ADS-AData Analysis Report

korte regel

Mai Linh Luong  
Mitchel Kuijpers  
Nikita Gavrilov  
Huy Nguyen  
  
1 December 2020  
V1.0  
Status: Complete

# Preface

The data analysis report before you has been made by following the business proposal which describes the steps chosen leading to the result of this report. The report will show the recommendations and conclusions we have made following the exploratory data analysis (EDA) and predictive analysis.

Before any analysis can be done, data has been changed by cleaning the data. This way the data can be used to make predictions with the help of machine learning and predictive analytics, which will support the beneficial recommendations and conclusions given with this data analysis report.

In addition, several other important elements, such as ethical considerations and approaching methods for machine learning, will be addressed in order to obtain the best results. Everything is being done to ensure that customers and their clients have a better business insight into their data.

# Table of Contents

[**Preface**](#_7vs8ivbgmpd2) **1**

[**Table of Contents**](#_34n7clfrye88) **2**

[**Version History**](#_bp1sujlea1k2) **4**

[**Introduction**](#_9wazzj2iwkkm) **5**

[**1. Method and Approach**](#_m00kagl1dyc) **6**

[**1.Predictive Analytics**](#_dgwb1hb8c1wn) **7**

[1.1.Age & Gender](#_9fwdg26mav59) 7

[1.2.Seasonal](#_sc6hl83xvlzq) 8

[1.3.Prices & Contribution](#_demd23uh3tql) 8

[**2.Prescriptive Analytics**](#_reitkzyjfe67) **10**

[2.1.Age & Gender](#_evo65i6v5i68) 10

[2.2.Seasonal](#_fpq8e9pq4try) 10

[2.3.Prices & Contribution](#_py6jbtrgbzpc) 11

[**3.Data Quality Improvements**](#_udr0xac5iy9l) **12**

[3.1.Data Cleaning](#_qcryjxk3nom2) 12

[3.1.1.Categorizing](#_5ij8ejz34viw) 13

[**4.Machine Learning Results**](#_lsy43j6y7k9u) **14**

[4.1.Age & Gender](#_9gcm8nrp81u5) 14

[4.1.1.Region 22](#_2prcbe6tid5j) 14

[4.1.2 Region 40](#_o2axz02k43v5) 17

[4.2.Seasons](#_40dzukd56rpy) 22

[4.2.1.Region 22](#_vj1mnwnk6o1g) 22

[4.2.2 Region 40](#_yovr5b65zak8) 25

[4.3 Products](#_ftwcn3ipxtkg) 26

[4.3.1 Region 40](#_zgqemjdabdn) 26

[4.3.2 Region 30 Product Pricing](#_hg64tvg3w4ex) 27

[4.3.3 Region 20: Product Revenue per Year](#_c8xllwrw3ixi) 28

[**5. Ethical Considerations**](#_1my046ligpqk) **30**

[5.1 Beneficial](#_jtkkbot74xt7) 30

[5.2 Progressive](#_g3wa2rm50bje) 30

[5.3.Sustainable](#_s8revfsixxvx) 30

[5.4.Respectful](#_xk11i4bz1ge6) 31

[5.5.Fair](#_h78g6vwwso0k) 31

[**6. Recommendations**](#_79eotlph8qbx) **32**

[6.1 Age & Gender](#_7mnzzwiq3wnl) 32

[6.1.1 Exploratory Data Analysis](#_6bdx6qac9cp2) 32

[6.1.2 Predictive Analytics](#_236gtn7ylesz) 32

[6.1.3 Final Recommendation](#_gikfmkftngdz) 32

[6.2 Seasons](#_jkpep7gbink4) 33

[6.2.1.Exploratory Data Analysis](#_dq87ympq1vhs) 33

[6.2.2.Predictive Analytics](#_j7p3l9uc5hwq) 33

[6.2.3.Final Recommendation](#_twtvp7jt4ozo) 33

[6.3 Products and Pricing](#_1wqkn75e2q48) 34

[6.3.1.Exploratory Data Analysis](#_14ww49xndgcu) 34

[6.2.2.Predictive Analytics](#_r24t62bn2k18) 34

[6.2.3.Final Recommendation](#_xppgjaewa3y9) 34

# 

# Version History

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Description** | **Name** |
| V0.1 | 05-11-2020 | * Setting up the document | Mitchel Kuijpers |
| V0.2 | 10-11-2020 | * Chapters 1-5 setup in the document. | Mitchel Kuijpers |
| V0.3 | 12-11-2020 | * Introductions written for chapters 1, 2, 3, 4, 5 | Mitchel Kuijpers |
| V0.4 | 27-11-2020 | * Chapter 1.1, 1.1.1, 1.1.2, 1.1.3 set up | Mitchel Kuijpers |
| V0.5 | 29-11-2020 | * Chapters 4.1.1, 4.2.1, 4.3.1, 5, 5.1, 5.2, 5.3, 5.4, 5.5 set up * Chapters 4.1.1, 4.2.1 and 4.3.1 made | Mitchel Kuijpers |
| V0.6 | 29-11-2020 | * Chapter 4.1.2, 4.2.2, 4.3.2 | Mai Linh Luong |
| V0.7 | 29-11-2020 | * Rewrote Introduction | Mitchel Kuijpers |
| V0.8 | 11-12-2020 | * Chapters 1.2 written | Mitchel Kuijpers |
| V0.9 | 17-12-2020 | * Chapters 4.1.3 and 4.3.3 written | Huy Nguyen |
| V0.10 | 08-01-2021 | * Preface partially written * Chapter 6.2, 6.2.1, 6.2.2 and 6.2.3 written | Mitchel Kuijpers |
| V0.11 | 09-01-2021 | * Chapter 5 written | Mai Linh Luong |
| V1.0 | 10-01-2021 | * Chapter 3 written * Finalization of document | Mitchel Kuijpers |

