Furni_EDA

July 25, 2021

Furniv - EDA

As part of any data analysis process, an initial exploratory phase is required to familiarize with the data and to find new information on the issue evaluated. That is how, by reviewing and cleaning the data and by plotting different variables in diferent formats, new insight is gathered and proper modeling aproximations are proposed.

As so, we started reviewing the data with the idea of getting initial information on how historical sales affected the behavior of the stock and the product portfolio.

1 Data consolidation

First we cleaned all tables to have a standarized, relevant data with no outliers, NaN values and precise working information. We also created some new fields for some tables to be able to work on differents asspects such as dates, quentities and categorical variables. Finally we merged some tables to have aditional information os the sales and current stock of the client's company.

1.1 Data composition

The data sent by the client is comprised by three tables `products_master` (renamed to `products`), `current_stock` (renamed to `stock`) and `sales`. From those tables, the tables `sales_prod` and `stock_prod` were created to relate the current stock and historical sales to the characteristics of each product. Finally a grouped table `groupby_sales` was created from the `sales_prod` table where each product has one row and the quantity values are aggregated for statistical analysis (`amount_sold`, `average_price`, `average_discount`, `sale_subtotal`, `sale_total` and `sale_average`). Here we can check the `.info()` values for each table.

1.1.1 `products`:

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ITEM              1285 non-null   int64
 1   REF               1285 non-null   object
 2   DESCRIPCION       1285 non-null   object
 3   CATEGORIA         1285 non-null   object
 4   SUBCATEGORIA      1285 non-null   object
 5   VIGENCIA          1285 non-null   object
 6   ORIGEN            1285 non-null   object
 7   ESTILO            1285 non-null   object
 8   MATERIAL          1285 non-null   object
 9   ACABADO           1285 non-null   object
 10  PUESTOS           1017 non-null   float64
 11  COLOR             1285 non-null   object
 12  ANCHO             1285 non-null   float64
 13  ALTO              1285 non-null   float64
 14  FONDO             1285 non-null   float64
 15  DESC_LARGA        1285 non-null   object
 16  SUBCATEGORIA_POS  1285 non-null   object
 17  COLOR_POS         1285 non-null   object
 18  MATERIAL_POS      1285 non-null   object
```

1.1.2 sales:

```
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   ID             36050 non-null  object
 1   NRO_DOCUMENTO  36050 non-null  object
 2   FECHA          36050 non-null  datetime64[ns]
 3   CODIGO_TIENDA  36050 non-null  int64
 4   TIENDA         36050 non-null  object
 5   PROD_REF       36050 non-null  object
 6   CANTIDAD       36050 non-null  int64
 7   PRECIO         36050 non-null  float64
 8   SUBTOTAL       36050 non-null  int64
 9   DESCUENTO(%)   36050 non-null  float64
 10  TOTAL          36050 non-null  int64
 11  ANIO           36050 non-null  int64
 12  MES            36050 non-null  int64
 13  DIA            36050 non-null  int64
```

1.1.3 stock:

```
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   ID            941 non-null    object
 1   REF           941 non-null    object
 2   CANTIDAD      941 non-null    int64
 3   CATEGORIA     941 non-null    object
 4   SUBCATEGORIA  941 non-null    object
 5   DETALLE_1     877 non-null    float64
 6   DETALLE_2     75 non-null     float64
```

1.1.4 sales_prod:

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ID                36180 non-null  object
 1   NRO_DOCUMENTO     36180 non-null  object
 2   FECHA             36180 non-null  datetime64[ns]
 3   CODIGO_TIENDA     36180 non-null  int64
 4   TIENDA            36180 non-null  object
 5   PROD_REF          36180 non-null  object
 6   CANTIDAD          36180 non-null  int64
 7   PRECIO            36180 non-null  float64
 8   SUBTOTAL          36180 non-null  int64
 9   DESCUENTO(%)      36180 non-null  float64
 10  TOTAL             36180 non-null  int64
 11  ANIO              36180 non-null  int64
 12  MES               36180 non-null  int64
 13  DIA               36180 non-null  int64
 14  ITEM              36180 non-null  int64
 15  REF               36180 non-null  object
 16  DESCRIPCION       36180 non-null  object
 17  CATEGORIA         36180 non-null  object
 18  SUBCATEGORIA      36180 non-null  object
 19  VIGENCIA          36180 non-null  object
 20  ORIGEN            36180 non-null  object
 21  ESTILO            36180 non-null  object
 22  MATERIAL          36180 non-null  object
 23  ACABADO           36180 non-null  object
 24  PUESTOS           29834 non-null  float64
 25  COLOR             36180 non-null  object
 26  ANCHO             36180 non-null  float64
 27  ALTO              36180 non-null  float64
 28  FONDO             36180 non-null  float64
 29  DESC_LARGA        36180 non-null  object
 30  SUBCATEGORIA_POS  36180 non-null  object
 31  COLOR_POS         36180 non-null  object
 32  MATERIAL_POS      36180 non-null  object
```

1.1.5 stock_prod:

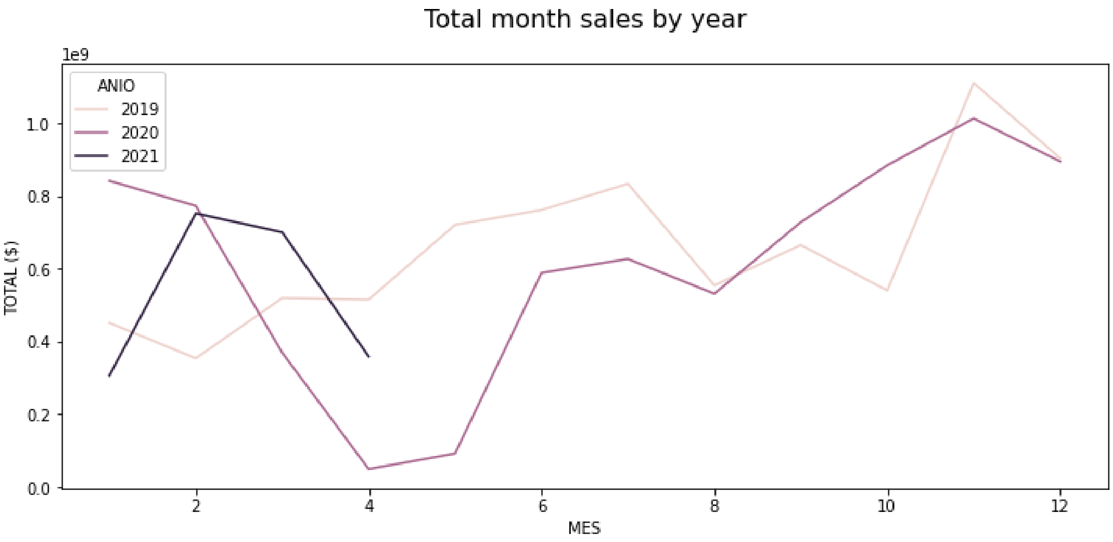
```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ID                943 non-null    object
 1   REF               943 non-null    object
 2   CANTIDAD          943 non-null    int64
 3   DETALLE_1         879 non-null    float64
 4   DETALLE_2         75 non-null     float64
 5   ITEM              452 non-null    float64
 6   DESCRIPCION       452 non-null    object
 7   CATEGORIA         452 non-null    object
 8   SUBCATEGORIA      452 non-null    object
 9   VIGENCIA          452 non-null    object
 10  ORIGEN            452 non-null    object
 11  ESTILO            452 non-null    object
 12  MATERIAL          452 non-null    object
 13  ACABADO           452 non-null    object
 14  PUESTOS           312 non-null    float64
 15  COLOR             452 non-null    object
 16  ANCHO             452 non-null    float64
 17  ALTO              452 non-null    float64
 18  FONDO             452 non-null    float64
 19  DESC_LARGA        452 non-null    object
 20  SUBCATEGORIA_POS  943 non-null    object
 21  COLOR_POS         943 non-null    object
 22  MATERIAL_POS      943 non-null    object
```

1.1.6 groupby_sales:

```
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   REF                  1285 non-null   object
 1   CANTIDAD             1285 non-null   int64
 2   TOTAL                1285 non-null   int64
 3   PRECIO_PROMEDIO      1285 non-null   float64
 4   DESCUENTO_PROMEDIO   1285 non-null   float64
 5   ITEM                 1285 non-null   int64
 6   DESCRIPCION          1285 non-null   object
 7   CATEGORIA            1285 non-null   object
 8   SUBCATEGORIA         1285 non-null   object
 9   VIGENCIA             1285 non-null   object
 10  ORIGEN               1285 non-null   object
 11  ESTILO               1285 non-null   object
 12  MATERIAL             1285 non-null   object
 13  ACABADO              1285 non-null   object
 14  PUESTOS              1017 non-null   float64
 15  COLOR                1285 non-null   object
 16  ANCHO                1285 non-null   float64
 17  ALTO                 1285 non-null   float64
 18  FONDO                1285 non-null   float64
 19  DESC_LARGA           1285 non-null   object
 20  SUBCATEGORIA_POS     1285 non-null   object
 21  COLOR_POS            1285 non-null   object
 22  MATERIAL_POS         1285 non-null   object
```

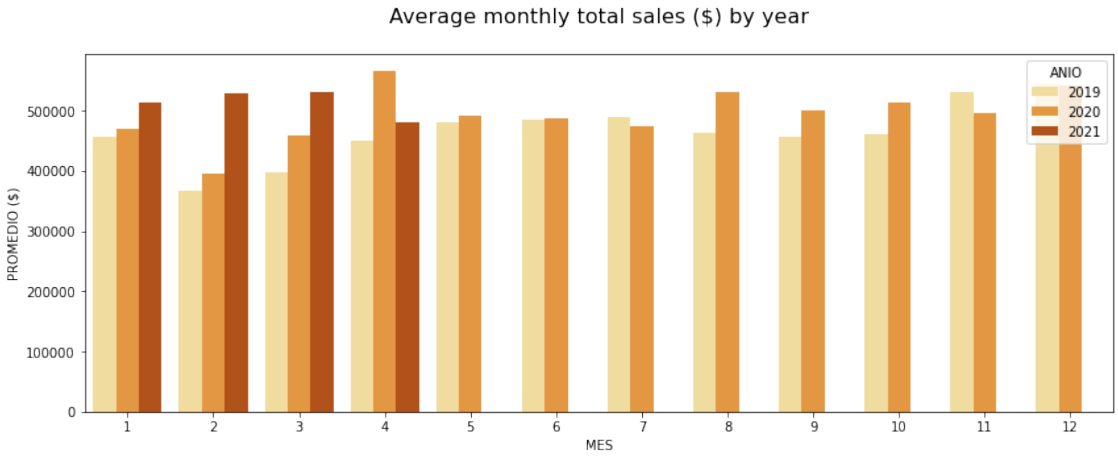
2 Sales analysis

To continue the analysis, we started by reviewing the sales behavior to find information about the most important products for the company according to the value of the total sales and the amount of items sold. Here we can see the total value of the items sold during the last tree years.



