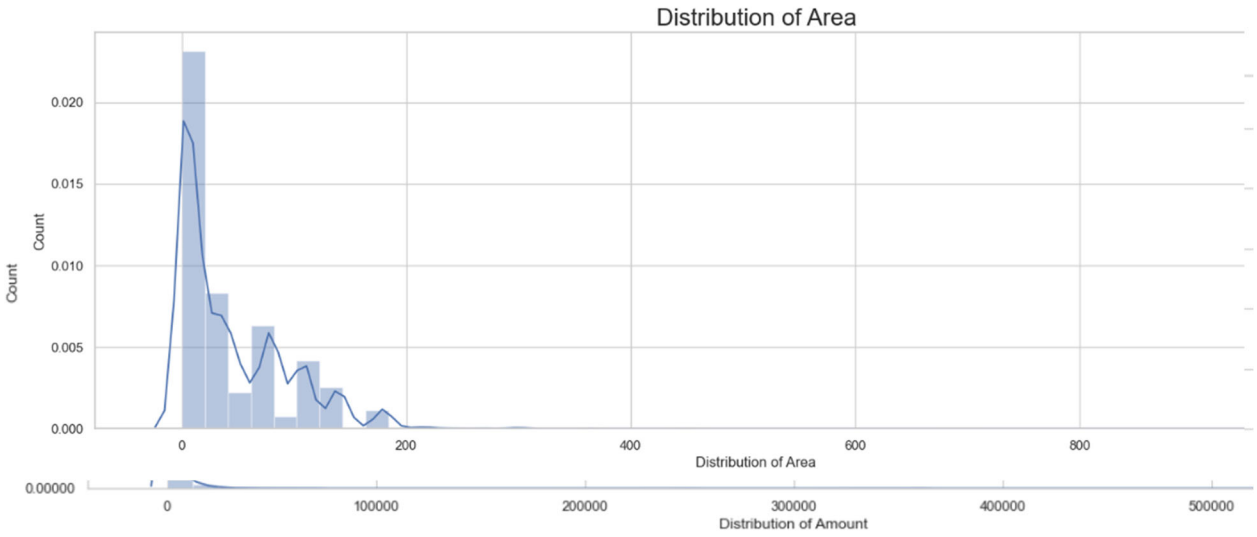
Part a: SalesData Exploration

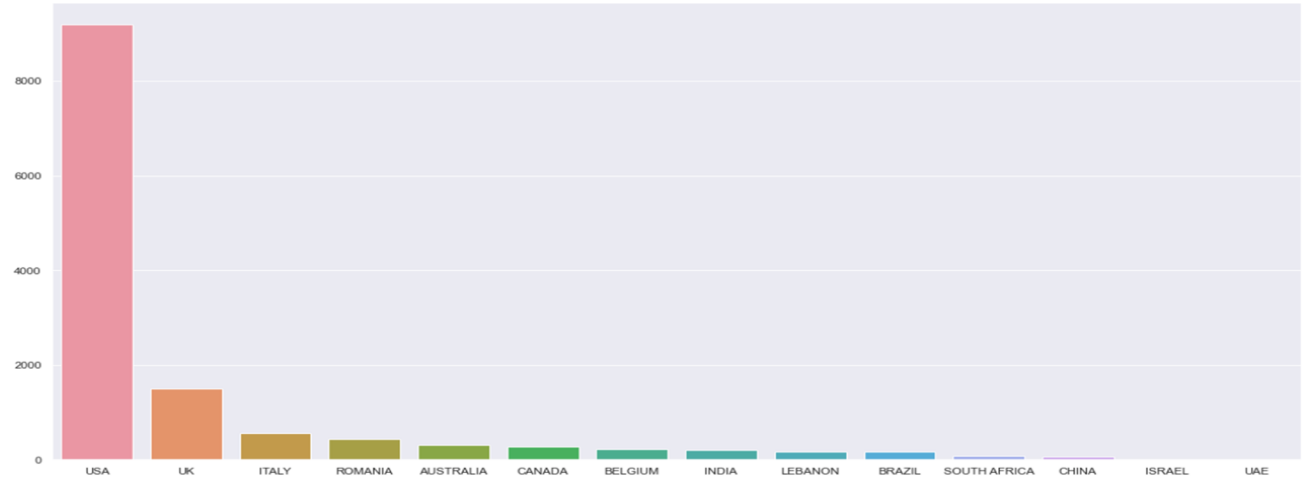
The sales dataset contains sample information of around 13,000 orders placed by about 45 different customers. Since Champo Carpets is currently facing the challenge of high sampling costs and low conversion rates, the approach we have taken to explore this dataset is to identify the key customers and most in-demand designs so that the company can focus its efforts in the most profitable buckets. That said, this dataset can be explored using multiple variables (such as 'TotalArea' and 'Amount') to see which customers are more important than others and which designs and colors of carpets are most popular. For starters, using the descriptive statistics we noticed the sales contribution of each order varies significantly:

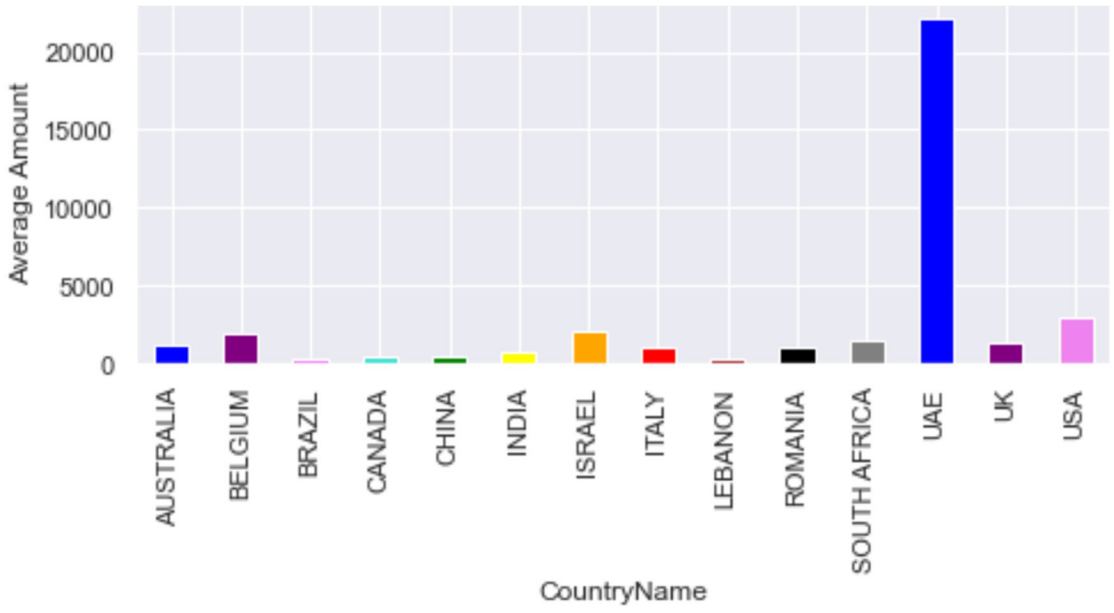
	QtyRequired	TotalArea	Amount	AreaFt	AreaMtr
count	13135.0000000	13135.0000000	13135.0000000	13135.0000000	13135.0000000
mean	44.4606014	44.7313699	2392.0401637	54.6223692	4.9517396
std	228.7495346	50.0995972	16832.0910796	49.0727674	4.4559331
min	1.0000000	0.0400000	0.0000000	0.4444000	0.0400000
25%	3.0000000	5.8000000	163.2000000	15.0000000	1.3500000
50%	8.0000000	24.0000000	590.6250000	40.0000000	3.6000000
75%	20.0000000	80.0000000	1540.0000000	80.0000000	7.2000000
max	6400.0000000	1024.0000000	599719.6800000	645.7222000	60.0000000

Then we looked at the distributions of the 'Total Area' and 'Amount' variables per order. This gave us insights into the spending patterns as well as the most popular carpet sizes.



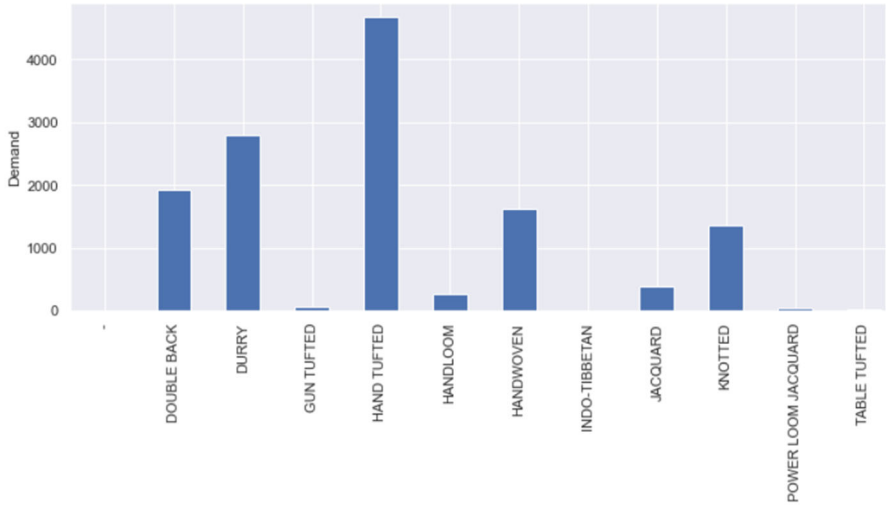
From the graphs above, we can see that the majority of orders require smaller carpets and multiple designs. This translates into high set-up cost for Champo Carpets. In order to improve the carpet sample conversion rate and reduce cost inefficiencies Champo Carpets should identify key customers - ones who order large quantities per order or those who contribute to the majority of revenue. According to the information provided in this dataset, the majority of the orders are coming for the USA customers. Whereas, the UAE customers are placing the largest orders.
The total of amount difference between the first highest and second highest is dramatic.We looked closely at the UK group, even through they only have two customers but they placed orders frequently which leads to highest amount.



