Data Science Intern Case Study

Emine Yiğit
emine.ygt@outlook.com

Bu projenin temel amacı, bir sağlık veri setini analiz etmek, temizlemek ve modellemeye hazır hâle getirmektir.

İş Akışı Özeti:

1. **Veri Yükleme ve Tanıma:** Veri seti yüklenir ve temel bilgileri (boyut, sütunlar, veri tipleri) ekrana basılır.
2. **Kapsamlı Keşifsel Veri Analizi (EDA):** Veri setinin yapısını anlamak için bir dizi analiz ve görselleştirme yapılır.
3. **Otomatik EDA Raporu:** ydata-profiling kütüphanesi ile tüm veri setinin detaylı bir HTML raporu otomatik olarak oluşturulur.
4. **Veri Ön İşleme:** EDA'dan elde edilen bulgular ışığında veri temizlenir, dönüştürülür ve modellemeye hazır hale getirilir.
5. **Sonuçların Kaydedilmesi:** İşlenmiş veri, processed_data.csv adıyla yeni bir dosyaya kaydedilir.

Keşifsel Veri Analizi (EDA)

Her sütundaki eksik değer sayısı ve yüzdesi hesaplandı.Eksik değerler grafik ve tablolarla görselleştirildi.Veri kalitesini değerlendirmek ve hangi sütunlarda eksik veri temizleme veya doldurma yapılacağını belirlemek için önemli bir adımdı**.**

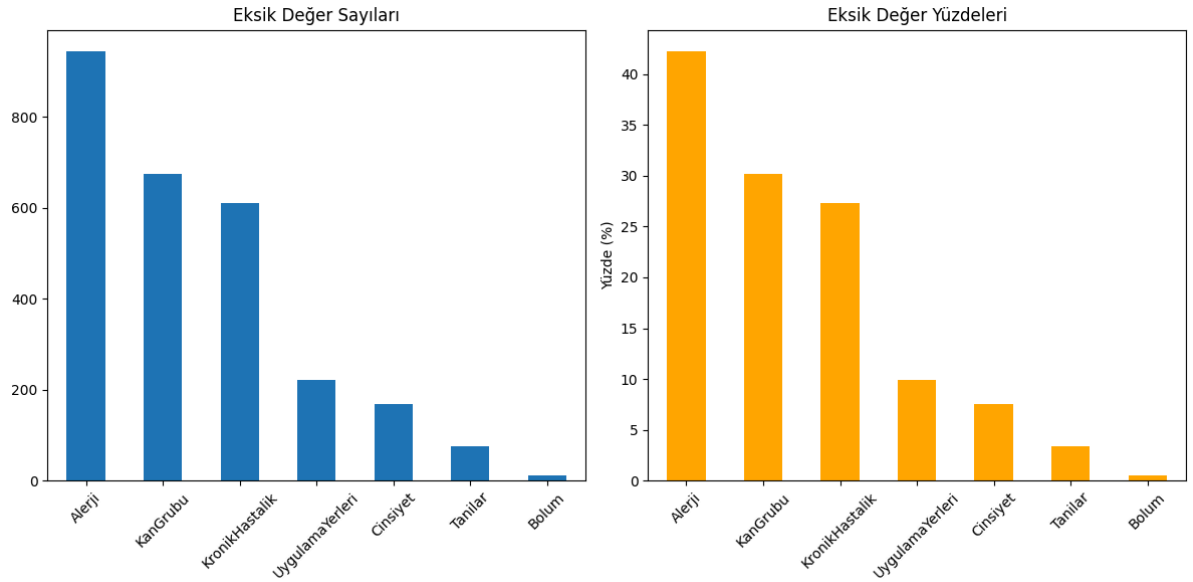
Bu adım sonrası kod çıktıları ve grafikler şu şekildedir.

```
✓ Veri yüklendi: 2235 gözlem, 13 özellik
============================================================
TEMEL VERİ BİLGİLERİ
============================================================
Boyut: 2235 satır x 13 sütun

Sütunlar:
   1. HastaNo                  (int64)
   2. Yas                      (int64)
   3. Cinsiyet                 (object)
   4. KanGrubu                 (object)
   5. Uyruk                    (object)
   6. KronikHastalik           (object)
   7. Bolum                    (object)
   8. Alerji                   (object)
   9. Tanilar                  (object)
  10. TedaviAdi                (object)
  11. TedaviSuresi             (object)
  12. UygulamaYerleri          (object)
  13. UygulamaSuresi           (object)

İlk 5 kayıt:
   HastaNo  Yas Cinsiyet KanGrubu  ...                       TedaviAdi TedaviSuresi UygulamaYerleri
UygulamaSuresi
0   145134   60    Kadın    0 Rh+  ...                     Ayak Bileği      5 Seans     Ayak Bileği      20
Dakika
1   145135   28    Erkek    0 Rh+  ...  Dorsalji -Boyun+trapez+skapular     15 Seans           Boyun      20
Dakika
2   145135   28    Erkek    0 Rh+  ...  Dorsalji -Boyun+trapez+skapular     15 Seans      Boyun,Sırt      20
Dakika
3   145135   28    Erkek    0 Rh+  ...  Dorsalji -Boyun+trapez+skapular     15 Seans           Boyun       5
Dakika
4   145135   28    Erkek    0 Rh+  ...  Dorsalji -Boyun+trapez+skapular     15 Seans      Boyun,Sırt      20
Dakika

[5 rows x 13 columns]

==================================================
EKSİK DEĞER ANALİZİ
==================================================
                Eksik_Sayı  Eksik_Yüzde
Cinsiyet               169         7.56
KanGrubu               675        30.20
KronikHastalik         611        27.34
Bolum                   11         0.49
Alerji                 944        42.24
Tanilar                 75         3.36
UygulamaYerleri        221         9.89
2025-09-06 21:13:30.429 Python[6957:95450] +[CATransaction synchronize] called within transaction
2025-09-06 21:14:00.995 Python[6957:95450] +[CATransaction synchronize] called within transaction

==================================================
TEKRARLI KAYIT ANALİZİ
==================================================
Toplam tekrar eden kayıt sayısı: 928
Aynı Hasta + Aynı Tedavi tekrar eden kayıtlar:
      HastaNo  Yas Cinsiyet  ... TedaviSuresi          UygulamaYerleri UygulamaSuresi
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
5      145135   28    Erkek  ...     15 Seans                    Boyun      20 Dakika
...       ...  ...      ...  ...          ...                       ...           ...
2230   145536   48    Erkek  ...     15 Seans     Sol El Bilek Bölgesi      10 Dakika
2231   145536   48    Erkek  ...     15 Seans     Sol El Bilek Bölgesi      20 Dakika
2232   145537   33    Kadın  ...     15 Seans  Sol Ayak Bileği Bölgesi      20 Dakika
2233   145537   33    Kadın  ...     15 Seans  Sol Ayak Bileği Bölgesi      15 Dakika
2234   145537   33    Kadın  ...     15 Seans  Sol Ayak Bileği Bölgesi       5 Dakika

[2196 rows x 13 columns]

==================================================
```

