

Customer Demographic Profiling

Katherine Fair
Matthew St. Peter
Holly Westerfield

April 27, 2007

Abstract ACO Hardware has contracted with ADVO to obtain detailed consumer spending profiles from a variety of census and economic data by zip code. Each profile is characterized by average household size, age breakdowns, economic conditions, education levels, and marital status. The focus of this project is to compare the gathered data with actual spending habits and dollars spent to determine a “successful customer profile” using a least-squares regression. The goal is to have a predictive model that uses the best demographic mix to determine profitable new store locations.

Table of Contents

Introduction.....	1
Phase I Revisited.....	1
Customer Targeting.....	2
Demographic Mix.....	4
Demographic Marketing.....	7
Conclusion.....	10
Future Work.....	10
References.....	11
Appendix.....	12
I. Demographic Profiles	12
II. Least-Squares Regression	15
III. Cluster Analysis	17
IV. Expected Values	18

Introduction

ACO currently operates 69 retail locations throughout Michigan, primarily situated in the metro Detroit area, though stores span as far east as Battle Creek and as far north as Bay City. Like many retailers, the company is in a transitional period. The expansion of “big box” hardware stores such as Home Depot and Lowe’s has created a new burden to remain competitive. In the face of this challenge, ACO has chosen to occupy a smaller niche, preferring to cater to the small home improvement market rather than compete directly with the larger retailers.

As part of their plan to remain a strong retail operation, ACO is in the process of a multi-faceted business review process, attempting to identify the factors that will contribute to continued success and expansion. A large amount of financial data, combined with site, demographic, traffic, consumer behavior and competition data has been gathered. In Phase I, traffic and competition data were examined in an effort to build a “successful store profile” to be used as a basis for expansion.

Now the focus has been turned toward demographic and consumer behavior data in order to build a “successful customer profile.” This profile will determine the type of customer that ACO should court in order to maximize sales and to further determine the best locations for expansion.

The process by which a successful customer can be determined is through the least-squares method. Numerical analysis can be utilized in determining a successful customer profile. The method of least-squares allows for an examination of the demographic data with store sales.

Phase I Revisited

In the first phase of the ACO project, a model was developed that focused on the success of a location based on geographic factors. However, the model generated in Phase I may be reworked to determine customer count as opposed to annual gross profit.

With the profit data and customer counts from fiscal year 2007 in hand, a correlation between the two may be calculated. The correlation is not surprising – more customers should naturally lead to higher profit. What is surprising is the strength of the correlation: initial models have an R^2 value of around 0.85, with each customer valued at around \$5. An R^2 value describes the goodness of the fit of a model. R^2 values close to 1 indicate a good model. Thus, the ten-factor model may be roughly translated to one that considers a customer count response by merely dividing the coefficients by five. In Tables 1 through 4 below, the factors of the rough customer model are shown. The coefficients have been rounded to the nearest tenth.

Table 1. Factors given continuously.

Factor Name	An additional...	Increase in Customer Count
Square footage of store	Square foot of retail space	10
Average \$ spent in repair	Average dollar spent per year on household repairs and maintenance	7.4
Sum of passing traffic	Car passing in front of store	0.1

Table 2. Yes / No Factors. The change in customer count is given if the factor carries a “Yes” value.

Factor Name	Change in Customer Count
Store lies on a “trunk line”	17,804.8
There is a grocery store within shopping center	- 23,877.8
There is a drug store within shopping center	- 14,229.6
There is a Home Depot within 4 miles	13,997.4
There is an ACE Hardware within 4 miles	2,915.2

Table 3. Visibility Rating.

Visibility Rating	Change in Customer Count
Good	13,990.4
Fair	17,969.8
Poor	0

Table 4. Directives Rating.

Ability to Follow Directives	Change in Customer Count
Good	13,302.8
Fair	4,965.2
Poor	0

Such a modification allows Phase II to better interpret Phase I. Now, instead of focusing on money, the model focuses on *people*, who may fall into distinct categories that may be separately analyzed. In Phase I, for example, it was determined that a grocery store in the same shopping center as an ACO Hardware location causes a loss of nearly \$120,000 in annual gross profit. Now it may be said that the same grocery store causes a loss of nearly 24,000 customers over the course of a year. But who are they?

Customer Targeting

Customer targeting is a well-established practice. Much information is available on the methods and models of customer targeting; however, the “foot work” of collecting the information necessary occurs prior to any implementation of a model, and accounts for the vast majority of total cost of implementation. Fortunately, much of this “foot work” has already been done by ADVO, a national distributor of shopping advertisements. ADVO has prepared a list of customer profiles and buying power indices that occur throughout the nation.

