Diana Amador

RETAIL ELECTRONICS

DATA ANALYSIS REPORT

November, 2014

Contents

Purpose

This is a Data Analysis Report based on the Customer Sales of Blackwell Electronics. The information was obtained by applying data mining and machine-learning techniques to make inferences about patterns in the data that will help the business make data-driven decisions about sales and marketing activities. Our goal was to investigate the patterns in customer sales to provide insight into customer buying trends and preferences.

Scope

The study focused on determining the following sales indicators:

1. Amount spent per transaction by Region
2. Correlation between customer age and amount spent
3. Correlation between customer age and online or in-store shopping

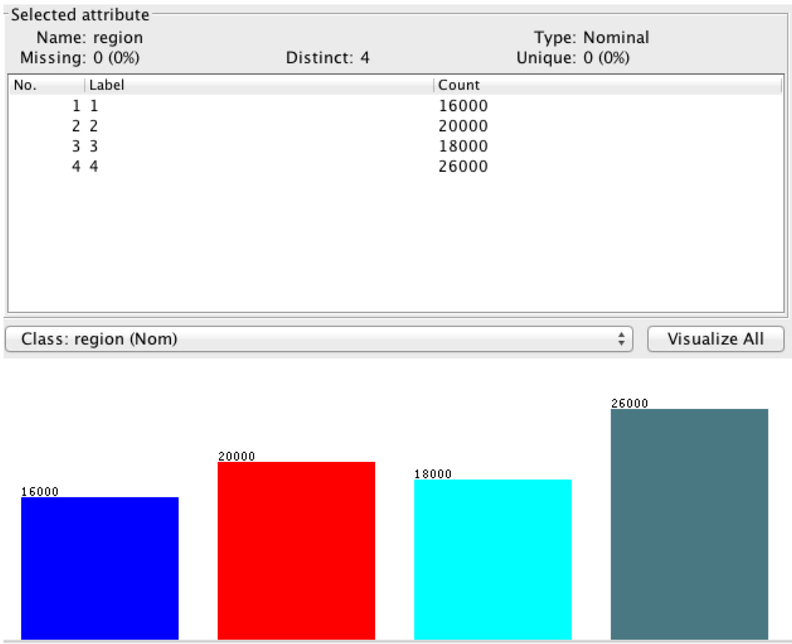
Region Color Code: Blue – Region 1 EAST; Red – Region 2 WEST; Light Blue – Region 3 SOUTH; Gray – Region 4 CENTRAL

Amount spent per Transaction by Region

Visualization Methodology

The amount spent per transaction by Region was determined using visualization tools and Histograms in WEKA (See graphs below). Histograms showed for each class of Region attribute the amount spent and we were able to explore how Region was related to amount and other attributes.

Figure 1 Amount spent by Region. Blue-Region 1, Red-Region 2, Light Blue-Region 3, Gray-Region4 (Region color code is used through all the graphs)



Observations

Customers in the Central Region (Gray/4) are the ones who spend the largest amount of money per transaction.

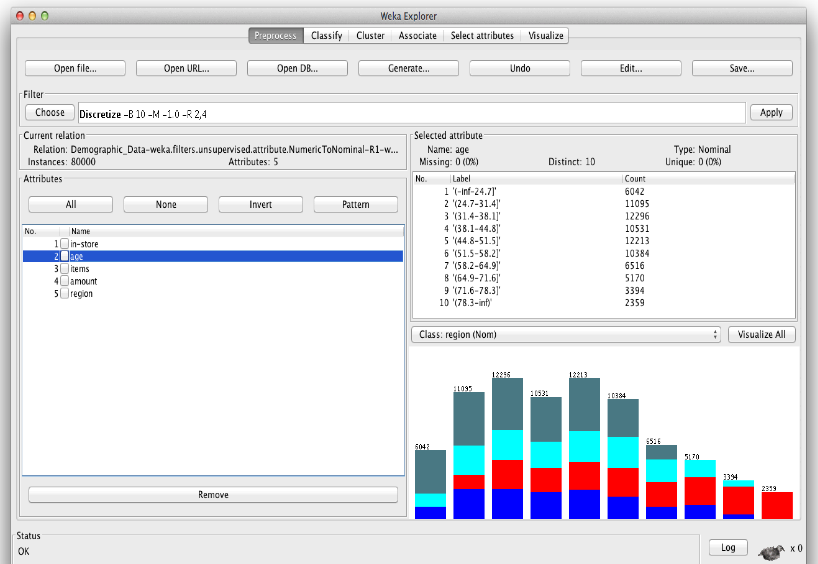
Customers in the West Region (Red/2) are the ones who spend the least amount of money per transaction.

