Dollarama Market Screening and Site Evaluation

Shakeeb Tahir

SA8912 - Spatial Technology in Strategic Planning

Toronto Metropolitan University

Dr. Shuguang Wang

February 28, 2024

Table of Contents

Introduction

The focus of this assignment is on market screening and site evaluation. Specifically, k-means cluster analysis will be conducted in order to do market screening for a new Dollarama store. As of 2020, Dollarama has 156 stores in the Toronto CMA. K-means cluster analysis using census variables will be able to tell us where “high” demand for a Dollarama store likely exists. The census variables used in the k-means analysis will all be variables that are associated with low income as that demographic is more likely to shop at Dollarama due to their low prices. The next part of the assignment will focus on site evaluation. This will be done for a Canadian retail store that operates 48 stores in the six major CMA’s of Toronto, Vancouver, Montreal, Ottawa-Gatineau, Edmonton and Calgary. In order to conduct the site evaluation, a multiple variable regression analysis will be used. Multiple independent variables will be used to evaluate their relationship with the average annual sales of the store. The independent variables that will be used represent the store attributes and trade area characteristics. This includes variables such as store size, number of direct and secondary competitors, average household income, etc. The multiple regression will compare two models, the first using the regular unstandardized variables, and the second using variables that have been standardized into log form. The two models can then be compared and evaluated using measures such as r-squared, f-score (with significance), standard deviation, range, etc., to see which model is more suitable and robust.

Cluster Analysis Literature Review

Cluster analysis is a powerful statistical technique that can be used as a way to group objects with similar characteristics together. This is particularly useful as it can help us to uncover and understand hidden patterns and relationships within the data. Cluster analysis is not just used in the realm of statistics, but can be applied to a variety of different disciples such as marketing, finance, biology, etc. This literature review will attempt to provide a comprehensive overview of cluster analysis, a brief history, some of its applications, and the strengths and weaknesses it offers.

To begin, cluster analysis has a long history which can be traced all the way back to 1939 when Robert Tryon published a book on it (Wilmink & Uytterschaut, 1984). However, it wasn't until the 1960's where a rapid interest in cluster analysis began, largely due to the invention of computers which allowed cluster analysis to be performed on much larger datasets than before (Blashfield & Aldenderfer, 1978). There are a variety of different cluster analysis algorithms that can be used, however two of the most common types of clustering methods used are hierarchical and k-means clustering (Wilmink & Uytterschaut, 1984). Hierarchical clustering groups items that are similar to each other and gradually combines these groups into larger clusters, forming a tree-like structure called a dendrogram. Each item is initially treated as an individual cluster and then the distance or similarity between each pair of clusters is calculated. The number of clusters can then be determined by cutting the dendrogram at points where there are gaps between the groups (Bridges, 1966). K-means clustering on the other hand receives a user-defined number of clusters and then defines a center point for each

cluster and groups points by minimizing the distance between the point to a center. The initial centroids for the clusters are randomly selected and then after each data point is assigned to a centroid, a new centroid is calculated using the average of the data points within that cluster. This process is repeated until there are no longer any significant changes to the centroids positions or after a fixed number of iterations defined by the user (Steinley, 2006).

To continue, cluster analysis has a variety of different applications and is used by many different disciplines. For instance, cluster analysis is a commonly used method when investigating climate change. Researchers will use cluster analysis in order to find the patterns and trends associated with things such as water levels or temperature fluctuations (Scitovski et al., 2021). By being able to group and rank these things it allows researchers to identify areas which are more affected by climate change and which need more attention or resources and should be prioritized above others. Another commonly used application of cluster analysis is in the medical field. Cluster analysis helps researchers in recognizing shapes, positions, and dimensions of human organs, as well as grouping medical data (Scitovski et al., 2021). In relation to strategic planning, cluster analysis helps analysts and planners conduct market screening and site selection. By running a cluster analysis, businesses can identify areas where their main customer base is located and tailor their strategic plans accordingly. This includes things like opening new stores in certain areas where they are most likely to profit. Cluster analysis allows businesses to group geographical areas based on similarities in customer demographics, behaviours, and preferences. By identifying these groups,

companies can gain insights into the concentration of their customers and allocate resources more effectively by tailoring marketing strategies to specific clusters (Prayag et al., 2012).