The buying power indices calculated by ADVO are designed to measure the propensity to purchase goods of a particular nature. A score is assigned as a percentage ratio, so a score of 100 is considered to be average, while a score above 100 indicates a higher ability to make a purchase. However, these are of little to no value as currently given. The buying power indices are calculated against a national average, while all ACO Hardware locations are located within southwest Michigan. Some of the variance within this local region is lost when considering a national average.

To more clearly illustrate this concept, simple scatterplots have been generated regression lines have been fit below in Figure 1. ADVO index scores are given as independent values and sales are given as dependent values for four separate departments, in an effort to ascertain any relationship between the index values and actual sales.

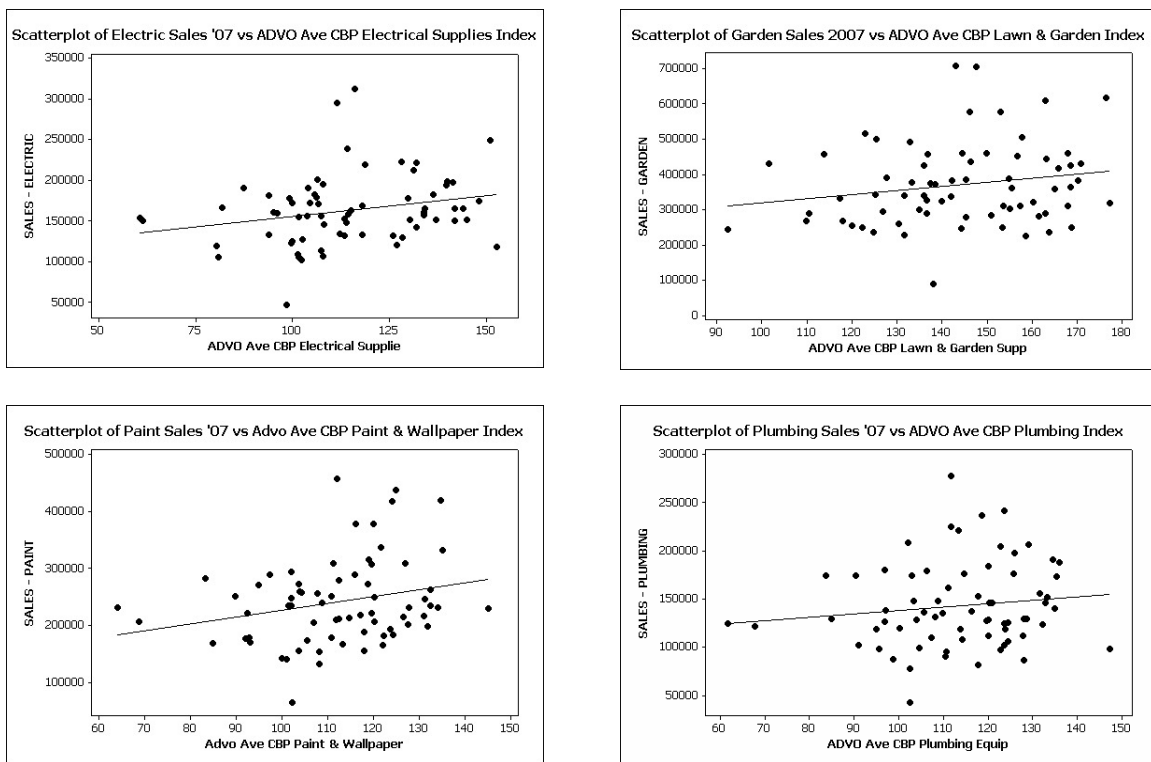


Figure 1. ADVO Indices versus actual sales. Note the low correlations.

An ideal graph would show the scattered points arranged in a nearly linear fashion, with a steeply sloped regression line. The graphs in Figure 1 show that there is, in fact, little to no relationship between the index values supplied and actual observed yearly sales. The points are arranged randomly and the regression lines have no slope, hence no explanatory power.

Sole reliance on the ADVO indices as a method of customer targeting is not fruitful; a better way must be sought. The raw data provided by ACO and ADVO also includes 53 demographic profiles that occur throughout zip codes across the nation.

These profiles may be used in order to better explain yearly store sales. First, it is useful to determine how much each profile is worth.

Demographic Mix

The data supplied by ADVO includes demographic distributions of 53 profiles for zip codes throughout the United States. However, in the limited subset of zip codes that surround ACO retail locations, not all profiles exist in all zip codes. In fact, there are ten profiles that have zero representation across all ACO Hardware stores. These profiles are removed prior to any further analysis. Then, there are a number of profiles that have insignificant representation. After removal and analysis of these, only 19 profiles remain. A list of these profiles is included in Appendix I.

