ADS-AExploratory Data Analysis

korte regel

Mai Linh Luong  
Mitchel Kuijpers  
Nikita Gavrilov  
Huy Nguyen  
Jak Benev  
1 December 2020  
V1.0  
Status: in process

# Version History

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Description** | **Name** |
| V0.1 | 17-09-2020 | * Setting up the document | Mitchel Kuijpers |
| V0.2 | 24-09-2020 | * 1.Question Brainstorm | Mitchel Kuijpers  Mai Linh Luong  Nikita Gavrilov |
| V0.3 | 16-10-2020 | **Setup of chapters:**   * 2.Exploratory Data Analysis * 2.1.Regions * 2.2.Genders & Age * 2.3.Season * 2.4.Medication Products | Mitchel Kuijpers  Mai Linh Luong |
| V0.4 | 16-10-2020 | **Setup of chapters:**   * 3. Region 22 * 4. Region 40 * 5. Region 30 * 6. Region 10 * 7. Region 20 | Mai Linh Luong |
| V0.5 | 18-10-2020 | * Region 22: Chapters 3.1 and 3.2 made | Mitchel Kuijpers |
| V0.6 | 20-10-2020 | * Regions, gender & age, medication products of region 40 | Mai Linh Luong |
| V0.7 | 21-10-2020 | * Region 40: Overall explanation * Region 40: Seasons | Mai Linh Luong |
| V1.0 | 1-12-2020 | * Last Check on document | Mai Linh Luong |

# 

# Table of Contents

[**Version History**](#_nh6u8xktfknd) **1**

[**Table of Contents**](#_yvoio28jfj8a) **2**

[**1.Data Cleaning**](#_plvge6qi6idi) **3**

[1.1.Data Cleaning](#_qcryjxk3nom2) 3

[1.1.1.Categorizing](#_5ij8ejz34viw) 4

[**2.Exploratory Data Analysis**](#_6tsqu22nigky) **5**

[2.1.Regions](#_ih2catktwl7y) 5

[2.2.Gender & Age](#_xc36z9l1tdlv) 5

[2.3.Season](#_ltxur68wkas8) 5

[2.4. Medication products](#_5jec9p8jldyh) 5

[**3. Region 22**](#_5e4mga2refav) **6**

[3.1.Analysis](#_nx99uccg9h35) 6

[3.1.1.Prices & Contribution](#_asd0a5wddpy) 6

[3.1.2.Seasonal](#_b8w11wso2xnq) 7

[3.1.3.Product Categories](#_1r8yohsjqid2) 8

[3.1.4.Age & Gender](#_tlm740eiqt6d) 9

[**4. Region 40**](#_ron4ewxpqywi) **11**

[Overall exploration](#_x1t63ufidu9v) 11

[4.1 Regions](#_g99jufbhacsa) 11

[4.2 Gender & Age](#_j7mv2exgjx3w) 12

[4.3 Season](#_vwe91we8jvuu) 13

[4.4 Medication Products](#_6h98hs6exusq) 13

[**5. Region 30**](#_f5l4zdn15go2) **15**

[5.1 Gender](#_ukfe7phsrouw) 15

[5.2 Season](#_gn11osvzz63f) 15

[5.3 Products](#_4rmpfb1ivhej) 16

[**7. Region 20**](#_c1f54azgi6bd) **17**

[**7.1 Product Type Sample**](#_cr48fc5lscvz) **17**

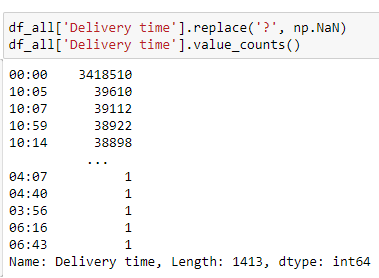
# 1.Data Cleaning

## 1.1.Data Cleaning

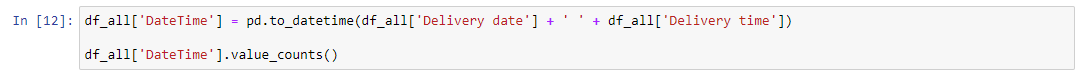
To be able to use the dataset in the best way possible, some data cleaning has been done before making the analysis. While cleaning the data there was some interesting data found that could be used for the analysis. These findings can be found below:

* **Delivery Data & Time**

In the image below we can see that there are alot of delivery times that are not in the correct time.

