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	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:

```
df.head()
```

	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag	Ntm_Speciality_Bucket	...
0	P1	Persistent	Male	Caucasian	Not Hispanic	West	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...
1	P2	Non-Persistent	Male	Asian	Not Hispanic	West	55-65	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...
2	P3	Non-Persistent	Female	Other/Unknown	Hispanic	Midwest	65-75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...
3	P4	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...
4	P5	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...

```
df.tail()
```

	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag	Ntm_Speciality_Bucket	...
3419	P3420	Persistent	Female	Caucasian	Not Hispanic	South	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...
3420	P3421	Persistent	Female	Caucasian	Not Hispanic	South	>75	Unknown	Others	OB/GYN/Others/PCP/Unknown	...
3421	P3422	Persistent	Female	Caucasian	Not Hispanic	South	>75	ENDOCRINOLOGY	Specialist	Endo/Onc/Uro	...
3422	P3423	Non-Persistent	Female	Caucasian	Not Hispanic	South	55-65	Unknown	Others	OB/GYN/Others/PCP/Unknown	...
3423	P3424	Non-Persistent	Female	Caucasian	Not Hispanic	South	65-75	Unknown	Others	OB/GYN/Others/PCP/Unknown	...

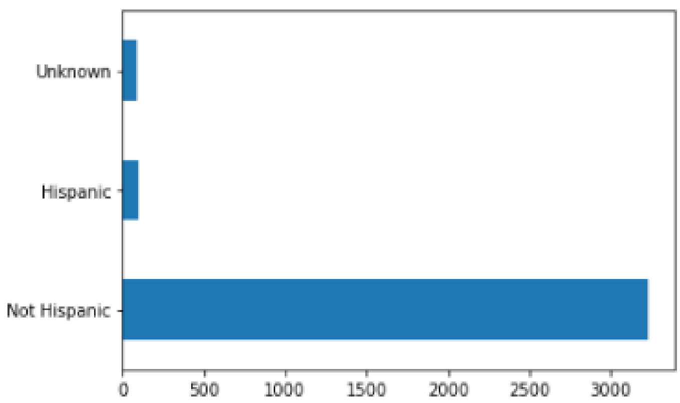
Totally we have 3424 observations and 69 features.

```
df.shape
```

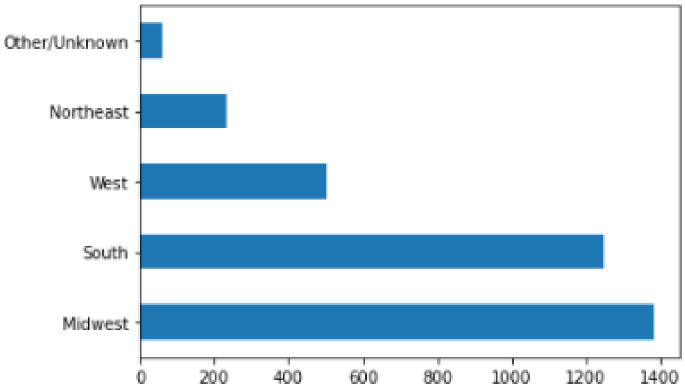
(3424, 69)

For Demographics, we have the followings:

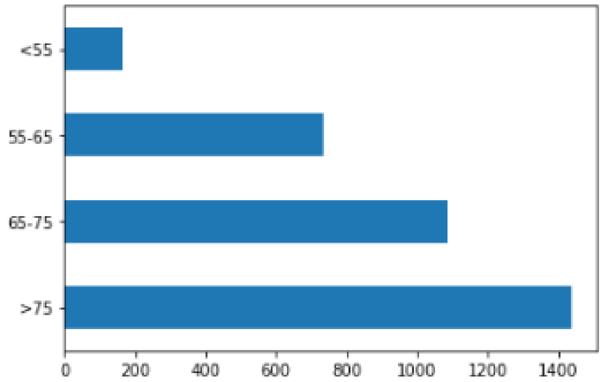
If we examine "Ethnicity", we see that "Non-Hispanic" people dominates the "Hispanic" people and also we have unknown values.