However, the analysis must be done on actual stores, not on zip codes. Thus, the demographic distributions for a number of zip codes surrounding each store must be accounted for. Typically, each store has between four and six zip codes immediately surrounding it that account for between 75 to 95 percent of all sales made.

In Figure 2 below, a hypothetical store is broken down into four distinct zip codes Z_1 through Z_4 . The blank space on the far right represents the portion of sales that do not correspond to surrounding zip codes. In addition, each zip code has an associated demographic profile breakdown determined by ADVO. This breakdown is simplified to D_1 through D_4 .

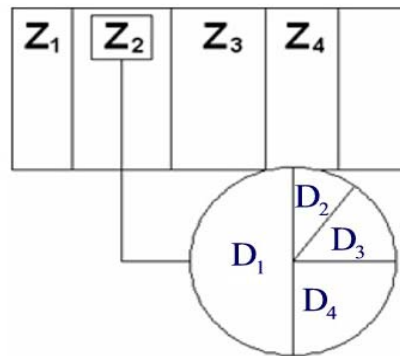


Figure 2. Space of customers broken down by zip code and demographic mix.

Further, nearly every transaction has an accompanying zip code recorded at the time of sale. By examining the transaction count for a particular zip code in relation to the total transaction count of a given store, it becomes possible to determine the effect that each surrounding zip code has on the overall demographic distribution of the store.

To arrive at a breakdown by store rather than by zip code, the percentage of transactions that come from each zip code are multiplied by their respective demographic distributions to arrive at reasonably accurate distribution for the store. For example, in

Figure 2, a certain percentage of total store sales may be found to originate from zip code Z_2 , say 20%. Then, in Z_2 , there is the demographic breakdown D_1 through D_4 , to which 20% of total store sales may be attributed. A multiplication of that percentage over the mix will yield a reasonable representation. So if

$$[D_1, D_2, D_3, D_4] = [0.5, 0.2, 0.2, 0.1]$$

the multiplication yields

$$20\% \times [0.5, 0.2, 0.2, 0.1] = [0.1, 0.04, 0.04, 0.02]$$

meaning that demographic D_1 in zip code Z_2 is responsible for 4% of total store sales. Now, by summing up the weighted zip codes, a total demographic breakdown by store is achieved.

With the demographic breakdown in hand for every store, the process of obtaining the expected value of each demographic may begin. These expected values are computed with the method used in Phase I – the least-squares method. In this case, the particular approach to solving the least-squares problem differs slightly from that of Phase I. The particular technique used to solve the least-squares problem is outlined in Appendix II, and the expected values generated for all 19 demographic profiles are in Appendix III. A snapshot of the results follow in Table 5.

Table 5. A snapshot of the calculated expected values.

Profile Name	Garden	Paint	Housewares	É	Total Sales
Town Council	3.6094	1.8954	1.7794	É	14.1822
Married with Homes	2.4284	1.7342	2.0327	É	12.9577
Suburban Society	3.0753	2.6375	0.9772	É	13.3804
Suburban Seniors	1.2715	0.6401	1.7958	É	9.4046
Suburban Success	3.3093	1.8701	1.8684	É	14.697
Suburban Starters	0.8666	2.091	0.429	É	9.3596
Senior Success	2.3446	1.9032	2.9042	É	13.9005
Hard Hats	2.2415	1.3263	1.7976	É	11.2132

The Total Sales column in Table 5 shows how much, on average, is sold to a customer of a given demographic each time a transaction occurs. So, while there may be a wide range of individual sale amounts from the “Suburban Seniors,” each additional transaction is expected to contribute about \$9.40 in total sales. Furthermore, each additional sale is expected to net an increase of about \$1.27 in Garden sales, \$0.64 in Paint sales, and so on.

The expected values calculated for each demographic profile may now be used to solve the forward problem; that is, project sales. Given two inputs; the number of transactions over an interval of time, and a demographic distribution for the surrounding area, the number of transactions may be multiplied across the demographic distribution to yield a distribution of customers. Each segment of the customer distribution may then be

multiplied by their respective expected value to yield an estimate of both total store sales and sales by department.

For example, if a prospective store has a percentage demographic distribution of

$$[20\% \quad 15\% \quad \dots \quad 2\% \quad 18\%]$$

and 5000 customers enter over a given week, then the number of customers may be broken down to their respective demographics by

$$\begin{aligned} D &= 5000 \times [20\% \quad 15\% \quad \dots \quad 2\% \quad 18\%] \\ &= [1000 \quad 750 \quad \dots \quad 100 \quad 900] \end{aligned}$$

which may then be combined with the expected values generated by solving the least squares problem to arrive at a sales projection. These computations may be automated using a Microsoft Excel spreadsheet with embedded formulas. A sample sales projection follow in Table 6.

Table 6. Recorded sales vs. projected sales for ACO Store #123 – Lansing, Frandor.