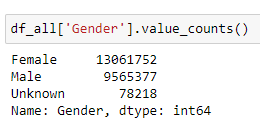
****

The Time and date have also been merged together in another column called Datetime. This makes time and date related analysis much easier to make. This is seen in the image below:



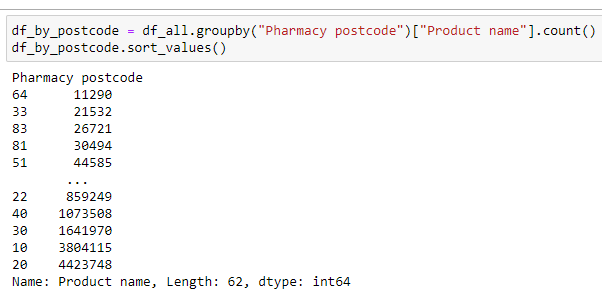
* **Amount of males vs females**

Another thing that was found is the amount of males and females that buy products at pharmacies. The dataset shows us that female customers buy more medications than males.



* **Top 5 regions**

We have decided to split the work on the dataset so that every project member could do their part of the analysis. The best way for us to split the dataset was by regional codes. This is why we took the top 5 regions with the most products sold.



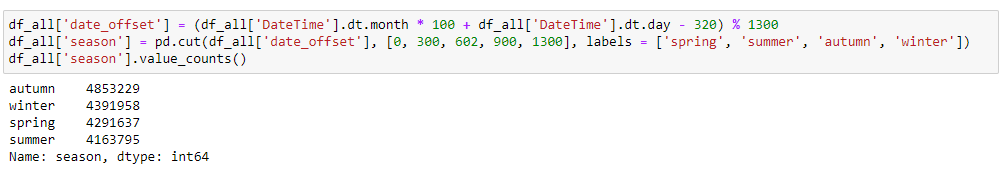
### 1.1.1.Categorizing

**Product Category**

With the help of a product category we can see what products are the most sold or what types of products are shown in the dataset.

**Season Category**

With the help of seasonal categories we can look at for example what season has the most sales or has the fewest sales.

****

With the code above we can conclude that the autumn season has the most products sold in total. After that we can see winter has the second most products sold in that season.

### 

# 2.Exploratory Data Analysis

In this chapter we will tell more about how we analyzed the data in an exploratory way. This way we can get to know the datasets we are going to use better for a better data story. With this analysis we are going to discover patterns, find anomalies, test hypotheses and check for assumptions with the help of statistics and graphical visualisations.

## 2.1.Regions

For region we want to take a look, which region sells the most medications. Based on the top 5 most sold regions we divide the region and see which products are sold the most in a certain region.

* What are the top 5 regions where products are being sold?
* By analyzing most sold medicine we can understand what illnesses appears the most in the region

## 2.2.Gender & Age

We want to see which gender would most likely buy the medications more. We can divide gender into age groups.

* Which gender would most likely buy medications more and which products
* Which gender in what age group would most likely buy medications more

## 2.3.Season

We want to see in which season, which product is sold the most. For this we want to see per region.

* Most sold medicine (In which season of the year does this medicine get sold the most)
* Is there a correlation between the current season (summer, winter, etc) and the types of medications that are being sold the most.

## 2.4. Medication products

We want to see which products are the most popular and divide the medications with honorarium.

# 3. Region 22

## 3.1.Analysis

The part that is analysed in this chapter is pharmacy postcode 22. This postcode lies in the region of Antwerp.

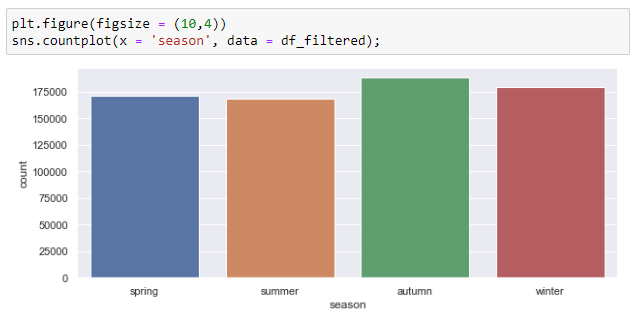
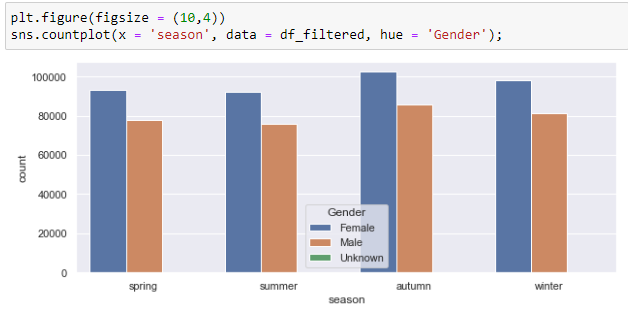
### 3.1.1.Prices & Contribution