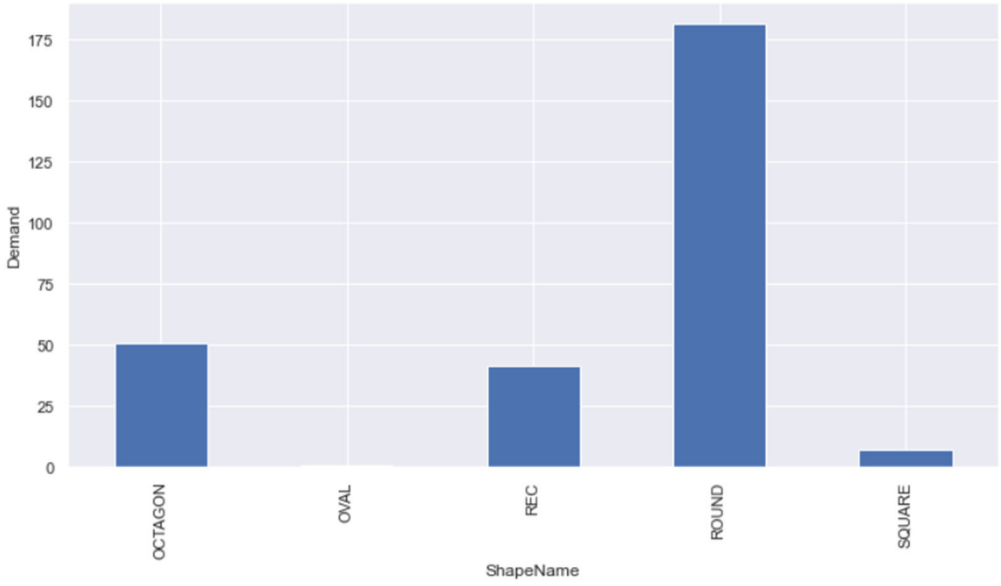


We also explored the most in-demand designs using raw material, carpet shape and color as the input variables. As can be seen in the charts below, Hand Tufted carpets are most popular. Additionally round shape and multi/gray/blue and silver colored carpets are preferred over the

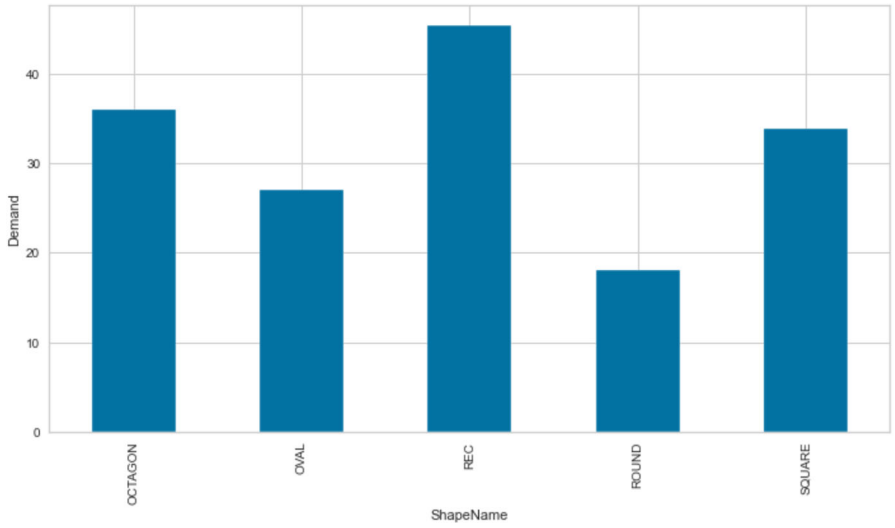
others. And interestingly the Round shape is popular one in terms of QtyRequired, but the REC shape has highest demand in terms of TotalArea.That said, the round shape carpet usually come with smaller total area per product.And the REC usually being produce larger area per product.



Demand of ShapeName in terms of QtyRequired:



Demand of ShapeName in terms of TotalArea:



These basic insights are very **general** and *might not be applicable to all customers*, but this analysis by and large gives some direction into customer preferences and might be helpful in reducing sample wastage and increasing conversion rate.

As a profit-making company it's also important for Champo Carpets to also know how much each of the popular items is making in revenue. The table below shows the ranking of order by Amount in terms of 'ITEM_NAME'.

ITEM_NAME	QtyRequired	TotalArea	Amount	AreaFt	AreaMtr	price_per_ft
HAND TUFTED	164288	195305.0495000	18504515.2700040	230565.7205000	20912.6474000	94.7467324
DURRY	319635	79552.2899000	5601134.4057920	114441.4166000	10401.2607000	70.4082109
KNOTTED	16459	87138.1218000	2196894.0736700	99814.9368000	9024.4551000	25.2116299
HANDWOVEN	39046	73885.4358000	1990592.7926000	90011.7700000	8146.7506000	26.9416127
DOUBLE BACK	18356	122285.1401000	1675588.3247500	140355.6842000	12712.8813000	13.7023053
HANDLOOM	8349	7901.1122000	746930.0835840	13574.8612000	1238.4006000	94.5348028
JACQUARD	4024	18594.1267000	461568.1300000	24451.9996000	2216.4842000	24.8233293
TABLE TUFTED	3564	729.0000000	132259.2000000	729.0000000	65.6100000	181.4255144
POWER LOOM JACQUARD	9753	481.8750000	84299.2500000	481.8750000	43.7112000	174.9400778
GUN TUFTED	366	1547.6721000	21591.4600000	2505.5485000	229.9838000	13.9509267
INDO-TIBBETAN	32	38.9800000	3571.1600000	382.7708000	35.3400000	91.6151873
-	118	87.7400000	503.4000000	149.2361000	13.5750000	5.7374060

The table below also shows the top 10 most important customers in terms of the amount of total area of carpet they ordered.