Furthermore, one of the main advantages of cluster analysis in contrast to other methods such as discriminant analysis is that a group structure does not need to be known prior to running the analysis (Wilmink & Uytterschaut, 1984). This flexibility is particularly advantageous in situations where researchers may not have a clear understanding of the underlying patterns or groupings within the data. As such, cluster analysis is a great method to use when dealing with large data sets as it can help simplify complex data into easier to understand sub groups. One of the reasons cluster analysis is a very popular method in strategic management is due to its ability to include multiple different variables in its analysis and then group them such that the statistical variance among the elements grouped together is minimized while between group variance is maximized (Ketchen & Shook, 1996). Prior to cluster analysis, other analysis techniques only allowed groups to be defined using a narrow set of variables, often one or two. This was too limited of an approach to capture the multidimensionality that was often involved in strategy research (Ketchen & Shook, 1996). Although cluster analysis has many benefits and use cases, it also has quite a few limitations and disadvantages. One of the most difficult problems facing cluster analysis is the assessment of the stability and validity of the clusters found by the numerical technique used (Everitt, 1979). Questions such as “Do the same types [of groups] emerge when new variables are used?”, “Do the same types [of groups] emerge when a new sample of similar

individuals is used?", "Do the members of different groups differ on variables other than those used in deriving them?" need to be asked (Everitt, 1979). Another important problem of cluster analysis has to do with choosing which clustering method to use. As there are numerous methods one can choose from, it has led to the problem of users questioning which is the best to use amongst them as the effectiveness of different methods can vary across a variety of data sets (Everitt, 1979).

Overall, even though cluster analysis may have some disadvantages or problems, it is still a very versatile and powerful statistical technique in data analysis that has a long history. As cluster analysis has been around since the 1930's and emerged throughout the 1960's with the invention of computers, it has become a tried and tested analytical tool. From being applicable to a variety of use cases such as analyzing trends and patterns in climate change to segmenting customers based on their purchasing behavior, cluster analysis has proven to be a useful tool across many diverse fields.

K-Means Cluster Analysis

The k-means clustering analysis was first conducted using the unweighted and unstandardized variables within the data set. These variables are as follows:

Variable	Variable Description	Assumption
im_11_16	Recent immigrants who arrived between 2011 and 2016	Tend to have low income
unemployed	Number of unemployed persons	Low/reduced income
renter	Home renter	Usually with low income
Avg_val	Average value of dwellings	Usually households with low income living in dwellings of low value
subsid	% of tenant households in subsidized housing	With low income
med_hh_inc	Median household total income $	
Hh_under_50k	Number of private households with household income below $50K	Low income household

Table 1: 2016 Toronto CMA Census Variables Used

In order to evaluate the cluster quality, a two-step cluster analysis was performed on cluster sizes of 4, 5, and 6. After this the variables were then weighted and standardized. A weight of 0.25 was applied to “im_11_16” and “Hh_under_50k” while the rest of the variables were given a weight of 0.10. This weight totals to 100 percent as 0.25+0.25+(0.1*5) = 1. A higher weight was given to “im_11_16” and “Hh_under_50k” as I believe they are more likely to represent areas where people have lower income than the other variables. The unemployed variable does not necessarily

mean an area is low income as the people in the area could have just temporarily lost their jobs for a variety of reasons when the census was taken and as such would not be more likely to shop at Dollarama. Renter and average value of dwellings are also not necessarily a great indicator of income as the Toronto CMA has very high housing prices and many people are forced to rent despite making high income. As such many people earning a medium to high income are probably also likely to live in areas with a lower average value of dwelling as they are not able to afford the extremely high prices in the main parts of the city. An area with a higher recent number of immigrants is much more likely to have a lower income population who would likely shop at Dollarama as they are still new to Canada and it takes a while for immigrants to establish themselves before they can afford higher quality products. The number of private households under $50,000 is also a definitive statistic that is likely not misleading like some of the other variables and is likely to represent areas of low income. As such these 2 variables were given higher weights than the rest of the variables.

After weighting the variables, they were standardized using z-scores. The two step cluster analysis was then run again on the z-scored variables for 4, 5, and 6 clusters. After viewing the results of the two step cluster analysis for the unweighted plus unstandardized and weighted plus standardized variables, the weighted and standardized variables with a cluster size of 4 showed the best cluster quality.

