Data Mining report

Hugo THIOLLIÈRE

Problem Understanding

At the time I was looking for a dataset and a problem to study for this data mining project, I happened to be listening to the news on the radio, and the news anchor was talking about the upcoming increase of the price of the cigarette pack in France. As an ex-smoker myself, I have an insider's view regarding tabagism and nicotine addiction, and I was wondering if there were any observable trends among the global population.

I believe smoking in France is truly a concerning issue with an ever increasing alarming rate of deaths and lung cancers, especially among women. It is funny to go and see how French people are depicted from the outside : most of foreigners will tell you that French people love their baguettes and their cigarettes.

At this time I remembered a podcast I listened to a few years ago describing how the US -despite suffering from other serious health-related issues (among which obesity)- managed to efficiently deglamorize cigarettes and have a real impact on the general population's behaviour.

For this project I decided to look for and work on raw data as I find it more challenging and closer to real life issues than an already well-polished dataset. I found three potential candidate datasets that I will present in further details in the Data Understanding part. As it turned out, the datasets I have chosen didn't fit well for a predictive data mining approach, and better fit a descriptive data mining approach.

Data Understanding

Here are the gitHub links to the three raw dataSets I found :

- Raw census data from 1995 to 2010
- Financial data from 1970 to 2006
- Health related data 2005-2009

To better understand the data, let's use the build-in R function `str()` that gives us insight regarding the structure of the data :

```
str(Raw_census_data_from_1995_to_2010)
```

```
$ Year      : int  2003 2009 2003 1998 1998 1995 2007 2002 2000 2006 ...
$ State     : Factor w/ 56 levels "Alabama","Alaska",..: 12 48 32 53 31 5 3 20 22 46 ...
$ Condition : Factor w/ 1 level "Adults who are current smokers": 1 1 1 1 1 1 1 1 1 1 ...
$ Yes       : Factor w/ 188 levels "10%","10.5%",..: 178 56 130 93 99 32 76 117 84 82 ...
$ No        : Factor w/ 194 levels "66%","67.4%",..: 1 130 50 92 86 153 110 64 102 104 ...
$ Location.1: Factor w/ 56 levels "Alabama\n(32.840569999605975, -86.63186000013877)",..: 12
```

A few remarks at first glance :

- The 'Location' column looks a bit messy. Indeed, the latitude and longitude values are prefixed by a string containing the name of the state. To be exploitable, we can use some regular expressions in order to only keep the numerical values.
- The 'No' (resp. 'Yes') column is redundant. Indeed, we can find it by using the complement of the 'Yes' (resp. 'No') column.
- The 'State' column is a factor with **56** levels ! Last time I checked, there were only 50 states in the United States. I was curious as to what the cause of this discrepancy was, so I used the `unique()` function. Upon further investigation, this column holds the values for **3** American territories (Guam, Puerto Rico and the Virgin Islands), the District of Columbia, and also **2** nationwide metrics in addition to the **50** regular states. This all adds up to 56.

I applied the same strategy on the health-centric data and appart from the fact that the gender column had **3** levels (Which turned out to be Male, Female and **Overall**) and the fact that a lot of colums held useless or inexpoitable data, nothing too much out of the ordinary came out of this analysis.

The last data set, holding financial information, is a big table of **14382** observations of **17** variables. It went through the same process and it seems that out of the 17 columns, an awful lot of them are unexploitable, either because all the lines have the same value for this attribute, or they simply don't carry significant information. These columns will therefore be removed.

The interesting information yielded by this analysis is that the attribute 'SubMeasureDesc' holds **6** distinct values :

- Average Cost per pack
- Cigarette Consumption (Pack Sales Per Capita)
- Federal and State tax as a Percentage of Retail Price
- Federal and State Tax per pack
- State Tax per pack
- Gross Cigarette Tax Revenue

These **6** attributes are observed for **51** entries (50 states plus the District of Columbia) over a period of **47** years (1970 - 2016).

We can even check that $6 \times 51 \times 47 = 14382$.
This underlying structure makes the data exploitable : we have one spatial dimension : *the state*, we have one temporal dimension : *the year*, and for each of these couples, we have **6** features.

Data Preparation

After we took time to understand the data, we shall now prepare it in order to make it exploitable.

Getting rid of inexploitable data

In order to do that, I will first delete all the columns that were deemed useless by our data understanding procedure, e.g. for the financial dataset,we can do the following :

```
columsToRemove = c("Datasource", "TopicDesc", "Source", "MeasureDesc", "TopicTypeId",
"TopicId", "MeasureId", "SubMeasureID", "SubMeasureIdDisplayOrder")

Financial_data = Financial_data[ , !(names(Financial_data) %in% columsToRemove)]
```

We create a vector containing all the columns we want to remove and then we only keep the ones whose names do not belong to the 'columnsToRemove' vector. This operation drastically reduces the width of our dataset (from **17** columns down to **8**). I applied the same procedure on the other two datasets.

