COMPANY		
Click here to access the company site	Project	Click here to access the project
	Key words	Pandas; data visualization techniques.
	Date	January - 2022
	Description	It was a data analysis project developed to developed with the aim of obtaining approval for an application to a data analyst vacancy at the Lookbox company.
	Index	1. Challenge proposed by the company 2. Report 3. Solution overview

1. Challenge proposed by the company

The challenge was proposed throw two steps. The first step was answering the questions below using Pandas.

1. How many products does the company have?
2. What are the 10 most expensive products in the company?
3. What sections do the 'BEBIDAS' and 'PADARIA' departments have?
4. Which store sold the most products in one day? Which day?
4. What was the total sale of products (in $) of each Business Area in the first quarter of 2019?

The second Step was developing a report about a dataset available throw a .csv file using python language. The visualizations chooses should be justified.

^Return to the index.

2. Report

Bellow, we have the answers to the first step of the challenge.

1.How many products does the company have?

```
df_dataProduct['PRODUCT_COD'].count()
```

2.What are the 10 most expensive products in the company?

```
df_dataProduct.sort_values('PRODUCT_VAL',ascending=False).head(10)
```

3.What sections do the 'BEBIDAS' and 'PADARIA' departments have?

```
filtro = df_dataProduct['DEP_NAME'].isin(['BEBIDAS','PADARIA'])
dfResultado = df_dataProduct[filtro].groupby(['DEP_NAME','SECTION_NAME'])
dfResultado.describe()
```

4.Which store sold the most products in one day? Which day?

```
df_dataStoreSales.sort_values('SALES_VALUE',ascending=False).head(1)
```

5.What was the total sale of products (in $) of each Business Area in the first quarter of 2019?

```
df_merge = pd.merge(df_dataStoreSales, df_dataStoreCad[['STORE_CODE','BUSINESS_NAME', 'BUSINESS_CODE']],
how = 'inner', on = 'STORE_CODE')
df_merge.dtypes
df_merge['DATE_CONVERT'] = pd.to_datetime(df_merge['DATE'])
filtro = (df_merge['DATE_CONVERT'] >= '2019-01-01') & (df_merge['DATE_CONVERT'] <= '2019-03-31')
df_merge[filtro].groupby(['BUSINESS_NAME','BUSINESS_CODE'])[['SALES_VALUE','SALES_QTY']].sum()
```

Fig. 2 shows the report developed as the second step of the report. It also may be accessed through the link: https://github.com/denise25maciel/Data-Analysis/blob/main/projetovisualizacaodados.py.

