Business Intelligence Data Challenge

Author: Bruno Pereira Barella

Exploratory Data Analysis for Conversions

We have two sets of data, one for conversion occurrences and another for channels that the user came to the site at another time. Thus, the objective of this analysis is to understand the behavior of users, stimulus channels and movements in the period from 2017-03-01 to 2018-03-26.

Preparation of the Development Environment

To start the analysis, it is necessary to prepare the development environment with the necessary packages for the execution of all the steps.

For this, we used a virtual environment (venv) with Python 3.8.10 and the description of the packages can be found in the requirements.txt file.

Analysis Steps

Initially, the consistency of data in the tables, such as types, number of nulls and descriptive statistics, was verified.

Then, the Revenue was evaluated over time, seeking to identify interesting patterns or explainable anomalies. The behavior of the time series was also analyzed, seeking to identify periods with trends to be analyzed. And finally, it verified the existence of outliers and the distribution of the Revenue.

Subsequently, the behavior of customers over the period was analyzed in order to identify their behavior and relate Revenue movements. The customer retention rate was also evaluated over the months, seeking to assess what percentage of customers make purchases for more than 1 month.

The channels of origin of the transactions were also analyzed, seeking to relate them to the impacts on customers and Revenue. For this, the channels with the most occurrences (sum of the percentages) were evaluated over time and in a macro way for the entire analyzed period.

The relationship between the variables was also evaluated in order to identify the channels that motivate the increase in Revenue and/or customers.

Finally, a customer segmentation was developed with the intention of creating groups with similar behaviors, allowing to carry out more punctual future analyses, create specific tools to increase conversions and among others.

Data Consistency

The inference of data types performed by pandas performed correctly, identifying all types of variables in the database.

The number of nulls was another point considered in the analysis, where you can see that there is about 2.88% missing data for `User_ID`. For users without identification, we will assign the value of `unidentified` to not lose data. We have 0.0352% conversions without the `IHC_Conv` values, that is, there was a conversion, but the model did not report the channel percentage. We can delete them due to their low percentage.

As for the descriptive analysis of the data columns was correct, values of at most 1 for each channel.

Check Revenue in Time

In order to analyze the Revenue in time, some auxiliary variables were created on the date column, in order to allow the analysis in daily, monthly, weekly scales.

The first Revenue analysis was performed on a daily scale, as can be seen in Figure 1.

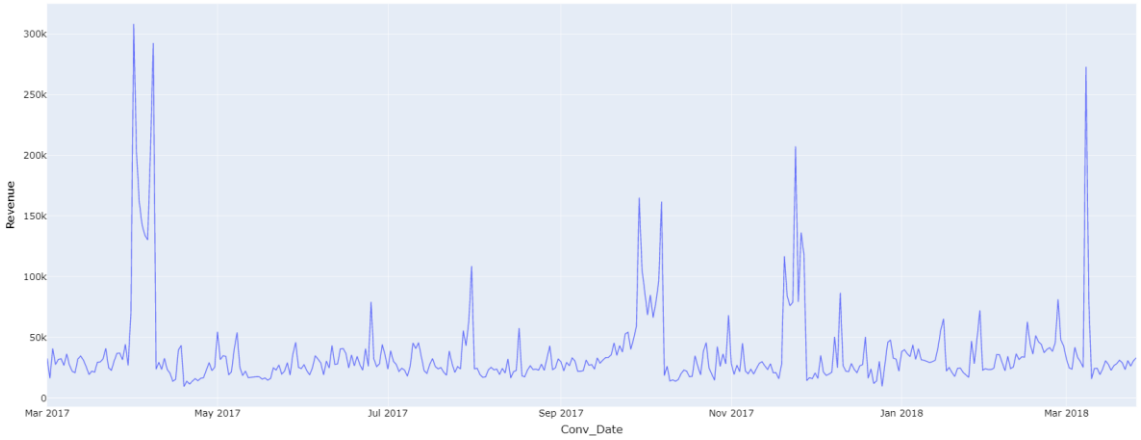

Figure 1 - Time series of Revenue daily

In figure 1 it is possible to identify anomalous peaks of Revenue, these are important points as they can be the effect of advertisements, events, promotions and among other tools. It would be interesting to evaluate these points with such possible events. If there are no justifications and come from few customers, it is interesting to evaluate their removal.

