# Data Glacier Internship – Week 07

## Group Name: Persistency of my own

| Name | E-mail | Country | College | Specialization |
|---|---|---|---|---|
| Nahari Terena | naharifterena@gmail.com | Brazil | La Sapienza University of Rome | Data Science |

# Healthcare - Persistency of a drug

## Problem Description

One of the challenges for all pharmaceutical companies is to understand the persistence of drugs as per the physician's prescription. To solve this problem, ABC pharma company approached an analytics company to automate this process of identification.

## Business Understanding

## Project Lifecycle

## Data Intake Report

Name: Healthcare – persistency of a drug
Report date: December 26th
Internship Batch: LISUM15
Version:1.0
Data intake by: Nahari Terena
Data intake reviewer: Data Glacier
Data storage location: local

**Tabular data details: Healthcare_dataset**

| | |
|---|---|
| **Total number of observations** | 3424 |
| **Total number of files** | 1 |
| **Total number of features** | 69 |
| **Base format of the file** | csv |
| **Size of the data** | 913359 KB |

## Data Understanding

The dataset corresponds to 69 variables about the 3424 patients. Our tag is "Persistency Flag" and the "Ptid" is the identification column. There is no duplicated row.
The description of the 69 input features is given below.

| Bucket | Variable | Type | Missing Value | Missing values (%) | Unique values |
|---|---|---|---|---|---|
| **Demographics** | Gender | Object | No | 0 | "Male", "Female" |

| | | | | | |
|---|---|---|---|---|---|
| | Race | Object | No | 0 | "Caucasian", "Asian", "African American", "Other/Unknown" |
| | Ethnicity | Object | Yes | 2.7 | "Not hispanic", "Hispanic", "Unknown" |
| | Region | Object | No | 0 | "Midwest", "West", "South",  "Northeast" |
| | Age_Bucket | Object | No | 0 | "<55", "55-65","65-75",">75" |
| | Idn_Indicator | Object | No | 0 | "Y", "N" |
| **Provider Attibutes** | Ntm_Speciality | Object | Yes | 9.1 | GENERAL PRACTITIONER', 'CARDIOLOGY', 'CLINICAL NURSE SPECIALIST', 'EMERGENCY MEDICINE', 'ENDOCRINOLOGY', 'GASTROENTEROLOGY', 'GERIATRIC MEDICINE', 'HEMATOLOGY & ONCOLOGY', 'HOSPICE AND PALLIATIVE MEDICINE', 'HOSPITAL MEDICINE', 'NEPHROLOGY', 'NEUROLOGY', 'NUCLEAR MEDICINE', 'OBSTETRICS & OBSTETRICS & GYNECOLOGY & OBSTETRICS & GYNECOLOGY', 'OBSTETRICS AND GYNECOLOGY', 'OCCUPATIONAL MEDICINE', 'ONCOLOGY', 'OPHTHALMOLOGY', 'ORTHOPEDIC SURGERY', 'ORTHOPEDICS', 'OTOLARYNGOLOGY', 'PAIN MEDICINE', 'PATHOLOGY', 'PEDIATRICS', 'PHYSICAL MEDICINE AND REHABILITATION', 'PLASTIC SURGERY', 'PODIATRY', 'PSYCHIATRY AND NEUROLOGY', 'PULMONARY MEDICINE', 'RADIOLOGY', 'RHEUMATOLOGY', 'SURGERY AND SURGICAL |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | SPECIALTIES', 'TRANSPLANT SURGERY', 'Unknown', 'UROLOGY', 'VASCULAR SURGERY' |
| | Ntm_Specialist_Flag | Object | No | 0 | "Others", "Specialist" |
| | Ntm_Speciality_Bucket | Object | No | 0 | OB/GYN/Others/PCP/Unknown', 'Endo/Onc/Uro', 'Rheum' |
| **Clinical Factors** | Gluco_Record_Prior_Ntm | Object | No | 0 | "Y", "N" |
| | Gluco_Record_During_Rx | Object | No | 0 | "Y", "N" |
| | Dexa_Freq_During_Rx | Integer | No | 0 | Median: 0<br>Mean: 3.016<br>Max: 146 |
| | Dexa_During_Rx | Object | No | 0 | "Y", "N" |
| | Frag_Frac_Prior_Ntm | Object | No | 0 | "Y", "N" |
| | Frag_Frac_During_Rx | Object | No | 0 | "Y", "N" |
| | Risk_Segment_Prior_Ntm | Object | No | 0 | "VLR_LR", "HR_VHR" |
| | Tscore_Bucket_Prior_Ntm | Object | No | 0 | ">-2.5", "<=-2.5" |
| | Risk_Segment_During_Rx | Object | Yes | 43.7 | "VLR_LR", "HR_VHR", "Unknown" |
| | Tscore_Bucket_During_Rx | Object | Yes | 43.7 | ">-2.5", "<=-2.5", "Unknown" |
| | Change_T_Score | Object | Yes | 43.7 | "No change", "Unknown", "Worsened", "Improved" |
| | Change_Risk_Segment | Object | Yes | 65.1 | "No change", "Unknown", "Worsened", "Improved" |
| | Adherent_Flag | Object | No | 0 | "Non-Adherent", "Adherent" |
| **Disease/Treatment Factor** | Injectable_Experience_During_Rx | Object | No | 0 | "Y", "N" |
| | NTM - Risk Factors | Object | No | 0 | "Y", "N" |
| | NTM - Comorbidity | Object | No | 0 | "Y", "N" |
| | NTM - Concomitancy | Object | No | 0 | "Y", "N" |
| | Count_Of_Risks | Integer | No | 0 | Median: 1<br>Mean: 1.239<br>Max: 7 |

