Dating Preferences Market Basket Analysis: Making the Data Interesting to Students

Stephen Larson
Slippery Rock University of PA, stephen.larson@sr.edu

Follow this and additional works at: http://aisel.aisnet.org/sigdsa2016

Recommended Citation
http://aisel.aisnet.org/sigdsa2016/3

This material is brought to you by the Special Interest Group on Decision Support and Analytics (SIGDSA) at AIS Electronic Library (AISeL). It has been accepted for inclusion in Proceedings of the 2016 Pre-ICIS SIGDSA/IFIP WG8.3 Symposium: Innovations in Data Analytics by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
Dating Preferences Market Basket Analysis: Making the Data Interesting to Students

Teaching Case

Stephen Larson
School of Business
Slippery Rock University of PA
Slippery Rock, PA 16057
Stephen.larson@sru.edu

Abstract

Learning data analysis can be boring for students; the data is often dry and unexciting, pulled from a firm some may have never heard of and dealing with a topic they care nothing about. This paper describes an exercise and lab that teaches rudimentary market basket analysis using the students’ own data: preferred characteristics for dating partners. The exercise and lab take the students through data gathering, data cleansing and normalization, data transformation, and analysis.

Keywords: association analysis, data analysis, teaching

Introduction

Data analysis can be dull for undergraduate students. Most data sets are usually boring and contain no characteristics that would excite and enthuse the students. After trying for years to engage students and keep them interested in yet another set of data from a real or fictional company they have never heard of or worked for, I developed a lab in which students provide the data, clean and normalize the data, transform the data, and run a rudimentary market basket analysis of the data using an MS Access database adapted with permission from Breimer, et al (2012). The result is a lab that is appealing to my students, many of whom are truly interested in the results of the analysis. The lab is in the appendix, and the database is on the USB drive or is available from the author. The database contains sample data that the instructor can use to become familiar with the lab, but when doing the lab in class the students are more engaged when using their own data.

Lab Description

This lab has 4 main parts: the survey, the data cleaning, the data transformation, and the data analysis. The survey is created by the class so the data pertains to a topic of their interest. The survey data my students gathered contains preferred characteristics of dating partners. The lab introduces the students to the data cleaning and data transformation parts of data analysis, and the importance thereof. The data analysis section introduces students to a rudimentary market basket analysis.

Survey

The goal for the first part of the lab is to help the students learn how to ask good questions; ie, questions that elicit the response type desired. The survey contains multiple choice and open-ended questions and was created using the online version of Excel 2016; this could also be done with Google forms. The survey created from either method can be shared with anyone and does not require a login to fill out. As we created our survey in Office 365 online and OneDrive, the following steps are for that method. To create
the survey, log into a OneDrive account, and within Excel Online (only the teacher needs a OneDrive account), create a new blank spreadsheet.

Once the blank spreadsheet was created, click the “Survey” button and select “New Survey” (Figure 1). The survey’s title is “Dating Preference Survey” with a description containing “Please take this survey on characteristics you look for in a dating partner. All answers are anonymous. All questions are optional.” (Figure 2).

As a class, the students then create the survey questions and possible answers, which are then added to the survey (example shown in Figure 2). Some of the questions are multiple choice (such as age or height) and some are short answers (such as major, favorite food, or hobbies). Our most recent survey includes the following questions: preferred age range, preferred height, preferred hair color, preferred eye color, preferred build (body type), preferred hobby or hobbies, preferred major(s), preferred favorite music genre, preferred favorite movie genre, preferred favorite food(s), and whether pets are preferred or not. Your students may come up with different questions, which is one of the goals of this lab.

Some of the questions have multiple choice answers and some are open-ended questions (such as hobby or hobbies, major, favorite food). The answers to these open-ended questions will teach the students the advantages (allows for unlimited and creative answers) and disadvantages of open-ended questions (such
as having the same hobby referred to in different ways). These answers are useful in the data cleaning portion of the exercise and provide material to show the difficulty of data cleaning.

