Data Science Intern Case Study

YAĞMUR GÖZEL
yagmurgozel@gmail.com

After thoroughly examining the provided dataset, I started working by converting it into a CSV file.

Step 1

First, we load the necessary libraries and our dataset.

```
In [2]: import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        import warnings
        warnings.filterwarnings("ignore")
```

```
In [3]: df=pd.read_csv("side_effect_data1.csv")
        df
```

Out[3]:

	Kullanici_id	Cinsiyet	Dogum_Tarihi	Uyruk	Il	Ilac_Adi	Ilac_Baslangic_Tarihi	Ilac_Bitis_Tarihi	Yan_Etki	Yan_Etki_Bildirim_Tarihi	Ale
0	107	Male	1960-03-01	Turkiye	Canakkale	trifluoperazine	2022-01-09	2022-03-04	Kabizlik	2022-02-19 18:28:43	
1	140	Male	1939-10-12	Turkiye	Trabzon	fluphenazine hcl	2022-01-09	2022-03-08	Yorgunluk	2022-02-03 20:48:17	

Step 2

At this stage, we will obtain basic information about the dataset.

```
In [6]: df.keys()
```

```
Out[6]: Index(['Kullanici_id', 'Cinsiyet', 'Dogum_Tarihi', 'Uyruk', 'Il', 'Ilac_Adi',
               'Ilac_Baslangic_Tarihi', 'Ilac_Bitis_Tarihi', 'Yan_Etki',
               'Yan_Etki_Bildirim_Tarihi', 'Alerjilerim', 'Kronik Hastaliklarim',
               'Baba Kronik Hastaliklari', 'Anne Kronik Hastaliklari',
               'Kiz Kardes Kronik Hastaliklari', 'Erkek Kardes Kronik Hastaliklari',
               'Kan Grubu', 'Kilo', 'Boy'],
              dtype='object')
```

```
In [7]: print(df.dtypes)
```

```
Kullanici_id                    int64
Cinsiyet                       object
Dogum_Tarihi                   object
```

```
In [9]: print(df.head())
```

```
In [11]: print(df.describe())
```

```
In [12]: print(df.info())
```

Step 3

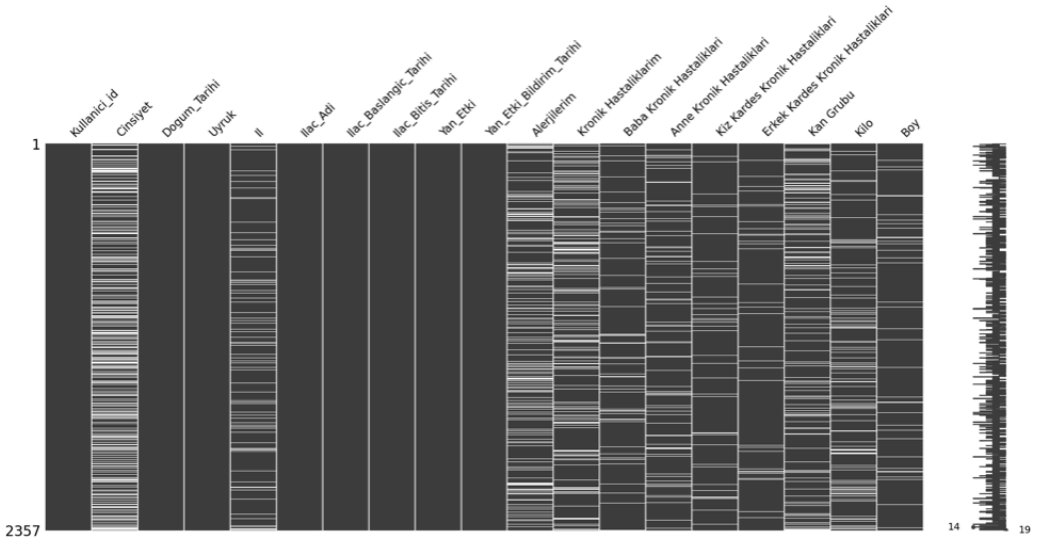
We will check for missing data in the dataset and then use the missingno library to visualize the distribution of missing values. We will analyze which columns have a higher concentration of missing data.

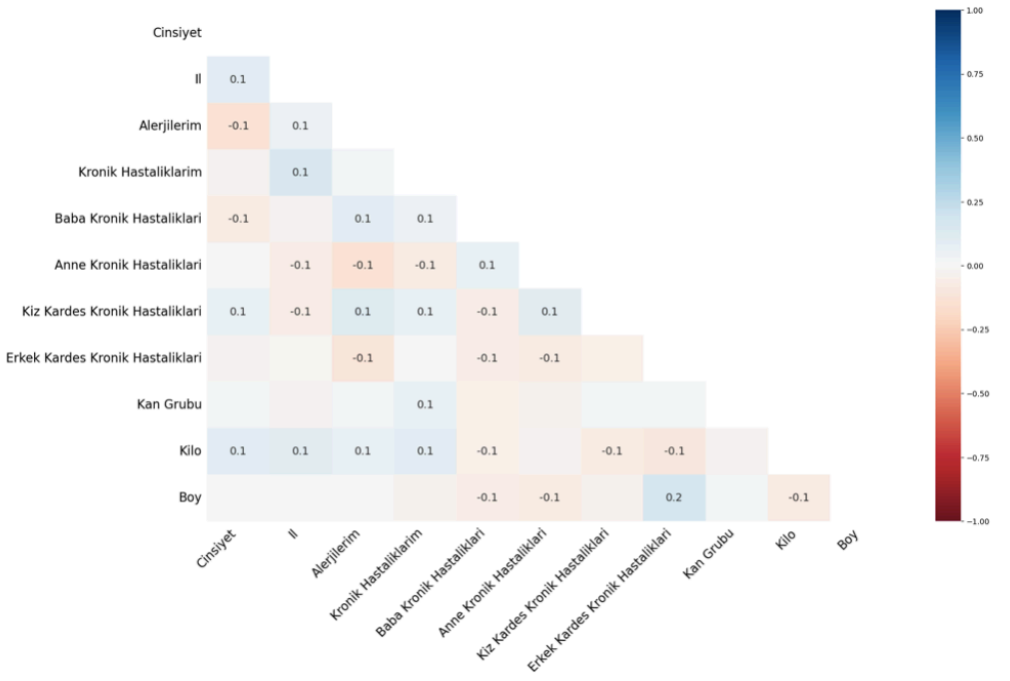
In [16]:
```
print(df.isnull().sum())
```

```
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
```

