BAN 620 Data Mining

Assignment 1

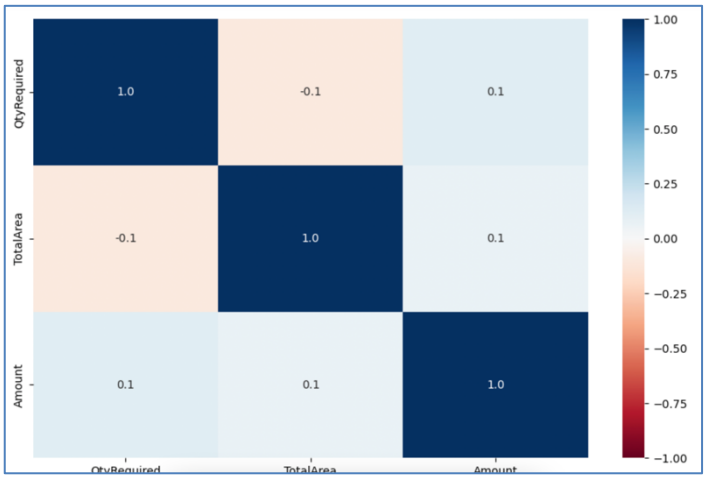
Submitted by: Group 4

Name	NETID
Jasvitha Buggana	js3225
Krupa Shah	yc4954
Manan Upadhyay	rs6739
Preksha Shah	cz2412
Shivani Agrawal	lw3758

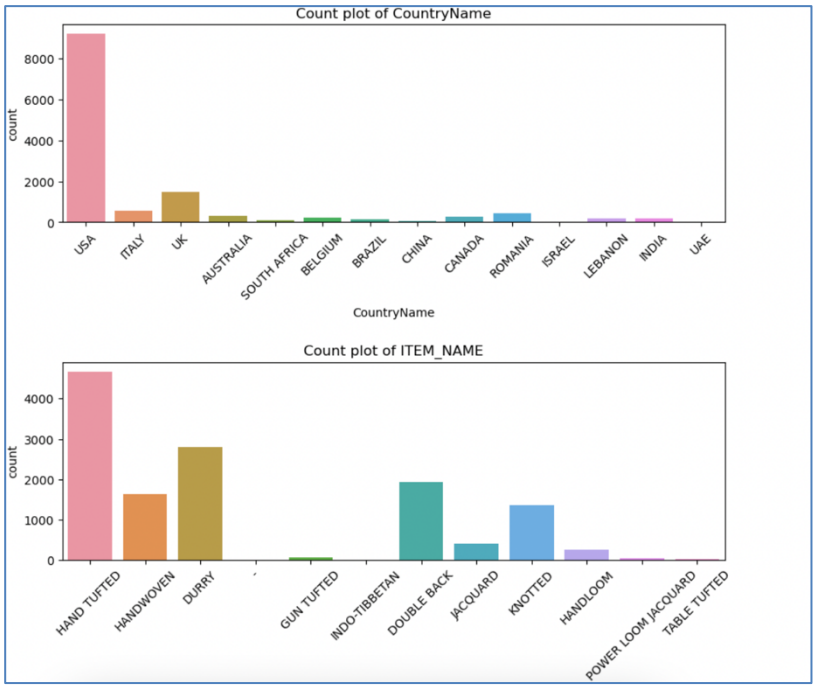
Champo Carpets Case:

Question 1: Visual exploration of Sales Data:

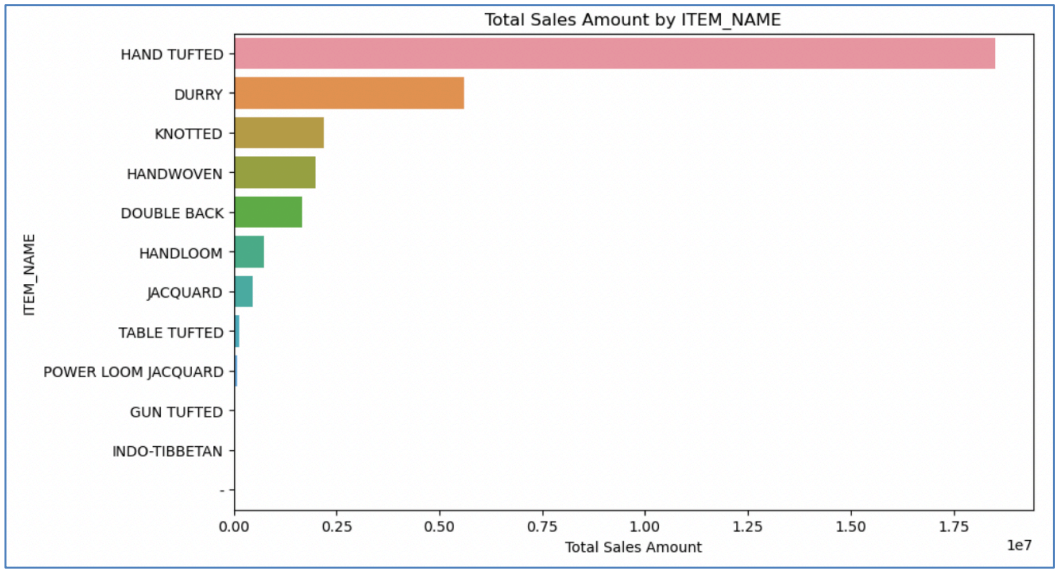
- The Champo Carpets Case data set has 13135 Rows and 12 variables. The name of the colu mns goes like ['CustomerCode', 'CountryName', 'QtyRequired', 'TotalArea', 'Amount','ITEM _NAME', 'QualityName', 'DesignName', 'ColorName', 'ShapeName', 'AreaFt', 'AreaMtr']
- There were no null values in any column
- **Redundant Columns Found:** AreaFt and AreaMt are redundant columns, we already have TotalArea column which is the area of the carpet in mts, so we are not taking them in our nu merical colums list¶
- The correlation graph of the numerical column shows no significant co-relation between them.

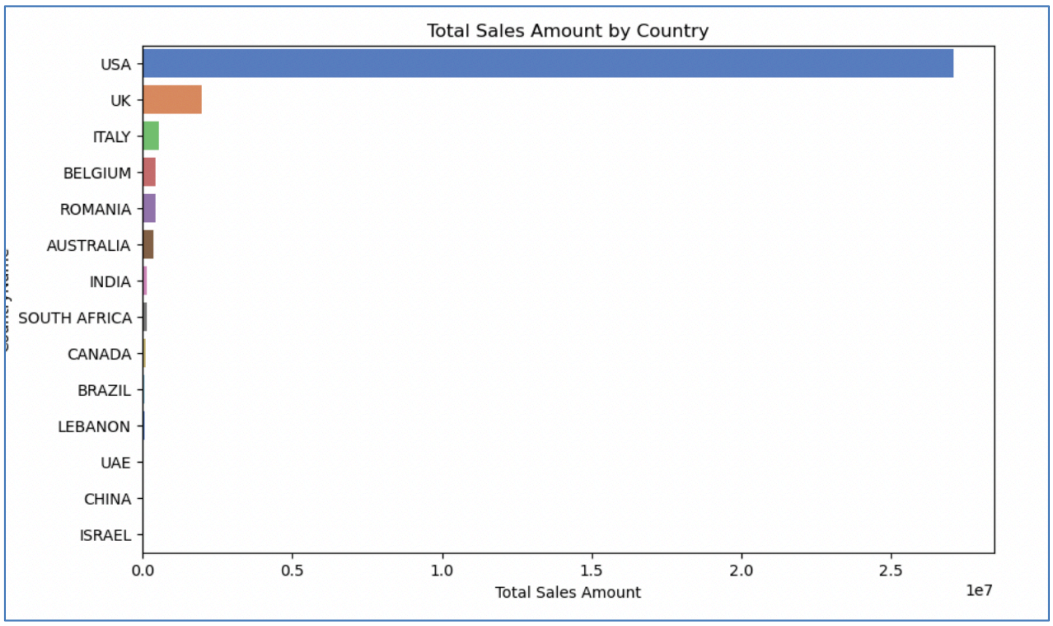


- The count plot told us which country is giving the orders to Champo Carpets most frequently and what is the most frequently bought item (USA and HAND TUFTED carpets)