In the graph we can see different behaviors. Fist we observe a normal year with a growing welling trend with peak during the last months of the year. The, for 2020 se wee the efects of the COVID pandemic restritions imposed by the colombian goverment during March, April and May, and then again a normal, thought slower, grouth trend to again gain a peak performance during October, November and December. And finally we have this years data that has a cut up to April 15 which explains the low behavior of that month compared to the, presumed growth, of 2019.

In the following second graph we have the average sale value of every month. in there we confirmed the suspission that even though the pandemic affected the amount of sales made during that period, the sales that were made had a similar mean value from the previous year and the following year. Which in turn translates on the general idea that the company has a similar average sales value every month and that what changed on those initial month of the pandemic was the anmount of sales made those periods.

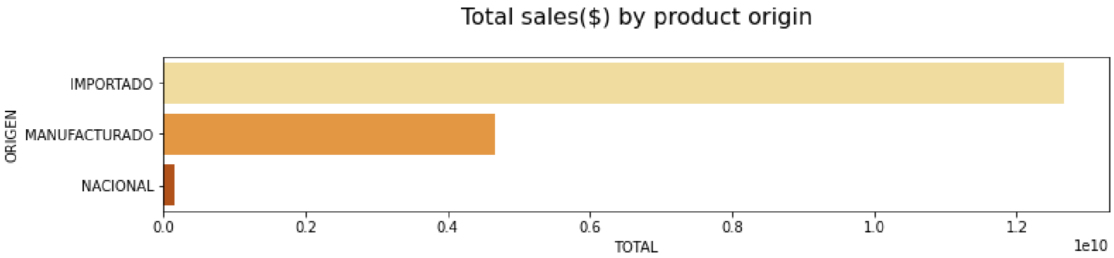


In this reggard we can see that the average sales per year for the company is arround 7,600 MM. In 2019 of 7,923MM, in 2020 7,387MM and up to Arpil 2021 of 2,116MM. All of which allows us to clonclude that 2020 was, nonetheless the efects of the pandemic, a really positive and profitable year for the company. In facty the company was able to considerably recover to almost get the same values from 2019.

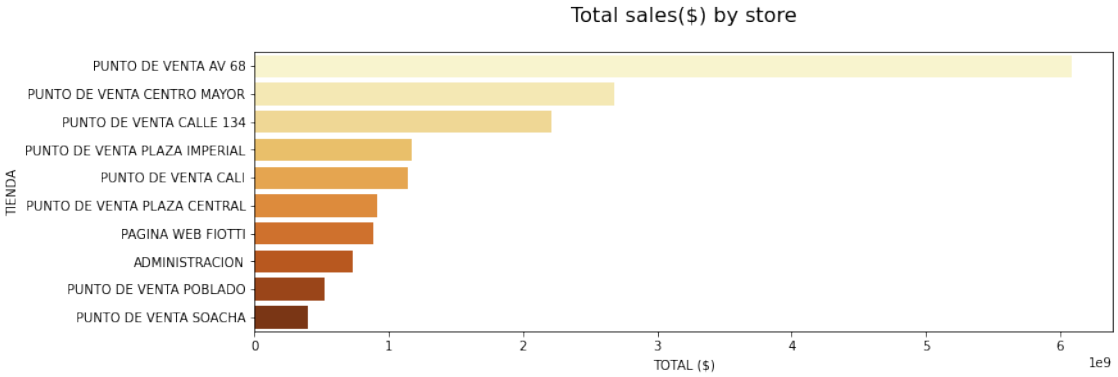
In relation to 2021 sales, we can also see that they have a 15% increase vs. 2019 del 15% and about 4% vs. 2020. A favorable anual company growth.

2.1 Characterized sold portfolio

Now we start to get information on some of the characteristicas of the sold portfolio of the company. Firts we see the distribution in relation to the product origin, the store it was sold and the main sub categories sold. Then we review the general information on the most sold styles and colors.

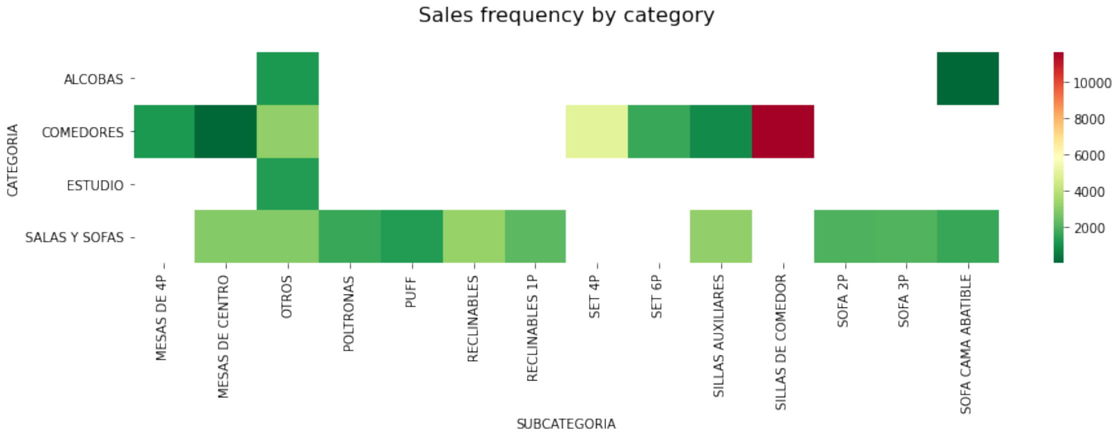


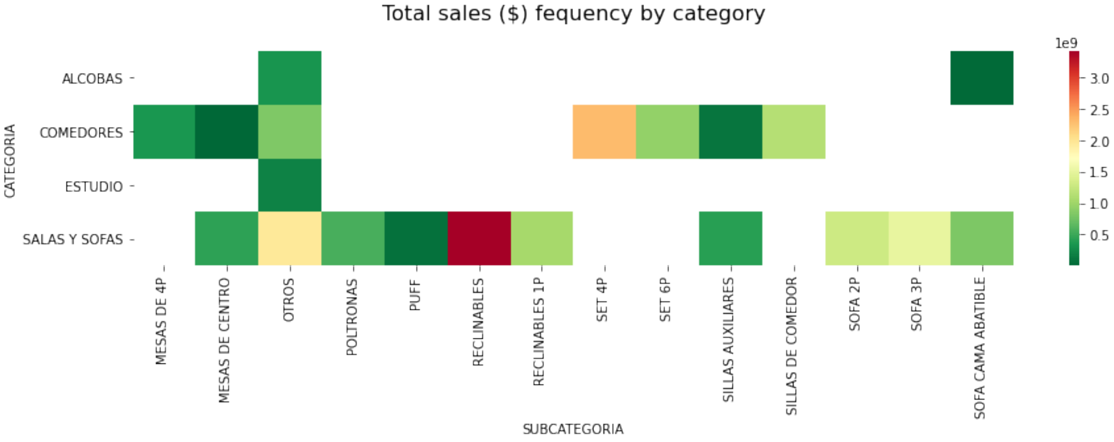
We can see here that the company's sold posrtfolio is comprised of imported products, with a 68% of the sold portfolio, and fabricated prodcuts with almost 30% of the portfolio; leaving the other categories with less than 2%.



