Emre Ataoğlu

**Side Effects Case Study Summary**

**As a first step, I called the ‘’pandas,numpy,matplotlib,seaborn and SimpleImputer’’ then I uploaded the dataset.**

**Then I started to explore dataset with the df.head(),df.describe(),df.info() commands. With the head command I wanted to see how data looks like. With ‘’describe’’ I saw maximum,minimum,mean values for each category. With ‘’info’’ command, I understand that there are 19 columns, we have 2357 user’s datas, and I saw how many data types we got and what data types each category is.**

**Then with ‘’isnull’’ command I wanted to see how many NaN values we got for each category. And started to handle missing values. For Gender I created category ‘’unknown.’’ And I wanted to see Gender Distribution. %36,99 Female, %33 Unknown, %29,9 Male .**

**For Chronic Diseases, I created category ‘’There is no known chronic disease found.’’ Because it’s important data for a side effects and not only are our data is missing, but there may also be users who do not really have chronic diseases. Users could have more than 1 diseases and in our dataset they were together with comma. I had to reshape the data so I split the answers after commas and explode column to have correct values for each diseases. I applied the same structure to sibling categories too.**

**Height and weight datas are different because if I use simple imputer and give mean values to missing datas it’ll effect value counts and may mislead us in our future measurements. Instead of imputer I used ‘’dropna’’ command.**

**After handling missing data, I did a table to see every drug’s side effects. Then I wanted to visualize it and created figure table. I created another cross table to see only the particular drug’s side effects. We can apply this to any other categories we want. For example height and side effects, allergies and side effects which is important. I wanted to carry these models to one step further and created a crosstable that make us able to see the particular report for the particular drug.**

**Then I thought side effects report date and duration of drug treatment is an important data. I used ‘’datetime’’ command to transform the date data. Then I created ‘’Tedavi Suresi’’ category which is the difference of the days between starting date and duration of the drug treatment. With the same approach I created ‘’Yan Etki Bildirim Suresi’’. And I wanted to see the mean value of the Side Effect Reportation. I find that the side effect reports averagely comes after 32 days starting to drug treatment.**

**With this new data, I created a application to see average reportation day of the particular drug. For example, for ‘’fluoxetine’’ average side effect reportation date is 32 days later.**

**After all this data I wanted to see correlation between Drug treatment duration and side effects reportation date. And the results were surprising. If user’s expected treatment is longer than 65 days, side effects don’t show themselves first 20 days. And if expected treatment is longer than 70 days, side effects don’t show themselves first 25 days. I wanted to see correlation coefficient and It was 0.35 which means they have low-positive correlation.**

**With the weight and height data I calculated Body Mesure Index and categorized as ‘’skinny,normal,overweight,obese’’. Then with the bar chart I showed ‘’BMI’’ value counts of our users.**

**Also I categorized age scales as ‘’(0-18),(19-30),(31-40),(41-50),(51-60),(61+)’’ and show side effects for the age scales with the cross table.**

**Finally I visualized all the data we have, with the pie charts, bar tables, cross tables.**