W9 - Data Cleansing and Transformation

Group details

Group Name: DS HealthCare group

Team member:

Name	Email	Country	College	Specialization
Yuchi Chen	ychen306@ur.rochester.edu	United States	University of Rochester	Data Science
Bolutife Akinlawon	bolu.akinlawon@gmail.com	United Kingdom	UWE, Bristol	Data Science
Alexis Collier	colliera75@gmail.com	United States	Emory University Bootcamp	Data Science
Han-Fu Lin	hanfu.lin@mail.utoronto.ca	Canada	University of Totonto	Data science
Runtian Wang	a446578875@gmail.com	United States	Clark University	Data science

Github Repo link: [Healthcare Persistency of a drug]

Problem description

A pharmaceutical company conducts a large number of clinical trials in order to study the durability of a new drug. These trials record a large number of different attributes of experimental subjects and the results of the experiment by means of control variables. The company wants to use the data to understand what properties affect the drug's durability.

Yuchi Chen

No missing value - no action needed for NA

Dexa_Freq_During_Rx

Distribution:

- right skew but no outliers
- zero for most frequencies and a normal distribution for other values

Methods:

- Log transformation
- Bin
- Capping

After transformation: feature can be divided by 0 and normal distribution