In [17]:
```
import missingno as msno
msno.matrix(df)
msno.heatmap(df)
```

Out[17]: <Axes: >





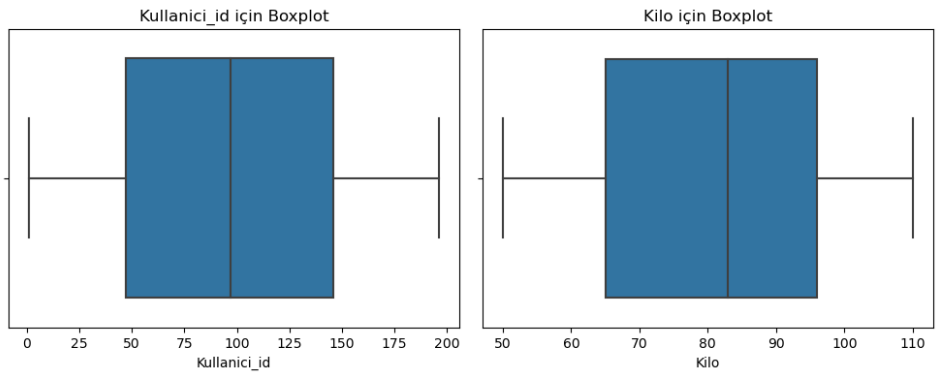
Step 4

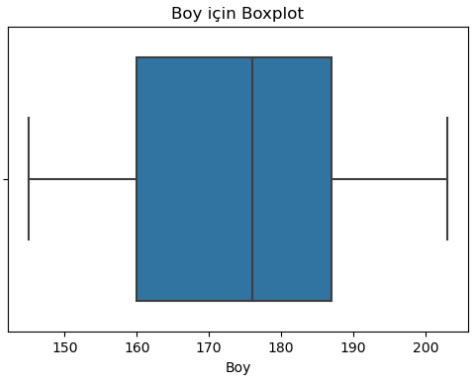
We will examine the categorical and numerical variables in the dataset. While analyzing the distribution of numerical variables, we will use visualizations.

In [18]:
```
categorical_columns = df.select_dtypes(include=['object']).columns

for col in categorical_columns:
    print(f"\n{col} sütunu frekans dağılımı:")
    print(df[col].value_counts())
```

```
Cinsiyet sütunu frekans dağılımı:
Female    872
Male      707
Name: Cinsiyet, dtype: int64

Dogum_Tarihi sütunu frekans dağılımı:
2002-04-15    28
1976-02-20    21
1996-09-10    20
1959-12-19    19
1989-03-06    19
              ..
1951-03-07     6
1965-09-11     6
2003-03-12     5
1992-03-20     5
2003-07-26     3
Name: Dogum_Tarihi, Length: 195, dtype: int64
```

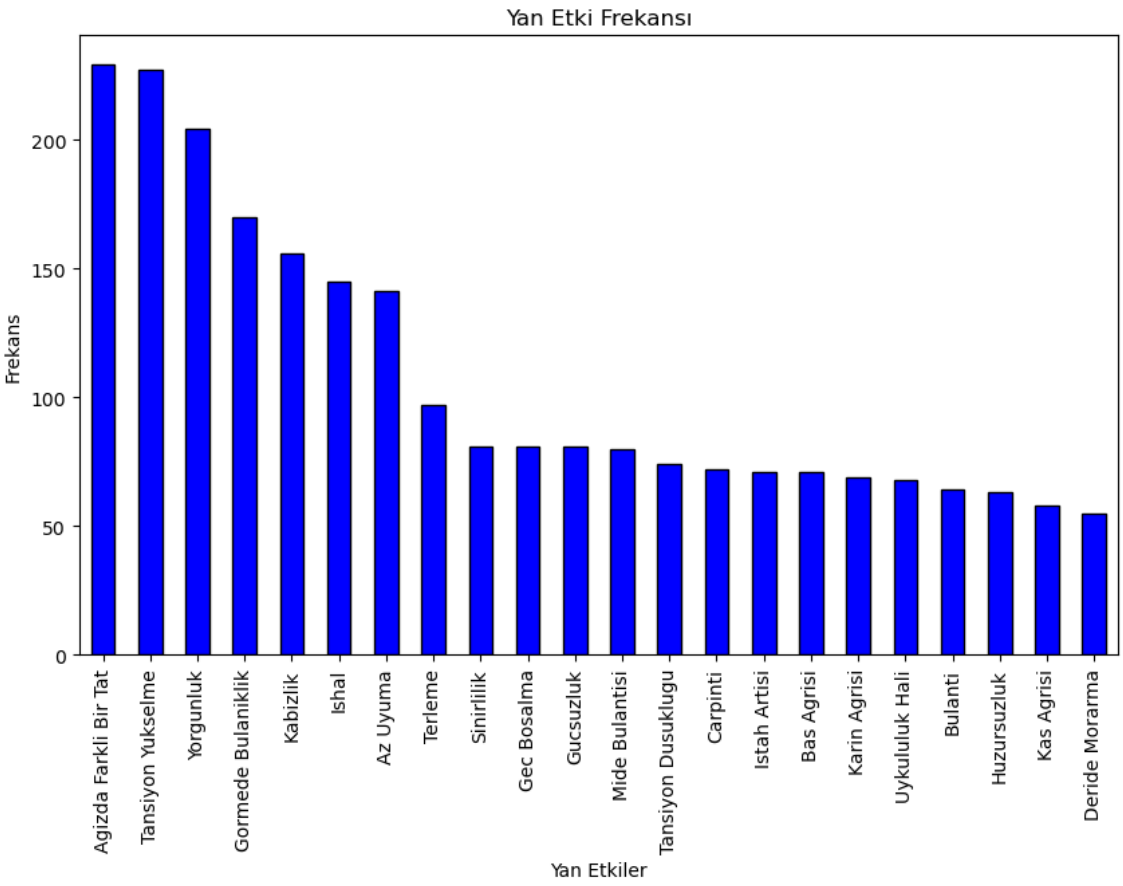
In [23]:
```
import matplotlib.pyplot as plt
import seaborn as sns

for col in numerical_columns:
    plt.figure(figsize=(6, 4))
    sns.boxplot(x=df[col])
    plt.title(f'{col} için Boxplot')
    plt.show()
```