Once the survey is finished, a link is created by selecting to share the survey, and the link is given to the students. The students then fill out the survey themselves and request their friends to do so to increase the number of records. As the survey is taken by participants, Excel Online adds the data to a spreadsheet (Figure 3).

![Figure 3: Spreadsheet with Survey Answers](image)

Data Cleaning

The goal of the data cleaning portion is to give the students experience cleaning and normalizing data. While the answers to the survey questions with multiple choice answers are easily cleaned, the answers to the open-ended questions may prove difficult. In the most recent iteration of administering the survey, we received numerous unique answers to the three open-ended questions. For example, the question about which major you would prefer in a dating partner, one respondent entered “$$$$$$$”, while another entered “anything that makes lots of money.” Other respondents entered answers such as “Business (something that will make him rich)” and “something that isn’t a waste of a degree.” Answers such as these make for an interesting discussion in class, including whether to change the question to multiple choice, which could limit the range of answers.

The data cleaning can take a few class periods or done as a homework assignment, depending on the number of respondents, but we found it important not to rush through this exercise as it can be used to teach how to ask good questions, how to come to a consensus on unusual responses, and what to do about outliers in the data.

Data Transformation

The goal of the data transformation is to show students that while we can analyze the data in Excel, oftentimes we receive data in a format that cannot be used in our current analysis tool, in this case MS Access. The database we will use for the analysis is an MS Access database adapted from Breimer et al
The transforming of the data from an Excel worksheet to the Access table can be accomplished manually, via the use of macros, or by using PowerQuery (which is part of Excel 2016). I let the students decide which method they want to use, and most chose the manual method. A few students chose to record a macro for the repetitive steps. None of the students used PowerQuery as it is a new feature to them and they did not want to spend the time to learn it.

The MS Access database table is set up so that each row in the table consists of an ID number and a characteristic (“item”). As such, each ID number is paired with its respective characteristics, each with its own row. Consequently, the ID number may appear several times. Thus, one row in the worksheet will become several rows or records in the table. For example, this record (row) in the spreadsheet:

<table>
<thead>
<tr>
<th>ID</th>
<th>Age range</th>
<th>Hair Color</th>
<th>Eye color</th>
<th>Build (body type)</th>
<th>Hobby or hobbies</th>
<th>Major</th>
<th>Favorite music genre</th>
<th>Favorite movie genre</th>
<th>Favorite food</th>
<th>Pets?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25+</td>
<td>blonde</td>
<td>Hazel</td>
<td>5’ 3” to 5’ 6”</td>
<td>Petite</td>
<td>travel</td>
<td>Accounting</td>
<td>Rock</td>
<td>Action</td>
<td>chocolate</td>
</tr>
</tbody>
</table>

is transformed as follows (to be imported into an MS Access table):

**ID ** **Characteristic (item)**

1 Age - 25+
1 Hair color - blonde
1 Eye color - Hazel
1 Height - 5’ 3” to 5’ 6”
1 Body type (build) - Petite
1 Hobbies - travel
1 Major - Accounting
1 Fave Music Genre - Rock
1 Fave Movie Genre - Action
1 Fave Food - chocolate
1 Pets - None

The records in the spreadsheet are then imported into a database with a table named BasketData (Figure 4), which has two fields, Basket (a number field) and Item (a short text field).

**Figure 4: BasketData table**

**Data Analysis**

The goal of the data analysis is to learn how to calculate and interpret support values and using a database. Before starting the data analysis, the students are given a pre-lab worksheet to study to learn...
how to calculate support values in a market basket analysis (see Appendix 1). The worksheet is adapted from Breimer et al (2102). Once the students understand the fundamentals of calculating support values, the database is opened and the students work through the market basket analysis lab, performing analysis tasks and answering questions (see Appendix 2).

Through the Dating Preferences Market Basket Analysis lab, the students learn how to do a market basket analysis using database queries, and how to interpret the results of the analysis (the characteristics that are most often preferred in a dating partner from their own population):

1. The single most preferred characteristic in a dating partner and its support value.
2. The two most often preferred characteristics that appear as a pair, and the pair’s support value.
3. The three most often preferred characteristics that appear as a triplet, and the triplet’s support value.

