JIGSAW ACADEMY
THE ONLINE SCHOOL OF ANALYTICS

Graded Assignment on Data Preparation and Exploration

Output Expected: Path to R code and output stored in lab, submitted in the LMS.
Slides stored in the lab and path specified in the LMS
Use the following data sets:

Campaign_File.txt
Customers_File.txt
Products_File.txt
Transactions_File.txt

The data sets are located at: **Data Science R/Assignments/Graded Assignments/Topic 8.2 Data Preparation**

Given below is the variable description for each data set

Campaign_File.txt

Variable	Description
Card_ID	Card id of the customer (unique)
Campaign_Reponce	Did the customer respond, T/F logical column

Customers_File.txt

Variable	Description
Card_ID	Card id of the customer (unique)
Registration Date	Date on which customer was registered with the store
Gender	Gender of customer
Birth Date	Date of birth

Products_File.txt (Master file of all products available)

Variable	Description
Product Code	Numeric code for each product
Product Category	Character code for category
Unit Price	Price in $

JIGSAW ACADEMY
THE ONLINE SCHOOL OF ANALYTICS

Transactions_File.txt

Variable	Description
Transaction_ID	Self Explanatory
Card_ID	Card id of the customer
Payment Method	Mode of payment
Timestamp	Time of transaction
Product Code	Numeric code for each product
Items_Number	Number of units purchased
Items_Amount	$ amount of purchase

Answer the following questions:

While importing files, make sure that column labels are read in properly also make sure you can using the right separator.

Q1**.** Based on the transactions, which product category dominates in terms of $ amount?

(Hint: You will need to merge Transactions and Product data sets and then look at Product category)

Q2. Perform a suitable age grouping and find out contribution of each of the age group in terms of $ amount spent.

(Hint: A merge between Customer and Transaction table will be required)

Q3. Find the response rate to the campaign. Also identify the age group of customers where response rate is high. Is there a consistent trend.

(Hint: Add age information from customer file to the campaign file, compute response rate by age. This can be done by binning age either in deciles or quartiles and seeing if there is a consistent trend)

Q4. Repeat the analysis above with “Tenure” of customer. (Tenure will be defined as the time period between the Date of Registration and 31/12/2002)

Q5. Create a cross tab of response rate between Age and Tenure of customers. Do you observe anything?
Q6. Which mode of payment is most popular? Is mode of payment affected by the time of transaction?
(Hint: Extract hour information from timestamp column by using appropriate date conversion function, based on the hour of the day extracted, you can do an appropriate classification and then look at the cross tab between payment mode and time of day)
Q7. Do you think, based on the data, that age and gender has any impact on $ amount spent?
(Hint: You'll need to merge customer and transaction tables appropriately and then do an age classification, post that you can create a cross tab between gender and age to arrive at an opinion)
Q8. Produce a histogram for "tenure of a customer" separately for male and female customers.