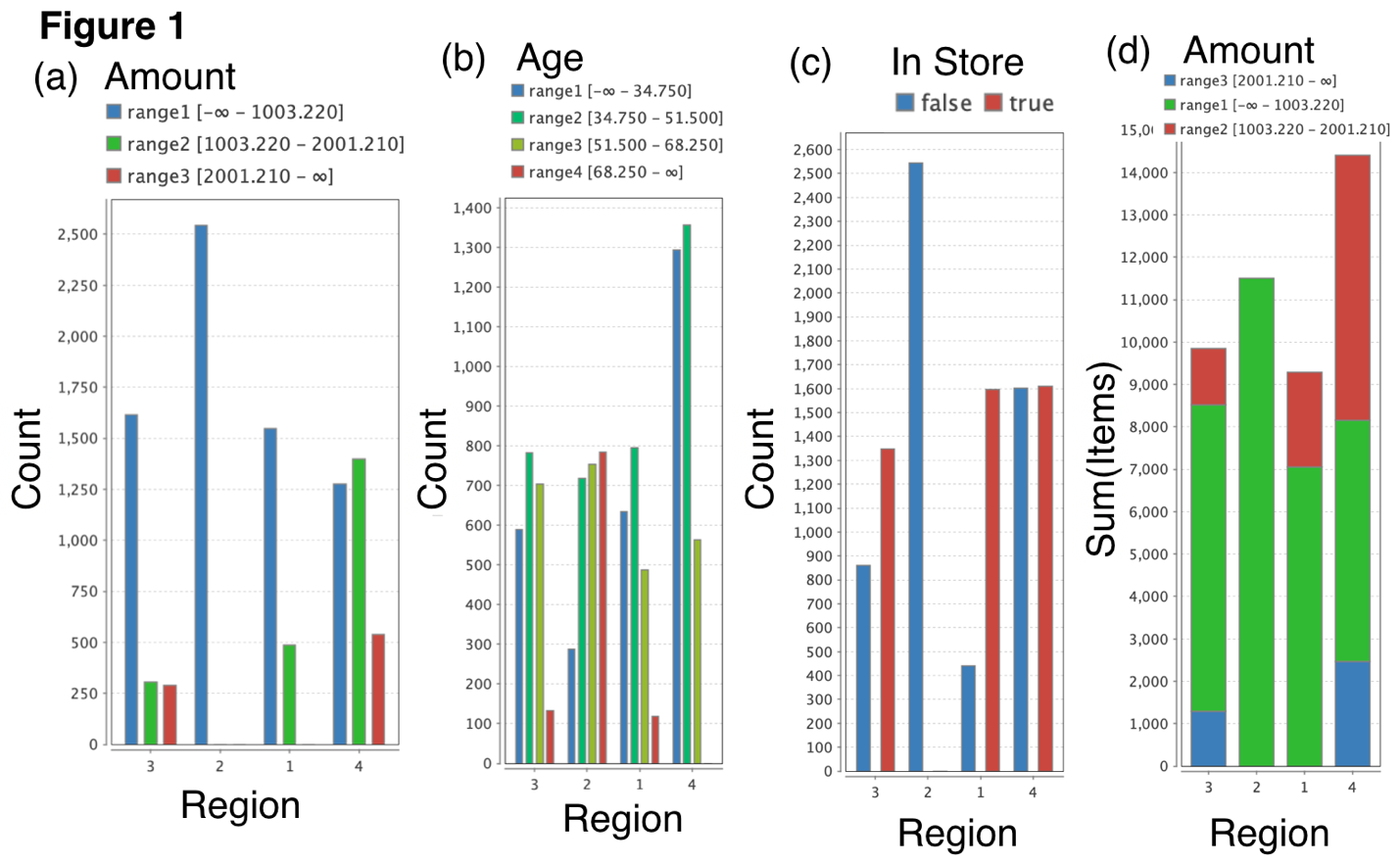
**Customer Buying Patterns Report**

John T. Leonard

**Abstract**

Here, we investigate customer buying patterns from four store regions. We begin by analyzing the data and trends using simple visualization methods, predominantly relying on bar charts and histograms. Following this analysis and general insights, we analyze the predictions from a decision tree model.

**Analytical Procedure**

The “Blackwell\_Hist\_Sample.csv” data set was used to gain insight into customer buying patterns. The data analysis was carried out using the “RapidMiner” software package. The data set analyzed contained information on the amount sold (“amount”), customer age (“age”), whether the purchase of interest was in-store or not (“in.store”), and the number of items purchased (“items”), for each region. To simplify the analysis, the “age” was discretized (binned) into four discrete ranges: < 35, 35-52, 52-68, and >68 years old (y.o.). Additionally, the “amount” was discretized into three bins: <1K , 1K-2K , and >2K.

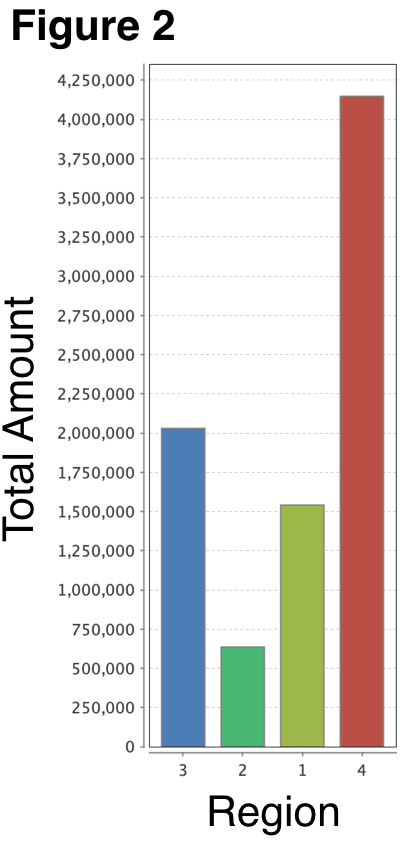
**Region & Amount Spent per Transaction**

Figure 1 summarizes the count/frequency of the amount (a), age (b), in-store (c) bins for each region, as well as the total (sum) of items for each region in histogram form. Focusing on the amount spent vs. region (Figure 1(a)), we see region 2 has the highest number of customers purchasing items in the low price range (<1K), while Region 4 has the highest number of customers purchasing items in the high price range (>2K). It is also of note that region 2 only has customers purchasing in the <1K price range, while region 4 has a more distributed customer base, with ~40% of the customers making purchases at the low and intermediate price ranges and ~20% of the customers making purchases at the high price range. Based on these trends of the frequency of customer buying in Region 2 and 4, it is not surprising that Region 2 has the lowest total amount spent by customers, while region 4 has the most. This is shown in Figure 2, where we plot the total (sum) amount for each region.