# 

# 

# Introduction

After the approval of the Business Proposal and the exploratory data analysis that has been made, phase 2 of the project can begin. Phase 2 is mostly focussed on the quality of the dataset, the subjects that are related to data quality are: Data Requirements, Data Collection, Data Understanding and Data Preparation.

This document will show the data analysis that has been done by Group B for Informa. In the first chapter we will discuss the predictive analytics that have been made with the dataset, chapter two will show the prescriptive analytics made, chapter 3 will show the improvements that had to be made, chapter 4 shows the results of machine learning for the different subjects. The last chapter we talk about the ethical considerations we have made during the predictive analysis.

# 1. Method and Approach

This chapter describes what methods the project group used in general.

**K-nearest neighbor (KNN)** - the k-nearest neighbor algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. KNN is a type of [instance-based learning](https://en.wikipedia.org/wiki/Instance-based_learning), or [lazy learning](https://en.wikipedia.org/wiki/Lazy_learning), where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then [normalizing](https://en.wikipedia.org/wiki/Normalization_(statistics)) the training data can improve its accuracy dramatically.

**Linear regression** - linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeller might want to relate the weights of individuals to their heights using a linear regression model.

**Decision tree** - an algorithm that we use in real life and which has influenced a wide area of machine learning. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. Thus, decision tree is one of the predictive modelling approaches used in statistics, data mining and machine learning. What makes decision trees special in the realm of ML models is really their clarity of information representation. The “knowledge” learned by a decision tree through training is directly formulated into a hierarchical structure. This structure holds and displays the knowledge in such a way that it can easily be understood, even by non-experts.

# 

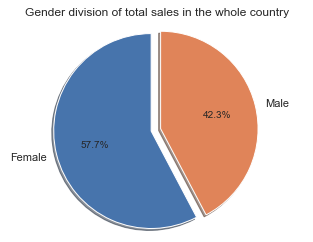
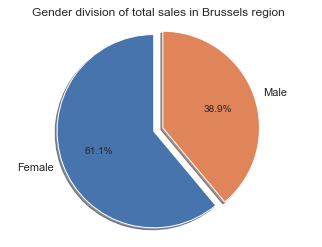
# 1.Predictive Analytics

This chapter will show the predictive analytics done by the project group. The predictive analytics are made with the help of the dataset given by Informa. This dataset has already been cleaned and explored by the project group, so that we get a better understanding of the dataset. In this chapter we will take a look at different subjects that are important to analyse for the stated requirements of the Business Proposal.

## 1.1.Age & Gender

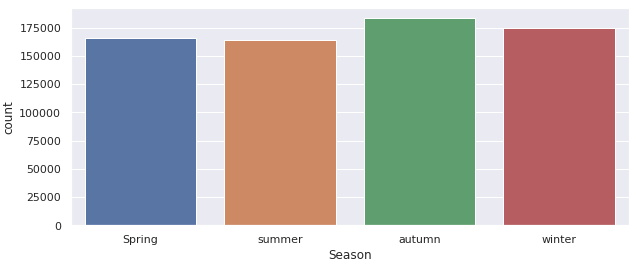
To start things off we’ll be looking into total sales according to gender division.

Initially we will look into the second biggest region in sales - Brussels (the capital) and the whole country.

A small difference of 3.4% towards equal division, however something to keep in mind as we’ll be looking into top clinics in sales in a few pages and the capital has one of the biggest. The following is the total number of sales in the whole country and in brussels, accordingly:

## 1.2.Seasonal

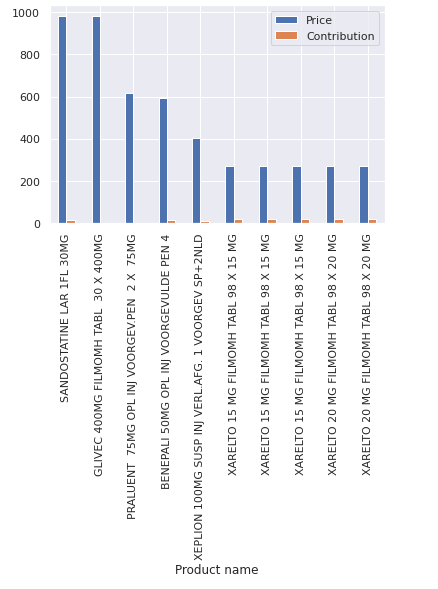
The following subject we will take a look at is the seasonal data from Informa. With this data we can make predictions based on seasonal information, which could give indications on what products or customers with certain ages buy the most in a given season.



According to the image above, the autumn season sells the most product in most of the regions, which has had 175000 sales in total. This could help the predictive analytics on making predictions about what seasons products are sold the most based on certain conditions like age or city.