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

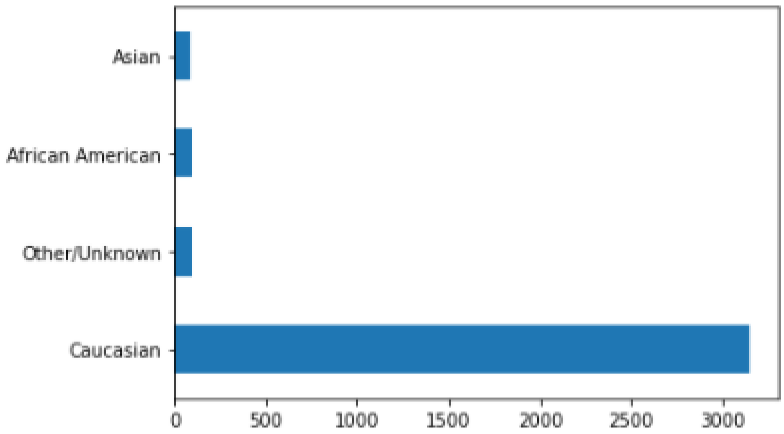


If we look at the “Age”, we see the following:

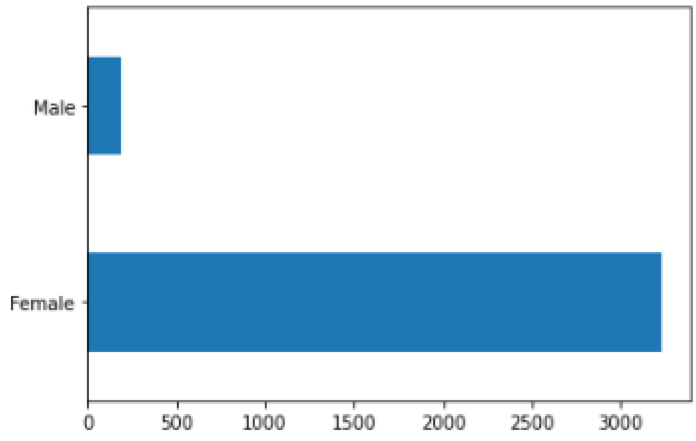


By looking at the above picture, it can be thought that being of age “>55” can be related to have persistency to drug.

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.



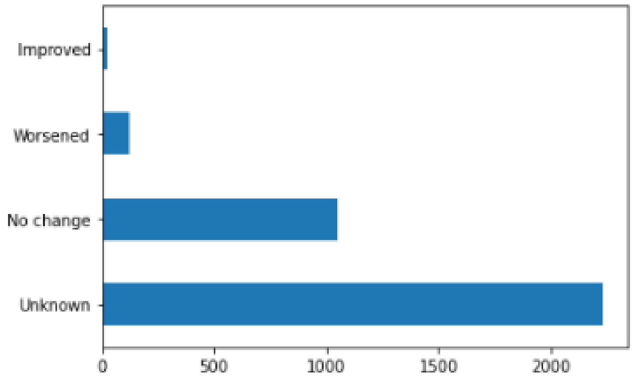
Ntm Speciality is the specialty of the HCP that prescribed the NTM Rx.

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

GENERAL PRACTITIONER	1535
RHEUMATOLOGY	604
ENDOCRINOLOGY	458
Unknown	310
ONCOLOGY	225
OBSTETRICS AND GYNECOLOGY	90
UROLOGY	33
ORTHOPEDIC SURGERY	30
CARDIOLOGY	22
PATHOLOGY	16
HEMATOLOGY & ONCOLOGY	14
OTOLARYNGOLOGY	14
PEDIATRICS	13
PHYSICAL MEDICINE AND REHABILITATION	11
PULMONARY MEDICINE	8
SURGERY AND SURGICAL SPECIALTIES	8
PSYCHIATRY AND NEUROLOGY	4
NEPHROLOGY	3
ORTHOPEDICS	3
GERIATRIC MEDICINE	2
HOSPICE AND PALLIATIVE MEDICINE	2
PLASTIC SURGERY	2
GASTROENTEROLOGY	2
VASCULAR SURGERY	2
TRANSPLANT SURGERY	2
OCCUPATIONAL MEDICINE	1
OPHTHALMOLOGY	1
PAIN MEDICINE	1