Department	Actual Sales	Simulated Sales	Difference Ratio
Paint	231,746.63	200,099.15	-13.66%
Tools	69,426.21	56,353.26	-18.83%
Electric	153,639.50	142,749.88	-7.09%
Plumbing	123,944.44	127,105.91	2.55%
Hardware	195,870.92	160,779.44	-17.92%
Housewares	167,681.95	204,907.03	22.20%
Garden	242,314.94	239,236.06	-1.27%
Sports	11,896.96	13,915.65	16.97%
Pet Supplies	25,949.28	33,795.74	30.24%
Seasonal	63,691.02	67,684.89	6.27%
Automotive	29,044.25	25,278.08	-12.97%
Gift	1,956.00	5,374.04	174.75%
Sundries	43,758.68	50,051.71	14.38%
Carpet Care	4,832.29	6,608.10	36.75%
Food	73,288.28	89,946.34	22.73%
Treasure Hunt	19,960.83	17,883.64	-10.41%
TOTAL SALES	1,459,002.18	1,441,768.94	-1.18%

There is a slight problem, however, with the projections based on expected values. Although the solution is generally very accurate for total store sales, it is less so for sales by department. The projected total sales are within a 10% error margin for most stores. On the other hand, the values for department sales exhibit a higher error than desired – most notably in the gift department in the figure above. This is caused by the relative unimportance of the smaller departments when considering yearly sales. Notice

that the departments that exhibit the largest relative error are also the departments that contribute the least to a store's total sales. So, while there may be a relative error of 174.75% in the gift department, the actual error in sales is less than \$3500. The computed expected values may now be used in further analysis.

Demographic Marketing

The expected values generated by the least-squares method may be used for far more than simple sales projections. They provide an insight into the spending habits of the population surrounding each and every store. An initial result is shown through the development of new indices, named the MSU indices, that indicate relative potential sales for various departments.

The indices are able to provide a "snapshot" view of a store's projected performance, relative to other stores, before a more rigorous analysis takes place. A score of 100 on the MSU scale indicates an average performance store in comparison to the other 68 stores, while a score above 100 indicates a higher expected performance in comparison with the other stores.

Using the expected values obtained through the least-squares method, the MSU indices are obtained by comparing the value for each profile with the value across the entire chain. Though the MSU indices computed do not have an extremely high correlation to the actual departmental store sales, they significantly outperform the ADV0 indices in every department.

Two examples are shown below as Figures 3 and 4; normalized sales for the garden department and for the plumbing department. Note that the range of the index scales are not the same: the ADV0 indices are developed by making comparisons across regions of the United States, while the ones developed by MSU are built on comparisons solely within the ACO Hardware region.

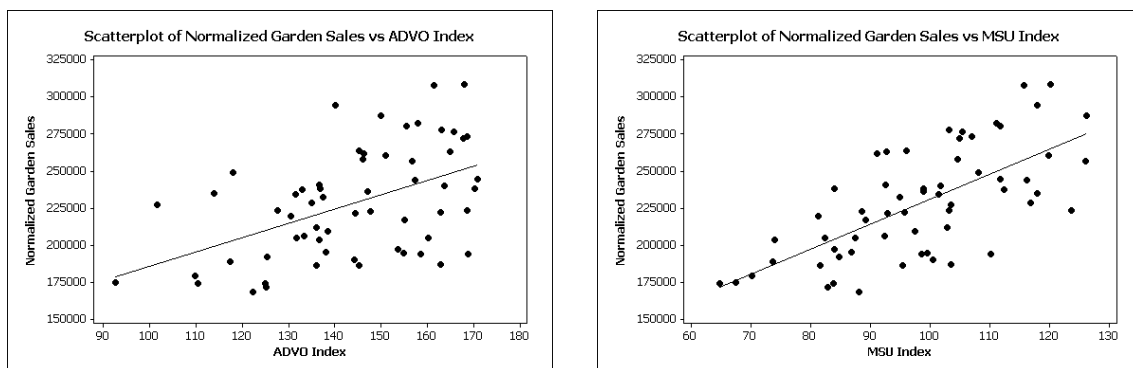


Figure 3. Comparison of given ADV0 index versus computed MSU index for predicting Garden sales.