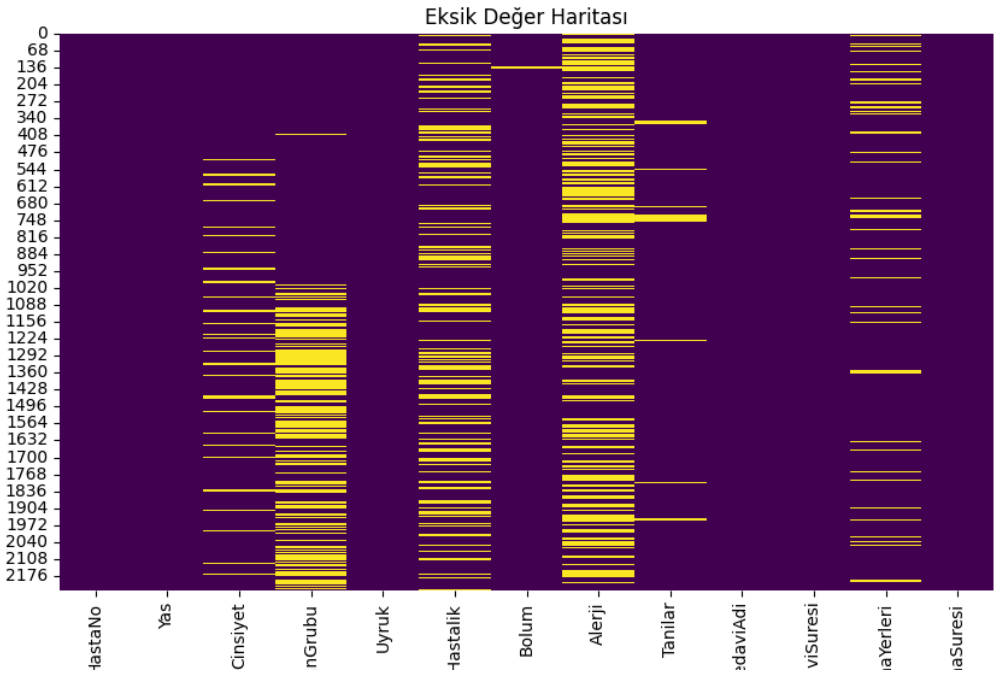
```
HEDEF DEĞİŞKEN ANALİZİ
==================================================
Tedavi Süresi Dağılımı:
TedaviSuresi
1 Seans        3
10 Seans     175
11 Seans       9
14 Seans       2
15 Seans    1670
16 Seans      27
17 Seans      36
18 Seans      20
19 Seans      10
2 Seans       45
20 Seans     113
21 Seans      20
22 Seans       5
25 Seans       5
29 Seans       5
3 Seans        7
30 Seans      12
37 Seans       5
4 Seans       35
5 Seans       17
6 Seans        3
7 Seans        5
8 Seans        6
Name: count, dtype: int64

Tedavi Süresi İstatistikleri:
Ortalama: 14.57 seans
Medyan: 15.00 seans
Standart Sapma: 3.73 seans
Min-Max: 1.0-37.0 seans
2025-09-06 21:14:16.682 Python[6957:95450] +[CATransaction synchronize] called within transaction

==================================================
YAŞ DAĞILIMI ANALİZİ
==================================================
count    2235.000000
mean       47.327069
std        15.208634
min         2.000000
25%        38.000000
50%        46.000000
75%        56.000000
max        92.000000
Name: Yas, dtype: float64
Aykırı değer sayısı: 41
2025-09-06 21:14:33.821 Python[6957:95450] +[CATransaction synchronize] called within transaction

==================================================
KATEGORİK DEĞİŞKENLER ANALİZİ
==================================================

Cinsiyet Dağılımı:
Cinsiyet
Kadın    1274
Erkek     792
Name: count, dtype: int64
Benzersiz değer sayısı: 2

KanGrubu Dağılımı:
KanGrubu
0 Rh+     579
A Rh+     540
B Rh+     206
AB Rh+     80
B Rh-      68
A Rh-      53
0 Rh-      26
AB Rh-      8
```

```
Name: count, dtype: int64
Benzersiz değer sayısı: 8

Uyruk Dağılımı:
Uyruk
Türkiye       2173
Tokelau         27
Arnavutluk      13
Azerbaycan      12
Libya           10
Name: count, dtype: int64
Benzersiz değer sayısı: 5

Bolum Dağılımı:
Bolum
Fiziksel Tıp Ve Rehabilitasyon,Solunum Merkezi    2045
Ortopedi Ve Travmatoloji                            88
İç Hastalıkları                                     32
Nöroloji                                            17
Kardiyoloji                                         11
Göğüs Hastalıkları                                   8
Laboratuar                                           7
Genel Cerrahi                                        6
Tıbbi Onkoloji                                       6
Kalp Ve Damar Cerrahisi                              4
Name: count, dtype: int64
Benzersiz değer sayısı: 10
2025-09-06 21:15:13.968 Python[6957:95450] +[CATransaction synchronize] called within transaction

================================================
KORELASYON ANALİZİ
================================================
2025-09-06 21:15:53.239 Python[6957:95450] +[CATransaction synchronize] called within transaction
Tedavi Süresi ile En Yüksek Korelasyonlar:
Uyruk_encoded       0.102957
KanGrubu_encoded    0.094389
Cinsiyet_encoded    0.022347
Yas                 0.020650
```