To analyze the behavior of the time series we can use a decomposition method that is used to isolate the trend, seasonality and noise in each period. We can combine this analysis with channel, in order to identify which channel impacted the increase or decrease in trends. The figure 2 shows this analysis.

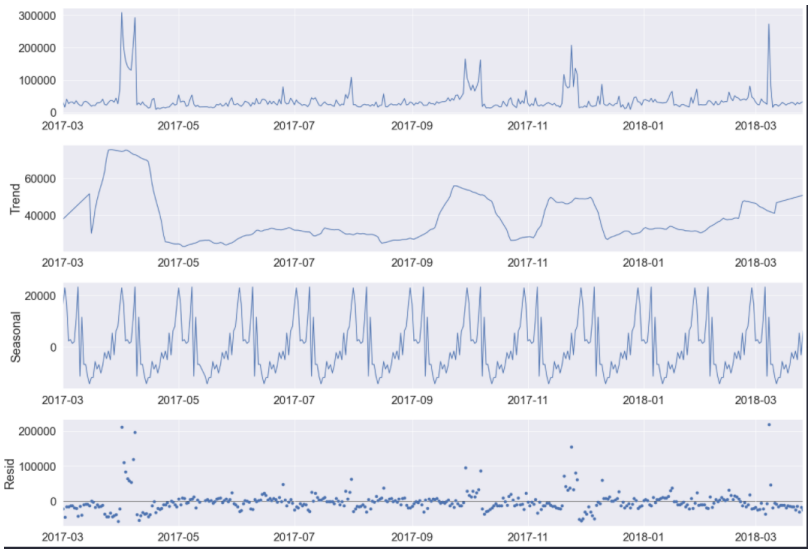

Figure 2 - Trend, seasonality and noise im Revenue daily

As can be seen in figure 2 a slight upward trend starting from 09/2017. This is a point to consider in future analyzes to identify what motivated this trend. Another point to be highlighted is that for the analyzed period the series does not present many patterns, this is observed in the large residues found.

For the evaluation of outliers, the anomaly identification method was used. For this, the Pycaret package was used together with the Isolation Forest model. Figure 3 presents the identified anomalies.

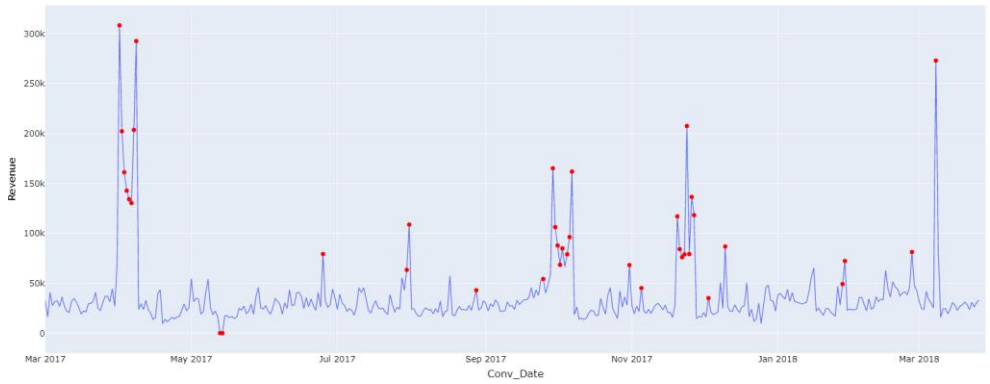

Figure 3 – Anomalies in Revenue

To consider the deletion of anomalies, it is important to confirm whether they are anomalous points or if there is any reason for these expressive movements. That way they weren't eliminated.

