WEB SCRAPPING PROJECT

ANALYSIS OF CARDMARKET OFFERS

CHRISTIAN GABRIEL CENTENO

LINKEDIN

GITHUB

INDEX

1. INTROUCCION

This is a personal project to realize a web scrap program. The aim is to extract information from an e-commerce website, transform and load it into csv files to make some analysis reports.

In this case, the choice of e-commerce site is *CardMarket*. *CardMarket* it's a European site founded in Germany where private and professional sellers negotiate with *TCG* (Trading Cards Games) products. In my case, I'm focusing on *Pokemon TCG* items.

The project is divided into two main areas:

- ***Data Engineering***: Extracting the data hosted in the e-commerce, transforming it into useful information and generating csv for loading.
- ***Data Analysis***: Use PowerBI to create analysis reports based on the csv.

And I always use the same card in the examples to be clearer, I chose the "Japanese's Charizard for the 151's collection".

This card is one of the most famous modern cards, and it's a nice product to analyze it (unfortunately, when I started this project, the "big boom" of *Pokemon TCG* just happened, and I can't register this big change. However, there is enough data to show some interesting trends).

So, let's go!

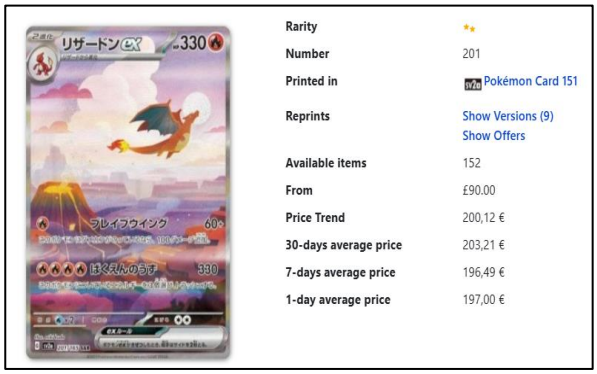
2. DATA ENGINEERING

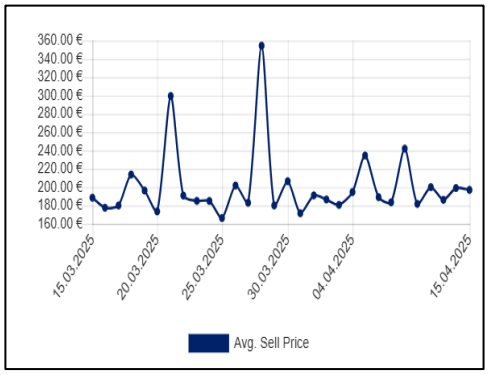
2.1 Knowing how the website works

First, I will show the basics of the website and what information we are interested in.

When you search a product; all the offers are from the same collection so, there are different between the Orient's collection and the Occident's collection. That's why only Japanese, Chinese and Korean languages are seen (not English, this being the most chase card languages).

At the top, there is the general information of the card. Like name, collection's name, a card's photo and **the last 30 days with average price of the confirm sells of the product** (If one day any cards it's sell, this day not show).





Below are all the available offers by seller. Where you can see:

- The seller's nick
- The seller's country
- The seller's status
- The seller's comments (optional)
- The language's card
- How many cards are in their stock
- The price in euros

Seller	Product Information		Offer	
3K [redacted]	GD		108,55 €	1
0 [redacted]	NM	Pack fresh, put in sleeve and toploa...	145,00 €	1
403 [redacted]	NM		164,90 €	1
579 [redacted]	NM		168,00 €	1
135 [redacted]	NM	close to MT more sv2a ask for photos	169,00 €	1
325 [redacted]	GD		170,00 €	1
81 [redacted]	EX	Front NM / Back EXC small dents	170,00 €	1
18 [redacted]	NM	Photo sur demande	170,00 €	1
2K [redacted]	EX		175,00 €	1

This is all the data that we'll need for our analysis.

2.2 What tools we will use

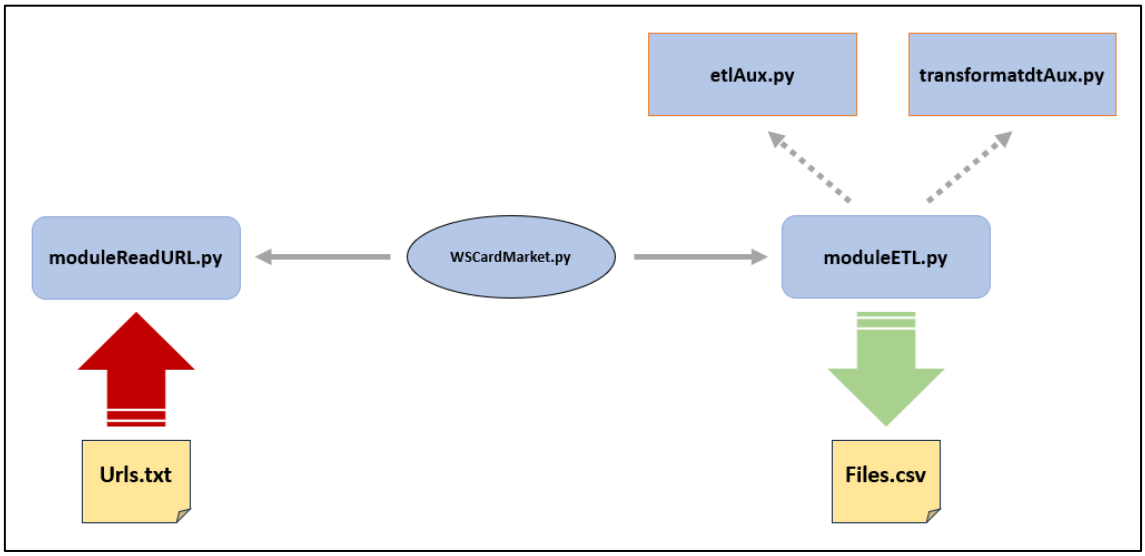
I created a **Python's** script using specific libraries for WebScrapping:

