PERSONAL PROJECT

SEGMENTING AND FINDING PATTERNS IN CUSTOMERS

Developer: Marcos Matheus de Paiva Silva

Barra Mansa - RJ

April 2022

Contents

1 Objective

With the growing demand of the current market, the interests and demands of the consumer have become increasingly relevant when choosing what to buy. Knowing which customer you are dealing with today has to be more and more relevant. Therefore in the current project, we fully contemplate the stages of a data science project, so that from the information collected on the Kaggle website and from the data visualization we can probe some questions that would help users, companies and retailers understand about:

- How graphically do sales change over time? What is the maximum sales peak? When did this peak sales peak occur?
- Which month does the highest sales volume occur?
- What are the other standards for higher sales volume, such as which product line sells the most?
- What are the companies or customers that sell the most?
- Is it possible to segment customers? If yes, what are the patterns of these customer groups?

Finally, after creating the model, we provide an online application that allows any user to predict what their group is (). This procedure can be performed based on information collected or informed by the user, such as which product line the user buys the most, which country the user resides in, among others.

2 Proposed Solution

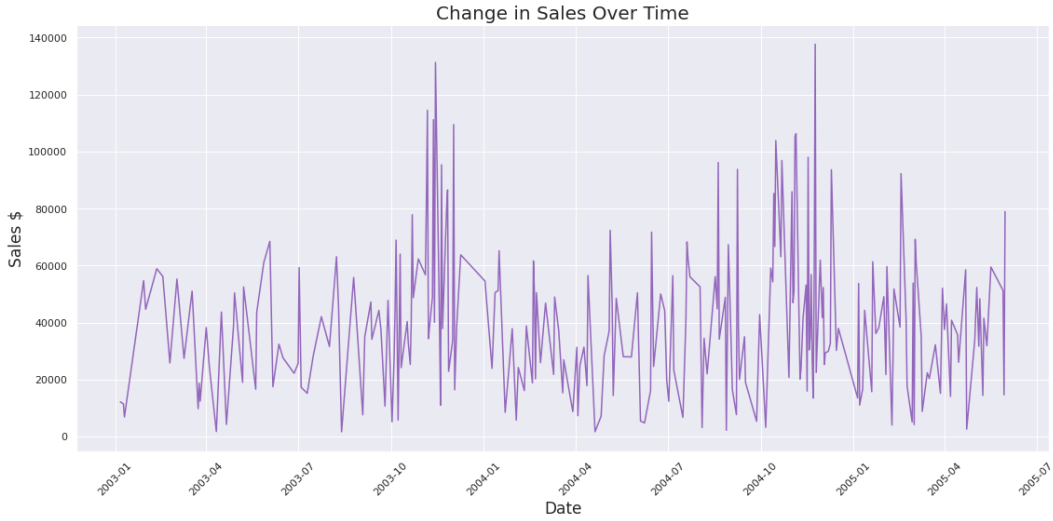
This project was developed following the steps in (GéRON, 2019), and also based on some phases in (MüLLER; GUID, 2016). To find the answers to the questions raised in the previous section, we first used the Google Colaboratory so that any individual can run the code in python and also better understand all the theoretical argument, insights and methodology used in the project.

The first step of the project was to collect data from the Kaggle website. After this procedure, a jupyter notebook was used in which the libraries to operate with pre-processing, data analysis, graphics, files, and machine learning were pandas, scikit-Learn, matplotlib, seaborn numpy, seaborn, imblearn.

Furthermore, after creating the machine learning model in Google Collaboratory and the library pickle was used to serialize the model and write it in a file. With the "trained_model.sav" file in hand, we implemented and locally tested the web application using the python language distribution anaconda and with the help of API streamlit we produce the app in a few lines of code. Finally, we host the web app on the streamlit platform where you can test our customer grouping app.