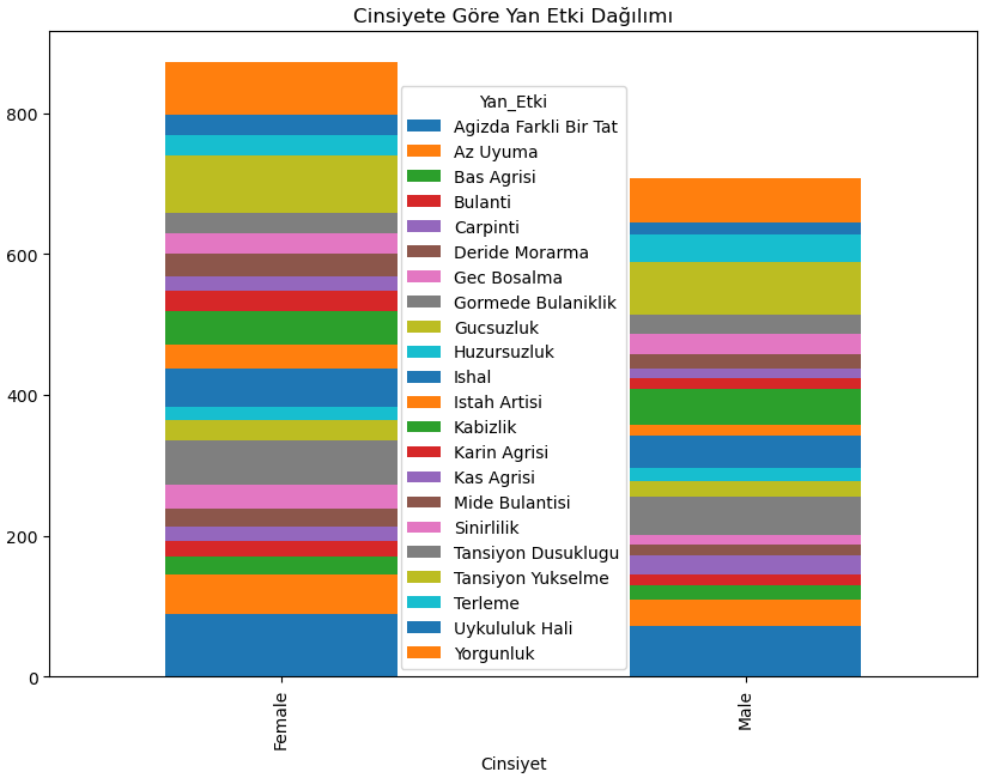


Step 5

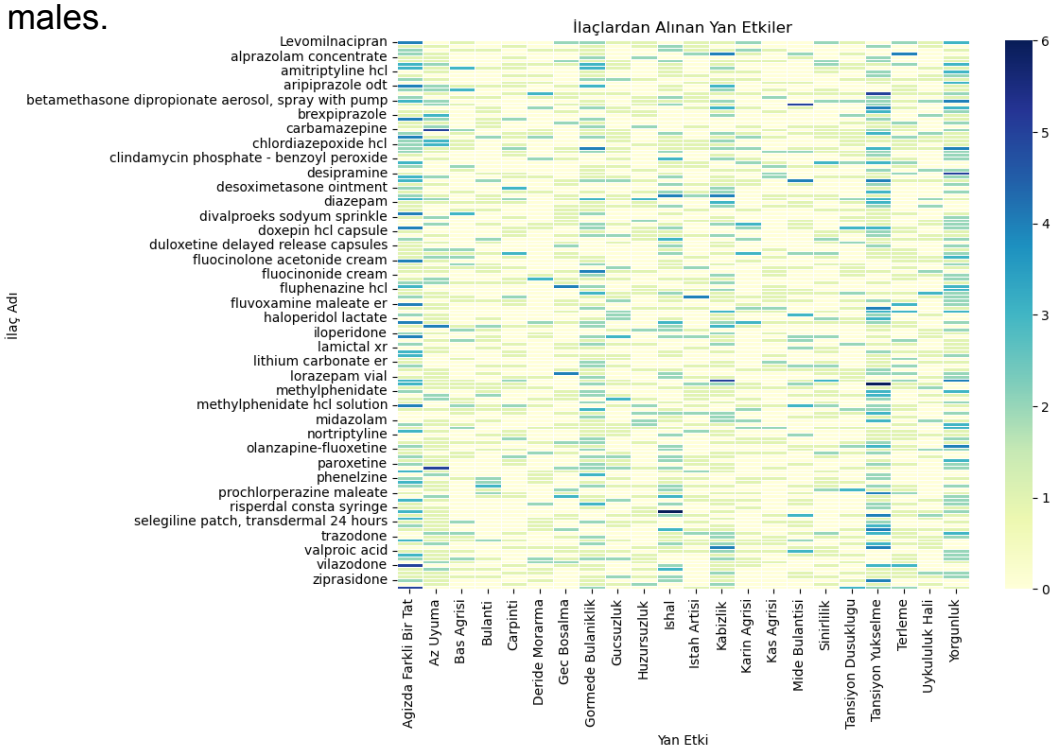
We are examining the relationships between variables. Since this is a study on drug side effects, we will elaborate on this section. We will investigate the relationships between variables that may influence drug side effects through data visualization.