The lab walks the students through the following market basket analysis steps:

1. Becoming familiar with the data by opening the BasketData table and answering a question.
2. Computing single item support by making a copy of the BasketData table with a make-table query named “MakeSingles.” The new table named Singles is used for analyses so the original data remains untouched. The SQL for the MakeSingles query is:
   ```sql
   SELECT BasketData.Basket, BasketData.Item INTO Singles FROM BasketData;
   ```

   Then single item support is calculated using a make-table query named ComputeSinglesSupport. This query computes the support for all characteristics (items) and creates a table called SinglesSupport. The SQL for this query is:
   ```sql
   SELECT Singles.Item, Count(Singles.Item) AS [Count], DLast("[Basket]","BasketData") AS Baskets, [Count]/[Baskets] AS Support INTO SinglesSupport
   FROM Singles
   GROUP BY Singles.Item, DLast("[Basket]","BasketData");
   ```

   The students open the SinglesSupport table (Figure 5) and are asked questions about different characteristics (items) in the table, how many baskets (students) are in the table, how many times was a certain characteristic preferred, support for a characteristic, etc.

   **Note:** all values in the figures are from a sample set of data; your data should be different.

3. The lab then teaches the students about the reasons for pruning the data and how to compute doubles (items pairs that occur in the data, such as blonde hair and blue eyes). Pruning is done by first making the item field a primary key in the SinglesSupport table, then running a delete query named PruneSingles, which deletes records from the Singles table where the support for the item (characteristic) in the SinglesSupport table is <0.01. The SQL for the PruneSingles query is:
DELETE Singles.*, SinglesSupport.Support
FROM Singles INNER JOIN SinglesSupport ON Singles.[Item] = SinglesSupport.[Item]
WHERE (((SinglesSupport.Support)<0.01));

Once the records are pruned, the make-table query named MakeDoubles is run. The SQL for the MakeDoubles query is:

SELECT Singles.Basket, Singles.Item AS ItemA, Copy.Item AS ItemB INTO Doubles
FROM Singles INNER JOIN Singles AS Copy ON Singles.Basket = Copy.Basket
WHERE (((Singles.Item)<[Copy].[Item]));

The resulting table is named Doubles (Figure 6) and lists pairs of characteristics preferred together. Students are then asked questions about pruning and pairs of characteristics, etc. The values for pruning (how many records did we start with, how many did we prune, what percent did we prune) are marked with # signs to show that these numbers will be different for each class and need to be changed to match the current data.

![Doubles Table](image)

4. The lab then continues the data analysis by computing support values for doubles. The make-table query named ComputeDoubleSupport is run to compute the support values for the pairs of characteristics in the Doubles table. The SQL for the ComputeDoubleSupport query is:

    FROM Doubles
    GROUP BY Doubles.ItemA, Doubles.ItemB, DLast("[Basket]","BasketData");

The resulting table named DoublesSupport (Figure 7) contains the pairs of characteristics and each pair’s support value. The students then open the DoublesSupport table and are asked questions about the support values for various characteristic pairs.
5. The lab then teaches the students how to compute support for triples (3 characteristics that are frequently preferred together). First the DoublesSupport table is opened in design view, and the ItemA and ItemB fields are set as a composite primary key, and the table is saved and closed. Then the delete query named PruneDoubles is run to eliminate all pairs with a support value is < 0.01. The SQL for the PruneDoubles query is:

```
DELETE Doubles.*, DoublesSupport.Support
FROM DoublesSupport INNER JOIN Doubles ON (DoublesSupport.ItemA = Doubles.ItemA) AND (DoublesSupport.ItemB = Doubles.ItemB)
WHERE (((DoublesSupport.Support)<0.01));
```

Once the infrequently preferred doubles are pruned, the make table query named MakeTriples is run; the resulting table is named Triples (Figure 8) which contains triplets of characteristics frequently preferred together. The SQL for the MakeTriples query is:

```
FROM Doubles INNER JOIN Doubles AS Copy ON Doubles.Basket = Copy.Basket
WHERE (((Doubles.ItemA)=Copy.ItemA) AND ((Doubles.ItemB)<Copy.ItemB));
```
The make-table query named ComputeTriplesSupport is run to compute the support values for the triplets of characteristics in the Triples table, and lists them in a table named TriplesSupport (Figure 9). The SQL for the ComputeTriplesSupport query is:

```
```

![TriplesSupport Table](image)

**Figure 9: TriplesSupport Table**

The resulting table contains a list of triples (3 characteristics preferred together) and each triple's support value. The students then open the TriplesSupport table and are asked questions about the data in the table.

