Elif Nida Kayaduman
kayaduman@sabanciuniv.edu

Documentation: Exploratory Data Analysis and Preprocessing Steps

The dataset contains a total of 2357 data entries and 19 columns. These columns contain various information such as user information, medication use, side effects, and chronic diseases. The general information of the columns is given below:

	Number of Missing Values	Percentage of Missing Values (%)
Kullanici_id	0	0.00
Cinsiyet	778	33.01
Dogum_Tarihi	0	0.00
Uyruk	0	0.00
Il	227	9.63
Ilac_Adi	0	0.00
Ilac_Baslangic_Tarihi	0	0.00
Ilac_Bitis_Tarihi	0	0.00
Yan_Etki	0	0.00
Yan_Etki_Bildirim_Tarihi	0	0.00
Alerjilerim	484	20.53
Kronik Hastaliklarim	392	16.63
Baba Kronik Hastaliklari	156	6.62
Anne Kronik Hastaliklari	217	9.21
Kiz Kardes Kronik Hastaliklari	97	4.12
Erkek Kardes Kronik Hastaliklari	121	5.13
Kan Grubu	347	14.72
Kilo	293	12.43
Boy	114	4.84

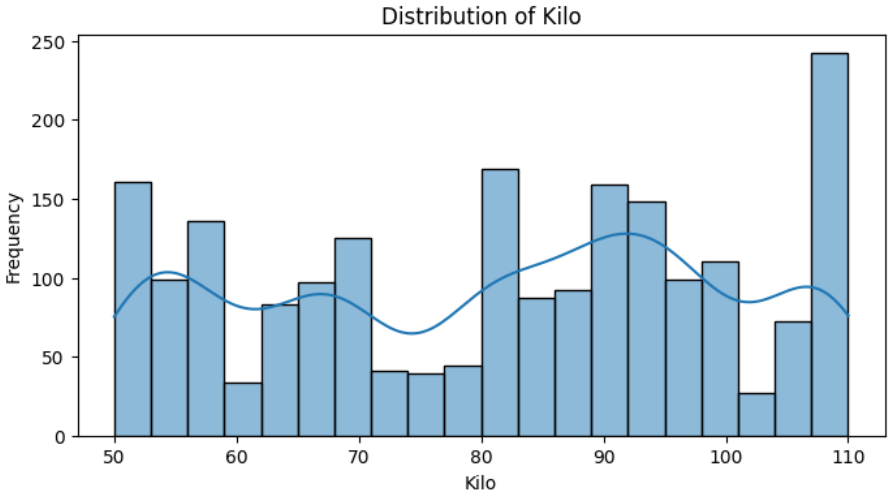
There are missing data in many columns. There is a significant amount of missing data, especially in columns such as Gender, Province, Allergies, Chronic Diseases, and Family Illnesses. The dataset contains numeric (int64, float64) and categorical data types, as well as historical data. Basic statistical analysis was performed on the numeric columns in the dataset. Below are some statistical summaries for the User_id, Weight, and Height columns:

```
Basic statistics for numerical columns:
       Kullanici_id         Kilo          Boy
count   2357.000000  2064.000000  2243.000000
mean      97.216801    80.863857   174.638431
std       57.017200    18.635269    16.516552
min        1.000000    50.000000   145.000000
25%       47.000000    65.000000   160.000000
50%       97.000000    83.000000   176.000000
75%      146.000000    96.000000   187.000000
max      196.000000   110.000000   203.000000
```

