Comic Center

Sagar Bansal

SECTION I: ANALYSIS SUMMARY

1) Understanding the Customer Behaviour:

After doing initial exploration, we found that the average total amount spent by a customer is roughly 25.94 dollars which is almost eight times the average amount of roughly 3.37 dollars spent by a customer on comic books (see Table 1 and 3 in Appendix). The mean amount of time spent in-store and mean number of items purchased by a customer were found to be 24.53 minutes and 3 items, approximately (see Table 2 and 5 in Appendix). From Table 4 and 6 in Appendix, it was found that most of the customers were 18 years old and only made a single visit over the last month. In addition, we can observe the distributions of all the recorded variables in Figure 1 to 6 in the Appendix.

Based on our statistical analysis in Appendix, we labelled different groups of customers. The first cluster represents **older customers who are primarily interested in saving their time and buying only what they need.** Hence, they visit the store once in a while, spend least amount of time in store and buy least number of items. The second cluster represents **younger customers who are primarily interested in browsing the products, buying latest and cheap items.** Hence, they spend most amount of time in store, visit the store very often and buy maximum number of items in least amount of total purchase. The third cluster represents **younger customers who are primarily interested in buying from their list to save some time but also open to browse a thing or two.** Hence, they spend decent amount of time in store, buy decent number of items, make decent number of visits and spend decent total purchase amount.

After labeling different customer groups, we went forward with discovering their interest in spending on comic books using comicBookAmount variable (see Table 7 and 9 in Appendix). We saw that **customers in cluster 1 are generally not interested in comic books at all** as they spent zero dollars on buying comic books ("non-readers" in terms of comic books). This made sense as usually older people are more inclined towards reading novels and other serious literature. Also, finding the right comic may take some time and these customers are focusing on saving their time in store. In contrast, **customers in cluster 2 seem to be really interested in comic books** as they spent maximum average amount on comic books among all other cluster ("nerds" in terms of comic books). This is reasonable as usually younger people are more inclined towards comic books as a source of entertainment. Also, usually comic book companies release new comic books very frequently. This could be a reason that customer in this cluster made high number of visits to get the latest comic books and spent high amount of time in store to browse from all the options.

Our last cluster had **customers who were somewhat interested in comic books** as they spent some amount on comic books ("hobby readers" in terms of comic books). This also made sense as some of the younger people are only interested in reading comic books in part of their pass time and spend rest of their pass time in other fun activities. Hence, they spend more

amount of time in store and made a greater number of visits compared to customers in cluster 1 (least interested). We also performed a statistical test to confirm that these customer groups are reasonable (details can be found in the Appendix).

As per these spending habits, our client should specifically target customers represented by cluster 1 and 2 to increase his comic books sales. He should work on maintaining customers in cluster 1. He can further try to understand and advertise those types of comic books that customers in cluster 2 may find interesting and are inclined to buy, which will increase the amount they spend on comic books. He can also investigate whether customers in cluster 2 are more inclined to come on particular days of a week or month since they make somewhat low number of visits and then initiate discounts/offers on comic books on those days. For customers in cluster 3, he can give them free comic books on their visit to see if their spending habits towards comic books change after a few weeks.

2) Limitations of the analysis:

Our client should be aware of some reservations before making any decision based on the analysis. First, these customer groups and their spending habits are specific to the data he provided. There may a possibility of other customer groups who did not visit the store last month when he collected the data. Second, it is not clear whether the dataset has all the customers who visited the store last month or only a subset of them. If it only represents a subset of them, the client should consider collecting more data for future analysis. Third, some of the variables in the dataset such as Time spent in store and Number of visits are subjective and their values can be inaccurate for few or many customers as they were dependent on the wisdom of customers. Customers might not have known the exact time spent in store or the exact number of visits they made in the last month while filling out the survey.