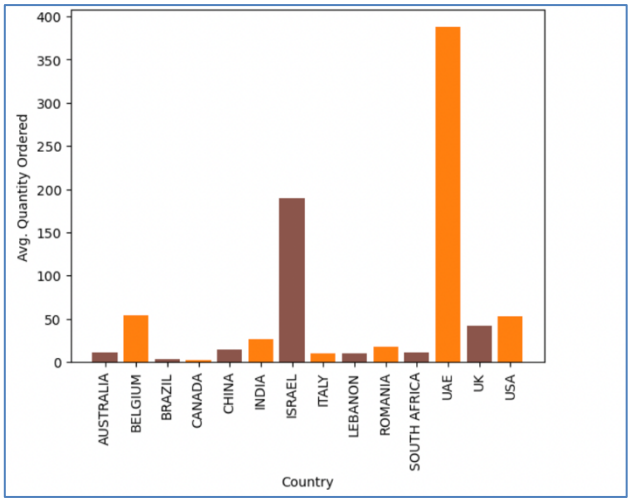
- Then we did an analysis **based on the revenue (Sales) and tried to found total sales by Item_name and Country.** It turned out that Hand Tufted carpets and orders from USA are the major chunck of revenues for Champo Carpets, so Champo should always be ready with the inventory as per the liking of US customers which is rectangular pattern (better explained in the code).



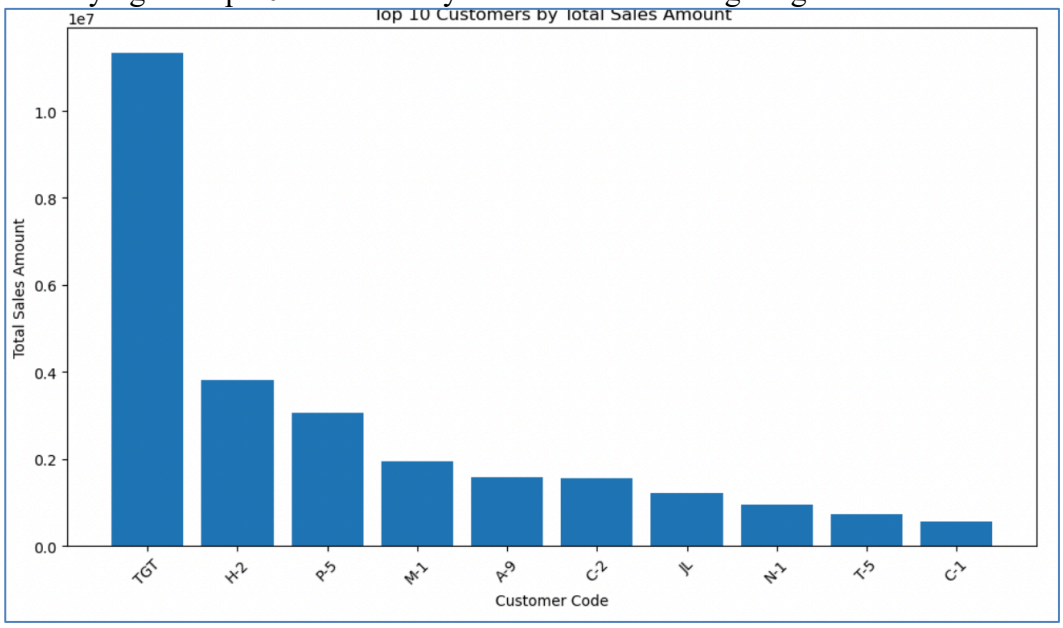


- <u>Till now, we have identified that USA is giving the maximum number of orders and their cumulative amount is also the largest, now lets see which country is giving the order with maximum items on an average, This will give us the information about which country places a big order and Champo should never miss this order</u>

- Next, we saw the average order quantity per country and **found that the average order quantity is highest for UAE**, like whenever they place an order, it will have the maximum items as compared to the orders by other countries



