Week 8 – Healthcare Project

**Group Name: Cool Data Scientists Team**

Team Members Details:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Email** | **Country** | **College / Company** | **Specialization** |
| Yousef Elbayoumi | yousefxelbayomi@gmail.com | Palestine | Bahçeşehir University | Data Science |
| Mukhammadjon Kholmirzev | kmukhammadjon@gmail.com | Uzbekistan | Ulsan National Institute of Science and Technology | Data Science |
| Jamila hamdi | jamila.hamdi90@gmail.com | Tunisia | Faculty of science and managment | Data Science |
| H. Melis Tekin Akcin | [meliss85@gmail.com](mailto:meliss85@gmail.com) | UK | Hacettepe University | Data Science |

**Problem Description**

One of the challenges for Pharmaceutical companies is to understand the persistency of drug as per the physician prescription. This issue results in a bad impact on the pharmacies for all the categories; patients, physicians, and administration. However, the team of data scientist is capable of discovering the analyzing the dataset and detecting the factors that are impacting the primary factor which is the "persistency". By building a classification machine learning model, we will be able to classify the dataset and find the variables that affect the target variables "Persistency Flag".

**Data understanding**

As a first step, we imported the dataset and copied it. Then we’ve looked at the first five and the last five entries.

The following pictures show how our dataset looks like:

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

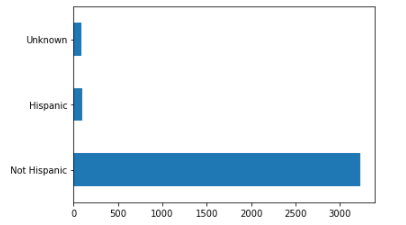
metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

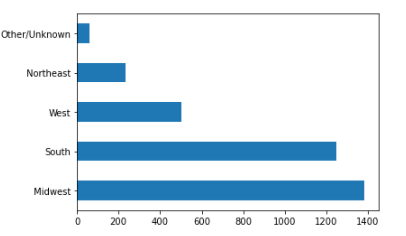
Totally we have 3424 observations and 69 features.

For Demographics, we have the followings:

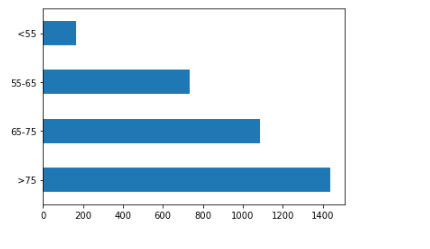
If we examine “Ethnicity”, we see that “Non-Hispanic” people dominates the “Hispanic” people and also we have unknown values.



If we examine the “Region”, we see that patients are mostly “Midwest” and “South” region:

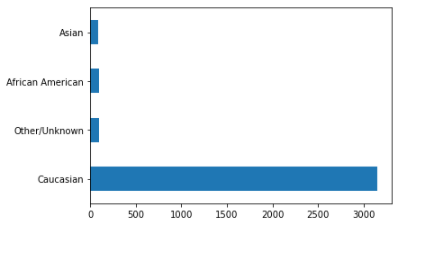


If we look at the “Age”, we see the following:

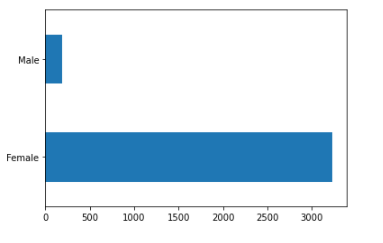


By looking at the above picture, it can be thought that being of age “>55” can be related to have persistency to drug.

If we look at the “Race”, we see that the Caucasians are dominated the other races.



If we look at the “Gender”, by the following picture , the female patients are more than the male patients.



**Ntm Speciality** is the specialty of the HCP that prescribed the NTM Rx.

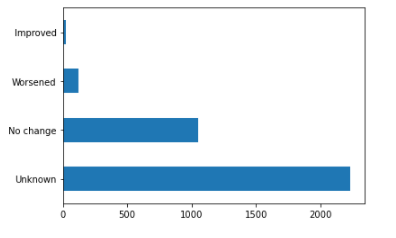
## We see that General Practitioner, Rheumatology, Endocrinology and Oncology specialists prescribed the NTM Rx most.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Clinical Factors:**

**Risk Segment:** We have compared the risk segments prior NTM and during NTM and examine how it changes:



We have done similar computations for all other clinical factors.