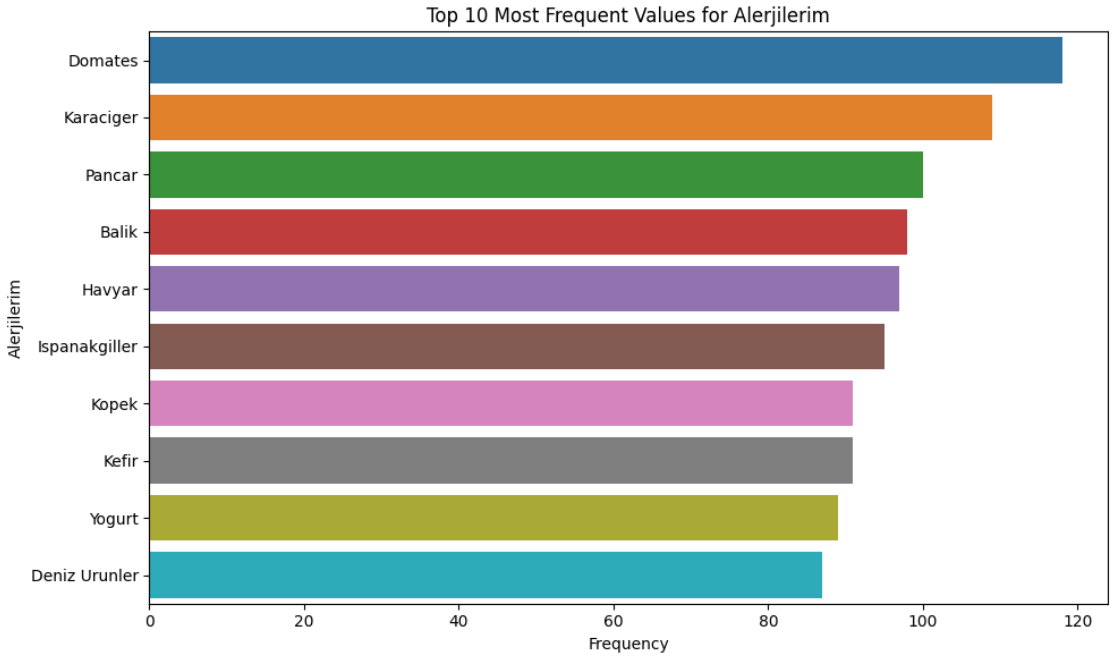
The average weight is 80.86 kg, the minimum value in the weight column is 50 kg and the maximum value is 110 kg. The average height is 174.64 cm, the minimum value is 145 cm and the maximum

Elif Nida Kayaduman
kayaduman@sabanciuniv.edu

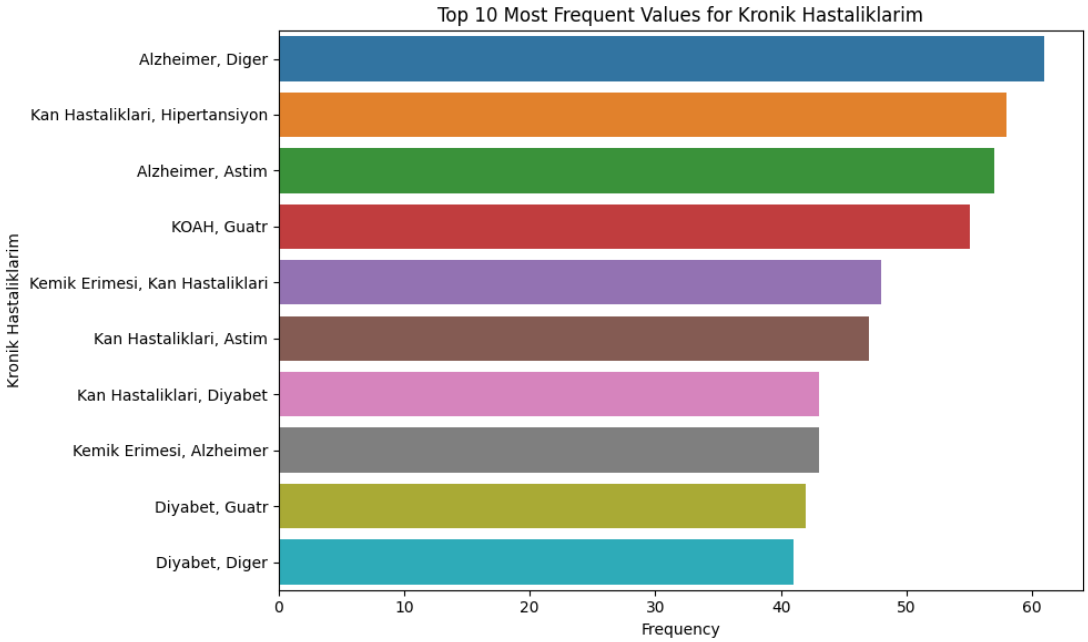
value is 203 cm. The height distribution is concentrated between 160 cm and 187 cm. Various visualization techniques were used to understand the distribution of numerical and categorical variables. Histogram graphs were created for each numerical column. The distribution of columns such as Weight and Height was examined. For example, the distribution of the Weight variable can be seen in the graph below:

