ISYE 6740 – Summer 2022
Final Project

Project by: Chinmay Pathare

Project Title: Optimization of 3PL Packaging Solutions Using Sales Data

Problem Statement:

Third party logistic (3PL) providers usually support their customers by providing optimal solutions for the warehousing of their inventory and deployment of it to requested site locations; saving the customer both time and money invested in managing inventory items. These services provided by 3PL often involve variable costs that depend on customer ordering habits, for example, the dunnage used in packaging, transportation costs of shipments sent per month, number of pallets required per month etc. While most of these costs can be back charged to the customer to avoid business losses, they end up increasing the customer costs which could potentially lead in market share loss in a competitive environment.

One such cost which is high and variable on customer orders is directly associated to packaging of outbound goods. In a parcel shipment environment, the total quantity of non-inventory items used is dependent on the type of goods the customer orders from their 3PL partners for each sales order. These costs are especially higher for customers with a Just in time business model of replenishment of their inventory.

This project aims to minimize the packaging costs associated with fulfilling customer sales order with Just in time business models by leveraging historic sales order data of their inventory.

Data Source:

The dataset format leveraged in this project was originally recorded in a 3PL providers customer sales order. However, since the original data is proprietary, the attributes that have been used in this project's dataset were kept similar but the numbers for each datapoint have been newly generated for showing proof of concept. The names of products, their product types, assigned product numbers and customer location sites to where the orders were deployed were also newly generated to replicate the format of the original dataset.

The mock dataset generated contains a business quarter's worth of data with 2000 datapoints. Each datapoint is a recorded line item from a sales order for the product that was ordered by the customer sites, the date it was ordered, total qty ordered and the category of the product type. The mock data is stored in a CSV file type and will be cleaned and preprocessed before using for analysis.

The following attributes were available to analyze the dataset:

Attribute	Information	Data Type	Data Format
Location	Customer site location where product was shipped	String	A, B, C
Product Type	The type of product deployed to customer site	String	Pen, Bag, Sofa
Product Level Identifier (PLI)	Each unique product is assigned an identifier called PLI in the system	String	123-123456
Ordering Qty	Total qty of product ordered for the given PLI	Int	5,25,35
Ordering Date	Date on which the sales order was placed	String	1/2/2021
Sales Order Number	Sales order number for each order placed	String	1234-567890

Methodology:

1) Data Cleaning

As mentioned earlier, the dataset was generated as a mock dataset using the attributes that would be similar to the information that a 3PL provider would collect during sales order placement in an ERP system. Since ERP generally stores data in a collective format under a single table, there usually are multiple lines for the same order with different order status like "Order Placed" , "Order Processing" , "Order Cancelled" and "Order Completed".

The first task would be to clean data during the query pull from the servers. Since our mock data was generated manually, this step was not required. However, while querying the data, the use case would define filters required to generate the data query. In this use case, the table would be queried to include all orders that with status of "Order Placed" and "Order Cancelled". Then resulting list would be grouped by the Sales order number (Unique Key) and any sales order number with more than one count would be dropped from the data to be used for analysis. This would clean any orders that were placed but were cancelled due to numerous unknown reasons. The rows would also be cleaned if there were any missing information for any of the columns as a result of a user input during entering a Sales order. This final list of cleaned data would then be used to conduct further analysis.

2) Grouping and Frequency of Orders

The cleaned data would then be further grouped. Since the companies for which this data was being collected used a Just in Time supply chain model, the ordering patterns could be irregular. What this indicates is that there could be multiple orders for the same product from the same site on any given day. To consolidate these instances, on the mock data, grouping was used to consolidate total quantity of a particular product that was ordered from a site on a particular day.

A few more decisions were made based on the usage of the data. The intention of this project was to leverage the available data to minimize the packaging costs associated with order fulfillment. This can be achieved from analyzing the data in two ways:

1. By identifying product types that usually get shipped together to a particular site and optimizing packaging solutions based on the product type grouping
2. By looking at Product level identifier (PLI), ordering qty, site and dates ordered and optimizing packaging solution based on the actual product

The packaging restriction that heavily impacts materials used for outbound shipments is the size of the products. From experience it is known that products with different PLI under the same Product Type have similar sizes. As a result, a decision was made to take the (1) approach for further analysis.