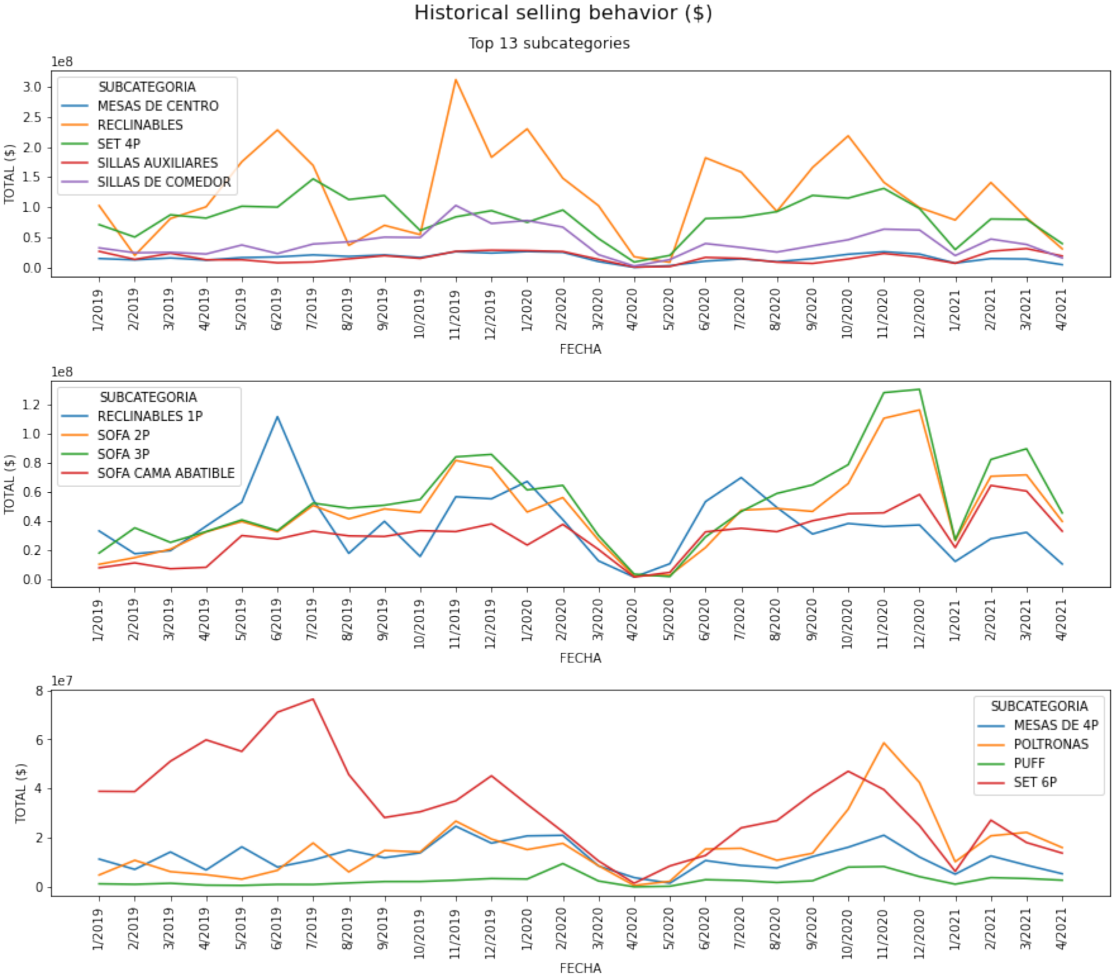
In this second graph we observe that most of the sales where made in Bogotá in phisical stores (the fisrt 4 stores comprise almost the 70% of the total sold portfolio), followed by sales from Cali and the webpage.

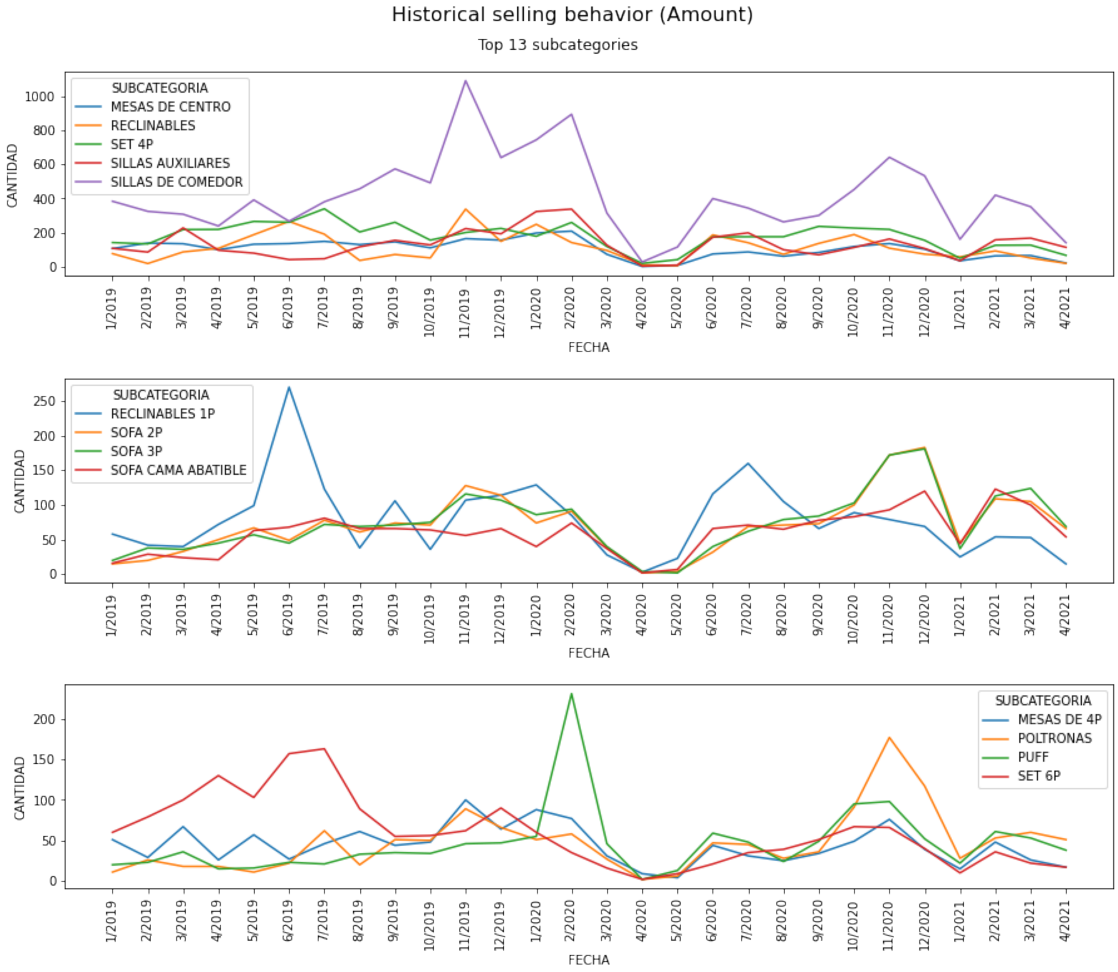
Next, we can review that the top 13 sub categories of products sold, wich is almost 80% of the total sold portfolio, are dining and living room elements with dinning room chairs with an exsesive amount sold.





As we can see, althoug the heat map informs us of the general composition of the sold portfolio, we are missing information on historical trends that could impact the way we decide on the proper model. As there could be other confounding variables not taken into account. For that we decided to look for possible historical trends in each of the 13 subcategories previously mentioned.



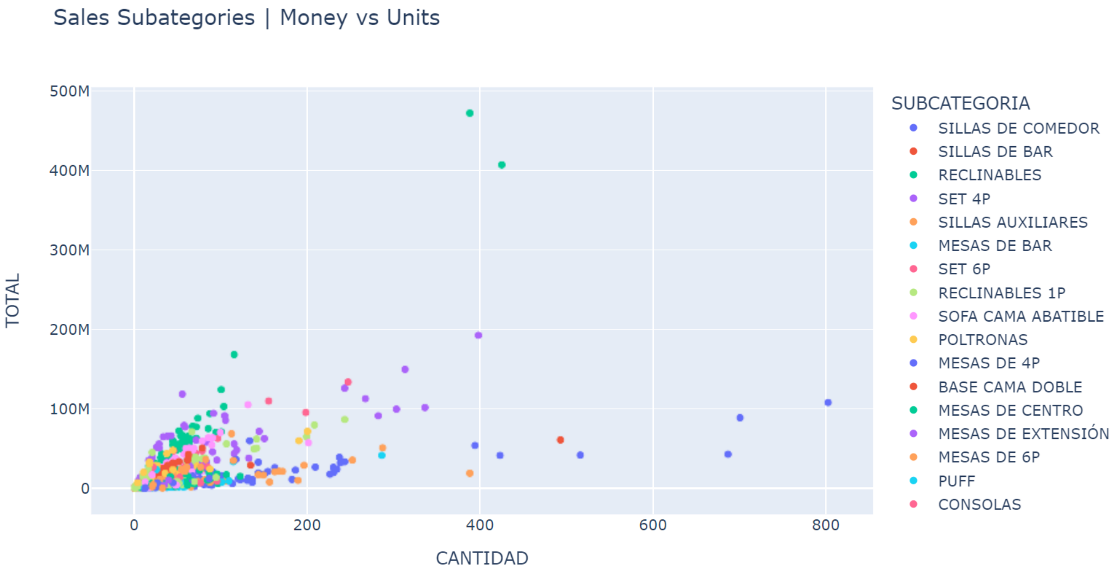


As of this point, it's still not clear if each element hsa particular trend or relevant pattern. Nonetheless, it seems that on september each year perople are more willing to buy furniture for the dinning room that on other months. On the graph we can also see selling peaks for some elemnts in november due to BlackFriday.

Here we have the top 5 selling products with their respective amount sold since 2019 and a plot tracing the relation between the value of the sold elements in the portfolio and the amount sold of each of those elements.

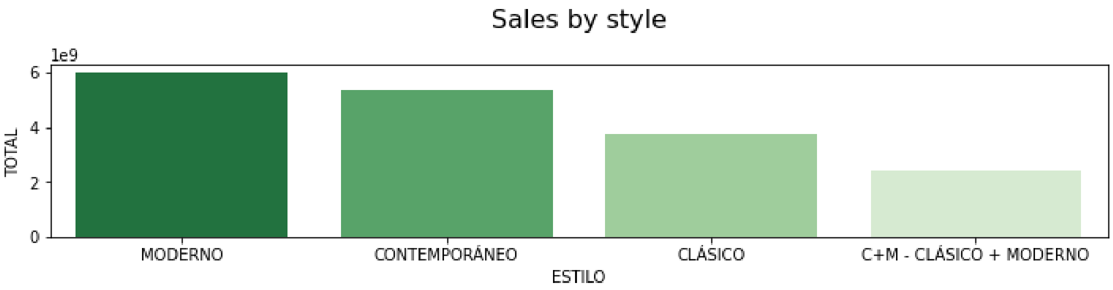
```
                   CANTIDAD           TOTAL
SUBCATEGORIA
RECLINABLES            3222  $3,427,202,692
SET 4P                 4982  $2,304,825,080
SOFA 3P                2041  $1,501,181,507
SOFA 2P                1995  $1,291,212,242
SILLAS DE COMEDOR     11657  $1,126,930,988
```

```
<Figure size 432x288 with 0 Axes>
```