### 

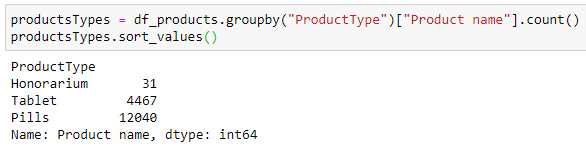
### 3.1.2.Seasonal

In this chapter we will take a look at all the data we can find that is related to the four seasons. As we can see there is not much of a difference between the total products that have been sold per season. in the data below.

The barplot below indicates that there is not much difference in the amount of products sold per season. It does show that in the autumn season more has been bought.This data can also be split per gender type as seen in the plot below. This plot indicates that female customers have been buying more than male customers in every season.  


### 3.1.3.Product Categories

The first thing we want to see for the product categories is what product is being sold the most. Because of the data not being fully filtered yet, there could be some data that is not correct yet.

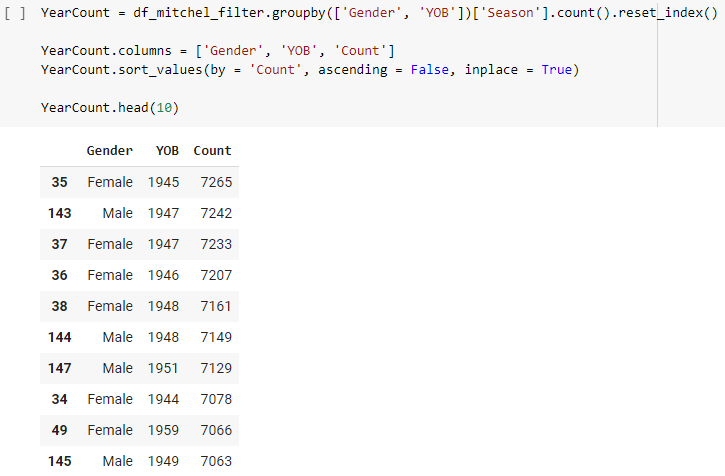




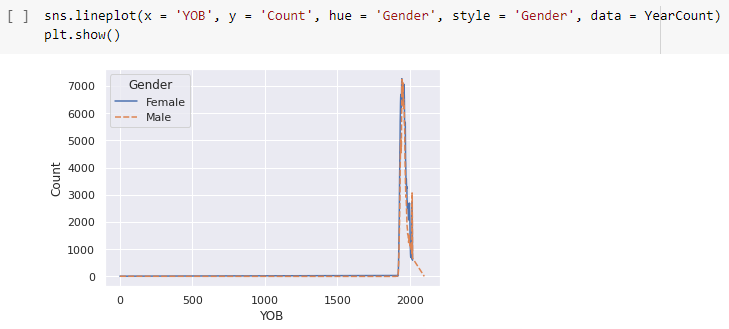
### 

### 3.1.4.Age & Gender

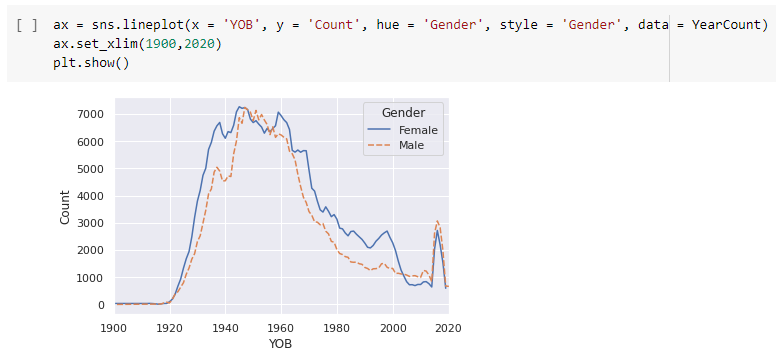
For the genders we would like to see how old the customers are. This could help us see the average age of the customers and predict what products are bought mostly by specific age groups.