Aside from Region 2 and 4, we can see region 3 has the 2nd highest total sales, followed by region 1. The predominant difference between Region 1 and 3 is the number of customers making high price range purchases. Specifically, Region 1 has no customers purchasing high price items, while Region 3 has ~13% of their customers purchasing these items.

In summary, we generally observe that having customers that purchase high value items has a strong impact on total amount sold.

**Region & Customers Age**

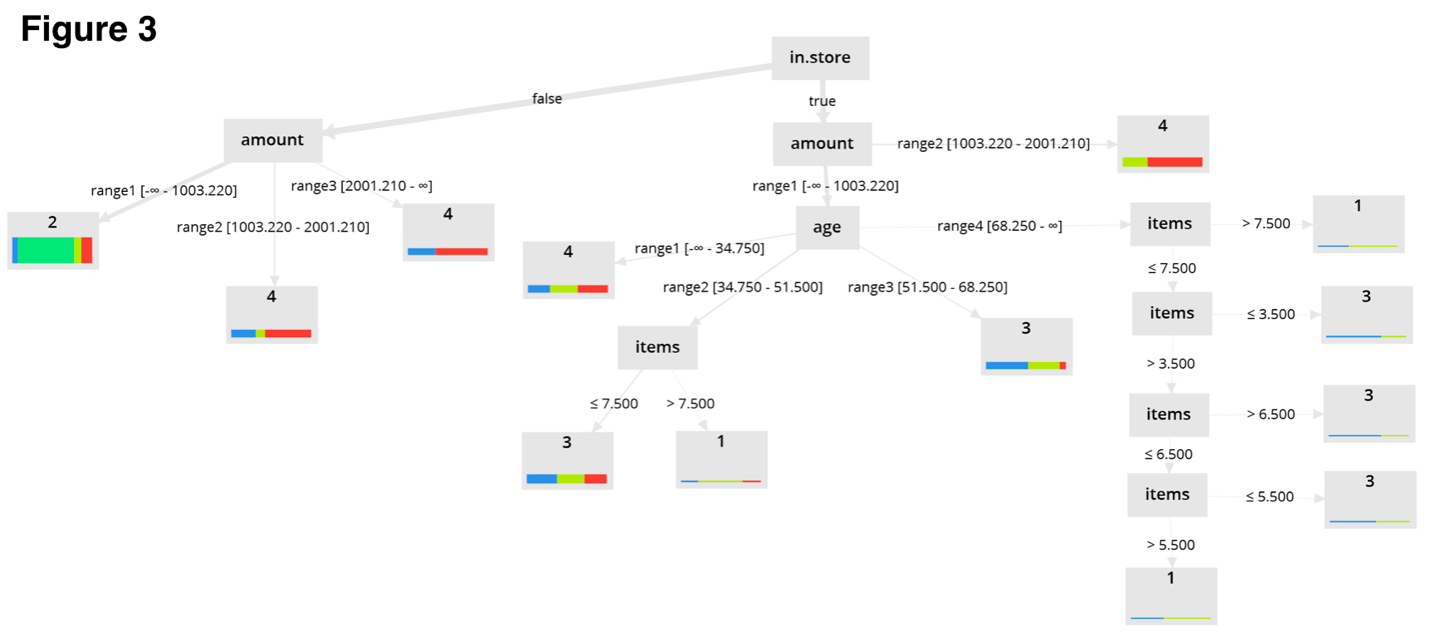
In Figure 1(b), we see the frequency/count of each age group in each region. We start by focusing on our two extreme regions from the previous section: Region 2, with the lowest sales, and region 4 with the highest sales. From Figure 1(b), we can see Region 2 has the highest number of >65y.o. customers, while Region 4 has the lowest amount (none). Also of note is the fact that Region 2 has a very small number of young customers in the <35y.o. range, while region 4 has the highest number of customers in this class and the 35-52 y.o. class. Across all regions, there is a comparably small variation in the number of customers in the 52-68y.o. range.

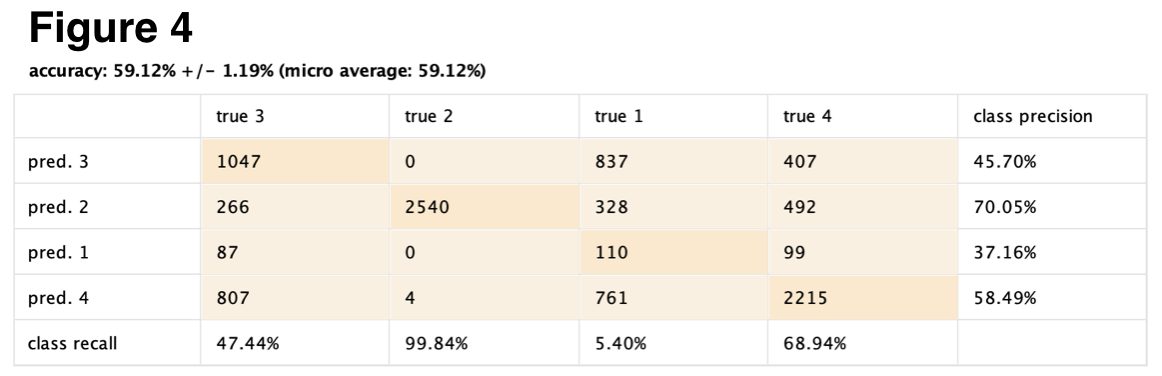
Taking these observations together with the trends we observed in our analysis of the amount per region, we hypothesize that customers >68y.o. have a preference for purchasing lower priced items, while younger customers, particularly those <52y.o., are more willing to purchase more expensive items. This implies that we can roughly predict the age of a customer, based on the amount they spend. This particular prediction will be discussed more quantitatively in the Decision Tree Model section below.

