Better in Bulk: What Consumer Habits of CAP 5771 Reveal Compared Alongside the Average Customer

Hayden Kirkeide and Alan Van Etten
CAP 5771
Professor Dong
Fall 2025

Introduction

Costco Wholesale is a membership-based wholesale retail chain that was founded in the early 1980s and has since grown into an international phenomenon, with locations in 14 different countries across 4 continents. Its main mission is to provide discounted products – ranging from produce, bakery items, sporting goods, furniture, and even tires and insurance – of the "best value" it can possibly provide to its membership holders (Costco Wholesale Corporation 2025).

Thus far, it has been successful, attracting 76.2 million members – a 7.3% increase from the previous year – as of the close of the 2024 fiscal year despite a $5 increase to membership prices in September (Costco Wholesale Corporation 2024). While the demographics of Costco memberships have traditionally skewed towards middle-aged adults with families, the demographics of new members have shown a steadily increasing growth of ages 30 - 35 and 20 - 29, belonging to the age groups known as millennials and Generation (Gen) Z, respectively (Foust 2025). Despite the fact that these age groups are the most likely to be enrolled in a degree program – including undergraduate, graduate, and doctoral programs – little research has been done regarding the frequency and density of membership status amongst students, most of whom are young adults that fall within the scope of those age ranges.

Better in Bulk: What the Consumer Habits of CAP 5771 Can Reveal Compared Alongside the Average Costco Customer is a survey-based data mining project that aims to examine the Costco habits of our class alongside those of the average American Costco member by utilizing traditional statistical methods in addition to both supervised and unsupervised learning techniques in order to establish an elementary link between transactional data, socioeconomic circumstance, and demographic characteristics of the modern-day college student.

Research Questions:

1. What do the motivations and benefits of shopping at Costco tell us about our demographic as students?
2. Do our class's Costco habits provide insight into the global trends of Costco consumers when compared against them?
3. What can the limited study of our consumer habits – while of varied socioeconomic circumstances – as students in this time of economic uncertainty reveal about consumer habits in the modern day?

Literature Survey:

In past studies and examinations of consumer behaviors, multiple different sources have found significant links between the demographic characteristics of consumers and their choice of retail store. Nilsson et al. (2014) sought to create a classification system of shoppers based on their choice of store and shopping method (larger infrequent trips versus shorter more frequent trips) and found significant differences in the demographics of each group, including age, income, household size, and access to transportation. Prasad's (2010) study of consumer's choice of supermarket in India and Bai et al. (2008) study in China both also found similar trends, providing more evidence of the correlations, as well as suggesting that these trends are not simply localized occurrences.

Additionally, the formats of retail stores can be an influential factor. Fox et al. (2004) compared different retail formats to see how consumers' behavior change between shopping at supermarkets versus smaller grocery stores and noticed that consumers tend to be more selective in their choice of grocery store but less selective in their choice of supermarket. Thus, supermarkets tend not to compete as directly

with each other compared to grocery stores. Costco, as a warehouse club-style retailer, doesn't fit either category, but still exists within the same competitive environment as other stores. Its subscription-based style leads to it facing a higher level of rivalry against other retailers, especially against other warehouse club stores (Crissone 2015). As a result, consumers may be more picky when choosing between Costco and other available options.

Methodology and Datasets:

After determining how we wanted to collect our data through survey methods, we first began by writing the questions for the Preliminary Survey, the survey that focused the most on demographic data. Determining the relative age, employment, and membership status of most of the respondents allowed us to decide which statistical and machine learning techniques to apply to the data. We then released this survey to members of our class during our Project Presentation. Based on the responses from the Preliminary Survey, we constructed the questions for the Trip-by-Trip Survey to align with the demographic data we gathered. The Trip-by-Trip Survey was a survey respondents filled out each time they went to the grocery store and included basic statistics such as the store, transaction cost, and Costco membership status of each grocery trip; it was available to submit for 5 weeks. Both sets of questions listed on the respective Google Forms we sent out and the responses we received are available to view in the appendix of this project.