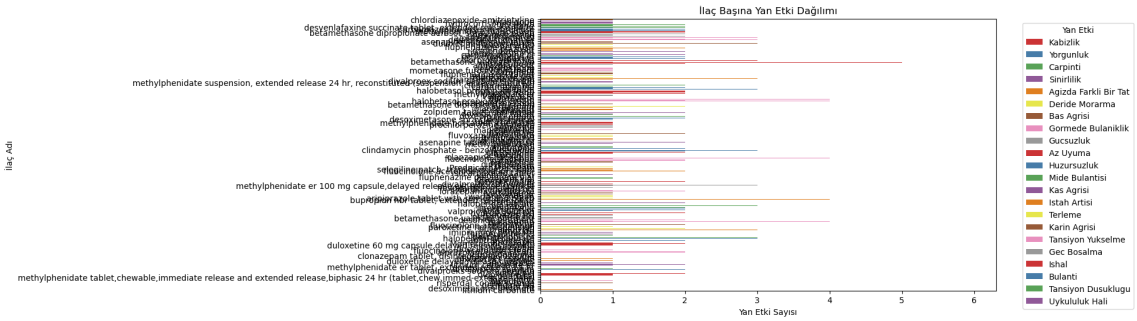
With this graph, we are examining the frequency of side effects.



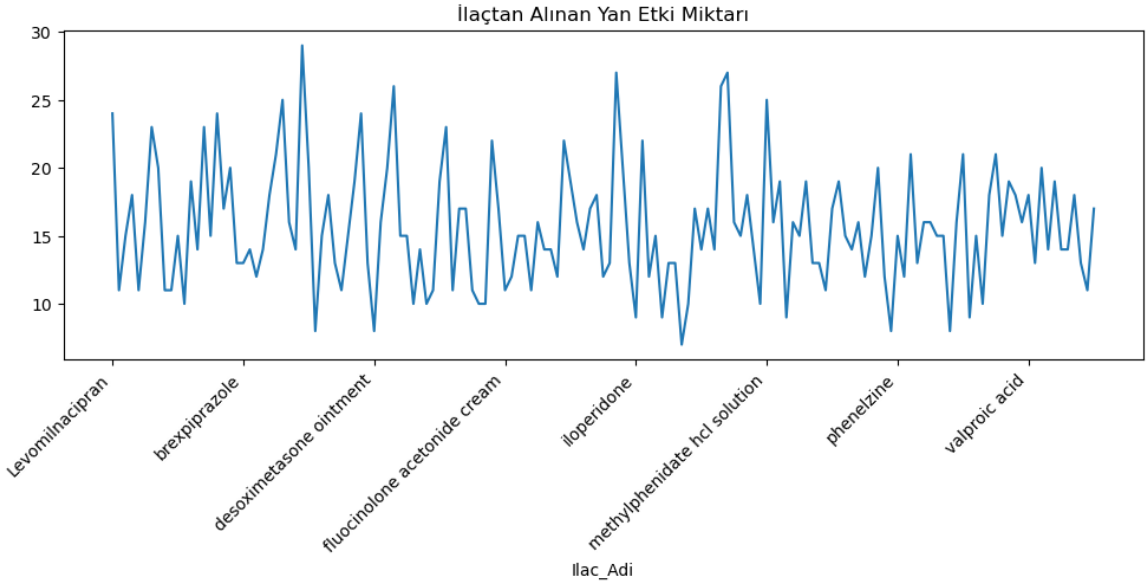
With this graph, you can analyze the effect of gender on side effects. You can compare the distribution of side effects between females and males.



