**Data Science HealthCare Project**

**Drug**  **Persistance**



**By : Sourour Cherif**

**Mail : Sourour.cherif@esprit .tn**

**Data Science Engineering Student**

**Esprit Tunisia**

# Problem description

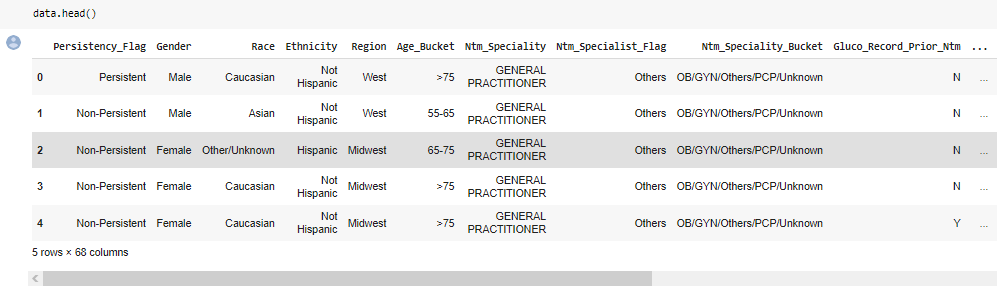


# Business understanding



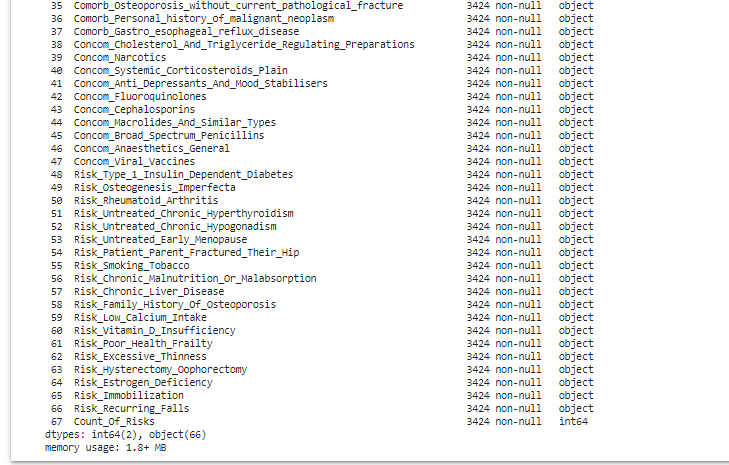
# Data understanding

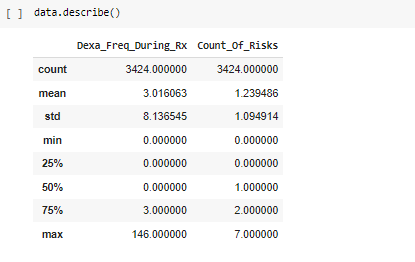
**To fit any predictive model on a dataset, we need to understand the complexity of the dataset before deciding which predictive model to use to get optimal performance.**



# Type of data







**Unique elements in each Column**

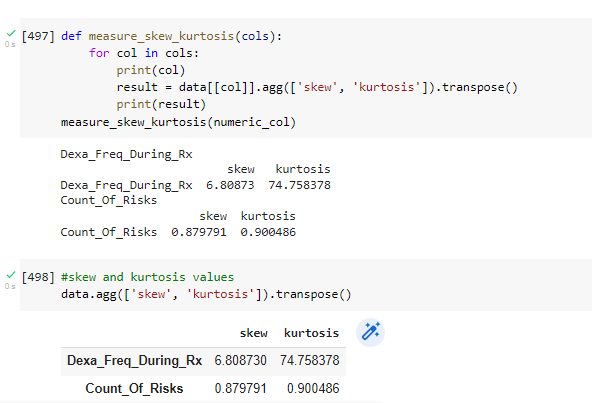


# Data Problems

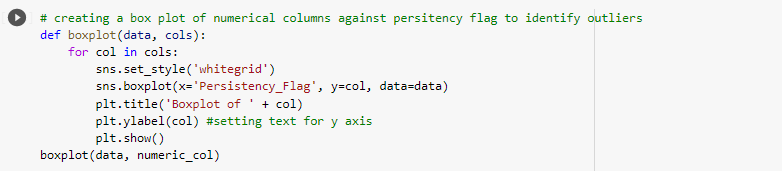
**Data problems such as irrelevant columns, Null values, duplicates, skewed data, outliers and many others may cause bad predictions …**

**So we need to check if we have one of them to know then how to overcome it .**

* **Skewed Data :**

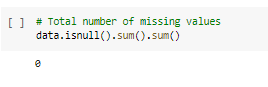


* **Outliers**



* **Duplicates**
* **There is no duplicates , Having duplicates leads often to overfitting**

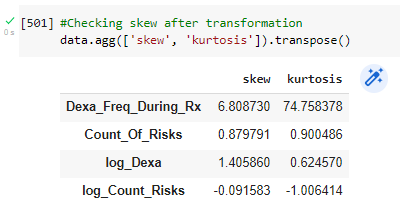
* **Missing Values**



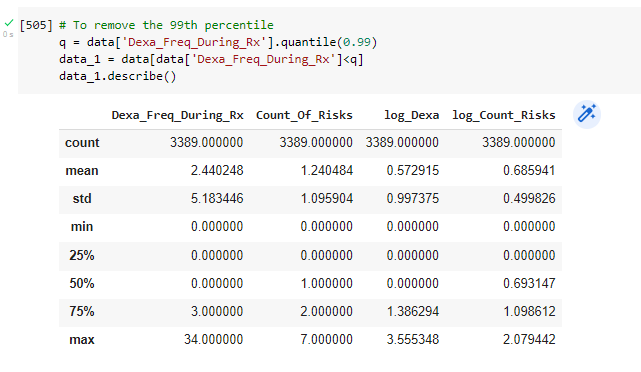
* **No missing Values**

**Solutions**

* **Removing duplicates if they exists**
* **Dropping unsignificant columns**
* **Eliminating Skewed data**



**Example of removing 99% Percentile**



* **Removing outliers**