The resulting subset from data cleaning was hence further grouped to find the frequency at which a particular Product Type was ordered by a site. To achieve this, grouping was done to count the total number of weeks in a year where the product type was ordered by a site. A new attribute column "Total Weeks per Year" was introduced to the dataset to represent this number. Lastly, the frequency number was calculated using total number of weeks in a year when the product is ordered by a particular site divided by the total weeks in a year. This attribute was named as "Yearly Order Frequency" and added as an attribute.

The resulting table after making the decisions and groupings above was consolidated to 198 rows with following attributes:

Attribute	Information	Data Type	Data Format
location	Customer site location where product was shipped	String	A, B, C
prod_type	The type of product deployed to customer site	String	Pen, Bag, Sofa
PLI	Each unique product is assigned an identifier called PLI in the system	String	123-123456
tot weeks	Total weeks for when a product was ordered by a particular site	Int	5,25,35
YOF	Yearly Order Frequency	Float	0.0 to 1.0

3) Preprocessing Data

The intention of pre-processing this data was to be able to run an unsupervised learning algorithm on it. Since these are primarily mathematical models that work on numerical data types, it was important to convert all variables to meaningful numerical data.

Attribute *PLI* has a data format of 123-123456 stored as a string. *PLI* is a unique identifier of a product. To ensure it can be processed, the " – " between the strings was removed and then the entire string was converted to data type float (float64).

Attribute's *location* and *product_type* are stored as a string in the dataset as well. The *location* attribute indicates the location to which the *PLI* was being sent to and are unique for a particular site. The *product_type* however is not unique to a particular product *(PLI)* and can be shared by multiple products *(PLI)*. These variables were handled by using categorical encoding. The attributes were first converted to the pandas datatype category and then using the inbuilt feature within pandas converted to assigned numerical values using the Label Encoding method. Each of the data points were then converted to datatype float (float64). This completed the pre-processing of data and was ready to be used for further analysis.

4) Correlation Analysis

The total features available for training were only 5, however, to minimize any chances of overfitting the data, correlation analysis was conducted. The results from the analysis are as shown in Fig(1)