- *Pandas*: For creating and manipulating DataFrames.
- *Selenium*: For making requests and automatic web browsing actions (press the "show more" button)
- *BeatifulSoup*: For managing Internet content (not really necessary because I used Selenium, but I wanted to try it)
- *ThreadPoolExecutor*: For making concurrency web scraping.

2.3 The architecture of the Python's project

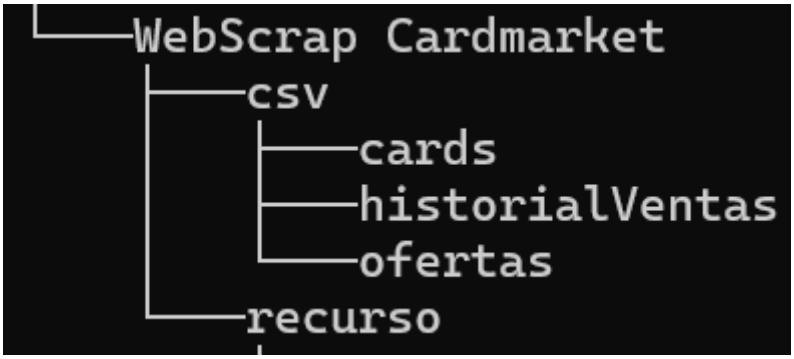
If you want to see deeper the code, it's available in my **Github** profile

2.3.1 Classes



- *WSCardMarket.py*: The main script, it's the *orchestrator*.
- *moduleReadUrl.py*: It's the module dedicated to read the urls::CardName file and apply format transformations to the other modules.
- *moduleETL.py*: It's the module dedicated to make the *WebScrapping* work and load the information in csv files.
- *etlAux.py*: It's an auxiliary module dedicated to do all the *etl's* process.
- *transformatdtAux.py*: It's an auxiliary module dedicated to manipulate and manage the *dataFrames*.

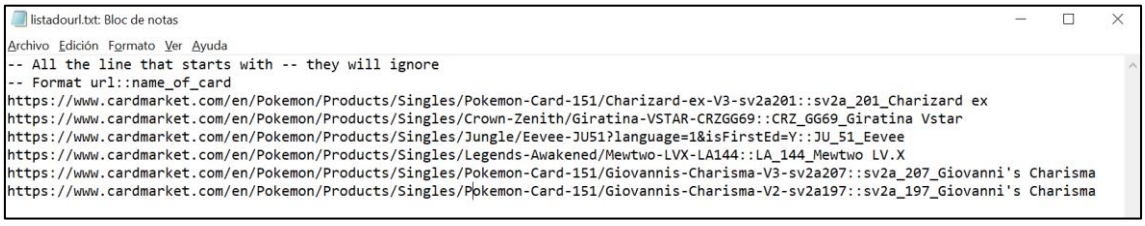
2.3.2 Folders



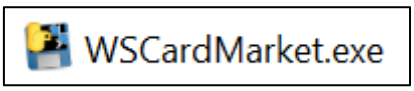
These folders are necessary to run the web Scrap Python:

- ***A main folder (WebScrap Cardmarket)***: where they will host the *executable* and the *PowerBI* project.
- ***Csv folder***: Where the csv is stored. In this folder there are two:
  - ***cards***: Where is host the summary data cards (photo, name, expansion and rarity)
  - ***historialVentas***: Where host the confirm sales price history.
  - ***ofertas***: Where host the offers of each card.
- ***recurso***: Is the resource folder, it's the place where is host the URL list txt (this file configure what cards are you want to study).

The URL file it's like:



Finally, on to portable targets. I've created a runnable script. You just have to download the architecture folder and run it (available in my GitHub repo with the whole architecture).



2.3.3 How to run

You just have to be sure to add all the URLs you want in the **.txt** and save. After that you just have to run the **.exe** and wait. If some csv already exists, they will be updated. If not, they will be created in each folder.

And that is all! You should have three csv for each card.

3. DATA ANALYST

3.1 Creating a Data Model

Once generated the csv it's moment to **PowerBI**.

First, it's necessary to access each directory and generate consults (tables), cleaning and apply ***powerQuery*** treatment and make the relationship between them.

Finaly, make complex calculations that I'll use in the graphs.