**Number of Items & Amount Spent**

Figure 1(d) shows the cumulative (sum) number of items vs. region. Here, we see Region 4 exceeds all other regions by ~24%. Observing the stacked columns within each region, we can see the correlation between the number of items and the amount spent by region. Here, we see that the least expensive items (i.e. <1K price range), are also the items that are bought in the highest volume with each purchase. In contrast, we see the most expensive items are bought in the lowest volume. These trends are consistent with general consumer behavior, however one important implication is that they highlight the fact that if the company needs to move volume, they should reduce prices to the <1K range, whereas if volume is low, then increasing prices to >2K will help increase total amount earned for a given region

**Decision Tree Model**

Now that we have an overview of the general trends in the customer buying patterns, we are ready to gain additional insight into these trends in a more quantitative manner by predicting the region of purchase using a decision tree classifier algorithm. In this algorithm, we run k-fold sampling, with a k value of 10. Additionally, the maximum depth for the tree is set to 10 and maximal depth pruning, as well as pre-pruning is applied to prevent over-fitting of the data. Figure 3 shows the resulting decision tree, while Figure 4 shows the confusion matrix for the tree.



Viewing Figure 3, we can see that the first split in the tree is on whether or not the purchase was made in-store. Below in-store, we see the next decision nodes occur for amount. Viewing this portion of the tree, we can see that if our purchase was made in-store, then we can guarantee that it did not occur in region 2. This particular observation is also evident in the bar charts in Figure 1(c), where we can see that region 2 had no in-store purchases. Furthermore, if our in-store purchase occurred in the intermediate price-range (1-2K), it is likely that it occurred in Region 4.

If the in-store purchase was for a low-priced item (<1K), then we must consider the age of the customer (the next branch). This branch indicates that in-store purchases of low price items (<1K) by a young person (<35y.o) are likely to have occurred in region 4. Considering Figure 1(b) and (c), this result is not that surprising since Region 4 has the largest frequency/population of <35y.o. costumers, as well as the highest number of in-store purchases (though only slightly more than Region 1). Similarly, in Figure 1(b) and (c), we can see the logic behind the decision tree prediction that in-store purchases of low price items (<1K) by 52-68y.o. customers are likely to occur in Region 3, since this region has a large number of in-store purchases, as well as a relatively large number of customers in the 52-68 y.o. price range.

Finally, the last two branches of our in-store purchases side of the tree are for the 34-52 and >68 y.o. cohorts. Here, we see these branches leading to nodes which split based on the number of items purchased. We will not described all these nodes in detail, but the general trend is that 34-52 y.o customers making in-store purchases of low priced items are likely to come from Region 3 if they are buying less than or equal to 7 items, while they are more likely to come from Region 1 if they are buying greater than 7 items. In contrast, if they are older than 68 y.o., then they are more likely to be from Region 1 if they buy great than 7 items, and are likely to be from Region 3 if they buy less than 7 items.

Viewing the out-of-store purchases side of the tree, we see many fewer layers, with only Region 2 and Region 4 being likely Regions for out-of-store purchases. This can be intuitively understand from Figure 1(c), where we see Region 1 and 4 having a significantly higher number of out-of-store purchases than the other regions. Additionally, as we observed in the previous sections, we see the decision tree predicts an out-of-store purchase is likely to come from Region 2 if it is a low priced item (<1K), while higher priced items, bought out-of-store, are more likely to have been purchased from Region 4.