- Identifying the top 10 customers by their codes who are giving maximum sales

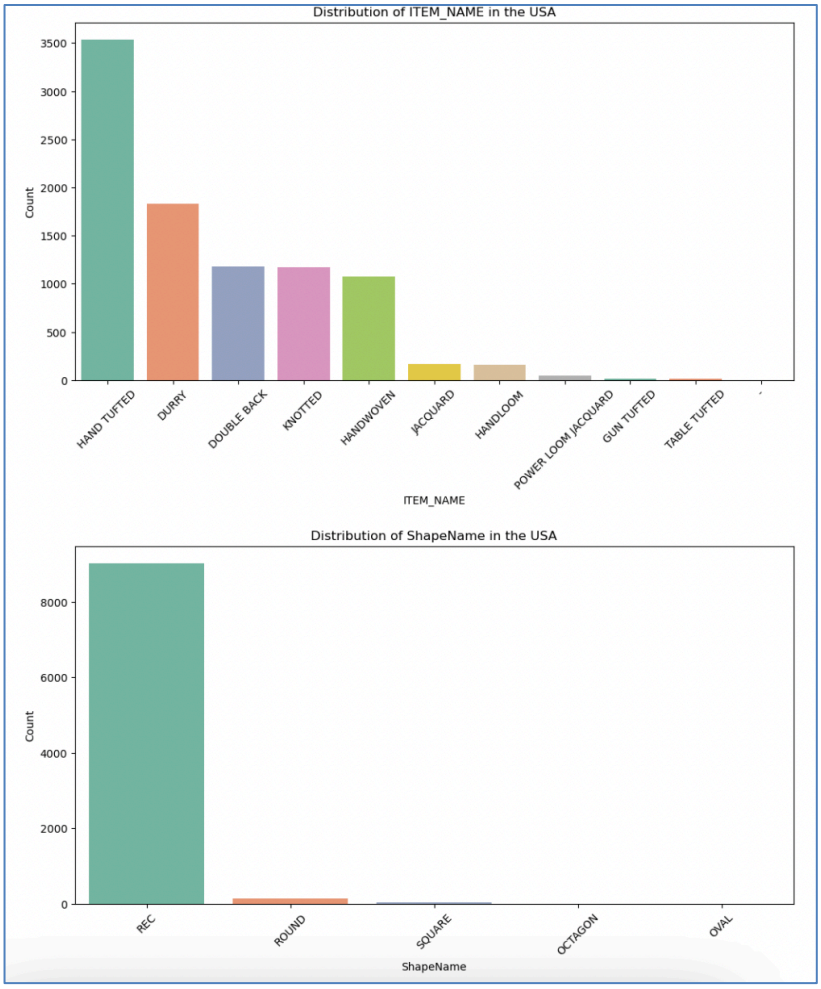


- Now, **identifying top customers in top 5 countries**. So that once a geography is targeted , those customers can be targeted too, and a high amount of sale can be pulled from each top 5 country