## 1.3.Prices & Contribution

The project group will take a look at prices and contribution. With data about the amount of money customers spend on the medication we can make a prediction on how prices affect the amount of sold 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.

# 

# 2.Prescriptive Analytics

This chapter is about the prescriptive analytics for the Informa dataset. With prescriptive analytics we will describe the recommended actions that can be taken to further help affect the wanted outcomes in the future.

## 2.1.Age & Gender

According to the dataset, it appears that in every region the females are the ones that would most likely buy medications more. However, if we take a look at region 40, males who are born in 1940 and 2000 would most likely buy medications more than females who are born in 1940 and 2000. It seems that males who are born between the year 1940 and 1970 would most likely buy medications more than females and even in the year 2000 we see a slight growth of the amount of males.

It seems that age plays a big role in whether or not males of females would buy medications. But overall, the pharmacy received more female customers than male customers. And these customers would most likely belong to the senior generation. Therefore, if the pharmacy wants to attract more customers, it’s important that the pharmacy is senior friendly but also the advertising could be more targeted to the older females.

## 2.2.Seasonal

From the dataset of Informa, we determined that the autumn season has the most products sold in total. 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.

With this information, we can use predictive analytics to get an idea of how many more products might be bought if we focus on producing pills during the autumn period of the next year.

Then, using prescriptive analytics, the company can look at scenarios where the costs for a certain type of medication increase, decrease or hold steady. These scenarios then allow them to make an informed decision about how to proceed in a way that’s both cost-effective and beneficial to their customers (for example, prepare more medication products for the autumn or winter). This could save costs on everything from medical supplies to transport fees, etc.

## 2.3.Prices & Contribution

The data set shows that in most of the regions people buy almost the same amount, types and manufacturers of products. Thus, the project group can not find the correlation between the region and diseases that are mostly common for that region. Moreover, people do not tend to save money when it affects their health.  
  
On the other hand, the project group can use prescriptive analytics to advise the company which pricing category of the product is the most popular and which are not. Thus, the company can adjust its pricing policy so it can sell more products of non popular pricing categories.

# 

# 3.Data Quality Improvements

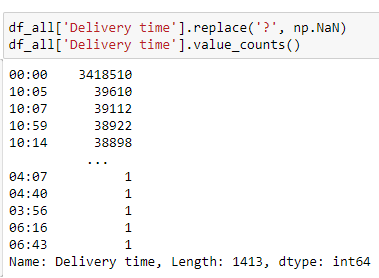
Chapter 3 of the data analysis report will describe the improvements we have made to the data quality, so that it can be used for predictive / prescriptive analytics.

## 3.1.Data Cleaning

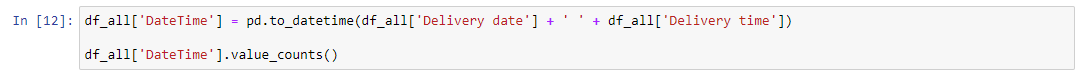
To be able to use the dataset in the best way possible, some data cleaning has been done before making the analysis. While cleaning the data there was some interesting data found that could be used for the analysis. These findings can be found below:

* **Delivery Data & Time**

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

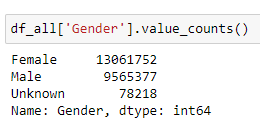
****

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:



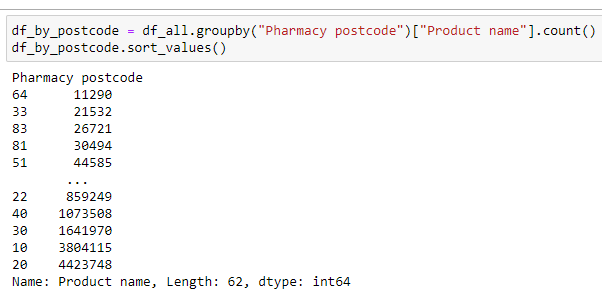
* **Amount of males vs females**

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.



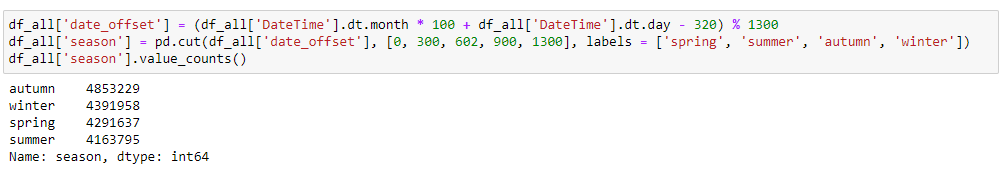
### 3.1.1.Categorizing

**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.

# 4.Machine Learning Results

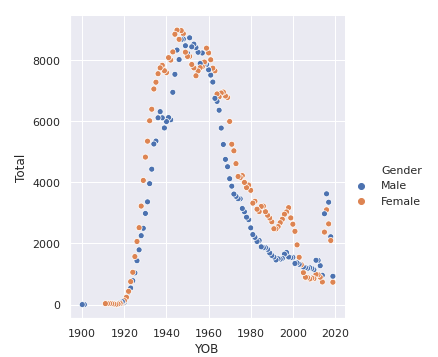
The fourth chapter is about the results we have made during the predictive and prescriptive analyses of the dataset. Here we will discuss the results we have found during these analyses. The analysis is split up into different subjects that have been analysed by each project member. This way we can find more information about the top 5 regions where the most medications are being sold.

