PBA Assignment #1

AS1: The Nuts and Bolts of Data Analysis Using R

This assignment uses data provided by the UC Irvine Machine Learning Repository, an open-source repository which boasts myriads of datasets that can be downloaded and used free of charge. Students will be using the "Online Retail Data Set" which can be downloaded from this website[1]:

- **Dataset**: Online retail transactions in CSV format [44.5Mb]
- **Description**: This is a transnational data set which contains all the transactions occurring between January 12, 2010, and September 12, 2011, for a UK-based and registered non-store online retailer. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers (B2B).

The following descriptions of the 8 variables in the dataset are taken from the UCI website:

1. **InvoiceNo**: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. **If this code starts with letter 'c', it indicates a cancellation (refund).**
2. **StockCode**: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
3. **Description**: Product (item) name. Nominal.
4. **Quantity**: The quantities of each product (item) per transaction. Numeric.
5. **InvoiceDate**: Invoice date and time. Numeric, the day and time when each transaction was generated.
6. **UnitPrice**: Unit price. Numeric, Product price per unit in sterling.
7. **CustomerID**: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
8. **Country**: Country name. Nominal, the name of the country where each customer resides.

Read this before you begin your assignment!

** The following instructions contain a set of questions (highlighted in purple) you will need to solve using the R statistical programming language. After completing the assignment, please submit the answers along with the script file used for the assignment.* ***You are highly encouraged to use R Markdown for the assignments.***

** InvoiceDate, as it is stored in the original data has the following format: mm/dd/yy.*

[1] Note that data is usually not 'given' to you! You'd have to go out of your ways to collect and clean data on your own most of the time.

INSTRUCTIONS

a) First, find out the path to the directory containing the data and load it onto R. The dataset has 541,909 rows and 8 columns. After importing the data, make sure you have imported it properly by using the `head()` and the `str()` functions.

b) We will be using data from 2011/07 to 2011/08. Convert **InvoiceDate** to date class and subset the data. You should see 3,664 unique transactions (i.e., **InvoiceNo**) left in the data. One alternative is to read the data from just those dates rather than reading in the entire dataset and subsetting to those dates.
   - You can use `length()` and `unique()` functions on **InvoiceNo** to find out the number of unique transactions.
   - You may find it useful to convert the Date and Time variables to Date/Time classes in R using the `strptime()` and `as.Date()` functions. `lubridate` package can also be used which will be helpful for more complex date operations.

c) Use for-loops to 1) compute the mean of **Quantity** and **UnitPrice**, 2) determine the types of each column, and 3) compute the number of unique values in each column.

d) Subset the data for which the transactions took place in the U.K., Netherlands, and Australia. Using the subset of data, 4) report the average and standard deviation (round them up to 3 decimal points) of the **UnitPrice** as well as 5) the number of unique transactions made in these countries. 6) How many customers residing in these countries made transactions in July and August of 2011?

e) Do we see any customers who made a refund? 7) If we do, how many customers made a refund (make sure to exclude the observations without the **CustomerID**)? Assign the IDs of the customers who made at least one refund during the period into a vector called **cust_refund**.

f) Some customers made purchases without logging into the e-commerce site. This would create records of transactions for which the **CustomerID** is missing (i.e., NA). These transactions cannot be traced since we do not know who ordered the products. Create a variable called **Sales** by multiplying the **Quantity** and the **UnitPrice**. 8) Then, calculate the total sales amount for those that are missing the **CustomerID**. 9) How many transactions were made without the customers logging into the e-commerce site?

Extra Credit:

g) EC1) Create a variable containing the monthly aggregate spending for each customer. EC2) Then, report the IDs and the monthly purchase amount of the five customers who have spent the most money in July 2011.