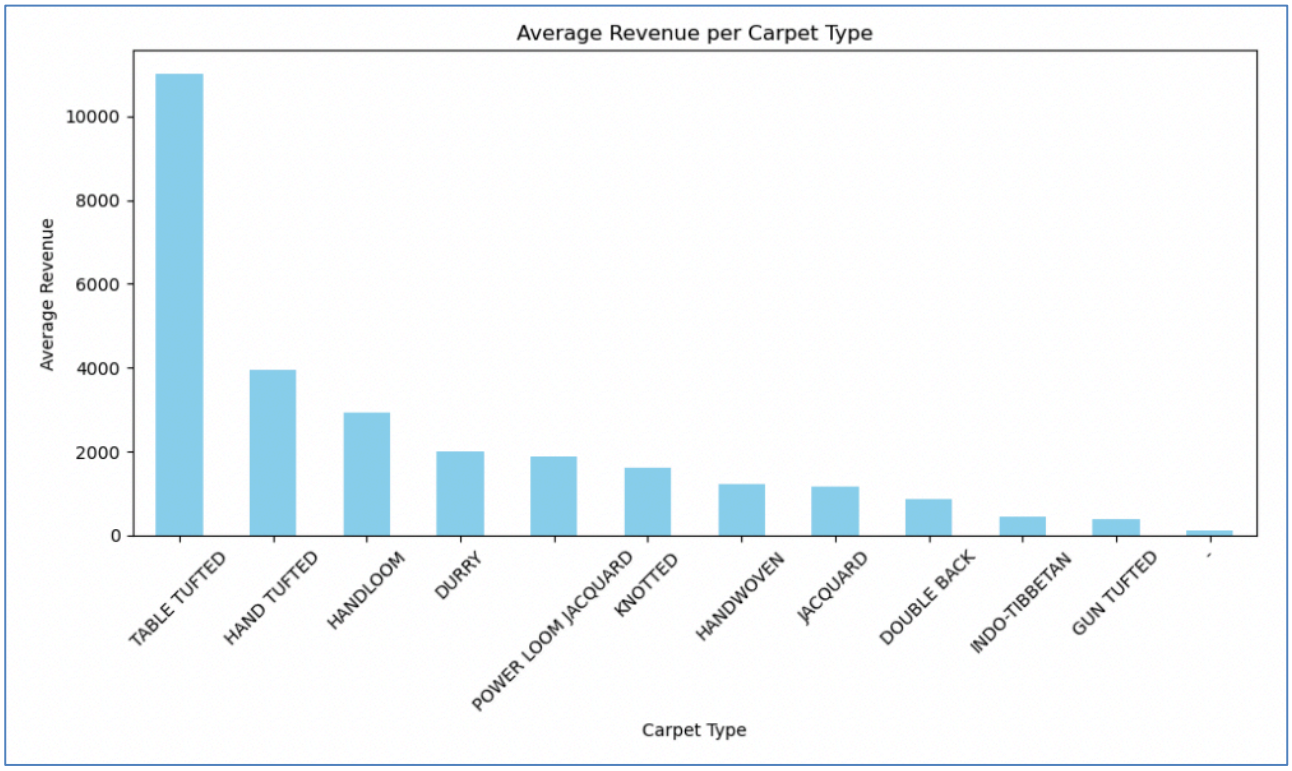
```
Top 5 customers in USA:
   CountryName CustomerCode     Amount
43         USA          TGT  1.134e+07
31         USA          H-2  3.805e+06
38         USA          P-5  3.067e+06
34         USA          M-1  1.960e+06
22         USA          A-9  1.592e+06
------------------------
Top 5 customers in UK:
   CountryName CustomerCode     Amount
18          UK           JL  1.232e+06
19          UK          T-5  7.338e+05
------------------------
Top 5 customers in ITALY:
   CountryName CustomerCode      Amount
12       ITALY          T-2  563098.848
------------------------
Top 5 customers in BELGIUM:
  CountryName CustomerCode      Amount
3     BELGIUM           PD  404528.945
2     BELGIUM         L-2    21503.495
4     BELGIUM         T-9      758.970
------------------------
Top 5 customers in ROMANIA:
   CountryName CustomerCode      Amount
14     ROMANIA          I-2  426626.048
------------------------

Segment Champo's customers
```

- **Doing some extra analysis on USA** as it is the top customer for Champo, so that Champo can have that inventory ready. We are finding which item USA orders the most and which Shape they are a fan of.



- Some Extra analysis to see **which carpet type is giving the highest revenue on an average**
    - **Answer**: We found that **it was not most selling - HAND TUFTED but rather TABLE TUFTED this may be** because it is costly because it was not something which is getting ordered the most