## 4.1.Age & Gender

### 4.1.1.Region 22

For the Age & Gender predictive analysis I have made use of linear regression. With this model we take a look at the total sales per year of birth divided by gender. As seen in the image 1 below, we can see the total sales made per age on the scatterplot.

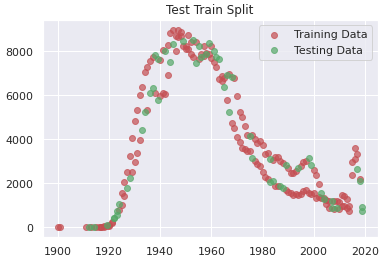
The scatterplot below indicates that males and females both have a high chance of buying medications when they are born between the years of 1930 and 1960, which could help pharmacists on focussing on a specific age group.



*Image 1: Scatterplot*

This data is then divided in testing and training data. In image 2 we can see by color what data is testing data and training data.

In the following image the data is divided by testing and training data. As seen on the image these have been split by two colors. Green being the testing data and red being the training data. To be able to make predictions the testing data will be used shown in the color green.



*Image 2: Test Train Split*

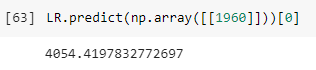
After splitting the data into training and testing data, we can start with the linear regression. For the linear regression we first fit the values of the training data. In image 3 below we can see what the actual test data is with the prediction line of the linear regression.

### 

*Image 3: Actual Test Data*

As we can see from the linear regression, the sales slightly go down depending on the age of the customer. This data shows that the customers born in the years between 1940-1960 would buy the most medications with a slight decrease the younger they are.

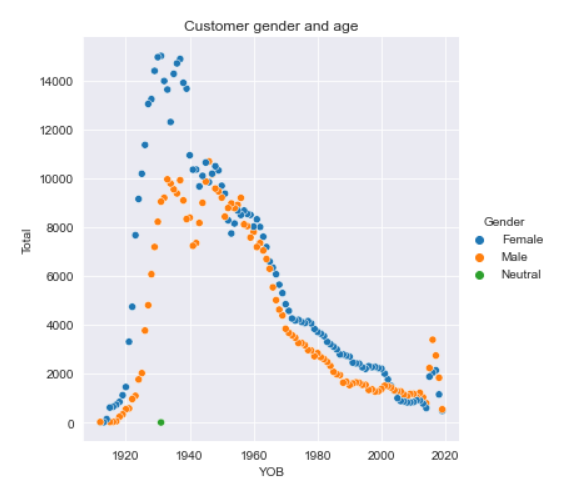
With this data we can look at different types of predictions per year of birth. In the image below we can see how much sales are predicted by customers that are born in the year 1960, which is 4054.



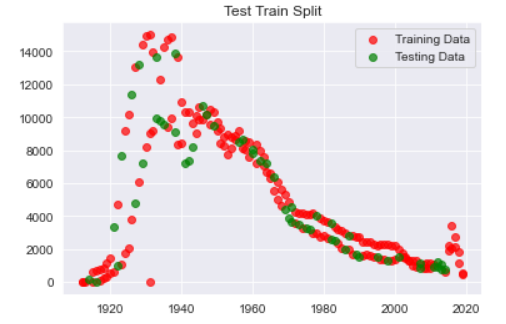
*Image 4: Prediction*

### 4.1.2 Region 40

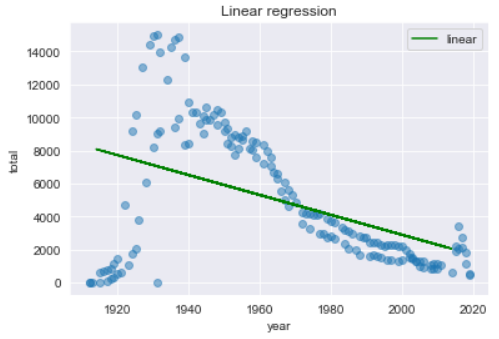
As you can see in image 1, the females that are born between 1920 - 1940 are most likely the ones that buy medications more than males. This can help the pharmacist to focus their advertisement on the females that are born between 1920 and 1940.

  
*Image 1: Custom gender and age*

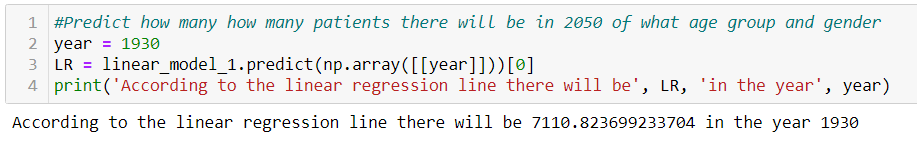
As you can see in image 2, the testing dataset is green and the training data is red. Everything in the red, so training dataset, will be used to create the linear regression model and that line will then be tested on everything in the green, ,so the testing data.

  
*Image 2: Test Train Split*

According to the regression line, the amount of customers will still be the "senior" probably because of the age it's not likely that there is a big population of people who are born in 1930 in the year 2020 for example, but there is a slight increase in for the people born in the year 2000. This is probably because in the future, this age group will be "seniors" and experience more health issues. Therefore they would most likely buy more medications (see image 3).