3 Results

To answer business questions, in the data visualization stage, we plot a graph:

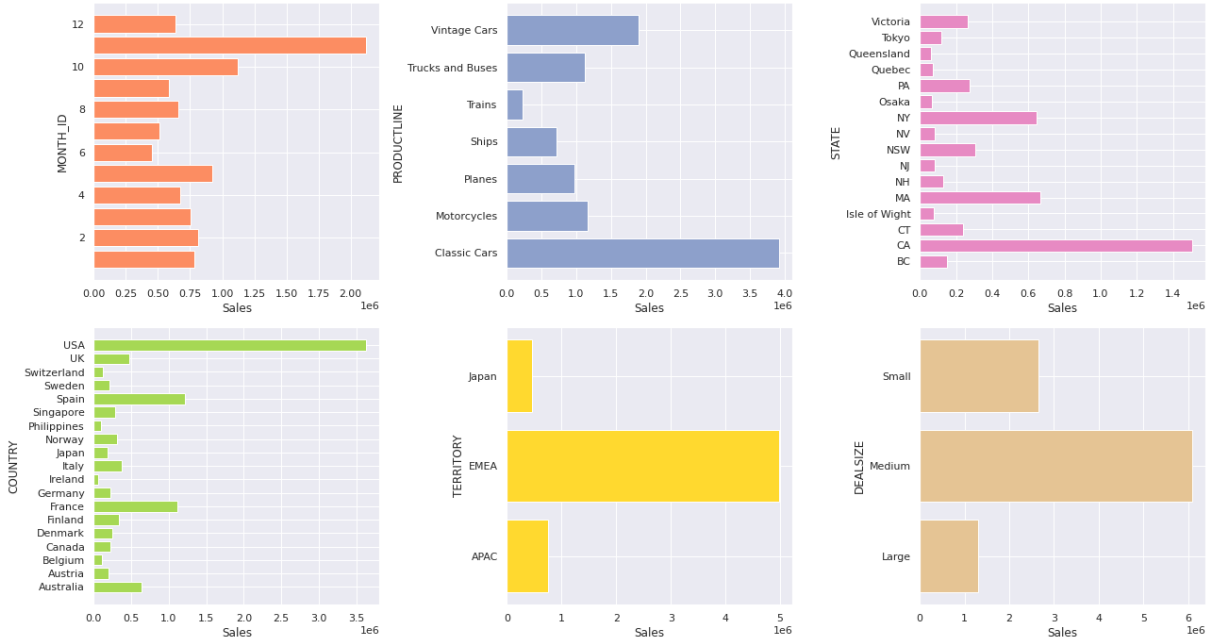


Source: Author

By this graph we understand how sales alternate with time, in addition we also note the maximum sales peak which is approximately 140000 at the end of the year 2004.

To understand the business patterns we plot the following chart:

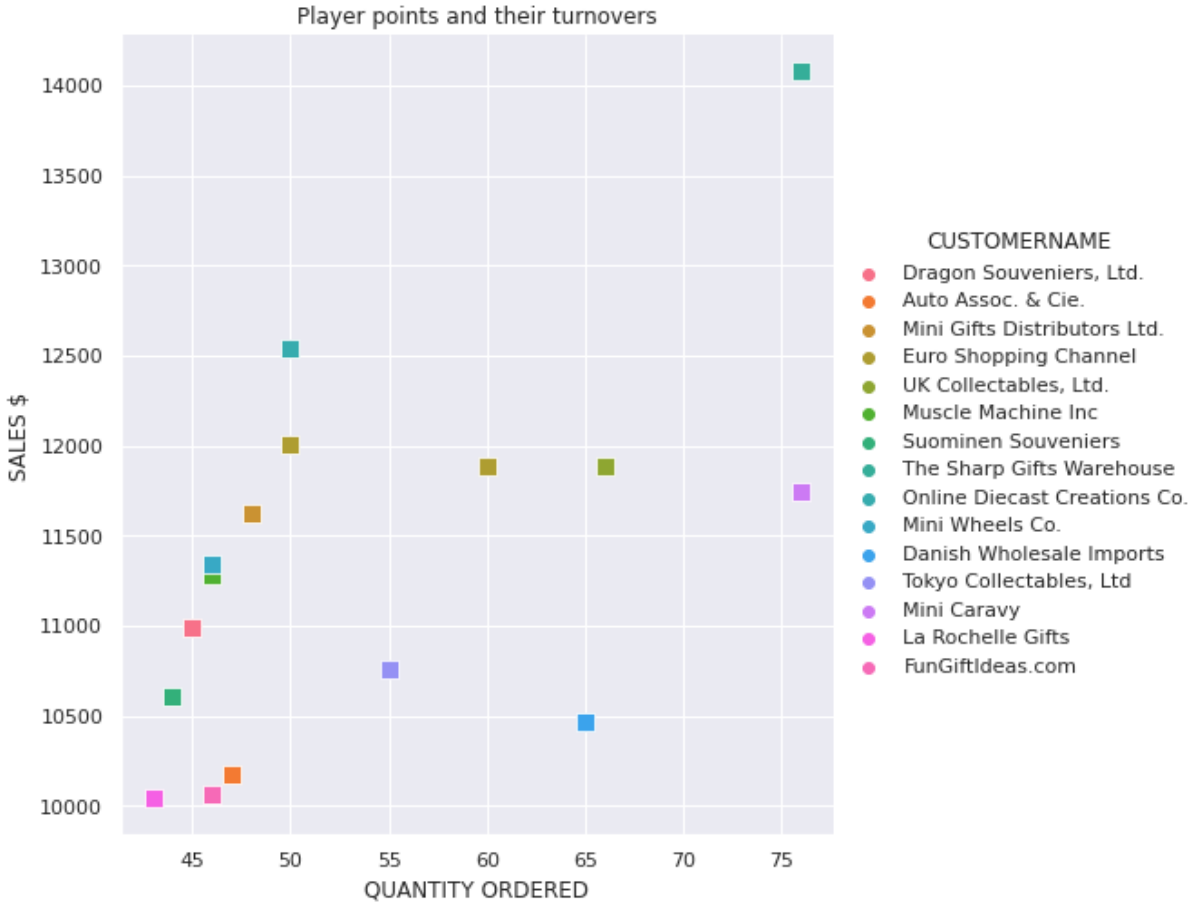
Sales Patterns



Source: Author

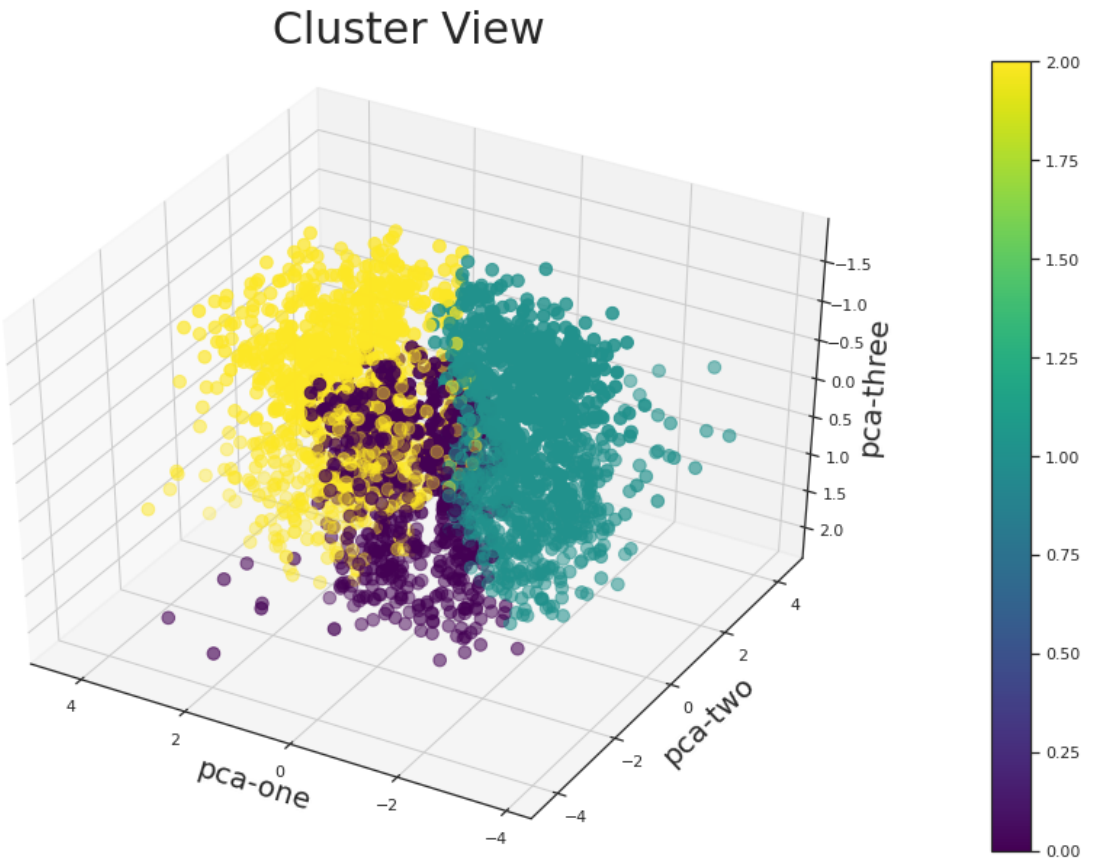
Such results indicate that the month, product line, state, country, Territory, business size that are most profitable are respectively: November, classic cars, CA(California), USA, EMEA (Europe, Middle East and Africa), Medium.