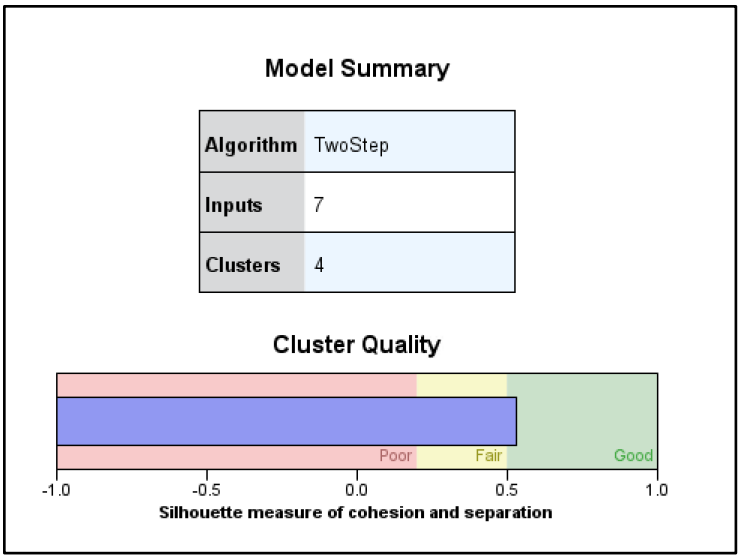

Figure 1: Best Cluster Quality (Weighted + Z-Scored variables with 4 clusters)

Based on figure 1, we can see that the cluster quality of the weighted and z-scored variables for 4 clusters falls into the “good” category. All the other two step clusters that were run for unweighted plus unstandardized and weighted plus standardized variables had a cluster quality in the “fair” range.Since this had the best cluster quality, a k-means cluster analysis with 4 clusters was then run using the weighted + z-scored variables.

Final Cluster Centers

	Cluster			
	1	2	3	4
Zscore(im_11_16_weighted)	-.49279	-.13757	1.60243	-.28043
Zscore(unempl_weighted)	-.67265	-.02696	1.29638	-.22453
Zscore(renter_weighted)	-.43086	.18264	1.77858	-.54306
Zscore(avg_val_weighted)	2.20531	-.29904	-.48551	-.11430
Zscore(subsid_weighted)	-.56974	.99445	.09632	-.53254
Zscore (med_hh_inc_weighted)	1.63072	-.52315	-.88588	.25609
Zscore (hh_under50k_weighted)	-.65703	.22994	1.78229	-.52832

Figure 2: Final Cluster Centers of Weighted + Z-scored variables

Based on figure 2, we can see that the cluster that is the best for new locations of Dollarama stores is cluster 3, followed closely by cluster 2. This is because they have higher z-scores in variables such as recent immigration, households under 50k, unemployed, subsidized housing, and renter which indicates these are clusters where these variables are more prevalent in the population. They also have lower average value of dwelling and median household income z-scores which indicates that these are clusters which are below the mean for these variables and as such are more likely to be where the target demographic for Dollarama is located.

In order to compare these results to the unweighted and unstandardized variables, the best cluster quality for those variables, which was also a size of 4, was run.

Final Cluster Centers

	Cluster			
	1	2	3	4
IM_11_16	212	267	385	195
UNEMPL	167	209	242	141
RENTER	553	489	783	355
AVG_VAL	1212323	764602	482869	1976033
SUBSID	4.6	10.7	13.4	5.8
MED_HH_INC	107658	90443	69263	156062
Hh_under50K	453	491	714	289

Figure 3: Final Cluster Centers of Unweighted + Unstandardardized variables

Based on figure 3, we can see that we get the same results as the weighted and standardized variables. Cluster 3, followed by cluster 2 are shown to be the best for locations of Dollarama as they have the lowest median household income and average household dwelling values as well as the highest values for the other variables.

These clusters were then plotted on a map with current Dollarama locations and 5 suitable locations were found that would be the best for opening new Dollarama stores. These are census tracts which fall into cluster 3 and cluster 2 and do not have any nearby Dollarama locations.

Figure 4: Toronto CMA Dollarama New Location Suggestions

Multiple Regression Analysis Literature Review

Multiple regression analysis is a statistical technique used to examine the relationship between two or more independent variables and a dependent variable. It is an extension of simple linear regression in which there is only one independent variable. Multiple regression analysis is a very useful tool as it allows us to understand how much of the variance in a dependent variable can be explained by the independent variables. This literature review will attempt to provide a comprehensive overview of multiple regression analysis, a brief history, some of its applications, and the strengths and weaknesses it offers.

To begin with, the modern notion of regression analysis is known to have originated around the late 1800's by Sir Francis Galton who was a British polymath and the cousin of Charles Darwin. Galton had an interest in genetics and introduced the term "regression" in the context of inheritance studies, where he observed that the children of parents who were tall, tended to regress closer to the average height of the general population. This led to the formulation of the concept of "regression toward the mean," which suggests that extreme values in a set of data are likely to be followed by values that are closer to the average (Barnes, 1998). Galton's findings were later expanded on by Karl Pearson who invented the Pearson Product Moment Correlation (PPMC), better known as the Pearson's correlation coefficient, which quantifies the strength and direction of a linear relationship between variables (Stanton, 2001). Since then regression analysis has been expanded on and there are now multiple different

methods of regression such as simple linear regression, multiple regression, logistic regression, polynomial regression, etc.

