EDA ve Preprocessing Raporu

1. Veri Seti Bilgisi

- Satır/sütun sayısı: 2235 x 13
- Sayısal değişkenler: `Yas`, `TedaviSuresi`, `UygulamaSuresi`
- Kategorik değişkenler: `Cinsiyet`, `KanGrubu`, `Uyruk`, `Bolum`, `Alerji`, `Tanilar`, `TedaviAdi`, `UygulamaYerleri`

Gözlem: Veri setinde çoğunlukla Türk hastalara ait bilgiler bulunmakta ve sayısal değişkenler tedavi ile ilgili süreleri temsil ediyor.

2. Eksik Değer Analizi

- `Cinsiyet`: %7.5 eksik
- `KanGrubu`: %30 eksik
- `KronikHastalik`: %27 eksik
- `Alerji`: %42 eksik
- `Tanilar`: %3.3 eksik
- `UygulamaYerleri`: %9.8 eksik

Gözlem: En çok eksik değer `Alerji` değişkeninde bulunuyor, veri ön işleme sırasında uygun stratejilerle doldurulduu.

3. Aykırı Değerler

- `Yas`: 41 aykırı değer → Winsorize ile sınırlandırıldı
- `TedaviSuresi`: 565 aykırı değer → clip yöntemiyle sınırlandırıldı
- `UygulamaSuresi`: 12 aykırı değer → clip yöntemiyle sınırlandırıldı

Gözlem: Aykırı değerler, veri dağılımını bozmayacak şekilde düzeltilmiştir.

4. Ön İşleme Adımları

- `TedaviSuresi` ve `UygulamaSuresi` sayısala çevrildi
- Eksik değerler uygun stratejilerle doldurulduu (mod, 'Yok' vb.)
- Metin verilerinden çıkarılan sayısal özellikler:
  - `KronikHastalik_sayisi`
  - `Alerji_sayisi`
  - `Tanilar_sayisi`
- NLP Feature Engineering: Tanılar ve kronik hastalık kelime sıklığı çıkarıldı
- Kategorik değişkenler OneHotEncode edildi
- Sayısal değişkenler StandardScaler ile standartlaştırıldı

Gözlem: Bu adımlar ile veri modellemeye hazır hâle getirildi.

5. EDA ve Görselleştirmeler

- Sayısal değişkenlerin histogram ve boxplotları
- Korelasyon heatmap: `TedaviSuresi` ile en ilişkili değişkenler görselleştirildi
- Encode edilmiş kategorik değişkenlerin frekans grafikleri
- NLP özetleri: En sık geçen 10 tanı ve kronik hastalık kelimesi
- Tedavi Süresi ilişkili görselleştirmeler:
    - Cinsiyet vs `TedaviSuresi`
    - Yaş vs `TedaviSuresi`

Gözlem: Tedavi süresi ile bazı kronik hastalık ve yaş değişkenleri arasında ilişkiler gözlendi.

6. Sonuç

- Veri temiz, eksiksiz ve modele hazır hâle getirildi
- NLP tabanlı özellikler ile metin verisi sayısala dönüştürüldü
- Görselleştirmeler ile `TedaviSuresi` ilişkili değişkenler belirlendi

Gözlem: Bu rapor, ileride modelleme veya istatistiksel analiz için güçlü bir temel sağlar.