User Analysis

In the total database there are 55333 unique customers and their behavior over time is like Revenue. This is an important aspect that implies that the increase in Revenue was impacted by the increase in customers.

Another analysis performed was the recurring customer by cohort analysis, which seeks to identify the customer retention rate in the analyzed months. Figure 4 presents this analysis.

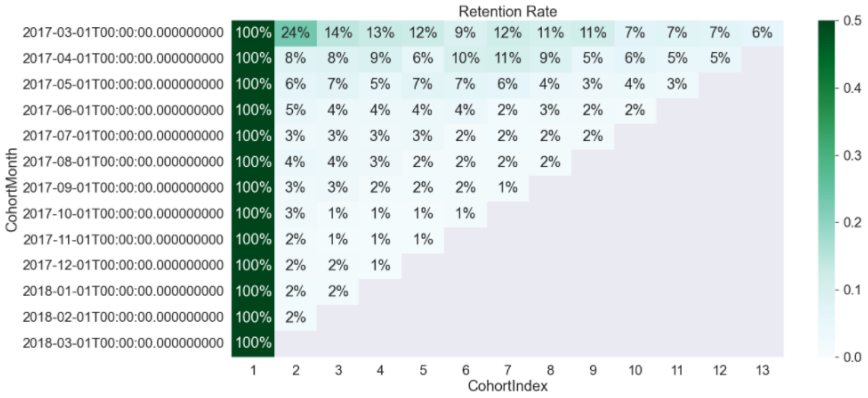

Figure 4 - Retention rate

It is noted that customer retention rates drop as the months go by. For example, from month 03/2017 to month 04/2017 only 24% of customers repurchase and this rate continues to drop. It is important to work in this area seeking customer loyalty, ensuring good profitability over time.

Channel Analysis

For the channels, the daily period and the sum of the IHC_Conv for each day were considered. It is possible to identify which channels were most used each day. Figure 5 presents this distribution.

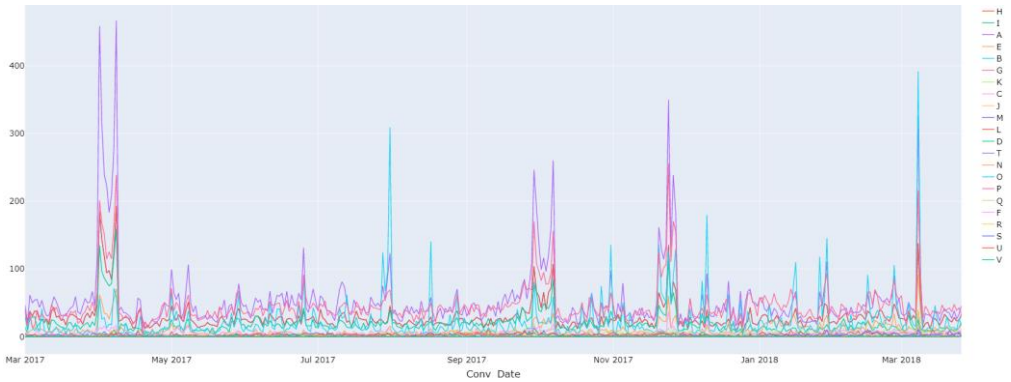

Figure 5 - Daily channels impact

In Figure 5 it is possible to identify which channels were most used over time. Points to be highlighted is that for the periods where there were possible anomalies there is also a significant increase in some channels. For example, the peaks in Revenue and User_ID quantity at the beginning of April were driven by channels A, G, H and I. Figure 6 shows a zoom in on this period.

