Data Science Intern Case Study Summary

Kübra Öztürk

kbrztrk.kbr@gmail.com

1. Exploratory Data Analysis (EDA):

At this stage, the dimensions of the data set, column names and data types, missing data were examined. Basic statistical summary was created. Categorical variables were identified and the unique values and their numbers in each categorical column were examined. Data that may have a relationship between them were examined with graphics. Pandas, Matplotlib and Seaborn libraries were used. More detailed visualization was applied after the preprocessing step.

```
Veri setinin boyutları: (2357, 19)
Sütunlar ve veri tipleri:
Kullanici_id                          int64
Cinsiyet                             object
Dogum_Tarihi                 datetime64[ns]
Uyruk                                object
Il                                   object
Ilac_Adi                             object
Ilac_Baslangic_Tarihi        datetime64[ns]
Ilac_Bitis_Tarihi            datetime64[ns]
Yan_Etki                             object
Yan_Etki_Bildirim_Tarihi     datetime64[ns]
Alerjilerim                          object
Kronik Hastaliklarim                 object
Baba Kronik Hastaliklari             object
Anne Kronik Hastaliklari             object
Kiz Kardes Kronik Hastaliklari       object
Erkek Kardes Kronik Hastaliklari     object
Kan Grubu                            object
Kilo                                float64
Boy                                 float64
dtype: object
```

```
Sayısal sütunlar için özet istatistikler:
       Kullanici_id                  Dogum_Tarihi  \
count   2357.000000                          2357
mean      97.216801  1974-11-25 04:06:12.677131936
min        1.000000           1939-10-12 00:00:00
25%       47.000000           1959-02-05 00:00:00
50%       97.000000           1973-09-09 00:00:00
75%      146.000000           1992-03-24 00:00:00
max      196.000000           2011-04-25 00:00:00
std       57.017200                           NaN

                Ilac_Baslangic_Tarihi             Ilac_Bitis_Tarihi  \
count                            2357                          2357
mean   2022-01-07 10:47:36.173101312  2022-03-10 16:25:27.365294848
min              2022-01-01 00:00:00           2022-03-02 00:00:00
25%              2022-01-04 00:00:00           2022-03-06 00:00:00
50%              2022-01-07 00:00:00           2022-03-11 00:00:00
75%              2022-01-11 00:00:00           2022-03-15 00:00:00
max              2022-01-14 00:00:00           2022-03-19 00:00:00
std                              NaN                           NaN

            Yan_Etki_Bildirim_Tarihi         Kilo         Boy
count                           2357  2064.000000  2243.000000
mean   2022-02-10 17:09:30.742044928    80.863857   174.638431
min              2022-02-01 04:34:33    50.000000   145.000000
25%              2022-02-04 05:29:20    65.000000   160.000000
50%              2022-02-09 20:53:54    83.000000   176.000000
75%              2022-02-17 07:08:01    96.000000   187.000000
max              2022-02-19 21:47:39   110.000000   203.000000
std                              NaN    18.635269    16.516552
```

```
Eksik veri analizi:
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
```

Total 2 different categories, Column: Cinsiyet
Total 1 different category, Column: Uyruk
Total 13 different categories, Column: Il
Total 151 different categories, Column: İlaç_Adi
Total 22 different categories, Column: Yan_Etki
Total 28 different categories, Column: Alerjilerim
Total 80 different categories, Column: Kronik Hastaliklarim
Total 92 different categories, Column: Baba Kronik Hastaliklari
Total 84 different categories, Column: Anne Kronik Hastaliklari
Total 85 different categories, Column: Kiz Kardes Kronik Hastaliklari
Total 90 different categories, Column: Erkek Kardes Kronik Hastaliklari
Total 8 different categories. Column: Kan Grubu

When the graphs were examined:
It was determined that people with the B rh(-) blood group had a different taste in their mouth as a side effect, people with the AB rh(-) blood group had high blood pressure, people with the B rh(+) blood group had a different taste in their mouth, people with the 0 rh(-) blood group had a different taste in their mouth, people with the AB rh(+) blood group felt tired, people with the 0rh(+) blood group had high blood pressure, people with the A rh(+) blood group had high blood pressure, and people with the A rh(-) blood group had a feeling of tiredness as a side effect. The most common side effect was bruising in Çanakkale, fatigue in Trabzon, high blood pressure in Adana, a different taste in the mouth in İzmir, a different taste in the mouth in Mersin, high blood pressure in Antalya, a different taste in the mouth in Eskişehir, high blood pressure in Samsun, high blood pressure in Ankara, fatigue in Bursa, fatigue in İstanbul, fatigue, a different taste in the mouth and weakness in Malatya, and finally blurred vision in Kayseri. And like these examples, the relationship between drug name and side effect, the relationship between chronic diseases and side effects, the relationship between allergies and side effects, and the relationship between blood type and

chronic diseases have also been made interpretable. These relationships should be tested with a detailed regression analysis and models can be created for specific groups.

2. Data Pre-Processing:

The numpy library was used at this stage. The missing values in the "Il" column were filled with the most common values. The column was deleted because the values in the "Uyruk" column were the same for everyone. The "Kullanıcı id" column was deleted. The column was deleted because there was too much missing data in the "Cinsiyet" column and it could be misleading. The missing values in the "Boy" and "Kilo" columns were filled by taking their averages. The "İlaç Kullanım Süresi" column was created using the drug start and end dates. The duration of side effects was calculated using the drug start date and side effect notification dates. The "Yan Etkilerin Süresi" column was created. The body mass index was calculated using the weight and height values and a new column named "Vki" was created. The missing values in the "Kan Grubu" column were filled with the most frequently used values. A new column named "Yaş" was created using the values in the date of birth column. For categorical values, the unique values examined in the EDA phase were listed in an ordered manner by creating a mapping dictionary. This process was applied for the columns "İlaç Adı", "Yan Etki", "Alerji", "Il", "Kan Grubu", “Kronik Hastaliklari”, “Baba Kronik Hastaliklar", “Anne Kronik Hastaliklari”, “Kiz Kardes Kronik Hastaliklari”, “Erkek Kardes Kronik Hastaliklari”. For chronic disease columns, all missing nan values were deleted at these stages. New columns were created and columns belonging to old values were deleted. New values consisting entirely of numerical data were converted to integers. Date columns were updated in date format. Finally a new dataset was created.

3. Data Visualization:

In order to better analyze the new dataset, the relationships between the values were examined with various graphs. The relationship between drug name and chronic diseases was visualized with the heat map, the relationship between blood group and side effects according to drug name was visualized with the facetgrid, the relationship between age and duration of drug use was visualized with scatterplot, the most common drug and side effect combinations were visualized with barplot, the relationship between weight and height and side effects was visualized with scatterplot, the most common allergy and side effect combinations were visualized with barplot, blood group and side effect combinations were visualized with barplot, the relationship between drug name and chronic diseases was visualized with heat map.

4. Future Plan:

In the next step; Comparisons can be made with the visualized relationships in the EDA stage and preprocessing stage, and they can be grouped and evaluated. Time series analysis can be performed. Future predictions can be made by interpreting with various machine learning and deep learning methods. In this way, the analyzed data can be made guiding with meaningful relationships.