3.1.1 Connect PowerQuery to us folders and create additional tables

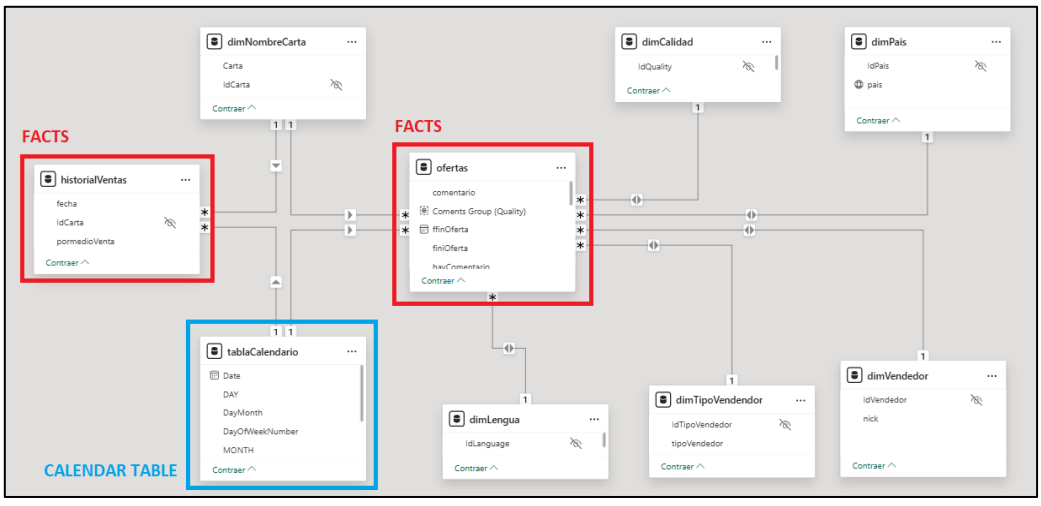
Connecting to *PowerBI* is easy, only it's necessary to use folders connectors. After that, *PowerQuery* generate a big consult of each folder concatenating all the files in.

These two big consults it's called "**Facts Tables**" and, from these, they were generating the "**Dimension Tables**".

I also created a "**Calendar Table**" to connect the two big tables.

3.1.2 Table Relationships

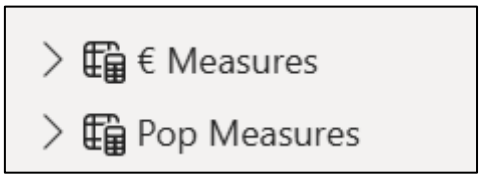
I created a snowflake relationship (**the Dimension Tables are only relation with the Facts Tables and the Facts Tables are not relation with each other**).



Now the filters will work.

3.1.3 Make Measures

I classified into two types of measures ***population*** (**POP**) and ***price*** (**€**) measures.



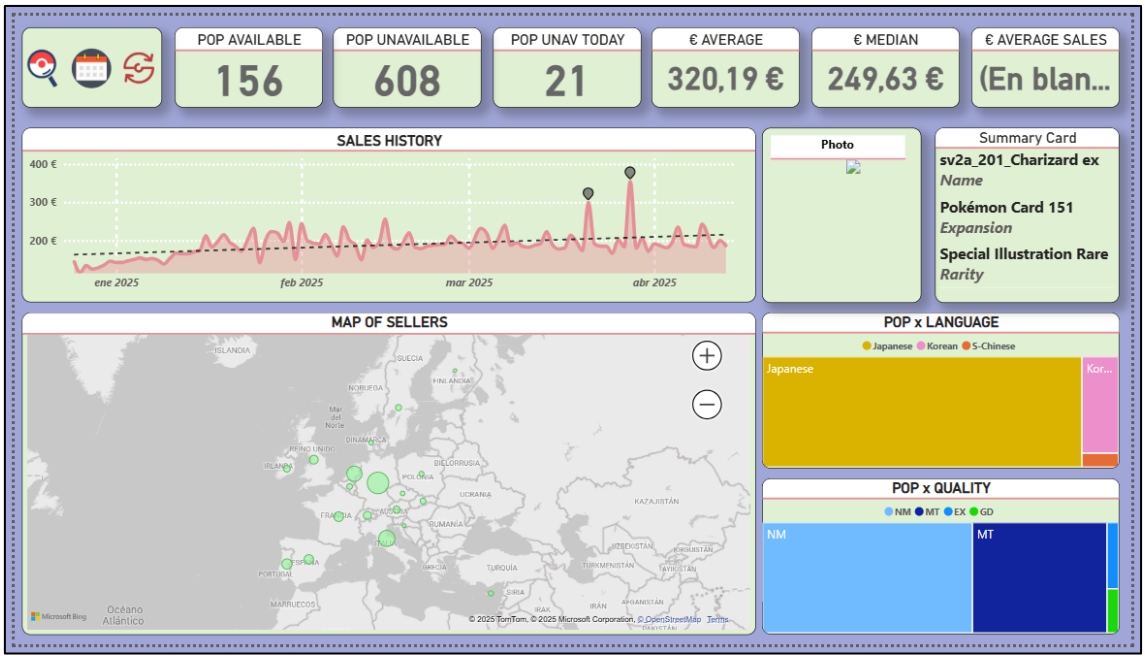
The POP measures are dedicated to measure the progress of the available population and how many offers aren't available day by day. This is controlled because in the python there are two types of dates: *the start offers date and the last day when the offer exists*. The more consistent you have been in running the exe day to day, the more realistic your measurements will be.

The € measures just measure the progress of the average price day by day (this is very interesting to compare with the official average selling price that *CardMarket* shows) also the minimum price day by day.