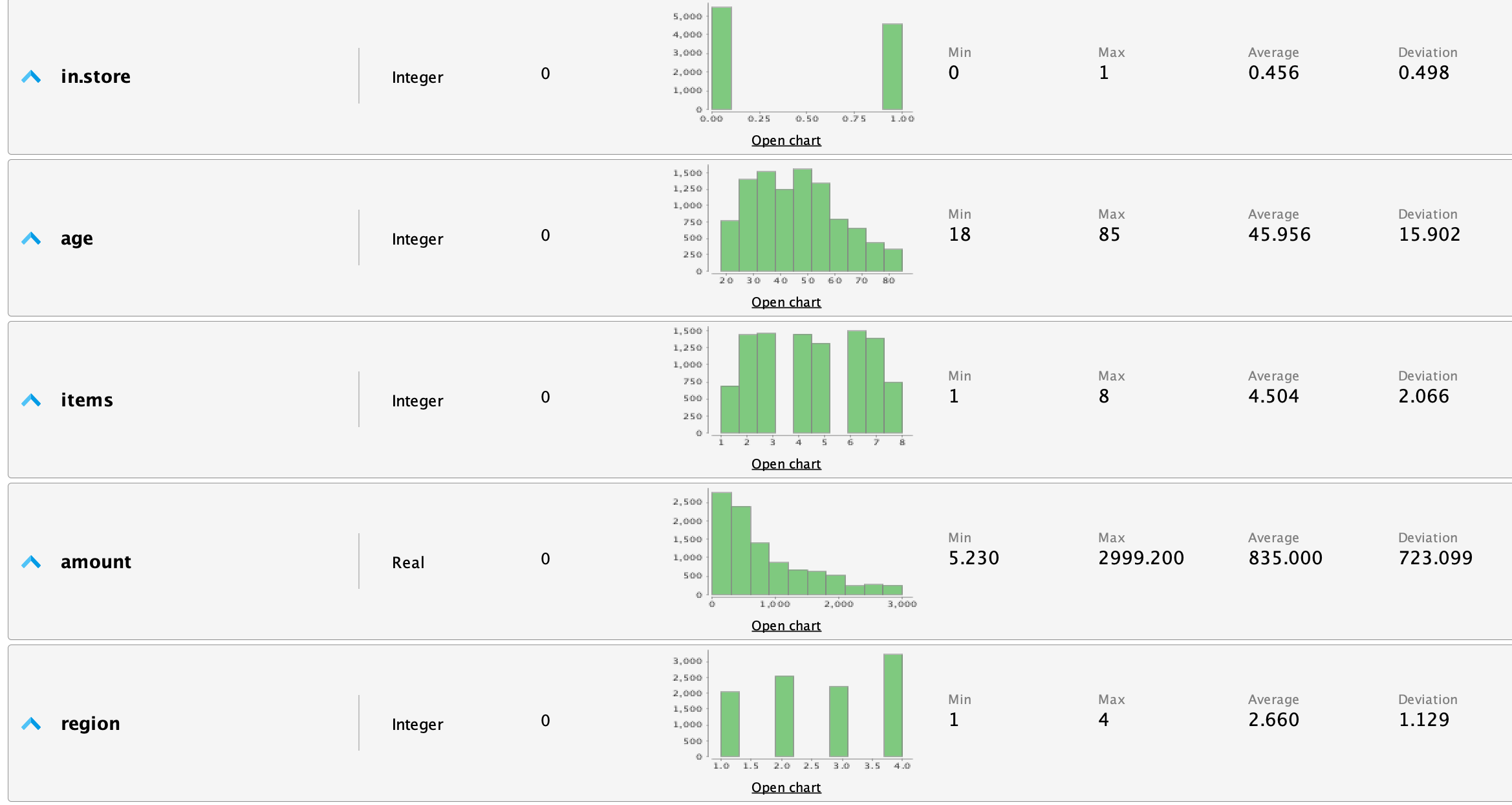
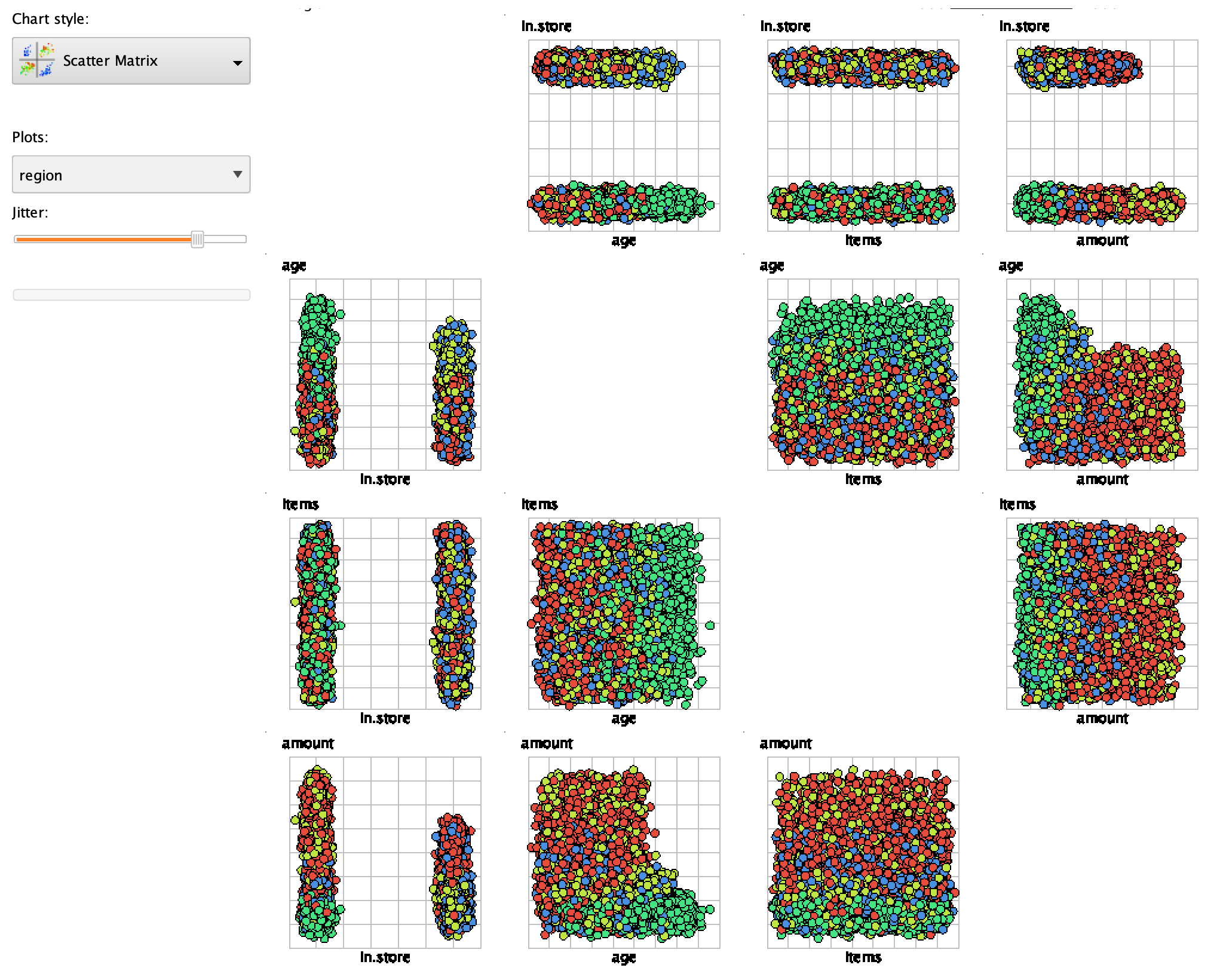
**Conclusions**

We have reviewed in detail the customer buying patterns both from a basic data visualization approach, primarily utilizing box plots/ histograms, and from a more sophisticated decision tree model. We observed that many of our intuitions from the data visualization approach are reflected in the decision tree. We can also see that some broad classificaitons, such as if an in-store purchase came from Region 2 or not, have nearly 100% probability of being true, while many more specific classifications can give general guidance, but do not have as high of a probability of being precise and/or accurate.