## Data Problems

1) Missing Values

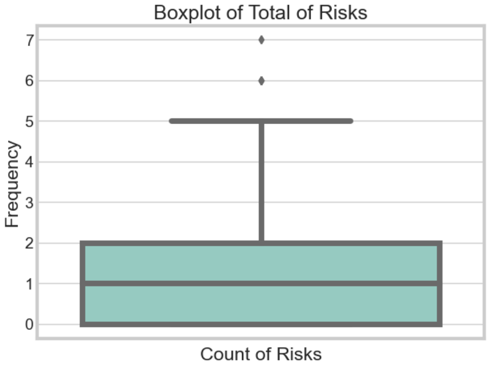
   About six features have a column with at least one missing value. Therefore, we have a specific approach to each one.

   a) “Ethnicity”: We considered “unknown” as a “not Hispanic” category. There 75% of which the respondent "Unknown" had their race classified as “Caucasian”, “Asian” or “African American”.
   b) “Ntm_Speciality”: we decided to consider “unknown” as a specific category.
   c) "Risk_Segment_During_Rx", "Tscore_Bucket_During_Rx", Change_T_Score" and “Change_Risk_Segment": These features were excluded as over 40% of their answer were “unknown”.

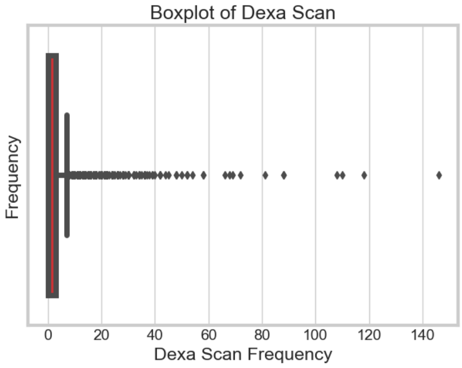
2) Outliers

   An outlier is when an observation differs significantly from other observations from other values. It can occur due to an error or data collection. Outliers can affect the mean of the distribution. There are two numeric columns and both have outliers.

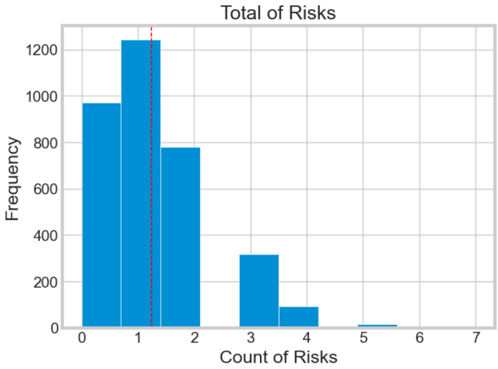
   a) “Count_Of_Risks”: we maintain the outliers (6 and 7) as they reflect the reality.

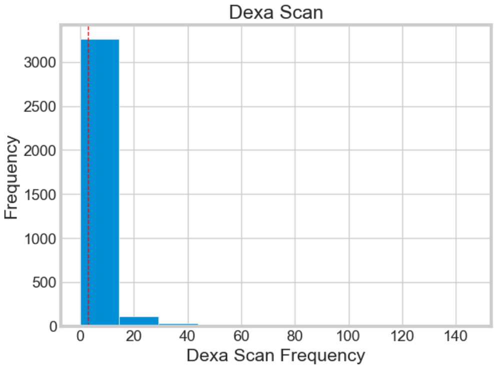


   b) “Dexa_Freq_During_Rx”: We decided to apply Tukey’s boxplot method which distinguishes between possible and probable outliers. A possible outlier is located between the inner and the outer fence, whereas a probable outlier is located outside the outer fence. For this method, only the probable outliers are treated. However, 272 observations are probable outlier and 460 are possible outliers. Nevertheless, we decided to maintain as we aim to classify the patients, it’s possible to have a group that take many scans during the year.



### 3) Skweness and Kurtosis





a) Skewness is a measure of asymmetry of a distribution. When the value of the skewness is negative, the tail of the distribution is longer towards the left hand side of the curve. When the value of the skewness is positive, the tail of the distribution is longer towards the right hand side of the curve.

b) Kurtosis is one of the two measures that quantify shape of a distribution. Kurtosis determine the volume of the outlier. If the distribution is tall and thin it is called a leptokurtic distribution (Kurtosis > 3). Values in a leptokurtic distribution are near the mean or at the extremes.

Along with skewness, kurtosis is an important descriptive statistic of data distribution. However, the two concepts must not be confused with each other. Skewness essentially measures the symmetry of the distribution, while kurtosis determines the heaviness of the distribution tails.

Github Repo