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

## Pre-Analysis

The original dataset was extracted from the US FDA webpage on the 13th February 2025. 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 and exclusivities 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 have revised the metadata provided by the FDA (see Metadata\_FDA\_Orange\_Book.docx) where we have made some notes and pointed out interesting questions to answer:

Products table:

* *Mode/median/mean number of API in the composition of FDA-approved drugs. Is it different between innovators and generics? Is it different between Rx and OTC?*
* *What percentage represent combination drugs (multiple APIs in their composition) with respect to the total FDA-Approved product? Is it different between innovators and generics? Is it different between Rx and OTC?*
* *Which are the most popular APIs in FDA-approved products’ composition? Is it different between innovators and generics? Is it different between Rx and OTC?*
* *Which are the most common dosage forms and administration routes in FDA-approved drugs? Is it different between innovators and generics?* *Is it different between Rx and OTC?*
* *Which are the firms owning the highest quantity of FDA-approved products (innovators/generics/Rx/OTC/)? And by Dosage Form? And by Route of administration?*
* *What percentage represent innovators and generics with respect to the total number of FDA-approved products? And within prescription drugs? Or within OTC? Which percentage represents Rx and OTC within innovators? And within generics?*
* *How many different new drug applications (NDA or ANDAs) are registered at the FDA orange book at this moment*?
* *How many different products (product numbers) are normally included in a new drug application that has been approved by the FDA (mode, mean, median)? Is it different for innovators and generics?* *Is it different between Rx and OTC?*
* *What percentage of FDA-approved generics are considered therapeutically equivalent to their reference listed drug (RLD)? Which percentage is not?*
* *Which has been the FDA approval tendency over the past years?*
* *What percentage of RLDs are considered RS by the FDA?*
* *Which firms have the highest numbers of FDA discontinued drugs?*

Patents table:

* *How many NDA’s are registered at the FDA Orange book at this moment and have an associated patent?*
* *Which is the total number of patents from FDA-approved drugs?*
* *What percentage of patents have been requested to be to delisted by the applicant?*
* *Mean/median/mode patent duration.*
* *Which has been the patent submission tendency throughout the years?*
* *Which has been the patent expiration tendency over the years?*

Exclusivities table:

* *Which is the percentage of innovators and generics with respect to the total new drug applications with a granted exclusivity?*
* *What’s the total number of new drug applications (NDA or ANDAs) with a granted FDA exclusivity?*
* *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 exclusivities are typical from innovator applications? And from generics?*
* *How many exclusivities expire each year in the dataset?*

## Data Cleaning and Transformation

The next step is to dive into the dataset. To explore the dataset we have 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)

Once we have cleaned and transformed our dataset, we will perform univariate and multivariate analysis on our columns. For this aim, we will choose the most popular data visualization libraries in python: matplotlib and seaborn.

Answers to my questions:

Products table (see EDA\_prod.ipynb):

* *Mode/median/mean number of API in the composition of FDA-approved drugs. Is it different between innovators and generics? Is it different between Rx and OTC?*  There is NO difference between the number of APIs in the composition of FDA-approved innovators and generic drugs. Same when comparing Rx and OTCs. The mean, median and mode is 1 in every group of drugs, which means the majority of FDA-approved drugs are composed by just one API and there is a gaussian distribution.
* *What percentage represent combination drugs (multiple APIs in their composition) with respect to the total FDA-Approved products? Is it different between innovators and generics? Is it different between Rx and OTC?* Only 11% of FDA-approved drugs are combination drugs. FDA-approved innovator combination drugs represent the 16% of all FDA-approved innovators. Within generics, combination drugs are less common (10%). Similarly, combination drugs are more common between FDA-approved OTCs (17%) than Rx (11%).
* *Which are the most popular ingredient in FDA-approved products? Is it different between innovators and generics? Is it different between Rx and OTC? The most popular ingredient within FDA-approved drugs is* ***Levothyroxine sodium*** *with 207 products. It is commonly used for treating hypothyroidism, an extended disease characterised by low secretion of thyroid hormone, essential for regulating several vital functions (cardiac frequency, calories consumption, etc.). However, this API is closely followed by* ***Pregabalin*** *(200 products) widely used for reducing neuropathic pain due to damaged nerves (very common in diabetic patients).*

*The most popular ingredient within FDA-approved innovators is still* ***Levothyroxine sodium*** *(121 products). However, the second most popular is the combination of dextrose, sodium chloride and potassium chloride (92 products), which corresponds to* ***typical electrolyte solutions*** *with carbohydrates, which are essential for feeding hospitalized patients. The third one is* ***Methylphenidate Hydrochloride*** *(42 products), a Central Nervous System stimulator commonly used for the treatment of Attention Deficit Hyperactivity Disease (ADHD).*