3.2 Sections of the Report

In this point I'll explain what's the topic of each section.

3.2.1 Summary Section



The first section is the summary, where you can find **KPI** like how many populations of the card already exists, how many populations not exit yet, how many cards was unable this day, the average/median price of the last day and the average confirm sale price of the last year.

On the second floor, there is a graph of the sales history with a **trend line** and **markers of anomalous** registers (more/less than expected by *PowerBI*). Also, there is a photo of each card with its summary

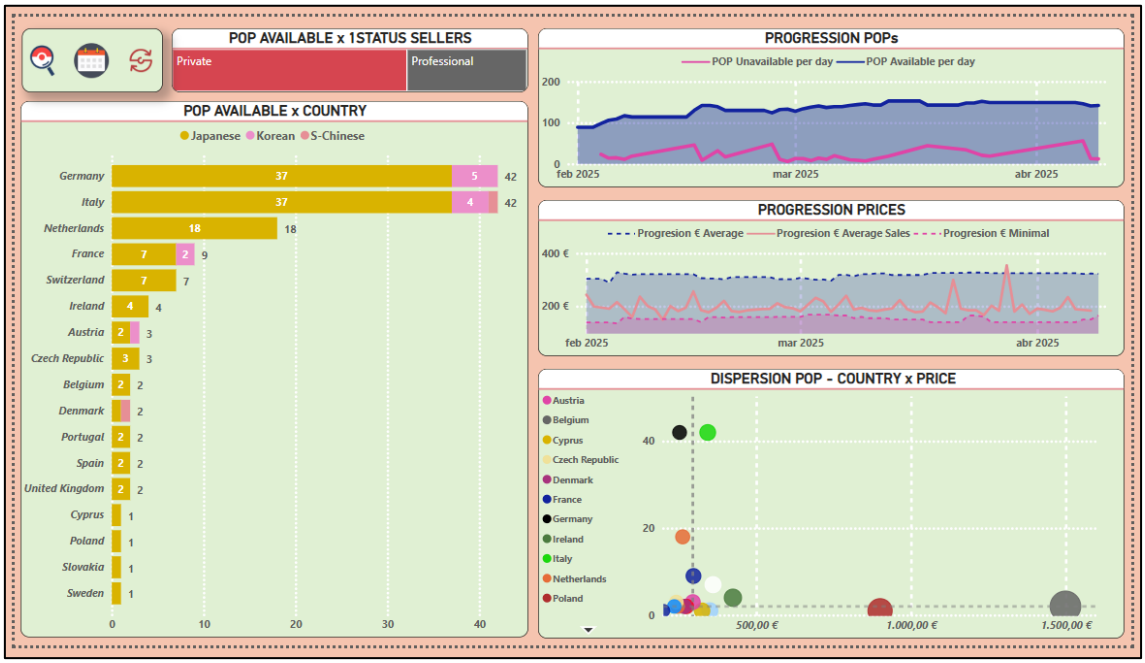
***Disclaimer**: The photo only shows where the Report is published in *Power Services*. In local mode you can see the photo only if you clicked on it and you are connected to Internet, then an extra window will show the card's photo.

In the last floor, there is a **map** of the seller's country and two **map trees**. One by languages of the cards and the other one by quality.

In all sections there are also three buttons:

- Cards features filter.
- Date filter.
- Restart filters.

3.2.2 Country Analyzer

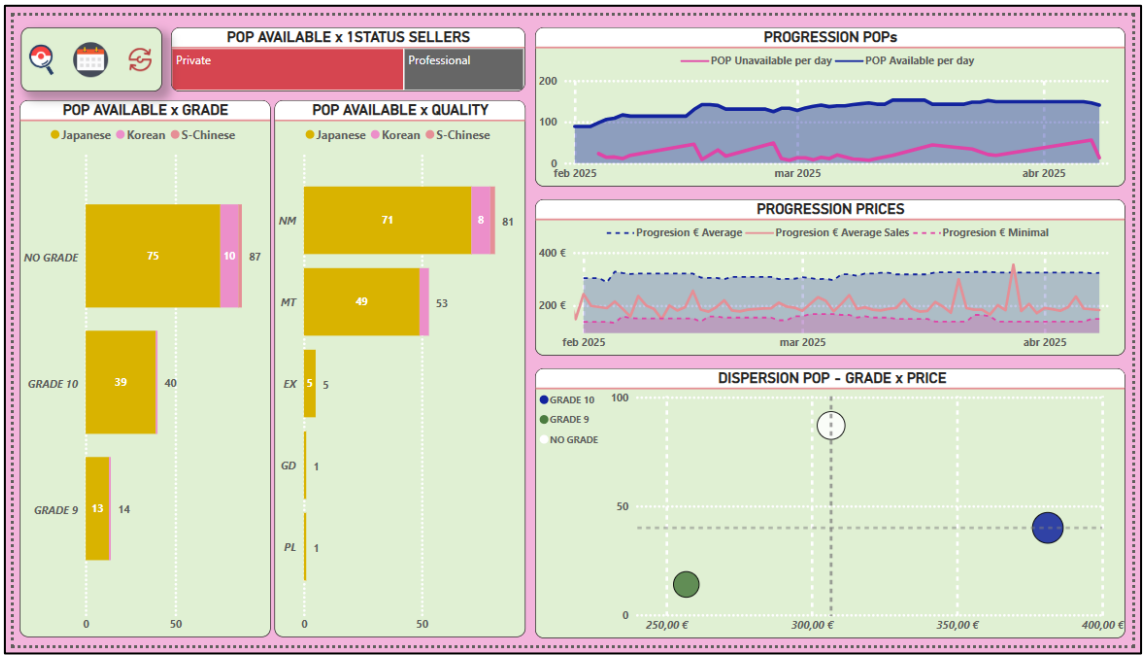


This section tries to analyze deeper the correlation between the price and the country of the seller.