  
*Image 3: Linear regression*

So as you can see in the prediction we made above, the amount of sales that is going to be for the people that are born in 1930, will be 7395 sales.

**

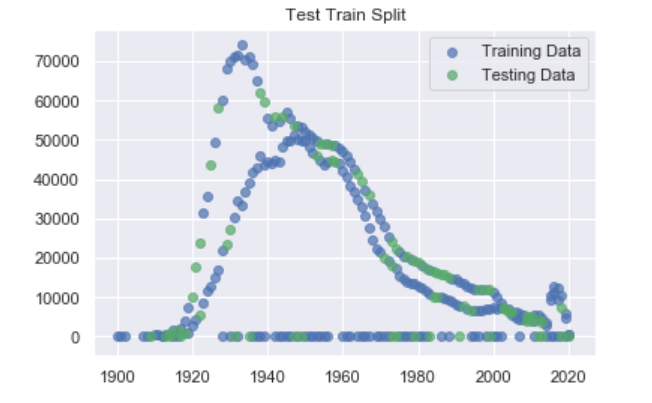
**4.1.3 Region 20**

For the Age & Gender Predictive Analysis, I used linear regression. With this model, we take a look at the total sales for each year of birth divided by gender. As shown in picture 1 below, we can see the total sales made on the scatterplot per age. As you can see, females born between 1920 and 1940 are most likely to buy more medications than males. This may help the pharmacist to focus their advertisement on females born between 1920 and 1940.

## 

*Image 1: Custom gender and age*

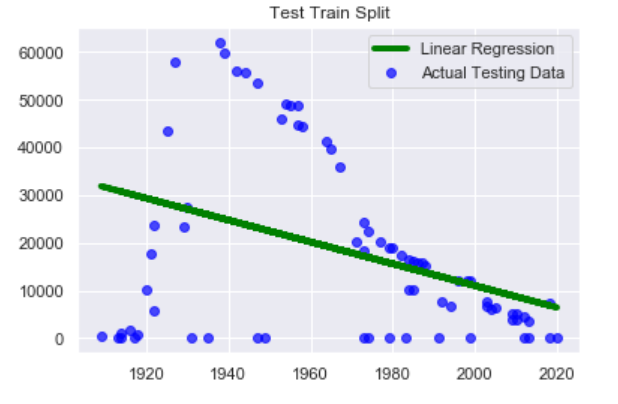
These data are then divided into the data on testing and training. In image 2, we can see by color what data is being tested and trained. The test dataset is green and the training data is red. Everything in the red, so training dataset, will be used to create a linear regression model, and the line will be tested on everything in the green, so the test data.



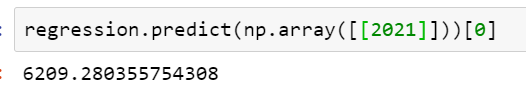
*Image 2: Test Train Split*

After dividing the data into training and testing data, we can start with a linear regression. We first fit the values of the training data to the linear regression. Image 3 displays the real test data with the forecast line.

According to the regression line, the number of customers will still be "senior" presumably because of the age, it is not likely that there is a large population of people born in 1940, but there is a small rise in the number of people born in 2000. This is likely because in the future, this age group will be "seniors" and will face more health problems. Therefore they will more likely buy more medications.



*Image 3: Linear regression*



*Image 4: Prediction for people born in 2021*

*So as you can see in the forecast we made above, the sum of revenue that is going to be for the people that are born in 2021, will be 6209 sales.*

## 

## 4.2.Seasons

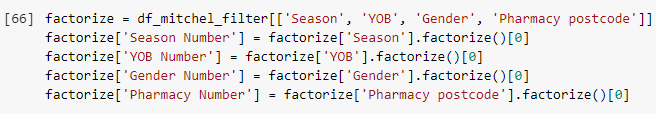
### 4.2.1.Region 22

For the seasonal data we have used the decision tree classifier method. The first thing that has been done is to drop all the null valued columns. This way we can factorize the data to be used in the decision tree.



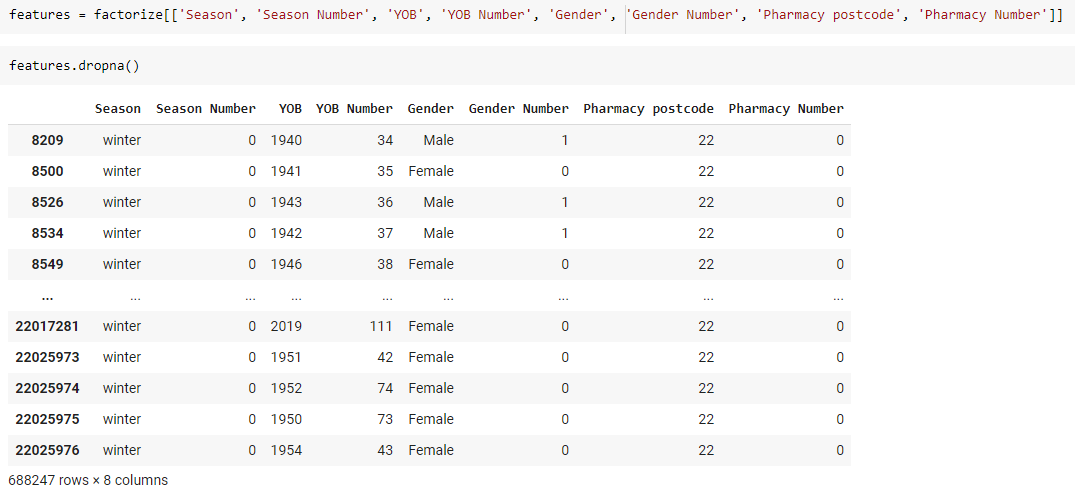
*Image 1: Drop Null Values*

The features used for this method are the Seasons, Year of birth, Genders and Pharmacy Postcodes. With this data we will look at how high the chance is that a specific season is chosen by the decision tree method. This data is also factorized and put into new columns as numbers as seen in the image below.