Correlation between customer age and amount spent

Visualization Methodology and Heuristics

In order to find the correlation between customer age and amount spent we used two approaches. The first one was direct observation of the amount spent by age groups using WEKA's visualization tools, mostly histograms. Ten age groups were created in order to visualize and process the information.

Figure 2 Amount spent in each Region by each Age group. Colors in each bar show the number of customers from each Region that belong to that particular Age Group



Machine Learning Methodology

Our first ML approach was using a Machine Learning algorithm J48 tree to classify transaction amounts for each Region and related them at the bottom of the tree to predefined age groups. We used 3 different J48 models to determine the age of a customer based on his/her spent amount. The first model was run on the Age attribute keeping Amount attribute values numerical and continuous. The J48 pruned was extremely complex and the relative absolute and root squared relative error rates were as high as 94% and 97% for unclassified instances.

The second J48 model on the Age attribute was run using a discrete Amount attribute to create amount ranges that theoretically would make instances' classification more compact. Once again, the J48 pruned tree generated by the algorithm was extremely complex and the relative absolute and root squared relative error rates were very high (101%) resulting on an unacceptable number of unclassified instances. Both Age-based J48 trees were overfitted and unpractical to use to get any inferences from them.

We changed our approach and run a third J48 model on Region Attribute in order to force the use of discrete nominal attributes with fewer classes at the top of the tree. This J48 pruned tree was much more readable and its relative absolute error (54%) and root relative squared error rates (76%) were significantly lower than previous models providing a large amount of correct classified instances to make conclusions more reliable (64%). Although this approach didn't provide direct answers on the customer age, it did provide a strong indicator on what region a customer may come from. By crossing this information with the already known age composition for each specific Region, we were able to create a more reliable prediction on the probable customer age.

Figure 3 J48 Pruned Tree run on Region and Amount.

