veri setindeki **toplam frekansına** göre kodlanmıştır. Bu, her bir hastalığın veya tedavinin ne kadar sık görüldüğünü gösteren bir oran elde etmemizi sağlamıştır.

Sayısal Değişkenlerin Ölçeklendirilmesi:

- "Yas" ve "TedaviSuresi_num" gibi sayısal değişkenler **StandardScaler** kullanılarak ölçeklendirilmiştir. Bu adım, farklı ölçeklerdeki sayısal verilerin model performansını olumsuz etkilemesini önlemek amacıyla uygulanmıştır.

8.Sonuç ve Gözlemler

- Veri seti genel olarak temizlenmiş ve analiz için hazırdır.
- Tedavi sürelerini etkileyen en güçlü faktörler bölüm, tanı ve bazı alerji türleri gibi klinik değişkenlerdir; yaş ve cinsiyet gibi demografik değişkenler tedavi süresinde anlamlı bir değişken olarak öne çıkmamaktadır.
- Kan grubu ve uyruk değişkenleri, bazı özel durumlar dışında tedavi süresi üzerinde sınırlı bir etkide bulunuyor.