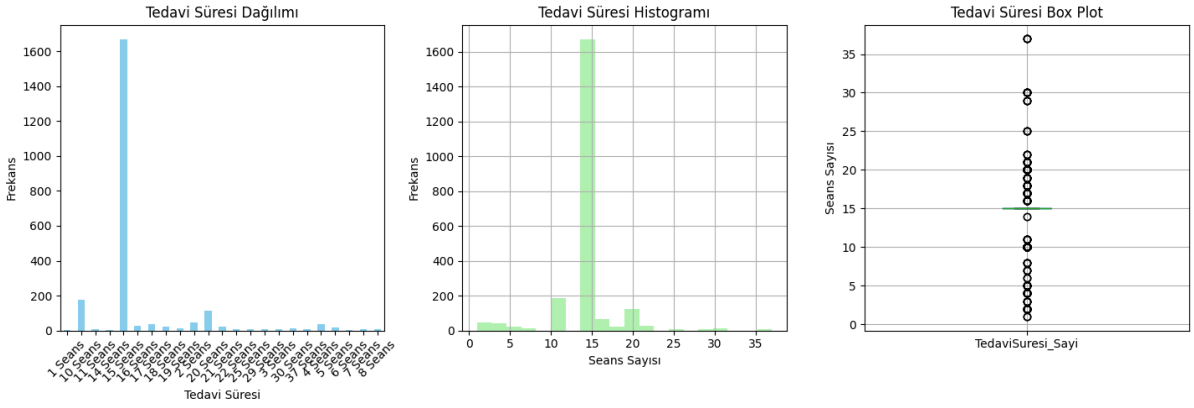
Kod çıktısında genel bilgiler verilmektedir.



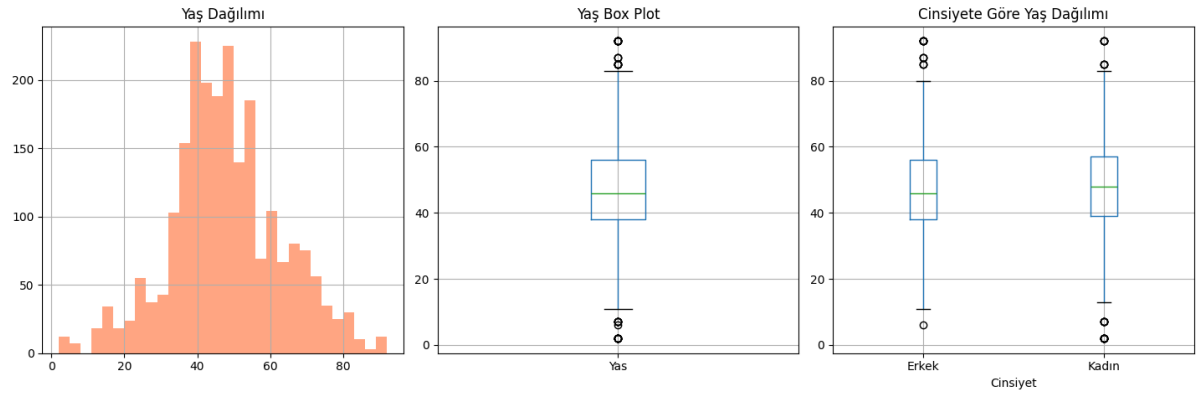
Bu grafik, Kronik Hastalık ve Alerji sütunlarında önemli miktarda eksik veri olduğunu göstermektedir. Kronik Hastalık sütununun yaklaşık %45'i, Alerji sütununun ise yaklaşık %38'i boştur.