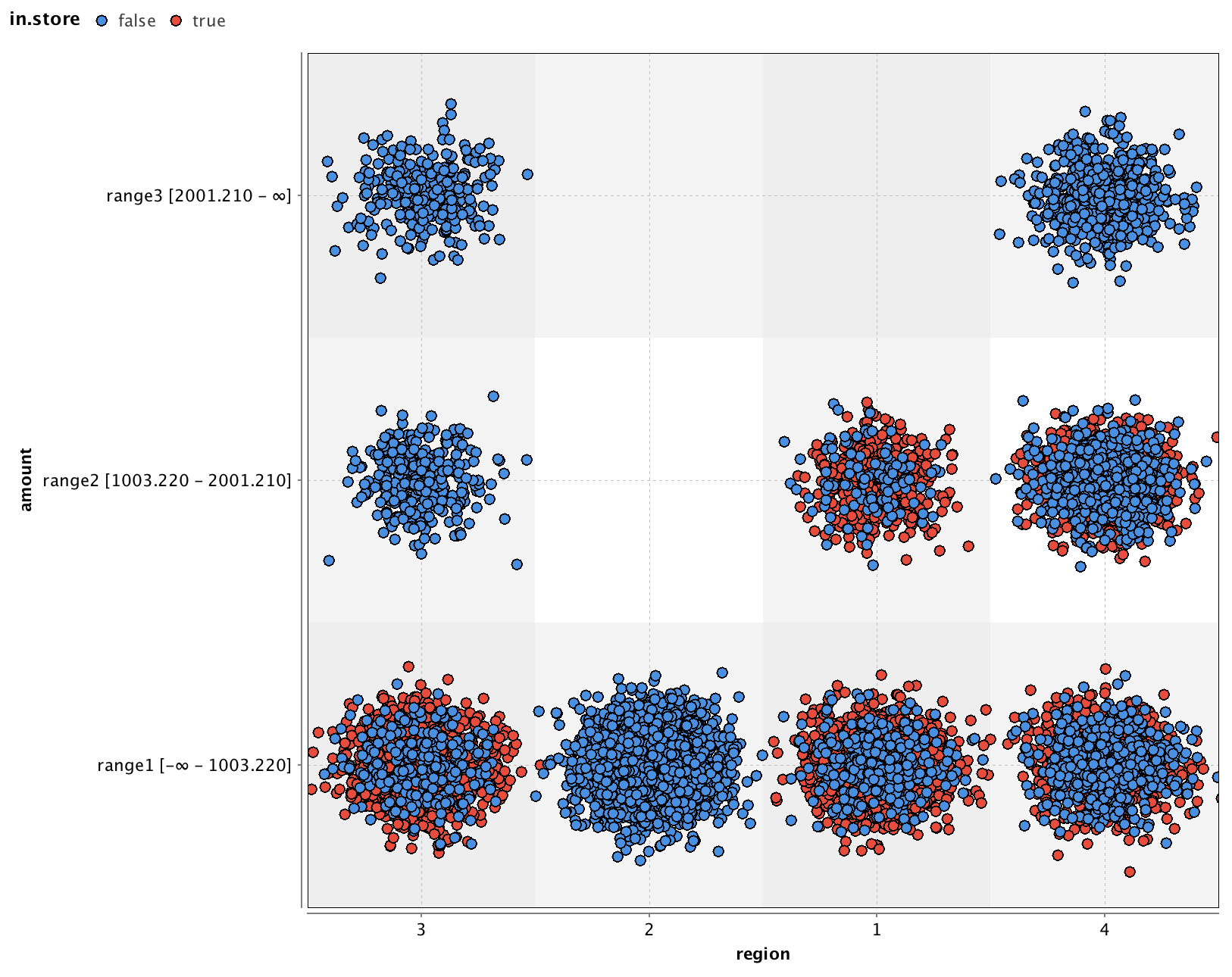
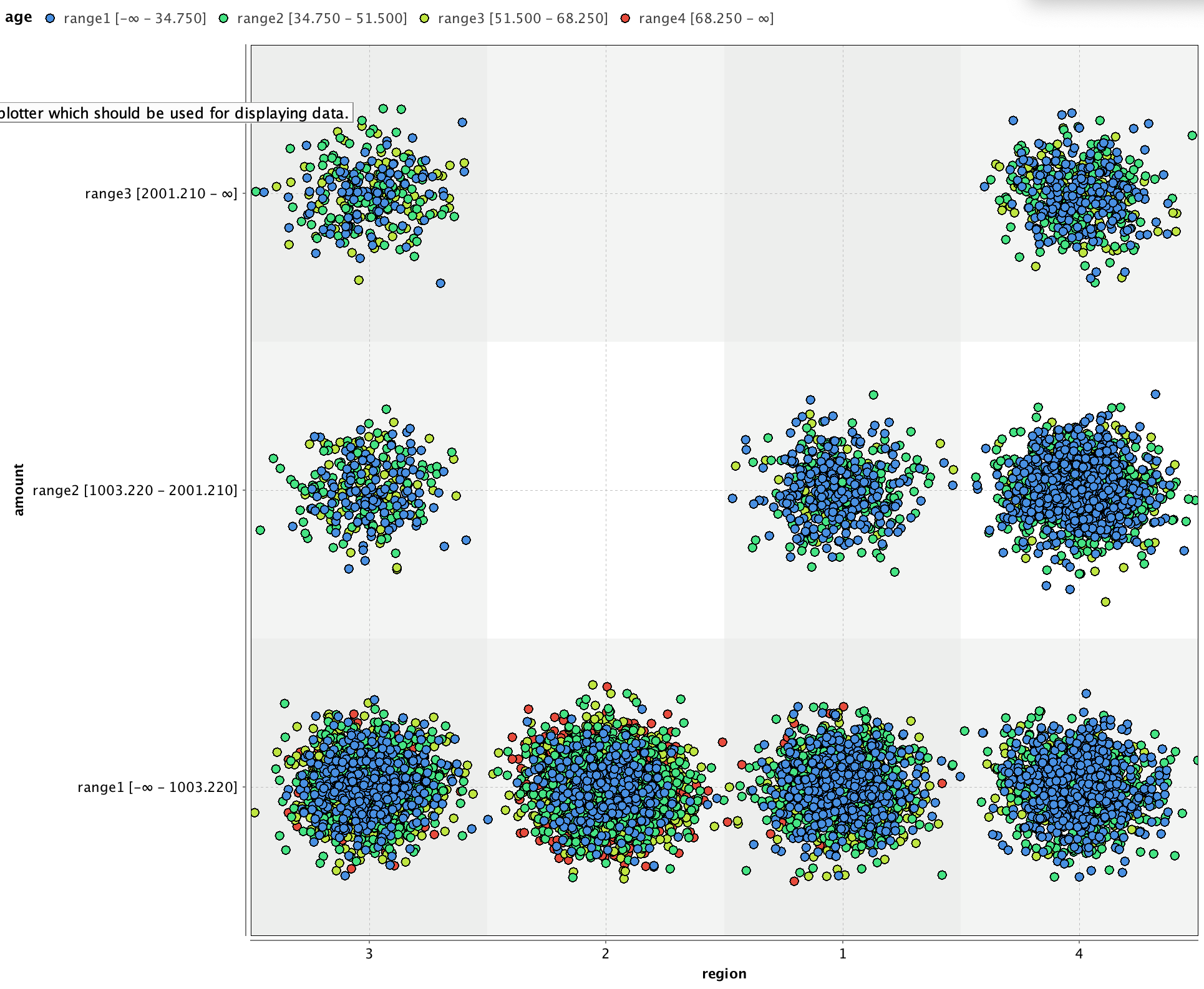
Appendix