*The most popular ingredient within FDA-approved generics is* ***Pregabalin*** *(188 products) very closely followed by a combination of* ***Amphetamines*** *(187 products) (Amphetamine aspartate, Amphetamine sulphate, Dextroamphetamine saccharate, Dextroamphetamine sulphate) used for treating Attention Deficit Hyperactivity Disorder (ADHD) and narcolepsy.*

*Finally, if we talk about FDA-approved Rx, the most popular ingredients are* ***Levothyroxine sodium*** *(207 products) and* ***Pregabalin*** *(200 products). However, within FDA-approved OTCs, the most popular ingredients are* ***Cetirizine hydrochloride*** *(93 products), which is indicated for the treatment of hay fever symptoms (pollen allergy, dust allergy, etc.),* ***Nicotine polacrilex*** *(70 products), which is a common component of medicated chewing gums for quitting smoking and* ***Ibuprofen*** *(62 products) a well-known anti-inflammatory substance.*

* *Which are the most common dosage forms and administration routes in FDA-approved drugs? Is it different between innovators and generics? Is it different between Rx and OTC?*

*In general terms,* ***tablets is the most common dosage form, followed by injectables.*** *However, the difference between both forms is more accused within FDA-approved generics (44.2% tablets vs 14.0% injectables) than within FDA-approved innovators (26.6% tablets vs 16.7% injectables). Why? Most probably because the development and manufacturing process of tablets is much simpler to transfer and replicate than injectables, which nowadays require more technical resources to ensure an aseptic manufacturing process. Generic manufacturers tend to prefer cost-effective, lower costs and easier regulatory paths when developing their products.*

*On the other hand,* ***FDA-approved Rx drugs follow the generality being tablets the most popular dosage form*** *and injectables the second one. However,* ***if we focus on FDA-approved OTCs****, logically,* ***injectables*** *disappear from the ranking. In fact, they* ***are substituted in the second place by extended release tablets****, which are specially formulated for maintaining stable therapeutic concentrations over longer periods of time. This means tablets occupy the first and second place within this group of drugs.*

*As regards administration routes,* ***orals are clearly the most popular FDA-approved drugs****. The main reasons for this is the well-established and deeply understood formulation of this kind of products (tablets, extended release tablets, capsules, syrups, etc.), their low manufacturing costs, their cost-effective profile and lower regulatory requirements. The second most popular is injection.*

*If we compare FDA-approved innovators and generics,* ***orals are much more popular within generics (73.9%) than within innovators (53.4%)*** *for the same reasons. The second place is occupied by injection products in both groups with approximately a 10% of the products.*

*If we compare FDA-approved Rx and OTCs,* ***orals are the most popular in both groups*** *with very similar share around the 70%.**On the other hand,* ***in OTCs injection drugs,*** *which appear at the second place within Rx (12.9%)* ***are substituted by topical products***, representing the 11.6% of OTCs.

* *Which are the firms owning the highest quantity of FDA-approved products (innovators/generics/Rx/OTC)? And by Dosage Form? And by Route of administration?*

***Baxter Healthcare*** *(227, 4.6%),* ***Pfizer*** *(141, 2.9%),* ***Hospira*** *(131, 2.7%),* ***B Braun Medical*** *(128, 2.6%) and* ***AbbVie*** *(115, 2.3%) are the top 5 firms owning the highest number of FDA-approved* ***innovators****. However, since Hospira is now part of Pfizer, in real terms, Pfizer is the top 1 owning the 5.6% of all FDA-approved innovators.*

*The Indian company* ***Aurobindo Pharma LTD****, with 855 products* ***clearly stands out within the U.S. generics market by number of FDA-approved generics.*** *Far away from this company we have* ***Zydus Pharmaceuticals*** *(556),* ***Dr. Reddy’s*** *(397),* ***Alembic*** *(390) and* ***Hikma*** *(384). Aurobindo, Zydus, Alembic are Indian companies. Therefore, there is a strong presence of Indian companies in this market, in fact, only these 3 companies own the 9.3% of the whole FDA-approved generics portfolio.*

***Aurobindo Pharma LTD*** *(799, 3.4%),* ***Zydus Pharmaceuticals*** *(556, 2.4%),* ***Hikma*** *(446, 1.9%),* ***Sandoz*** *(417, 1.8%) and* ***Fresenius Kabi*** *(404, 1.7%) are the top 5 firms owning the highest number of FDA-approved* ***Rx*** drugs.