The Dating Preferences Market Basket analysis portion of this exercise can be worked together in class or can be used as a homework assignment.

**Discussion and Conclusion**

After having used a large data set to introduce market basket analysis to students who had no interest in or relationship to the data, using a data set (albeit small data set) that the students created worked very well. Several students were eagerly searching the TriplesSupport table for different combinations of 3 dating preferences they had entered and checking the support values.

I continue to gather data each subsequent semester and use the new data as a new dataset to compare with older datasets in addition to adding them to an aggregate dataset. Through creating and using this lab in the classroom for several semesters, I have found that the students are more engaged and interested in data analysis. Though the sample size is small, the students appreciate that the data is their own data; some students even search the data for their survey entry. Working with their own data made the analysis more meaningful to the students.
After going through this lab with my students, I then assign them the full lab from Breimer et al (2012), which has a data set with over 900,000 records in an MS Access database. This lab can be obtained from the authors.

Acknowledgements

Many thanks to Dr. Eric Breimer, Dr. Scott Vandenberg and Dr. Robert Yoder for letting me adapt their work to fit my course and teaching style. Dr. Breimer also invites those interested in his original workshop (from which this workshop is derived) to contact him at ebreimer@siena.edu.

References


Appendices

Appendix 1: Market Basket Worksheet
Appendix 2: Dating Preferences Market Basket Analysis
**Appendix 1**

Market Basket Worksheet

Name: ______________________

Goals

Use Market Basket Analysis (also known as affinity analysis) to find items consistently purchased together.

Market Basket Data & Support

Consider the following data:

<table>
<thead>
<tr>
<th>BasketID</th>
<th>Item</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>Beer</td>
<td>BasketID represents the unique “market basket” of a customer who purchased more than one Item. Think of it as the items listed on a sales receipt.</td>
</tr>
<tr>
<td>101</td>
<td>Salsa</td>
<td>Support is the ratio of number of times two or more items occur together to the total number of transactions. Support of a product or product bundle indicates the popularity of the product or product bundle in the transaction set. The higher the support, more popular is the product or product bundle. This measure can help in identifying driver of traffic to the store.</td>
</tr>
<tr>
<td>101</td>
<td>Chips</td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>Water</td>
<td></td>
</tr>
<tr>
<td>201</td>
<td>Chips</td>
<td></td>
</tr>
<tr>
<td>201</td>
<td>Beer</td>
<td></td>
</tr>
<tr>
<td>301</td>
<td>Soda</td>
<td></td>
</tr>
<tr>
<td>301</td>
<td>Water</td>
<td></td>
</tr>
<tr>
<td>301</td>
<td>Beer</td>
<td></td>
</tr>
<tr>
<td>301</td>
<td>Chips</td>
<td></td>
</tr>
<tr>
<td>401</td>
<td>Beer</td>
<td></td>
</tr>
<tr>
<td>401</td>
<td>Soda</td>
<td></td>
</tr>
<tr>
<td>501</td>
<td>Chips</td>
<td></td>
</tr>
<tr>
<td>501</td>
<td>Soda</td>
<td></td>
</tr>
<tr>
<td>501</td>
<td>Water</td>
<td></td>
</tr>
</tbody>
</table>

Questions:

1. How many items did customer 301 purchase? ______
2. How many customers are shown in the table? ______
3. How many baskets contain chips? ______
4. How many baskets contain soda and beer? ______
5. What is Support(Soda, Beer)? ______
6. What is Support(Chips, Salsa)? ______
7. What is Support(Water, Chips)? ______
Appendix 2

Dating Preferences Market Basket Analysis

Note: In the sections “How much pruning did we do” and “Understanding big data” there are values marked as [count]. These should be replaced with the numbers that fit your data set.

