MAR 653 Project

Fall 2018 Team S

Catherine Devine
Lauren Foltz
Smit Patel

Executive Summary

Executive Summary

The objective was to analyze household grocery transactions over two years to understand the characteristics of customers and develop a predictive model for future grocery sales. This analysis was based on the “Complete Journey” dataset that was sourced from the Learning Management System (LMS).

Cluster analysis was conducted on demographic information in the dataset to understand the profiles and characteristics of customer segments.

Linear regression was conducted to understand the variables that drive total sales from each household.

In summary, our findings were that the highest income households spent the most on food. This is possibly not surprising, except that you would expect everyone has a base need for groceries regardless of income.

Research Objectives

Key **research objectives** were twofold:

1. Identify **profiles** of customer segments.

1. Create a **predictive sales model** by understanding the variables influencing sales, based on historical buying patterns.

Background on Shopper Cards

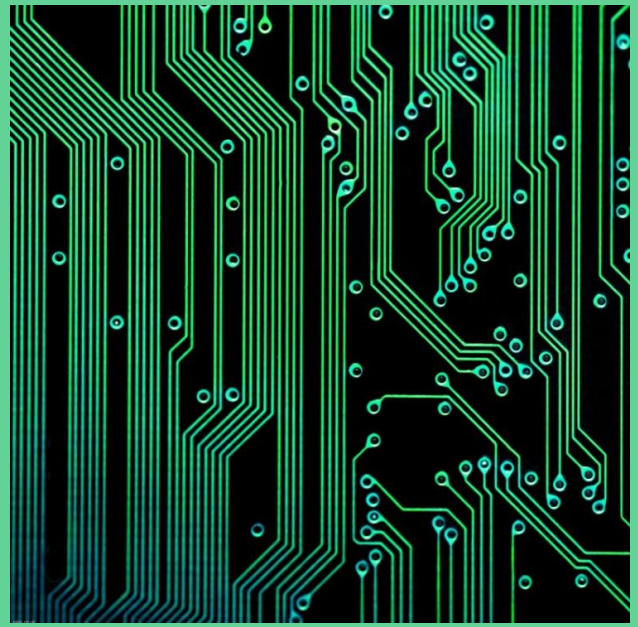
Shopper Cards

- Prevalence - It is estimated that over 85% of US retailers have some form of loyalty card.
- Data Capture-
  - When a customer signs up for a loyalty card, demographic information is captured.
  - Each time a shopper checks out of the grocery store, data for that transaction is captured. This includes the products sold, the quantity, and the actual price paid. It can also include payment method, timestamp, and the checkout lane used. Some retailers also capture coupon usage, and whether the item purchased was in a regular stocking location, or was part of a special display.
- Analytics - Shopper card data can be used in many ways, such as identifying price sensitive customers and sending them coupons via mail, a kind of personalized pricing.

About the Data

Complete Journey Data:

The dataset comprised 46,757 transactions over two years representing the grocery transactions of 801 households.



Complete Journey Data - Demographic Subset

The data contained 7 discrete demographic variables for 801 households.
Each variable had between 3 and 12 levels.
For example, Marital Status had 3 levels: Married, Single, Unknown.
Using R, a subset of demographic data was created, and the levels were converted to numeric factors to allow for cluster analysis.

- **Age Group**: bands of consumer age groups
- **Marital Status**: marital status - married, single etc.
- **Income Level:** bands of income levels
- **Homeowner Status:** scale from owners to renters
- **Household Composition:** children/adult mix in the household
- **Household Size:** number of people in the household
- **Kid Category:** number of children in the household

Complete Journey Data - Transaction Portion

The following 9 transaction variables were available to be utilized for the linear regression model. Subsequent analysis only leveraged those that were significant. The final linear regression model is detailed later.

- **Household ID:** Household number identifier
- **Year** (data covered a 2 year period): Year 1 or 2
- **Category**: Deli, Meat, Pastry, etc.
- **Quantity**: Quantity purchased in each transaction category
- **Sales Value (price)**: Total value of transaction for category
- **Coupon Discount:** Coupon discount used
- **Num_National**: Private label or national brand identifier
- **Coupon Redeemed**: Number of coupons
- **Coupon Value**: Total amount of coupon

Complete Journey Data - Data Cleaning

Each transaction row included the associated demographic data. This resulted in 46,757 rows of demographic data although there were only 801 distinct households.

We segmented the demographic data, identified the unique (distinct) instances, converted the variables to numeric factors, and ran a cluster analysis where each distinct household was represented only once in the dataset.

For the 46k+ transactions, we eliminated one of the coupon related columns (coupon discount) from the dataset. This was because 39k transactions had missing values for this column which negated ability to run regression analysis. We also converted all the categorical data to numeric factors for linear regression analysis.

