Task 1: Investigate Customer Buying Patterns



**FROM:** Danielle Sherman  
**Subject:** Investigating customer buying patterns

Hello,

As CTO and head of Blackwell's eCommerce Team, I'd like to welcome you aboard. I'm excited to get started on this project, but I'd first like to give you a bit of background to get you up to speed. Blackwell has been a successful electronics retailer for over 40 years, with over 30 stores in the Southeast. A little over a year ago we launched our eCommerce website. We are starting to build up customer transaction data from the site and we want to leverage this data to inform our decisions about site-related activities, like online marketing, enhancements to the site and so on, in order to continue to maximize the amount of revenue we generate from eCommerce sales.

For example, our VP of Sales, Martin Goodrich, thinks that customers who shop in the store are older than customers who shop online and that older people spend more money on electronics than younger people. He is considering some marketing activities and potentially some design changes to the website to attract older buyers. I, on the other hand, believe that the differences in transactions and customer demographics may be regional. Before we even consider any additional activities related to the website, we want to gain insight into any factors that can explain how our customers shop and how much they spend.

To that end, I would like you to explore the customer transaction data we have collected from recent online and in-store sales and see if you can infer any insights about customer purchasing behavior. Specifically, I am interested in the following:

* Do customers in different regions spend more per transaction? Which regions spend the most/least?
* Are there differences in the age of customers between regions? If so, can we predict the age of a customer in a region based on other demographic data?
* We need to investigate Martin’s hypothesis: Is there any correlation between age of a customer and if the transaction was made online or in the store? Do any other factors predict if a customer will buy online or in our stores?
* Finally, is there a relationship between number of items purchased and amount spent?

To investigate this, I’d like you to use data mining methods to explore the data, look for patterns in the data and draw conclusions. I have attached a data file of customer transactions; it includes some information about the customer who made the transaction, as well as the amount of the transaction, and how many items were purchased. Once you have completed your analysis, please create a brief report of your findings and conclusions and an explanation of how you arrived at those conclusions so I can discuss them with Martin.

Thanks,

Danielle

**Danielle Sherman**

Chief Technology Officer

Blackwell Electronics

www.blackwellelectronics.com

**Attachments:**

https://common.socraticarts.com/lib/images/windowsIcons/csv.gif [Blackwell\_Hist\_Sample](https://s3.amazonaws.com/gbstool/emails/2895/Blackwell_Hist_Sample.csv?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20200117T231916Z&X-Amz-SignedHeaders=host&X-Amz-Expires=28800&X-Amz-Credential=AKIAJBIZLMJQ2O6DKIAA%2F20200117%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=338fdb26825ffedeaab35bb2906136d1b9e63051d575ef2b1ce45adb97bc9ea1" \t "_blank)

# **Task 1:** Investigate Customer Buying Patterns

[INTRODUCTION](https://codeacademy.ubiqum.com/mc/poa?productID=6637&taskID=4241#introduction)

# Your Task

Blackwell Electronics’ CTO Danielle Sherman has asked you to use data mining methods to explore the customer transaction data collected from recent online and in-store sales to infer any insights and conclusions about customer purchasing behavior, specifically:

* Do customers in different regions spend more per transaction? Which regions spend the most/least?
* Are there differences in the age of customers between regions? If so, can we predict the age of a customer in a region based on other demographic data?
* Is there any correlation between age of a customer and if the transaction was made online or in the store? Or do other factors correlate to an online or in-store transaction?
* Is there a relationship between the number of items purchased and amount spent?

This task requires that you prepare one deliverable:

**Customer Buying Patterns Report.**A zip file that includes:

* A two to three page Word or PowerPoint document in which you:
  + Summarize your findings and conclusions regarding the questions posed about customer purchasing behavior
  + Explain how the results of each classifier you ran support your conclusions
* A narrative explaining the results of each algorithm you ran in RapidMiner, but the technical output is not needed in the report itself. Please only include it in a separate file so the mentors may review your work.

The steps in the following tabs will walk you through this process.