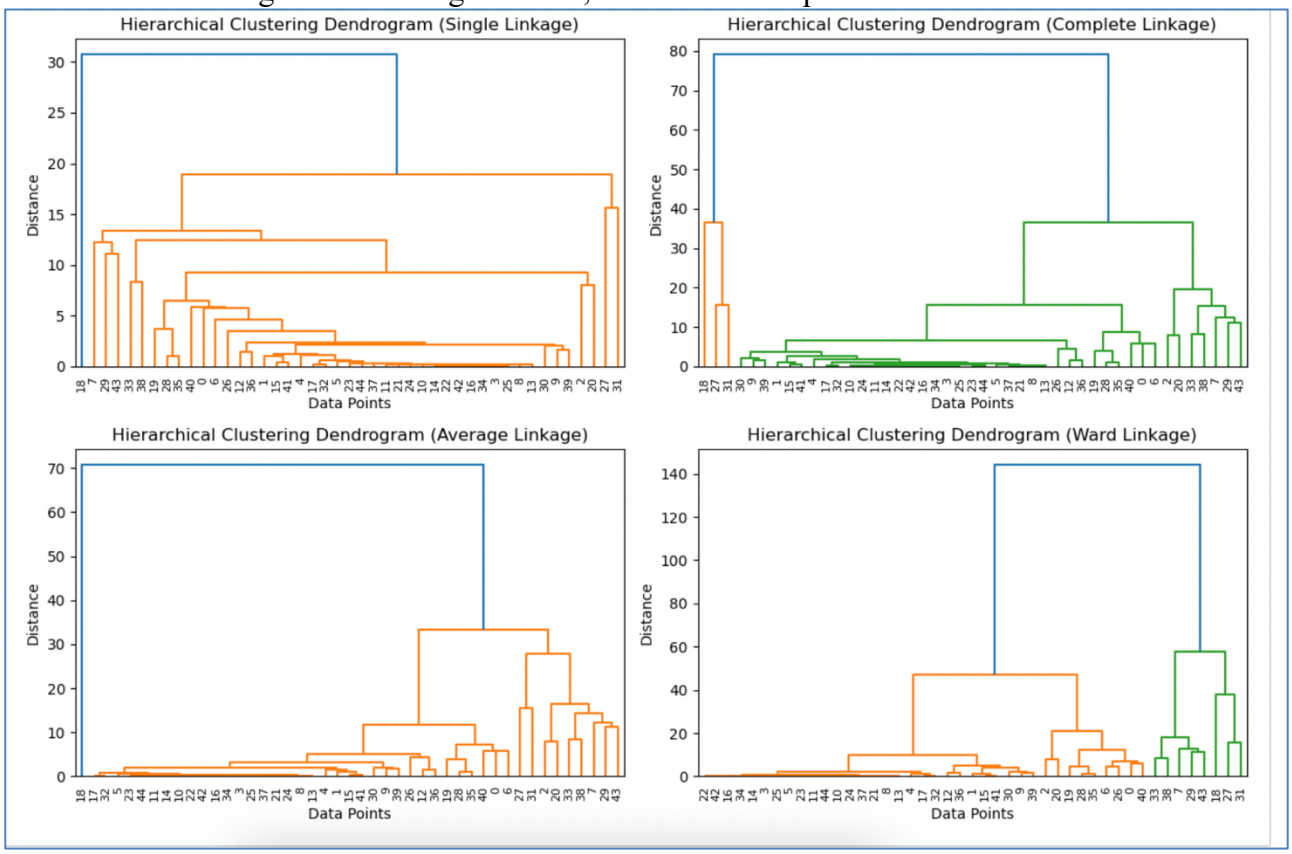
Question 2: Segmentation of Customer Data:

Approach:

- Converting all the numerical data types to float first, so that normalization could be done easily
- Normalizing the dataframe using StandardScaler()
- Calculating distance between data points using Euclidian Matrix
- Then plugging the above Euclidian matrix dataframe into Linkage function and experimenting with different methods, namely: ['single', 'complete', 'average', 'ward'] to plot dendrograms
- Choosing the number of clusters from the dendograms
- Then assigning cluster labels to the datapoints using the linkage_methods = ['single', 'complete', 'average', 'ward']
- Finally, making the cluster line chart by grouping data points belonging to the same cluster to identify each cluster's characteristic.

Observations:

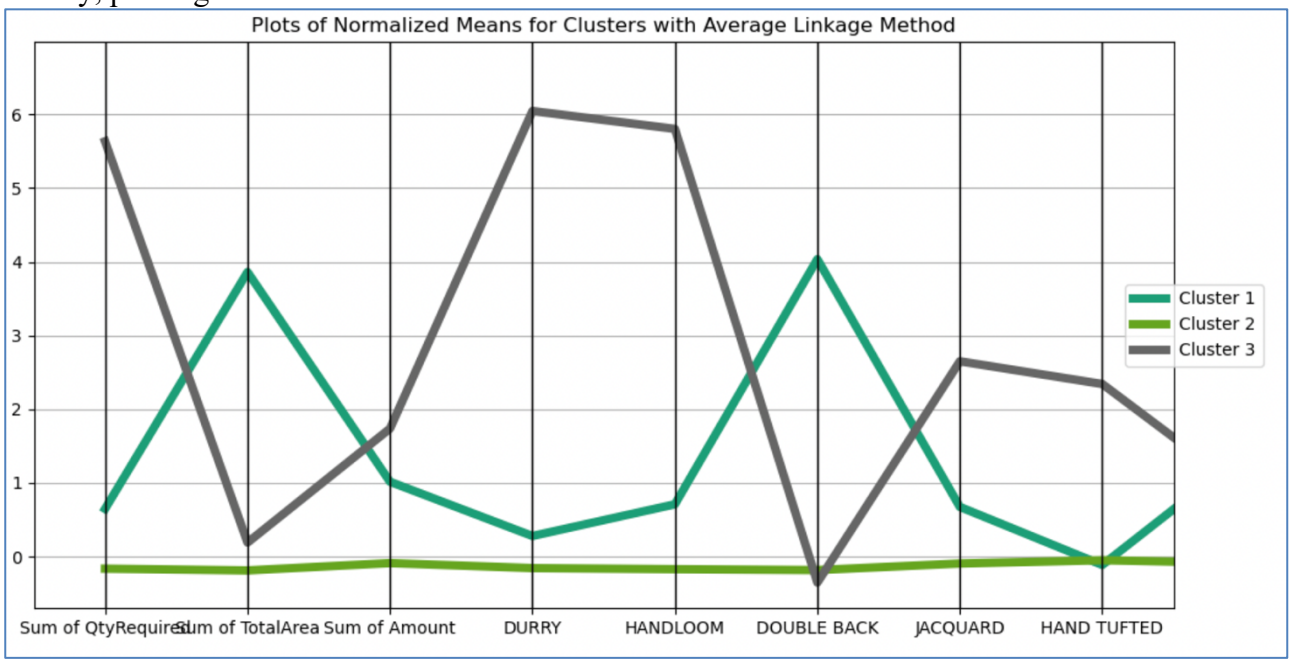
- From the dendrograms we are making s to choose maximum 3 clusters, Ideally no bar in the dendrogram should not have a single point in it because it would mean that only that datapoint belongs to that cluster. But here, in all the cases there is at least one cluster with a single datapoint
- So, we have opted 3 clusters as maximum limit and now assigning data labels to these three clusters using various linkage method, as asked in the question



- Cluster membership using various linkage methods, there is so such method which does not give us a single datapoint in either cluster.

```
Cluster membership for 'single' linkage:
Cluster 1: A-11, A-6, A-9, B-2, B-3, B-4, C-1, C-2, C-3, CC, CTS, DR, E-2, F-1, F-6, G-1, G-4, H-1, I-2, JL, K-2, K
-3, L-2, L-3, L-4, L-5, M-2, N-1, P-4, P-5, PC, PD, R-4, RC, S-2, S-3, T-2, T-4, T-5, T-6, T-9, TGT, V-1
Cluster 2: M-1
Cluster 3: H-2
----------------------
Cluster membership for 'complete' linkage:
Cluster 1: M-1, P-5
Cluster 2: A-11, A-6, A-9, B-2, B-3, B-4, C-1, C-2, C-3, CC, CTS, DR, E-2, F-1, F-6, G-1, G-4, H-1, I-2, JL, K-2, K
-3, L-2, L-3, L-4, L-5, M-2, N-1, P-4, PC, PD, R-4, RC, S-2, S-3, T-2, T-4, T-5, T-6, T-9, TGT, V-1
Cluster 3: H-2
----------------------
Cluster membership for 'average' linkage:
Cluster 1: A-11, A-6, A-9, B-2, B-3, B-4, C-1, C-2, C-3, CC, CTS, DR, E-2, F-1, F-6, G-1, G-4, H-1, I-2, JL, K-2, K
-3, L-2, L-3, L-4, L-5, M-2, N-1, P-4, PC, PD, R-4, RC, S-2, S-3, T-2, T-4, T-5, T-6, T-9, TGT, V-1
Cluster 2: M-1, P-5
Cluster 3: H-2
----------------------
Cluster membership for 'ward' linkage:
Cluster 1: M-1, P-5
Cluster 2: A-11, A-6, A-9, B-2, B-3, B-4, C-1, C-2, C-3, CC, CTS, DR, E-2, F-1, F-6, G-1, G-4, H-1, I-2, JL, K-2, K
-3, L-2, L-3, L-4, L-5, M-2, N-1, P-4, PC, PD, R-4, RC, S-2, S-3, T-2, T-4, T-5, T-6, T-9, TGT, V-1
Cluster 3: H-2
----------------------
```

Finally, plotting the cluster means for various attributes:



Our points based on cluster characteristics:

Cluster 2: Customers in this group prefer buying fewer carpets, and they tend to choose smaller-sized carpets with less emphasis on specific features like durry, handloom, double back, jacquard, and hand-tufted styles.

