EXPLORATORY DATA ANALYSIS (EDA) SUMMARY

Dataframe'in ilk ve son beş satırını görüntüleme:

```
In [4]: df.head()
Out[4]:
   Kullanici_id Cinsiyet Dogum_Tarihi    Uyruk        Il                      Ilac_Adi Ilac_Baslangic_Ta
0           107     Male   1960-03-01  Turkiye  Canakkale               trifluoperazine           2022-0
1           140     Male   1939-10-12  Turkiye    Trabzon              fluphenazine hcl           2022-0
2             2   Female   1976-12-17  Turkiye  Canakkale               warfarin sodium           2022-0
3            83     Male   1977-06-17  Turkiye      Adana                 valproic acid           2022-0
4             7   Female   1976-09-03  Turkiye      Izmir  carbamazepine extended release           2022-0
```

```
In [5]: df.tail()
Out[5]:
      Kullanici_id Cinsiyet Dogum_Tarihi    Uyruk        Il                          Ilac_Adi Ilac_Baslang
2352             9      NaN   1957-01-04  Turkiye       NaN  desoximetasone spray, non-aerosol            2
2353           101   Female   2004-11-09  Turkiye    Mersin              olanzapine-fluoxetine            2
2354           127   Female   1951-11-29  Turkiye    Mersin                          trazodone            2
2355           178     Male   1980-01-30  Turkiye   Kayseri          duloxetine hydrochloride            2
2356           174   Female   1986-11-07  Turkiye  Istanbul                       valproic acid            2
```

Dataframe'in boyutunu görüntüleme:

```
In [6]: df.shape
Out[6]: (2357, 19)
```

Dataframe'in bilgilerini görüntüleme:

```
In [7]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2357 entries, 0 to 2356
Data columns (total 19 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Kullanici_id                    2357 non-null   int64
 1   Cinsiyet                        1579 non-null   object
 2   Dogum_Tarihi                    2357 non-null   datetime64[ns]
 3   Uyruk                           2357 non-null   object
 4   Il                              2130 non-null   object
 5   Ilac_Adi                        2357 non-null   object
 6   Ilac_Baslangic_Tarihi           2357 non-null   datetime64[ns]
 7   Ilac_Bitis_Tarihi               2357 non-null   datetime64[ns]
 8   Yan_Etki                        2357 non-null   object
 9   Yan_Etki_Bildirim_Tarihi        2357 non-null   datetime64[ns]
 10  Alerjilerim                     1873 non-null   object
 11  Kronik Hastaliklarim            1965 non-null   object
 12  Baba Kronik Hastaliklari        2201 non-null   object
 13  Anne Kronik Hastaliklari        2140 non-null   object
 14  Kiz Kardes Kronik Hastaliklari  2260 non-null   object
 15  Erkek Kardes Kronik Hastaliklari  2236 non-null   object
 16  Kan Grubu                       2010 non-null   object
 17  Kilo                            2064 non-null   float64
 18  Boy                             2243 non-null   float64
dtypes: datetime64[ns](4), float64(2), int64(1), object(12)
```

Sütun isimlerini görüntüleme:

```
In [8]: df.columns
Out[8]: Index(['Kullanici_id', 'Cinsiyet', 'Dogum_Tarihi', 'Uyruk', 'Il', 'Ilac_Adi', 'Ilac_Baslangic_Tarihi', 'Ilac_Bitis_Tarihi', 'Yan_Etki', 'Yan_Etki_Bildirim_Tarihi',
```

Dataframe'in indeksini görüntüleme:

```
In [9]: df.index
Out[9]: RangeIndex(start=0, stop=2357, step=1)
```

Sayısal verilerin istatistiksel özetini görüntüleme:

```
In [10]: df.describe().T
Out[10]:
```