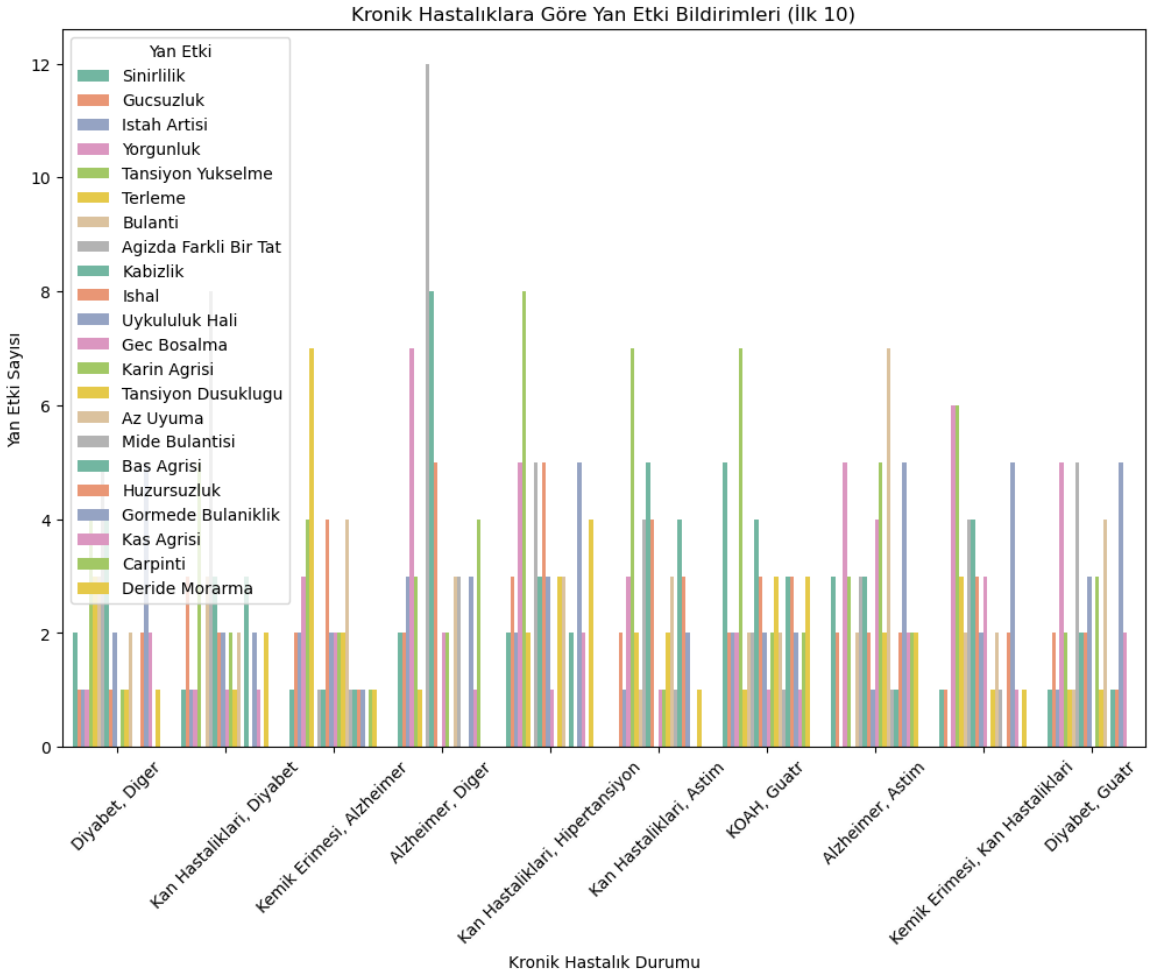
This graph visualizes which side effects are reported more frequently for each drug. The axes will feature drug names and side effects. Side effects that are reported in higher numbers will be represented with darker colors on the map. This way, we can easily analyze which side effects are associated with each drug.