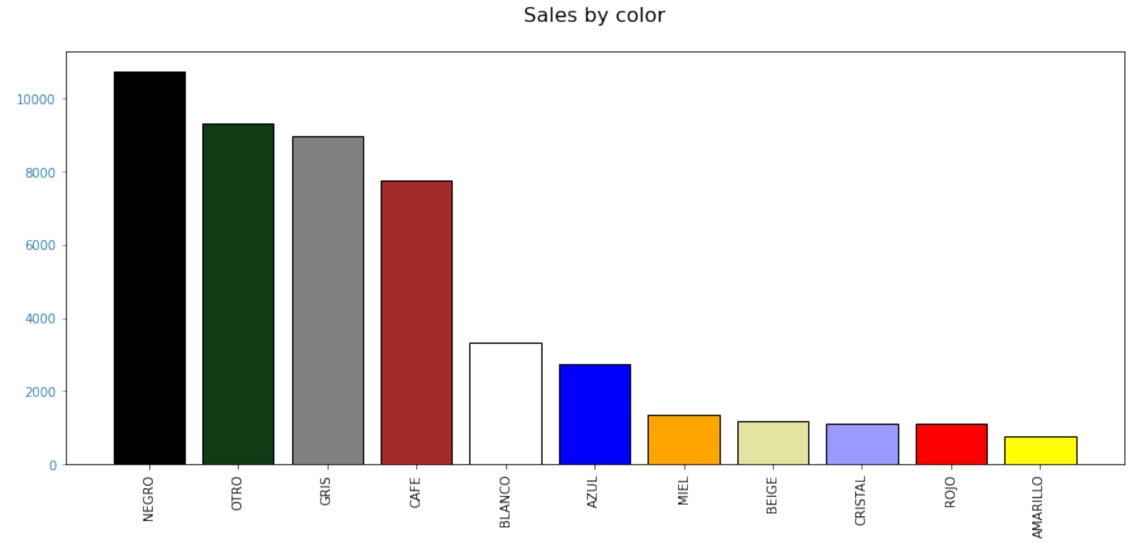


Sales Subategories | Money vs Units

As mentioned, we now present the most representative styles and colors in the sold portfolio, and the historical analysis on the chosen color as we want to find if the color is an important variable and has a specific impact on the products being sold. Decision also based on the fact that the predictive demand tool to develop will need to account for color variations on each product.



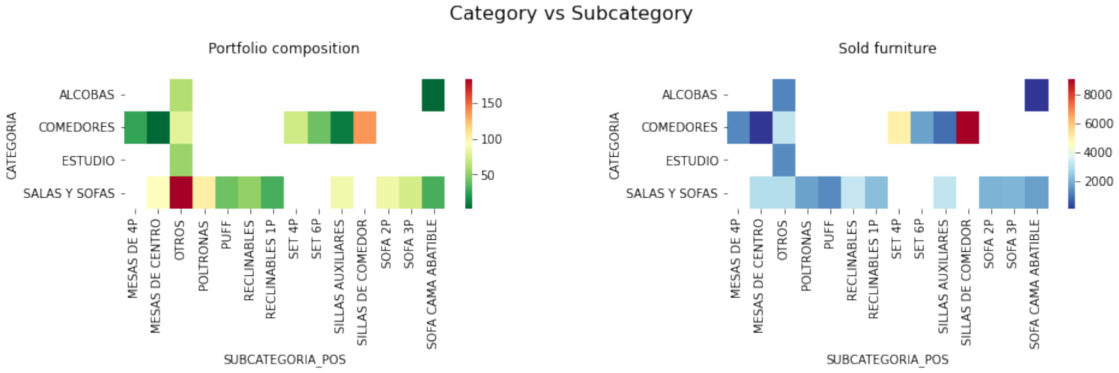
Sales by style

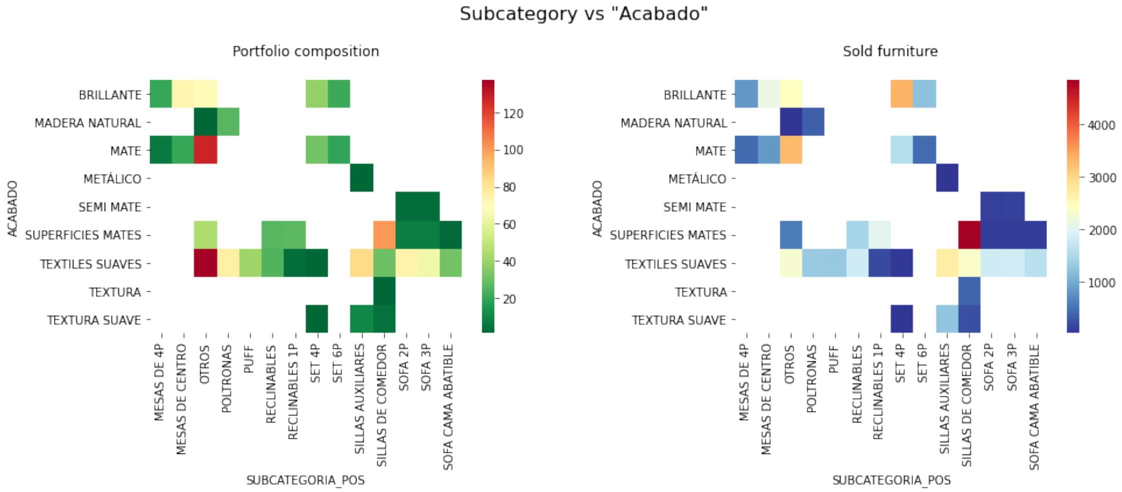


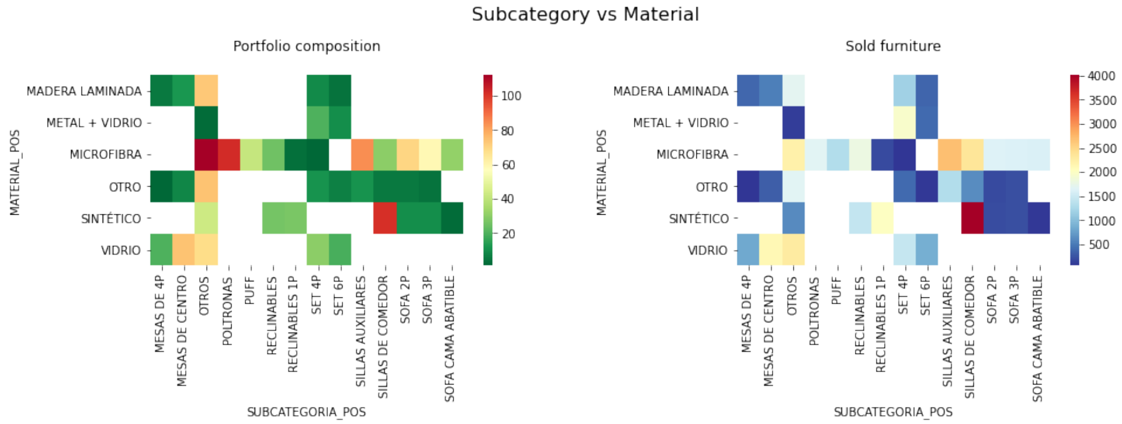
2.2 Products to sales comparison analysis

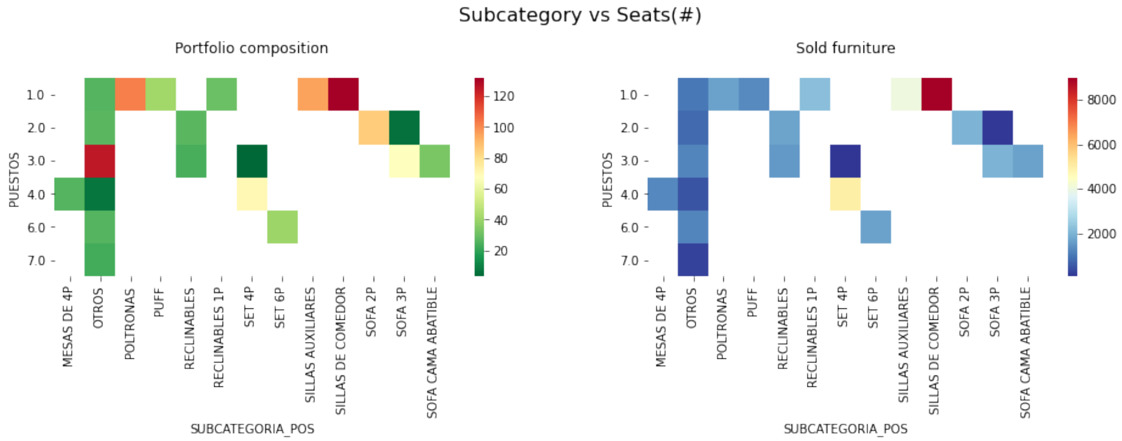
After reviewing the behavior of the sales during th past years we saw the oportunity to check whether the finding had relevant aditional information in relation to the actual composition of the whole portfolio, and the way each subcategory realtes to the "acabado", main material, number of seats and color.

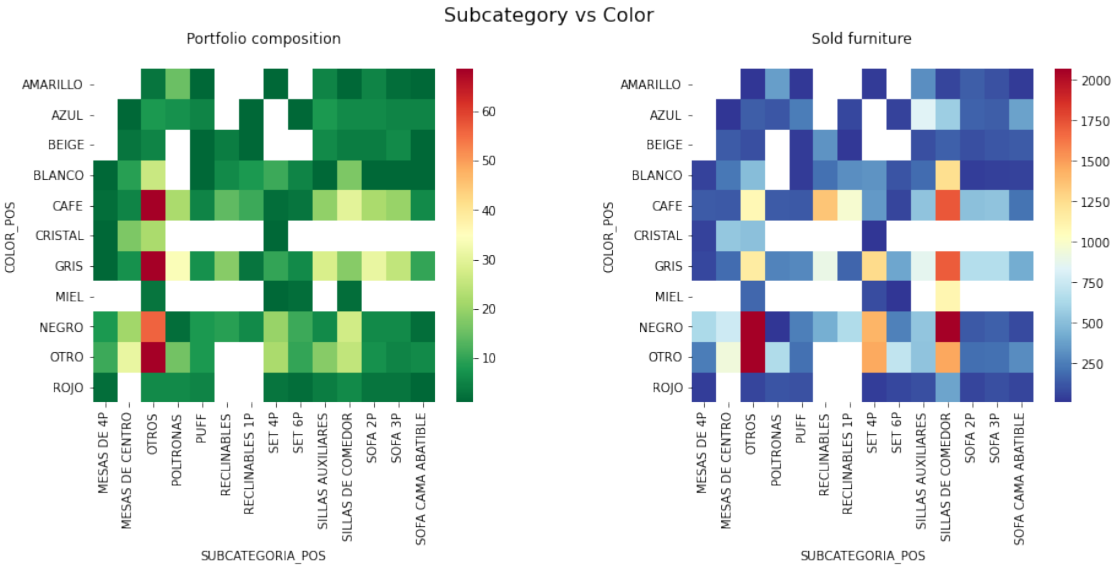
Here we have the comparable plots between the composition of the sold portfolio and the actual portfolio of the company. For each plot we can check in red the composition of the portfolio in terms of the selected variables, and in blue the composition of sold furniture during the past years in terms of the same variable.