All the customers have been grouped per year and gender. Now we are able to put this in a plot and see what years the customers are born the most frequently.



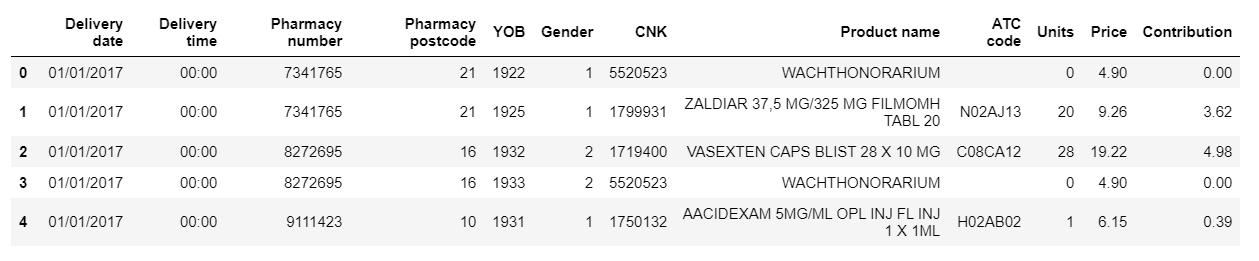
As we can see, the ages are not entirely correct. This means we will only look at the ages that could be alive at the moment. This can be seen in the plot below:



# 

# 4. Region 40

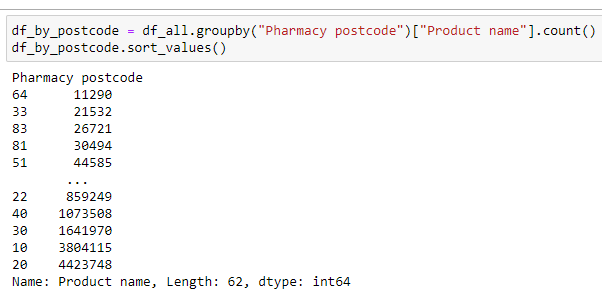
## Overall exploration



The dataset, focused on region 40 has 12 columns. There are 348134 null values in delivery time. The delivery date and delivery time’s type are objects so this has to be changed to date-time. Price and contributions are floats.

## 4.1 Regions

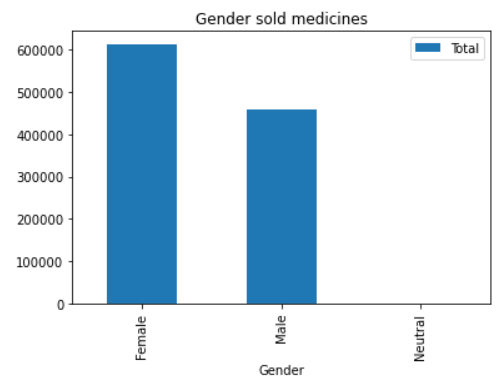
To divide the work, we first had to take a look where most of the medications are being sold in which region. In this case, the top 5 regions are: number 20, 10, 30, 40 and 22. I had to do my EDA on region 40.



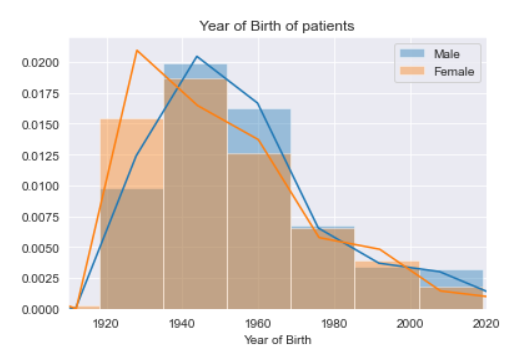
The postcode (region) 40 is from the province of Liege, the capital city Liege. The population is quite high, Liege is ranking 4th as the most populated city in Belgium with 197355 people living in the city in 2019. That is the reason why the amount of sold products is quite high in this region.

## 4.2 Gender & Age

Females would more likely buy medications than males according to the dataset. The amount of products sold to females is 614182, while the amount of products sold to males were only 458473.