For this mission there is a ranking of the countries spit in by language's card, two temporal graph to sow the population progression and how close are the country with them price in comparison with the confirmed average sales.

In last position, there is a dispersion graph that try to explain the correlation between the population, the price and the country.

3.2.3 Grade Analyzer



The final section tries to analyze deeper the correlation between the price and the quality of the cards.

There are two types of quality scale:

The quality condition: It's a value determined by the seller and exists different symbols:

- **MT – Meant**: It's the best condition
- ***NM – Near Mint***: It's almost the best condition
- **EX – Excellent**: The card has only few visible marks.
- ***GD – Good***: The card has some marks
- **LP – Light Played**: The card has moderate marks.
- ***PL – Played***: The card has several marks.
- **PO – Poor**: It's the worst condition.

The grade condition: Exists companies that specialize in grading cards. This consists in evaluating the card condition and determining a score from 0 -10. It's the way to get the maximum possible value from a card (generally only if it's a 10 and it also depends on the type of card). The problem with the grade it's, that there's no field in *CardMarket* to determine it. It can only be verified by the seller's photos or by the seller's comments. For this project I analyzed the comments and I made groups in function of them. The more the seller explains, the more consistent the groups are.

Taking all these considerations, there are two rankings of quality spit in by language card, two temporal graph to sow the population progression.

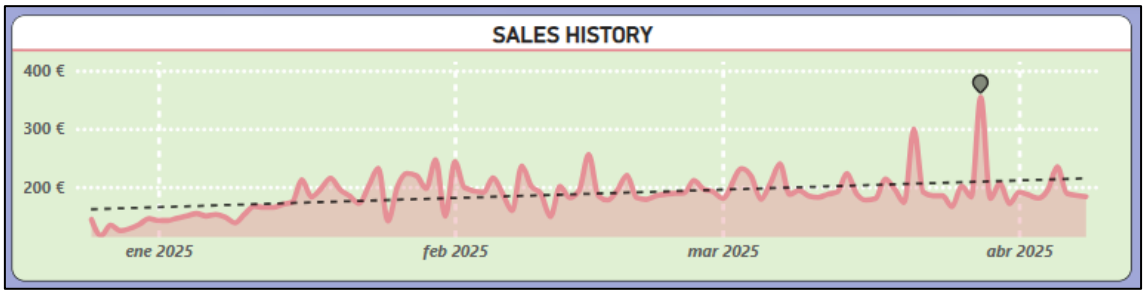
In last position, there is a dispersion graph that try to explain the correlation between the population, the price and the **grade condition**.

4. EXAMPLES

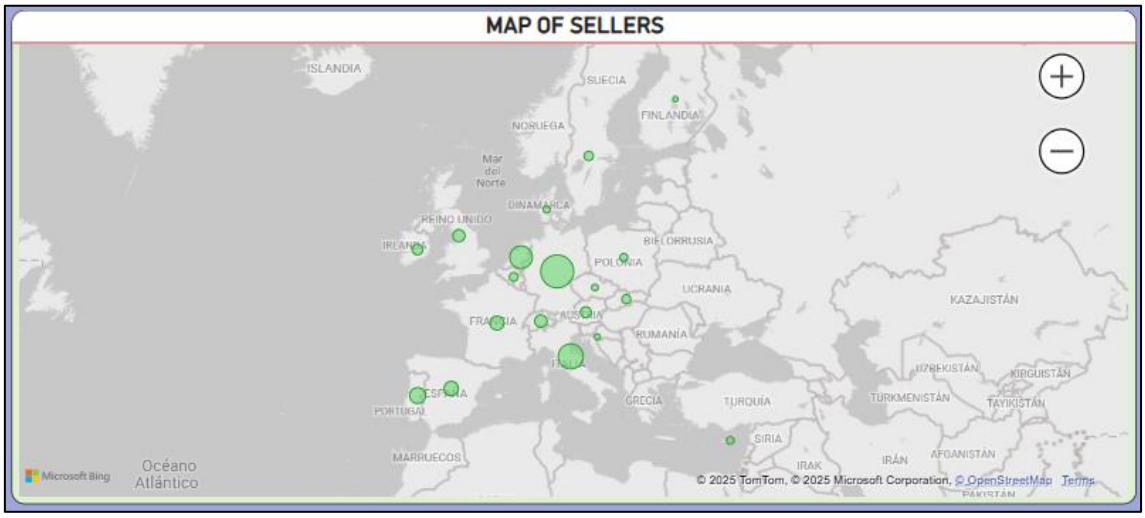
Now we will see some examples using the *PowerBI* Report