#### **[1. Get Started](https://codeacademy.ubiqum.com/mc/poa?productID=6637&taskID=4241" \l "collapsepoa12295)**

1. **Review the email from Danielle to make sure you understand the details of this task.**
2. **Download and install RapidMiner. RapidMiner** is a Data Science Platform you will use to perform data mining tasks in this course. Refer to the link in the ***Learning* *Resources*** tab to access the download.
3. Read Section One in ***Data Mining for the Masses***
4. Work through the Operators and Processes exercises in the ***RapidMiner Learn***tutorials located within RapidMiner
5. Read the article on ***variables*** in the resources
6. **Download and review Blackwell\_Hist\_Sample*.csv***.
   1. Each row in this data file represents a single transaction that was made at either the bricks & mortar Blackwell Electronics store or on its website.
   2. The "in-store" column indicates where each transaction was made online (0), or in-store (1).
   3. The "age" column indicates the age of the customer who made the transaction.
   4. The "items" column tracks the number of items the customer purchased.
   5. The "amount" column records the amount of money spent on the transaction.
   6. The "region" column indicates in which of the four regions the purchase was made (1 = East, 2 = West, 3 = South, 4 = Central).

**Definitions:**

* Columns with known values in data sets can be referred to in several ways that are relatively synonymous: Attributes, features, independent variables, X-values and predictors.
* The unknown value of a column that you are trying to predict can also be referred to in several ways: Dependent variable, Y-value, y-Hat, output variable, Y-variable.
* Rows in a data set can be referred to as instances or observations

#### **[2. Import, Visualize, and Preprocess the Data](https://codeacademy.ubiqum.com/mc/poa?productID=6637&taskID=4241" \l "collapsepoa12297)**

**Importing and Visualization**

1. **Import your data into RapidMiner.**To analyze the data in RapidMiner, you will first need to import it into the tool. To do this, click the "Add Data" button under the "Repository" tab, navigate to the ***Blackwell\_Hist\_Sample*.csv**file on your computer, and click "Open."
2. After importing the data click on the 'Statistics'  tab on the left side of the RapidMiner screen.
   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?
   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.
3. 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.

**Pre-processing**

1. After data has been imported, it is often necessary to apply filters to the data so RapidMiner interprets it in the way the you expect prior to running an algorithm on the data. Two helpful filters you will use are the "Numeric To Binominal" filter and the "Numerical to Polynominal" filter. Apply these filters to the data using the following steps:

1. Drag the imported data into the Design view; this will initialize the Retrieve Data Operator
2. Next, Use the "Select Attributes" Operator and set the filter type to all after connecting the 'out' port of the retrieve operator to the 'exa' port of the select attributes operator
3. Convert the "in-store" purchase data from a numeric range to nominal classes using the "Numeric to Binominal" operator.
4. The form our data from the "in-store" column takes—0s and 1s—may look like a numeric range to RapidMiner if we were to run an algorithm without applying any filter. But, in this case, we actually want our "in-store" data to be seen by RapidMiner as nominal labels rather than a numeric scale. This is because numbers were used to represent two classes of nominal data:  "1" to represent an in-store transaction and a "0" to represent an online transaction.
5. In the "Operators" tab, click in the search box and type 'numeric'; the find the 'Numerical to Binominal' operator and drag it into the design view.
6. Connect the 'exa' from 'select attributes' to the 'exa' of the "Numerical to Binominal" operator
7. In the 'Numerical to Binominal' operator parameters change the attribute filter type to single and the attribute to the in-store attribute of the data set.
8. Repeat the same process using the "Numerical to Polynominal" operator and the "region" attribute in the dataset.

**TIP:**

At this time, you can connect the example port from the last operator in your process to the results port in RapidMiner and run the process. After doing this you can look at the statistics tab in the results view and verify that your changes have been made correctly.

2. After you have applied filters to "region" and "in-store", turn your attention to the filter called Discretize.

1. If you click on "age" you may notice that the Statistics information on the right shows the age as a continuum from 18 to 85. We could leave it like this, but it is often much more useful to "bin" the continuous data, like ages or scores, into more manageable groups. This is called *discretizing* the data.
2. To Discretize the data you will pick up where you left off after the last operator you added to the design process. You will visit the operators sections and look for the Discretization operator.
3. Just like you have previously done with converting datatypes you will add the Discretization operator to the design view at the very end of the process you have already built.
4. Next you will connect the Discretization operator to the last operator you added and then change the parameters within the Discretization operator to meet your needs.
5. Starting out with 4 bins is a good starting point for dividing the ages into more manageable chunks.
6. Now that you'e discretized the "age" attribute, go ahead and do the same for the other numeric feature by dividing the "amount" attribute into 3 bins.