	count	mean	min	25%	50%	75%	max	std
Kullanici_id	2357.0	97.216801	1.0	47.0	97.0	146.0	196.0	57.0172
Dogum_Tarihi	2357	1974-11-25 04:06:12.677131936	1939-10-12 00:00:00	1959-02-05 00:00:00	1973-09-09 00:00:00	1992-03-24 00:00:00	2011-04-25 00:00:00	NaN
Ilac_Baslangic_Tarihi	2357	2022-01-07 10:47:36.173101312	2022-01-01 00:00:00	2022-01-04 00:00:00	2022-01-07 00:00:00	2022-01-11 00:00:00	2022-01-14 00:00:00	NaN
Ilac_Bitis_Tarihi	2357	2022-03-10 16:25:27.365294848	2022-03-02 00:00:00	2022-03-06 00:00:00	2022-03-11 00:00:00	2022-03-15 00:00:00	2022-03-19 00:00:00	NaN
Yan_Etki_Bildirim_Tarihi	2357	2022-02-10 17:09:30.742044928	2022-02-01 04:34:33	2022-02-04 05:29:20	2022-02-09 20:53:54	2022-02-17 07:08:01	2022-02-19 21:47:39	NaN
Kilo	2064.0	80.863857	50.0	65.0	83.0	96.0	110.0	18.635269
Boy	2243.0	174.638431	145.0	160.0	176.0	187.0	203.0	16.516552

Boş veri kontrolü:

```
In [12]: df.isnull().values.any()
Out[12]: True
```

```
In [13]: df.isnull().sum()
Out[13]:
Kullanici_id                      0
Cinsiyet                        778
Dogum_Tarihi                      0
Uyruk                             0
Il                              227
Ilac_Adi                          0
Ilac_Baslangic_Tarihi             0
Ilac_Bitis_Tarihi                 0
Yan_Etki                          0
Yan_Etki_Bildirim_Tarihi          0
Alerjilerim                     484
Kronik Hastaliklarim            392
Baba Kronik Hastaliklari        156
Anne Kronik Hastaliklari        217
Kiz Kardes Kronik Hastaliklari   97
Erkek Kardes Kronik Hastaliklari 121
Kan Grubu                       347
Kilo                            293
Boy                             114
dtype: int64
```

Kategorik, sayısal ve tarihsel sütunları ayırma:

```
In [14]: cat_cols = [col for col in df.columns if str(df[col].dtypes) in ["object"]]
    ...: num_cols = [col for col in df.columns if str(df[col].dtypes) in ["int64", "float64"]]
    ...: date_cols = [col for col in df.columns if str(df[col].dtypes) in ["datetime64[ns]"]]
    ...:
In [15]: cat_cols
Out[15]:
['Cinsiyet',
 'Uyruk',
 'Il',
 'Ilac_Adi',
 'Yan_Etki',
 'Alerjilerim',
 'Kronik Hastaliklarim',
 'Baba Kronik Hastaliklari',
 'Anne Kronik Hastaliklari',
 'Kiz Kardes Kronik Hastaliklari',
 'Erkek Kardes Kronik Hastaliklari',
 'Kan Grubu']
```

```
In [16]: num_cols
Out[16]: ['Kullanici_id', 'Kilo', 'Boy']
```

```
In [17]: date_cols
Out[17]:
['Dogum_Tarihi',
 'Ilac_Baslangic_Tarihi',
 'Ilac_Bitis_Tarihi',
 'Yan_Etki_Bildirim_Tarihi']
```

DATA PRE-PROCESSING SUMMARY

Kategorik verilerde, veri dağılımını bozmamak için en sık kullanılan değerle doldurma (Mode Imputation) yöntemini seçtim. Metin verilerinde, eksik değerlerin belirgin olması için "None" ekledim. Sayısal verilerde ise, eksik bilgileri daha doğru tahmin etmek için KNN Imputer kullandım.

```
In [18]: from sklearn.impute import SimpleImputer
    ...: from sklearn.impute import KNNImputer
    ...:
In [19]: mode_imputer = SimpleImputer(strategy='most_frequent')
    ...: df[['Cinsiyet', 'Il', 'Kan Grubu']] = mode_imputer.fit_transform(df[['Cinsiyet', 'Il', 'Kan Grubu']])
    ...:
In [20]: df[['Alerjilerim', 'Kronik Hastaliklarim', 'Baba Kronik Hastaliklari',
    ...:     'Anne Kronik Hastaliklari', 'Kiz Kardes Kronik Hastaliklari', 'Erkek Kardes Kronik Hastaliklari']] = df[[
    ...:     'Alerjilerim', 'Kronik Hastaliklarim', 'Baba Kronik Hastaliklari',
    ...:     'Anne Kronik Hastaliklari', 'Kiz Kardes Kronik Hastaliklari', 'Erkek Kardes Kronik Hastaliklari'
    ...: ]].fillna('None')
In [21]: knn_imputer = KNNImputer(n_neighbors=5)
    ...: df[['Kilo', 'Boy']] = knn_imputer.fit_transform(df[['Kilo', 'Boy']])
```