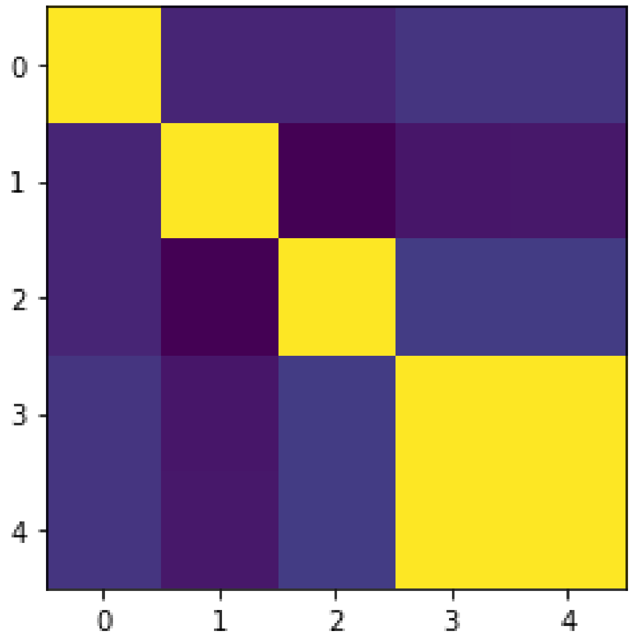


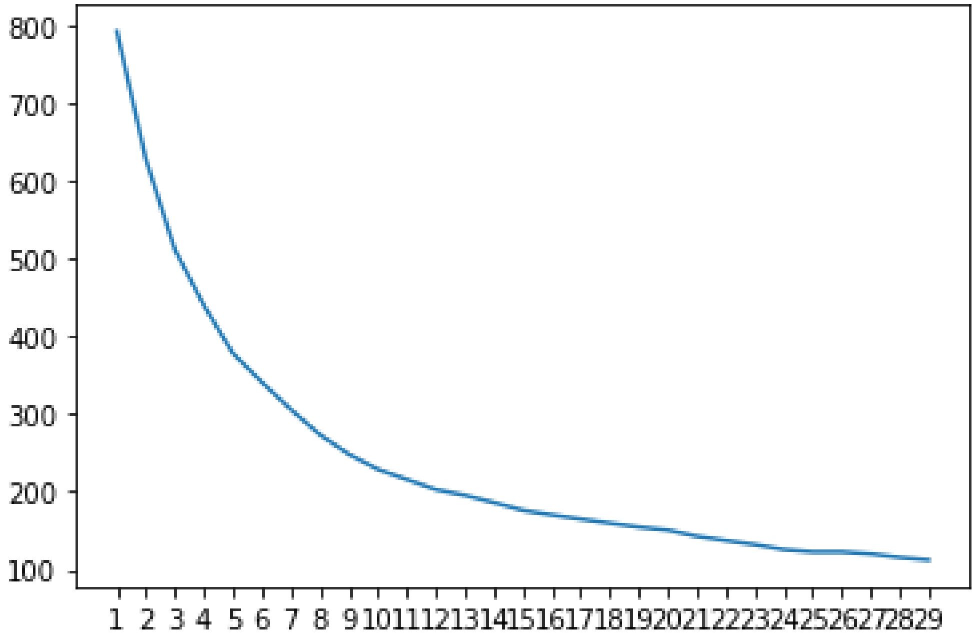
Fig (1): Correlation Analysis

The results indicate that there is a heavy correlation between variables *tot_weeks (3)* and *YOF (4)* . This was anticipated since the metric YOF was derived from a similar methodology to generate a frequency variable like the *tot_weeks*. As a result of this analysis, it was decided to remove *tot_weeks* from the dataset. This decision was made at random by removing one of the two heavily correlated variables.

5) Clustering

In order to find groupings of products that usually get shipped together, a meaningful pattern needs to be identified from the available data. Since there are no clear response variables that would help indicate as such, the need for an unsupervised clustering model was felt necessary. KMeans clustering algorithm was selected for this purpose since it is an iterative algorithm that tries to partition the dataset in k defined non-overlapping clusters. This helps ensure that product and site combinations are not mapped to two different clusters at the same time.

In order to find the best suitable k value, the elbow method was used to find the optimal number of clusters. In this method, the Kmeans algorithm is run over multiple iterations with different number of clusters. For each complete iteration, the Sum of Squared Errors (SSE) is calculated and recorded. A plot is then mapped with y-axis as SSE and x-axis as number of clusters. The resulting elbow plot was as shown in Fig(2).



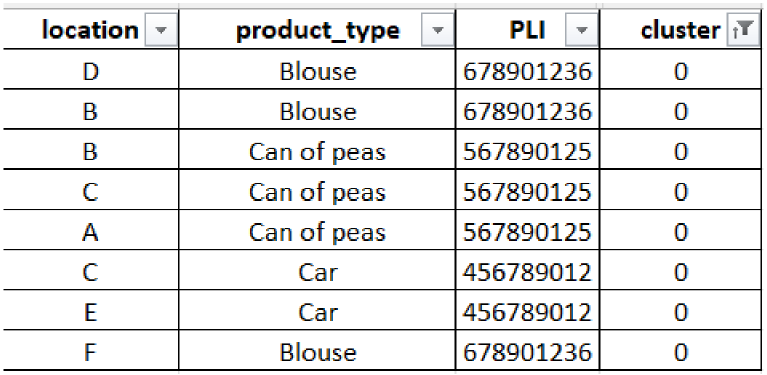
Fig(2): Elbow Plot

Based on the results, the slope of reduction in SSE evidently slows down roughly near 23 clusters. This indicates that the benefit to cost ratio in terms of continuing to add more clusters reduces at this point. As a result, a k value of 23 was considered for the final clustering run.

A new model was generated based on the k value of 23 and ran for a maximum of 300 iterations. This model was then fitted to the preprocessed dataset from earlier. The resulting clusters where then exported to an excel sheet to interpret information from them.