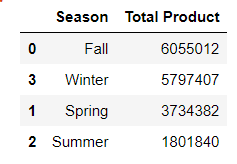


But looking at the age of these patients, male patients in the year of birth of 1940 and 1960 will most likely buy medications more than female from this year of birth in this region. Another remarkable thing is that females who are born in the 1920 would, would most likely buy medicines more than males.



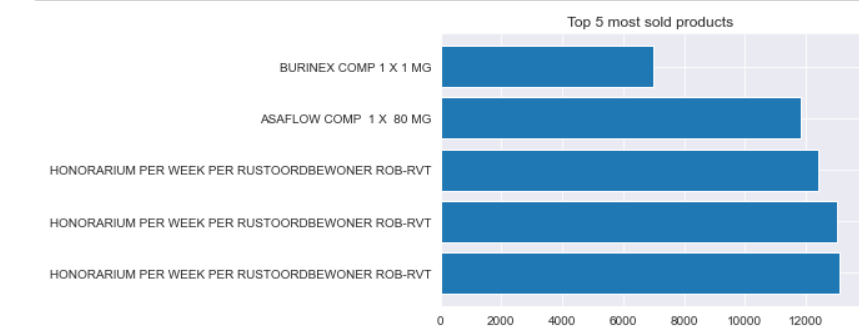
## 4.3 Season

We want to see in which season, the medications are being sold the most. In this case, it’s the fall season where the pharmacy sells most of its products. As you can see in the table below, the hotter the weather, the less likely people would buy medications. In the summer, the pharmacy doesn’t sell a lot of medications compared to the fall. But then again, in the fall season people would most likely buy medications more than winter.

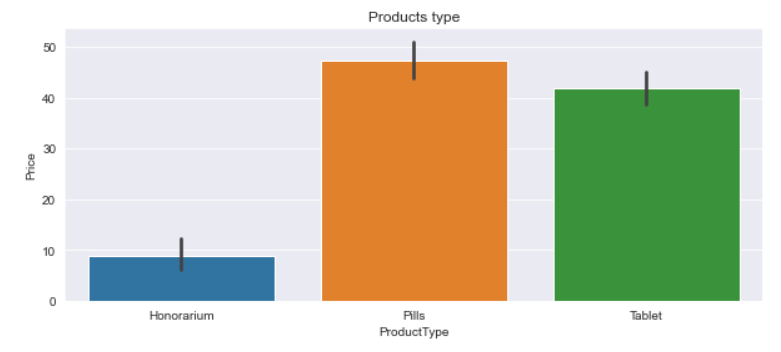


## 4.4 Medication Products

Looking at the top 5 popular products within the pharmacy that the pharmacist sells the most may help us to see why in particular these products.



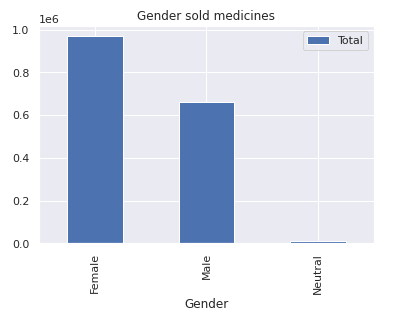
To dive more into the medications, we want to see what kind of medications are popular in this pharmacy region. It seems that pills are the most popular and brought more than tablets.



# 5. Region 30

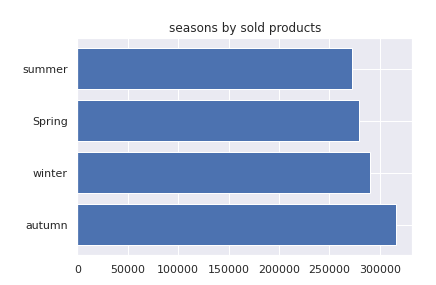
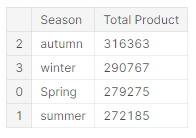
In this chapter I am going to describe the postcode of 30 which is Leuven which is the capital and largest city of the province of Flemish Brabant in the Flemish Region of Belgium.