***Aurobindo Pharma LTD*** *(56, 7.2%),* ***P&L Development*** *(43, 5.5%),* ***Perrigo*** *(39, 5.0%),* ***Haleon*** *(38, 4.9%) and* ***Dr. Reddy’s*** *(34, 4.4%) are the top 5 firms owning the highest number of FDA-approved* ***OTC*** drugs.

Top firms by dosage forms and administration routes will be something to be analysed through the dashboard applying different filters.

* *What percentage represent innovators and generics with respect to the total number of FDA-approved products? And within prescription drugs? Or within OTC? Which percentage represents Rx and OTC within innovators? And within generics?*

*Within the total FDA-approved drugs there is an 80% - 20 % ratio between generics and innovators, being generics the most abundant type of drug application. This ratio keeps unaltered when analysing just Rx products. However, when focusing on OTC, the ratio between generics and innovators changes into 70% - 30%, meaning that meaning that* ***OTC drugs have a higher proportion of innovator products compared to prescription drugs.***

***Within the total FDA-approved drugs there is an 97% - 3 % ratio between prescription (Rx) and OTC****, being Rx the most abundant. This ratio keeps practically unaltered when analysing innovators and generics separately.*

* *How many different new drug applications (NDA or ANDAs) are registered at the FDA orange book at this moment*?

*At the moment, there are* ***26,122*** *drug applications registered at the FDA Orange book.*

* *How many different products (product numbers) are normally included in a new drug application that has been approved by the FDA (mode, mean, median)? Is it different for innovators and generics? Is it different between Rx and OTC?*

***Most of the FDA-approved drug applications include just 1 product*** *(Mode = 1). The mean number of products included in a drug application is equal to 1.8 products. However, the mean is not truly representative of the whole population due to the existence of extreme values (drug applications with 16 different products, for example). In this case, the median is more representative, which coincides with the mode (1.0).*

***These values remain consistent across the different product categories (Generic, Innovator, Rx, OTC).***

* *What percentage of FDA-approved generics are considered therapeutically equivalent to their reference listed drug (RLD)? Which percentage is not?*

*The majority of FDA-approved generics are considered therapeutically equivalent to their RLD (94.6%). However, there is a non-rated 5.1%. In addition, there is a remaining 0.3% of FDA-approved generics which are not considered therapeutically equivalent to their RLD. How is this possible and which are these generics?*

*The* ***0.3% of FDA-approved generics without therapeutic equivalence*** *mostly belong to categories where* ***bioequivalence is harder to establish*** *due to formulation differences, drug complexity, or regulatory concerns. These drugs are* ***still FDA-approved but may not be automatically substitutable*** *at the pharmacy level without additional verification.* ***They are not fully interchangeable.***

*These generics are the following:*

* + *EPIFOAM*
  + *PROCTOFOAM HC*
  + *TACROLIMUS*
  + *DIFLORASONE DIACETATE*
  + *TERCONAZOLE*
  + *VITAMIN K*
  + *METHYLPHENIDATE HYDROCHLORIDE*
  + *PALIPERIDONE*
  + *ACETAZOLAMID*
  + *ALBUTEROL SULFATE*
  + *AMABELZ*
  + *ANDROID 25*
  + *ATORVASTATIN CALCIUM*
  + *BACLOFEN*
  + *CARBAMAZEPINE*
  + *DEXAMETHASONE*
  + *E.E.S. 400*
  + *ERYTHROMYCIN ETHYLSUCCINATE*
  + *FEBUXOSTAT*
  + *FENOFIBRATE*
  + *METHYLTESTOSTERONE*
  + *PREDNISOLONE*
  + *PREDNISONE*
  + *PROPYLTHIOURACIL*
  + *RISEDRONATE SODIUM*
  + *SILDENAFIL CITRATE*
  + *TADALAFIL*
  + *BLISOVI FE 1.5/30*
* *Which has been the FDA approval tendency over the past years?*

*Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.*

*Over the past 5 years, FDA approval of innovator drugs (orange) has been more or less constant, with an average of 10 innovator drug approvals per month. April 2020, May 2021, April 2023 and March 2024 have been the top months while December 2023 and April 2024 have been the worst months.*

*As regards generic drugs (blue), FDA approvals show much more variability month-to-month. On average, the FDA has approved 63 generic drugs per month. February 2023 and July 2023 have been the top months by far. The worst months in terms of generic approvals have been August 2021 and November 2021.*

*What about the FDA approval tendency over the last 20 years?*

*Interfaz de usuario gráfica, Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.*