SECTION II: APPENDIX

1) Statistical Analysis:

We used Agglomerative Hierarchical clustering technique with Average linkage Criterion to identify relationships between different customers and understand their purchase behavior. Based on our domain knowledge, we decided to perform cluster analysis using all the variables except comiBookAmount i.e., timeInStore, purchaseAmount, Age, numItems and numVisits variables. We kept comicBookAmount variable to validate the final clusters later on. After running the cluster analysis using SAS 9.4 version, we observed the resulting dendrogram to examine jumps (see Figure 7). We noticed two clear jumps from 3 to 2 clusters and 2 to 1 cluster and identified two candidate cluster configurations that were the beginning of those jumps. The candidate configurations had three clusters in one and two clusters in the other configuration.

As shown in Table 7 and 8, we ran the MEANS procedure that gave us the mean, standard deviation, minimum and maximum value of all variables in each cluster for both cluster configurations. We then tried to comprehend different clusters in each of the cluster configurations and found that configuration with three clusters made more sense practically. The cluster configuration with two clusters had wide range of values for each variable in the second cluster. Moreover, the lowest and highest values for each of numVisits and numItems variables in second cluster was exactly the same as the smallest and largest values for

each of numVisits and numItems variables in the whole dataset. Hence, we were not satisfied enough to use this configuration as our final configuration.

In the configuration with three clusters, the first cluster has customers with minimum average amount of time spent in store, average number of items purchased and average number of visits but maximum average total purchase amount and average age, among all three clusters. The second cluster has customers with lowest mean total purchase amount but highest mean amount of time spent in store, mean number of items purchased and mean number of visits, among all clusters. The mean age of customers in this cluster is around 17 years that is almost similar to the mean age of customers in our third cluster. The third cluster has average amount of time spent in store, average total purchase amount, and average number of items purchased around 20.78 minutes, 24.68 dollars and 3 items, respectively. These values lie somewhat in between the corresponding average values of cluster 1 and cluster 2. The average number of visits for cluster 3 is around 1 that is similar to the average number of visits in cluster 1. Hence, we used this configuration as our final configuration for further validation.

To validate our final clusters statistically, we used comicBookAmount variable because client wanted to know spending behavior of each group on comic books. As seen in Table 9, we performed ANOVA test to confirm that the means of comicBookAmount are statistically different for the three different clusters at 5% significance level. The p-value (0.0010) came out to be less than 0.05. Hence, we reject the null hypothesis i.e., the test confirms that the mean values of comicBookAmount are statistically different for the three clusters at 5% significance level and the clusters are reasonable.

2) Supporting figures and tables:

Figure 1: Histogram – Comic Book Amount ($)

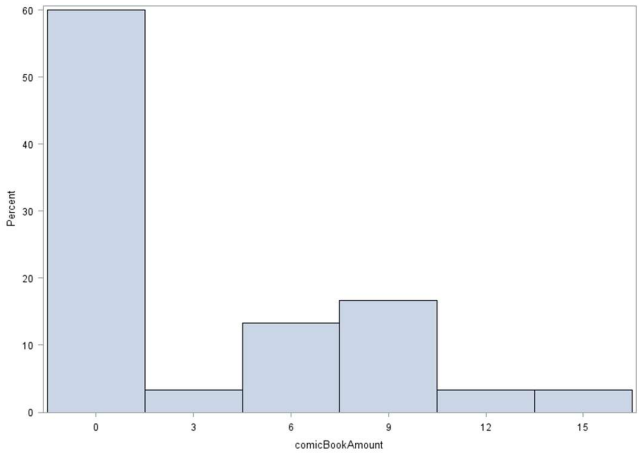

Table 1: Numerical Summary of Comic Book Amount ($)

Basic Statistical Measures			
Location		Variability	
Mean	3.373667	Std Deviation	4.67945
Median	0.000000	Variance	21.89729
Mode	0.000000	Range	15.49000
		Interquartile Range	6.31000

Figure 2: Histogram – Time Spent in Store (min)