**TIP:**

What just happened? You won’t know this just yet since you haven’t run your current process, but applying discretization to a continuous numeric attribute actually changes the attribute to a nominal data type. Why do you think this occurs? Feel free to discuss this with your classmates or reach out to your mentor if you have an idea about this. (Hint: it is very important that you know this, but we are asking you to investigate it further).

Segmenting the data into blocks that you care about may be more useful than specific numeric values**.** In the case of age, for example, separate marketing campaigns that target 18.5 year olds, 24.6 year olds, 43.6 year olds and so on would not be realistic, however separate marketing campaigns that target teens, young adults and middle aged adults would make sense. Conversely, knowing the specific value of a diagnostic test may be more important than a range of values.

3. Finish Up Preprocessing

1. Now that you have finished preprocessing the data go ahead and connect the last operator to the results port and run the process. Check the statistics tab to make sure that all the data is in the correct format.
2. This is also a great time to go ahead and save your process by clicking on file> save process in RapidMiner.

#### **[3. Data Mining: Investigate the Relationship Between Region and the Amount Spent Per Transaction](https://codeacademy.ubiqum.com/mc/poa?productID=6637&taskID=4241" \l "collapsepoa12300)**

1. After verifying that all of your features are in the correct data types, revisit the chart section and investigate the specific relationship between the region of purchase and the amount spent.

**TIP:**

If you are not currently in the results view and it is not available at the top of the screen you may need to reconnect your last operator to the results port and run the process again.

1. Create, explore, and interpret scatter plots in RapidMiner. By selecting 'Region' as the X-Axis and 'Amount' as the y-Axis you can explore how a particular attribute (region) is related to the amount spent in each section. Note: You might need to adjust the jitter to see the observations in each region. Here is an example what your chart should look like (you can click on the image to enlarge it):

**TIP:**

**The scatter plot matrix is very useful tool for inferring relationships between attributes.**It is possible to change the size of each individual 2D plot and the point size, and to randomly "jitter" the data (to uncover obscured points). It also possible to change the attribute used to color the plots, to select only a subset of attributes for inclusion in the scatter plot matrix, and to sub-sample the data.

1. **Keep track of your observations about the transaction amount for each region to include in your report.**

#### **[4. Data Mining: Investigate the Relationship Between the Region of Purchase and a Customer's Age](https://codeacademy.ubiqum.com/mc/poa?productID=6637&taskID=4241" \l "collapsepoa12302)**

The next step in most data mining or data science processes usually involves a deep investigation of the relationships between all the independent variables (features with known values) and the dependent variable (the variable whose value you want to predict). This process is referred to as feature engineering; we will be covering it in depth a little bit later in the program, but let’s dive into our first machine learning algorithm, which is in a family named decision trees.

**TIP:**

**A decision tree is a versatile tool that has wide applications**. Decision trees can be applied to almost any domain of questions: they can solve both classification and regression problems, they are computationally efficient (i.e., you can build one quickly, which is important when looking for patterns in large data sets), they tend to perform well on most types of questions, and the results can be easily visualized and intuitively interpreted.

**Pre-work:**

Before diving into the next step, it is highly recommended that you visit the resources and watch/read the following selections regarding decision trees:

* [**Introduction to Decision Trees**](http://www.youtube.com/watch?v=WBebD49pKFA)**-**This video provides a brief introduction to decision trees by using a simple business example.
* [**Decision Trees**](http://www.youtube.com/watch?v=b77DykV-L28)**-**This follow-up short video to *Introduction to Decision Trees* explains the terminology associated with decision trees, e.g., root node, splitting, etc.
* [**Decision Trees**](http://www.youtube.com/watch?v=WOOTNBxbi8c) **–**In this video, Alexander Ihler of UC Irvine, explains the use of decision trees in machine learning. If you want to learn more about how decision trees are constructed for both classification and regression, this 5 minute video is worth watching.
* [**Decision Tree**](http://en.wikipedia.org/wiki/Decision_tree)**-** Wikipedia article which provides a basic overview of decision tree analysis, the components of decision trees, and the advantages and disadvantages. It includes several examples.