## 5.1 **Gender**



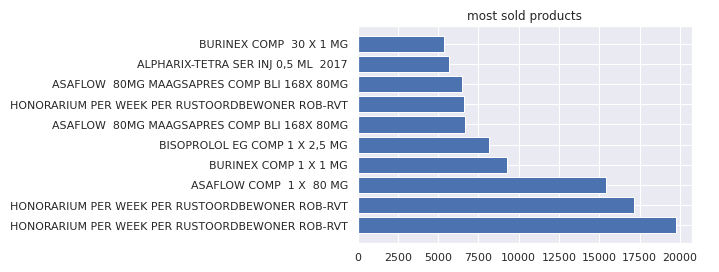
The graph shows the distribution in purchasing the medicine between Female, male and neutral genders. As you can see in the region of Leuven women purchased 30 thousand more medicine than men.

## 5.2 Season



The above graph shows when the purchase of the medicine is usually accurate. The figures are similar for each of the seasons. However, during the summer people purchase the least amount of medicine. We can assume that people are less likely to get ill because the weather is warm and dry compared to another season.

## 5.3 Products

****

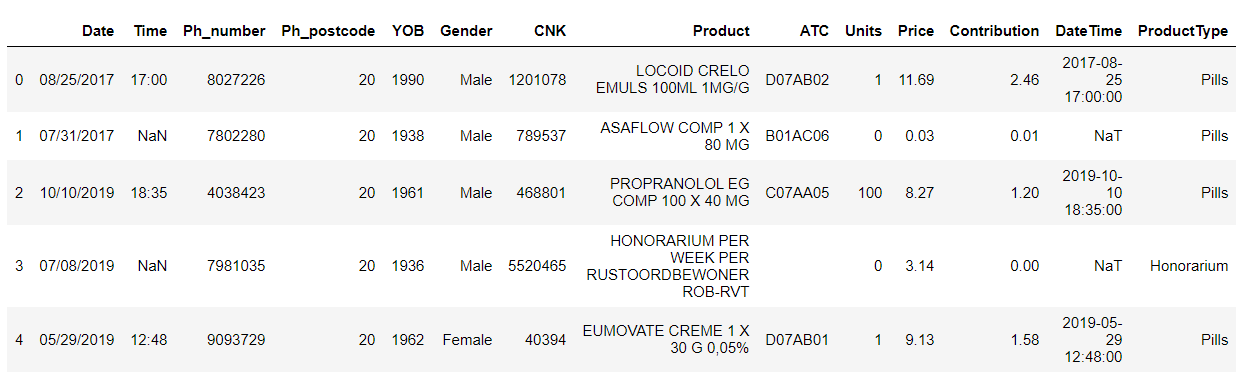
If you look at the graph you can clearly see that the most sold medicine by far is “Honorarium per week per rustoord bewoner”, however, it is not particularly a medicine. “Honorarium” is a fee per week for the rest home. Thus, it is not in the scope of this chapter.

In the region of Leuven, the most sold medicines are Bisoprolol for heart diseases, Asaflow (in different variations)which is also preventing some cardiovascular diseases, Burinex is a medicine used to treat swelling and high blood pressure, Alphrix-tetra medicine which is used against flu.

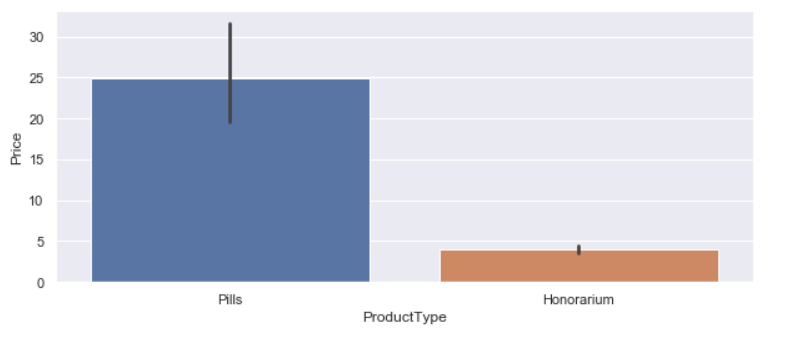
# 7. Region 20

# **7.1 Product Type Sample**

I had to take the region 20 and since it contains too much data so I only filter by the first 1000 samples and sort it by the Product Type.