In terms of multiple regression analysis, it is simply using multiple explanatory variables to predict the outcome of a response variable. When performing multiple regression, there are a variety of things to consider and conditions that must be met in order to determine if the model is robust. The formula for multiple regression is as follows:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$$

Figure 5: Multiple regression formula

The y represents the dependent variable, x1…xn represent the independent variables, B0 is the value of y when all independent variables are equal to 0, B1…Bn are the regression coefficients, and ε represents the error term (Uyanık & Güler, 2013). The regression coefficients are most commonly estimated using ordinary least squares (OLS) which attempts to minimize the sum of the squared residuals. In order to run multiple regression analysis, it is assumed that the data is normally distributed, has linearity, freedom from extreme values, and that there are not multiple ties between independent variables (Uyanık & Güler, 2013). To evaluate the results of the regression model, there are a few things to consider. The R^2 value tells us the correlation of the

model and how much of the variance of the dependent variable can be explained by the independent variables. The skewness statistics tells us if the data is close to normally distributed or if it is asymmetric. A value of 0 tells us that the data is perfectly normally distributed, positive values indicate it is right skewed as the the right tail of the distribution is longer than the left tail, and negative values indicate left skewness. Usually, if the skewness is less than -1 or greater than 1 then it is considered highly skewed (Gupta et al., 2019). Kurtosis on the other hand tells us about the peakedness of a distribution. A mesokurtic distribution is considered to be similar to a normal distribution and is when the kurtosis value equals 0. A leptokurtic distribution is when the kurtosis value is greater than 0 and the tails of the distribution are longer which indicates there are more extreme values in the data while a platykurtic distribution is when the kurtosis statistic is less than 0 and the tails are much thinner which indicates that there are few extreme values and the data is more spread out (Gupta et al., 2019). If the data is not normally distributed then it must be normalized by logging the values. In order to know which variables in the model are significant we look at the p-value. P-values of less than 0.05 are usually considered significant and any variables with higher values are eliminated from the model. The durbin watson statistic tells us if there is any autocorrelation in the residuals. If the statistic is outside of the 1.5-2.5 range then the residuals are considered to be autocorrelated. The VIF value tests for multicollinearity and a VIF greater than 10 indicates that there is multicollinearity between the variables (Slinker & Glantz, 2008). Lastly, the root mean square error (RMSE) tells us the average difference between the model's predicted values and the actual values.

In terms of applications, multiple regression is used across many fields as it is a great way to understand the relationship between multiple independent variables and a dependent variable. In fields such as marketing, it helps assess the impact of multiple variables on consumer behavior and purchasing decisions, while in the medical field multiple regression can be applied to study the influence of multiple factors on patient outcomes or disease progression. Other fields such as the social sciences use it to understand the relationships between various socio-economic, demographic, and psychological variables (Paul & Das, 2023). There are many different advantages and disadvantages of multiple regression. It is considered to be both a very simple and effective tool at establishing relationships between variables, it is able to provide a quick benchmark for more advanced methods, and it works reasonably well even when the model is not perfectly specified (Verbeek, 2017). However, it is not without limitations as the specification of the model is not always straightforward because there is no simple, hard rule on how to choose an appropriate specification (Verbeek, 2017). Also, interpreting a linear regression model as a causal relationship is challenging and requires strong assumptions as it is often difficult to establish empirically (Verbeek, 2017).

Overall, multiple regression is a great statistical tool that can be used to explore the relationship between a dependent variable and multiple independent variables. By running a multiple regression we can establish models that allow us to understand the combined impact of various factors on a variable and eliminate any factors which are not significant in helping us predict our dependent variable.

Multiple Regression Analysis

Model	F-Score (with sig.)	R^2	Average Error	Minimum Error	Maximum Error	Range	Standard Deviation
Model 1	14.054 (sig <0.001)	0.811	2,118,858	-6,587,820	8,018,048	14,605,868	2,856,528
Model 2	1.938 (sig 0.116)	0.587	2,356,041	-5,265,723	9,272,593	14,538,316	3,335,369

Table 2: Multiple Regression Model Comparison

Independent Variable	**Model 1 Significance Level**	**Model 2 Significance Level**
Store Size	0.001	0.133
Number of direct competitors in trade area	0.237	0.898
Number of secondary competitors in trade area	0.593	0.897
Total number of households	0.013	0.197
Population of 25-45 years of age	0.749	0.651
Average household income	0.063	0.303
Number of home owners	<0.001	0.002
Number of persons working at home	<0.001	0.024
Number of persons who moved in the past 5 year	0.105	0.280
Number of new homes constructed between 2011 and 2016	0.847	0.771
Number of recent immigrants	0.713	0.685