M-1 has the most purchase record, but it ranked fourth in terms of the total amount. We can observe that this customer prefers to purchase a lower average price per foot.

CustomerCode	QtyRequired	TotalArea	Amount	AreaFt	AreaMtr	avgprice_per_ft
M-1	16649	209725.2220000	1959794.0300200	209725.2220000	18907.1236000	9.3445796
P-5	48373	79666.7905000	3066518.2278920	79698.1405000	7183.5158000	38.4918008
C-1	5137	62763.0555000	567620.7209640	62763.0555000	5664.0957000	9.0438669
A-9	18923	53625.6544000	1592079.7900000	56266.7776000	5064.1216000	29.6887713
TGT	15045	37630.3318000	11341052.5133400	37630.3318000	3420.5622000	301.3806143
H-2	183206	19505.3958000	3804801.2957000	19505.3958000	1756.7906000	195.0640394
E-2	581	18878.0000000	116778.3000000	18878.0000000	1699.0200000	6.1859466
C-2	55172	9510.0000000	1557123.0000000	9510.0000000	855.9000000	163.7353312
H-1	1137	9327.0625000	65383.7950000	9327.0625000	841.6621000	7.0101165
T-5	42967	9221.3750000	733832.9500000	9221.3750000	831.8560000	79.5795584

The following table shows the top 10 consumers based on the amount of money they spent. TGT spent far more money than other consumers, and it seems that they prefer larger carpets.

CustomerCode	QtyRequired	TotalArea	Amount	AreaFt	AreaMtr	avgprice_per_ft
TGT	15045	37630.3318000	11341052.5133400	37630.3318000	3420.5622000	301.3806143
H-2	183206	19505.3958000	3804801.2957000	19505.3958000	1756.7906000	195.0640394
P-5	48373	79666.7905000	3066518.2278920	79698.1405000	7183.5158000	38.4918008
M-1	16649	209725.2220000	1959794.0300200	209725.2220000	18907.1236000	9.3445796
A-9	18923	53625.6544000	1592079.7900000	56266.7776000	5064.1216000	29.6887713
C-2	55172	9510.0000000	1557123.0000000	9510.0000000	855.9000000	163.7353312
JL	18861	2980.6500000	1231578.2775000	32420.5738000	2980.6500000	413.1911756
N-1	72888	919.6505000	949375.6757500	3182.8978000	296.4773000	1032.3222526
T-5	42967	9221.3750000	733832.9500000	9221.3750000	831.8560000	79.5795584
C-1	5137	62763.0555000	567620.7209640	62763.0555000	5664.0957000	9.0438669

In order to accurately target the potential customers, the sample selection should be able to address the needs in terms of both design and costing.

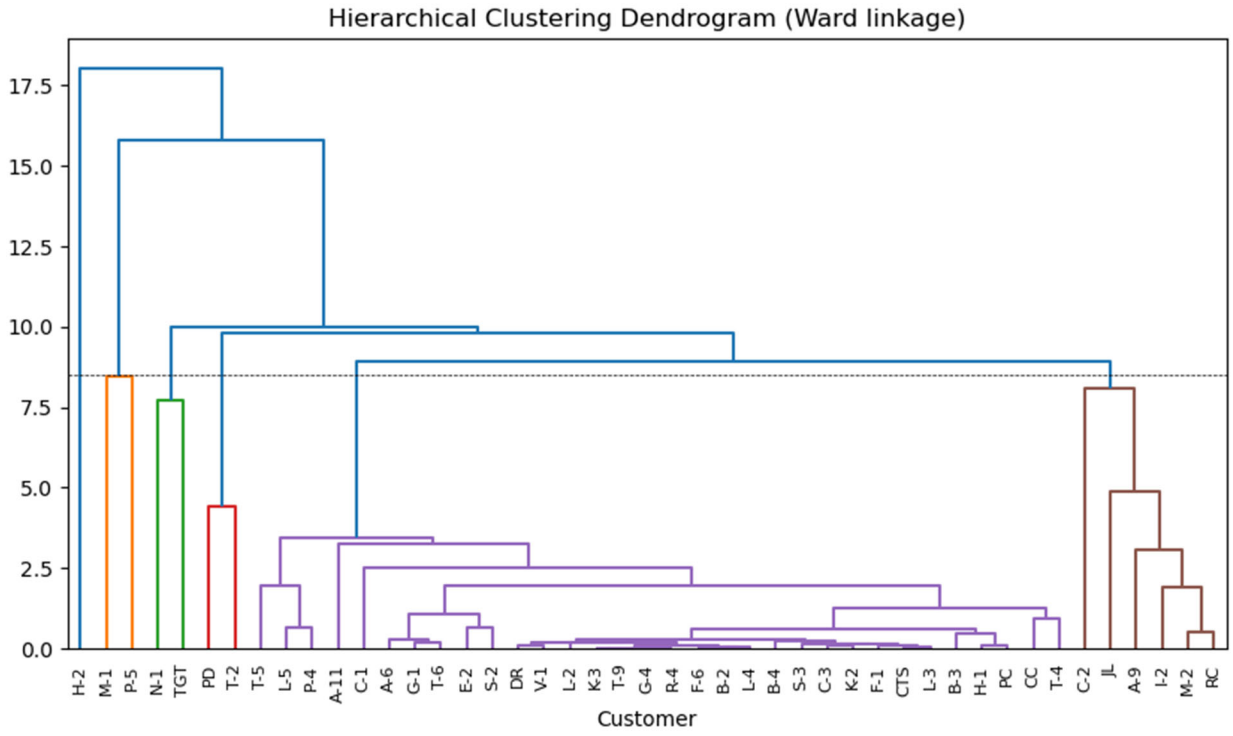
To look at each type of carpet, we were able to spot the top customers by the total area they ordered. For Hand tufted carpet, the majority have been purchased by customers with the code C-1, TGT, and A-9. And interestingly, these customers all come from the USA.

```
HAND TUFTED order dy TotalArea
              QtyRequired     TotalArea           Amount        AreaFt  \
CustomerCode
C-1                  4176 51599.0555000   454005.3034640 51599.0555000
TGT                 15045 37630.3318000 11341052.5133400 37630.3318000
A-9                 11716 24320.8850000   959040.2500000 26028.5000000
H-2                 26612 15056.0625000  2458060.8000000 15056.0625000
M-1                  2697 11572.0000000   133665.2500000 11572.0000000
```