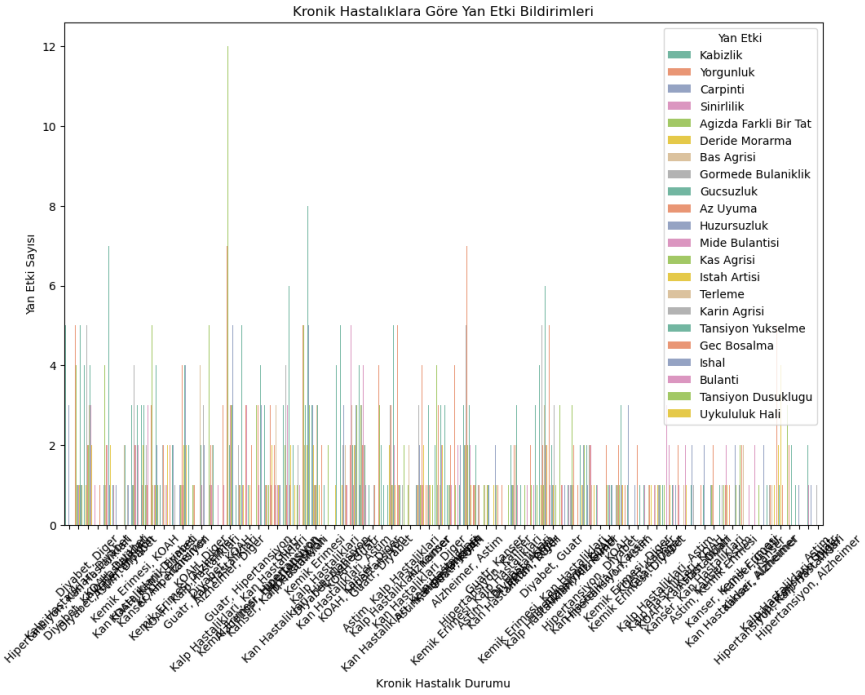
İLAÇTAN ALINAN YAN ETKİ SIKLIĞI

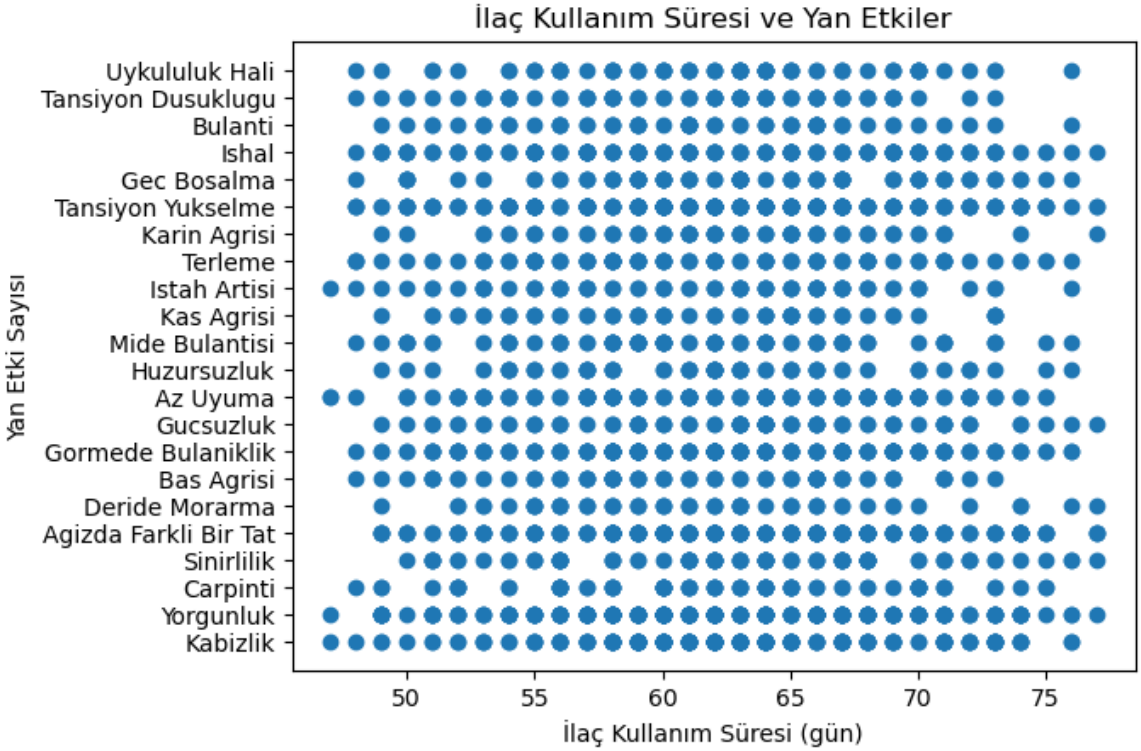


These graphs provide alternative visualizations of side effects associated with the drug.

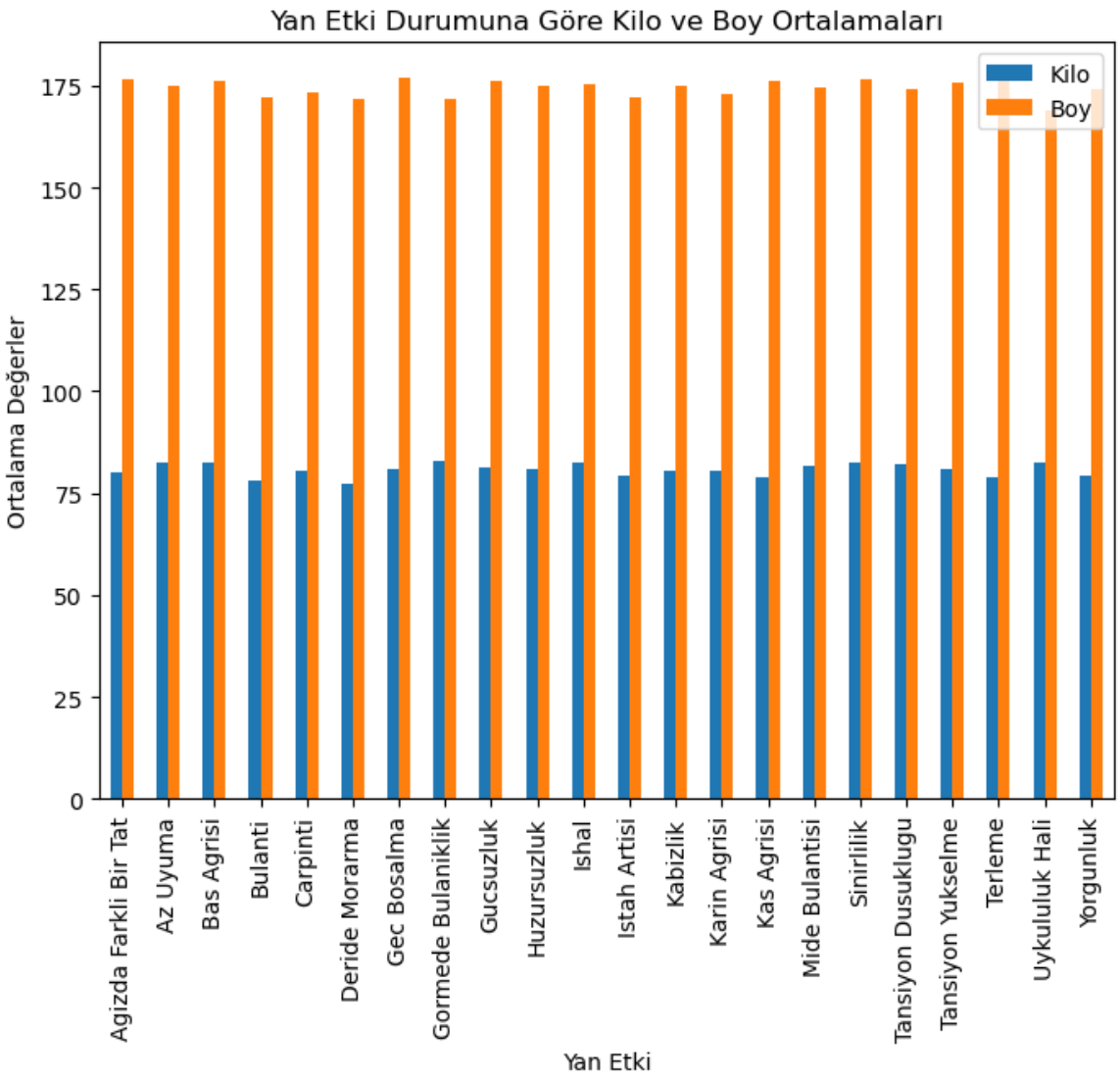


This graph visualizes side effects according to chronic diseases. By interpreting the results of the analyses, we can gain a better understanding of the impact of chronic diseases on side effects.