Subcategory vs "Acabado"
Portfolio composition
Sold furniture
ACABADO
BRILLANTE
MADERA NATURAL
MATE
METÁLICO
SEMI MATE
SUPERFICIES MATES
TEXTILES SUAVES
TEXTURA
TEXTURA SUAVE
MESAS DE 4P
MESAS DE CENTRO
OTROS
POLTRONAS
PUFF
RECLINABLES
RECLINABLES 1P
SET 4P
SET 6P
SILLAS AUXILIARES
SILLAS DE COMEDOR
SOFA 2P
SOFA 3P
SOFA CAMA ABATIBLE
SUBCATEGORIA_POS
120
100
80
60
40
20
4000
3000
2000
1000


Subcategory vs Material
Portfolio composition
Sold furniture
MATERIAL_POS
MADERA LAMINADA
METAL + VIDRIO
MICROFIBRA
OTRO
SINTÉTICO
VIDRIO
MESAS DE 4P
MESAS DE CENTRO
OTROS
POLTRONAS
PUFF
RECLINABLES
RECLINABLES 1P
SET 4P
SET 6P
SILLAS AUXILIARES
SILLAS DE COMEDOR
SOFA 2P
SOFA 3P
SOFA CAMA ABATIBLE
SUBCATEGORIA_POS
100
80
60
40
20
4000
3500
3000
2500
2000
1500
1000
500

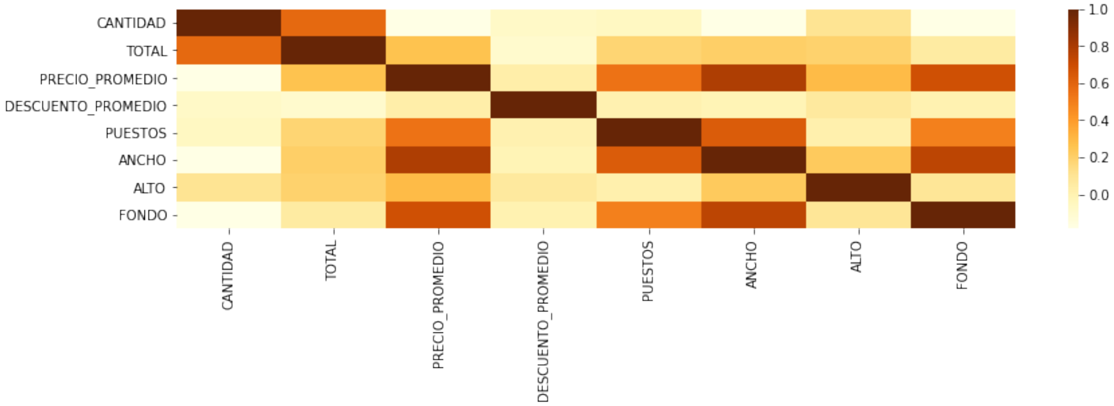

Subcategory vs Seats(#)
Portfolio composition
Sold furniture
PUESTOS
1.0
2.0
3.0
4.0
6.0
7.0
MESAS DE 4P
OTROS
POLTRONAS
PUFF
RECLINABLES
RECLINABLES 1P
SET 4P
SET 6P
SILLAS AUXILIARES
SILLAS DE COMEDOR
SOFA 2P
SOFA 3P
SOFA CAMA ABATIBLE
SUBCATEGORIA_POS
120
100
80
60
40
20
8000
6000
4000
2000



By checking the previous tables we see that for the most trending subcategories, classified by colors, materials, “acabados” and the number of seats; the portfolio behaves similar to the number of sales. We could argue that the seen correlation between both behaviors is explained due to the fact that the probability of buying an element of a certain color, material, “acabado” or amount of seats, depends much on how many possibilities the customer has when choosing one of those characteristics. That is why the most sold characteristics are the most diverse.

2.3 Extra correlations

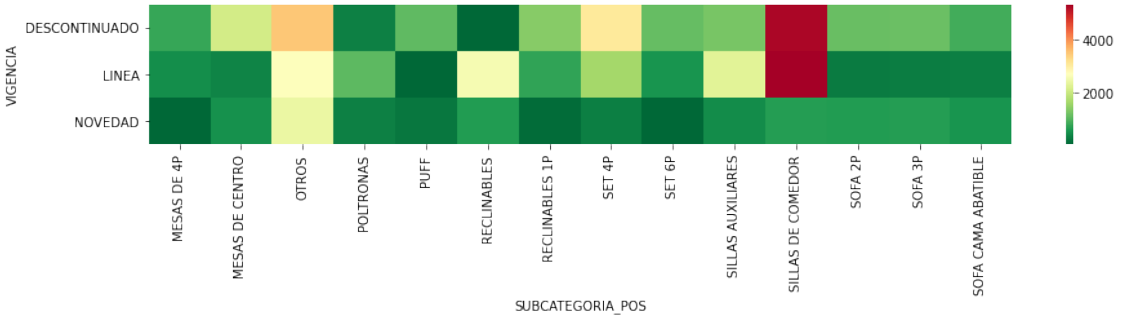
Following, we have some other data of the portfolio such as the correlation matrix between amount sold, value sold, price, discount, measures, and the number of seats to assert the considered relation between the price and the measurements and number of seats.



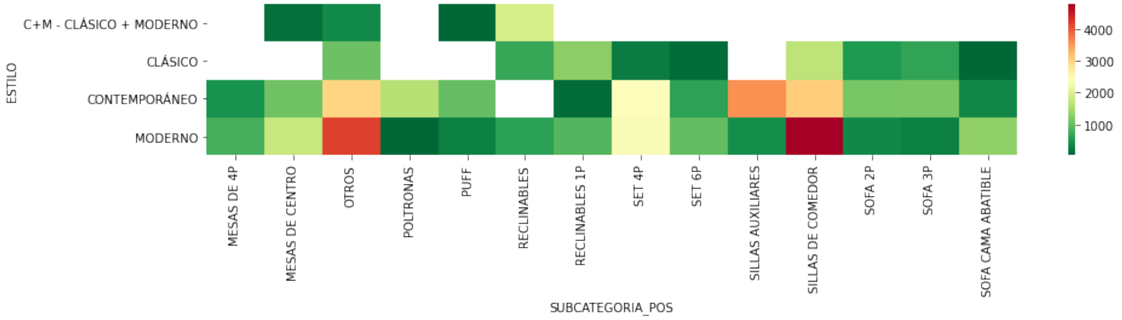
In here we can see, as expected, that there is a high correlation between total sale price and the amount of products sold,the average price and the dimensions of the product. Making it clear that even though those variables affect the price they are not relevant to further analysis.

Other variables, on the contrary, seem to have certain impact on the sales behavior and should be taken into consideratin for modeling a proper statistical predictive tool. Here we have some:

2.3.1 Validity (`vigencia`)

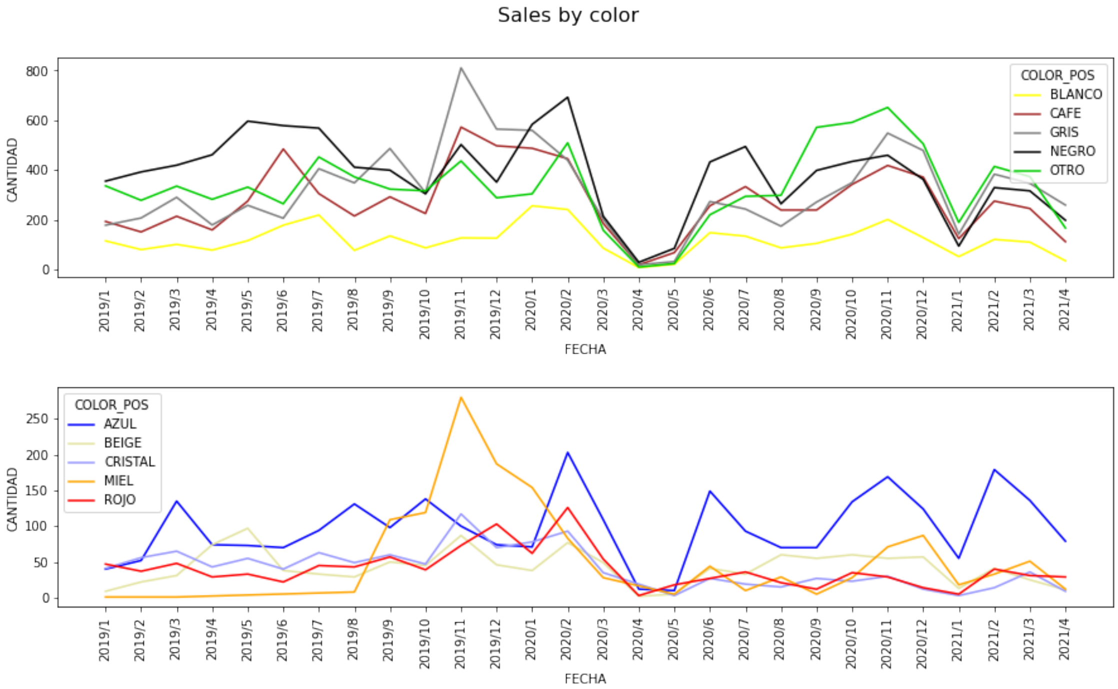


2.4 Style (`estilo`)



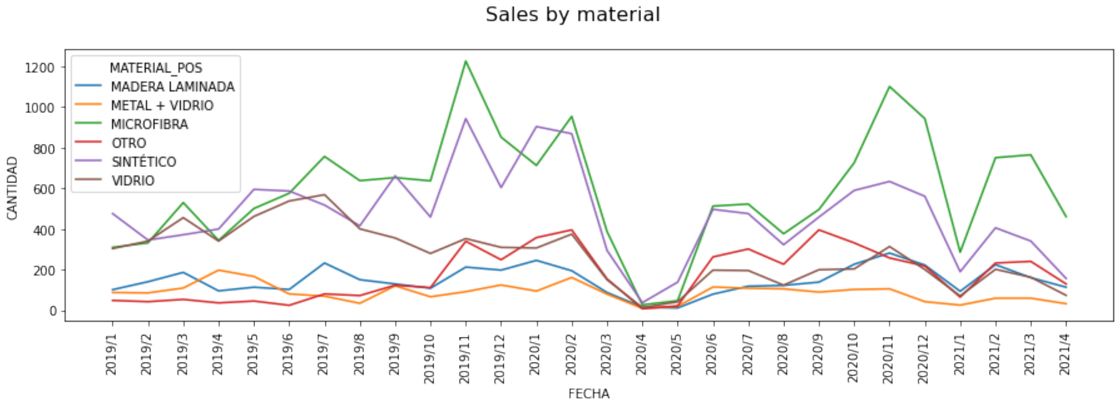
2.5 Historical sales analysis by features

For the next part of the analysis we created a time series to review the behavior of certain features of the furtinute over time. First we start by reviewing colors, then materials, then “acabados” and finally number of seats. For each of these features we only use the 13 top selling categories that represent 80% of the sold portfolio. Finally is worth to mention that for each plot we are only reviewing the amount of units sold and not the total sales value.



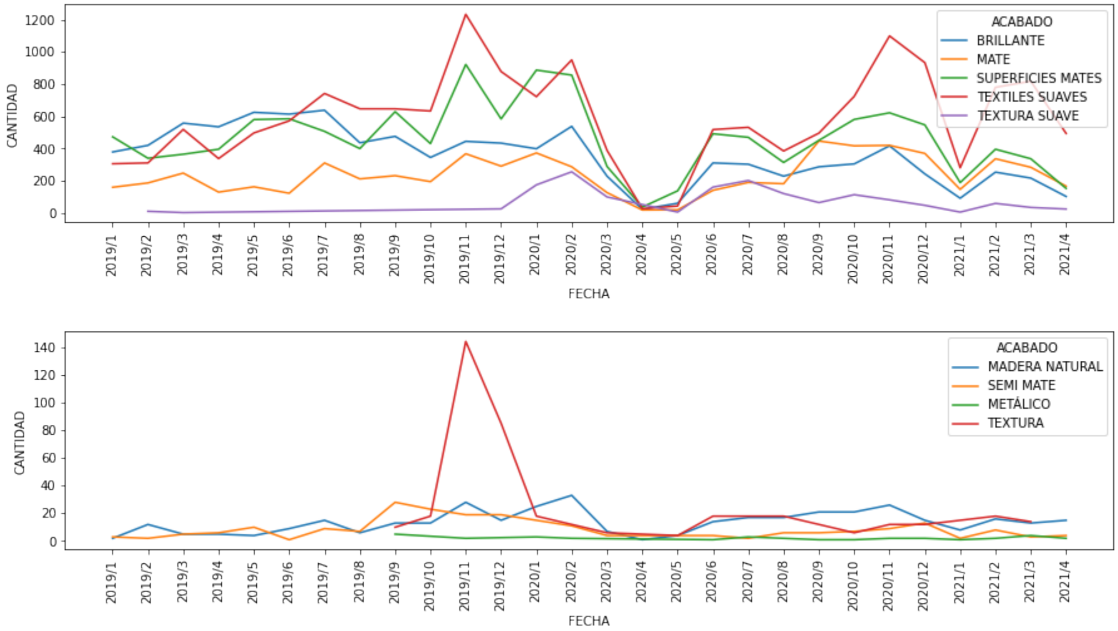
In the previous graph we can see a similar behavior on all colors with peaks over the last months of the year. Nonetheless no further information is provided that could indicate a prevalence of a specific color in a certain period of time. most colors behave the same way with black, gray and brown being the most sold colors. Here we can also confirm the efects of the confinement decreed during the first months of the pandemic and their efect over the sales.

To further investigate the relevance of certain features over diferent periods of time we can see the historical sales of different materials, where `laminated wood` and `metal + glass` are the top selling materials, with a mean sales average over 500 units per month, and the rest of the colors falling behind bellow the 200 line.



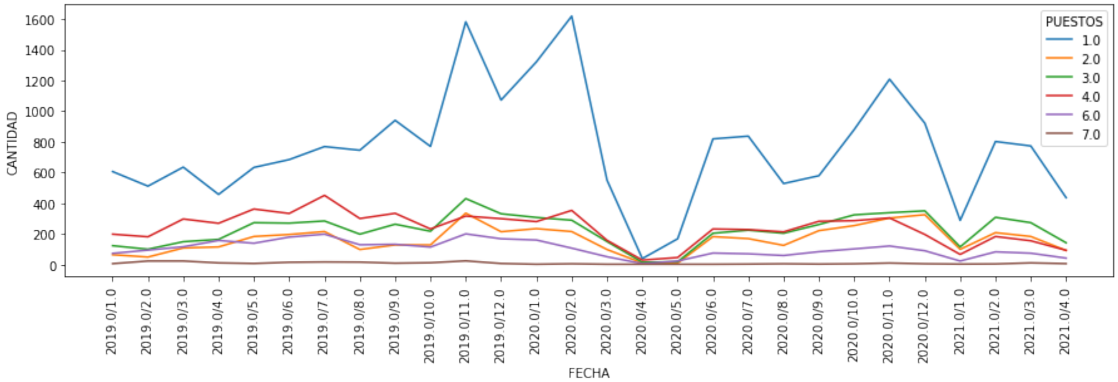
Next we have the historical behavior of the finishing materials of the different units. In here the range is really broad with `soft textures` having sold units over the 600 average, and `metal` texture bellow the 10 average sold units.

Sales by "acabado"



Finally we have the historical behavior of the `number of seats` feature with a similar behavior as the other features. Although they have diferent sale amounts on every posible value, that is explained by the composition of the portfolio and not attributable to a certain relation between a specific month and the values of the feature. No single value of any of these features has a month where it's behavior changes in a relevant manner to identify a substancial change in the analysis.

Sales by seats



Due to the previous findings, we opted for a different approach by evaluating a classification procedure to review possible patterns that cannot be easily identifiable by graphing the different numeric and categorical variables as we previously did. For this, we decided on a Demand Classification method.

2.6 Demand classification

The Forecast accuracy of any model strongly depends on your product forecastability, and we can find evidence of it in the demand history characteristics previously reviewed in the sales analysis section.

In this case then, to be able to determine a product forecastability, we use the demand calssification process. For this we apply two coefficients to the different products to check whether the product can be classified as high/low variable. These coeficients are:

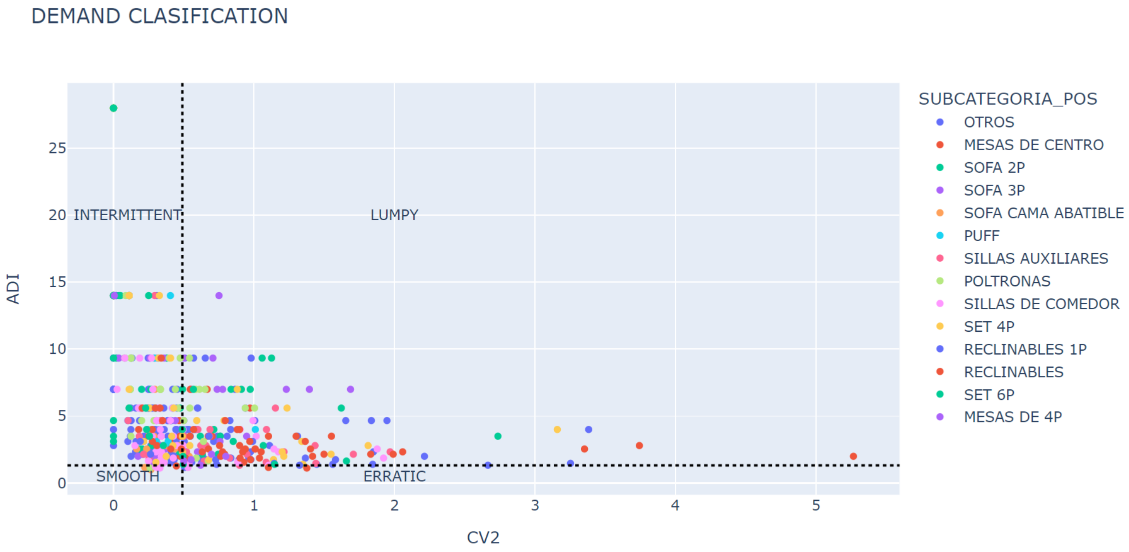
- the Average Demand Interval **(ADI)**. It measures the demand regularity in time by computing the average interval between two demands.
- the square of the Coefficient of Variation **(CV^2)**. It measures the variation in quantities.