However, this is true for the US customers only and Champo Carpets needs something more robust, something that can be applied to more customers. Therefore, later in our analysis we formed clusters of customers who were similar in nature and different from the remaining lot.

Part b: Customer Segmentation

In clustering the customer data, we tried different configurations of methods and distances between measurements. We settled on Ward's method. Ward's method considers the loss of information that occurs when records are clustered together. Ward's method measures the loss of information with "error sum of squares" that measures the difference between individual records and a group mean. We tried various thresholds, one of the thresholds we tried was 8.5, this yielded 6 clusters as seen below. The dendrogram is seen below, the vertical lines in the dendrogram reflect the distance between the records. Blue lines indicate single clusters, in this case we only have one cluster made up of one record – customer h-2. Visually the clustering below made intuitive sense as we had 5 clusters populated with multiple records and only one singleton cluster.

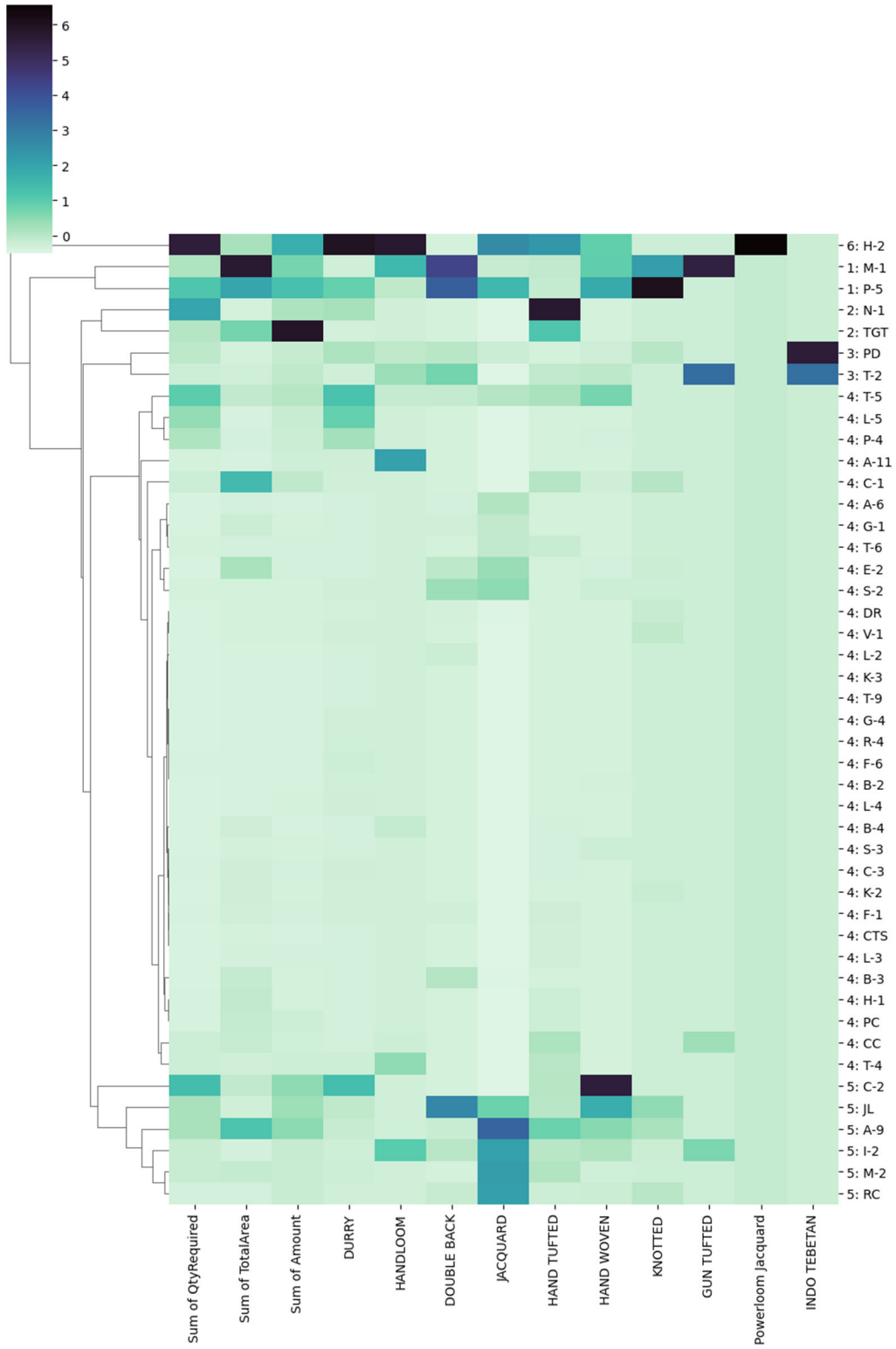


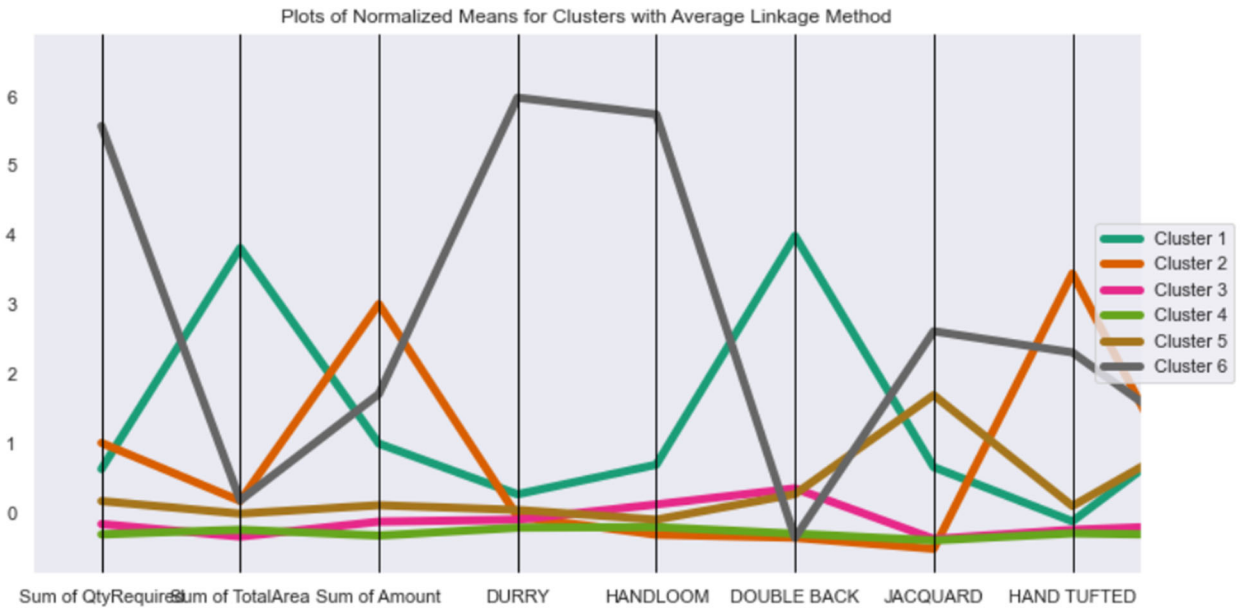
The clustering results are as follows:

```
1 :  M-1, P-5
2 :  N-1, TGT
3 :  PD, T-2
4 :  A-11, A-6, B-2, B-3, B-4, C-1, C-3, CC, CTS, DR, E-2, F-1, F-6, G-1, G-4, H-1, K-2, K-3, L-2, L-3, L-4, L-5, P-4, PC, R-4, S-2, S-3,
T-4, T-5, T-6, T-9, V-1
5 :  A-9, C-2, I-2, JL, M-2, RC
6 :  H-2
```