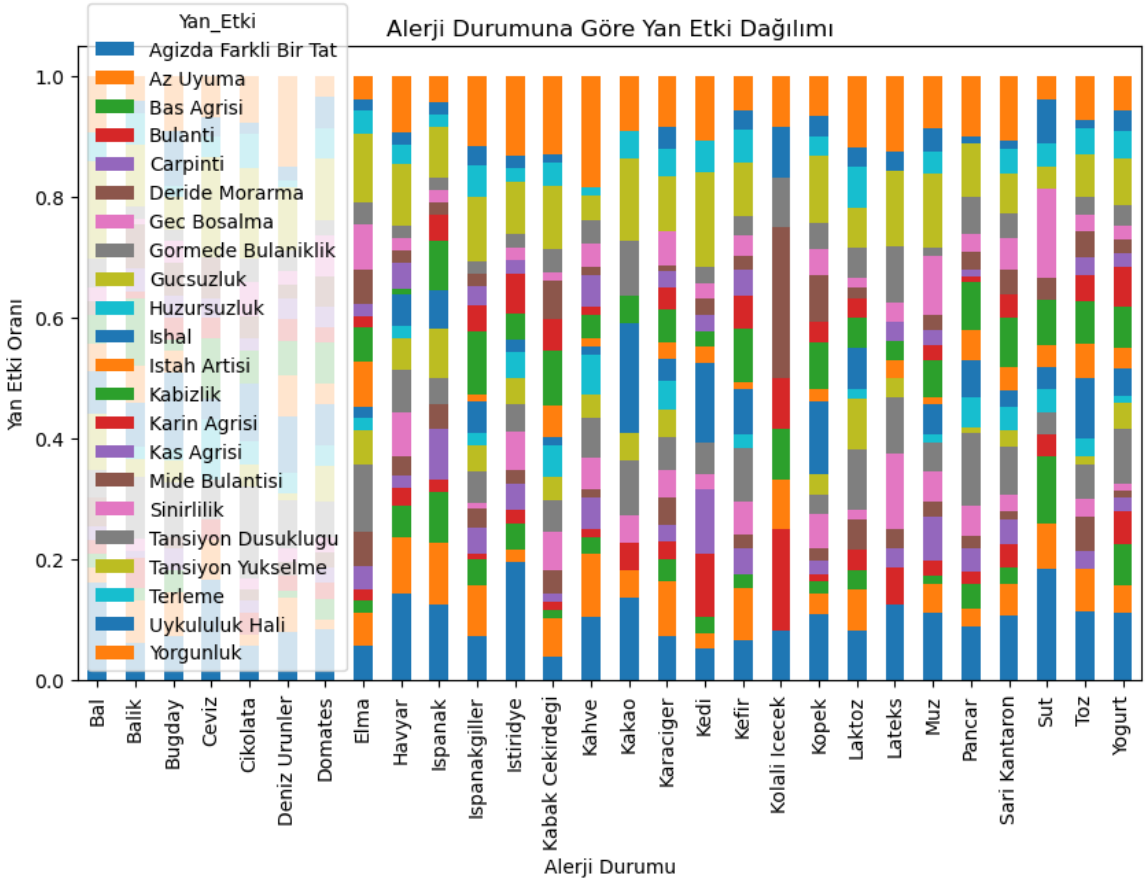




To determine whether there is a relationship between the duration of drug use and side effects, we will first calculate the duration of drug use and then visualize it.



To examine the relationship between numerical variables such as weight and height and side effects, we need to perform numerical analyses. We should calculate the average weight and height for each side effect condition. This way, we can visually analyze the effects of side effects on individuals' weight and height values.



In this graph, we show the distribution of side effects based on allergy status, visualizing the proportions of side effects for each allergy condition using a stacked bar chart.

The analyses we conducted allow you to understand the relationships between side effects and various variables. Such analyses can reveal which groups, drugs, regions, or conditions exhibit more frequent side effects, helping us identify potential risk factors.