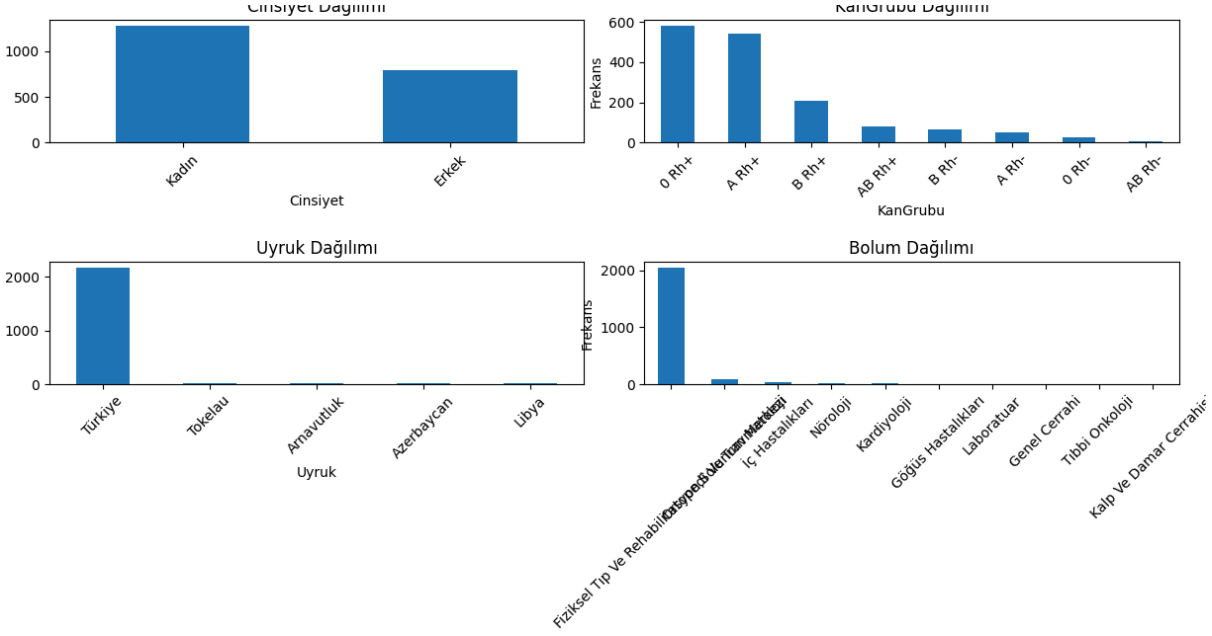
Bu grafik eksikliklerin veri setindeki konumunu haritalandırır. Sarı çizgiler eksik verileri temsil eder.



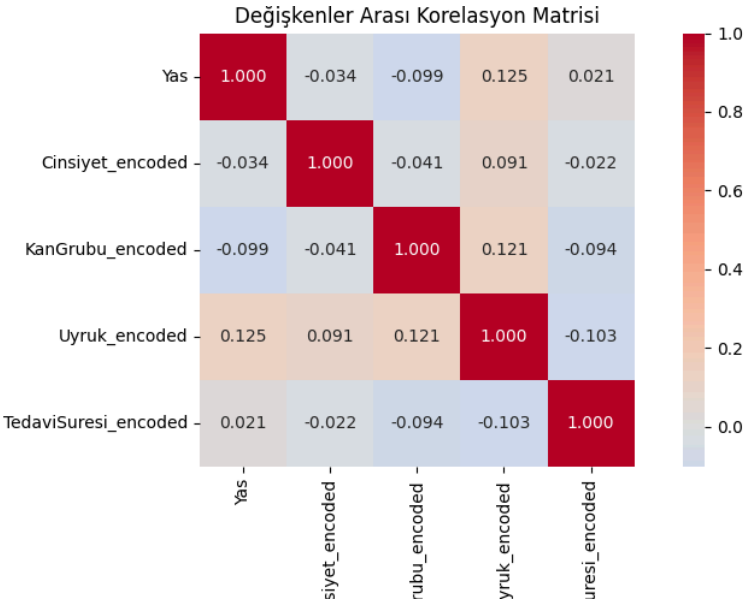
Tedavi süresi histogramı hastaların çoğunluğunun daha kısa süreli tedaviler alırken, az sayıda hastanın çok uzun süren tedaviler aldığını gösterir. Box plot grafiğinde ise aykırı değerler gösterilir. Bu aykırı değerler, ortalamanın üzerinde seans sayısına sahip özel durumları temsil eder.