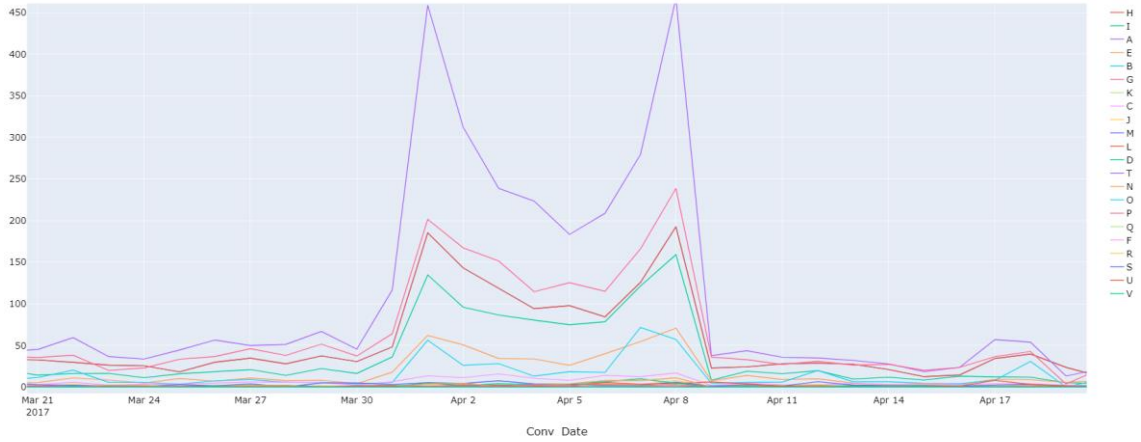

Figure 6 - Peaks in channels A, G, H and I

Relationship of Variables

To initially evaluate the relationships between the variables, Pearson's correlation method was used in the search to understand the motivators of growth or decline in Revenue and the number of customers. Figure 7 presents the results with the correlation values.

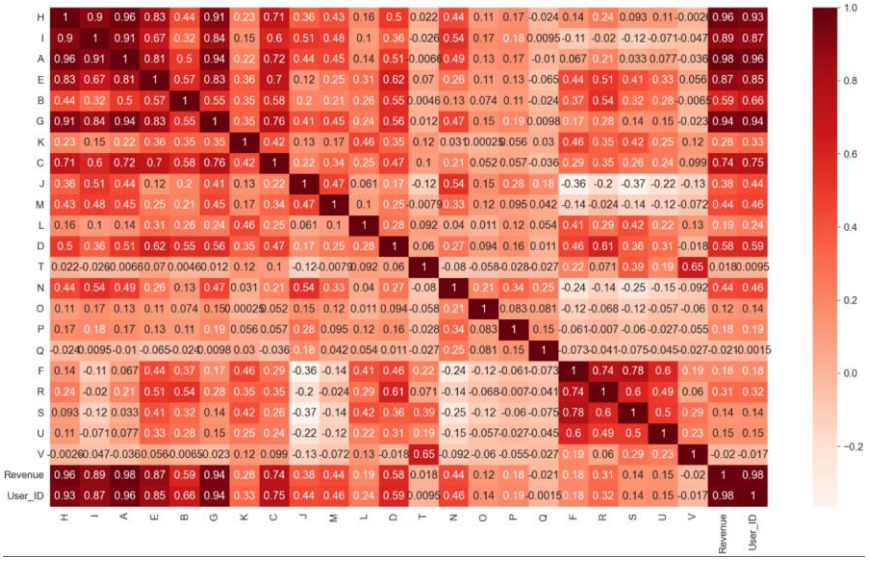

Figure 7 - Correlations in data

Through Figure 7, we can identify that the increase in Revenue and in the number of customers is explained by the increase in channels H, I, A, E, B and G. An interesting strategy is to evaluate sociodemographic variables in order to identify patterns and increase conversions on the other channels.

Customer Segmentation

Customer segmentation is important for businesses to understand their target audience. Different advertisements can be curated and sent to different audience segments based on their demographic profile, interests, and affluence level.

There are many unsupervised machine learning algorithms that can help companies identify their user base and create consumer segments. We will be looking at a popular unsupervised learning technique called K-Means clustering. This algorithm can take in unlabeled customer data and assign each data point to clusters.