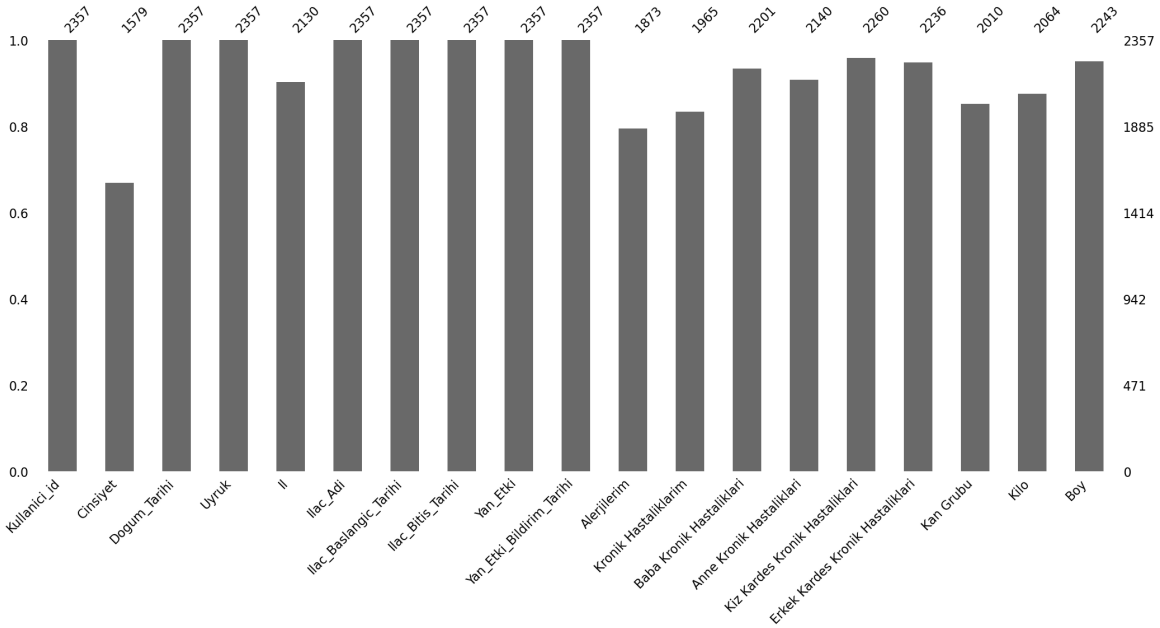
Step 6

During the Data Pre-Processing stage, the dataset is prepared for analysis. In this phase, we will perform tasks such as filling in missing values, transforming variables, adding new features, and cleaning the dataset.

First, we will identify the missing values. The purpose of this is to decide how to fill in the variables in the dataset.

```
print(df.isnull().sum())
```

```
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
```



Step 7

In this phase, we are performing data cleaning and filling in missing values. Missing gender and city information is filled with the mode, chronic disease information and blood type is filled with 'Bilinmiyor', while numerical values like weight and height are filled with the median. These steps will help prevent errors during modeling or analysis due to missing data.

In the output below, we can see that the data cleaning was successful and the missing values have been filled in. This means that the dataset is now suitable for analysis and modeling.

```
In [72]: print(df.head())
```

```
   Kullanici_id Cinsiyet Dogum_Tarihi    Uyruk         Il  \
0           107     Male   1960-03-01  Turkiye  Canakkale
1           140     Male   1939-10-12  Turkiye    Trabzon
2             2   Female   1976-12-17  Turkiye  Canakkale
3            83     Male   1977-06-17  Turkiye      Adana
4             7   Female   1976-09-03  Turkiye      Izmir

                        Ilac_Adi Ilac_Baslangic_Tarihi Ilac_Bitis_Tarihi  \
0                 trifluoperazine            2022-01-09        2022-03-04
1                fluphenazine hcl            2022-01-09        2022-03-08
2                 warfarin sodium            2022-01-11        2022-03-12
3                   valproic acid            2022-01-04        2022-03-12
4  carbamazepine extended release            2022-01-13        2022-03-06

                Yan_Etki Yan_Etki_Bildirim_Tarihi Alerjilerim  \
0              Kabizlik      2022-02-19 18:28:43       Ceviz
1             Yorgunluk      2022-02-03 20:48:17         Toz
2              Carpinti      2022-02-04 05:29:20         Muz
3             Sinirlilik      2022-02-08 01:01:21      Pancar
4  Agizda Farkli Bir Tat      2022-02-12 05:33:06  Bilinmiyor

              Kronik Hastaliklarim  Baba Kronik Hastaliklari  \
0  Hipertansiyon, Kan Hastaliklari      Guatr, Hipertansiyon
1                      Bilinmiyor            Guatr, Diger
2      Kalp Hastaliklari, Diyabet           Diyabet, KOAH
3                  Diyabet, Diger  Kalp Hastaliklari, Diger
4      Diyabet, Kalp Hastaliklari  Alzheimer, Hipertansiyon

           Anne Kronik Hastaliklari    Kiz Kardes Kronik Hastaliklari  \
0                              KOAH  Kemik Erimesi, Kalp Hastaliklari
1  Hipertansiyon, Kalp Hastaliklari
2           Kemik Erimesi, Diyabet            Diyabet, Kemik Erimesi
3                      Bilinmiyor                             Astim
4  Kan Hastaliklari, Kemik Erimesi                   Diyabet, Diger
```

```
In [73]: print(df.isnull().sum())
```

```
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
```

Step 8

We are trying to create some variables that will make the data set meaningful.I wanted to create age and medication duration values in this dataset to facilitate analysis.

```
   Dogum_Tarihi
0    1960-03-01
1    1939-10-12
2    1976-12-17
3    1977-06-17
4    1976-09-03
5    1982-01-05
6    1997-01-10
7    1997-01-15
8    1973-08-05
9    1941-10-16
10   1955-10-07
11   1992-03-24
12   2001-06-01
13   1964-05-22
14   1981-03-01
15   1973-09-09
16   2002-04-15
17   1969-07-23
18   2007-06-13
19   2010-07-23
20   1996-09-10
21   1981-03-01
22   2000-10-06
23   1962-06-30
24   1945-12-06
25   1986-03-27
26   1986-11-07
27   2007-06-13
28   1954-01-20
29   1976-02-20
```

```
[79]: print(df['Ilac_Kullanım_Suresi'].head(30))

0     54
1     58
2     60
3     67
4     52
5     71
6     61
7     56
8     68
9     62
10    65
11    58
12    51
13    66
14    60
15    50
16    69
17    52
18    73
19    51
20    61
21    65
22    65
23    57
24    52
25    70
26    74
27    59
28    61
29    72
Name: Ilac_Kullanım_Suresi, dtype: int64
```

Step 9

In order to simplify the analysis and ensure easier access to the data, I removed the 'Birth_Date', 'Medication_Start_Date', and 'Medication_End_Date' columns from the dataset. Finally, I saved the cleaned dataset.