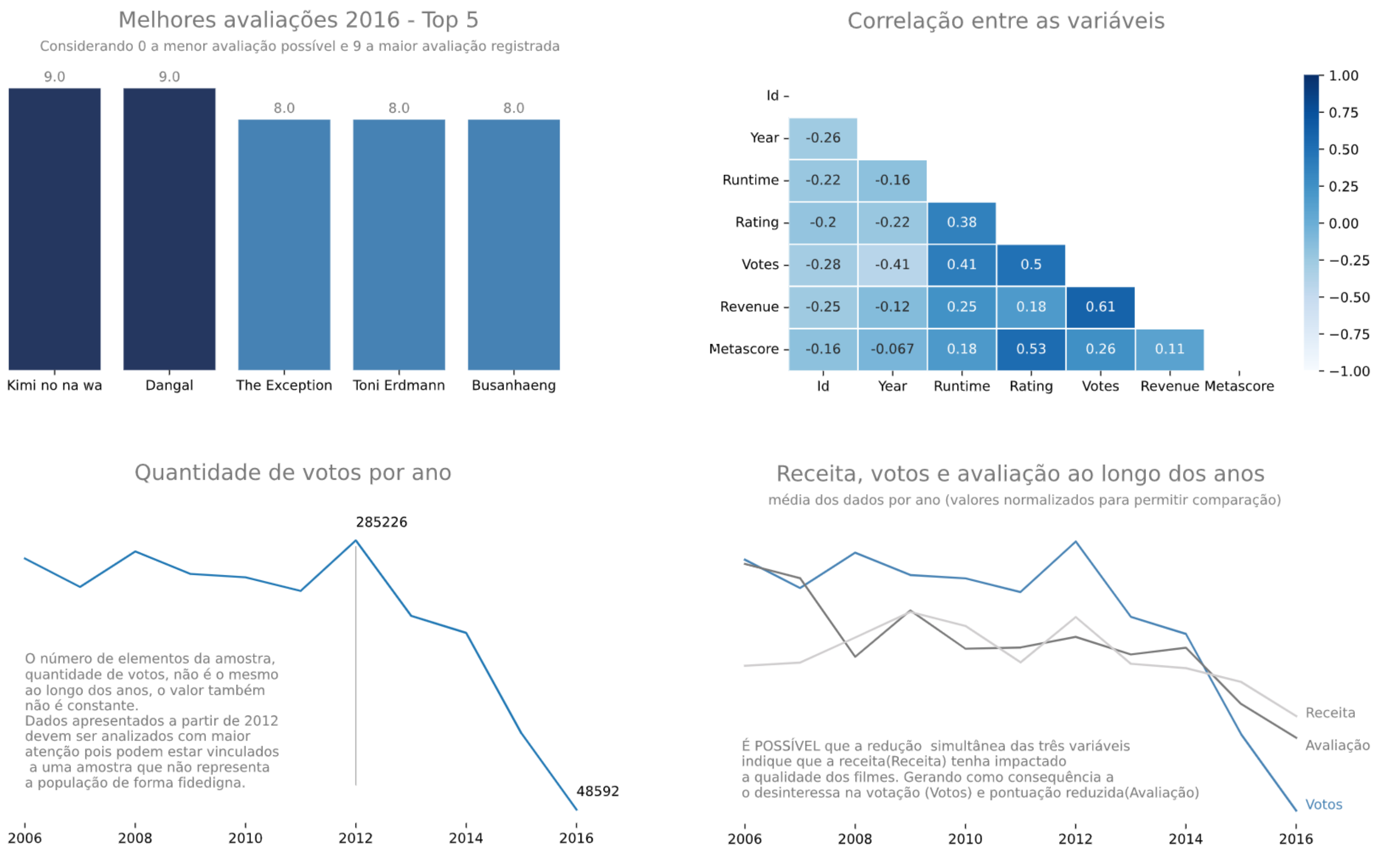

Fig. 2 – Final Report

The challenge required justification to the visualizations proposals. It will be presented at the next section.

^Return to the index.

3. Solution overview

Sequentially will be presented the justified to any visualizations choose presented at the Fig.2. The three mainly references were heurísticas de usabilidade de Nielsen, leis da Gestalt normas da engenharia semiótica e orientações provenientes da literatura sobre visualização de dados. Especificamente, o livro “Storytelling com dados”, Fig. 3.

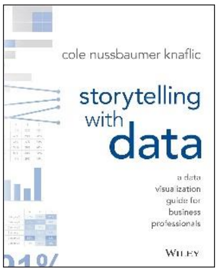

Fig. 3 – Storytelling with data

Among the viewing choices, there are:

- Minimalist design aimed at presenting elementary information was used;

- Although the company Looqbox uses the color green, the colors blue, gray and white were chosen because of color blind users. If green were used, the palette of green, gray and white would be used, which could be mistakenly interpreted by colorblind users as green and red (tones that create problems for this audience). In Brazilian culture, green and red are commonly associated as semiotic symbols of good and bad respectively. As this is not the intended information, the colors blue and white were chosen;

- Sans serif fonts were chosen because it was considered that the data would be presented in a digital environment. If the analysis were carried out in printed form, the serif version would be used, which is aimed at this purpose;

- The presentation of text labels was presented aligned to the right in respect to the fact that both reading and writing, in notebooks, for example, is carried out from the right margin towards the left margin;

- The presentation of the graphs followed the Z reading model. In other words, it is expected that the data reading follows the order presented at Fig. 4.

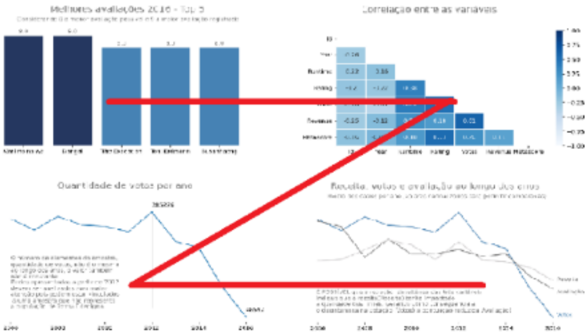

Fig. 4 – Z reading model.

- The choice of data to be represented in the graphs occurred because the data were obtained from samples of different sizes, Votes column. In this scenario, some graphical representations could lead the user to decisions that may be compromised by the sample size. Consequently:

Graphic 1 - Best reviews 2016 - Top 5

We chose to analyze only one year's data to avoid comparisons with data from different samples.

Important: the column chart is not best suited for data of a continuous, discrete nature. Even so, it was used because the respective column contains only discrete data and the representation is well known.

Graph 2 - Correlation between variables

As the frequency analysis could be compromised by the sample size, we decided to analyze how much the sample growth impacts on the other variables. The relationship between Votes, Rating and Revenue was verified and subsequently explored.

Graph 3 - Number of votes per year

The graph was designed for the user to acquire temporal contextualization of the sample size.
Important: In this graph, the presentation of the average line was initially designed, as there are examples on the Looqbox company page. However, the average deviation value for the column was greater than 100%, which implies that comparisons with the average value do not provide meaningful information for the end user.

Graph 4 - Revenue, votes and rating over the years

In this graph, the results of the strongest correlations obtained through Graph 2 were compared. As the measurement units are different between the Votes, Rating and Revenue variables, the data were normalized. The color blue was used for Votes, because it is the variable with which the comparison is primarily intended. The remaining variables are shown in gray scale.

Important: although the analysis began by analyzing the impact of the number of votes, graph 4 raises the hypothesis that the reduction in the Votes and Rating variables may be related to the reduction in the Rating variable.

^Return to the index.