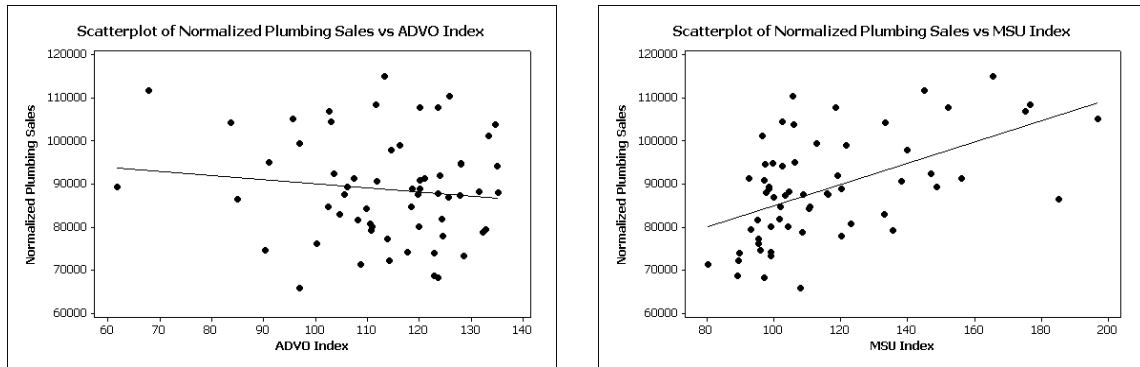


Figure 4. Comparison of given ADVO index versus computed MSU index for predicting Plumbing sales.

However, even given the improved index, it would be useful to group the departments together in a meaningful way, so any underlying commonalities may be exploited. For example, the profile “Suburban Starters” has a high expected value in the paint department. Are there any other departments for which “Suburban Starters” will also have a high expected value? If this question can be answered, then targeted advertising becomes a real possibility. Additionally, such insight can aid in the layout of stores.

In order to answer this question, a cluster analysis may be performed on the expected values obtained through the least squares method. Cluster analysis is used to build taxonomic trees, assigning each variable to a particular “cluster” if it shares some similarity with the other variables in that cluster. In marketing applications, cluster analysis assists in group identification and segmentation.

A distance metric must be employed by the clustering algorithm to ascertain which variables are “near to” or “far away” from each other. In this case, the correlation coefficients between every department are computed and analyzed as the distance metric. A precise mathematical statement of the method may be found in Appendix II. However, the clustering algorithm is automatically implemented in MINITAB, yielding the results in Figure 5 below.

The expected value is broken down by demographic for each department, allowing the most profitable demographic for each department to be identified. For example, each sale to the demographic profile “Suburban Starters” is expected to yield about \$1.92 in sales in the paint department. This information is useful, as ACO is currently introducing Benjamin Moore brand paint, a premium brand, into some of their stores. With the knowledge that the “Suburban Starters” profile is the most profitable demographic with respect to paint, ACO can identify stores with a high percentage of “Suburban Starters” shoppers and introduce Benjamin Moore into those stores.

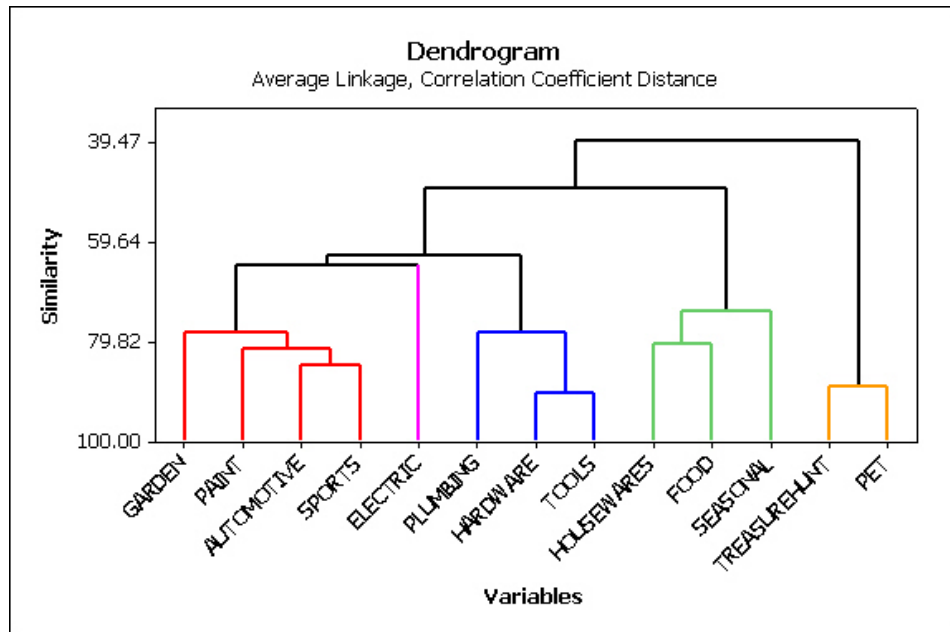


Figure 5. Expected values by department clustered into distinct groups.

Once the cluster analysis has been performed, it is up to the researcher to determine the commonalities that the clustering indicates. A cluster analysis can only group the variables, it can not determine what the relationship between elements in the same cluster. Human ingenuity is required.

