# DRGCODES.csv

### RangeIndex: 297 entries, 0 to 296

**Data columns (total 8 columns):**

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | row_id | 297 non-null | int64 |
| 1 | subject_id | 297 non-null | int64 |
| 2 | hadm_id | 297 non-null | int64 |
| 3 | drg_type | 297 non-null | object |
| 4 | drg_code | 297 non-null | int64 |
| 5 | description | 297 non-null | object |
| 6 | drg_severity | 168 non-null | float64 |
| 7 | drg_mortality | 168 non-null | float64 |

## 🗁 Tablo İncelemesi: DRGCODES.csv (Tanımlı Klinik Gruplar)

DRGCODES.csv tablosu, hastaların yatışlarına atanan DRG (Diagnosis Related Group) kodları ve bu kodlara ilişkin klinik şiddet/mortalite skorlarını içerir. Klinik segmentasyon, mortalite riski ve finansal analizler gibi konular için temel referans niteliğindedir.

## 📊 Sütun Açıklamaları ve Proje Bağlantısı

| Sütun Adı | Açıklama | Proje ile Bağlantısı |
|---|---|---|
| row_id | Satır ID | Teknik amaçlı, modelde kullanılmaz |
| subject_id | Hasta ID | Hasta bazlı birleştirme için anahtar |
| hadm_id | Hastane yatışı ID | Kayıt birleştirme |
| drg_type | DRG türü (APR, MS, HCFA) | Klinik segmentasyon |
| drg_code | DRG kodu | Klinik/finansal sınıflandırma |
| description | DRG tanımı | Klinik gruplar hakkında bilgi |
| drg_severity | Şiddet (1-4, yüksek değer daha ağır) | Mortalite/klinik risk tahmini |
| drg_mortality | Mortalite riski (1-4, yüksek değer ölümcül) | Ölüm riski tahmini/modellemesi |

# 📈 Kategorik Dağılımlar

🎯 **DRG_TYPE değeri dağılımı:**

- APR : 168
- MS : 72
- HCFA: 57

🎯 **DRG_SEVERITY değeri dağılımı (168 kayıt):**

- 4.0: 93
- 3.0: 57
- 2.0: 18

🎯 **DRG_MORTALITY değeri dağılımı (168 kayıt):**

- 4.0: 82
- 3.0: 57
- 2.0: 25
- 1.0: 4

## Modelde Kullanılabilir Özellikler (Feature Engineering)

| Özellik | Kullanım Türü | Açıklama |
|---|---|---|
| drg_type | Kategorik | Klinik risk segmentasyonu |
| drg_code | Kategorik | Detaylı sınıflandırma |
| description | Kategorik | Klinik bilgi |
| drg_severity | Sayısal | Şiddet skoru (feature) |
| drg_mortality | Sayısal | Mortalite skoru (feature) |

# ⚠ Sınırlamalar & Eksik Veri

- **drg_severity** ve **drg_mortality** sütunlarının %43,4’ü eksik (129/297). Modelde sadece dolu satırlar kullanılmalı veya doldurma yapılmalı.
- **row_id, subject_id, hadm_id** sadece teknik eşleşme amaçlı kullanılmalı.
- Kategorik sütunlar çoklu unique değer içerdiği için gruplama veya kodlama gerekebilir.

# 💼 Kullanılabilirlik Özeti

| Alan | Modelde Kullan | Gerekçesi |
|---|---|---|
| drg_type | Evet | Klinik ayrım, kategorik |
| drg_code | Evet | Kategorik/feature müh. |
| description | Evet | Kategorik/feature müh. |
| drg_severity | Evet | Sayısal risk skoru |
| drg_mortality | Evet | Sayısal ölüm riski |
| row_id | Hayır | Teknik alan |
| subject_id | Hayır | Birleştirme için |
| hadm_id | Hayır | Birleştirme için |

# 🏁 Sonuç

DRGCODES.csv, klinik risk, mortalite ve hasta gruplandırması için modellemede güçlü bir tablodur. Sadece teknik/ID sütunları hariç tutulmalı, kategorik alanlar uygun şekilde kodlanmalı ve eksik değer yönetimine dikkat edilmelidir.