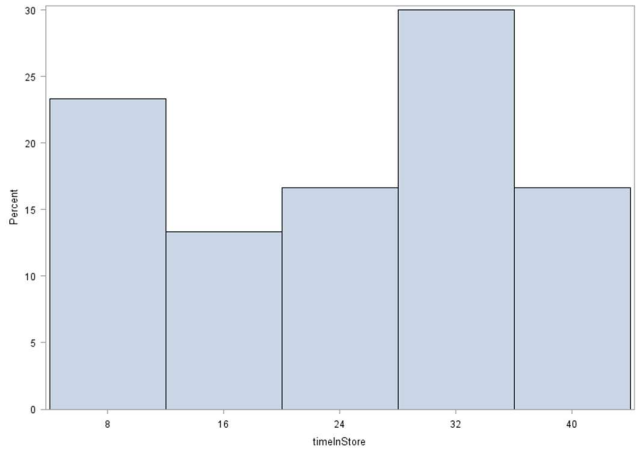

Table 2: Numerical Summary of Time Spent in Store (min)

Basic Statistical Measures			
Location		Variability	
Mean	24.53333	Std Deviation	11.40700
Median	24.00000	Variance	130.11954
Mode	35.00000	Range	35.00000
		Interquartile Range	18.00000

Figure 3: Histogram – Total Purchase (dollars)

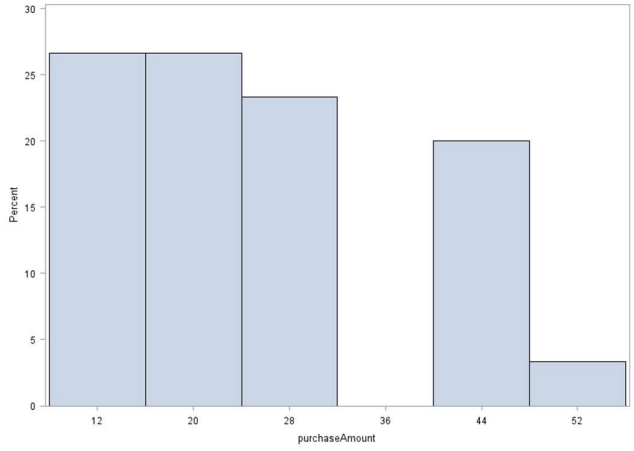

Table 3: Numerical Summary of Total Purchase (dollars)

Basic Statistical Measures			
Location		Variability	
Mean	25.94100	Std Deviation	12.19766
Median	22.97500	Variance	148.78300
Mode	.	Range	37.03000
		Interquartile Range	12.93000

Figure 4: Histogram – Age (years)

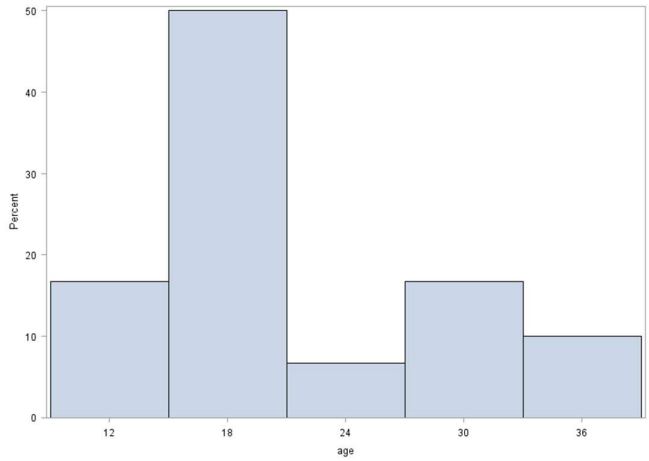

Table 4: Numerical Summary of Age (years)

Basic Statistical Measures			
Location		Variability	
Mean	20.90000	Std Deviation	7.24140
Median	18.50000	Variance	52.43793
Mode	18.00000	Range	24.00000
		Interquartile Range	11.00000