In Table 7, we see four clusters with two or more elements. There is a single department, electric, not listed in the table below. This is due to the fact that the electric department has nearly the same distance from Cluster 1 as from Cluster 2.

Table 7. Clusters indicated in Figure 5 above with commonalities listed.

Cluster Number	Comprised of	Commonalities
1	Garden Paint Automotive Sports	Appearance-focused Activity Outdoors
2	Plumbing Hardware Tools	Handiness Work Less visible
3	Housewares Food Seasonal	Kitchen Convenience
4	Treasure hunt Pet supplies	Bargain Esoteric

The top five departments (based on sales) are garden, housewares, paint, electric, and hardware. Focusing on the top spenders per department allows the identification of important demographics. As seen in Table 8 in Appendix I, the three profiles of “Just Getting By”, “Lots of Tots”, and “Ethnic Elders” are clustered into a group called “On

the Bubble.” This group is expected to spend the most in the garden department. With lower median incomes, they are driven by price and function. “Power Players” is another cluster comprised of “Established Elite” and “Influential Elders” that spends the most in the electric department. “African American Success” spends the most in housewares. “Suburban Starters” are the most influential demographic in both the paint and hardware departments. This is logical since “Suburban Starters” are characterized as young, low-income homeowners with a high need for paint and hardware to make minor repairs to their new home. With this information, ACO will be able to analyze demographic breakdowns around individual stores and decide what departments could be expanded. The marketing department can determine what products to put on sale to attract these profiles into the store.

Conclusion

The Phase I model may be modified to fit a customer count response rather than an annual gross profit response with virtually no loss in explanatory power. Then, given a demographic distribution surrounding a store, the least-squares method may be employed to determine what each customer profile is worth per transaction. These expected values may be combined into various metrics that assist ACO when determining store location, store layout, and product selection.

- It is naïve to claim that there exists one demographic that is clearly superior to all of the rest. Every store has “demographic weaknesses” when compared to other stores.
- Since the vast majority of store sales can be determined by customer count alone, the demographic analysis does not provide ACO Hardware with tools to determine optimum store location based purely on demographics. A store with a favorable demographic distribution but with very few households near, and hence very few projected customers, is destined to fail.

Rather, the analysis serves to detail which demographics are superior in certain settings, so that ACO may tailor each store to its surrounding distribution in order to better lure customers.

Future Work

ACO management has been analyzing data surrounding unemployment rates and their influence on store profits. This is a possible future project that would provide another method for ACO to determine strong store locations and future store sites.

Other possible projects include a study of how seasonal changes affect store profit, and if there is any link to weather patterns. This project might have some correlation to geographic location and customer demographics, as it seems unreasonable that customer from a given demographic is worth the same amount throughout the entire

year. A seasonal approach should further refine the results discussed here, thus building upon both Phase I and II.

References

Berry, Michael J. A. and Linoff, Gordon. Data Mining Techniques for Marketing, Sales, and Customer Support. New York: John Wiley and Sons, Inc, 1997. 1-5.

Cabena, Peter, et al. Discovering Data Mining. Upper Saddle River: Prentice Hall, 1998. 42-46.

Hallberg, Garth. All Consumers are Not Created Equal. New York: John Wiley and Sons, Inc, 1995. 1-5.

Kamakura, Wagner and Wedel, Michel. Market Segmentation. 2nd es. Boston: Kluwer Academic Publishers, 2000. 1-5.

Ratner, Bruce. Statistical Modeling and Analysis for Database Marketing. London: Chapman and Hall, 2003. 1-5.

Sabor, Michael, Silva, Ana Rita, and St. Peter, Matthew. Geographic Determination of a Successful ACO Hardware Store. Michigan State University, 2006.