Sonuçları gözden geçirme:

```
In [22]: print(df.head())
   Kullanici_id Cinsiyet Dogum_Tarihi    Uyruk        Il                     Ilac_Adi Ilac_Baslangic_Tarihi Ilac_Bitis_Tarihi
0           107     Male   1960-03-01  Turkiye  Canakkale               trifluoperazine           2022-01-09       2022-03-04
1           140     Male   1939-10-12  Turkiye    Trabzon               fluphenazine hcl           2022-01-09       2022-03-08
2             2   Female   1976-12-17  Turkiye  Canakkale                warfarin sodium           2022-01-11       2022-03-12
3            83     Male   1977-06-17  Turkiye      Adana                  valproic acid           2022-01-04       2022-03-12
4             7   Female   1976-09-03  Turkiye      Izmir  carbamazepine extended release           2022-01-13       2022-03-06  Agi
```

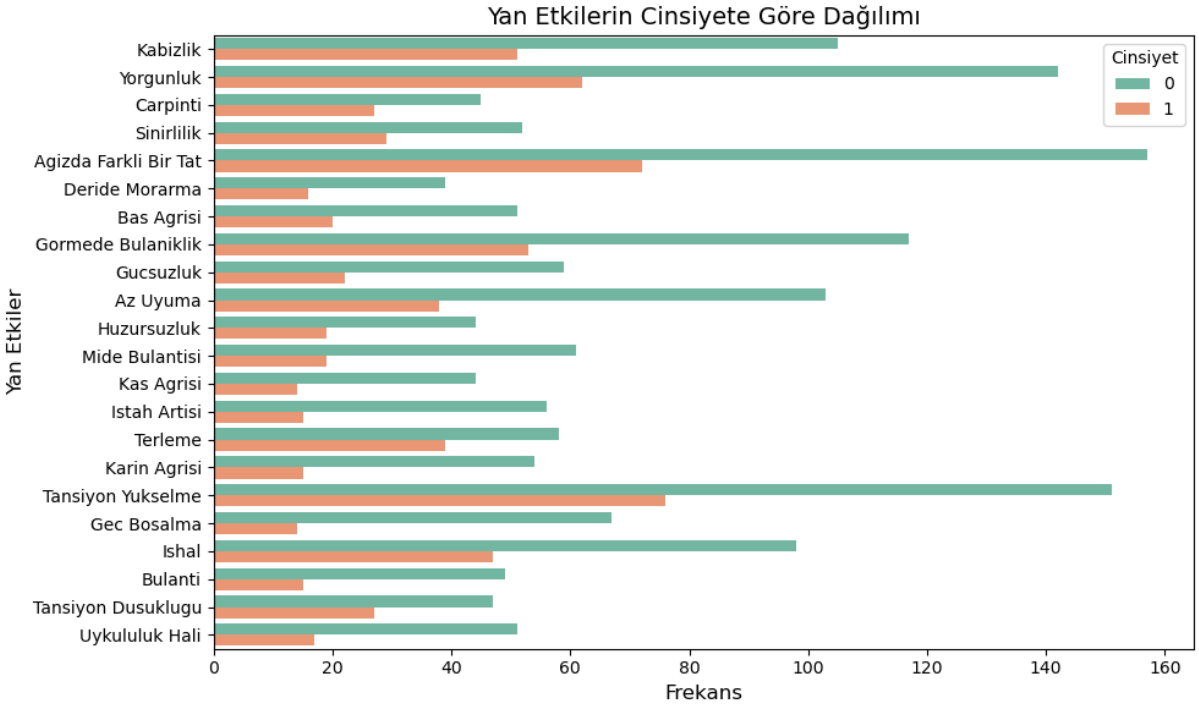
Boş verilerin kontrolü:

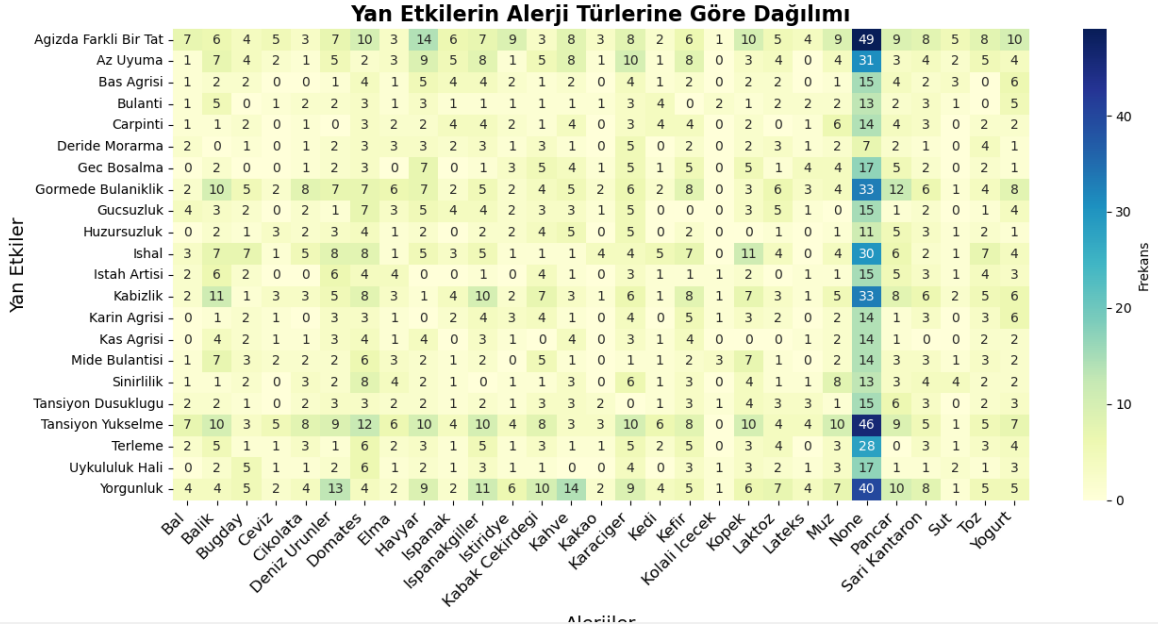
```
In [23]: df.isnull().sum()
Out[23]:
Kullanici_id                    0
Cinsiyet                        0
Dogum_Tarihi                    0
Uyruk                           0
Il                              0
Ilac_Adi                        0
Ilac_Baslangic_Tarihi           0
Ilac_Bitis_Tarihi               0
Yan_Etki                        0
Yan_Etki_Bildirim_Tarihi        0
Alerjilerim                     0
Kronik Hastaliklarim            0
Baba Kronik Hastaliklari        0
Anne Kronik Hastaliklari        0
Kiz Kardes Kronik Hastaliklari  0
Erkek Kardes Kronik Hastaliklari 0
Kan Grubu                       0
Kilo                            0
Boy                             0
dtype: int64
```

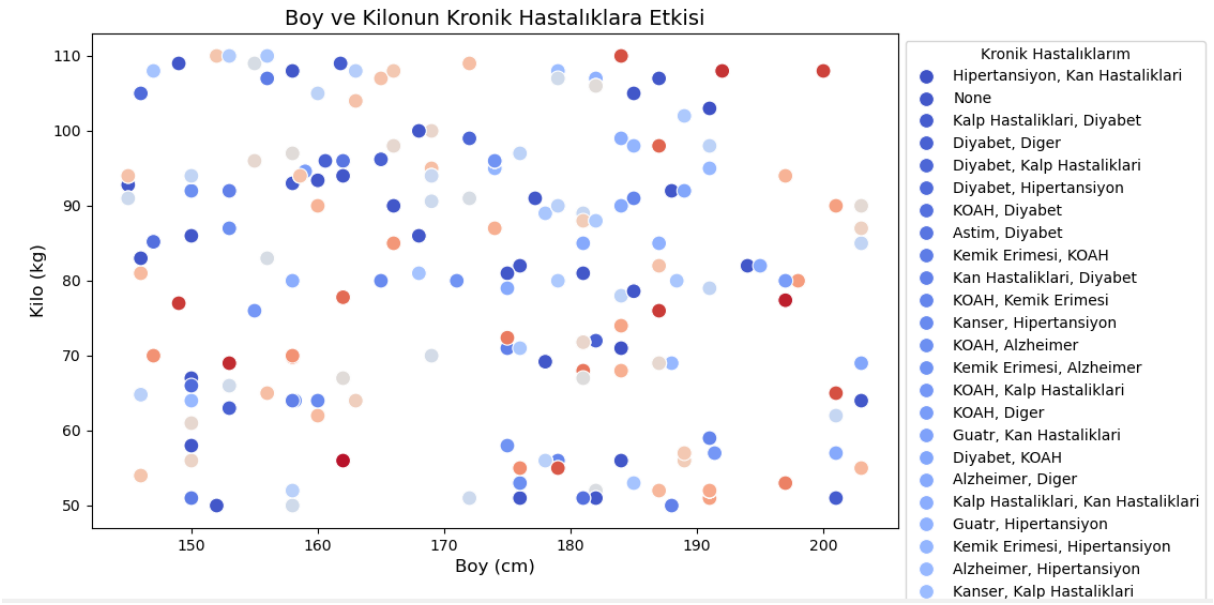
Label Encoding ile “Cinsiyet” gibi kategorik bir sütunu sayısal değerlere çevirdim. Bu sayede makine öğrenmesi algoritmalarında kullanılabilir hale getirmiş oldum. Sonuçlarını da kontrol ettim.