The goal of K-Means is to group all the data available into non-overlapping sub-groups that are distinct from each other.

The elbow method was used to select the number of clusters. When executing the method, we obtained a value of 9 clusters. To assess the quality of the groups created, we used silhouette coefficient, or a silhouette score is a metric used to evaluate the quality of clusters created by the algorithm. Silhouette scores range from -1 to +1. The higher the silhouette score, the better the model. The silhouette score measures the distance between all the data points within the same cluster. The lower this distance, the better the silhouette score. This grouping resulted for silhouette coefficient is 0.42.

To try to increase the similarity of the groups, the methodology was used Principal Component Analysis (PCA) that helps us reduce the dimension of a dataset. When we run PCA on a data frame, new components are created. These components explain the maximum variance in the model. And then we can create other clusters with more separate variables and with more aggregated information.

With this change, he obtained a quantity of 10 clusters and a result for silhouette coefficient of 0.45. To continue the segmentation evaluation, it is important to interpret the groups created looking for relevant insights for the business. Table 1 presents the sums and averages for each group found.

	cluster	Revenue		count_purchase		A		G		H		I		B		E		
		mean	sum	mean	sum	mean	sum	mean	sum	mean	sum	mean	sum	mean	sum	mean	sum	count
0	0	183.3048	1262053.6213	1.0983	7562	0.0299	205.9077	0.0122	84.1697	0.8239	5672.4189	0.0666	458.4489	0.0090	61.9965	0.0085	58.7217	6885
1	1	185.8162	1918923.8480	1.0827	11181	0.0113	116.9928	0.8577	8857.4811	0.0077	79.1605	0.0410	423.2533	0.0082	84.6535	0.0226	232.8916	10327
2	2	205.7172	2156738.7512	1.1465	12020	0.8198	8594.8486	0.0246	257.8763	0.0210	220.0334	0.0443	464.9177	0.0206	216.2391	0.0202	212.1982	10484
3	3	888.1653	3086374.2657	4.2098	14629	0.3366	1169.6238	0.2036	707.4933	0.1250	434.4556	0.0888	308.5304	0.1328	461.3621	0.0475	165.1994	3475
4	4	151.0347	2718.6240	1.0556	19	0.0238	0.4280	0.0000	0.0000	0.0537	0.9665	0.0246	0.4422	0.0000	0.0000	0.0000	0.0000	18
5	5	184.7718	1398352.6579	1.1132	8425	0.0789	597.1748	0.0446	337.3128	0.0313	236.9469	0.0179	135.8245	0.6032	4564.6678	0.0264	200.0111	7568
6	6	29117.3031	29117.3031	111.0000	111	0.9027	0.9027	0.0136	0.0136	0.0034	0.0034	0.0092	0.0092	0.0285	0.0285	0.0267	0.0267	1
7	7	184.7943	2907738.1416	1.1273	17738	0.1480	2328.6697	0.1730	2721.7834	0.1013	1594.5857	0.2372	3731.6583	0.0255	401.5876	0.1240	1951.4771	15735
8	8	173.9276	52004.3556	1.1037	330	0.0521	15.5924	0.0544	16.2786	0.0195	5.8208	0.0372	11.1099	0.0228	6.8197	0.0097	2.9046	299
9	9	2341.5920	1264459.6773	9.8222	5304	0.3956	213.6107	0.1734	93.6413	0.0880	47.5411	0.0721	38.9537	0.1654	89.3314	0.0412	22.2727	540

table 1 - Interpretation of segmentation

Here we can see that cluster 7 has the highest number of customers and the most stimulated channels on average were I, G and A respectively. We have cluster 6, which has the highest average Revenue, the highest number of average purchases in the period, but only 1 customer and the most stimulated channels is A. We also have cluster 3 with the second highest average purchases.

It is important to verify that better results can be obtained using other algorithms, adding new variables for the development of segmentation.