Cluster 1: Customers in this group typically buy a moderate quantity of carpets, opting for larger-sized ones. They show a moderate interest in durry and handloom styles, a high preference for double back, and lower interest in jacquard and hand-tufted styles.

Cluster 3: There's only one customer in this group who buys a high quantity of carpets. However, they prefer slightly smaller carpets compared to Cluster 2. These carpets are priced higher, and this customer shows a strong preference for durry and handloom styles, very little interest in double back, and a moderate interest in jacquard and hand-tufted styles.

Question3: Other Machine Learning Tool which Champo Carpts could use:

1. **PCA:** Since there are many features, ML engineers can do PCA analysis to identify the importance of each feature and then use only the important features to build a parsimonious model
2. **Comparison between and within Clusters:** This will give us an idea of how similar customers are within each cluster, and also what are the factors which make the difference between clusters, ex: if any cluster likes the material quality of Tufted 60C while another cluster might like DB 60 C more, so that customer targeting within and between clusters could happen as per customer's liking
3. **K-means Clustering:** To identify the optimal number of clusters, instead of identifying from the dendograms, it is wiser to do a K-means clustering before and see what the elbow-point in the graph is, this might save the computation time getting wasted on more number of clusters.

Question 4: Final Recommendation for Champo Carpets:

1. Since USA is contributing max to the sales **(more than 86.2%)**, Champo should always have inventory ready for the likings of USA orders, **Example**: they order Hand Tufted carpets most of the time and these carpets are having rectangular shaped designs, so Champo should be ready with these carpets, or to say the least should have the procurement partners on-board who could supply raw-material for the carpets of USA likings.
2. The big-orders are placed by UAE, meaning whenever UAE places an order, they order in bulk with maximum average quantity than other orders, so maintain good business-relations with UAE and try to predict their next big order items beforehand.
3. **Don't lose cream of the cream**: Since it is impossible for any company to cater to all the customers in all its geographies, to prevent revenue losses happening, we have identified the top customers in top 5 countries, **offer them discounts or deals to earn** their loyalty.

```
Top 5 customers in USA:
   CountryName CustomerCode     Amount
43         USA          TGT  1.134e+07
31         USA          H-2  3.805e+06
38         USA          P-5  3.067e+06
34         USA          M-1  1.960e+06
22         USA          A-9  1.592e+06
----------------------
Top 5 customers in UK:
   CountryName CustomerCode     Amount
18          UK           JL  1.232e+06
19          UK          T-5  7.338e+05
----------------------
Top 5 customers in ITALY:
   CountryName CustomerCode      Amount
12       ITALY          T-2  563098.848
----------------------
Top 5 customers in BELGIUM:
  CountryName CustomerCode      Amount
3     BELGIUM           PD  404528.945
2     BELGIUM          L-2   21503.495
4     BELGIUM          T-9     758.970
----------------------
Top 5 customers in ROMANIA:
   CountryName CustomerCode      Amount
14     ROMANIA          I-2  426626.048
----------------------
```

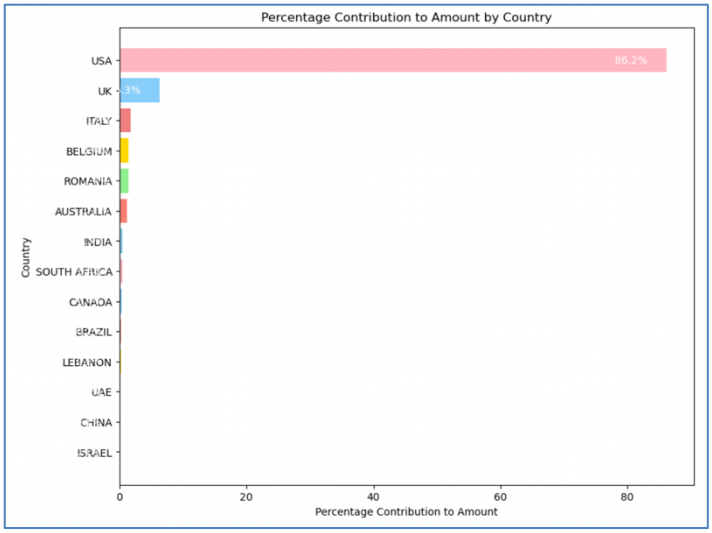


Fig explaining point 3 and 1 respectively