*Image 2: Factoring Data*

The data now looks as follows after factorizing the columns that have been chosen as target and variables.



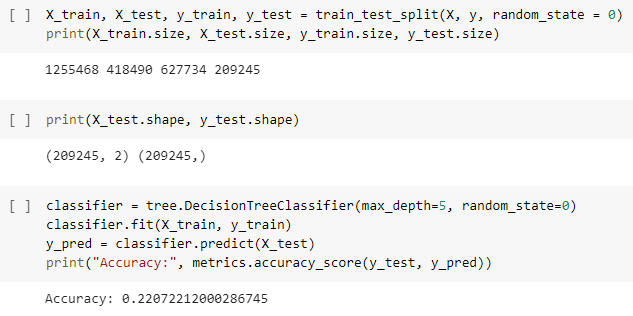
*Image 3: Factorized Data Table*

The next thing we do is to select the target and variables. The target is chosen with “iloc”. The features are shown in the X variable and the target is selected in the y variable.

### 

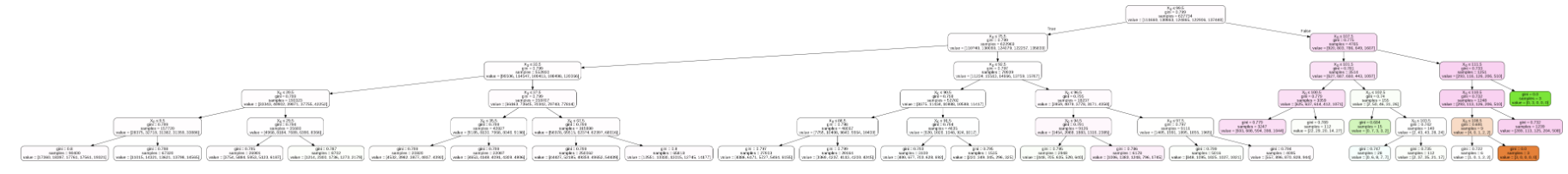
*Image 4: Splitting Features and Target*

After the selection we take a look at the size and the accuracy of the chosen data. This can be seen in the image below:



*Image 5: Test Size and Accuracy*

As shown in the image above, the accuracy is rather low and could indicate that the decision tree will show inaccurate data. The decision tree below shows that there are a lot of branches which indicate that there are alot of different paths the data can take.

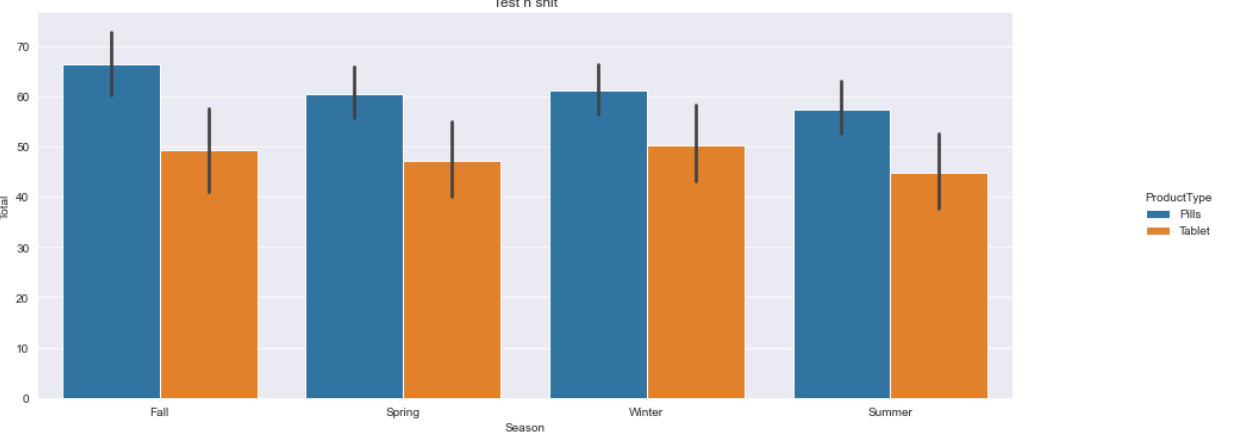
**

*Image 6: Decision Tree Result*

### 

### 4.2.2 Region 40

It seems that pills are sold more in every season especially in the fall. Tablets are the most common type of ill. They are inexpensive, safe and effective ways to deliver oral medication while capsules include medication that's enclosed in an outer shell. This outer shell is broken down in the digestive tract and the medication is absorbed into the bloodstream and then distributed and metabolized in much the same way as medication from a tablet.