Clinical Factors:

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



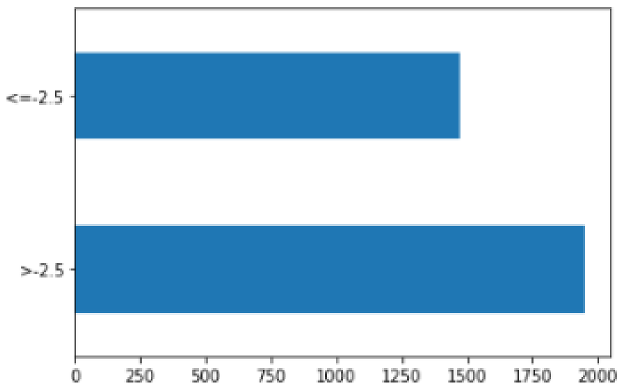
We have done similar computations for all other clinical factors.

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

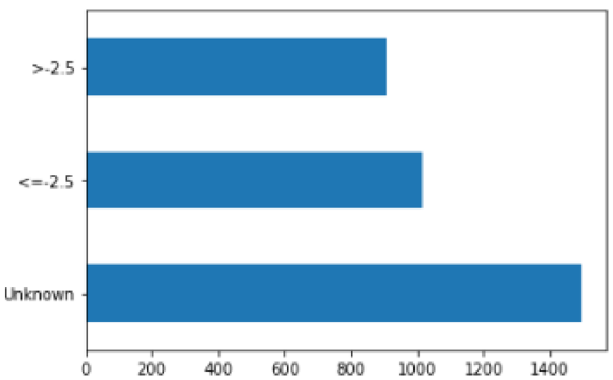
Frag_Frac_During_Rx	N	Y
Frag_Frac_Prior_Ntm		
N	2691	181
Y	316	236

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:

1) Comorb_Encounter_For_Screening_For_Malignant_Neoplasms
2) Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias
3) Comorb_Encounter_For_Immunization.

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:

```
df.info()
```

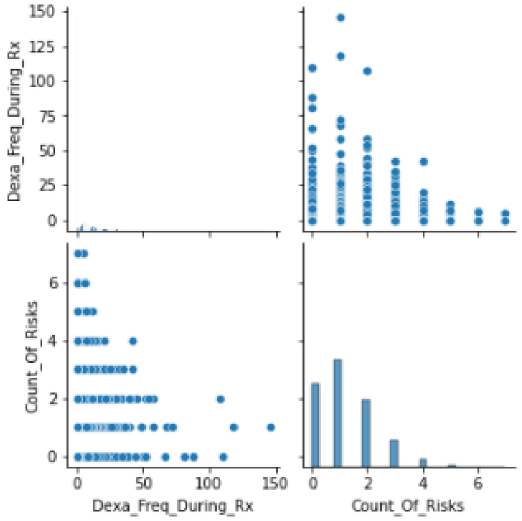
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3424 entries, 0 to 3423
Data columns (total 69 columns):
 #   Column                                                    Non-Null Count  Dtype
---  ------                                                    --------------  -----
 0   Ptid                                                      3424 non-null   object
 1   Persistency_Flag                                          3424 non-null   object
 2   Gender                                                    3424 non-null   object
 3   Race                                                      3424 non-null   object
 4   Ethnicity                                                 3424 non-null   object
 5   Region                                                    3424 non-null   object
 6   Age_Bucket                                                3424 non-null   object
 7   Ntm_Speciality                                            3424 non-null   object
 8   Ntm_Specialist_Flag                                       3424 non-null   object
 9   Ntm_Speciality_Bucket                                     3424 non-null   object
 10  Gluco_Record_Prior_Ntm                                    3424 non-null   object
 11  Gluco_Record_During_Rx                                    3424 non-null   object
 12  Dexa_Freq_During_Rx                                       3424 non-null   int64
 13  Dexa_During_Rx                                            3424 non-null   object
 14  Frag_Frac_Prior_Ntm                                       3424 non-null   object
 15  Frag_Frac_During_Rx                                       3424 non-null   object
 16  Risk_Segment_Prior_Ntm                                    3424 non-null   object
 17  Tscore_Bucket_Prior_Ntm                                   3424 non-null   object
 18  Risk_Segment_During_Rx                                    3424 non-null   object
 19  Tscore_Bucket_During_Rx                                   3424 non-null   object
 20  Change_T_Score                                            3424 non-null   object
 21  Change_Risk_Segment                                       3424 non-null   object
 22  Adherent_Flag                                             3424 non-null   object
```