Notes From Plan of Attack

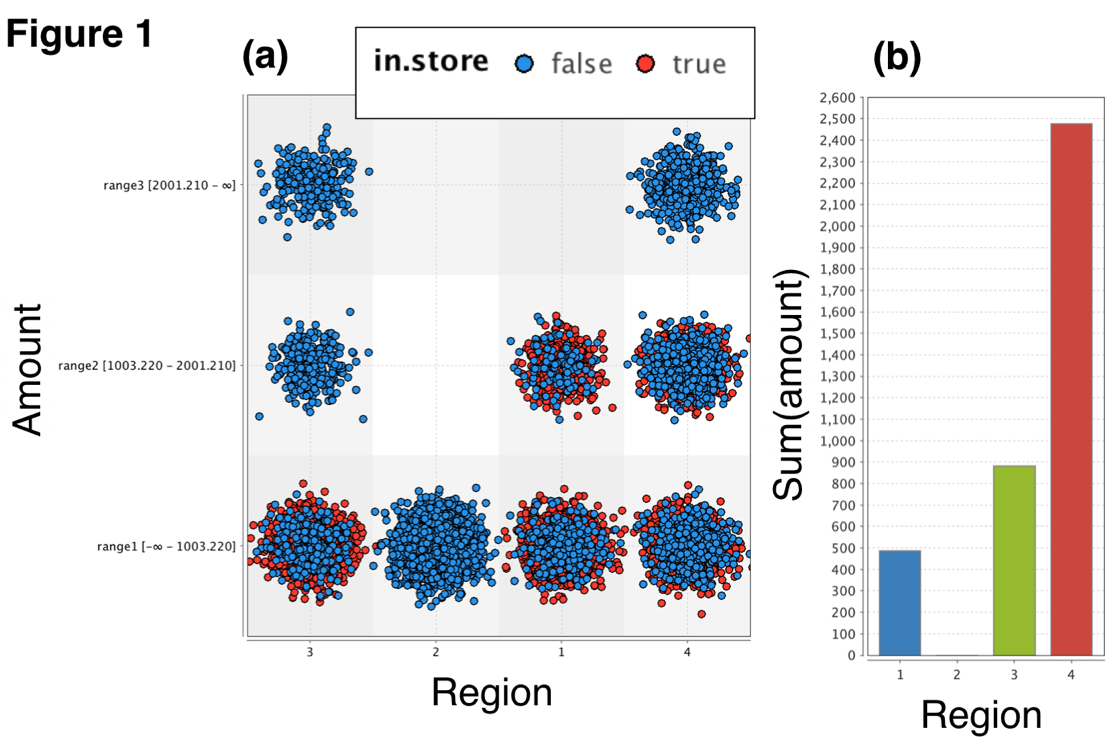
Import, Visualize, and Pre-Process Data

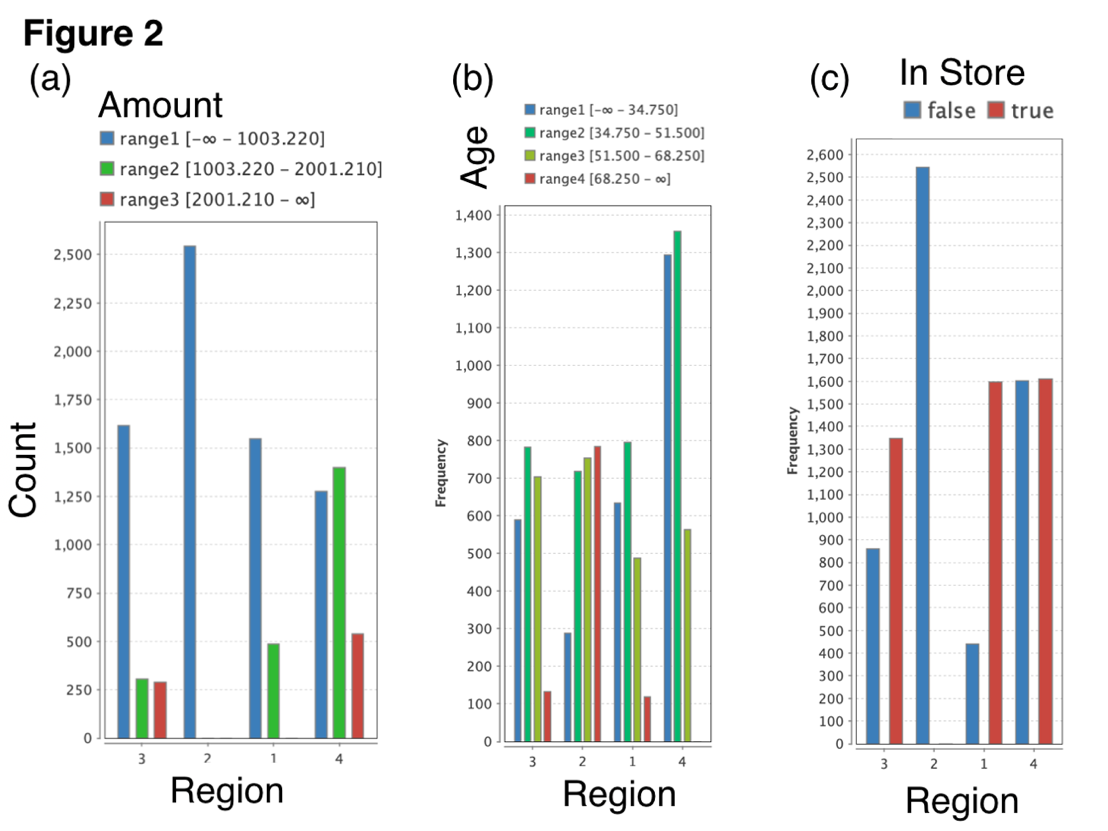
1. Note the Min, Max and Average Values for the Numeric data.
2. Note the values for Nominal data.
3. Are there any missing values in any of the features?
   * No missing data
4. Do you see anything that might be unusual? Is any of the data more heavily distributed with respect to any particular feature? Why or why not? Make a note of your findings for your report and for further analysis.
   * “amount” column has a large variation in the data, however, the histogram does not look very unusual. Almost like half a normal distribution.
5. Next, click on the 'Charts' tab on the left side of RapidMiner.
   1. Using a histogram visualize the various distributions of each attribute in the data.
   2. This is also a good time to use scatter plots to compare the relationships between any two features.
   3. Finally, familiarize yourself with other charts in RapidMiner and feel free to experiment to see what other findings you can glean in the data.
      * See scatter matrix below:
        + No string continuous trends in the data, however a number of classifications can be observed:
          - Strongest relationship seems to be between amount, age, and region