Throughout our data collection, we researched the necessary Python scripts to complete our analytics techniques. Once data collection was completed, we anonymized both sets of data by removing personal identifiable information and then created an additional optimized dataset designed to be more manageable to code with (concatenating feature names, globally updating responses for binary analysis, etc.) in Google CoLab. Additionally, we utilized generative artificial intelligence (AI) to generate a synthetic dataset of 500 entries based on the anonymized demographic and transactional data we received from members of the class to provide data on a larger scale from which to draw conclusions. We generated this data according to the global percentage of Costco membership holders aged 18 - 29, which was 67%.

Below are the characteristics of both datasets:

CAP 5771		**SYNTHETIC**	
Descriptive Statistic	**Resulting Figure**	**Descriptive Statistic**	**Resulting Figure**
Total Samples	12	Total Samples	250
Membership (%)	33.33	Membership (%)	57.72
Average Transaction Cost ($)	118.94	Average Transaction Cost ($)	171.49
Standard Deviation of Time Spent (mins)	34.69	Standard Deviation of Time Spent (mins)	36.48
Variance of Time Spent ($mins^2$)	1203.43	Variance of Time Spent ($mins^2$)	1330.82

For statistical analysis, we used Pearson's Chi-Square Test of Independence to evaluate the presence of a relationship between age demographic and Costco membership status. We then used ANOVA tests to compare the previously observed membership and demographic data with time spent in-store and amount of money spent per trip. For our machine learning techniques, we employed k-means clustering to evaluate the relationship between demographic, trip frequency, and money spent per trip as well as linear regression to attempt to explain the relationship between time spent in-store and the amount spent per visit. Also, a perceptron trained to use the demographic and behavioral characteristics of consumers to predict Costco membership will be created, using the previous tests to determine the classifiers to be used and their starting weights.

Finally, we completed both our statistical and machine learning analysis in Python. Working in CoLab, we uploaded the anonymized collected and synthetic datasets to perform our analysis. Once we had finished coding, we wrote down our results in the Final Project Presentation and eventually wrote them down formally in this paper. One of the limitations of our project was that only students present during our project proposal presentation and those that checked Canvas submitted the forms necessary to participate in the experiment. Next, the short time frame reduced the amount of data collected. Ideally, we would wish to employ a full year or multi-year analysis of student habits to see how they change throughout the year and their academic career, but we were restricted to about 4 weeks of actual data collection. Lastly, our experiment was restricted by the narrowed scope of our data collection stemming from focusing on our class specifically. In the future, were we to expand on our research, we would expand our demographic analysis to include larger, more robust student types and demographics (i.e. freshmen, first generation students, etc.).

Assumptions and Additional Implementations:

One of the assumptions we had for both our collected (class) data and synthetically generated data was that all respondents to the survey were, in fact, students. For the class data, this assumption was easily made as only people within our class had access to the survey forms, so the assumption was most relevant to the synthetic data. The next assumption we established was that everyone who filled out the Preliminary Survey did it once. Again, this was easily managed with the collected data, so this assumption was most applicable to the synthetic data. Another assumption we made was that all respondents for both datasets were located in the Tallahassee area for the duration of the survey. This aided in restricting to a local context. Finally, we assumed that the Trip-by-Trip Survey responses were as in-depth and truthful as possible. For the class dataset, while we knew filling out the form was tedious at times, we assumed that respondents who signed up to participate in our surveys wished to contribute as much as possible to our project.

Results:

Statistical Analysis

Pearson's Chi-Square (χ^2) *Test of Independence*

We established our null and alternative hypotheses applying to our class and synthetic datasets to state that there is no observed relationship between age demographic and Costco membership status versus the presence of a statistically significant relationship between age demographics and Costco membership status, respectively. As is standard, we established the alpha value threshold of statistical significance to be 0.05. Adapting Python code for chi-square test for independence from Python for Data Science, we first explored the data using the built-in ResearchPy function and iteratively developed our

code according to the example provided in the blog. Eventually, we received an output with results as follows:

CAP 5771		**SYNTHETIC**	
Chi-Square Test	**Results**	**Chi-Square Test**	**Results**
Pearson Chi-Square	1.98385	Pearson Chi-Square	5.8679
p-value	0.3794	p-value	0.1182
Cramer's V	0.3721	Cramer's V	0.1532

In terms of the data collected from the class, the chi-square value is quite small, meaning that the outcome of the data is close to the expected value of the chi-square. Cramer's V – which indicates the association between variables – has several tiers of association, with 0.15 indicating a strong association and anything above a 0.25 indicating a very strong association. The class data's Cramer's V is 0.3721, quite large in comparison to the benchmark and thus indicating that there is a very strong association between the two variables. For the synthetic data, the chi-square figure is a bit larger than that of the class data's, understandably so since the size of the former is much larger than that of the latter, yet still indicates an outcome close to the expected value of the chi-square. The synthetic data's Cramer's V is that of a 0.1532 and, while not as large of a value as that of the class data's, indicates a strong association between age demographic and membership status. Ultimately, the p-value for neither the class data nor the synthetic data was statistically significant enough to reject the null hypothesis.

One-way ANOVA

Our next experiment was utilizing a one-way ANOVA test to gauge the statistically significant difference between various demographics, membership statuses, and trip frequency in the two datasets. We once again established a null hypothesis of there being no relationship between Costco membership status and money spent per transaction while our alternative hypothesis was the presence of a statistically significant relationship between membership and money spent per transaction. We adapted our Python code from Python for Data Science's article "One-way ANOVA" and once again began by preliminary data exploration and optimization. Following those steps, we received an output with results as follows:

CAP 5771		**SYNTHETIC**	
ANOVA Statistic	**Results**	**ANOVA Statistic**	**Results**
F-statistic	0.3652	F-statistic	702.4277
p-value	0.5567	p-value	≈ 0.000001
Effect size (η^2)	0.0295	Effect size (η^2)	0.7391

The experiment yielded vastly different F-statistics on the data collected from the class versus the synthetic data. The ANOVA test performed on the class data yielded an F-statistic that was significantly

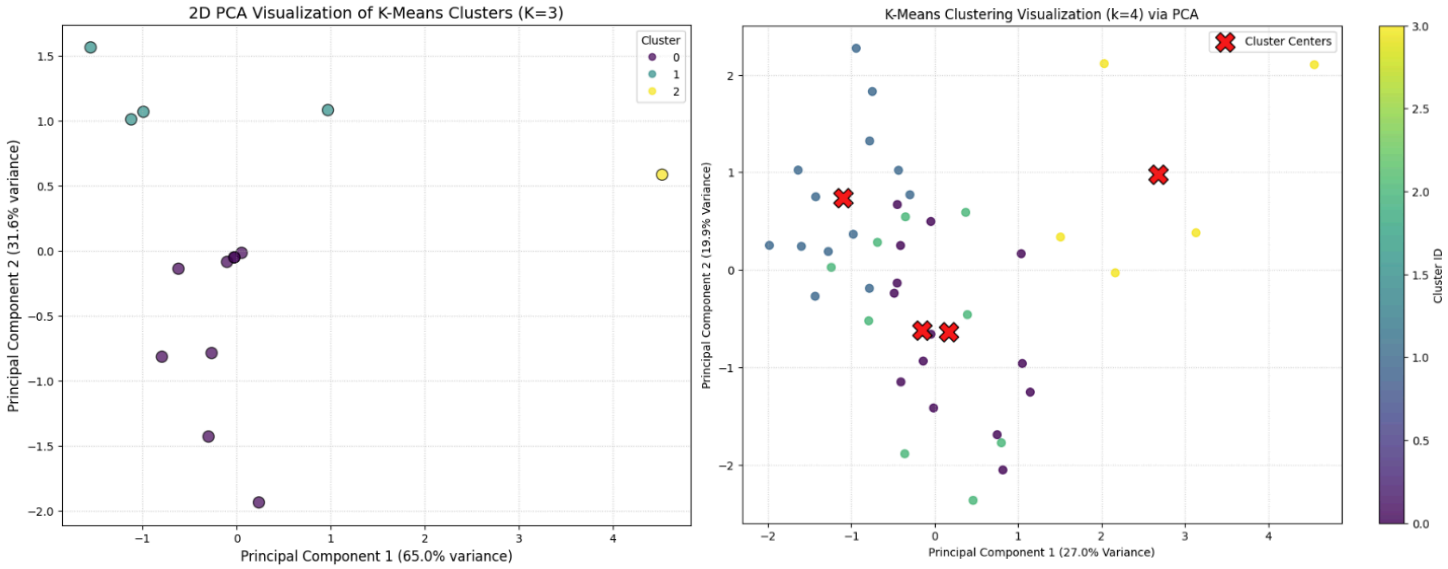
smaller than on that of the synthetic data, indicating much less variance in the class dataset versus the synthetic. While the F-statistic for the synthetic data seems alarmingly high, this is in keeping with the characteristics of the dataset as the number of samples within the synthetic dataset is much more than that of the class dataset. In terms of the effect size, the effect of the variables on the synthetic data is a staggering 73%, much stronger than the 2% of the class data and reaffirming the importance of variance within data. In the end, the p-value of the class data is about a 0.56, making the findings not statistically significant enough to reject the null hypothesis; however, the ANOVA test performed on the synthetic data was so small that the data had to be rounded to the millionth, uncontestedly confirming the statistical significance of the test and leads to the adoption of the alternative hypothesis for that dataset.