Instructions

Goals

Learn to use Market Basket Analysis to find characteristics of dating partners that are consistently grouped or preferred together.

I. Understanding the Data

a. Open DatingPreferences.accdb
b. Open the BasketData table
c. The Basket field represents the unique “market basket” of students who filled out the dating preferences survey; each number is an anonymous ID.
d. The Item field is a characteristic that a person seeks in a dating partner. Examine the table to better understand the data.

Question 1: Why might you see the same item (characteristic) many times?

a. Because an item (characteristic) can be preferred by many students
b. Because a student might choose several characteristics
c. Both a and b are correct

II. Computing Single Item Support

The BasketData table is raw data with over 2600 rows. No human could intelligently process this data quickly. It is not easy to see which characteristics are preferred together frequently. To figure this out, we first have to try to figure out which single characteristics (items) are preferred frequently.

a. Close the BasketData table.
b. Run the query MakeSingles.
c. Click yes twice.
   i. This creates a copy of BasketData called Singles so we do not lose the original data.
d. Run the query ComputeSinglesSupport.
e. Click yes twice.
   i. This computes the support for all characteristics (items) and creates a table called SinglesSupport.
f. Open the SinglesSupport table to answer these questions.

Question 2: How many characteristics (items) are in the table?

Question 3: How many baskets or students are in the database?
Question 4: How many times did a student prefer someone whose height is 5' 3" to 5' 6"?

Question 5: Sort the characteristics (items) alphabetically. What is the Support for Pets - Cats?

Question 6: What is another way to describe support?
   a. The probability that an item is in a basket
   b. The number of baskets that contain an item divided by the total number of baskets
   c. The percentage of baskets that contain the item
   d. All of the above are correct descriptions

III. Pruning and Computing Doubles

Item pairs are called Doubles. For example, many people buy Peanut Butter and Jelly as a pair.

Question 7: If you have 1000 different characteristics (items) in your store how many different characteristics (items) pairs exist? Hint: You should not pair the same item with itself, i.e. Jelly and Jelly is not a valid pairing.
   a. 1000 * 1000 = 1 million
   b. 1000 * 999 = 999,000
   c. 1000 * 2 = 2000
   d. 1000 + 999 = 1999

Too much data:
The problem with our data is that we have 267 different characteristics (items) in our database. If you correctly answered the above question, then you know that the number of pairs is 267 * 266 = 71,022 pairs. Marketers are very interested in characteristics (items) that are preferred together as a pair, but this is too much data to analyze.

Pruning:
Marketers do not care about infrequently preferred characteristics (items), nor do we care about infrequently preferred dating characteristics. Before we compute all the pairs (i.e. doubles), we can remove or prune characteristics that are preferred infrequently (ie, less than one percent of the time)

1. Open the SinglesSupport table in the Design View
2. Right click on Item and make it a Primary Key
3. Hit Ctrl + S to save the table
4. Close the table
5. Run the PruneSingles query
   a. Then click yes
6. Run the MakeDoubles query to find pairs of characteristics preferred together
   a. Then click yes

1. Open Singles in the Datasheet View.

Question 8: How many rows are in the Singles table?

2. Open BasketData in the Datasheet View.
Question 9: How many rows are in the BasketData table?

How much pruning did we do?
BasketData contained the data to compute the Doubles, so we made a copy of the data (Singles) and we pruned out characteristics (items) that are infrequently preferred. This reduced the number of rows from [count] to [count] which is about a [count]% reduction in the amount of data.

Question 10: Why is it OK to prune out [count]% of the data?
   a. The pruned data is redundant
   b. The pruned data is not important because the items are preferred infrequently
   c. The pruned data is not needed to compute frequently preferred doubles
   d. Both answers b and c are correct

Question 11: Why are we doing all this work anyway?
   a. We want information about which characteristics are frequently preferred together
   b. Because our teacher is making us do the work.
   c. We are removing data redundancy
   d. We want information about which characteristics are not preferred

IV. Computing Support for Doubles
   a. Close all open tables and queries
   b. Run ComputeDoubleSupport query and click yes twice.
   c. Open DoublesSupport in the datasheet view
   d. Sort the table by Support (Largest to Smallest)

Question 12: What two characteristics (items) are preferred together most frequently?