Reshaping data

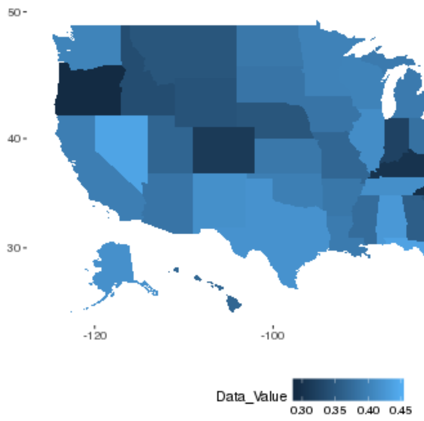
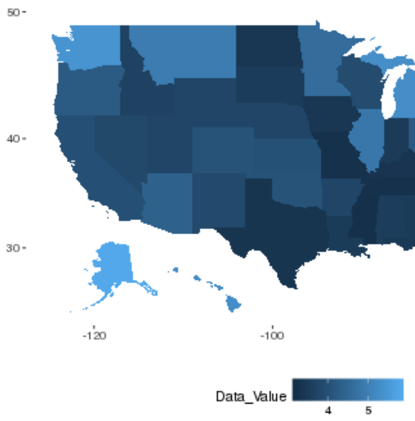
As I mentioned before, the 'Location' column of the Raw census data was a bit messy and so, I used some Regular Expression magic to reshape the data accordingly to the use I intended to make of it :

```
location = Raw_census_data_from_1995_to_2010$Location.1
matches = regexpr("\\(-?[0-9]+.[0-9]+, -?[0-9]+.[0-9]+\\)",location, perl = TRUE)
regmatches(location,matches)
```

This is a simple regex match in which we want to keep couples of floating point numbers in between parenthesis, separated by a comma. We only have to keep in mind that we must escape the parenthesis characters (as they serve a very specific purpose in the regex world) and to add an optional negative sign right before our numerical values. Indeed, we are working west of the Greenwhich meridian, so we expect all of the longitude values to be negative.

Displaying information

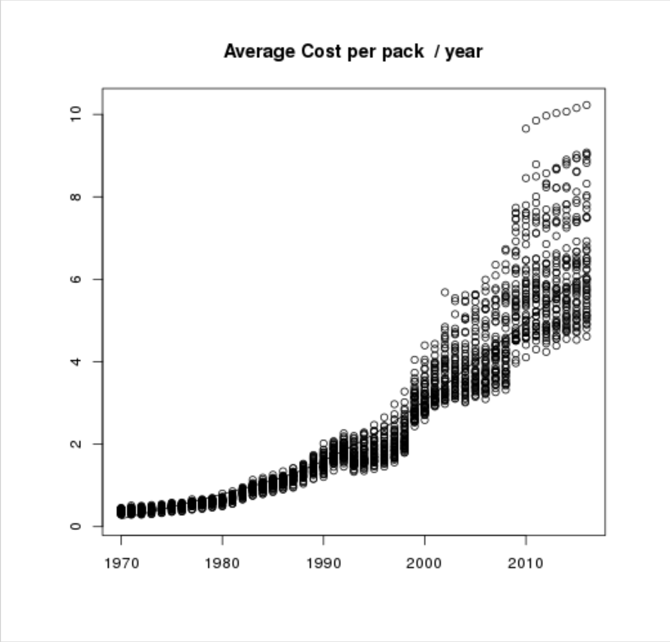
As it turns out, I didn't need to use regex at all! Still, I decided to showcase it nevertheless, to show that there almost never exists only one way of overcoming an obstacle. As a matter of fact, I decided to try and display my data in an easier-to-understand way.

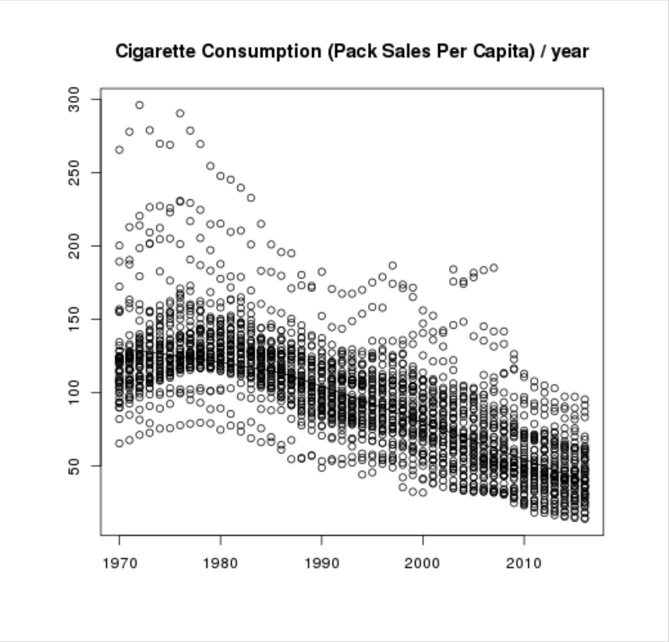
Average Cost per pack in 1970 :	Average Cost Per Pack in 2006 :
 50 40 30 -120 -100 -80 Data_Value 0.30 0.35 0.40 0.45	 50 40 30 -120 -100 -80 Data_Value 4 5