Classifier output

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Demographic_Data-weka.filters.unsupervised.attribute.NumericToNominal-R1,5-weka.filters.unsupervised.attribute.Discretize
Instances:    80000
Attributes:   5
              in-store
              age
              items
              amount
              region
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===
```

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

in-store = 0
|   amount <= 500.13
|   |   age = '(-inf-24.7]': 4 (360.0/130.0)
|   |   age = '(24.7-31.4]': 2 (1655.0/440.0)
|   |   age = '(31.4-38.1]': 2 (2917.0/418.0)
|   |   age = '(38.1-44.8]': 2 (2424.0/374.0)
|   |   age = '(44.8-51.5]': 2 (2828.0/434.0)
|   |   age = '(51.5-58.2]': 2 (2934.0/458.0)
|   |   age = '(58.2-64.9]': 2 (2414.0/264.0)
|   |   age = '(64.9-71.6]': 2 (2432.0)
|   |   age = '(71.6-78.3]': 2 (2425.0)
|   |   age = '(78.3-inf)': 2 (2359.0)
|   amount > 500.13: 4 (17252.0/6017.0)
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       51339               64.1738 %
Incorrectly Classified Instances     28661               35.8263 %
Kappa statistic                          0.5146
Mean absolute error                      0.2193
Root mean squared error                  0.3312
Relative absolute error                 59.1583 %
Root relative squared error             76.934  %
Total Number of Instances            80000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.569     0.204      0.411     0.569     0.477      0.827    1
                 1         0.04       0.893     1         0.944      0.989    2
                 0.178     0.044      0.538     0.178     0.268      0.778    3
                 0.732     0.194      0.645     0.732     0.686      0.823    4
Weighted Avg.    0.642     0.124      0.636     0.642     0.614      0.855

=== Confusion Matrix ===

     a     b     c     d   <-- classified as
  9100     0  2707  4193 |     a = 1
     0 20000     0     0 |     b = 2
  7672   854  3206  6268 |     c = 3
  5390  1534    43 19033 |     d = 4


in-store = 1
|   amount <= 999.9
|   |   age = '(-inf-24.7]': 4 (1664.0/683.0)
|   |   age = '(24.7-31.4]': 1 (4660.0/2760.0)
|   |   age = '(31.4-38.1]': 1 (4652.0/2738.0)
|   |   age = '(38.1-44.8]': 1 (4154.0/2408.0)
|   |   age = '(44.8-51.5]': 1 (4770.0/2907.0)
|   |   age = '(51.5-58.2]': 1 (3806.0/2187.0)
|   |   age = '(58.2-64.9]': 3 (2364.0/1096.0)
|   |   age = '(64.9-71.6]'
|   |   |   items <= 2: 3 (557.0/220.0)
|   |   |   items > 2
|   |   |   |   items <= 3: 1 (384.0/182.0)
|   |   |   |   items > 3
|   |   |   |   |   items <= 7: 3 (1591.0/706.0)
|   |   |   |   |   items > 7: 1 (206.0/99.0)
|   |   age = '(71.6-78.3]': 3 (969.0/432.0)
|   |   age = '(78.3-inf)': 1 (0.0)
|   amount > 999.9: 4 (10223.0/3632.0)

Number of Leaves  :     25

Size of the tree :      33


Time taken to build model: 0.67 seconds
```

Observations

The customers who spend the least average amount of money per transaction are Region 2 customers (Red/2) except for the oldest customer age group where they are the only customers in that Age group. Region 2 also accounts for the oldest population (above 51 years), which are customers who buy exclusively online. These observations don't support the idea that mostly young people buy online, but support the idea that the youngest and the oldest customers may be less wealthy and buy less per transaction than all the other age groups.

Age groups between 24.7 and 58.2 years old are the ones who typically spend the largest amount per transaction. Almost 80% of the sales are generated by these age groups regardless of the Region or the shopping method.

Correlation between amount spent and online or in-store shopping

Visualization Methodology and Heuristics

In order to find the correlation between amount spent on online or in store shopping we used two approaches again. The first one was direct observation of the online/in-store shopping habits on defined age groups using WEKA's visualization tools, mostly histograms. The second approach was using a Machine Learning algorithm M5P tree that works with numerical attributes, such as amount to determine the amount spent on online or in-store transactions.

Figure 4 Online/ In-store transactions composition by region. Color code: Blue-Online transactions; Red-In-store transactions.