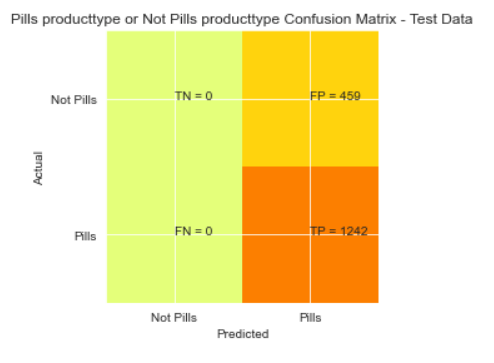
## 

## 4.3 Products

### 4.3.1 Region 40

As you can see in the confusion matrix, the amount of pills that were predicted as pils are correct. There were no false negatives, which means that all the products that were predicted as pills were not seen as pills.

One problem here is that all the products were predicted as pills, as you can see in the confusion matrix (image 1) we have quite a lot of false positives. This model is not ideal because all the products are being predicted as pills.

  
*Image 1: Evaluation Matrix Pills producttype*

### 

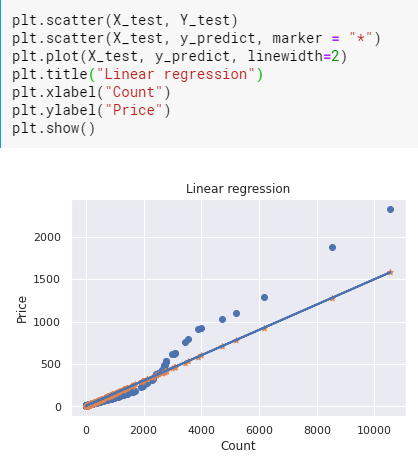
### 

### 

### 

### 

### 4.3.2 Region 30 Product Pricing



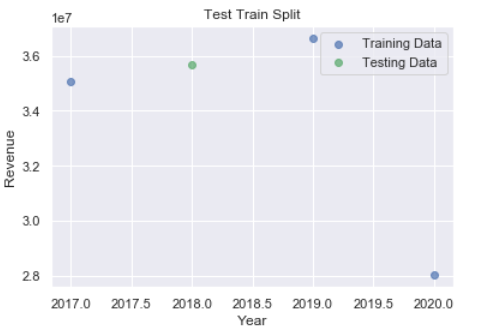
As we can see a large number of products are located at a low price point. While only a few of the products are located at a high price point. The explanation for that can be that people don’t tend to buy expensive medication or buy medication for common diseases such as flu which are relatively not very expensive. On the other hand, a medication that is used for some rare and specific diseases can be very expensive but the number of sales for such medication can be very low because not many people have them.

### 4.3.3 Region 20: Product Revenue per Year



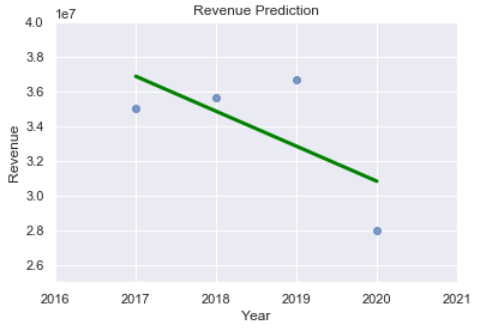
*Image 1: Total revenue per year*

These data are then divided into the data on testing and training. In image 2, we can see by color what data is being tested and trained. The test dataset is green and the training data is red. Everything in the red, so training dataset, will be used to create a linear regression model, and the line will be tested on everything in the green, so the test data.

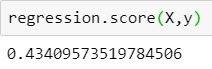


*Image 2: Test train split*

Since the pharmacy data for the year 2020 just updated, it seems not enough to show that the overall revenue is absolutely right. However, based on what we have, the income is declining (could be because of the Coronavirus) and people possibly reduced their time buying medication at the pharmacy. The store should have some measure in regarding this problem.



*Image 3: Linear regression*

**

*Image 4: Scoring*

43.4% is not a bad score. In conclusion, we could use this number to predict the future development for the pharmacy.

# 

# 5. Ethical Considerations

In this chapter the ethical considerations are discussed. During the stages of the project questionable situations could come to light, which have to be taken into account. The questions we ask ourselves are the following:

* Beneficial: Does our use of data benefit consumers as much as it benefits us?
* Progressive: Do we have a culture of continuous improvement and data minimization?
* Sustainable: Are the insights we identify with data sustainable over time?
* Respectful: Have we been transparent and inclusive?
* Fair: Have we thought through the potential impacts of our data use on all interested parties?

## 5.1 Beneficial

First ethical principle of big data states that all concerned parties such as individuals that generate data and organizations that collect data should receive the same amount of benefits.

The use of data could benefit the pharmacy in many different ways. The goal of our use of data is to give the pharmacy to be prepared for the illness season. This could help the pharmacy to be prepared for their stock and improve their customer service. On top of that, with the help of the findings of the dataset, the pharmacy can focus their advertisement on a certain group of people and therefore increase their amount of customers.

Not only does the use of data benefit the customers, but us as well. We increase our knowledge as a result to increase our advice to the customer. This knowledge will have a huge advantage for us, in order to help our customers

## 5.2 Progressive

With the help of big data we are going to improve and innovate services. In other words we as an organization learn from applying big data and it helps to deliver better and more valuable results. The amount of data that we are going to use promotes more sustainable and less risky analysis.