Table 3: Significance Level for coefficients of independent variables

Based on table 2, we can see that both of the multiple regression models were significant as they have a significance value <0.001 and normally a value <0.05 is considered statistically significant. The R^2 value of the model that was not logged (model 1) is higher which indicates that the unstandardized variables were better at explaining the variance of the average annual sales. Model 1 also has a lower standard deviation which indicates that there is less variation within the model and model 2 likely has more extreme values. In Table 3, we can see that many of the independent variables are statistically significant and should be eliminated from the model. The only significant independent variable for model 2, based on an assumed alpha of 0.05 is number of home owners, while number of home owners, number of persons working at home, store size, and total number of households are statistically significant for model 1.

Conclusion

In conclusion, we can see how cluster analysis is a useful technique in order to conduct market screening for potential store locations and how multiple regression analysis is a useful technique for retail site evaluation. Cluster analysis helps us identify potential locations within the study area that are suitable for building new Dollarama stores as the census tracts are grouped into clusters based on low income based census variables. By knowing which census tracts have more people living in them with these low income attributes we can choose potential locations to build Dollarama's as the people in these areas are more likely to be customers as they fit Dollarama's target demographic. Furthermore, by conducting a multiple regression analysis for site evaluation, we can identify which store attributes are the best at explaining the variance

within the average annual sales of the store and can base decisions off of it. For instance, if a model shows that average annual sales are most influenced by population 25-45 years of age, then the retail chain could build more stores in areas with that target demographic or it could conduct internet marketing targeted towards that age group. Overall, cluster analysis and multiple regression analysis are both useful techniques that can help make businesses make logical and informed decisions.

References

Barnes, T. J. (1998). A history of regression: Actors, networks, machines, and numbers. *Environment and Planning A: Economy and Space*, *30*(2), 203–223. https://doi.org/10.1068/a300203

Bridges, C. C. (1966). Hierarchical Cluster Analysis. *Psychological Reports*, *18*(3), 851–854. https://doi.org/10.2466/pr0.1966.18.3.851

Blashfield, R. K., & Aldenderfer, M. S. (1978). The literature on Cluster Analysis. *Multivariate Behavioral Research*, *13*(3), 271–295. https://doi.org/10.1207/s15327906mbr1303_2

Everitt, B. S. (1979). Unresolved Problems in Cluster Analysis. *Biometrics*, *35*(1), 169–181. https://doi.org/10.2307/2529943

Gupta, A., Mishra, P., Pandey, C., Singh, U., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia*, *22*(1), 67. https://doi.org/10.4103/aca.aca_157_18

Jeffrey M. Stanton (2001) Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors, Journal of Statistics Education, 9:3, , DOI: 10.1080/10691898.2001.11910537

Ketchen, D. J., & Shook, C. L. (1996). The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. *Strategic Management Journal*, *17*(6), 441–458. http://www.jstor.org/stable/2486927

Paul, R., & Das, K. N. (2023). Trends of optimization algorithms from supervised learning perspective. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewjcce32021049

Prayag, G., Landré, M., & Ryan, C. (2012). Restaurant location in Hamilton, New Zealand: Clustering patterns from 1996 to 2008. *International Journal of Contemporary Hospitality Management*, *24*(3), 430–450. https://doi.org/10.1108/09596111211217897

Scitovski, R., Sabo, K., Martínez-Álvarez, F., & Ungar, Š. (2021). *Cluster Analysis and Applications*. https://doi.org/10.1007/978-3-030-74552-3

Slinker, B. K., & Glantz, S. A. (2008). Multiple linear regression. *Circulation*, *117*(13), 1732–1737. https://doi.org/10.1161/circulationaha.106.654376

Steinley, Douglas. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 1–34. https://doi.org/10.1348/000711005x48266

Uyanık, G. K., & Güler, N. (2013). A study on multiple linear regression analysis. *Procedia - Social and Behavioral Sciences*, *106*, 234–240. https://doi.org/10.1016/j.sbspro.2013.12.027

Verbeek, M. (2017). Using linear regression to establish empirical relationships. *IZA World of Labor*. https://doi.org/10.15185/izawol.336

Wilmink, F. W., & Uytterschaut, H. T. (1984). Cluster analysis, history, theory and applications. *Multivariate Statistical Methods in Physical Anthropology*, 135–175. https://doi.org/10.1007/978-94-009-6357-3_11