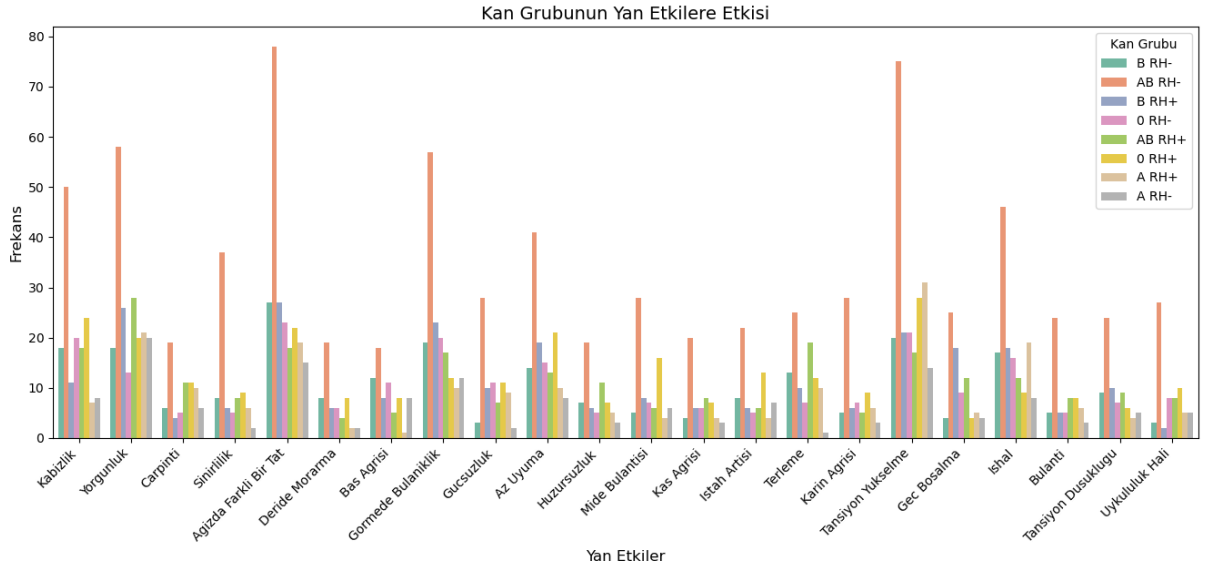
```
In [24]: from sklearn.preprocessing import LabelEncoder
In [25]: label_encoder = LabelEncoder()
    ...: df['Cinsiyet'] = label_encoder.fit_transform(df['Cinsiyet'])
    ...:
In [26]: print(df[['Cinsiyet']].head())
   Cinsiyet
0         1
1         1
2         0
3         1
4         0
In [27]: print(df.head())
   Kullanici_id  Cinsiyet Dogum_Tarihi    Uyruk        Il                     Ilac_Adi Ilac_Baslangic_Tarihi Ilac_Bitis_Tarihi             Yan_Etki Yan
0           107         1   1960-03-01  Turkiye  Canakkale               trifluoperazine           2022-01-09       2022-03-04             Kabizlik
1           140         1   1939-10-12  Turkiye    Trabzon               fluphenazine hcl           2022-01-09       2022-03-08            Yorgunluk
2             2         0   1976-12-17  Turkiye  Canakkale                warfarin sodium           2022-01-11       2022-03-12              Carpinti
3            83         1   1977-06-17  Turkiye      Adana                  valproic acid           2022-01-04       2022-03-12            Sinirlilik
4             7         0   1976-09-03  Turkiye      Izmir  carbamazepine extended release           2022-01-13       2022-03-06  Agizda Farkli Bir Tat
```