## 5.3.Sustainable

As data specialists, we are aware that our insight will not always provide sustainable value after a reasonable lifetime since certain events might change over time. With the world changing and especially technology that is supposed to make our daily lives easier, numbers and results could differ and therefore our insight will not provide the same sustainable value.

## 5.4.Respectful

The customer has a transparent view on the data that we use for the analysis. Moreover, the data analysts respect the interests of all different parties. Those parties include individuals and organization. In this context respectful relates to contractual or notice related restrictions on how the data might be applied. For example, the group does not use data to affect certain individuals if they do not expect it to be used.

## 5.5.Fair

We as the data analysts are fair to the individuals to whom the data pertains. The data that we are using and analyzing is NOT going to discriminate against individuals based on gender, race, physical characteristics, genetics or age and made sure we included everyone equally as well as making sure that the way we described our findings is not offensive comes across unequal to certain parties.

# 6. Recommendations

## 6.1 Age & Gender

In this section, we are going to discuss the recommendations regarding age and gender, that are based on exploratory and predictive analyses.

### 6.1.1 Exploratory Data Analysis

It’s most likely females that would buy more medications from the pharmacy. However, according to the exploratory data analysis that has been made for different regions, we can see that males who are born in the year between 1945 and 1965, would most likely buy more medications than females. What is also remarkable is that males who are born around the year 2000> also have a higher purchase of medications than females. Nonetheless, looking at numbers, it’s still females that have a higher purchase of medications than males and are therefore the target group for the pharmacy.

### 6.1.2 Predictive Analytics

According to our predictive analytics, the largest group of customers will still be the “seniors” in the future so we see an increase in the year of birth line. This is probably because of the age, it’s not likely that there is a big population of people who are born in 1930 in the future but rather a slight increase of purchase for the people who are born in the year 2000. There is a simple explanation for this because in the future, this age group will be the “seniors” and would experience more health issues and therefore would most likely buy more medications. And according to the predictive analytics, it would still be the females who purchase more medications than males.

### 6.1.3 Final Recommendation

The overall recommendation that can be given to the pharmacy is to focus their advertisement on mostly seniors since this is the most important customer group with the highest amount of purchase. The result is not surprising since seniors tend to experience more health issues than younger people. If we want to zoom in more, it’s most likely females who buy more medications than males and therefore advertisements can be more focused on females who belong in the senior age group. With this information, the pharmacy can focus more on the senior group and how to improve their customer service more for her customers.

## 6.2 Seasons

In this part of the recommendations chapter, the recommendations will be discussed that are based on the seasonal exploratory and predictive analyses made.

### 6.2.1.Exploratory Data Analysis

Based on the exploratory data analysis the season that has the most sales in most of the regions is the autumn season. This would be the best season to focus on when looking at a specific season to sell more medical products. The exploratory data analysis also showed that female customers bought the most in this season, which could help with focussing on a specific customer group.

### 6.2.2.Predictive Analytics

Based on the predictions made with the seasonal data. We have seen that pills are being sold the most, especially in the autumn season as predicted given the exploratory data analysis results. Pills are probably the best way to sell medication because of their safe and effective ways to deliver medication to a human body. Based on the predictions these products are going to be sold more in the future in all seasons, but especially more in the autumn season because of the high sales currently in this season, which gives this season a high chance of sales compared to the other seasons.

### 6.2.3.Final Recommendation

The overall recommendation that can be given is to focus more on selling pills in the autumn season because of the successful sales of this season and product at the moment. This could also help reduce costs of overstocked medications that are not bought as much as pills in this season.

## 

## 6.3 Products and Pricing

### 6.3.1.Exploratory Data Analysis

In the exploratory data analysis we compared different regions to find out what are the most sold medicines and does correlation between popular products and regions exist. When we compared the regions we found out that the most popular medicines are used to treat cardiovascular diseases such as Bisoprolol, Asaflow, Burinex and Coversyl.

### 6.2.2.Predictive Analytics

As we can see a large number of products are located at a low price point. While only a few of the products are located at a high price point. The explanation for that can be that people don’t tend to buy expensive medication or buy medication for common diseases such as cardiovascular diseases which are relatively not very expensive. On the other hand, a medication that is used for some rare and specific diseases can be very expensive but the number of sales for such medication can be very low because not many people have them.

Moreover, since the pharmacy data for the year 2020 just updated, it seems not enough to show that the overall revenue is absolutely right. However, based on what we have, the income is declining (could be because of the Coronavirus) and people possibly reduced their time buying medication at the pharmacy. The store should have some measure in regarding this problem.

### 6.2.3.Final Recommendation

It is hard to give a particular recommendation in this section because of the ethical consideration of pricing for medication. We can easily remember the scandal that happened to pharmaceutical company Valeant and its pricing strategy. The company raised prices on all its brand name drugs by 66% in 2015, five times more than its closest industry peer. The cost of Valeant flucytosine was 10,000% higher in the United States than in Europe. The decision was very controversial because the company produced essential life drugs.

On the other hand it would be a good option to increase the amount of selling cardiovascular types of medicines as they are the most popular medication types among all regions.