4.1 The Charizard's Summary

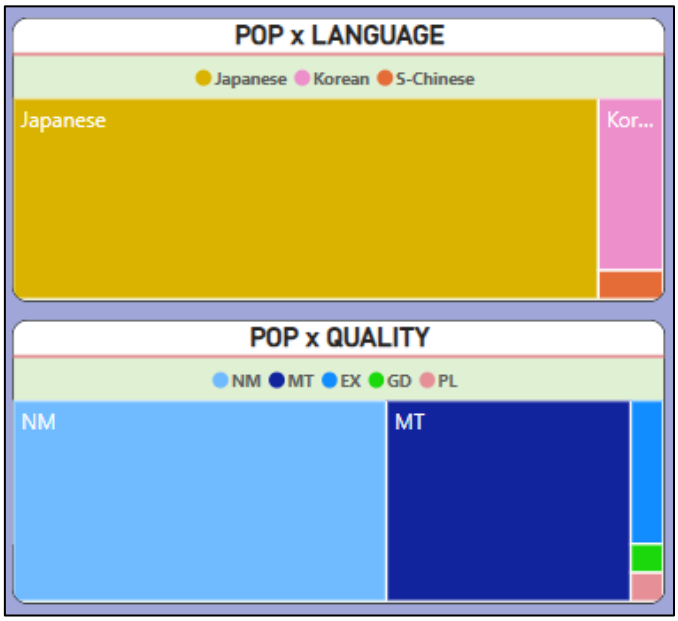
Apart from the KPI and the summary card information, with the sales history graph we can confirm that the trend of the card is in continuous growing. The last price is superior that all the first register month. We see only an anomaly in April, it's possible that only really expensive copies (like 9/10 grades) were sold that day.



With the map we can also confirm that it's a popular card, because there are many different countries with stock. Highlighting **Germany**, **the Netherland** and **Italy**.

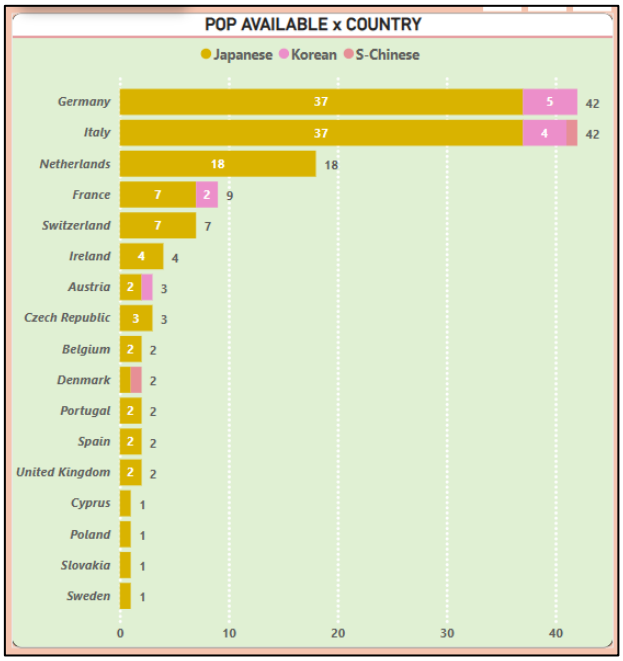


With the tree map we consider **Japanese** as the most popular Asian language. In quality, the most common are **NM** and **MT** (It's normal because it's a modern card, this tree map change totally with "*vintage cards*").

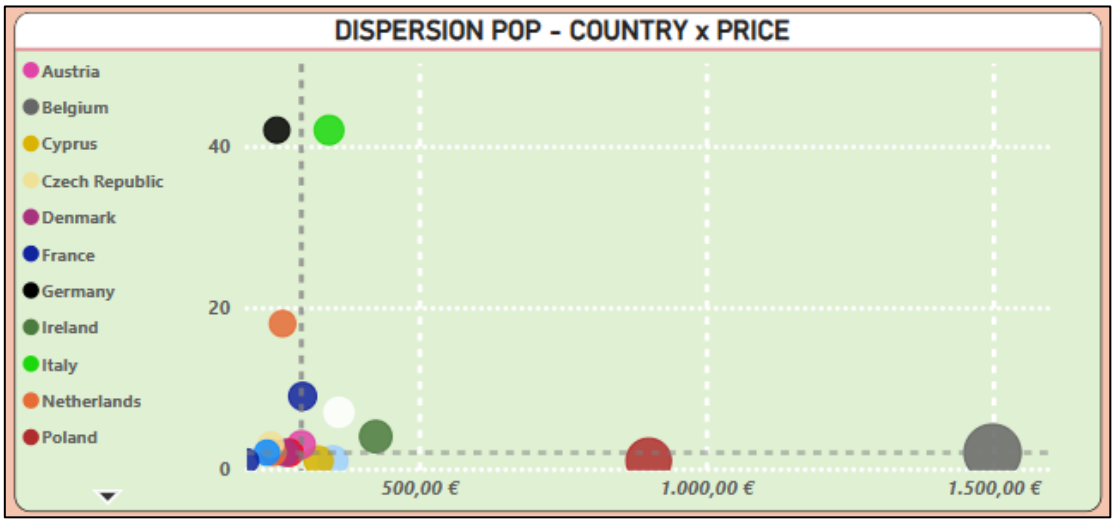


4.2 How the price changes among the countries

First, with the ranking we can determine the Top 3 and what are their proportions by languages. **Germany** and **Italy** are the winners with two and three languages and in the third position is **the Netherlands** with only Japanese cards.

