Assignment – Online Retail Analytics

Purpose

This assignment integrates your knowledge from the practice activities, readings, and narrated presentations in Modules 1-4 to show your mastery of the module outcomes.

Learning Outcomes

- Apply different transformation to the data to make it suitable for modelling. (CLOs 1, 6)
- Use R to implement data wrangling and transformations. (CLO 5)
- Analyze the role of descriptive statistics in data exploration phase of analytics projects. (CLOs 1, 6)

Instructions

Please answer all questions. You should use R to solve the questions and include the screen shots or pdf of your R-notebook in your submission. The Golden questions are optional and carries additional marks. This means that you will not lose marks if you do not answer that question.

For this assignment, you need to use the 'Online Retail' dataset which can be downloaded in CSV format from the Dataset folder. This is a transnational data set which contains all the transactions occurring between 01 Dec 2010 and 09 Dec 2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers. The data contains the following attributes:

- 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.

- StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.

- Description: Product (item) name. Nominal.

- Quantity: The quantities of each product (item) per transaction. Numeric.

- InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.

- UnitPrice: Unit price. Numeric, Product price per unit in sterling.

- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

- Country: Country name. Nominal, the name of the country where each customer resides.

Download the dataset, and use the read.csv() command to load the file into a R dataframe and answer the following questions. Note that Questions 4 and 9 are optional and carry only extra points.

1. Show the breakdown of the number of transactions by countries i.e., how many transactions are in the dataset for each country (consider all records including cancelled transactions). Show this in total number and also in percentage. Show only countries accounting for more than 1% of the total transactions. (10% of total points)

2. Create a new variable 'TransactionValue' that is the product of the exising 'Quantity' and 'UnitPrice' variables. Add this variable to the dataframe. (10% of total points)

3. Using the newly created variable, TransactionValue, show the breakdown of transaction values by countries i.e. how much money in total has been spent each country. Show this in total sum of transaction values. Show only countries with total transaction exceeding 130,000 British Pound. (15% of total points)

4. This is an optional question which carries additional marks (golden questions). In this question, we are dealing with the InvoiceDate variable. The variable is read as a categorical when you read data from the file. Now we need to explicitly instruct R to interpret this as a Date variable. "POSIXlt" and "POSIXct" are two powerful object classes in R to deal with date and time. Click here for more information. First let's convert 'InvoiceDate' into a POSIXlt object:

   ```
   Temp=strptime(Online_Retail$InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')
   ```

   Check the variable using, head(Temp). Now, let's separate date, day of the week and hour components dataframe with names as New_Invoice_Date, Invoice_Day_Week and New_Invoice_Hour:

   ```
   Online_Retail$New_Invoice_Date <- as.Date(Temp)
   ```

   The Date objects have a lot of flexible functions. For example knowing two date values, the object allows you to know the difference between the two dates in terms of the number days. Try this:

   ```
   Online_Retail$New_Invoice_Date[20000]- Online_Retail$New_Invoice_Date[10]
   ```

   Also we can convert dates to days of the week. Let's define a new variable for that

   ```
   Online_Retail$Invoice_Day_Week= weekdays(Online_Retail$New_Invoice_Date)
   ```

   For the Hour, let's just take the hour (ignore the minute) and convert into a normal numerical value:

   ```
   Online_Retail$New_Invoice_Hour = as.numeric(format(Temp, "%H"))
   ```

   Finally, lets define the month as a separate numeric variable too:

```
Online_Retail$New_Invoice_Month = as.numeric(format(Temp, "%m"))
```

Now answer the flowing questions.

a) Show the percentage of transactions (by numbers) by days of the week (extra 1% of total points)

b) Show the percentage of transactions (by transaction volume) by days of the week (extra 1% of total points)

c) Show the percentage of transactions (by transaction volume) by month of the year (extra 2% of total points)

d) What was the date with the highest number of transactions from Australia? (extra 2% of total points)

e) The company needs to shut down the website for two consecutive hours for maintenance. What would be the hour of the day to start this so that the distribution is at minimum for the customers? The responsible IT team is available from 7:00 to 20:00 every day. (extra 4% of total points)

5. Plot the histogram of transaction values from Germany. Use the hist() function to plot. (5% of total points)

6. Which customer had the highest number of transactions? Which customer is most valuable (i.e. highest total sum of transactions)? (15% of total points)

7. Calculate the percentage of missing values for each variable in the dataset (5% of total points). Hint colMeans():

8. What are the number of transactions with missing CustomerID records by countries? (10 % of total points)

9. On average, how often the costumers comeback to the website for their next shopping? (i.e. what is the average number of days between consecutive shopping) (5% of total points!) Hint: 1. A close approximation is also acceptable and you may find diff() function useful.

10. In the retail sector, it is very important to understand the return rate of the goods purchased by customers. In this example, we can define this quantity, simply, as the ratio of the number of transactions cancelled (regardless of the transaction value) over the total number of transactions.

With this definition, what is the return rate for the French customers? (10% of total points). Consider the cancelled transactions as those where the ‘Quantity’ variable has a negative value.

11. What is the product that has generated the highest revenue for the retailer? (i.e. item with the highest total sum of ‘TransactionValue’). (10% of total points)

12. How many unique customers are represented in the dataset? You can use unique() and length() functions. (10% of total points)

General Submission Instructions:

All work must be your own. Copying other people’s work or from the Internet is a form of plagiarism and will be prosecuted as such.

You may submit a Microsoft Word (.doc/.docx) document as an attachment using the Canvas Assignment tool, or you may copy and paste your answer into the provided box within the Assignment tool. If you attach a document for your assignment, be sure to include your name in the text of the document and in the name of the document.

- You can only submit once, so make sure you are completely finished before submitting and that you attach the correct word .doc/.docx file.
- Submissions sent by email will NOT be accepted.

Due dates are listed in the Assignment Schedule document.