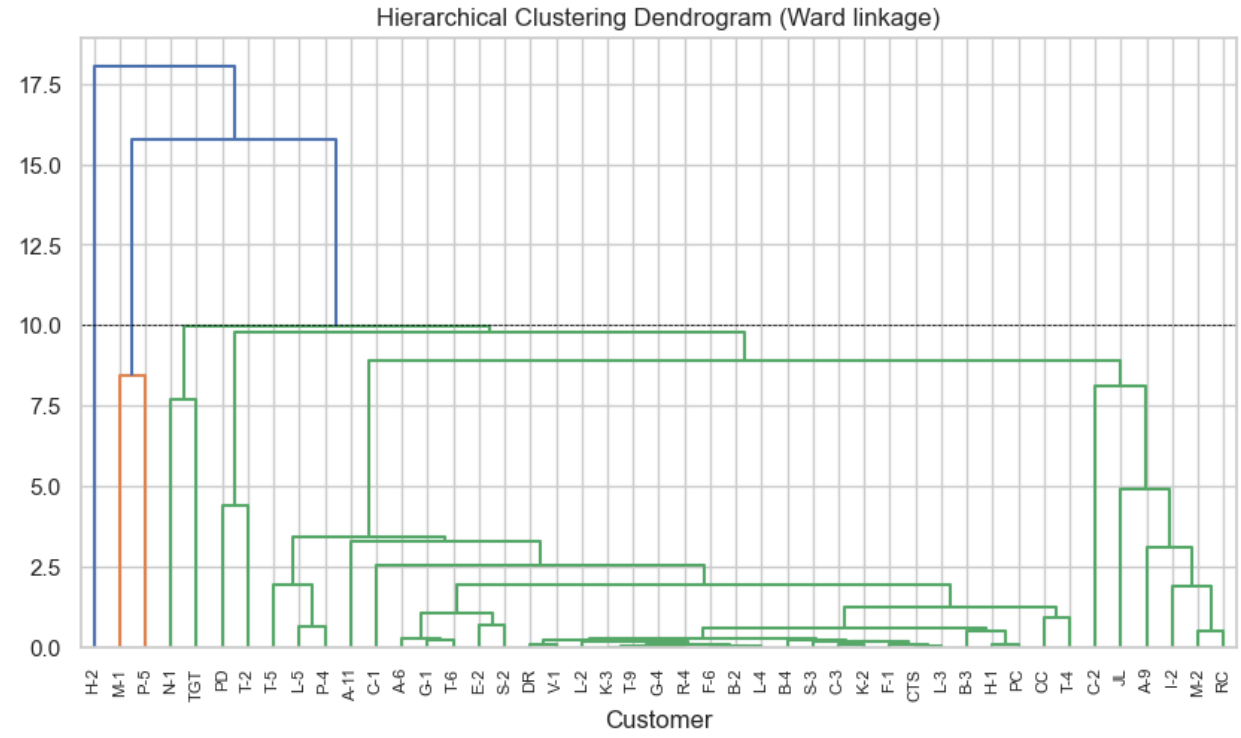
To understand the clustering and dendrogram better, we created a heatmap with the above clusters. The heatmap is reproduced below. Darker color indicates a strong influence of a given variable on the cluster. Starting at the top of the heatmap we see cluster 6, which is a singleton cluster of H-2. Cluster 6 and customer H-2 standout by requiring large total quantities, as well as a preference for Durry, Handloom, and Powerloom Jacquard style carpets. Cluster 6 could also be distinguished by significant order amounts, and a weaker but still noticeable preference for Jacquard and Hand tufted style carpets. Moving further down the dendrogram, we arrive at cluster 1, which is composed of customers M-1 and P-5. This cluster is distinguished by a strong preference for Double Back style carpets. Next is cluster 2, which is made up of customer N-1 and TGT, and is most clearly distinguished by a customer preference for the Hand