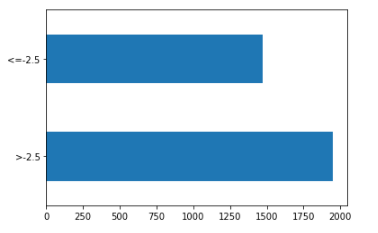
For instance, we have examined the Fragility and we have obtained the following cross- table:

metin içeren bir resim

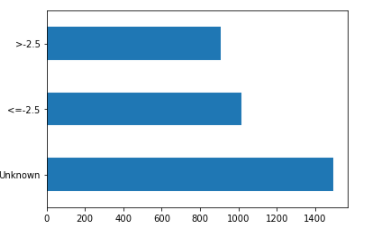
Açıklama otomatik olarak oluşturuldu

**T-scores:**

We have compared the “T-scores”. The following picture shows the prior to NTM:



The following shows the “T-scores” during the Rx:



Besides, we have examined the Disease and Treatment Factors. They are all Yes/No information and we have decided which of the variables can affect the persistency to drug:

By comparing the results, we see that the followings can affect the target variable:

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu.

Similarly, we think that Vitamin D-insufficiency can affect the target variable.

**What type of data you have got for analysis?**

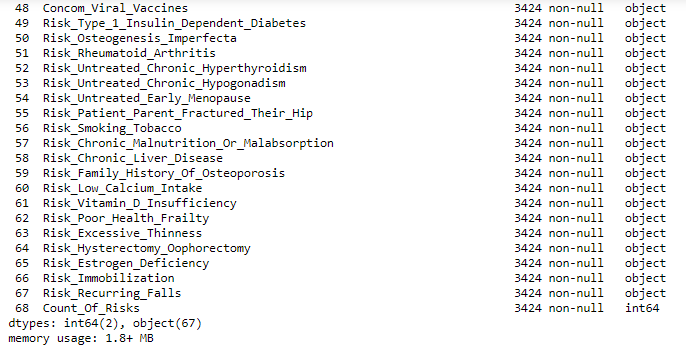
When we’ve checked the types of the variables, we obtained the following result:

metin içeren bir resim

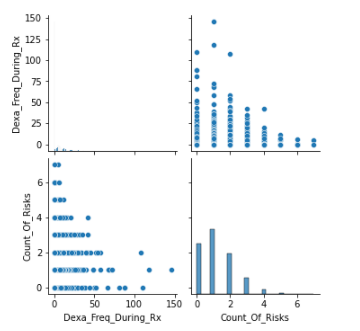
Açıklama otomatik olarak oluşturuldu

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu



We have that those 67 features are of object type and just 2 of them are int64 type.

And we have determined the relation between these two numerical variables:

**What are the problems in the data ( number of NA values, outliers , skewed etc):**

**NA Values:**

When we checked that whether there is any NA value, we have obtained the following:

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Even if we don’t have any NA values, we have “Unknown” variables. The followings are only

examples of some of them:

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin içeren bir resim

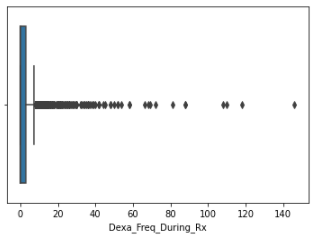
Açıklama otomatik olarak oluşturuldu

metin içeren bir resim

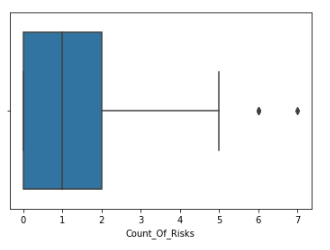
Açıklama otomatik olarak oluşturuldu

**Outliers:**

To detect the outliers, we’ve used boxplot.



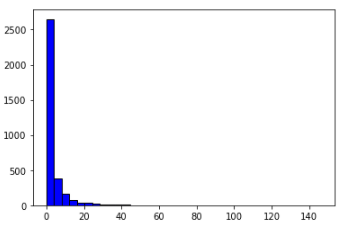
We have 460 outliers in “Dexa\_Freq\_During\_Rx” variable.



We have 8 outliers in “Count\_Of\_Risks” variable.

**Skewed Data:**

We have the following histogram graphs:



As seen in the above, since the tail is on the right side, we can say that “Dexa\_Freq\_During\_Rx” variable has right-skewed distribution. Hence, we can conclude that the mean value is greater than the mode.

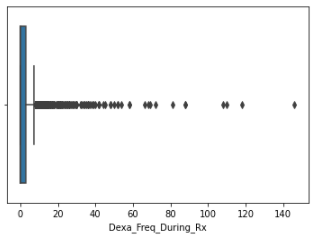
**What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?**

**For NA values**: Since all of the NA values are in object types we prefer to ignore these values.

For instance, we have NA values in “Ethnicity”. If we change the Unknown values with “Hispanic” or “Non-Hispanic” it can change the result of the dataset.

**For Outliers:**

As seen in the following picture, the outliers of the “**Dexa\_Freq\_During\_Rx**” variable are place on the right-hand side of the upper bound. So, if we replace them with the mean value can change the type of the dataset. But instead, we have discussed on suppressing them with the upper bound.



On the other hand, the number of the outliers of the “**Count\_Of\_Risks**” variable is just 8. So, we can use mean value or suppress them with the upper bound.

**Github Repo link**

https://github.com/melis-ta/Healthcare