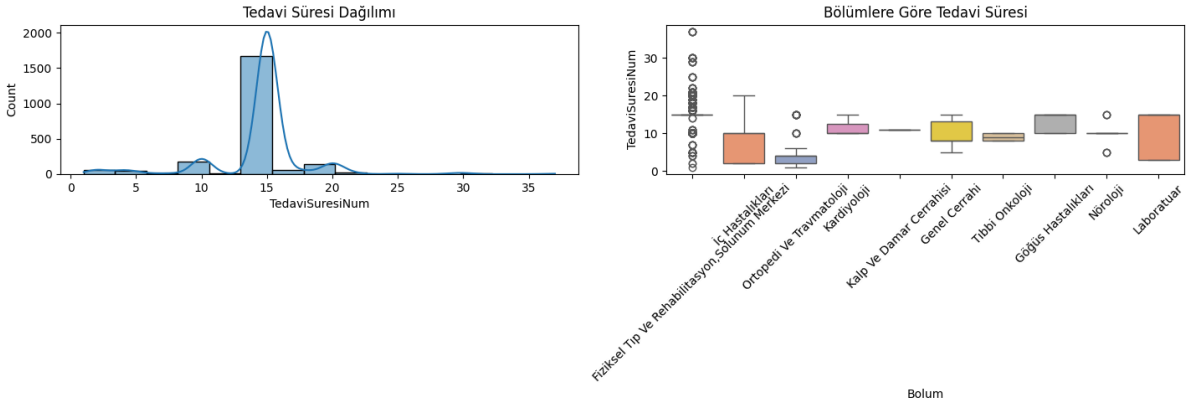
Bu grafiklere göre Hasta yoğunluğunun orta yaşlarda (40-60 yaş arası) daha fazla olduğu görülmektedir.Cinsiyete göre yaş dağılımı incelendiğinde, kadın ve erkek hastaların medyan yaşları ve yaş aralıkları arasında belirgin bir fark olmadığı görülmektedir.



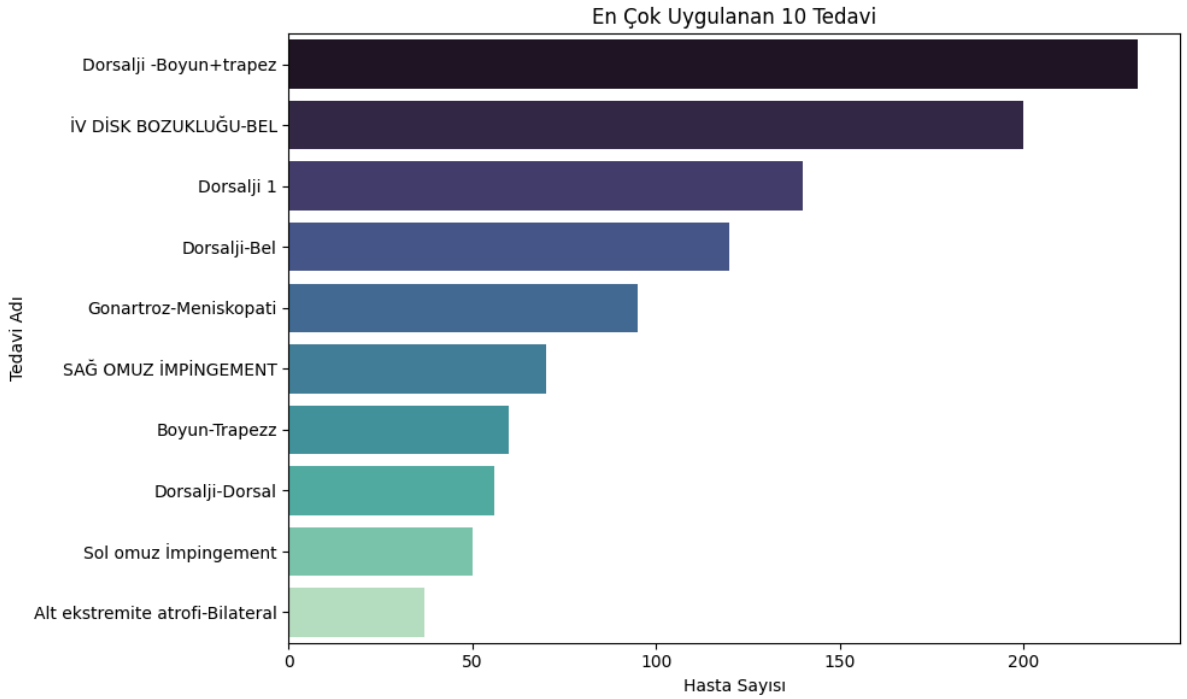
Bu tablolarda ise cinsiyet, kan grubu, uyruk gibi kategorik özelliklerin dağılımı görülmektedir.