Research Plan & Methodology

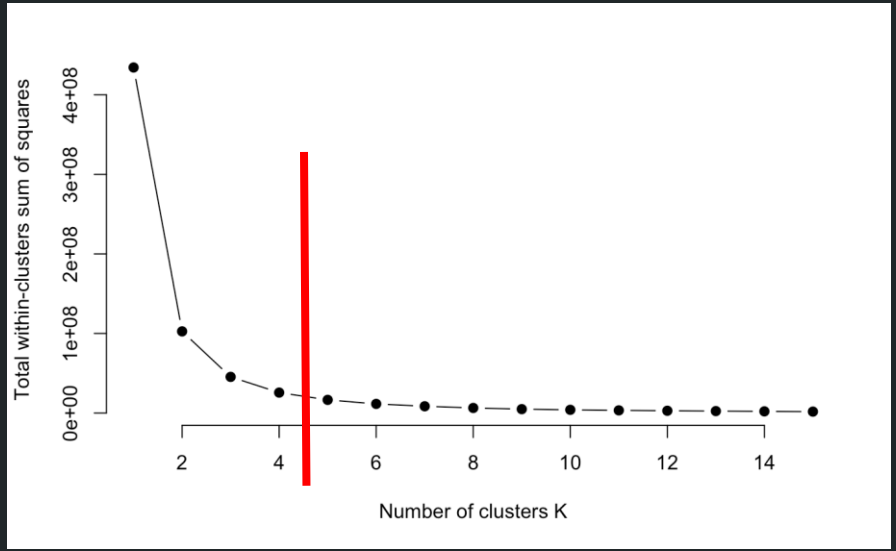
Cluster Analysis to identify customer segments

Performed cluster analysis on demographic data for the 801 households in the dataset.

Clustering of customer demographics including total sales per household

To conduct cluster analysis, we took the 801 unique household demographic data. In addition, we aggregated total sales for each household to include in the cluster analysis.

Running the “elbow plot” analysis identified optimum number of clusters for k-means clustering analysis of 4.



K-means cluster analysis on demographic data

A K-means clustering analysis was run with k = 4, as identified in an elbow plot, which resulted in four clusters of the 801 households with sizes 124, 56, 336, 285.

Results of the cluster means in that analysis are shown in the below table.
The legend is on the following page, followed by profile development of each cluster.

Cluster Means

agelevel	maritalstatus	incomedesc	homeownerdesc	hhcomp	hhsize	kid	sales.value
3.37	1.99	5.10	2.35	3.24	2.40	3.26	8476.07
3.41	1.75	6.48	1.43	2.82	2.61	2.98	13653.10
3.57	2.09	4.30	2.46	3.47	2.10	3.43	2466.88
3.45	1.96	4.68	2.44	3.33	2.14	3.34	5017.41

Legend - Demographics Clustering

Age
(agelevel)

1 - 19-24

2 - 25-34

3 - 35-44

4 - 45-54

5 - 55-64

6+ 65+

Marital Status
(maritalstatus)

1 - Married

2 - Single

3 - Unknown

Income
(incomedesc)

1 -Under 15k

2- 15-24k

3-25-34k

4-35-49k

5-50-74k

6-75-99k

7-100-124k

8-125-149k

9-150-174k

10-175-199k

11-200-249k

12-250k+

Homeowner

(homeownerdesc)

1 - Owner

2 - Probable Owner

3 - Probable Renter

4 - Renter

5- Unknown

Household Comp

(hhcomp)

1 - 1 Adult, kids

2 - 2 Adult, kids

3 - 2 Adults no kids

4 - Single female

5 - Single male

6 - Unknown

Household Size

(hhsize)

1 - 1

2 - 2

3 - 3

4 - 4

5 - 5+

Children

(kid)

1 - 1

2 - 2

3 - 3+

4 - None/unknown

Cluster profiles

Cluster 1

Mid range spenders, mid income, young renters

Second highest spending cluster at 3x lowest cluster in aggregate spend

Youngest of the four clusters skewing to 35-44 year olds

Likely to be single

Income at 50-74k

Skewed to renters

Skewed to 2 adults no kids

Household size of 2

Cluster 2

High spenders, high income, young homeowners

Highest aggregate spending cluster with 5x lowest cluster

Second youngest of the four clusters skewing to 35-44 year-olds.

Likely to be single

Highest income at close to 100k

Skewed to homeowner

Skewed to 2 adults with no children

Household size of 2

Cluster profiles

Cluster 3

Low spenders, low income older renters

Lowest spending cluster.

Oldest of the four clusters skewed to 45-54 year olds.

Likely to be married

Lowest Income at < 50k

Skewed to renters

Skewed to single female

Household size of 1-2

Cluster 4

Mid spenders, mid income, older renters

Second lowest spending cluster at 2x lowest cluster

Second oldest of the four clusters skewed to 45-54 year olds.

Likely to be single

Income close to 50k

Skewed to renters

Skewed to single female

Household size of 1-2

Cluster Profiles Summary

Our analysis identified four distinct groups of customer segments.