**The stakeholder question to be considered in this section can be addressed in two ways:**

* By looking at the histogram and scatter plot data
* By running a decision tree operator on the "region" attribute.  
  You should try both methods.

**TIP:**

As in the previous step, when looking at the histogram and scatter plot data, look for relationships between how a particular attribute is related to all other attributes, then record your observations to include in your report.

**TIP:**

You can use RapidMiner to explore the structure of data, make predictions, or derive specific conclusions. People frequently use machine learning techniques to gain insight into the structure of the data rather than use it to predict new cases. In this task, the decision tree will summarize the data and express it in a concise way, so that you can derive conclusions about the relationships between various attributes.

**TIP:**

You may have read that decision trees only work on nominal data, but that’s not really the case. While they are most often used with nominal data many decision tree algorithms can also be used in problems where the dependent variable is a continuous number (also called a numeric variable). Since this is a classification problem we will select one of the nominal attributes for the dependent variable.

Follow the steps below to run the decision tree operator (The operator is also referred to as an "algorithm."):

1. Continuing with the same process you have been building, you now must select one of the attributes in the data set as a dependent variable. In many machine learning classification problems, dependent variables are often referred to as labels so this next step involves choosing one of the attributes and assigning it to a new role called ‘label’.
2. To do this, you will visit the operator section and look for the ‘set role’ operator. After finding it, connect it to the end of your process and change its parameters to reflect that 'region' is the attribute name and ‘label’ is its target role. This will establish that the region attribute is going to be used as a dependent variable in your decision tree analysis. **Important note:** after you finalize this process and run it you can easily go back and change the parameters of the set role operator and use a different dependent variable for analysis. In this case as long as the dependent variable is a nominal data type the decision tree algorithm will function as expected.
3. Now that you have identified the dependent variable in the process, return to the operator section and search for a decision tree algorithm. You’ll see many different trees in the results, but selecting the first one in the list (‘Decision Tree’) is a good place to start for this analysis.
4. You can see the decision tree operator has ports that look a little different than many of the operators you’ve already used. This is because decision trees can be used in both data mining, which is very close to what you’re doing now, and machine learning which involves training a model and testing it, which is something we will get into very soon, but for now let’s stick to our analysis.
5. Connect the example port from the set role operator to the train (tra) port of the decision tree operator, then connect the model (mod) port from the decision tree operator to the results port, and run the process.
6. If everything worked correctly, you should have a new tab called 'Decision Tree' in your results. By clicking on this tab, you can see what the decision tree looks like and the various features within your analysis.

#### **[5. Machine Learning: Classifying Where a Transaction Took Place](https://codeacademy.ubiqum.com/mc/poa?productID=6637&taskID=4241" \l "collapsepoa12305)**

**Pre-work:**

Before beginning this step, it is highly recommended that you visit the resources and watch/read the following selections regarding decision trees and the machine learning process:

* [**Machine Learning: Model Selection and Cross Validation**](http://www.youtube.com/watch?v=hihuMBCuSlU)**-**This short, well-illustrated video explains how to use the technique of Cross-Validation to estimate machine learning algorithm's performance and to choose between different models.
* [**Decision Tree Learning**](http://en.wikipedia.org/wiki/Decision_tree_learning)**-**Wikipedia article which explains how decision trees are used in machine learning.
* Data Mining for the Masses: Cross-Validation (pages 221-227)

**Step Details**

As you can see from the following machine learning workflow this step will be broken into two parts: Training and Prediction (We'll often refer to Prediction as testing):

In the first part, you will use your already preprocessed data and a decision tree algorithm, along with a new tool called cross validation, to construct a trained model and assess its performance. The term model refers to the output of the machine learning algorithm, and it is the object that you will use during the second step in which you will make predictions.

It is very important for you to understand that the training process should be considered iterative and that part of this process involves revisiting things like preprocessing and feature engineering to get the best results out of your trained model(s). It is not out of the ordinary to build multiple models and even to use multiple algorithms to get the best result. Let’s begin!