To know exactly who are the customers who sell the most (or who bought, who gave the data was vague in this information) we plot the number of orders per sales and highlight the name of the customers:



Source: Author

the client or company that bought the most was The Sharp Gifts Warehouse. Finally, as in our dataset we do not have predetermined target values, the problem in question to be addressed is clustering. In this way, we implemented an algorithm called K-means Clustering, such a model groups the data from k-centers in distinct areas of the feature space, which he thinks portray very different groups. Using the Elbow analysis, we found three groups and using dimensionality reduction we plot the graph with the groups:



Source: Author

1. In this group the quantity of goods ordered is about 27.5, the price of each good purchased is approximately 100 dollars.In addition, total purchases are approximately 3000 dollars, the largest purchases occur in December and the most consumed product line is classic cars, and the agreements with these individuals are medium to small;

2. In the current group, the quantity of goods ordered is about 45, purchases more products in the range of 100 dollars, as well as having total purchases of about 4000 dollars. Their purchases occur more frequently in December and the most consumed product line is classic cars.The agreements with this group are medium-sized;

3. In this group, the amount of merchandise ordered is about 30, buys more products in the range of 65 dollars, as well as having total purchases of about 1600 dollars. This group buys more in December and the most purchased product line is vintage cars. Agreements with this group are small.

4 Conclusion

In the course of this project, we developed relevant concepts in data science and analysis, in order to suggest the solution to know which group a given customer belongs to, to know what their preferences are, how that customer spends and what he spends on. Thus, in future applications it will be possible to make specific advertisements and discounts for these customers.

We went through several precise steps to extract as much information as possible from our dataset. As a result, we found that the month, product line, state, country, territory, deal size that are most profitable are respectively: November, Classic Cars, CA (California), USA, EMEA (Europe, Middle East and Africa), Medium. Using time series, we find that the maximum sales peak is approximately 140,000 at year-end 2004. Furthermore, we understand that the biggest buyer for this company was Sharp Gifts Warehouse.

Finally, we employ two techniques that can be verified in jupyter notebook in order to both segment K-meins clients and to download the PCA data dimension. That way we were able to divide customers into three types, those who spend around 3000, 4000 and 1600 dollars.

Bibliography

GéRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2ª. ed. GCanada: O'Reilly, 2019. Citado na página 3.

MüLLER, A. C.; GUID, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. 1ª. ed. United States of America: O'Reilly, 2016. Citado na página 3.