Investigate the Relationship Between Region and Amount Spent Per Transaction

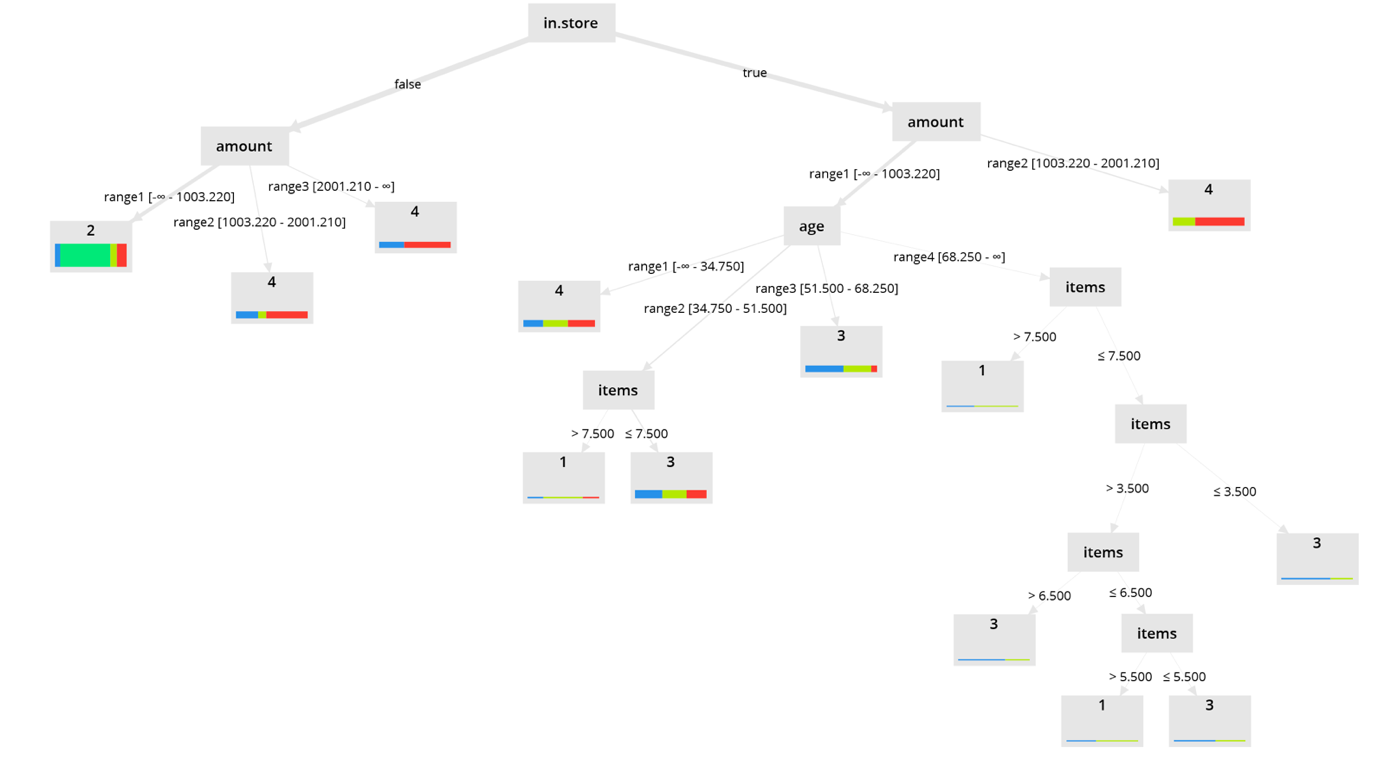
1. Investigate the specific relationship between the region of purchase and the amount spent:
   * All regions roughly show highest volume for lowest price class.
     + This price class also contains the largest distribution in age cohort (color), with region 2 having the largest volume of the eldest age cohort (>68ys)
     + Switching color from “age” to “in.store” reveals in.store purchases do not occur at all for the most expensive price class. Furthermore, in-store purchases are not occurring in region 2. This suggest that there may be a preference for elder people, or people in region 2, or both, to prefer out of store purchases.