## [2] D_ICD_PROCEDURES.csv

### RangeIndex: 3882 entries, 0 to 3881

**Data columns (total 4 columns):**

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | row_id | 3882 non-null | int64 |
| 1 | icd9_code | 3882 non-null | object |
| 2 | short_title | 3882 non-null | object |
| 3 | long_title | 3882 non-null | object |

# 🗂️ Tablo İncelemesi: D_ICD

# _PROCEDURES.csv (Tıbbi İşlem Kodları Sözlüğü)

**D_ICD_PROCEDURES.csv** tablosu, MIMIC-III veri tabanında yer alan tüm prosedürlerin (tıbbi işlemlerin) ICD-9 kodları, kısa ve uzun açıklamaları ile birlikte tutulduğu referans sözlüktür.
Her hastaya uygulanan işlem kayıtları ile birleştirilerek, hasta bazında hangi işlemlerin uygulandığı tespit edilir.
Klinik sınıflandırma, risk modellemesi ve işlem sıklığı analizleri için temel kaynak tablodur.

# 📊 Sütun Açıklamaları ve Proje Bağlantısı

| Sütun Adı | Açıklama | Proje ile Bağlantısı |
|---|---|---|
| row_id | Satır ID | Teknik amaçlı, modelde gerekmez |
| icd9_code | ICD-9 prosedür kodu | Klinik işlem türü (feature) |
| short_title | Kısa işlem adı | Kategorik öznitelik/analiz |
| long_title | Uzun işlem açıklaması | Kategorik/metin özniteliği |

# 📈 Kategorik Değer Dağılımları

- **icd9_code:** Yüzlerce benzersiz işlem kodu barındırır.
- **short_title:** Sıklıkla tekrar eden kategorik başlıklar (ör. “ENDOTRACHEAL TUBE”, “ARTERIAL CATH” gibi).
- **long_title:** İşlemin uzun açıklaması, metinsel analizler için uygundur.

# Modelde Kullanılabilir Özellikler (Feature Engineering)

| Özellik | Tür | Açıklama |
|---|---|---|
| icd9_code | Kategorik | Yapılan işlem türü (binary/one-hot encoding ile) |
| short_title | Kategorik | İşlem grupları |
| long_title | Kategorik | Detaylı işlem tanımı, metinsel analizde eklenebilir |

- Hasta bazında **prosedür çeşitliliği, işlem sayısı, belirli işlem var/yok** gibi derived feature’lar üretilebilir.
- Özellikle bazı işlemler (örn. solunum desteği, cerrahi, invaziv işlemler) sepsis veya ölüm riskiyle ilişkili olabilir.

# ⚠ Eksik Veri ve Sınırlamalar

- Tüm alanlar **tam dolu** (eksik yok).
- Yalnızca **row_id** sütunu teknik amaçlıdır, modelde kullanılmaz.
- **Kategorik sütunlar** (icd9_code, short_title, long_title) çoklu benzersiz değer içerdiği için dummy encoding veya grouping gerekebilir.
- Bu tablo **referans sözlüğüdür**; doğrudan model datası değildir, hasta-prosedür eşleşme tabloları ile birlikte kullanılmalıdır.

## 💼 Kullanılabilirlik Özeti

| Alan | Modelde Kullan | Gerekçesi |
|---|---|---|
| icd9_code | Evet | Klinik işlem türü, kategorik feature |
| short_title | Evet | İşlem grubu, kategorik |
| long_title | Evet | Detaylı açıklama, metin analizi için |
| row_id | Hayır | Sadece teknik/indeks amaçlı |

## 🏁 Sonuç

D_ICD_PROCEDURES.csv, klinik prosedürlerin türünü ve çeşitliliğini anlamak için güçlü bir referans tablosudur.
Risk modellemesi ve hasta profili çıkarmada, yapılan işlemlerin etkisini analiz etmek için **diğer hasta-prosedür tabloları** ile birleştirilerek kullanılmalıdır.
Tek başına modelde feature olarak yer almaz; ancak yapılan işlemlerin **varlığı/sıklığı/çeşidi** derived feature olarak modellenebilir.

