SOLVOYO CASE STUDY

CUSTOMER SUCCESS CONSULTANT

Name: CAN UMUR AKMAN

27.05.2024

Preliminary Observations

- Data is for the year 2021.
- Open date ranges from year 2006 to 2021.
- Close date columns include many null entries. In fact, 99% of the entries take NULL value.
- Revenue ranges from 243,876 to 0.
- Average price ranges from 0 to 45,5.
- Number of stores are 304. For a histogram, it would be suitable to use 5 to 10 many bins (one-dimensional clusters), grouped by their revenue or average price or sales quantity or option capacity or date of order. To incorporate all of these, we need higher dimensions.
- Option Capacities are exceeded numerous times. These surpluses are substantial as well.

The total revenues for stores are clustered around 100 million to 200 million. The corresponding frequency histogram looks like a normal distribution with a small variance, if we are to think that the bin ranges are sufficiently sensitive for nuances in the levels of the total revenues. This is one of the assumptions I make: Grouping total revenues that are within an interval of 50 million $ is an appropriate approach.

Further, I took the average of the average prices for each store and plotted a frequency histogram for these prices. Mostly, the average price for each store cluster within the interval [41, 54]. This interval encapsulates 266 values, whereas the rest of the intervals only encapsulate 34 values. I believe the deviations in average prices across different stores are not very high.

Methodology Selection & Further Analysis of Data

I plan on clustering by stores according to the average additional sales that exceed the capacity of each store. Also, as additional dimensions, I will incorporate Sales Quantity, and either Revenue or Average Price. I will use only one of them since feeding both into a model would mean feeding the same information twice (given that Revenue = Average Price x Sales Quantity, and that we have already included Sales Quantity). There may exist time or seasonality effects, thus another dimension may be included. To decide, further analysis of data is required.

I want to see whether there is correlation between quantity ordered and prices. The reason is that SalesQty takes values from a wide range of numbers. I want to first observe/compute this for all of the stores together, and then for only specific stores in order to purify “store-effects” if there exists any. If it exists, then we may want to increase the number of clusters.

- The correlation for the formal (general case) came out to be 0,12.
- The correlation for the two stores (91st and 481st) that receive orders most frequently (i.e., most number of entries), came out to be -0,18 and -0,26 respectively. Just to make sure, I also computed the correlation

for the 139th store (51 observations), which came out to be -0,20. I assume that these 3 stores fairly represent all of the stores.

- The difference from the general case implies that "pooled" store data is not a good representative of the individual stores, and thus should be ignored. The general case ignores the time/date of the order as well, which makes it even less comprehensive.

Taking the store-specific correlations into account, this shows that there is little to no negative correlation between Sales Quantity and Average Price for all of our stores. My interpretation is that the **price elasticity of quantity demanded (i.e., Sales Quantity) is low**. In other words, the demand is not very sensitive to price changes of the products. However, this must be further analyzed, as correlation only captures linear relationships, and does not indicate any direction of causality. To check causality -even though it is very difficult to confidently argue causality due to imperfections of analyses-, regression analysis can be done. Moreover, to observe whether there is any nonlinear relationship, nonlinear methods should be employed.

Moreover, seasonality effects may exist. I analyze whether there is quarterly seasonality, since many products have seasonality in that manner (e.g., swimsuits, ice cream, boots, etc.). I will compute the quarterly AvgPrice and SalesQty for the afore-mentioned 3 stores.

- As can be seen in the "Correlation Seasons" Sheet, prices dramatically increase in the 4th quarter of the year 2021. Furthermore, the sales quantities are higher in Q2 and Q3 compared to Q1 and Q4.
- I believe that the Open Date of the store would not be very informative. On the other hand, location, district population of the location and other indicators may improve the clustering of the stores.

Considering all of the analysis above, I can suggest that the clustering can include the **Revenue** data instead of separately including Sales Quantity and Average Price due to price inelasticity of demand. In other words, dissecting Revenue into these two components would not add significant value, and also would redundantly increase the dimension of the clusters by one. Secondly, including the **average excesses** on option capacities for each store must be another important indicator (I am still not completely sure if this is the capacity of the store). Note that average excess being greater than 0 implies capacity was NOT exceeded and there is empty storage whereas lower than 0 implies the Option Capacity is exceeded due to overflow of sales/demand (I assume these sales are realized by ordering the products from nearby stores?). Thirdly, the clusters may take into account the **quarter** of the orders so as to capture seasonality.

Clustering Model

The clustering of 298 stores, for the year of 2021, according to 1) Average Revenue, 2) Average Excesses, and 3) Average Quarter of the Year (i.e., the fractional quarter value that receives orders most frequently). The third one is

included with the aim of observing specifically when a store is visited more frequently, which may account for its higher/lower Average Revenue or Average Excesses.

I employ centroid-based clustering, i.e., K-Means Algorithm. The biggest downside of this type of clustering is that the number of clusters are not decided by the algorithm and thus is a parameter. To pick the right one(s), I employ Silhouette Analysis on Python. K=3 has the highest score, and K=5 has the third highest score. I implemented the algorithm for both of the scenarios. The cluster assignments can be found in the folders created for each of these scenarios.

The plots show how the clustering is done as per the "Average Revenue" and "Average Excess" features. Bearing in mind that these plots do not provide information on how "Average Quarter" feature is utilized in the clustering, I think it would be wrong to say K=3 is more suitable than K=5 for the clustering of these stores. Further analysis such as computation of other clustering evaluation metrics may be helpful on deciding the optimal number.

Policy Implications

For each cluster, a different set of policies can be implemented. For example, the number of staff, Option Capacity (surplus/shortage), pricing of products, and even sale quantities can be adjusted according to these results, leading to cost reduction, rise in customer satisfaction, higher revenue, and more utilization of resources (e.g., capacity of the stores). Further, the stores would be more prepared for the fluctuation in demand that occur due to the seasonality effects.

REFERENCES

Thompson, J. (2022, September 7). Choosing the right clustering algorithm for your dataset. KDnuggets. https://www.kdnuggets.com/2019/10/right-clustering-algorithm.html

Kmeans. scikit. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html