*The FDA approval of generic drugs (blue line) has shown a consistent upward trend over the past 20 years. A significant shift is observed in the second quarter of 2015, where the number of ANDA approvals per month increases sharply. To quantify this change, we compare the average monthly approvals across two periods:*

*January 2000 - March 2015: 33 ANDA approvals/month*

*April 2015 - December 2024: 64 ANDA approvals/month*

*This nearly twofold increase suggests a structural shift in the approval process, potentially influenced by regulatory changes or increased industry activity.*

* *What percentage of RLDs are considered RS by the FDA?*

*The 40% of the total Reference Listed Drugs have been selected by the FDA as Reference Standards for bioequivalence testing of generic drugs.*

* *Which firms have the highest numbers of FDA discontinued drugs?*

Watson, Norvium, Hospira and Chartwell are the applicants with more FDA- discontinued drugs.

Patents table (see EDA\_pat.ipynb):

* *How many NDA’s are registered at the FDA Orange book at this moment and have an active patent?*

*There are 1,215 different NDA’s registered at the FDA Orange book with active patents.*

* *Which is the total number of patents from FDA-approved drugs registered at this moment?*

*There are 3,715 patents from FDA-approved drugs registered at this moment at the FDA Orange Book.*

* *Gráfico, Gráfico de barras

  El contenido generado por IA puede ser incorrecto.Which has been the patent submission tendency over the last years? With the available data, it is not possible to answer this question. The patents table from the FDA Orange Book just includes those patents that remain active. Therefore, as we can see in the following graphic, there is a present-day bias: the most frequent submission years are the most recent because of the removal of inactive patents, with less recent submission dates.*
* *How much patents will expire over the next years? As well as with patent submission rates, when we analyse patent expiration rates there is a present-day bias. As we can see in the following graphic, the patent expiration rate follows a negative trend: in general, recent expiration dates are more frequent than distant expiration dates.*

*Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.*

* *What percentage of patents have been requested to be to delisted by the applicant? Which applicants have requested to delist patents the most (take into account the total number of patents per applicant)?*

*Only the 0.44% of the registered patents were requested to be delisted from the database.*

* *Mean/median/mode patent duration.*

*The mean patent duration is equal to 11.9 years. The median is 12.3 years and the most repeated patent duration is equal to 13.9 years.*

Exclusivities table (see EDA\_ex.ipynb):

* *Which is the percentage of innovators and generics with respect to the total new drug applications with a granted exclusivity?*

Gráfico, Gráfico circular

El contenido generado por IA puede ser incorrecto.*The majority of drug applications with a granted exclusivity are Innovators (96.1%). Generics represent almost the 4% of drug applications with granted exclusivity.*

* *What’s the total number of new drug applications (NDA or ANDAs) with a granted FDA exclusivity?*

*There are 652 different drug applications with associated exclusivities.*

* *Which are the most popular post-approval exclusivities granted by the FDA?*

*Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.The most popular post-approval exclusivity granted by the FDA is the* ***Orphan Drug*** *Exclusivity with 660 exclusivities. It is followed by the* ***New Chemical Entity*** *Exclusivity and the* ***Pediatric*** *Exclusivity (278). The least frequent exclusivities are the Rx to OTC Switch or OTC use, the New Chemical Enantiomer and the New Combination with just one exclusivity granted*

* *Do certain exclusivities tend to be granted together?*

*The* ***Pediatric*** *and* ***Orphan Drug*** *exclusivities are very commonly granted together. In fact, it is the most recurrent combination of exclusivities being granted for a drug application with 978 co-occurrences, with a considerable difference from the rest of possible combinations.*

*This insight is particularly relevant for understanding patterns in exclusivity grants. The****Pediatric Exclusivity****is typically granted to incentivize studies on drug effects in children, while the****Orphan Drug Exclusivity****is aimed at encouraging the development of treatments for rare diseases. The frequent co-occurrence may reflect a trend where drugs targeting rare diseases are also studied for pediatric use, possibly due to overlapping regulatory incentives or market needs.*

*Imagen que contiene Calendario

El contenido generado por IA puede ser incorrecto.*

* *Which exclusivities are typical from innovator applications? And from generics?*

*All exclusivities are typically granted to new drug applications (innovator drugs) except from the* ***Competitive Generic*** *exclusivity and the* ***Patent Challenge*** *exclusivity, which are generic-specific.*

* *How many exclusivities expire each year in the dataset?*

*The majority of exclusivities have expired during 2024 (up to 500 of them). Moreover, the majority of the remaining exclusivities will expired between 2025 and 2027.*

*Gráfico, Gráfico de barras, Histograma

El contenido generado por IA puede ser incorrecto.*