# D_ICD_DIAGNOSES.csv

### RangeIndex: 14567 entries, 0 to 14566

**Data columns (total 4 columns):**

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | row_id | 14567 non-null | int64 |
| 1 | icd9_code | 14567 non-null | object |
| 2 | short_title | 14567 non-null | object |
| 3 | long_title | 14567 non-null | object |

## 🗂️ Tablo İncelemesi: D_ICD_DIAGNOSES.csv (Tanı Kodları Sözlüğü)

D_ICD_DIAGNOSES.csv, MIMIC-III veri tabanında yer alan **tüm tanıların** ICD-9 kodları, kısa ve uzun açıklamaları ile birlikte tutulduğu referans tablodur.
Hasta-tanı eşleşmelerinde, **ICD-9 kodlarının ne anlama geldiği** bu tablodan çekilir. Klinik risk sınıflandırması, hastalık grupları analizi ve etiket mühendisliği için temel kaynaktır.

# 📊 Sütun Açıklamaları ve Proje Bağlantısı

| Sütun Adı | Açıklama | Proje ile Bağlantısı |
|---|---|---|
| row_id | Satır ID | Teknik amaçlı, modelde gerekmez |
| icd9_code | ICD-9 tanı kodu | Klinik tanı grubu (feature) |
| short_title | Kısa tanı adı | Kategorik öznitelik/analiz |
| long_title | Uzun tanı açıklaması | Kategorik/metin özniteliği |

# 📈 Kategorik Değer Dağılımları

- **icd9_code:** 14567 farklı kod, büyük kısmı tekil.
- **short_title:** 14567 benzersiz kısa başlık.
- **long_title:** 14567 benzersiz uzun başlık.
- *Her tanı kodu yalnızca bir kez geçiyor. (Tekil referans sözlüğü.)*

### En Sık Geçen İlk 5 Tanı Kodu ve Kısa Başlık

- Her tanı kodu ve başlık **sadece bir kez geçiyor** (tekil referans sözlüğü olduğu için).
- Örnek:

| ICD-9 Kodu | İngilizce Açıklama | Türkçesi |
|---|---|---|
| 0010 | Cholera due to Vibrio cholerae | Vibrio cholerae kaynaklı kolera |
| 0011 | Cholera due to Vibrio eltor | Vibrio eltor kaynaklı kolera |
| 0019 | Cholera NOS | Kolera, başka şekilde tanımlanmamış (NOS) |
| 0020 | Typhoid fever | Tifo (Salmonella Typhi enfeksiyonu) |
| 0021 | Paratyphoid fever A | Paratifo A (Salmonella Paratyphi A) |

# Modelde Kullanılabilir Özellikler (Feature Engineering)

| Özellik | Tür | Açıklama |
|---|---|---|
| icd9_code | Kategorik | Klinik tanı grubu (binary/dummy encoding) |
| short_title | Kategorik | Tanı adı (feature/etiket mühendisliği) |
| long_title | Kategorik | Detaylı tanı açıklaması, metin analizi |

# ⚠ Eksik Veri ve Sınırlamalar

- Tüm sütunlar **eksiksiz** (14567/14567 dolu).
- **row_id** sadece teknik amaçlıdır, modelde kullanılmaz.

- Tablo **referans sözlüğü** olduğu için tek başına feature değildir, hasta-tanı eşleşme tabloları ile birlikte kullanılır.
- **Çok fazla kategori:** Kategorik kodlar binlerce benzersiz değer içerir, gruplama/kümeleme gerekebilir.