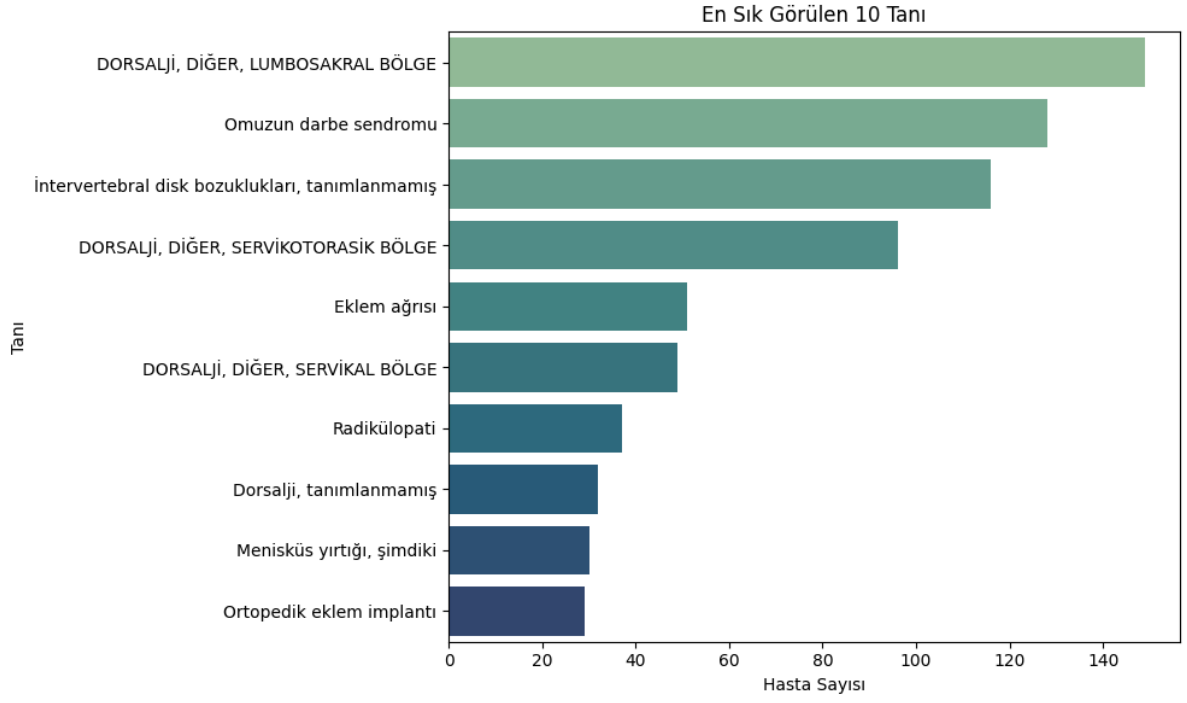


Korelasyon analizi sayısallaştırılmış değişkenler arasındaki doğrusal ilişkiyi ölçmeye yardımcı olur.Seçilen değişkenler arasında güçlü bir doğrusal korelasyon olmadığını göstermektedir. Değerlerin çoğu sıfıra yakındır. Bu, doğrusal bir modelin (örn: Lineer Regresyon) bu değişkenlerle tek başına etkili sonuçlar veremeyebileceğini, daha karmaşık ve doğrusal olmayan ilişkileri yakalayabilen ağaç tabanlı modellerin (örn: Random Forest, Gradient Boosting) daha başarılı olabileceği anlamına gelebilir.



Bu grafik tedavi süresinin başvurulan bölüme göre nasıl değiştiğini analiz etmeyi amaçlar.Bolum özelliğinin Tedavi Süresi'ni tahmin etmede önemli bir gösterge olabileceğini ortaya koymaktadır. Farklı bölümlerin medyan tedavi süreleri ve seans sayısı dağılımları belirgin şekilde farklıdır. Örneğin, "Solunum Merkezi" bölümündeki tedavilerin süresi diğerlerine göre daha geniş bir aralığa yayılmış ve daha yüksek aykırı değerlere sahip gibi görünmektedir.

