# Pre-Analysis, Data Cleaning and Transformation, EDA Notes

## Pre-Analysis

The original dataset was extracted from the US FDA webpage. It consists on 3 files in txt format:

* exclusivity\_raw.txt
* patents\_raw.txt
* products\_raw.txt

Being my first health data project, it seems an interesting dataset to start with as it feels manageable and sufficiently challenging. I hope this analysis brings impactful insights on patents associated with FDA-approved drugs.

In order to come up with insightful questions to answer when analysing the dataset is crucial to understand all data fields. For that reason, we first have revised the metadata provided by the FDA (see Metadata\_FDA\_Orange\_Book.docx) where we have made some notes and have come up with interesting questions to answer:

* How many NDAs are associated with innovators / generics products?
* How many different NDA are registered at the FDA orange book?
* How many different products (product numbers) are normally included in NDA (mode)? It is different for innovators or generics?
* Which are the most popular post-approval exclusivities granted by the FDA?
* Do certain exclusivities tend to be granted together? (e.g., do some applications have multiple exclusivity types?)
* Which exclusivity codes are found only in innovator applications?
* Which exclusivity codes are found only in generic applications?
* How many exclusivity periods expire each year in the dataset?
* What percentage of applications receive exclusivity? (for this I need those applications not appearing at the exclusivities table)

## Data Cleaning and Transformation

The next step was to dive into the dataset. To explore the dataset we opted to use pandas, a python library very useful for data manipulation, cleaning and transformation. See:

* pre\_Analysis\_cleaning\_transformation\_ex.ipynb
* pre\_Analysis\_cleaning\_transformation\_pat.ipynb
* pre\_Analysis\_cleaning\_transformation\_prod.ipynb

## EDA (Exploratory Data Analysis)