NOVA IMS Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

Visual Analytics for Decision Support: A Supply Chain Perspective

Nuno Penim - 20210998

Gonçalo Gomes - 20211007

Paulo Oliveira - 20211002

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

1. DATASET DESCRIPTION

In this final visualization project, we decided to work on a dataset for a superstore located in the United States, which is responsible for the sale of technology products, furniture and office supplies, and has resale facilities in all US states except Alaska. This dataset had data from 2014 to 2018.

Regarding the Dataset itself, although we have access to a diverse range of variables, we will emphasize those that we consider important to analyze the Delivery Process of this company.

As far as the numerical variables are concerned, we decided to focus our attention on the metrics that measure the economic performance of the company, such as "Profit", "Sales" and "Discounts". All of these have decimal places and will be the main objects of study of our visualization application.

In terms of spatial data, we consider that the "State" variable, which represents the North American state where the purchase order was made, would be the most relevant for what we want from our application. Thus, we decided not to use variables like "Region" for being less specific, and "City" for the opposite reason.

Furthermore, there are variables that allow us to categorize the products of a given purchase order, which we consider to be one of the main focuses of our project, being these, "Category" (already described above) and "Sub-Category" which is a specification of the previous one.

Finally, we also agreed to use the variable "Segment", which represents the type of client that made the purchase in question, that is, if it is from the "Consumer", "Corporate" or "Home Office" class. We will have the opportunity later on to better describe the importance of this variable in our visualizations.

At last, it is important to point out that each observation in our dataset represents a purchase order, by a certain client, for one or more products from the superstore.

2. VISUALIZATION AND INTERACTION CHOICES

2.1. PROJECT INSPIRATION AND CONTEXTUALIZATION

In this project, our inspiration lay in the theme we chose for our paper: "Visual Analytics for Decision Support: A Supply Chain Perspective". However, our focus this time was directed towards the delivery process, which is characterized as one of the most important processes in supply chains.

Thus, we chose this topic not only because it is a continuation of our previous work, but also, as our paper states, due to the huge usefulness that analytical visualization techniques offer to the retail industry, namely, for identifying the best sales strategy, improving the delivery of the respective products to customers' hands, and improving customers' shopping experiences.

In this regard, our main goal is to provide the company in question with all the visual tools it needs to improve its sales strategy and, consequently, be more successful in the retail market.

2.2. VISUALIZATION GRAPHS AND INTERACTION USED

We decided to incorporate six different visualizations that, in our opinion, allow us to achieve the objective initially outlined.

The first visual analysis we developed was a choropleth map, which is a type of statistical thematic map that uses color intensity to correspond to an aggregate summary of a geographic feature within spatial enumeration units. In our case, this map represents the profit per state associated with each product sub-category, with blue color intensity associated with higher or lower profit obtained, red color related to losses, and white color representing a certain neutrality regarding this topic. Along with this graphic, we have a RadioItems component that allows the user to choose the sub-category that he wants to see represented. We decided to create this chart in order to try to understand which states contribute most to the profitability of the company and, conversely, which states have the highest losses. This way, the company's managers can make the decision, for example, to maximize their facilities in the more profitable states and decrease their participation in the less profitable ones.

Next, we thought it would be interesting to establish the relationship between sales and profit for each subcategory in the superstore. To do this, we created a bar chart that represents this very idea, that is, for each subcategory we have two bars represented: a yellow bar representing total sales and a purple bar symbolizing the profit generated. The user is allowed to omit both the bars referring to sales and those referring to profit, in order to allow a better observation of the parameter that one wants to see represented. This view is extremely important because it allows the selection of the company's most profitable subcategories.

Subsequently, we decided to add a temporal dimension to the analysis of the company's profit, in this case, for each product category. Thus, we created a line graph that would allow us to visualize this same relationship over the 4 years of complete data (2014 to 2017). In this graph, furniture is represented by a yellow line, office supplies by the purple line, and technology by the blue line. As in the previous chart, there is also the standard plotly option to omit the desired categories. With all this, the company will be able to get a better sense of the evolution of profitability over the last few years of operation, which can be a huge advantage in identifying the factors that may have contributed to this oscillation.

For the next graph, we decided that it might also be interesting to show the total amount of discounts, in dollars, which the superstore applied to each of its categories and sub-categories. To do this, we decided to create a sunburst where the first variable described would be represented in the first layer of the graph, and the second in the last layer. We used it because we thought it would be the ideal visualization to represent the hierarchy present between the variable "Category" and "Sub-Category", and also due to the fact that this type of graphics supports a large number of variables. This way, the company can use this type of information to relate the gross amount of discounts applied, with the profitability that a particular category or sub-category presents.

Finally, we wanted to focus on the characteristics of each type of customer who shops at the superstore. To do this, we decided to create two types of visualization: a stacked bar chart and a violin/box plot.