Machine Learning

K-Means Clustering

When applying k-means clustering algorithms to the class and synthetic datasets, we first perform principal component analysis (PCA) on the data to compress and streamline the clustering process. For both datasets, principal component (PC) 1 is a combined metric of both transaction costs and shopping frequency while PC 2 is time spent in the store; data points to the far right side of the graph indicate high frequency shoppers making high value transactions and vice versa. Below are the clustering graphs for both datasets:

CAP 5771 **SYNTHETIC**

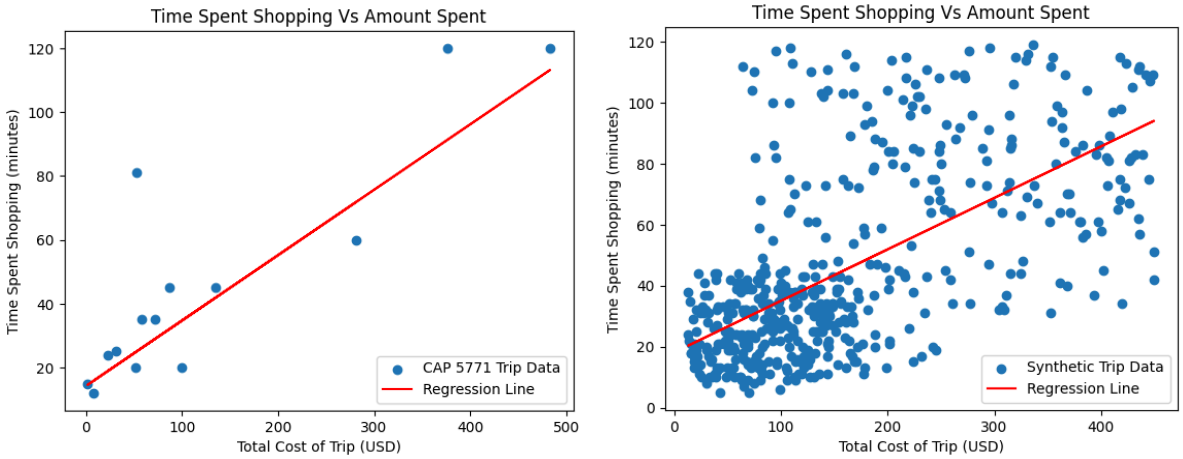


The graph showing data collected from the class depicts a series of distinctly different clusters with low variance while the graph showing data from the synthetic dataset depicts reasonably different clusters showcasing more variance than that of the class data. The main conclusion we can draw from both of these k-means clustering graphics is that high frequency shoppers tended to spend more time in the store and thus spent more money per transaction; the opposite is also true: low frequency shoppers spent less time in store and thus spent less money. Within the context of our research questions, this could also indicate age differences within the datasets, for example younger students aged 18 to 21 don't shop for themselves much because of dorm food or cafeterias and thus go infrequently, buy only specific items, and therefore spend less money. Summarily, what both the class and synthetic datasets show is the

conclusion that students tend to be less frequent shoppers and thus spend less money on grocery shopping transactions.