- Mid spenders mid income young renter
- High spenders high income young homeowners
- Low spenders low income older renters
- Mid spenders mid income older renters

This clustering allows us to understand the types of customers we have for groceries which can inform marketing, purchasing etc. without each cluster being over or under-represented based on its transaction frequency.

Interestingly, the addition of aggregate spend by household significantly changed the cluster profiles which suggests that selection of variables included in cluster analysis is important.

Linear Regression
to develop predictive sales model

Performed linear regression to develop a model that would allow us to predict total sales value.

Predictive Sales Model Approach

- We conducted an initial multivariate linear regression on all variables to identify those that were significant. This model included both demographic and transaction data elements as unlike cluster analysis we wanted to understand how demographics influenced sales.
- We chose to eliminate the **coupon_discount** variable from the analysis as it had 39k missing values (of the 46k+).
- We iterated through multiple regression models eliminating variables that were not statistically significant in the mode.
- The final model is detailed on following page.

Linear Regression Model

Finalized model to predict sales value using independent variables of:

- Category
- Quantity
- Num_national
- DH_coupon_redeemed
- Agelevel
- Incomedesc
- Homeownerdesc

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.37187	1.57240	18.044	< 2e-16 ***
category	-0.22407	0.03333	-6.723	1.80e-11 ***
quantity	0.86128	0.01123	76.709	< 2e-16 ***
num_national	1.81916	0.02029	89.665	< 2e-16 ***
dh_coupon_redeemed	-2.11552	0.54818	-3.859	0.000114 ***
agelevel	-0.63322	0.26498	-2.390	0.016867 *
incomedesc	-0.60300	0.10157	-5.937	2.92e-09 ***
homeownerdesc	-1.68787	0.18437	-9.155	< 2e-16 ***

Adjusted R^2 value of 71.78%, indicated that 71.78% variability in Sales value can be explained by Independent Variables

Residual standard error: 70.8 on 46749 degrees of freedom

Multiple R-squared: 0.7178, **Adjusted R-squared: 0.7178**

F-statistic: 1.699e+04 on 7 and 46749 DF, p-value: < 2.2e-16

Linear Regression Model

Predicted Sales Value

= 28.37187

- 0.22407 * Category

+ 0.86128 * Quantity

+ 1.81916 * Num_National

- 2.11552 * Coupon Redeemed

- 0.63322 * Age Level

- 0.60300 * Income Level

- 1.68787 * Homeowner Status

71% of sales value for a transaction can be explained by this model.

Linear Regression Model

Interpretation

The most significant variables are

- Whether or not you own or rent. Renters spend less than homeowners.
- National or private brand. National brands spend more.
- How old you are - younger customers spend less than older ones.
- Higher income levels spend more than lower income levels
- Redemption of coupons reduces total sales
- The higher the quantity the higher the sale. This is as expected.
- Food category does not have significant impact on total sales.

The variables that were not significant were

- Number of children
- Size of household
- Composition of household
- Marital status of household
- Year of transaction
- Coupon value if used

Second Linear Regression Model

Interpretation

- We ran a second regression model to try and predict aggregate sales for each household rather than just sales in any single transaction of a household.

- The overwhelming driver of aggregate sales is similar to what we saw in our cluster analysis. That is, income level significantly drives aggregate spend. However, the model concludes that these variables only explain approximately 8% of the sales variance.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4091.3137   987.2407   4.144 3.78e-05 ***
household_key    0.3311     0.1530   2.164   0.0308 *
agelevel      -110.8774    92.6945  -1.196   0.2320
maritalstatus   49.7453   151.3210   0.329   0.7424
incomedesc     350.2657    51.7262   6.772 2.48e-11 ***
homeownerdesc  -81.4217    73.2289  -1.112   0.2665
hhcomp        -209.8979   128.1772  -1.638   0.1019
hhsize          98.3076   149.5374   0.657   0.5111
kid            -29.8470   133.9078  -0.223   0.8237
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3170 on 792 degrees of freedom
Multiple R-squared:  0.09379,   Adjusted R-squared:  0.08463
F-statistic: 10.25 on 8 and 792 DF,  p-value: 1.08e-13
```

Findings, Conclusions, & Recommendations

Findings / Conclusions / Recommendations

- ★ Clustering analysis revealed that while everyone purchases groceries, the majority of our customers buying groceries fall into the 35-54 age group . In addition, while everyone needs groceries, aggregate spend is highly influenced by income. Higher incomes spend up to 5 x what lower incomes spend on groceries.
- ★ Linear modeling revealed:
  - ○ Redemption of coupons, younger ages, lower income and renter status has a negative impact on the sales value of any single transaction.
  - ○ Quantity has positive effect on sales value as you would expect. The higher the quantity the higher the sales value.
- ★ Recommendations
  - ○ Target marketing to higher income customers to encourage overall aggregate sales volume.
  - ○ Encourage purchasers to buy increased quantity of any single food item. For example, promotions for buying more than 1 (4 for $5 offers as an example). More food quantity = More sales = Money


Questions?