Question 13: What is the Support for the most frequently preferred item pair (i.e. double)?

Question 14: How is the Support value computed?
   a. Number of baskets that contain the item pair multiplied by the total number of baskets.
   b. Number of baskets that contain the item pair divided by the total number of baskets.

Question 15: What does a 0.1 Support value mean for a double?
   a. 10% of customer baskets contain the double (item A and Item B)
   b. The probability that a random customer will buy both Item A and Item B is 1 out of 10
   c. 100% of the customers who bought Item A also bought Item B
   d. Both answers a and b are correct.

Question 16: According to the DoubleSupport table, which two characteristics (i.e. double) are preferred together least frequently? Note: It may be a tie between 2 or more doubles.
**Question 17:** Consider your answer to the previous question. How often was the least supported double actually preferred?

**Question 18:** What table is the best example of raw data?
- a. BasketData
- b. DoublesSupport
- c. SinglesSupport

**Question 19:** What table is the best example of useful information at this time?
- a. Doubles
- b. DoublesSupport
- c. SingleSupport

**Pruning and Computing Triples**
A customer might recognize that a sale on Jelly is a marketing ploy to sell Bread, which has been marked up. But it is harder to recognize associations between three or more characteristics (items). These complex or hidden associations are very valuable because marketers can leverage complex product relationship to upsell and cross-sell products. To find complex associations, we first compute the frequent triples i.e. three characteristics (items) that are often preferred together.

**Understanding Big Data**
The problem with computing triples is that we have \([\text{count}]\) characteristics (items). We can pair each of these characteristics (items) with the \([\text{count}]\) other characteristics (items) to create \([\text{count}]\) pairs. And then we can group each pair with each of the remaining characteristics (items), so the number of triples is calculated as follows:

\[
[\text{count}] \times [\text{count}] \times [\text{count}] = [\text{count}]
\]

In general, if we have \(N\) characteristics (items), then the number of triples equals \(N \times (N-1) \times (N-2)\)

**Question 20:** How many triples exist at a store that sells 21,000 different characteristics (items)?
- a. About 9 billion
- b. About 4 billion
- c. About 9 trillion
- d. About 4 trillion

Most databases cannot efficiently process tables with trillions of rows. Luckily, we pruned infrequent singles before we computed doubles. And, we can also prune infrequent doubles before we compute triples.

**Why Pruning Works**
If two characteristics (items) are rarely preferred as a pair they will rarely be preferred as part of a trio.

**V. Computing Triple Support**

1. Open **DoubleSupport** in the design view
2. Hold the shift key and select **Item A** and **Item B**
3. Click the primary key button
4. Save and then close all tables and queries
5. Run the **PruneDoubles** query and click yes twice
   - i. Run the **MakeTriples** query and click yes twice
6. Run the **ComputeTriplesSupport** query and click yes twice
7. Open the **TriplesSupport** table in datasheet view
8. Sort the table by **Support** (largest to smallest)
**Question 21:** How many triples have a support higher than 0.1?

**Question 22:** Is it possible that we pruned some triples with support greater than 0.01?
   a. Yes, because we pruned out a lot of doubles.
   b. No, because the doubles that we pruned (support less than 0.01) could never be part of a triple with support greater than 0.01.
   c. It is impossible to tell because there is too much data.

**Question 23:** What are the three most preferred characteristics that appear together?

**Question 24:** What is the support for the top 3 preferred characteristics?

**Question 25:** What does this number tell us?
   a. This is the percent of people who took the survey that likes these three characteristics combined together in a dating partner
   b. It’s better to be a blonde than a brunette if you want dates
   c. Many people prefer to date a brunette that likes comedy movies about dogs.

**Question 26:** What is the support for three preferred characteristics that you entered into the survey (if listed)?