4. Investigate the Relationship Between the Region of Purchase and a Customers Age

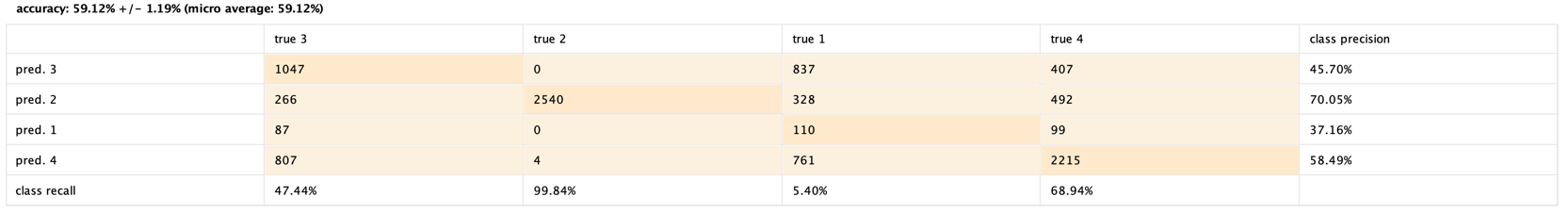
1. The stakeholder question to be considered in this section can be addressed in two ways:
   1. By looking at the histogram and scatter plot data
      1. Stakeholder question: Do customers in different regions spend more per transaction? Which regions spend the most/least?
         * Yes, in Figure 1 below, we see two ways of visualizing the relationship between region and amount spent per transaction. In Figure 1(a) we see a scatter plot of region vs amount with the color denoting whether the item was bought in-store (red) or not (blue). Here, we can see that regions 3 and 4 have low, intermediate, and high price range costumers, while region 1 only has intermediate and low, and region 2 only has low price range customers.
         * To gain more insight into the quantitative value of these distributions, we can view the bar chart in Figure 1(b), which shows the total sum of amount sold per region. Here, it is evident region 4 is selling the most, followed by region 3, then 1, then 2. Notice in Figure 1(a) that region 2 does not have any in-store purchases, and does not have any purchases in the intermediate and high price ranges. This suggests that we would generally predict a low total sales amount from stores with no in-store purchases, and no intermediate and high price range purchases

* We can further investigate the features responsible for variation in spending in each region by analyzing the count/frequency per region, grouping by amount, age, and in-store/out-of-store. These histograms are shown in in Figure 2. Focusing on Figure 2(a), we see, region 4, which has the highest total sales, also has the highest number of sales in the low + intermediate range. Region 3, with the 2nd highest total sales, has the 2nd highest cumulative sales of low + intermediate priced items. Region 1 also has significant sales in the low + intermediate, however it does not have any high price range sales. Finally, region 2 with the lowest volume of sales, has no intermediate or high volume sales. These trends imply that we would generally predict a higher probability of total sales for regions with higher total sales of intermediate + high price range items.
* Next, in Figure 2(b) we see the frequency of sales grouped by age class (blue: <35 y/o, red: >68y/o). In most of the regions we see fairly similar relative counts of <35y/o and 35-52y/o cohorts. Furthermore, across all the regions, there is a relatively similar frequency/count for the 52-68 y/o cohort. What is most noteable about the age distribution is the variance in the >68y/o cohort purchasing frequency. Specifically, we see region 4, which had the highest sales volume, has no sales from the 68y/o cohort, while region 2, with the lowest sales volume, has the highest number of >68y/o customers. This implies that we would generally predict a higher number of sales for a region with more of the costumers below the age of 68y/o.
* Finally, we consider Figure 2(c), where the frequency/count per region is grouped by in-store (red) vs. not-in-store (blue). Here the most notable data point is the fact that region 2, with the lowest amount of sales, also has the lowest the highest amount of not-in-store sales. There may be a number of reasons for this: >65y/o cohort may prefer to not purchase in-store; not-in-store purchases may be cheaper generally; region 2 may have it’s store(s) in a bad location, leading to more customers making not-in-store purchases.

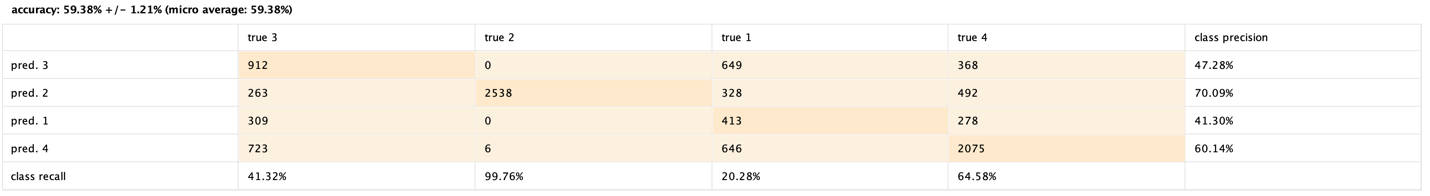
1. By running a decision tree operator on the “region” attribuite.

* To quantify the relationship between region and the various features within each region in a more rigorous manner, we use a decision tree analysis on the region label, with splitting determined by the gain ratio to determine the optimal splitting criteria.
* Here we see the largest split occurs on the in-store criteria. If the purchases were not-in-store then the strongest parameter determining which region the purchase occurred in is the amount, with region 2 having the lowest amount, and region 4 being most likely to have the intermediate and high price ranges, as seen previously.
* For the case of in-store purchases, we can focus of 2 price classes: low (range 1) and intermediate (range 2). If the purchase is in the intermediate price range, than it is likel to come from region 4. If it is in the low range, then our decision tree splits into more branches, namely, age, then items. The relationships can be seen in the decision tree, and we won’t explicitly state them here. The main take-away from these relationships is that for in-store purchases in the lowest price range/class, age, followed by number of items, are the factors that most heavily influence the performance of region.

5. Machine Learning: Classifying Where a Transaction Took Place

* What do you think the depended variables are for predicting the region of a sale
* Amount, age, in.store, items
* Examine the performance metrics. Does accuracy indicate a good or poor model fit?
* I would say the accuracy indicates a poor model, since it is ~59%, though there are some recall cases (true 2) which have high probability (99% and 69%)

6. Machine Learning: Classificaiton – Understand Items and Amount Spent

* Examine the decision tree. What, if anything, is the role of amount?
* Amount is the 2nd most important / strongest feature, aside from in-store.
* Examine the performance metrics. Does the accuracy indicate a good or poor model fit
* Applying random forest (and the other ML models variants suggested) generally yields and accuracy of ~59%, so there isn’t really an improvement over the single decision tree. This isn’t that surprising since there aren’t too many classes
* What business insights might you glean regarding the relationship between the number of items and amount spent
* Essentially, for out-of-store purchases, the items do not really impact the amount spent in each region. There may be some unknown feature unique to different regions that impact the amount people spend on out-of-store purchases, since customers seem to prefer to spend more in region 4. For in-store purchases, the number of items predominately affects the amount spent for the low price range items. More specifically, middle aged