Figure 5: Histogram – Number of Items

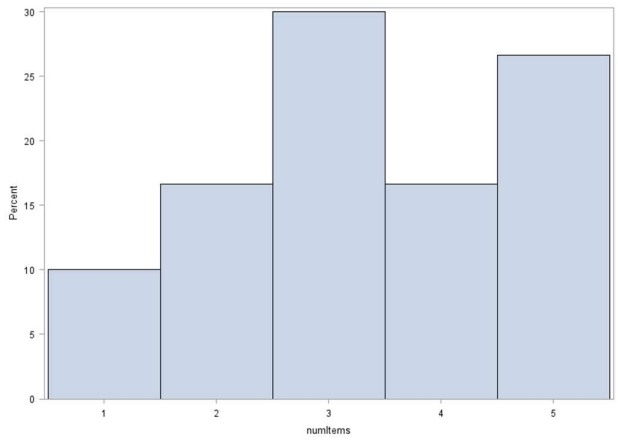

Table 5: Numerical Summary of Number of Items

Basic Statistical Measures			
Location		Variability	
Mean	3.333333	Std Deviation	1.32179
Median	3.000000	Variance	1.74713
Mode	3.000000	Range	4.00000
		Interquartile Range	3.00000

Figure 6: Histogram – Number of Visits

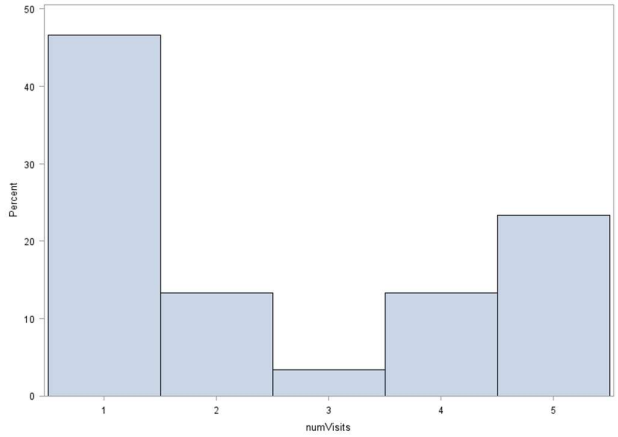

Table 6: Numerical Summary of Number of Visits

Basic Statistical Measures			
Location		Variability	
Mean	2.533333	Std Deviation	1.71672
Median	2.000000	Variance	2.94713
Mode	1.000000	Range	4.00000
		Interquartile Range	3.00000