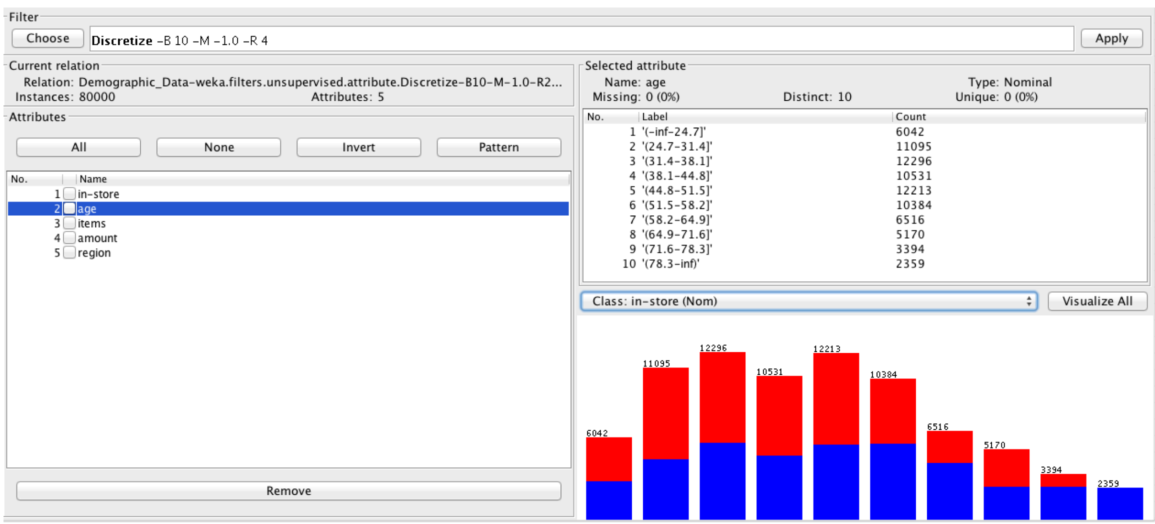
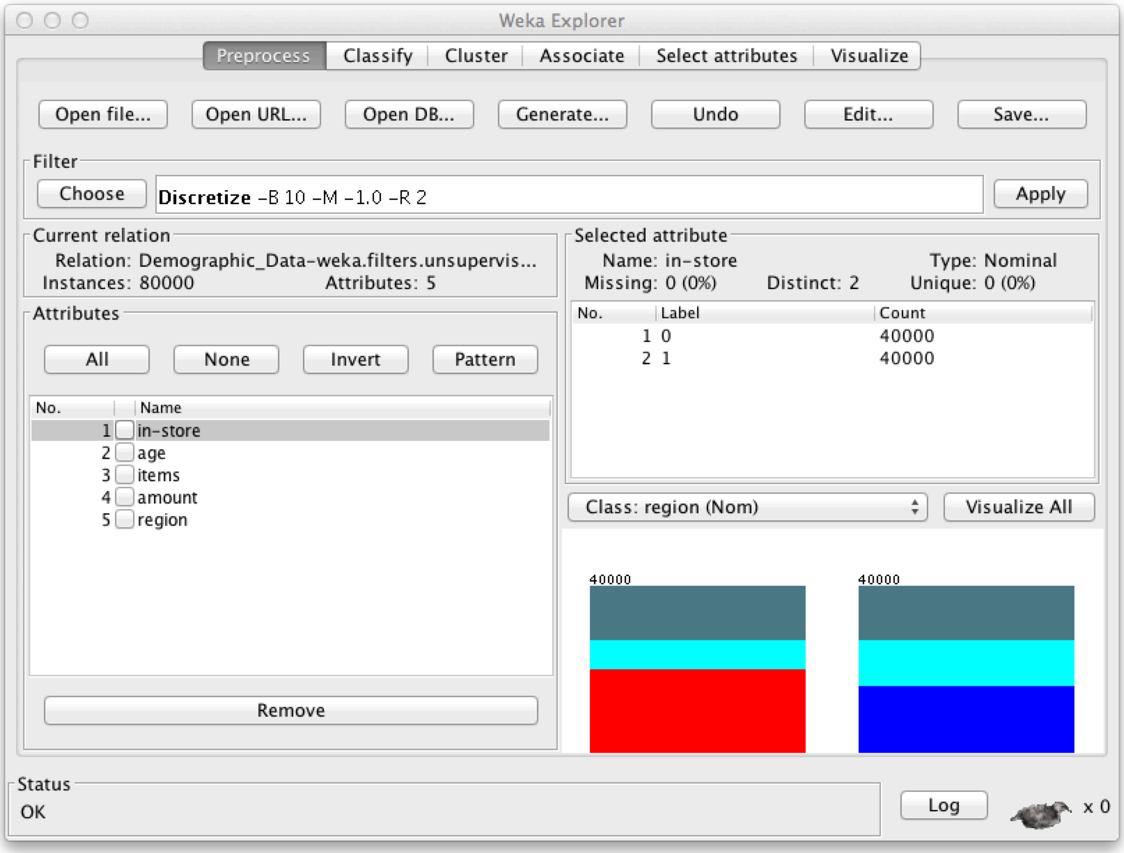