Appendix I – Demographic Profiles

Town Council	Older, town couples with & without children. Age 45+. College graduates; employed in a variety of blue & white collar jobs. Median household income approximately \$59,250. Mid-market shopping behavior, driven by value & function.
Affluent Asian Families	Rich, middle-aged, suburban families with children. Highly educated, they are employed in professional, management & Federal government jobs. Median household income \$108,000+. Home is owner occupied. Predominantly Asian; 45-64 years of age. Their upscale shopping behaviors are driven by service & comfort.
Affluent Town Boomers	Upscale, boomer homeowners living in smaller towns. Predominantly Asian & white; age 35-54. Mostly married, mix of households with and without children. College+ education; employed in well-paying, white collar occupations. Median household income of over \$73,700. Upscale shopping behaviors, driven by service & comfort.
Affluent Town Families	Affluent, mobile town families with children. Predominantly white, age 35-54. They are very well educated and are employed in a variety of well-paying white collar occupations. They possess high median incomes of over \$102,000. Their upscale shopping behaviors are driven by service & comfort.
African American Success	Mix of African-American singles & families with children, living in suburbia. Mostly homeowners. Age 45-64. Some college; employed in decent paying blue & white collar jobs. Median household income is approximately \$59,100. Mid-market shoppers driven by service & comfort.
Country Boomers	Exurban homeowners. Mix of married couples with & without children. Age 45-64. High school/some college; employed in a variety of well-paying, blue collar occupations. Median household income of approximately \$54,300. Discount shoppers, driven by price & function.
Country Success	Upscale, exurban homeowners. Predominantly white; age 45-64. Mostly married, mix of households with and without children. College+ education; employed in well-paying, white collar jobs. High median household incomes of approximately \$80,400. Their mid-market shopping styles are driven by service & function.
Established Elite	Prosperous suburban families with children. Median household income \$165,000+, highly educated professionals & executives. Home is owner occupied. Predominantly white and Asian; 45-64 years of age. Upscale shopping behavior with service & comfort purchasing triggers.
Ethnic Elders	Disadvantaged, older African-Americans living in their own suburban homes. Income <\$28,000. Elementary/some high school education; few high school graduates. Age 55+. Those still working are employed in blue collar & service occupations. Mid-market shoppers, they are driven by value & function.
Ethnic Success	Suburban, ethnic blend of couples with & without children. Mostly age 25-44. Ethnically diverse with a very strong Asian presence. Well educated; employed in a variety of white collar occupations. Median household income approximately \$61,200. Upscale shopping behavior, driven by service & style.
Golden Years	Aging, white empty nesting couples. Age 55+, living in their own homes in smaller towns. Very well-educated & working well-paying white collar jobs. High median household incomes of over \$82,600. Upscale shopping behaviors, driven by service & comfort.
Hard Hats	Middle-class, white couples with & without children. Small town homeowners. Age 35-54. High school/some college or Associate's degrees; employed in a mix of blue & white collar occupations. They have slightly above average median household incomes of \$48,100. Their discount shopping style is driven by price & function.
Influential Elders	Wealthy older couples without children, living in suburbia. Highly educated professionals & executives with median household income of \$115,000+. Home is owner occupied. Predominantly white & Asian, age 55+. Upscale shopping behavior with service & comfort purchasing triggers.
Just Getting By	Underprivileged, Gen-X town singles with children. Mix of homeowners & renters. Predominantly African-American; age <35. Some high school education, employed in blue collar & service occupations, with median incomes less than \$26,500. Their mid-market shopping behaviors are driven by price & function.
Kids on Decks	Upscale, town families with children living in their own homes. Predominantly white; age 35-54. College graduates; employed in a variety of white collar occupations. Median household income over \$77,000. Mid-market shopping behaviors, driven by value & function.
Lots of Tots	Suburban mix of African-American singles & families with children. Mostly homeowners. Age <45. High school graduates; employed in decent paying blue collar jobs. Median household income is approximately \$44,200. Mid-market shoppers driven by price & function.

Married with Homes	Suburban, white couples with & without children, living in their own homes. Age 25-44. High school graduates; employed in decent paying blue collar & service occupations. Median income is nearly \$43,000. Discount shoppers, driven by price & function.
Middle America	Middle-class, white singles & married couples without kids. Small town homeowners. Median age 41 with strong presence of residents <35. Some college/Associate degree level education, working a mix of white & blue collar jobs, with median household incomes modestly above average at approximately \$51,100. These mid-market shoppers are driven by value & style.
Senior Success	Mature couples, living in suburbia. Age 55+. Well educated; employed in white collar occupations. Above average median household incomes of approximately \$57,400. Mid-market shopping behaviors, driven by service & function.
Smart Renters	Suburban singles, ethnic mix. Mobile renters without kids. Median age 40 with strong presence of residents age 25-34 and 15-24. College education; employed in a mix of blue collar & white collar occupations. Median household income just shy of \$36,000. These mid-market shoppers seek value & style.
Suburban Seniors	Mature suburban singles without children. Mobile renters. Predominantly white & Asian. Age 55+ with strong presence of residents age 65+. High school graduates; median household income is nearly \$33,000. These mid-market shoppers are driven by service & comfort.
Suburban Society	Upscale, suburban homeowners. Predominantly white, age 45-64. Mostly married with children. College+ education; employed in well-paying, white collar jobs. High median household incomes of approximately \$82,000. Mid-market shopping behaviors, driven by service & comfort.
Suburban Starters	Low income, younger suburban homeowners. Mostly single, with & without children. Predominantly white; age <35. High school graduates; employed in blue collar, service & production/transportation/material moving occupations. Median household income approximately \$33,200. These discount shoppers are driven by price & function.
Suburban Success	Suburban homeowners, white families with children. Age 35-54. College education; working a mix of blue & white collar jobs. Median household income of approximately \$60,400. Mid-market shoppers, driven by value & function.
Town Elite	Wealthy and stable town families with children. Median household income \$113,000+. Highly educated, they enjoy management, executive & professional occupations. Home is owner occupied. Predominantly white, 45-64 years of age. Their upscale shopping behaviors are driven by service & comfort.
Upward Mobility	Middle-class, single suburban renters without children. Predominantly Asian & white; age <45. College+ educations; working decent paying white collar jobs. Median household income approximately \$55,500. Upscale shopping behaviors, driven by service & style.