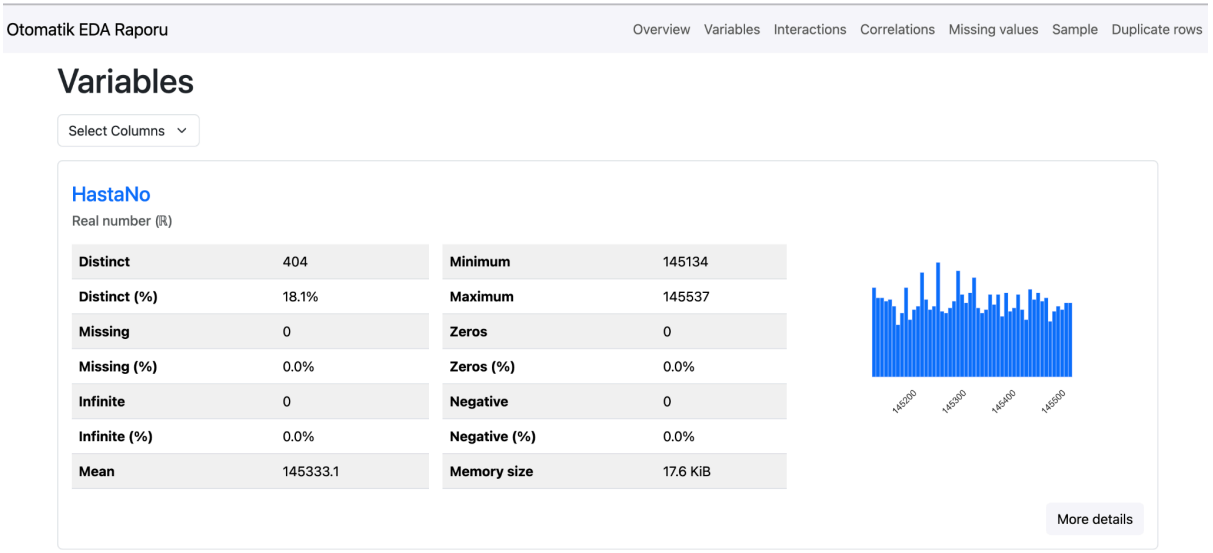


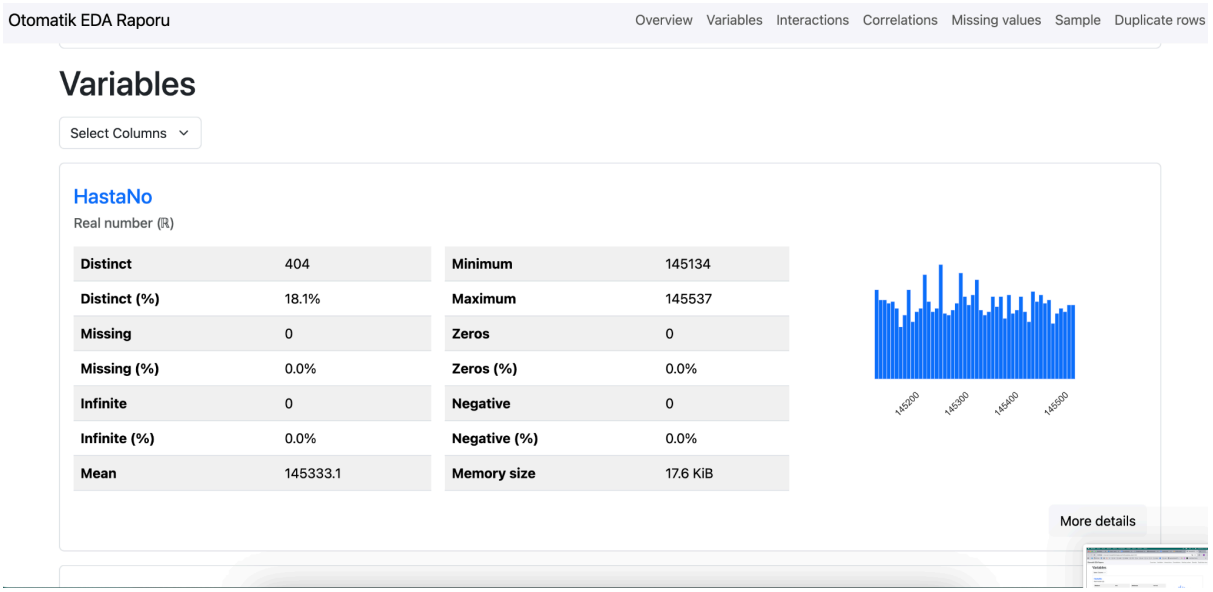


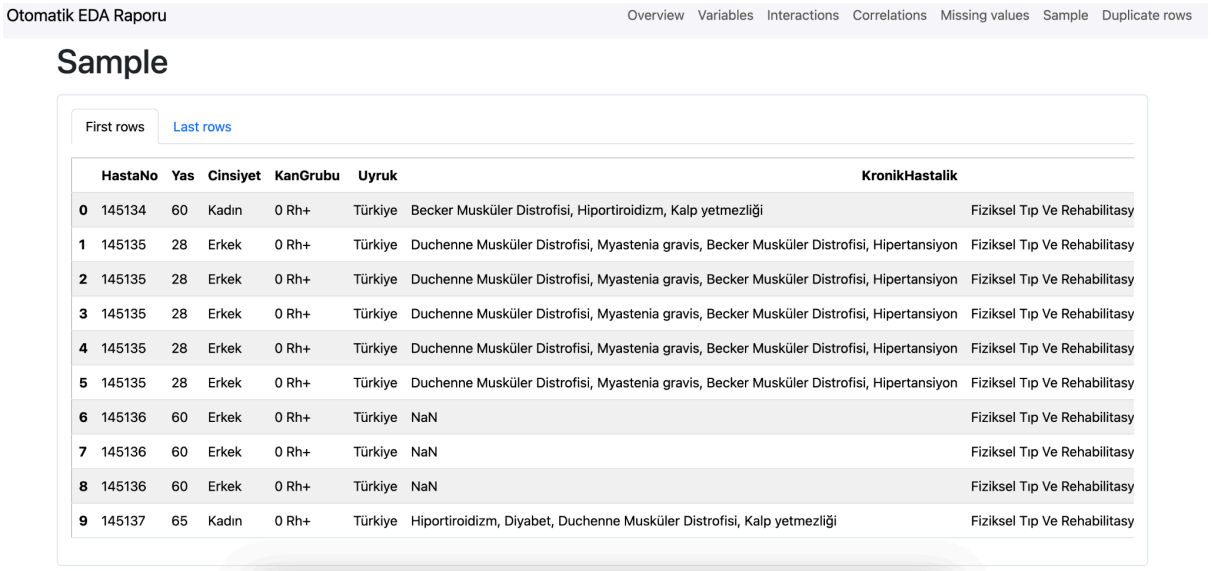
Bu grafikler ise klinikte en sık karşılaşılan durumları ve uygulanan prosedürleri belirlemeyi amaçlar.

EDA HTML Raporu Oluşturma

ProfileReport fonksiyonu ile otomatik bir HTML EDA raporu oluşturuldu. Bu rapor, veri setinin boyutu, sütun bilgileri, eksik değerler, dağılımlar, korelasyonlar ve istatistiksel özetleri detaylı biçimde görselleştirilir. Bu sayede veri analizi sürecinde hızlı ve görsel bir şekilde veri setinin genel durumu incelenebildi. Aşağıda eda_report.html sayfasından örnekler bulunmaktadır.