Figure 5 Amount spent online/in-store by each Age group. Online (Red), In-store (Blue)



Machine Learning Methodology

M5P is a ML algorithm that has two parts. The first part is a decision tree that splits in two to create a minimal number of sub-layers. The leaves are linear regressions that run on the assumption that some attributes are fixed or have a slight difference that is deemed irrelevant for reclassification. Knowing from the previous direct observations and algorithms that Region, Items and In-store attributes had little impact on determining the spent amount, we used this assumption at the top of the M5P tree and simplified the problem to two variables: Amount and Age.

Figure 6 M5P Pruned Tree run on Age

```
=== Run information ===

Scheme:weka.classifiers.trees.M5P -M 4.0
Relation:     Demographic_Data-weka.filters.unsupervised.attrib
Instances:    80000
Attributes:   5
              in-store
              age
              items
              amount
              region
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

M5 pruned model tree:
(using smoothed linear models)

region <= 2.5 :
|   in-store <= 0.5 : LM1 (20000/19.778%)
|   in-store >  0.5 : LM2 (16000/66.273%)
region >  2.5 :
|   in-store <= 0.5 : LM3 (20000/115.629%)
|   in-store >  0.5 : LM4 (24000/62.688%)

LM num: 1
amount =
```

```
M5 pruned model tree:
(using smoothed linear models)

region <= 2.5 :
|   in-store <= 0.5 : LM1 (20000/19.778%)
|   in-store >  0.5 : LM2 (16000/66.273%)
region >  2.5 :
|   in-store <= 0.5 : LM3 (20000/115.629%)
|   in-store >  0.5 : LM4 (24000/62.688%)

LM num: 1
amount =
        0.2913 * in-store
        + 0.018 * age='(71.6-78.3]','(64.9-71.6]','(58.2-64.9]','(51.5-58.2]','(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]',
        + 0.0139 * age='(64.9-71.6]','(58.2-64.9]','(51.5-58.2]','(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
        + 0.0974 * age='(58.2-64.9]','(51.5-58.2]','(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
        + 0.0781 * age='(51.5-58.2]','(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
        - 3.153 * age='(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
        + 0.0129 * age='(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
        + 0.0065 * age='(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
        + 0.0375 * age='(24.7-31.4]','(-inf-24.7]'
        + 0.25 * age='(-inf-24.7]'
        + 0.088 * region
        + 253.2357
```

```
LM num: 2
amount =
	0.3762 * in-store
	+ 0.0123 * age='(71.6-78.3]','(64.9-71.6]','(58.2-64.9]','(51.5-58.2]','(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]',
	+ 0.0077 * age='(64.9-71.6]','(58.2-64.9]','(51.5-58.2]','(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
	+ 0.0974 * age='(58.2-64.9]','(51.5-58.2]','(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
	+ 174.1632 * age='(51.5-58.2]','(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
	+ 65.0396 * age='(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
	+ 36.0995 * age='(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
	+ 0.0065 * age='(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
	+ 0.0411 * age='(24.7-31.4]','(-inf-24.7]'
	+ 199.7341 * age='(-inf-24.7]'
	+ 0.088 * region
	+ 520.5206

LM num: 3
amount =
	-0.571 * in-store
	+ 0.0336 * age='(71.6-78.3]','(64.9-71.6]','(58.2-64.9]','(51.5-58.2]','(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]',
	+ 0.0316 * age='(64.9-71.6]','(58.2-64.9]','(51.5-58.2]','(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
	+ 0.1574 * age='(58.2-64.9]','(51.5-58.2]','(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
	+ 0.0467 * age='(51.5-58.2]','(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
	+ 0.0225 * age='(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
	+ 17.7414 * age='(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
	+ 0.006 * age='(24.7-31.4]','(-inf-24.7]'
	+ 0.0987 * age='(-inf-24.7]'
	+ 0.2692 * region
	+ 1533.522


LM num: 4
amount =
	-0.4825 * in-store
	+ 0.0336 * age='(71.6-78.3]','(64.9-71.6]','(58.2-64.9]','(51.5-58.2]','(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]',
	+ 0.0316 * age='(64.9-71.6]','(58.2-64.9]','(51.5-58.2]','(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
	+ 0.1445 * age='(58.2-64.9]','(51.5-58.2]','(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
	+ 0.0436 * age='(51.5-58.2]','(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
	+ 0.0202 * age='(38.1-44.8]','(44.8-51.5]','(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
	+ 0.0187 * age='(31.4-38.1]','(24.7-31.4]','(-inf-24.7]'
	+ 0.0082 * age='(24.7-31.4]','(-inf-24.7]'
	+ 0.0932 * age='(-inf-24.7]'
	+ 503.2068 * region
	- 988.2531

Number of Rules : 4

Time taken to build model: 384.6 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.6708
Mean absolute error                    399.4745
Root mean squared error                534.9465
Relative absolute error                 68.784  %
Root relative squared error             74.167  %
Total Number of Instances             80000
```