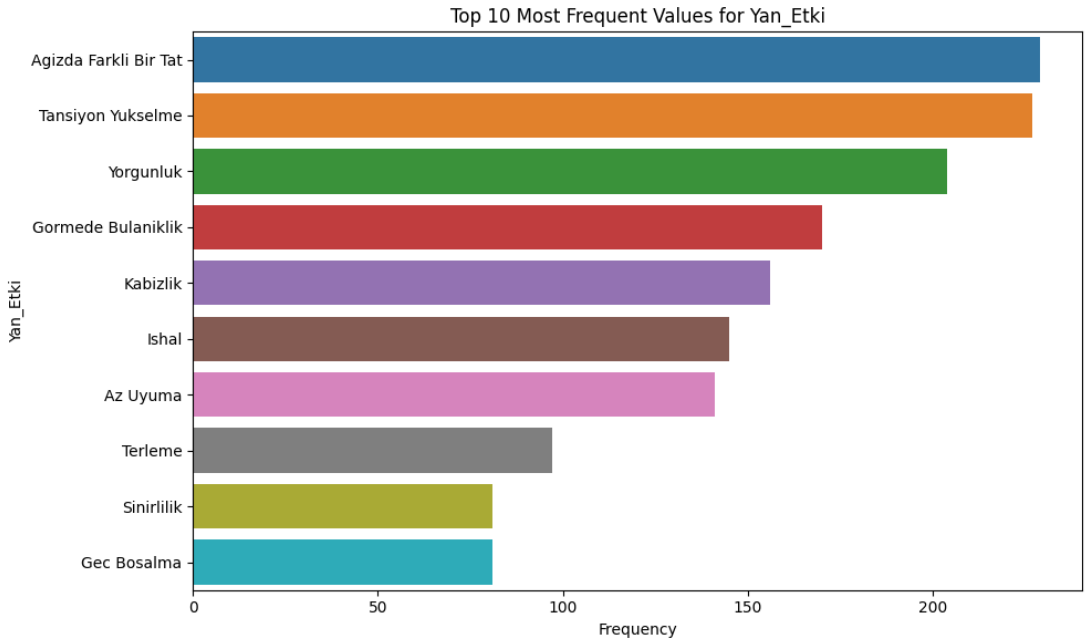


In the Distribution of Categorical Variables, bar charts were created to show the first 10 most frequently occurring categories. In categorical variables, especially in variables such as Drug Name, Side Effect, and Chronic Diseases, the most frequently occurring values were visualized. For example:



Elif Nida Kayaduman
kayaduman@sabanciuniv.edu

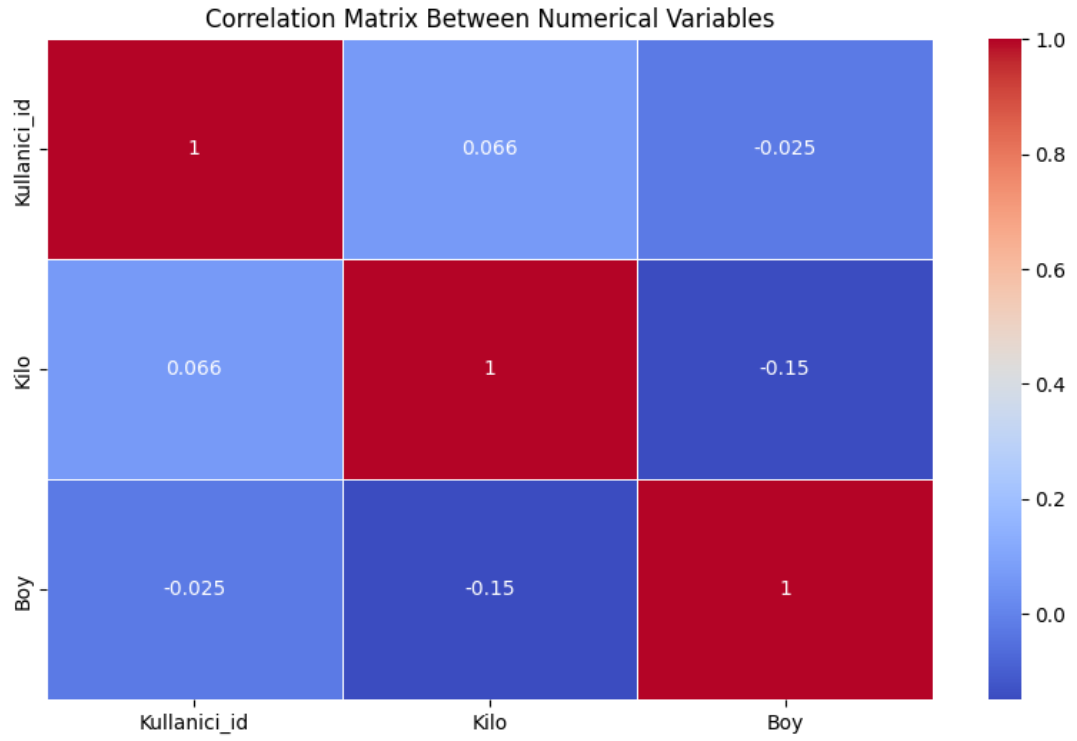




A correlation matrix was created to understand the relationship between numerical variables. The correlations between the variables User_id, Weight and Height are quite low. The correlation between the variables Weight and Height was found to be -0.15, which indicates a very weak negative relationship between these two variables. Since the other correlation values are also very low, there is

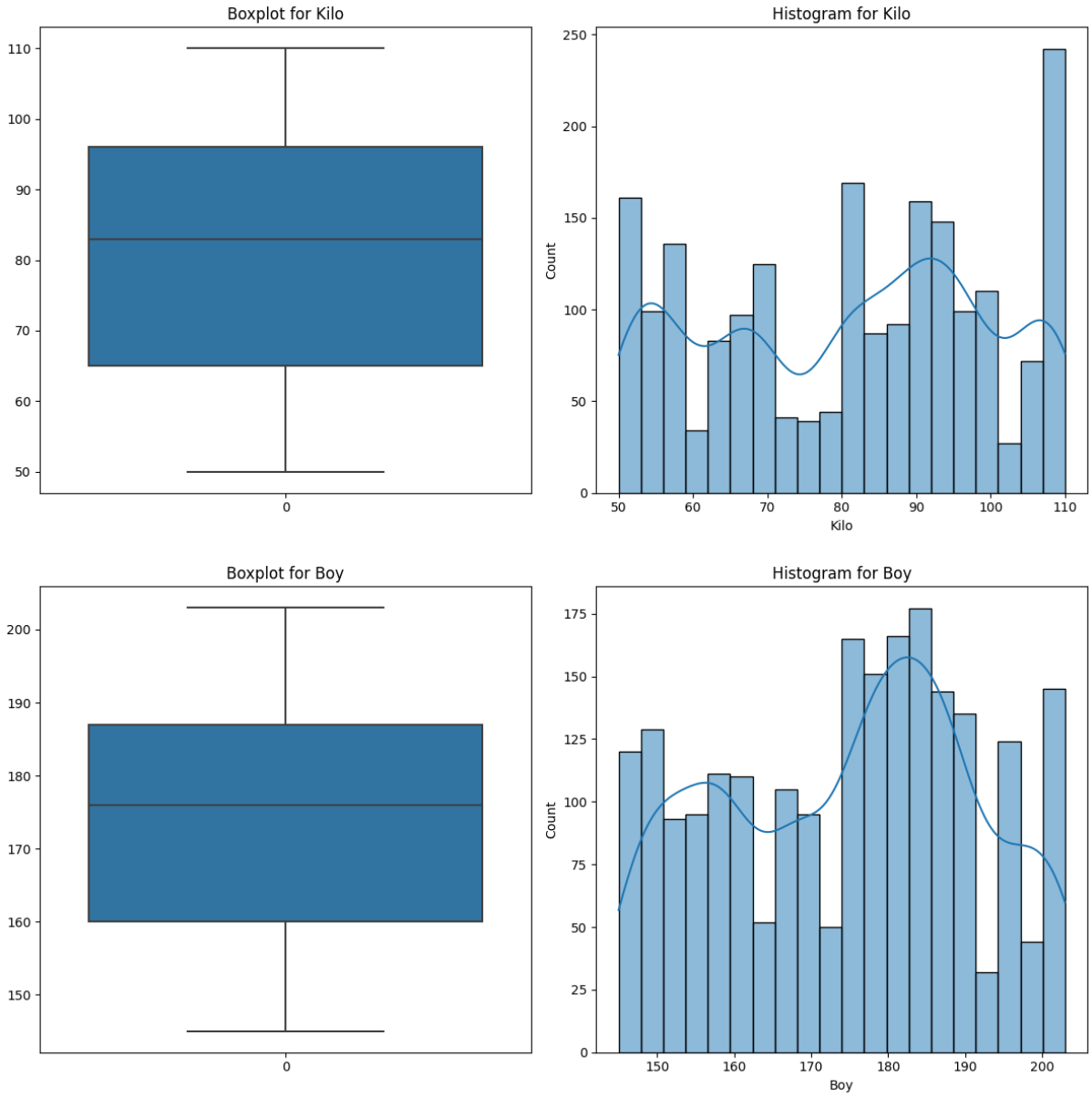
Elif Nida Kayaduman
kayaduman@sabanciuniv.edu