Based on these 2 dimensions, the literature classifies the demand profiles into 4 different categories:

- **Smooth demand (ADI $<$ 1.32 and $CV^2 < 0.49$)**. The demand is very regular in time and in quantity. It is therefore easy to forecast and you won't have trouble reaching a low forecasting error level.
- **Intermittent demand (ADI $>=$ 1.32 and $CV^2 < 0.49$)**. The demand history shows very little variation in demand quantity but a high variation in the interval between two demands. Though specific forecasting methods tackle intermittent demands, the forecast error margin is considerably higher.
- **Erratic demand (ADI $<$ 1.32 and $CV^2 >= 0.49$)**. The demand has regular occurrences in time with high quantity variations. Your forecast accuracy remains shaky.
- **Lumpy demand (ADI $>=$ 1.32 and $CV^2 >= 0.49$)**. The demand is characterized by a large variation in quantity and in time. It is actually impossible to produce a reliable forecast, no matter which forecasting tools you use. This particular type of demand pattern is unforecastable.

Lets see the results applyed to our historical product sales on the 16 pareto subcategories previously chosen:

```
<Figure size 432x288 with 0 Axes>
```



In the graph we can visualice the deman calsification for Furni, where most of the products of the 16 pareto subcategories are clasified between intermitent and lumpy. This in turn indicates a patterns of high unstable variation of the amount of sales, and low and high variation on the interval frequency where the products are sold. With such informaton a simple objective to fulfill would be to migrate the behavior of the product sales to a more smoth clasification, with low variation in frecuency and amounts.

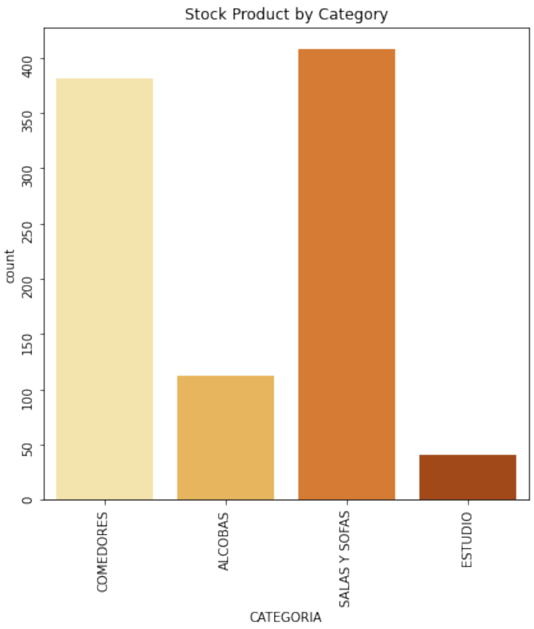
For the intemittent products we can evaluate possible strategies of periodicity and buy back. As the variability in amount is some how stable a stock buffer could be created on a stable demand basis. instead, for the lumpy products we could evaluate the portfolio clasification and the validity of the products to determine if some products could be removed from the portfolio or if they should sell on an on-demand or seasonal basis.

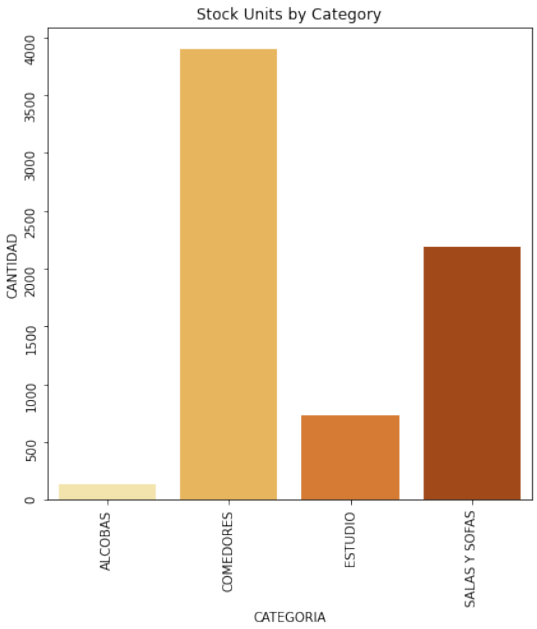
3 Stock sales comparative analysis

Part of the analysis finally relates to the clients' requirement to predict the demand of products to enable a better stock composition and reduce costs by optimizing stock management. To relate the previously gathered information with the way the stock should behave and be composed monthly, we also require a compared analysis between the historical sales and the current stock.

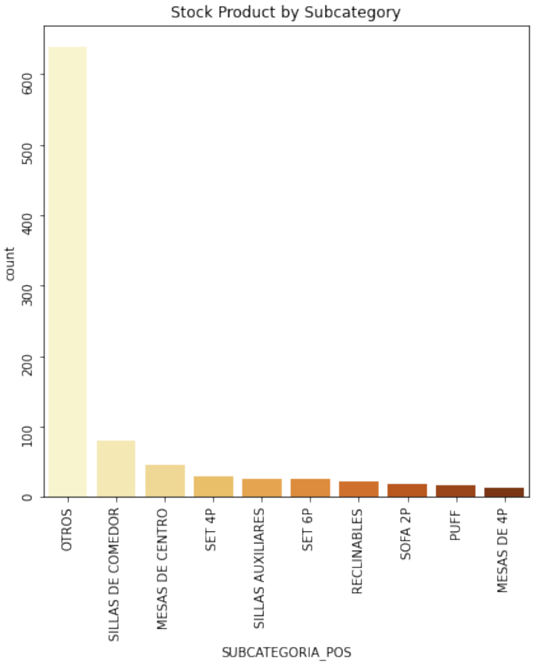
For this, we will evaluate the behavior of sales versus the current inventory of the company. In there we will try to identify if the inventory is consistent with the composition of the sales for these characteristics. The idea is to identify which features and characteristics are understocked, overstocked, or on point. And review if some of those results are due to the variation in demand over time.