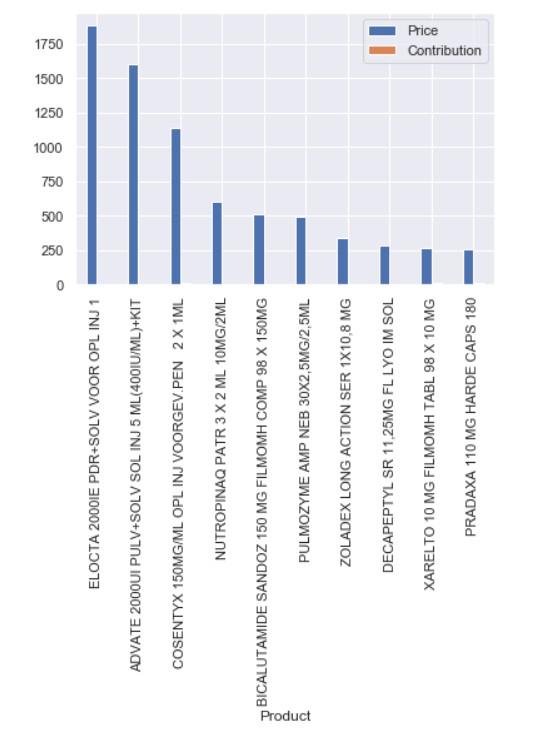
**7.2 Comparison between Product Type**



According to the graph, we can see that the price of pills is higher than the honorarium product.

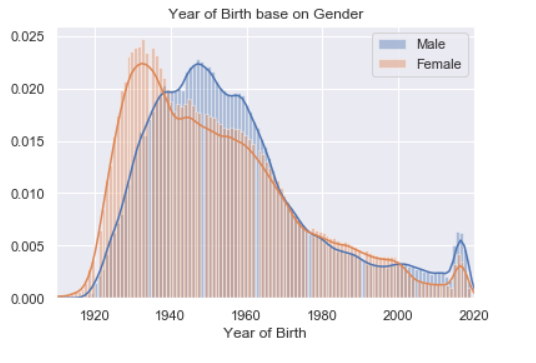
**7.3 Most expensive products:**

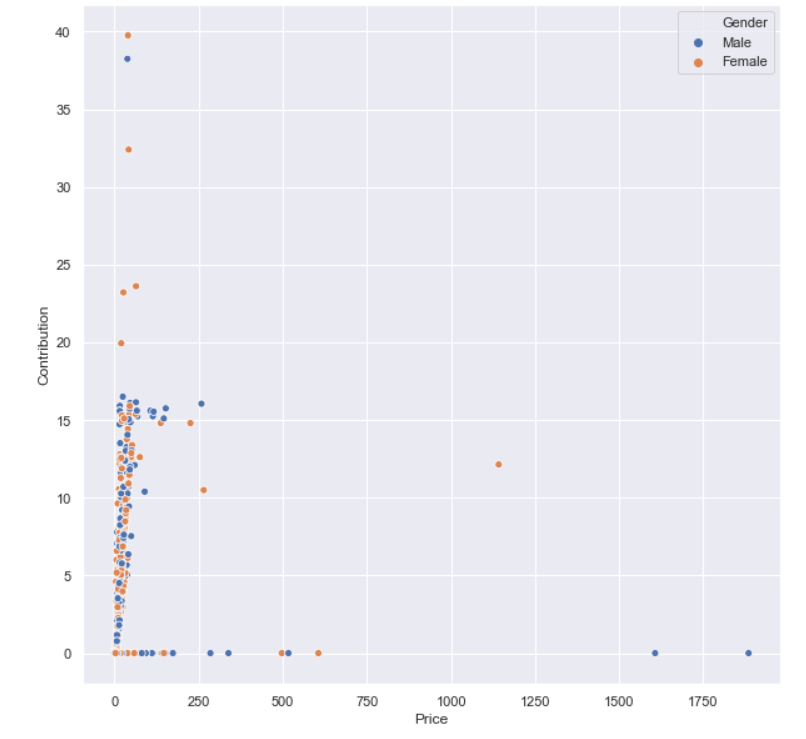
The chart below noted on which product has the highest income. The ten most expensive products have a price range between 250 and 1750.



**7.4 Year of birth base on Gender**

Looking at the age of these patients from region 20, male patients in the year of birth of 1940 and 1960 will most likely buy medications more than female from this year of birth in this region. Another remarkable thing is that females who are born in the 1920 would, would most likely buy medicines more than males.





**7.5 Timelines**

There is a significant rise in sales of products in 05-2019. Another huge change in selling products is between May 2017 and September 2017.