```
 23  Idn_Indicator                                                    3424 non-null   object
 24  Injectable_Experience_During_Rx                                  3424 non-null   object
 25  Comorb_Encounter_For_Screening_For_Malignant_Neoplasms           3424 non-null   object
 26  Comorb_Encounter_For_Immunization                                3424 non-null   object
 27  Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx  3424 non-null   object
 28  Comorb_Vitamin_D_Deficiency                                      3424 non-null   object
 29  Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified             3424 non-null   object
 30  Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx 3424 non-null   object
 31  Comorb_Long_Term_Current_Drug_Therapy                            3424 non-null   object
 32  Comorb_Dorsalgia                                                 3424 non-null   object
 33  Comorb_Personal_History_Of_Other_Diseases_And_Conditions         3424 non-null   object
 34  Comorb_Other_Disorders_Of_Bone_Density_And_Structure             3424 non-null   object
 35  Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias  3424 non-null   object
 36  Comorb_Osteoporosis_without_current_pathological_fracture        3424 non-null   object
 37  Comorb_Personal_history_of_malignant_neoplasm                    3424 non-null   object
 38  Comorb_Gastro_esophageal_reflux_disease                          3424 non-null   object
 39  Concom_Cholesterol_And_Triglyceride_Regulating_Preparations      3424 non-null   object
 40  Concom_Narcotics                                                 3424 non-null   object
 41  Concom_Systemic_Corticosteroids_Plain                            3424 non-null   object
 42  Concom_Anti_Depressants_And_Mood_Stabilisers                     3424 non-null   object
 43  Concom_Fluoroquinolones                                          3424 non-null   object
 44  Concom_Cephalosporins                                            3424 non-null   object
 45  Concom_Macrolides_And_Similar_Types                              3424 non-null   object
 46  Concom_Broad_Spectrum_Penicillins                                3424 non-null   object
 47  Concom_Anaesthetics_General                                      3424 non-null   object
```

```
 48  Concom_Viral_Vaccines                                              3424 non-null   object
 49  Risk_Type_1_Insulin_Dependent_Diabetes                             3424 non-null   object
 50  Risk_Osteogenesis_Imperfecta                                       3424 non-null   object
 51  Risk_Rheumatoid_Arthritis                                          3424 non-null   object
 52  Risk_Untreated_Chronic_Hyperthyroidism                             3424 non-null   object
 53  Risk_Untreated_Chronic_Hypogonadism                                3424 non-null   object
 54  Risk_Untreated_Early_Menopause                                     3424 non-null   object
 55  Risk_Patient_Parent_Fractured_Their_Hip                            3424 non-null   object
 56  Risk_Smoking_Tobacco                                               3424 non-null   object
 57  Risk_Chronic_Malnutrition_Or_Malabsorption                         3424 non-null   object
 58  Risk_Chronic_Liver_Disease                                         3424 non-null   object
 59  Risk_Family_History_Of_Osteoporosis                                3424 non-null   object
 60  Risk_Low_Calcium_Intake                                            3424 non-null   object
 61  Risk_Vitamin_D_Insufficiency                                       3424 non-null   object
 62  Risk_Poor_Health_Frailty                                           3424 non-null   object
 63  Risk_Excessive_Thinness                                            3424 non-null   object
 64  Risk_Hysterectomy_Oophorectomy                                     3424 non-null   object
 65  Risk_Estrogen_Deficiency                                           3424 non-null   object
 66  Risk_Immobilization                                                3424 non-null   object
 67  Risk_Recurring_Falls                                               3424 non-null   object
 68  Count_Of_Risks                                                     3424 non-null   int64
dtypes: int64(2), object(67)
memory usage: 1.8+ MB
```

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:

```
In [8]:  df.isnull().values.any()
Out[8]:  False

In [9]:  df.isnull().sum()
Out[9]:  Ptid                              0
         Persistency_Flag                  0
         Gender                            0
         Race                              0
         Ethnicity                         0
                                          ..
         Risk_Hysterectomy_Oophorectomy    0
         Risk_Estrogen_Deficiency          0
         Risk_Immobilization               0
         Risk_Recurring_Falls              0
         Count_Of_Risks                    0
         Length: 69, dtype: int64
```

Even if we don't have any NA values, we have "Unknown" variables. The followings are only examples of some of them:

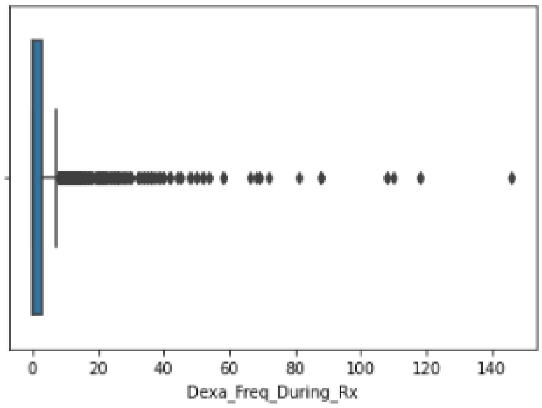
```
In [9]: df["Ethnicity"].value_counts()
Out[9]: Not Hispanic    3235
        Hispanic          98
        Unknown           91
        Name: Ethnicity, dtype: int64
```

```
In [11]: df["Region"].value_counts()
Out[11]: Midwest          1383
         South            1247
         West              502
         Northeast         232
         Other/Unknown      60
         Name: Region, dtype: int64
```

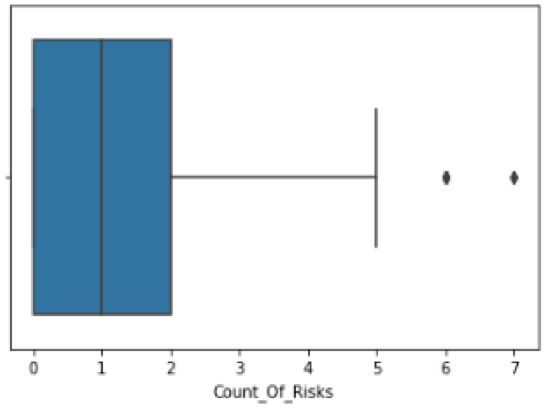
```
In [20]: df["Risk_Segment_During_Rx"].value_counts()
Out[20]: Unknown    1497
         HR_VHR      965
         VLR_LR      962
         Name: Risk_Segment_During_Rx, dtype: int64
```

Outliers:

To detect the outliers, we've used boxplot.



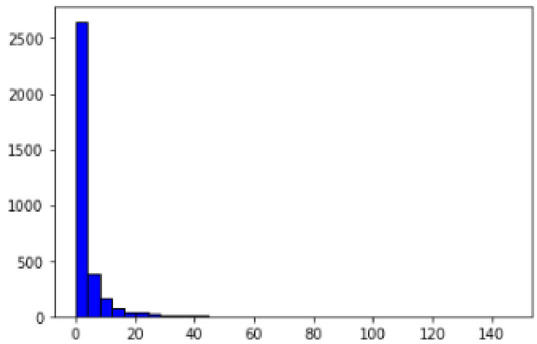
We have 460 outliers in "Dexa_Freq_During_Rx" variable.



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.

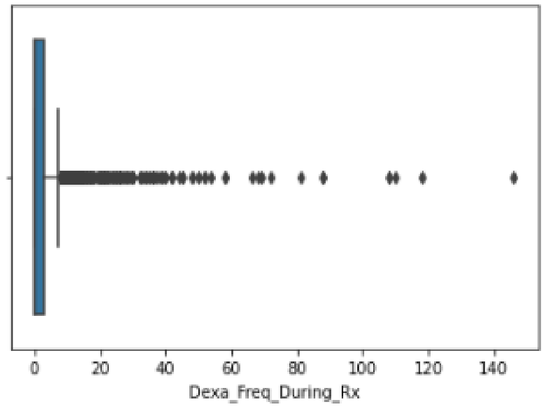
What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?

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