no significant correlation between these numerical variables.



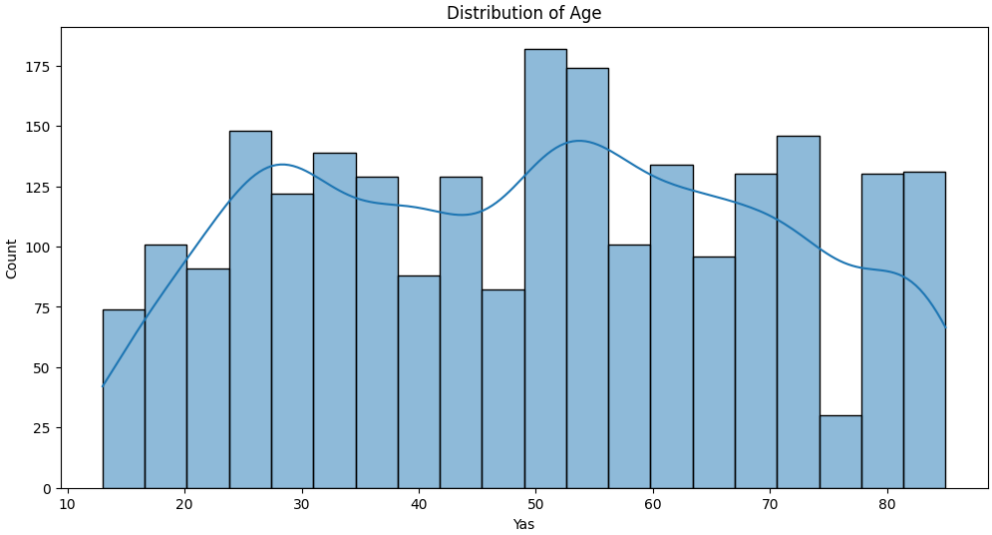
Boxplot and histogram visualizations were made to see the distribution and extreme values of numerical variables. Weight distribution generally varies between 50 and 110 kg and there are no extreme values. A distribution between 150 cm and 200 cm was observed in the height variable. The averages are generally concentrated around 170 cm.