Tufted style carpets. There are other preferences for this cluster that could be gleaned from the heatmap, however they are not as clear, since each of the customers has a varying preference across groups. We could say however, that both customers in cluster 2, order larger amounts than customers in cluster 4. Continuing down the heatmap, cluster 3 seems to be distinguished by a strong preference for Indo Tibetan style carpets. Cluster 4, the largest of clusters, is distinguished by the fact that none of the variables seem to play a prominent role in this cluster. Many of the customers in cluster 4 do not seem to have strong relationships with any of the measurements.

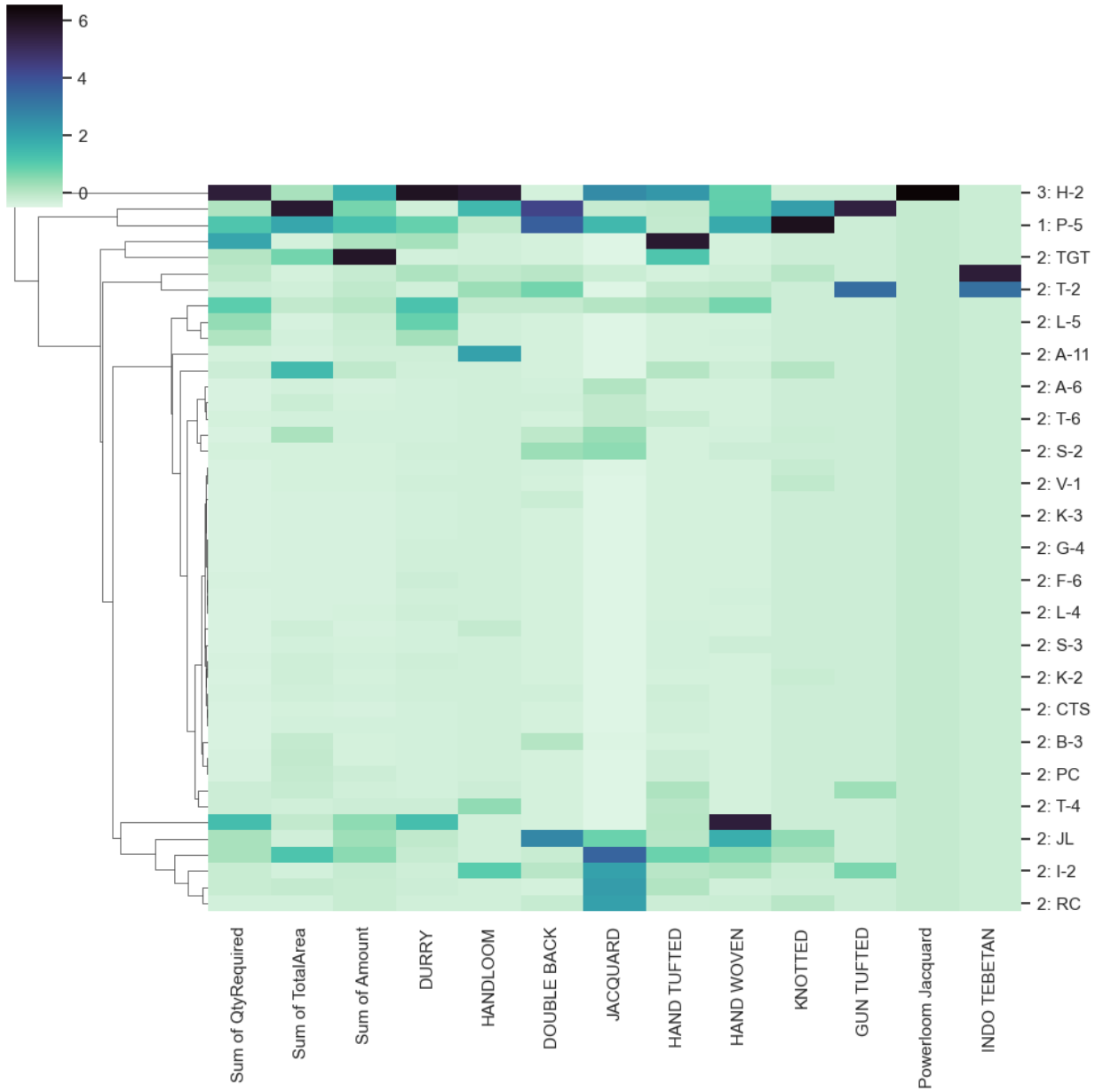
Finally, cluster 5 seems to be influenced by the Jacquard style carpets (with the exception of customer C-2). Each of the aforementioned clusters has other interactions, and possibly more complicated relations than can be gleaned from this heatmap.