Otomatik EDA Raporu

Overview Variables Interactions Correlations Missing values Sample Duplicate rows

Sample

First rows Last rows

	HastaNo	Yas	Cinsiyet	KanGrubu	Uyruk	KronikHastalik	
0	145134	60	Kadın	0 Rh+	Türkiye	Becker Musküler Distrofisi, Hiportiroidizm, Kalp yetmezliği	Fiziksel Tıp Ve Rehabilitasy
1	145135	28	Erkek	0 Rh+	Türkiye	Duchenne Musküler Distrofisi, Myastenia gravis, Becker Musküler Distrofisi, Hipertansiyon	Fiziksel Tıp Ve Rehabilitasy
2	145135	28	Erkek	0 Rh+	Türkiye	Duchenne Musküler Distrofisi, Myastenia gravis, Becker Musküler Distrofisi, Hipertansiyon	Fiziksel Tıp Ve Rehabilitasy
3	145135	28	Erkek	0 Rh+	Türkiye	Duchenne Musküler Distrofisi, Myastenia gravis, Becker Musküler Distrofisi, Hipertansiyon	Fiziksel Tıp Ve Rehabilitasy
4	145135	28	Erkek	0 Rh+	Türkiye	Duchenne Musküler Distrofisi, Myastenia gravis, Becker Musküler Distrofisi, Hipertansiyon	Fiziksel Tıp Ve Rehabilitasy
5	145135	28	Erkek	0 Rh+	Türkiye	Duchenne Musküler Distrofisi, Myastenia gravis, Becker Musküler Distrofisi, Hipertansiyon	Fiziksel Tıp Ve Rehabilitasy
6	145136	60	Erkek	0 Rh+	Türkiye	NaN	Fiziksel Tıp Ve Rehabilitasy
7	145136	60	Erkek	0 Rh+	Türkiye	NaN	Fiziksel Tıp Ve Rehabilitasy
8	145136	60	Erkek	0 Rh+	Türkiye	NaN	Fiziksel Tıp Ve Rehabilitasy
9	145137	65	Kadın	0 Rh+	Türkiye	Hiportiroidizm, Diyabet, Duchenne Musküler Distrofisi, Kalp yetmezliği	Fiziksel Tıp Ve Rehabilitasy

Veri Ön İşleme

1. Eksik Değer Temizleme

clean_missing_values fonksiyonu ile veri setindeki eksik değerler işlendi. Boş stringler NaN ile değiştirildi ve ardından Kronik Hastalık, Alerji ve Uygulama Yerleri sütunlarındaki eksik değerler sırasıyla "Yok", "Yok" ve "Belirtilmemiş" olarak dolduruldu. Bu adım, modelleme sırasında eksik verilerden kaynaklanabilecek hataların önüne geçilmesini sağladı.

2. Sayısal Özellik Çıkarımı

extract_numerical_features fonksiyonu ile bazı sütunlardan sayısal değerler elde edildi. TedaviSuresi ve UygulamaSuresi metin sütunlarından yalnızca sayı değerleri çekildi ve TedaviSuresi_Sayi ile UygulamaSuresi_Sayi sütunları oluşturuldu. KronikHastalik ve Tanilar sütunlarındaki çoklu değerlerin sayısı ayrı sütunlarda (KronikHastalik_Sayisi ve Tanilar_Sayisi) tutuldu. Ayrıca Yas ile TedaviSuresi_Sayi kullanılarak Yas_TedaviSuresi_Ratio sütunu üretildi ve yaş ile tedavi süresi arasındaki oranlar veri setine eklendi.Bu, modele daha fazla nicel bilgi sağlamak için yapılmıştır.

3. Kategorik Özellikleri Kodlama ve Sayısallaştırma

encode_categorical_features fonksiyonu ile kategorik sütunlar sayısal hâle getirildi. Cinsiyet ve Uyruk sütunları LabelEncoder ile etiketlendi, KanGrubu sütunu ise one-hot encoding yöntemi ile ayrıştırıldı. Alerji ve KronikHastalik sütunları için varlık-yokluk bilgisi 0 ve 1 değerleri ile çıkarıldı.

4. Sayısal Özellikleri Standartlaştırma
scale_numerical_features fonksiyonu ile sayısal sütunlar standartlaştırıldı. StandardScaler kullanılarak farklı ölçeklerdeki sayısal değişkenler, ortalaması 0 ve standart sapması 1 olacak şekilde StandardScaler ile standartlaştırılmıştır. Bu, modelin daha hızlı ve kararlı çalışmasına ve farklı büyüklüklerdeki değişkenlerin modele orantısız etkisi önlenmesine yardımcı olur.

Tüm bu işlemler, run_pipeline fonksiyonu içinde ardışık olarak çalıştırılarak veri setini eksiksiz, sayısal, modellemeye hazır hâle getirdi. Veri yüklendikten sonra temel bilgiler (show_basic_info), eksik değer analizi (analyze_missing_values), tekrarlı kayıt kontrolü (analyze_duplicates), hedef değişken analizi analyze_target_variable, yaş dağılımı analizi (analyze_age_distribution), kategorik değişken analizi (analyze_categorical_features) ve korelasyon analizi (analyze_correlations) ile detaylı şekilde incelendi. Temizlenmiş, sayısallaştırılmış ve standartlaştırılmış veri CSV dosyası olarak kaydedildi.