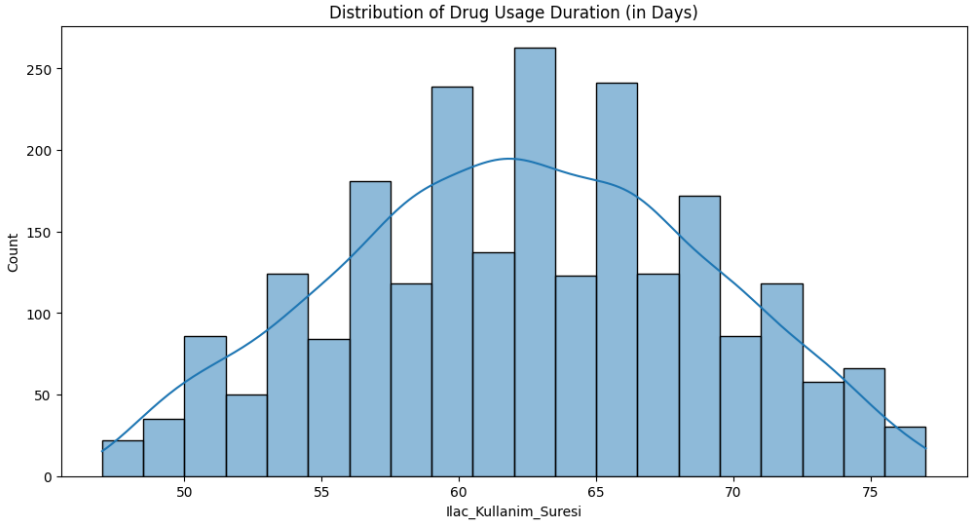
Elif Nida Kayaduman
kayaduman@sabanciuniv.edu



Two new features were added to the dataset. We calculated the user's age based on the Date_of_Birth column. Age was calculated by subtracting the year 2024 from each user's year of birth. The duration of use of each drug was calculated in days using the Drug_Start_Date and Drug End Date columns.