## 💼 Kullanılabilirlik Özeti

| Alan | Modelde Kullan | Gerekçesi |
|---|---|---|
| icd9_code | Evet | Klinik tanı grubu, kategorik feature |
| short_title | Evet | Tanı adı, kategorik |
| long_title | Evet | Detaylı açıklama, metin analizi için |
| row_id | Hayır | Sadece teknik/indeks amaçlı |

## 🏁 Sonuç

D_ICD_DIAGNOSES.csv, tanı kodları ve açıklamaları ile **hastalık grubu çıkarımı ve risk segmentasyonu için vazgeçilmez bir referans tablosudur**.
Doğrudan feature olarak kullanılmaz; ancak hasta-tanı eşleşme tabloları ile birleştirilerek modelde tanı bazlı analizlerde kullanılır.
Kategorik çeşitlilik yüksek olduğundan, gruplama veya alt kategori kullanımı tavsiye edilir.

# D_CPT.csv

### RangeIndex: 134 entries, 0 to 133

**Data columns (total 4 columns):**

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | row_id | 134 non-null | int64 |
| 1 | code | 134 non-null | object |
| 2 | short_title | 134 non-null | object |
| 3 | long_title | 134 non-null | object |

# 📂 Tablo İncelemesi: D_CPT.csv (CPT İşlem Kodları Sözlüğü)

D_CPT.csv, MIMIC-III ve benzeri sağlık veri tabanlarında kullanılan **Current Procedural Terminology (CPT)** işlem kodlarının kısa ve uzun açıklamalarının bulunduğu referans tablosudur. Tıbbi hizmetlerin faturalandırılması, klinik prosedür grupları ve işlem bazlı analizler için temel kaynak niteliğindedir.

# 📊 Sütun Açıklamaları ve Proje Bağlantısı

| Sütun Adı | Açıklama | Proje ile Bağlantısı |
|---|---|---|
| row_id | Satır ID | Teknik amaçlı, modelde gerekmez |
| code | CPT kodu | Klinik işlem türü (feature) |
| short_title | Kısa işlem adı | Kategorik öznitelik/analiz |
| long_title | Uzun işlem açıklaması | Kategorik/metin özniteliği |

# 📈 Kategorik Değer Dağılımları

- **code:** 134 farklı CPT kodu mevcut. Her biri genellikle tekil ve benzersizdir.
- **short_title:** 134 farklı kısa başlık.
- **long_title:** 134 farklı uzun başlık.

### En Sık Geçen İlk 5 CPT Kodu ve Kısa Başlık

- 90791 | Psych diagnostic evaluation
- 90471 | Immunization admin (single)
- 93000 | Electrocardiogram complete
- 36415 | Collection of venous blood
- 99213 | Office/outpatient visit est

## *Modelde Kullanılabilir Özellikler (Feature Engineering)*

| Özellik | Tür | Açıklama |
|---|---|---|
| code | Kategorik | Klinik işlem türü, feature olarak kullanılabilir |
| short_title | Kategorik | Kısa açıklama, işlem grubu olarak |
| long_title | Kategorik | Detaylı açıklama, metin analizi |

## ⚠ *Eksik Veri ve Sınırlamalar*

- **Tüm sütunlar eksiksiz** (134/134 dolu).
- **row_id** yalnızca teknik amaçlıdır, modelde kullanılmaz.
- **Çoklu kategorik değer:** code/short_title/long_title alanları yüksek çeşitlilik içerir; gruplama veya dummy encoding ile analiz yapılmalıdır.
- **Referans tablosudur:** Doğrudan hasta verisi değil, hasta-CPT eşleşme tablolarıyla anlamlı olur.

## 🗂 *Kullanılabilirlik Özeti*

| Alan | Modelde Kullan | Gerekçesi |
|---|---|---|
| code | Evet | Klinik işlem kodu, feature |
| short_title | Evet | İşlem grubu, feature |
| long_title | Evet | Detaylı açıklama |
| row_id | Hayır | Sadece teknik amaçlı |

.

## 🏁 Sonuç

D_CPT.csv, tıbbi işlemler ve hizmetlerin CPT kodları ile tanımlandığı, modellemelerde işlem grubu ve çeşitliliği açısından önemli **referans tablosudur**.
Doğrudan feature olarak değil; hasta-CPT eşleşmesi ile birlikte, örneğin “hastada bu işlem var mı?” veya “toplam kaç CPT işlemi uygulanmış?” gibi türev feature’lar yaratmak için kullanılmalıdır.