Observations

In-store/online or region attributes don't have a significant impact on the spent amount or the number of products a customer buys. It is Age attribute the main factor in determining the amount a customer spends.

GENERAL CONCLUSIONS

Based on our findings for the main three sales indicators we addressed the management and marketing questions as follows:

- Do customers in different regions spend more per transaction? Amount Spent depends mostly on customer's Age. Moreover, the amount per transaction tends to be consistent throughout age groups regardless of the customer's region of origin. Therefore, region doesn't have a significant impact on determining customer expenditure per transaction except for the oldest age group. Based on these findings age-oriented marketing strategies would be more effective in increasing customer's spent amount than any other kind of marketing strategies.

- Which regions spend the most/least? Region 4 spends the most, while Region 2 spends the least. These findings are also consistent with the age composition of both regions. Region 2 has the oldest population which spends the least per transaction. While Region 4 has the largest population of middle-aged customers ($44.8 \leq$ Customers ≤ 58.2) who typically spend more per transaction than any other age groups.

- Are there differences in the age of customers between regions? Each region shows a particular age composition that largely determines its current and potential sales. If so, can we predict the age of a customer in a region based on other demographic data? We can predict with certain level of confidence the age of a customer knowing his/her region's age composition and expenditure level.

- Can we predict the amount a customer will spend per transaction based on other data we have collected about that customer? By knowing the customer age and region we can predict what he/she is likely to spend since we can determine a typical spent amount for each age group in each region.

- Is there any correlation between age of a customer and if the transaction was made online or in the store? Region 1 customers only buy in-store while Region 2 customers only buy online. Further analysis is required to understand extrinsic factors that may be causing this shopping behavior. However, for Region 3 and Region 4 there is not a strong correlation between age of a customer and online or in-store shopping.

- Do any other factors predict if a customer will buy online or in our stores? Together, the amount of items, expenditure level and age would be able to predict in a very limited way if the customer is going to buy online or in-store because these variables are not strongly correlated. Age and amount are the most strongly correlated attributes.

RECOMMENDATIONS

Our findings support the hypothesis that Blackwell's main growth opportunities rely on developing age-targeted marketing strategies. According to the available data and its analysis we cannot ascertain the impact of future website improvements on the overall sales. Our analysis is inconclusive about that specific matter. However, the fact that some significant populations only buy online or in-store gives room to work on regional-age group marketing strategies. This is the case for Region 2 where further data analysis would lead us to understand the specific factors that created this region's online habits. In general, we recommend conducting a deeper data analysis to gain insight on online/in-store customer behavior.