Elif Nida Kayaduman
kayaduman@sabanciuniv.edu



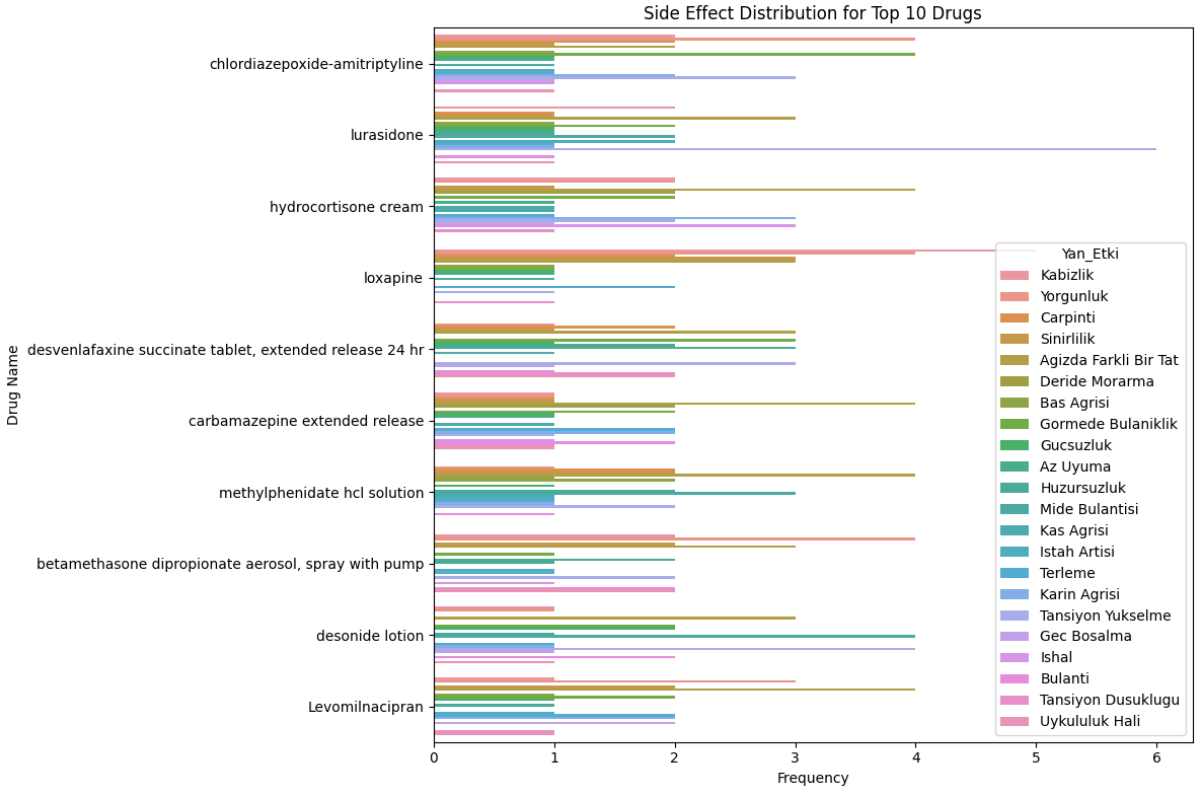


The age distribution is widely spread between 20 and 80 years old. The most common age range is between 30 and 60 years old. The duration of use of drugs mostly varies between 50 and 75 days. The most frequent duration of use is around 60 days.

A visualization was made to analyze the relationship between the drugs in the dataset and the side effects caused by these drugs. The graph below shows the side effect distributions of the 10 most

Elif Nida Kayaduman
kayaduman@sabanciuniv.edu

commonly used drugs. The distribution of different side effects for each drug is indicated in color.



In the graph, it is possible to examine which side effects are more common for each drug. In particular, some drugs exhibit a more intense distribution of certain side effects, while other side effects are less frequent. For example:

• For the drug Clordiazepoxide-amitriptyline, Constipation is one of the most common side effects.

• For Loxapine, Constipation is again in the foreground.

The relationship between the drugs in the dataset, their side effects and the duration of use of these drugs (in days) was analyzed. The table below shows the average values of the side effects and the duration of use of each drug in days. This analysis allows to understand which drugs are associated with which side effects and for how long.

Relationship between drugs, side effects, and usage duration (average in days):

	Ilac_Adi	Yan_Etki	Ilac_Kullanim_Suresi
0	Levomilnacipran	Agizda Farkli Bir Tat	64.500000
1	Levomilnacipran	Bas Agrisi	53.000000
2	Levomilnacipran	Gec Bosalma	63.000000
3	Levomilnacipran	Gormede Bulaniklik	64.000000
4	Levomilnacipran	Gucsuzluk	63.000000
...	...	...	...
1602	zolpidem tablet, sublingual	Kabizlik	73.000000
1603	zolpidem tablet, sublingual	Tansiyon Dusuklugu	54.333333
1604	zolpidem tablet, sublingual	Tansiyon Yukselme	67.000000
1605	zolpidem tablet, sublingual	Terleme	57.500000
1606	zolpidem tablet, sublingual	Yorgunluk	59.000000

Elif Nida Kayaduman
kayaduman@sabanciuniv.edu

For example:

- For the drug Levomilnacipran, the side effect of Delayed Ejaculation is associated with an average treatment duration of 63 days.
- Zolpidem tablets are associated with an average treatment duration of 73 days with the side effect of Constipation.

Data Preprocessing: Preparation for Modeling

A series of data preprocessing steps were performed to prepare the dataset for the modeling phase. These steps included filling in missing data, quantifying categorical variables, and standardizing numerical variables.

1. **Filling in Missing Data:**

The dataset contained missing data in both numerical and categorical columns. The missing data in numerical variables were filled with their mean values. This method ensured that the missing numerical data in the dataset were included in the modeling process. The missing data in categorical variables were filled with the value "unknown". This process ensured that the model was not affected by the missing data and all missing data were successfully filled.

2. **Coding of Categorical Variables:**

Categorical variables in the dataset (e.g., Gender, Nationality, Province, Drug Name, Blood Group) were converted to numerical format for modeling. This process was performed with the One-Hot Encoding method. A separate column was created for each unique value of the categorical variables and these columns were represented in binary (0 and 1) format. For example, the Gender variable was divided into two separate columns as female and male and each category was represented by 0 or 1. As a result of this process, the model was able to process categorical variables correctly.

3. **Standardization of Numerical Variables:**

Numerical variables in the dataset (e.g., Weight, Height, Duration of Medication Use) were standardized for better performance in the modeling phase.

4. **Saving the Dataset:**

After the completion of the data preprocessing steps, the processed dataset was saved to a CSV file. This stage made it possible to prepare the data to be used in the modeling step. All missing data were filled, categorical variables were made numerical, and numerical variables were standardized. In this form, the dataset is ready for any machine learning algorithm.