## 🗁 Tablo İncelemesi: DIAGNOSES_ICD.csv (Hasta-Tanı Eşleşmeleri)

DIAGNOSES_ICD.csv, hastalara yatış bazında atanan **ICD-9 tanı kodlarının** tutulduğu temel klinik tablodur.
Her satırda bir hastanın bir hastane yatışına ait bir tanı kodu bulunur.
Sepsis, ölüm riski, komplikasyon analizleri gibi modelleme ve analizlerin en kritik klinik girdilerinden biridir.

## 📊 Sütun Açıklamaları ve Proje Bağlantısı

| Sütun Adı | Açıklama | Proje ile Bağlantısı |
|---|---|---|
| row_id | Satır ID | Teknik amaçlı, modelde gerekmez |
| subject_id | Hasta ID | Hasta bazlı birleştirme için anahtar |
| hadm_id | Hastane yatışı ID | Yatış bazında eşleşme için |
| seq_num | Tanının yatış içindeki sırası | Ana tanı / ikincil tanı ayrımı, öncelik analizi |

| Sütun Adı | Açıklama | Proje ile Bağlantısı |
|---|---|---|
| icd9_code | ICD-9 tanı kodu | Klinik tanı grubu, feature olarak önemli |

## 📈 Kategorik Değer Dağılımları

- **subject_id:** 6.517 farklı hasta
- **hadm_id:** 7.886 farklı hastane yatışı
- **icd9_code:** 5.997 farklı tanı kodu
- **seq_num:** 1’den başlayıp hastaya atanan tanı adedine göre artan değerler

### En Sık Geçen İlk 5 Tanı Kodu:

- 41401 | Coronary atherosclerosis (Koroner arter hastalığı)
- 4280 | Congestive heart failure (Konjestif kalp yetmezliği)
- 42731 | Atrial fibrillation (Atriyal fibrilasyon)
- 4019 | Unspecified essential hypertension (Hipertansiyon, tanımlanmamış)
- 5849 | Acute kidney failure, unspecified (Akut böbrek yetmezliği)

## Modelde Kullanılabilir Özellikler (Feature Engineering)

| Özellik | Tür | Açıklama |
|---|---|---|
| subject_id | Kimlik | Diğer tablolara hasta bazlı bağlamak için |
| hadm_id | Kimlik | Yatış bazlı bağlamak için |
| seq_num | Sayısal | Ana tanı (1), ikincil tanı (2,3...) gibi öncelik sıralaması |
| icd9_code | Kategorik | Klinik tanı, binary/one-hot encoding ile feature olabilir |

## ⚠ Eksik Veri ve Sınırlamalar

- Tüm sütunlar eksiksiz (65147/65147 dolu).
- **row_id** teknik amaçlıdır, modelde kullanılmaz.
- Çok sayıda tanı kodu içerir; bu yüzden kod gruplama veya “var/yok” (binary flag) olarak kullanılabilir.
- **Tek başına referans değil**, D_ICD_DIAGNOSES.csv ile birleştirilerek tanı başlığı/metni eklenir.

## 💼 Kullanılabilirlik Özeti

| Alan | Modelde Kullan | Gerekçesi |
|---|---|---|
| subject_id | Evet | Hasta birleştirme, segmentasyon için |
| hadm_id | Evet | Yatış segmentasyonu için |
| seq_num | Evet | Tanı öncelik sırası, feature olarak kullanılabilir |
| icd9_code | Evet | Klinik tanı grubu, outcome üretimi, feature |
| row_id | Hayır | Sadece teknik amaçlı |

## 🏁 Sonuç

DIAGNOSES_ICD.csv, hasta-yatış bazında tanı kodlarını barındıran ve **sepsis başta olmak üzere birçok klinik riski etiketlemek ve modellemek için en kritik tablolardan biridir**.
icd9_code üzerinden doğrudan etiket üretilebilir, “ana tanı/ikincil tanı” ayrımı seq_num ile yapılabilir, diğer sözlük tablosu ile birleştirilerek metinli analiz de yapılabilir.