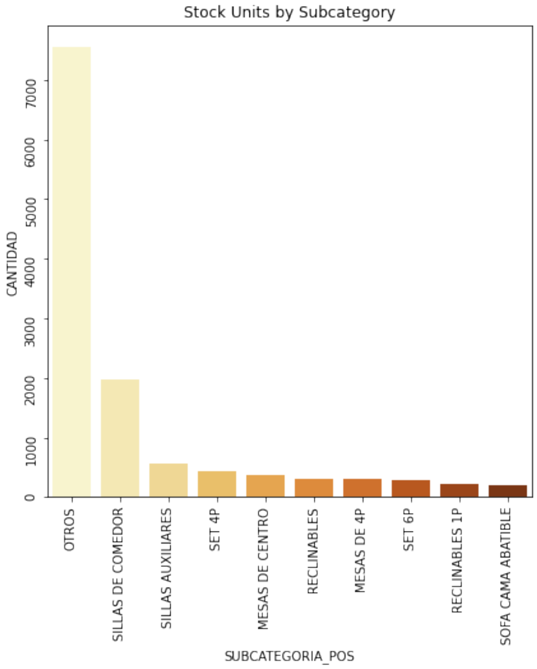
1. Category



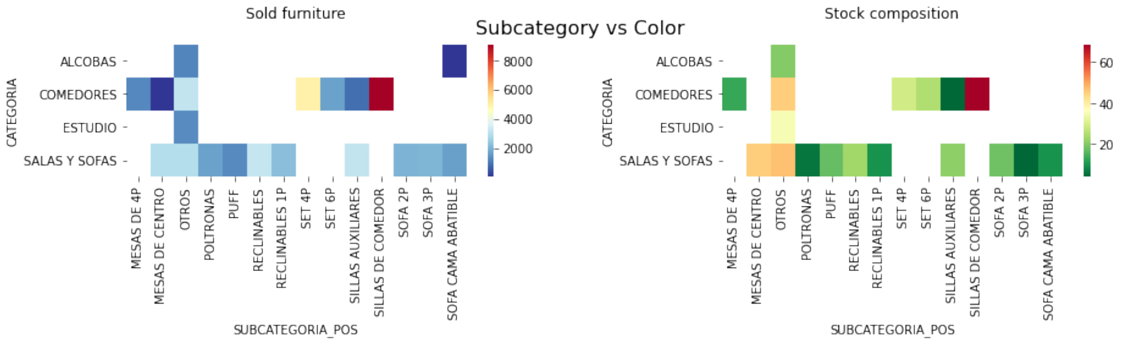


2. Subcategory

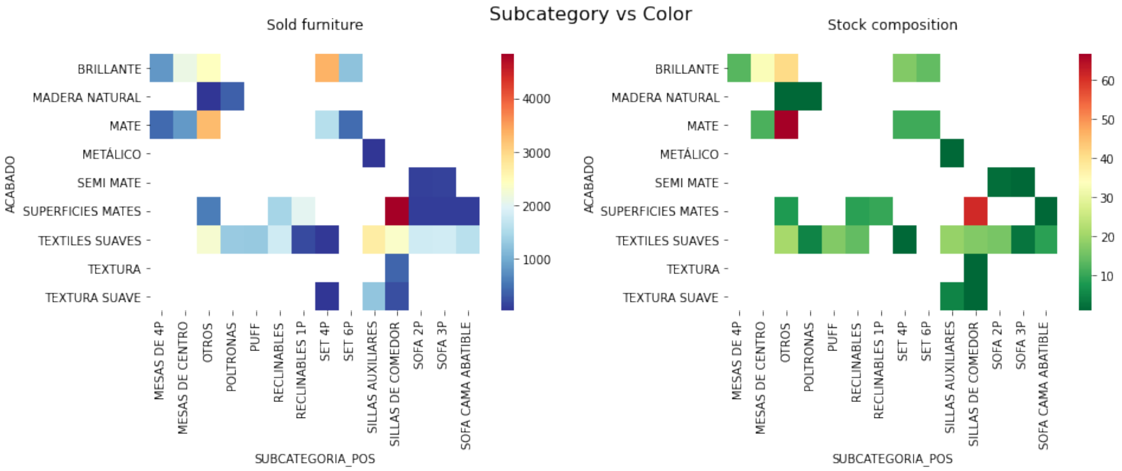


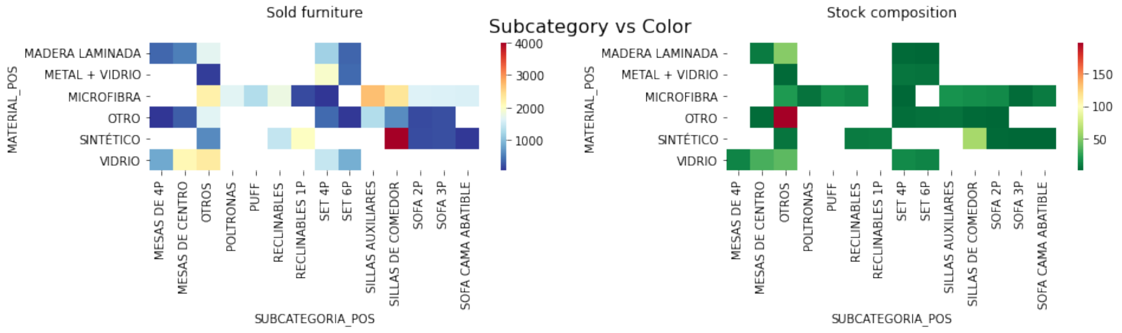


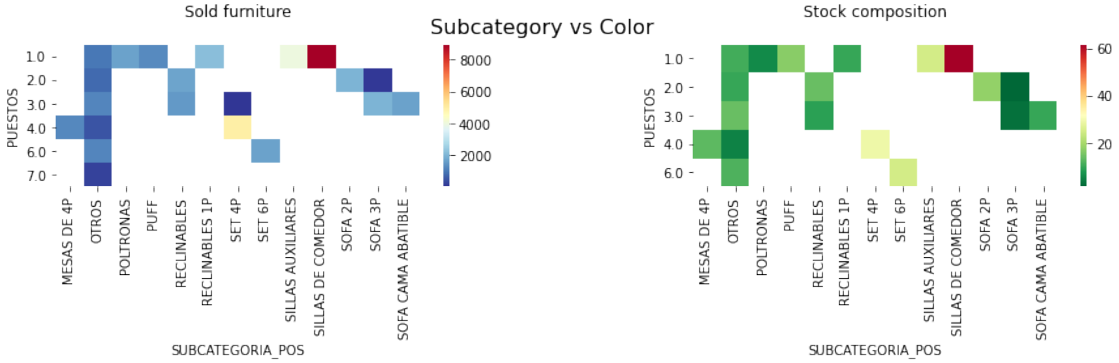
3. Frequency Tables Category|Subcategory

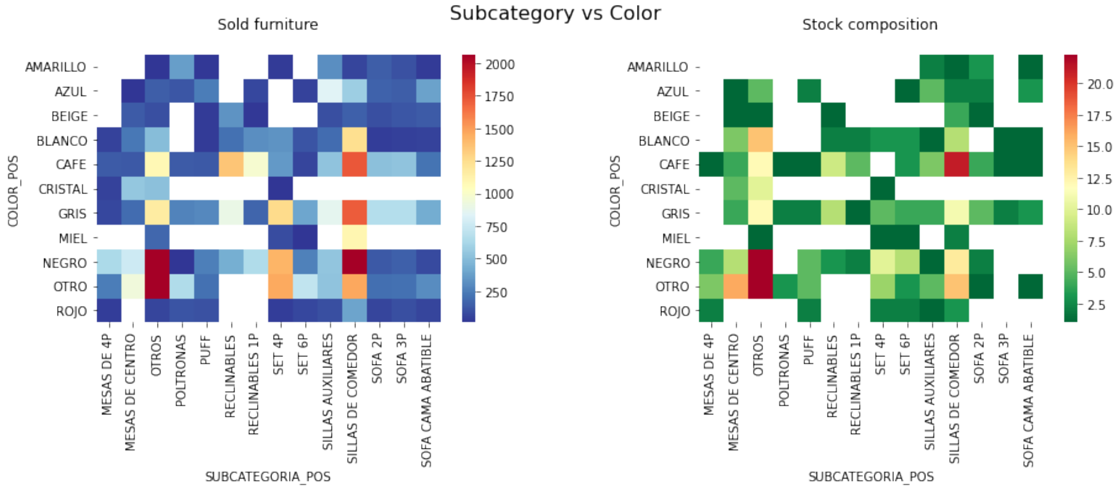


4. Subsubcategories_list vs other variables.









4 Conclusions

During this EDA analysis we explored the data given by the client, we created extra tables by merging the information we needed from the initial tables, we then cleaned the information to have a standardized format for all tables and values and then started the analysis to find additional information on the data to be able to gather insights on the problem at hand.

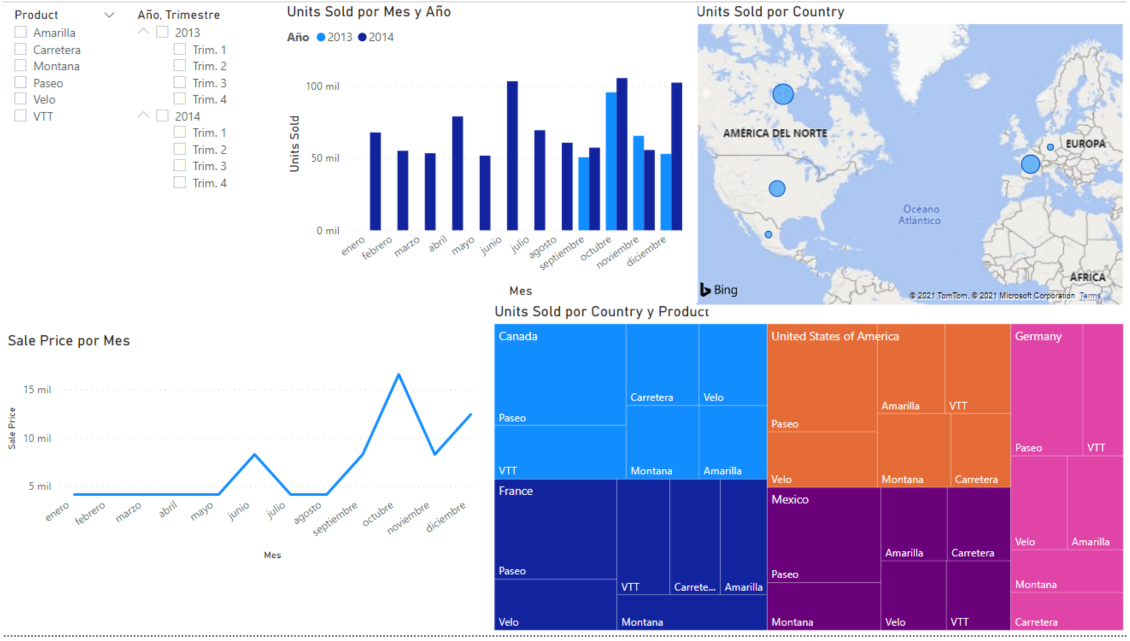
From the information, we plotted several line plots, scatter plots, and heatmaps to identify possible patterns and relations of the different variables, categorical or numerical. Some of that related information between the historical sales and the portfolio, and, the historical sales and the stock. From there we identified that there is a close relationship between the sales and the portfolio due mainly to the diversity of the products offered by the company. To higher diversity in colors, materials, texture, number of seats and category; higher number of sales of said subcategories (positive correlation). Finally, we determined the relevant features to take into consideration for the predictive model to implement.

From there, we decided to review possible trends of the relevant features in the historical sales, to check whether or not the different characteristics had a particular effect on the total amount of sales of the Pareto subcategories that comprised 80% of the sold portfolio. We grouped sales by different variables with no clear patterns observed. On the contrary, the only new information gathered was the relation between the number of sales by characteristic and the portfolio: to higher diversity of items for each feature the higher amount of sales. To improve this information we decided to reduce the dimensionality of several variables by conducting a Pareto test according to the influence on the sales. Similar results were observed thus confirming the previous results and excluding possible confounding variables on the process, such as dimensions -though they do affect the price average of the products-.

We then identified the possibility to generate a classification strategy on the portfolio concerning the sales behavior between sales frequency and sale quantity fluctuation. We decided to go on with the Demand Driven methodology to inspect whether the products behave on a certain pattern -Intermittent, Lumpy, Smooth, Erratic- or not. From those results, we hope to properly classify the portfolio and be able to generate marketing and stocking strategies. We also identified possibilities to improve the validity of the portfolio by reviewing the composition of discontinued and new products and it's relation to the whole portfolio.

Therefore, we expect to have created a stable base to correctly predict the demand and optimize the stock of the client.

5 Mockup frontend


Product
□ Amarilla
□ Carretera
□ Montana
□ Paseo
□ Velo
□ VTT
Año, Trimestre
□ 2013
□ Trim. 1
□ Trim. 2
□ Trim. 3
□ Trim. 4
□ 2014
□ Trim. 1
□ Trim. 2
□ Trim. 3
□ Trim. 4
Units Sold por Mes y Año
Año ● 2013 ● 2014
100 mil
50 mil
0 mil
Units Sold
enero
febrero
marzo
abril
mayo
junio
julio
agosto
septiembre
octubre
noviembre
diciembre
Mes
Units Sold por Country
AMÉRICA DEL NORTE
EUROPA
Océano Atlántico
ÁFRICA
Bing
© 2021 TomTom, © 2021 Microsoft Corporation Terms
Units Sold por Country y Product
Canada
Carretera
Velo
Paseo
VTT
Montana
Amarilla
United States of America
Amarilla
VTT
Paseo
Velo
Montana
Carretera
Germany
Paseo
VTT
Velo
Amarilla
Montana
Carretera
France
Paseo
VTT
Carrete...
Amarilla
Velo
Montana
Mexico
Amarilla
Carretera
Paseo
Montana
Velo
VTT
Sale Price por Mes
15 mil
10 mil
5 mil
Sale Price
enero
febrero
marzo
abril
mayo
junio
julio
agosto
septiembre
octubre
noviembre
diciembre
Mes