1. During the training process you will continue to use the same data set that you have already preprocessed in previous data mining steps. If you have saved that data or the process, go ahead and open it now.
2. Beginning where you left off in the previous data mining step, your last operator should be a decision tree operator. In the following steps you are going to add a new process to the end of your workflow that will allow you to properly train and on test this data for building a machine learning solution that will be used to satisfy the objectives of this step which include classifying where a transaction took place. *Knowing this, what do you think the dependent variable should be?*
3. To add cross validation to your current process, you must insert the cross validation operator just after the set role operator that you have already used. Disconnect any operators after the set role operator, locate the cross validation operator, and drag it into the design view. You can also connect the example port from set role to the example port on the cross validation operator.

After adding the cross validation operator double-click on it to open the training and testing windows, which look like the following image without the operators. (You will add the operators next.)

Now it's time to add and connect the three operators to the cross validation process. Let’s review the connections that are shown in image above (check the RapidMiner documentation for more information on each):

**Training Process**

Decision Tree Operator:

* Input: training from cross validation operator
* Output: training model to testing model

**Testing (Prediction) Process**

Apply Model Operator

* Input: training model to apply model input
* Input: testing values to unlabeled
* Output: label to label of Performance Operator

Performance Operator

* Input: Label from Apply Model
* Output: Performance to Performance Results
* Output: Examples to Testing Results

**Assessing Performance**

Now that cross validation has been established, it is time to proceed with training the model and assessing its performance. Since this is a classification problem we need to specify three different performance metrics to correctly assess the performance of the model and the accuracy of its predictions. The three metrics are (more information about each can be found in the resources):

* Accuracy
* Kappa
* Confusion matrix (this occurs by default)
* Weighted Mean Recall
* Weighted Mean Precision

1. While still in the cross validation view, select the performance operator and choose the accuracy and kappa parameters.

**Let’s train the model!**

1. If you are already in the design view click on the 'Process' link to return to the main design view.
2. Next, you will need to connect the outputs of the Cross Validation operator to the results ports in RapidMiner as shown below and run the process.

**Checking Output**

If your process ran correctly RapidMiner will generate the following:

* A tab containing the Decision Tree
* A tab for Performance Vectors that contains the Confusion Matrix (shown below), Accuracy, kappa, weighted mean recall, weighted mean precision
* A tab containing cross validation example set with the actual and the predicted values (in green columns)

Example of the Confusion Matrix

**Deeper Analysis**

Sometimes different algorithms require different preprocessing methods. You are now tasked with using the same process you have already saved, training and testing three new classification models, and making any modifications needed to your process to make each algorithm perform as needed. This iteration is part of the learning process and experimentation is highly recommended. Above all, have fun, enjoy this exploration, and always consult your mentor if you get stuck!

Here are the three additional classification algorithms to try (you can find more information about each in the resources):

* Random Forest
* Gradient Boosting Trees
* ID3

1. After performing your analysis with all four algorithms, choose the one that produces the best results and include its results in your report.

**TIP:**

Remember, you will need to present your results to someone that has very little  or no experience with data mining or machine learning, so frame your presentation in business, rather than technical, terms.

1. **Optionally explore other algorithms.**If you have time, try other Classification algorithms. As you try each one, read about it in RapidMiner and do additional research on your own to understand how it is different from the others. Keep in mind that some will only work for numeric prediction, while others will only work for nominal prediction. There are many that will work for both numeric and nominal prediction.

#### **[6. Machine Learning: Classification - Understand Items and Amount Spent](https://codeacademy.ubiqum.com/mc/poa?productID=6637&taskID=4241" \l "collapsepoa12316)**

1. Using the same process you set up in step 5, train a decision tree algorithm on the items attribute.

Tip: If a numerical label is not supported for classification, binning may help. Which operator creates bins?

1. Examine the decision tree. What, if anything, is the role of amount?
2. Examine the performance metrics. Does accuracy indicate a good or poor model fit?
3. Now try a second algorithm and compare the results.
4. What business insights can you glean regarding the relationship between the number of items and amount spent?
5. **https://s3.amazonaws.com/gbstool/pub/images/mentor_review.pngWrite up and submit your *Customer Buying Patterns Report*.** Once you have completed your analysis, create a two or three page Word or PowerPoint document in which you summarize your findings. Be sure to include the results from RapidMiner. After you have completed your report, upload the deliverable (a zip file) by clicking on **SUBMIT CUSTOMER BUYING PATTERNS REPORT** in the ***Submit Your Work*** tab.