Starting with the first, we had the clear objective of trying to understand if the profitability of the company differed according to the type of customer selected. To consummate this idea, we then chose the stacked bar chart where each bar would be divided into the three product categories previously announced. Thus, we have represented three bars in which each one represents the total profit associated with that same customer class, as well

as the portion of profit related to each product category. The interactivity of this graph is limited to the selection of the categories we want to be observed, just like in the previous graphs.

As for the violin/box plot, our goal was to provide the company with the opportunity to analyze the total sales distribution, by sub-category and by customer type. In this chart, we incorporated two types of interactivity: a dropdown component that allowed the choice of the desired product sub-category and a RadioItems component that allowed the change between the violin plot and the box plot. To some extent, are just two different ways of visualizing the same purpose.

3. TECHNICAL ASPECTS

For the implementation of this final project, we mainly used methodologies and techniques that were taught during the practical and theoretical classes. For graphical projection we used plotly, a powerful python library that allows users to create complex interactive web-based visualizations, which allows us to make all the necessary graphs. For the presentation of these graphs in the context of a web page, we used the Dash framework. At a lower level, for data processing, we used Data Science oriented libraries such as pandas, numpy and scipy.

Initially, we started by downloading our dataset, which was in csv format. Thus, we started by performing data exploration in order to understand the characteristics of the information present in the dataset. Afterwards, we realized that the data for the year 2018 was incomplete, so we chose to delete all entries for that year. Depending on the visualization we wanted to represent, the method of analysis and processing was different, however, the concept of data manipulation was almost similar in all situations. Here, we mostly used the pandas library to group, sort and select the desired data. At this stage, when implementing the violin/box plot, we quickly realized that our visualization products were highly distorted. This phenomenon was due to the existence of a huge amount of outliers, so we decided to use the z-score method to remove them (z-score greater than 3). With this, we did not intend to remove legitimate information at all, however, we consider the removed observations to be anomalies. In addition, in order to allow for more consistency in the visualization, we decided to round the numerical variables of interest to two decimal places. It is also important to mention that, in our dataset, there was no need to perform any data encoding techniques, due to the nature of the visualizations we decided to implement.

The next required step would be to apply HTML and CSS in order to develop our web page. To do this, we used the techniques that were imparted to us during the practical lessons. In addition, we also used a resource from existing projects, which can be found in the references chapter. From here we mainly extracted a CSS container to wrap our HTML Divs around, to make the web page aesthetically pleasing and organized.

Finally, it is important to mention that we implemented one callback for the selection of the sub-category in the choropleth map and another that incorporates the choice of the sub-category and the visualization type in the violin/box plot. Thus, the structure of these callbacks did not vary much from what was taught during the classes, with only minor modifications in the inputs and outputs in order to meet what we wanted.

4. DISCUSSION AND LIMITATIONS

In order to conclude this project, we thought it was important to make a short retrospective of the work developed.

The clear objective of this application was to provide a set of visual tools that would allow the company to draw conclusions on how to better manage its delivery process. It should be noted that our visualization tools are based on key factors that are fundamental to this phase of the supply chain, such as sales and profitability.

Therefore, we consider that our main objective has been successfully completed, despite the existence of some associated limitations. One of them, for example, concerns the lack of data regarding the supply chain return process. This bottleneck prevents us from making possible associations with customer satisfaction regarding the products they purchase and, consequently, makes it impossible for the company to make more conscious decisions in this regard.

Finally, and regarding possible improvements, we could increase the complexity of the project by trying to implement a clustering algorithm in sales for example. Thus, if we obtained distinct clusters, we could try to understand which characteristics made them different. It would be an interesting tool for the company, as it could apply different marketing strategies depending on the group in question.

5. REFERENCES

[Retrieved 28th of March of 2022] A. Khakpour, R. Colomo-Palacios and A. Martini, "Visual Analytics for Decision Support: A Supply Chain Perspective," *in IEEE Access*, vol. 9, pp. 81326-81344, 2021, doi: 10.1109/ACCESS.2021.3085496, https://ieeexplore.ieee.org/document/9445829

[Retrieved 4th of April of 2022] Superstore Dataset, https://www.kaggle.com/datasets/vivek468/superstore-dataset-final

[Retrieved 9th of April of 2022] The CSS pretty container, https://github.com/plotly/dash-sample-apps/tree/main/apps/dash-oil-and-gas

[Retrieved 6th of April of 2022] *Porcentagem: entenda como calcular desconto e lucro!* https://www.altoastral.com.br/entretenimento/porcentagem-desconto-lucro/

[Retrieved 6th of April of 2022] Nicole Madison, "What is the Relationship Between Revenue and Profit?", https://www.smartcapitalmind.com/what-is-the-relationship-between-revenue-and-profit.htm

[Retrieved 8th of April of 2022] Plotly documentation, https://plotly.com/python/

[Retrieved 10th of April of 2022] Dash documentation, https://dash.plotly.com/

Class materials

6. LINKS

Heroku > https://dv-grupo12.herokuapp.com/

GitHub > https://github.com/ColdFoxMaster/Data-Visualization