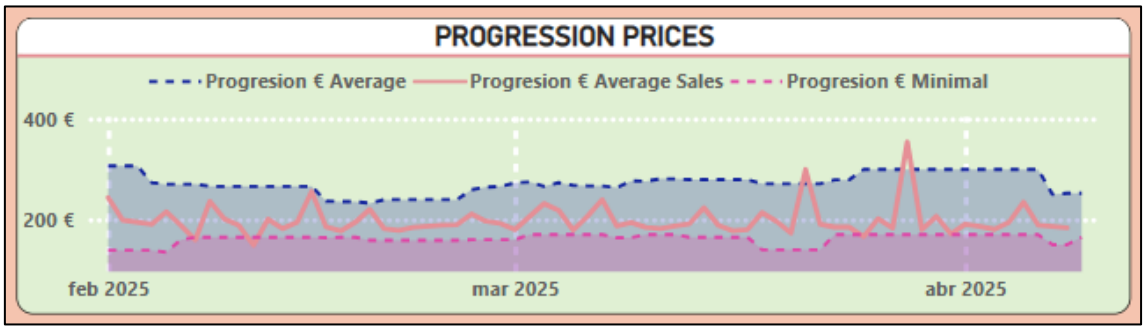


Now, we see the correlations with the price and it´s interesting. The top 3 are the most relevant in terms of population value, but there are differences. Germany and Netherlands are close in price but Italy is far above the median.

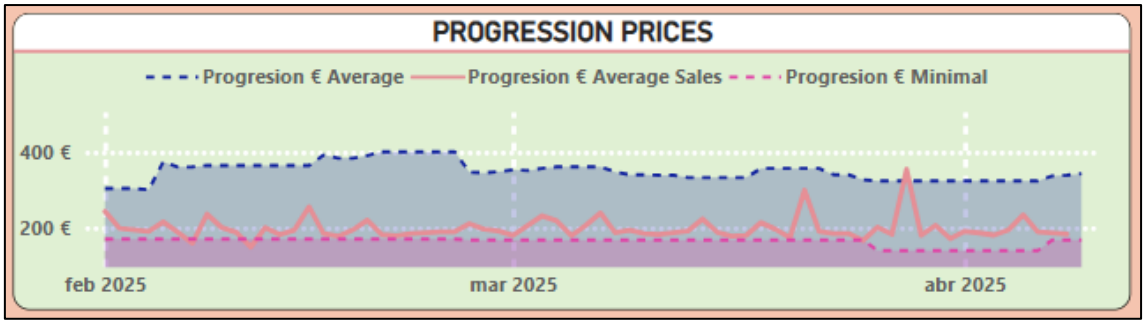


To be sure of the correlations, lets analyze the evolution of the prices of Germany and Italy.

In the case of Germany, we check that the average price fits to the average selling price. So, this could mean that the **German people have a "fair'" price**.

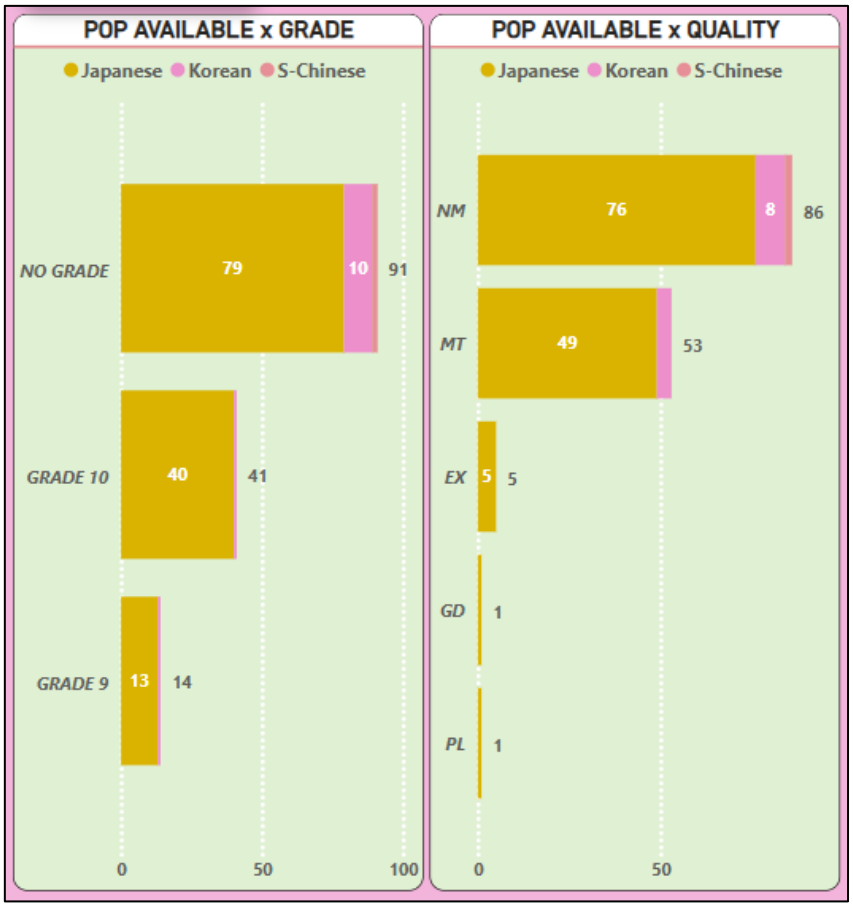


On the other hand, in Italy it's clear that the average price is not as fit as in Germany. Usually, it's more expensive that the average selling price. This can confirm the last correlation graph.