Linear Regression

To confirm the existence of trends similar to those seen in past studies, we examine the relationship between the amount of time spent shopping and the amount of money spent per trip, using the trip-by-trip survey data and corresponding synthetic data. The results are shown below:



Using the null hypothesis that the two attributes have no relationship and the alternative hypothesis that a relationship exists, we obtain a p-value of 4.60e-05 for the survey data and 3.81e-53 for the synthetic data, proving that there is a relationship, which is consistent with the previous research.

Logistic Regression, Random Forest and Multi-Layer Perceptron

In order to attempt to predict Costco membership status with the survey information presented, we create three model types and compare the accuracies of these models with each other, using the survey and synthetic data. The models take average spending per trip, trip duration, visit frequency, age group and employment status as inputs and predict Costco membership status. The class survey data used a 50-50 split of training and testing data, needed due to the lower amount of data, while the synthetic data used a 70-30 split. The average accuracies over 10 iterations for each model and dataset are provided in the table below:

Average accuracies of model predictions over 10 iterations	CAP 5571	Synthetic Data
Logistic Regression	0.3857	0.882
Random Forest	0.3714	0.8648
MLPerceptron	0.3667	0.866

All models provide similar accuracies, with logistic regression performing the best on average, but the models show high variance in the accuracies for each iteration, especially with the class survey data. As a result, it's not appropriate to claim any of these models as better than the others.

Conclusion:

The findings of these experiments offer some insight into how the demographics of our class relates to Costco membership and general shopping trends, though it is dependent on the use of synthetic data, which makes the validity of these results somewhat questionable. Synthetic data can copy trends in the datasets it's based on, but has the issue of interpreting random noise as trends. The higher significance in the one-way ANOVA test and higher accuracies in the predictive models is not necessarily caused by faults in the synthetic data, but further analysis of the synthetic data and more survey data would help alleviate these concerns for continuations to this study. The linear regression plot shows similarity between the two datasets, having significant correlation in both the original data and synthetic data, as well as having similar point distributions, both of which provide evidence for the validity of the data. As for the other examinations, a lack of samples can result in the data appearing as random, with not enough information to identify a trend. The analysis failing to reject the null hypothesis doesn't mean the alternative hypothesis is wrong, just that there isn't enough evidence to the contrary.

As far as the relations found within the synthetic data go, they fall roughly in line with what was expected, that being similar findings to the previous research and general assumptions. The link between money spent per trip and Costco membership makes intuitive sense, shopping habits would be different due to the difference in how Costco and similar stores operate compared to more general stores. The relation between time spent in the shop and money spent also makes intuitive sense, as a customer spending more time shopping is likely using that time to buy more goods. The predictive models were able to reach an accuracy above 80% on the synthetic data, which indicates that the model was able to find some trends connecting the chosen variables with Costco membership, though it wasn't especially consistent. More data, the inclusion of untested features and model parameter adjustments may result in improvements to the prediction accuracy for future studies. Additional analysis of feature importance may be something to include for studies seeking to incorporate new variables.

Future Research:

In considering the applications of our admittedly limited research on student grocery habits, we should first consider how to evaluate students as consumers. To build a foundational basis from which to expand, there should first be an in-depth examination of the value Costco consumers receive as a result of their membership to the chain through comparative, year-over-year analysis of chain revenues, member spending, and discounts provided through the membership. While more focused on consumer research, accomplishing this will provide a baseline to compare students who hold Costco memberships to.