For comparison with our k-MEANS model, we also tried a 3 cluster model. The dendrogram and heatmap is reproduced below. Customer H-2 once again populates a singleton cluster and clusters M-1 and P-5 also populate the same cluster as before. The rest of the customers are clustered into one big cluster. This large cluster is composed of customers that formerly made up clusters 2, 3, and 4 in the 6 cluster model.





Part c: Other Machine Learning tools

In identifying the most important customers and the most important products just generally and for a particular customer, Champo's carpets can employ one of several machine learning techniques to understand the segmentation of a particular customer and subsequently be able to present more relevant carpet samples that can lead to better conversion rates.

In terms of unsupervised learning, one method Champo's Carpets can use based on their available data are neural networks. Although the explainable relationships between the predictors and the outcome groupings are hidden in the case of neural networks, it can potentially be very

effective in for example using the data to form segmentations and groupings of what carpet material a customer is likely to buy. Based on this, Champo's Carpets would be able to send a more relevant sample by material and/or color or carpet type to achieve better conversion rates.

However, with the data collected, Champo's Carpets also has the option of using supervised learning models using one of several outcomes- the amount a customer is likely to spend or quantity they are likely to order which will show the potential importance of a customer, as well as what carpet type or carpet material a customer is most likely to order. This can be done through a prediction analysis that uses mathematical models such as a multiple linear regression. However, in order to do this, it would be helpful for Champo's Carpets to collect slightly more information about each of its orders such as the Company size, for example, in order to have more predictive metrics than just what country an order is from. Or, Champo's Carpets would be able to use machine learning methods that **do not rely on prior models such as k-NN**, which will predict the desired outcome variables based on other similar customer orders to tell what carpet a customer is likely to order, or a CART analysis to segment based on specific characteristic groupings, different potential outcomes of a desired outcome variable.

For supervised learning, Logit regression analysis is also a good approach for Champo. But before we implement this method, more customer data need to be collected, such as demographics, geographic and purchase behavior.We can find the factors that differentiate the customer preferences. We can also apply this approach to classify if a new customer will like a specific carpet item. In addition, it provides a better solution in examining the differences between or among clusters, without considering the normality issue.

Part d:Final recommendations

Based on looking at the results of clusters from various methods above- we are able to observe a few important customers as well as their tastes, which Champo's Carpets should cater its samples for accordingly.

For example, customer H-2 from the USA, which is subject to its own cluster based on the Ward Method, is one of the largest orders in terms of the total quantity of carpets ordered. Based on H-2's segmentation and characteristics, we recommend that the samples being sent to H-2 are either Durry or Handloom carpet types since they are the most in demand for this cluster. Additionally, over 35% of H-2's orders are made up of a plain color scheme such as navy, blue, gray, and black which is important to know when catering to this specific cluster.

Similarly, customers M-1 and P-5 make up a common cluster. This cluster is distinct in that it buys fewer large carpets, giving it one of the highest total carpet areas relative to all the orders.

This cluster has a strong affinity for Double Black, Knotted, Jacquard, and Hand Woven carpet types and we recommend that their samples should align with these types.

In terms of the total amount spent on carpets, customers N-1 and TGT form a cluster and should be prioritized due to the total amount they spend on carpets. We recommend that the majority of samples sent to members of this cluster are Hand Tufted, since this cluster is distinct in having a strong preference for that carpet type.