Assignment 1

1.) Download the Acquire Valued Shoppers Challenge data from Kaggle or from my dropbox.

The data is available at these locations:

Kaggle: https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data

(You can read more about this data on the Kaggle competition data page linked above.)

2.) Decompress the contest data and read it into Python, R, or SAS. Your choice.

The gzipped files can be unzipped into normal CSV files using a standard application, such as 7Zip (http://www.7-zip.org/) or the unix command line application gunzip. Storing the *.csv files may require more than 22 GB of disk space.

You should now have 4 data sets:

offers – 6 columns, 37 rows
testHistory – 5 columns, 151484 rows
trainHistory – 7 columns, 160057 rows
transactions – 11 columns, 349655789 rows

3.) The target variable for the contest was repeater. Remove repeattrips from the trainHistory set.

In the Acquire Valued Shopper Challenge data the trainHistory and testHistory data sets contain a large number of customers. Each customer is uniquely identified by the id variable. In the contest, the labeled training data was used to build a model to predict whether each customer in the test data would be a repeat shopper. The two extra columns in the training data are target variables. There are two possible target variables in the training set:

repeater - a binary variable which signifies whether a shopper is a repeat shopper.
repeattrips - an ordinal variable name which indicates the number of times a shopper repeats.

After dropping the repeattrips variable, the trainHistory set should now have 6 columns and 160057 rows.

4.) Left join the offers set onto the trainHistory and testHistory sets.

Predicting repeat shoppers is difficult. You will need as much data as possible. The type of offers a customer has received may influence their repeating behavior. As a rule in predictive modeling, any transformation that is applied to the training data must also be applied to the test data. Join the information in the offers table onto the training data and onto the test data using a common variable in all three sets.

The new training set should contain 11 columns and 160057 rows. The new test set should contain 10 columns and 151484 rows.

5.) Determine the number of items and the dollar amount a shopper has spent on items in the same category as the item for which they received an offer in the trainHistory or testHistory sets.

Another important type of information is customers past behavior. It is logical to conclude that a customer who has bought an item many times in the past may continue to buy that item, or a similar item, in the future. To include this information in the training and test data, you must summarize the transactions set. Group the transaction set by id and category and summarize the groups by summing the purchasequantity and purchaseamount variables. Then left join the summarized set onto the trainHistory and testHistory sets by id and category. (This operation took about 5 minutes on my laptop.) After the join, the new training and test sets may contain missing values. It is preferable to replace these missing values with 0, as missingness in this case simply indicates the customer has purchased 0 items, costing 0 dollars.

After the join operation, the new training set should contain 13 columns and 160057 rows. The new test set should now contain 12 columns and 151484 rows.

6.) Create small subsets of the training and test data, export them to CSV, and turn them in.

Subset the training data so that it contains only id numbers less than 14000000. Export this small file to CSV format.

Subset the test data so that it contains only id values greater than 4810000000. Export this small file to CSV format.

Zip your code and your CSVs into a single archive file and submit to blackboard.