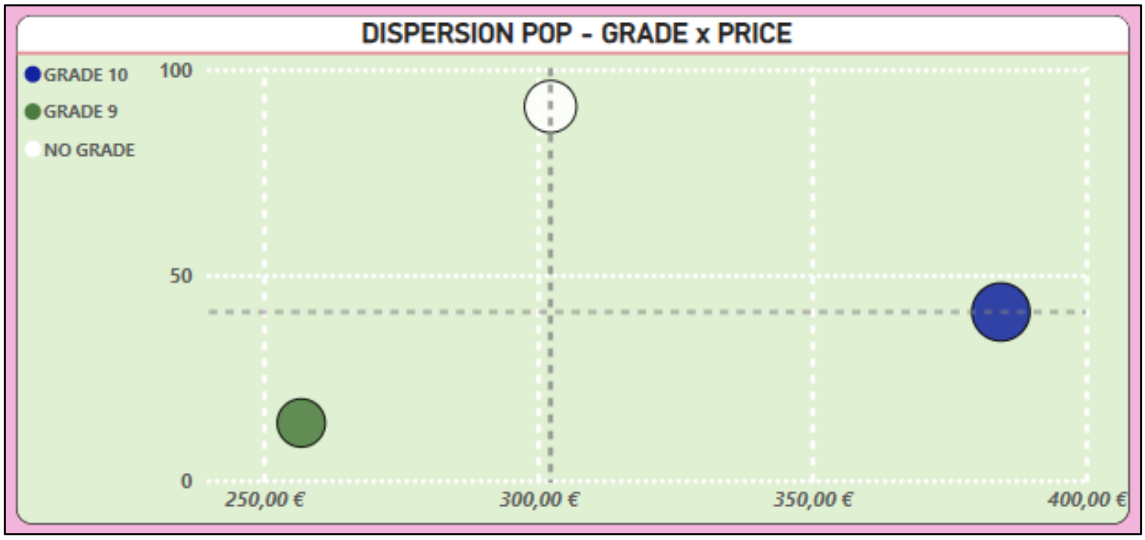
4.3 How the quality changes the price

When we do the ranking, we can confirm that the clear winner is No Grade and NM. The no grade is usually the most common offer due to grade is expensive and hard to sell. This also means that it's not necessary to have a grade card to classify a card with a NM (MT is stranger that this happened because the seller could be reclaimed if the customer see imperfections).

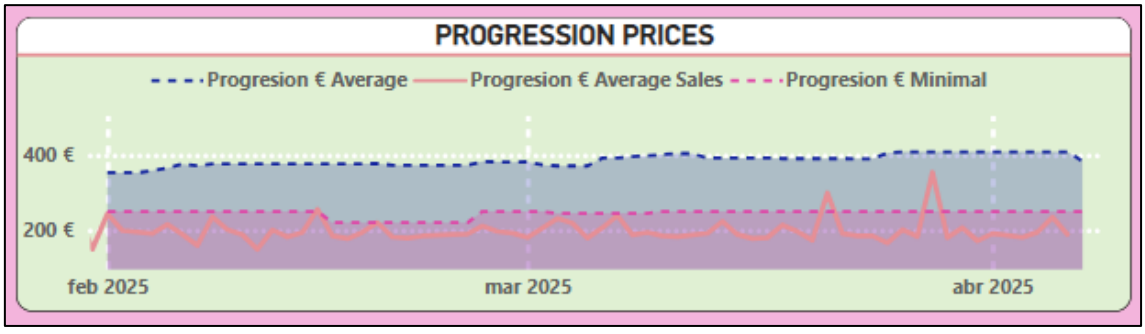


In the correlation graph, exists a great price influent in the cases of 10 Grades, and we can see an unexpectable relationship: The no grade group is the most numerous yes (and it's the reason why this group remark the trend), but is more expensive that the group of Grade 9. There are exist three possible reasons for this:

- **The comments are not so accurate**: It's possible than many offers have a grading card but the sellers didn't comment it.
- **Some of these cards have a "special format"**: Like bundles of more cards. It isn't too common but there are some cases.
- **Finaly, it could be a real trend**: In the vintage cards can be a little different, but generally exists a trend where the cards with no grade can be more expensive that other grades' cards where it isn't a 10 grade (in special in case of 8 grade or less).



If we focused in 10 grade group, we can check that as the average price, as the minimum prices are usually more expensive than the average sales prices. Only in the peak prices the minimum was exceeded. So, **people are not in the habit of buying the perfect grade card**.



5. FINAL CONCLUSIONS

We can check how powerful are the data when you can model and treatment to make the most information value possible. Exploring a website (and despite the limited scope) we can extract analysis reports about customers, sellers and price trends than can help us to understand the market and make more profited and confident decisions.

I hope this project interested the reader and we see us in future projects.