We also believe that there is a connection to be explored between increasing rates of inflation and the increasing numbers of young adults enrolling in a Costco membership. In considering the value for money, Costco memberships are some of the most effective ways to spend money effectively, whether it be on the $5.00 rotisserie chicken or $1.50 hot dog combo where you get multiple meals for pennies on the dollar. As students – many of whom are on a fixed income or lack an income at all – in an increasingly expensive economy, utilizing money effectively is something that is constantly on your mind, and Costco may aid in alleviating some of your worries.

Furthermore, more research into how students behave as consumers in the face of their socioeconomic circumstance would lay the groundwork for expansion into fields beyond groceries like hobbies, dining out, engaging in nightlife, travel, and much, much more. As students are becoming young adults, socializing and engaging in recreational activities are just as important as completing their education and working their first job. Not only does student spending equally contribute to the economy,

it also provides them with a preview of their post-educational life, teaching them the importance of budgeting, life experience, and work-life balance. While several counterarguments can be made about the difference between spending as a student versus a post-graduate young adult, the spending occurs regardless and stimulates the economy in the same way. Summarily, as everyone is a student at least once in their lifetime, their contributions and spending habits are just as socioeconomically important as those in their post-graduate and adult lives. By expanding the scope of consumer research to focus on students specifically, social, scientific, and economic research may be able to yield some much-needed conclusions that have evaded these sectors for as long as their insights have been recorded.

Bibliography and References

Bai, Junfei, et al. “Consumer Choice of Retail Food Store Formats in Qingdao, China.” Journal of International Food & Agribusiness Marketing, vol. 20, no. 2, 2008, pp. 89–109, https://doi.org/10.1080/08974430802186217.

Bobbitt, Zach. “K-Means Clustering in Python: Step-by-Step Example.” Scientific Blog. Statology, https://www.statology.org/k-means-clustering-in-python/. Accessed 19 Nov. 2025.

Chen, Jiangpei. Marketing Strategy Management of Costco : Analysis and Comparison to S-Group. 2021, Theseus, https://www.theseus.fi/bitstream/handle/10024/504238/Chen_Jiangpei.pdf?sequence=2&isAllowed=y. Centria University of Applied Sciences.

“Chi-Square Test of Independence.” Informational Blog. Python for Data Science, https://www.pythonfordatascience.org/chi-square-test-of-independence-python/. Accessed 17 Nov. 2025.

Crissone, Chloe Danica. Costco: A Strategic Analysis. 2015, https://louis.uah.edu/cgi/viewcontent.cgi?article=1274&context=honors-capstones. University of Alabama, Honor’s.

Foust, Lyden. “Costco Brand Teardown: Seizing A Cultural Moment.” Blog. Spatial.Ai, 21 Aug. 2025, https://www.spatial.ai/post/costco-brand-teardown-seizing-cultural-moment.

Fox, Edward J., et al. “Consumer Shopping and Spending across Retail Formats.” The Journal of Business, vol. 77, no. S2, 2004, pp. S25-60, https://doi.org/10.1086/381518.

Nilsson, Elin, Tommy Gärling, et al. “Who Shops Groceries Where and How? – The Relationship between Choice of Store Format and Type of Grocery Shopping.” The

International Review of Retail, Distribution and Consumer Research, vol. 25, no. 1, 2014, pp. 1–19, https://doi.org/10.1080/09593969.2014.940996.

"One-Way ANOVA." Scientific blog. Python for Data Science, https://www.pythonfordatascience.org/anova-python/. Accessed 17 Nov. 2025.

Prasad, Ch. Jayasankara. "Effect of Consumer Demographic Attributes on Store Choice Behaviour in Food and Grocery Retailing - An Empirical Analysis." Management and Labour Studies, vol. 35, no. 1, 2010, pp. 35–58, https://doi.org/10.1177/0258042X1003500103.

Wholesale Corporation, Costco. "About Us." Retail. Costco.Com, https://www.costco.com/about.html.

———. "ANNUAL REPORT ON FORM 10-K FOR THE FISCAL YEAR ENDED SEPTEMBER 1, 2024." Costco Wholesale Corporation, 8 Oct. 2024.