Figure 7: Dendrogram – Cluster Analysis

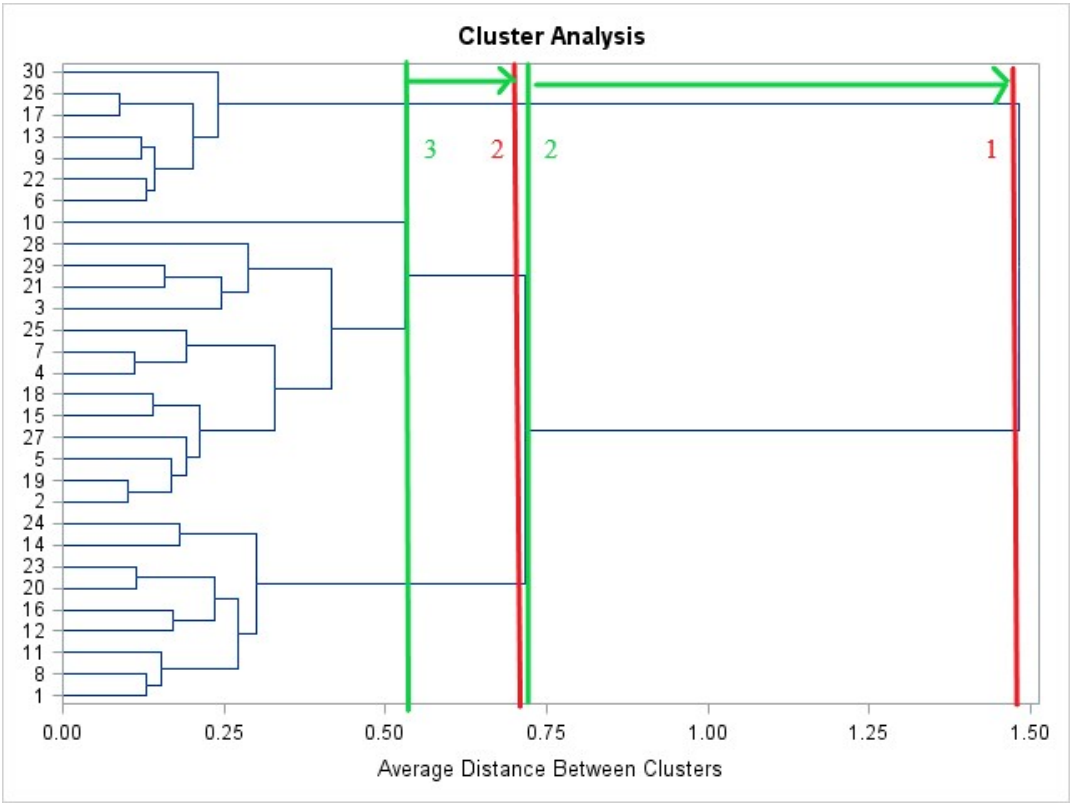

Table 7: The MEANS Procedure for Configuration with Three Clusters

CLUSTER	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
1	7	timeInStore	7	8.2857143	1.9760470	5.0000000	11.0000000
		purchaseAmount	7	45.9600000	2.3050380	42.9800000	48.8700000
		age	7	32.4285714	1.7182494	30.0000000	35.0000000
		numItems	7	1.8571429	0.6900656	1.0000000	3.0000000
		numVisits	7	1.0000000	0	1.0000000	1.0000000
		comicBookAmount	7	0	0	0	0
2	14	timeInStore	14	35.0714286	3.4743961	30.0000000	40.0000000
		purchaseAmount	14	16.7407143	3.9786323	11.8400000	24.4700000
		age	14	16.5000000	4.1648714	11.0000000	27.0000000
		numItems	14	4.0714286	0.9168748	3.0000000	5.0000000
		numVisits	14	4.0714286	1.2688145	1.0000000	5.0000000
		comicBookAmount	14	6.4478571	4.8035983	0	15.4900000
3	9	timeInStore	9	20.7777778	2.8185891	17.0000000	25.0000000
		purchaseAmount	9	24.6822222	3.0088901	19.9200000	28.2900000
		age	9	18.7777778	2.1081851	15.0000000	22.0000000
		numItems	9	3.3333333	1.3228757	1.0000000	5.0000000
		numVisits	9	1.3333333	0.5000000	1.0000000	2.0000000
		comicBookAmount	9	1.2155556	3.1850201	0	9.6300000

Table 8: The MEANS Procedure for Configuration with Two Clusters

The SAS System

The MEANS Procedure

CLUSTER	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
1	7	timeInStore	7	8.2857143	1.9760470	5.0000000	11.0000000
		purchaseAmount	7	45.9600000	2.3050380	42.9800000	48.8700000
		age	7	32.4285714	1.7182494	30.0000000	35.0000000
		numItems	7	1.8571429	0.6900656	1.0000000	3.0000000
		numVisits	7	1.0000000	0	1.0000000	1.0000000
		comicBookAmount	7	0	0	0	0
2	23	timeInStore	23	29.4782609	7.8036679	17.0000000	40.0000000
		purchaseAmount	23	19.8482609	5.3245210	11.8400000	28.2900000
		age	23	17.3913043	3.6274099	11.0000000	27.0000000
		numItems	23	3.7826087	1.1263990	1.0000000	5.0000000
		numVisits	23	3.0000000	1.7056057	1.0000000	5.0000000
		comicBookAmount	23	4.4004348	4.9133505	0	15.4900000

Table 9: The ANOVA Procedure using Comic Book Amount

Dependent Variable: comicBookAmount

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	253.8974387	126.9487194	8.99	0.0010
Error	27	381.1240579	14.1157058		
Corrected Total	29	635.0214967			