After solving the least squares problem using these demographics displayed above, many values were obtained that must be discounted out of hand. For example, it is not possible to have a negative expected sales volume, and it is highly improbable that a demographic has an expected sales value of one hundred dollars.

To aid in analysis and to create a more realistic model, some of the demographics may now be clustered using a stepwise cluster analysis (See Appendix III). Of the 26 profiles, 11 profiles were clustered into the five groups below. The 11 profiles chosen to be clustered generated expected sales that were unrealistic. By grouping them together, they have a more logical, and hence analyzable, expected value. Upon inspection, the demographic profiles clustered together also have similar characteristics, such as purchasing motives and age.

Table 8. Clusters formed prior to expected value analysis.

Cluster Name	Power Players
Comprised of profiles Description	Established Elite, Influential Elders Highly educated, median household income over \$115,000. Home is owner occupied. Driven by service and comfort purchasing.
Cluster Name	On the Bubble
Comprised of profiles Description	Lots of Tots, Just Getting By, Ethnic Elders Mid-market shoppers that have low incomes and low education levels. Blue collar jobs. Primary motivations are price and function.
Cluster Name	On the Rise
Comprised of profiles Description	Middle America, Upward Mobility Still young, these shoppers have a higher educational level and slightly higher shopping levels. With a median income of ~\$52,000, they are motivated by style.
Cluster Name	Prime Time
Comprised of profiles Description	Affluent Town Families, Kids on Decks Middle-aged shoppers with mid-market shopping behaviors. Very well educated, with a median salary of ~\$59,000. Driven by value, function, and style.
Cluster Name	Still Going Strong
Comprised of profiles Description	Golden Years, Town Elite Older shoppers, passing middle age and moving into senior status. Upscale shopping behaviors with a high median salary of ~\$80,000 to boot. Motivated by service first, then comfort.

where U and V are orthogonal matrices. To solve the system shown in the Figure for x , the *pseudoinverse* of A , denoted A^+ may be computed by

$$A^+ = V\Sigma^{-1}U^T.$$

and the expected value vector can be computed by

$$x = A^+b = V\Sigma^{-1}U^Tb.$$

This methodology is applied to an entire year's worth of data. Once the expected values for each demographic are calculated, the process can be run again using departmental sales totals. The matrix setup will look the same as Figure 4 but will use departmental sales totals instead of total store sales. There are sixteen departmental equations, one for each department.

Then, since the expected value operator is linear, the expected values for each department will sum to the total expected value for each customer demographic.

$$\begin{aligned} E(\text{Total Sales}) &= E(\text{Sum of Department Sales}) \\ &= E(\text{Paint}) + E(\text{Garden}) + E(\text{Food}) + \dots + E(\text{Treasure Hunt}) \end{aligned}$$

The values that result from these computations follow in Appendix IV.

Appendix III – Cluster Analysis

The distance metric used in the cluster analysis is derived from the correlation coefficient between two variables,

$$\text{Similarity} = \frac{1}{N} \sum_{i=1}^N \left(\frac{X_i - \bar{X}}{\sigma_X} \right) \left(\frac{Y_i - \bar{Y}}{\sigma_Y} \right)$$

where \bar{X}, \bar{Y} are the means of the each variable, σ_X, σ_Y are the standard deviations, and N is the number of instances.

The algorithm proceeds in steps, gradually relaxing the notion of similarity. The first step will join the two variables that have the highest correlation coefficient, and thus the highest similarity. Each additional step will join either two single variables or an additional variable to an already constructed “cluster” until all variables have been added.

It should be noted that once a cluster is formed, variable additions proceed from the “center” of the cluster. This leads to the concept of linkage; the method by which variables or clusters are added to already existing clusters rather than single variables.

There are multiple linkage methods, the three most popular being single linkage, complete linkage, and average linkage. Each linkage method will produce a different clustering. After consideration of all three, average linkage was determined to be the preferred method. It creates a “center” of a cluster and computes distances from the cluster as taking the average distance of all points within the cluster.

Appendix IV – Expected Values

Due to the sensitive nature of the results in this appendix, it has been redacted from public view.