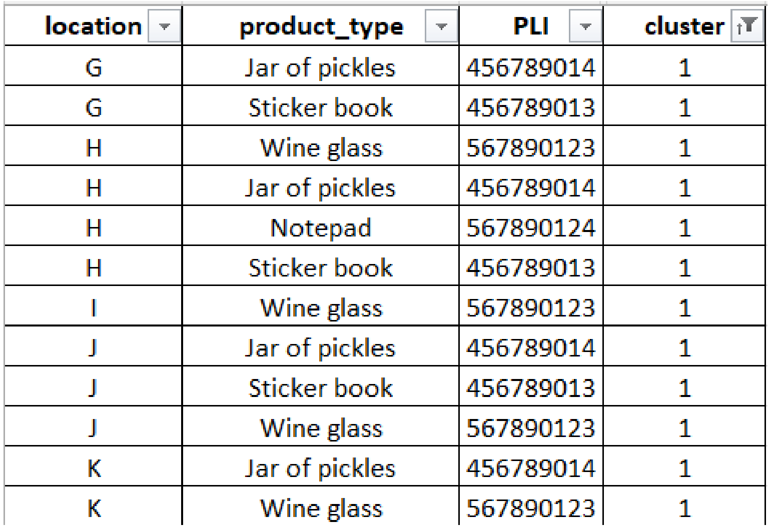
Results and Interpretation:

The resulting clusters from the implementation of the algorithm provide insights on how the products are usually ordered by customer sites. One way to interpret the results are by looking at the clusters themselves. Example outputs are as show in Fig (3) and Fig(4):



location	product_type	PLI	cluster
D	Blouse	678901236	0
B	Blouse	678901236	0
B	Can of peas	567890125	0
C	Can of peas	567890125	0
A	Can of peas	567890125	0
C	Car	456789012	0
E	Car	456789012	0
F	Blouse	678901236	0

Fig(3) : Cluster 0

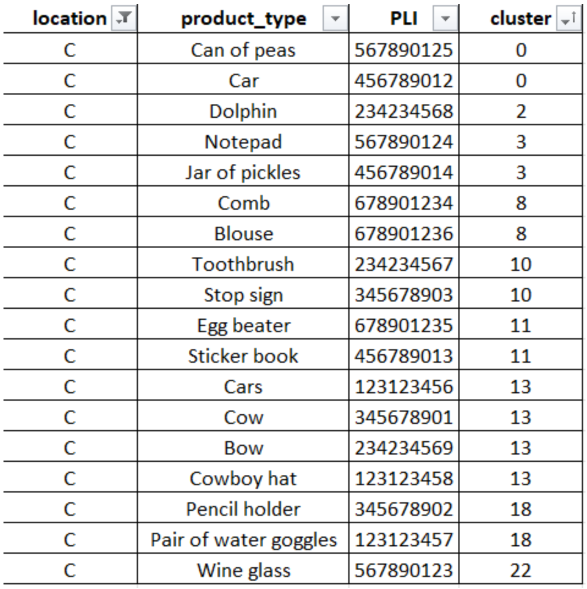


location	product_type	PLI	cluster
G	Jar of pickles	456789014	1
G	Sticker book	456789013	1
H	Wine glass	567890123	1
H	Jar of pickles	456789014	1
H	Notepad	567890124	1
H	Sticker book	456789013	1
I	Wine glass	567890123	1
J	Jar of pickles	456789014	1
J	Sticker book	456789013	1
J	Wine glass	567890123	1
K	Jar of pickles	456789014	1
K	Wine glass	567890123	1