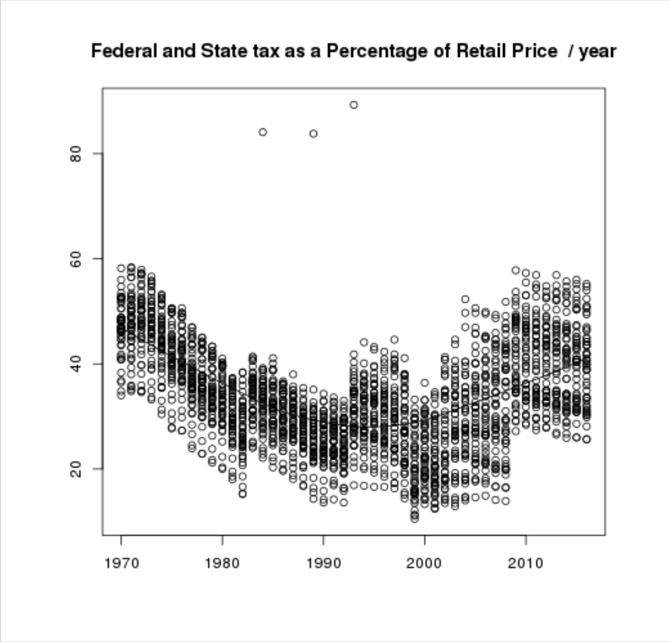
I generated the maps for the 6 features, for each of the 50 states and for each year from 1970 to 2006. When I skim through the generated pictures chronologically, I can visually see some trends that I would have been completely unable to detect if I only have the raw data at my disposal.

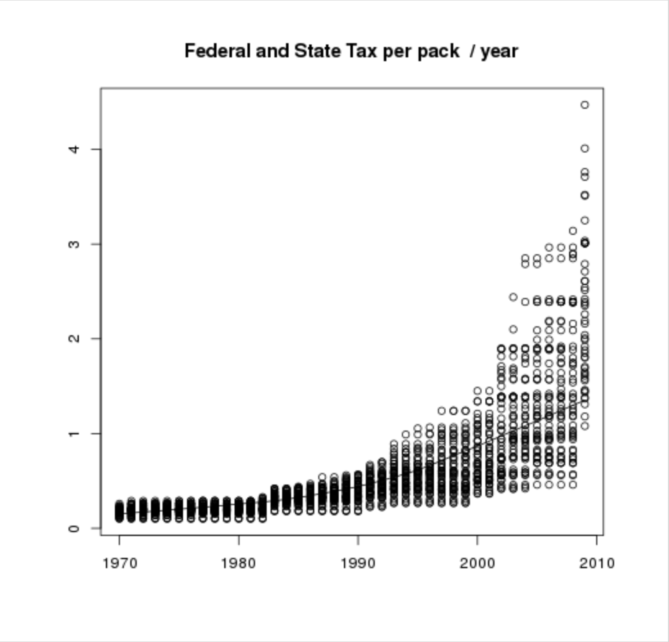
What arises from this pre-process is that it seems the northern parts (and the western parts, to a smaller extent) of the territory have a more radical approach than the south-east. Cigarette packs cost more and we see a clear correlation with the Gross Cigarette revenue in these states.

Modeling


Average Cost per pack / year


Cigarette Consumption (Pack Sales Per Capita) / year


Federal and State tax as a Percentage of Retail Price / year


Federal and State Tax per pack / year

Using a custom function to parse the data according to the different axes that were discovered in the data preparation part, we can display the data in a readable and exploitable way.

We can observe that the taxes per pack and the price of the pack are greatly correlated, of course, and that it seems that they are negatively correlated with the cigarette consumption over the years. Of course, we have to keep in mind that correlation doesn't imply causality.

Evalutation And Deployment

Here, we took a rather descriptive data mining approach in order to find some structure in raw messy data. We could go through the same steps on a new data set, e.g. coming from another country and find some interesting underlying structure as well.