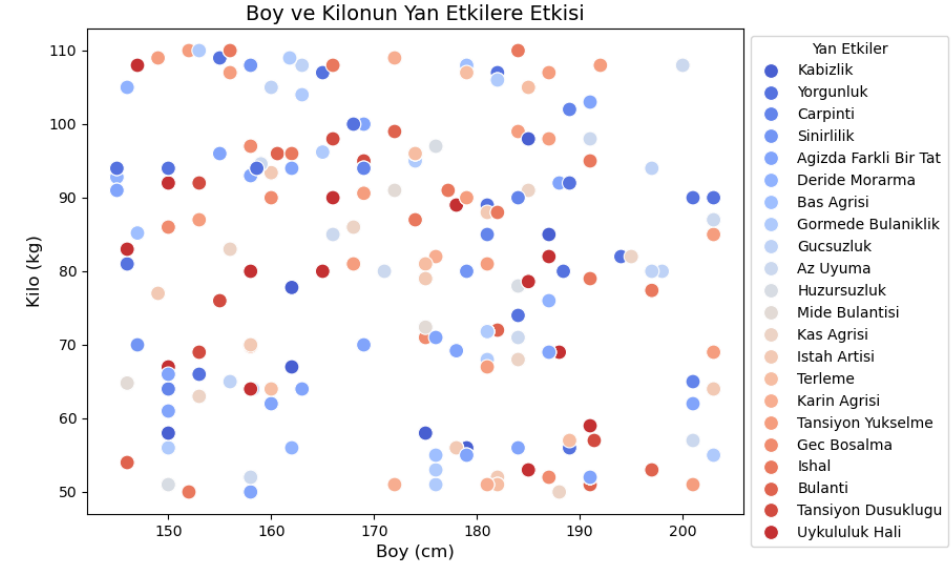
VISUALIZATIONS






Boy ve Kilonun Kronik Hastalıklara Etkisi
Kilo (kg)
Boy (cm)
Kronik Hastalıklarım
Hipertansiyon, Kan Hastaliklari
None
Kalp Hastaliklari, Diyabet
Diyabet, Diger
Diyabet, Kalp Hastaliklari
Diyabet, Hipertansiyon
KOAH, Diyabet
Astim, Diyabet
Kemik Erimesi, KOAH
Kan Hastaliklari, Diyabet
KOAH, Kemik Erimesi
Kanser, Hipertansiyon
KOAH, Alzheimer
Kemik Erimesi, Alzheimer
KOAH, Kalp Hastaliklari
KOAH, Diger
Guatr, Kan Hastaliklari
Diyabet, KOAH
Alzheimer, Diger
Kalp Hastaliklari, Kan Hastaliklari
Guatr, Hipertansiyon
Kemik Erimesi, Hipertansiyon
Alzheimer, Hipertansiyon
Kanser, Kalp Hastaliklari


Kan Grubunun Yan Etkilere Etkisi
Frekans
Yan Etkiler
Kan Grubu
B RH-
AB RH-
B RH+
0 RH-
AB RH+
0 RH+
A RH+
A RH-
Kabizlik
Yorgunluk
Carpinti
Sinirlilik
Agizda Farkli Bir Tat
Deride Morarma
Bas Agrisi
Gormede Bulaniklik
Gucsuzluk
Az Uyuma
Huzursuzluk
Mide Bulantisi
Kas Agrisi
Istah Artisi
Terleme
Karin Agrisi
Tansiyon Yukselme
Gec Bosalma
Ishal
Bulanti
Tansiyon Dusuklugu
Uykululuk Hali


Boy ve Kilonun Yan Etkilere Etkisi
Kilo (kg)
Boy (cm)
Yan Etkiler
Kabizlik
Yorgunluk
Carpinti
Sinirlilik
Agizda Farkli Bir Tat
Deride Morarma
Bas Agrisi
Gormede Bulaniklik
Gucsuzluk
Az Uyuma
Huzursuzluk
Mide Bulantisi
Kas Agrisi
Istah Artisi
Terleme
Karin Agrisi
Tansiyon Yukselme
Gec Bosalma
Ishal
Bulanti
Tansiyon Dusuklugu
Uykululuk Hali