Fig(4) : Cluster 1

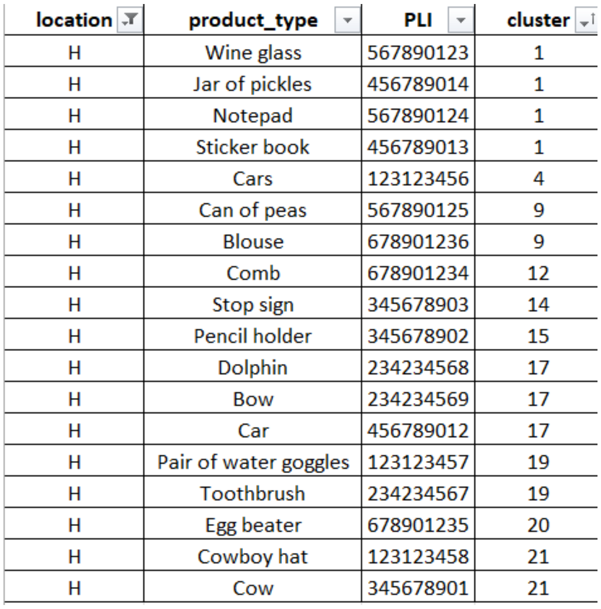
The products in cluster 0 and cluster 1 are usually ordered together based on the results data. However, these are not always going to the same site. This shows that the frequency of ordering influences the clustering and we have to be careful to use this for our packaging needs. Based on the PLI's for each of the product types being ordered in a cluster, it looks like sites tend to order similar products at a given time. This could be due to business consumption for the customers or seasonal requirements of these products. The analysis however, is not enough by itself to comment on either of the possibility. Based on the analysis it only indicates that these PLI's are usually ordered together from multiple locations.

Another way to interpret the results are by looking at a specific location and identifying clusters within that location. Examples as show in Fig(5) And Fig(6):



location	product_type	PLI	cluster
C	Can of peas	567890125	0
C	Car	456789012	0
C	Dolphin	234234568	2
C	Notepad	567890124	3
C	Jar of pickles	456789014	3
C	Comb	678901234	8
C	Blouse	678901236	8
C	Toothbrush	234234567	10
C	Stop sign	345678903	10
C	Egg beater	678901235	11
C	Sticker book	456789013	11
C	Cars	123123456	13
C	Cow	345678901	13
C	Bow	234234569	13
C	Cowboy hat	123123458	13
C	Pencil holder	345678902	18
C	Pair of water goggles	123123457	18
C	Wine glass	567890123	22

Fig(5) : Location C Clusters



location	product_type	PLI	cluster
H	Wine glass	567890123	1
H	Jar of pickles	456789014	1
H	Notepad	567890124	1
H	Sticker book	456789013	1
H	Cars	123123456	4
H	Can of peas	567890125	9
H	Blouse	678901236	9
H	Comb	678901234	12
H	Stop sign	345678903	14
H	Pencil holder	345678902	15
H	Dolphin	234234568	17
H	Bow	234234569	17
H	Car	456789012	17
H	Pair of water goggles	123123457	19
H	Toothbrush	234234567	19
H	Egg beater	678901235	20
H	Cowboy hat	123123458	21
H	Cow	345678901	21

Fig(6): Location H Clusters

Looking at the clustering results by sorting for a location type provides information for a particular site and their ordering patterns. In Fig(5), based on the results, it indicates that a Notepad and Jar of Pickles, as grouped together in cluster 3, are usually ordered together by location C. Similarly we can say that that a Pair of Water googles and Toothbrush are usually ordered together by location H. This provides insight on what products can be expected to be ordered by a location at time.

More insight can be obtained by using method 2 of interpretation for our use case. Using the clustered products for a particular location, we can estimate the required packaging materials for that group. An optimized packaging solution can then be created for that cluster specific to that location. This includes work on reducing the dunnage by creating customized box sizes and bulk payment for parcel Vs pallet shipment based on product groupings. The optimized outbound box sizes would also reduce total volume of packages and add to further cost savings. As a result of this optimized packaging, the company can be charged less for the services provided by the 3PL vendor.

Project Shortcomings and Potential Improvements:

The clustering based on initial grouping may result in a loss of critical data that could end up in creating wrong clusters. The detailed information for Date Time that was omitted as a part of the few decisions made early on could add value to the training parameters. Instead the date time parameter can be converted from string in a usable data format for further analysis. This could also be leveraged in analyzing any trends or seasonality's in the ordering using Holt-Winter's method.

The categorical encoding of the attributes using Label encoding method has potential drawbacks of creating hierarchical weights for each of the two attributes: *location* and *product type.* Where 3>2>1 could influence the clustering if the model interprets the weight as a decision parameter. Instead, One-hot encoding method might be more beneficial to remove this bias.

The overall accuracy of the model proposed is also dependent on the number of datapoints available. Since the mock data generated was a manual process, the resulting solutions were as comparable as the quality of the mock data which was randomized. In real datasets this would be much improved that the data would not entirely be as random.