Data Science Challenge

A well-known online e-commerce marketplace is operating in the Middle East market, selling a multitude of products belonging to several different brands. Snapshots of the products that are being sold and of the registered users are provided as two data files in csv format. Information about the events when the user has viewed an item and, eventually, has decided to add it to the shopping basket is also provided. The list of orders/transactions done on the marketplace platform is also available as a csv file. All four files can be downloaded from the following link:

http://tamanna.s3.amazonaws.com/data.zip

Dataset name	description
user_table.csv	Information about the registered online store users.
product_data.csv	Information about the products that are being sold online.
events_data.csv	Information about the products on which the user has clicked to get a detailed information. The column 'action' provides information if the product was added to the user's shopping basket or not.
transactions.csv	Information about the product purchase events.

The goal of this challenge is to accomplish the following:

1. Determine the quality of the data (missing value, mixed value type in a column and so on)
2. Are there any relationships between the features?
3. Figure out 3 possible uses cases that can be used to derive value to the customers or to the company (classification, predictions, clustering, statistics, etc.)
4. What is the probability that a product type from a given brand is sold depending on the age of the customer? Does this value change in time?
5. Given that the manager wants you to develop a model that predict sales amount for whole marketplace on daily-bases.
    a. What type of problem you should solve?
    b. What are the evaluation function (loss) and the metric you are going to use and why?
    c. What steps are you going to follow to solve this problem?
6. Use the given dataset to develop a prediction model to predict the weekly sales for a certain brand. Think about those points:
    a. how are you going to split the data?
    b. how are you going to evaluate the performance of the model giving the current dataset?
    c. how are you going to evaluate the performance of the model for unseen data in future?
7. If you have an access to external data, which data could be useful to improve the performance of the model?
8. Based on the historical data and your model, you noticed that one brand has a poor sales performance. Which feature do you suggest to change in order to increase the sales in the future?
9. You are asked to present your results in front of your team and send a report to your manager. Prepare a self-explained visualization and report.
10. Optional question:
    If you are asked to determine the best price for a certain product for each season of the year:
    a. what type of problem are you going to solve?
    b. explain shortly (no calculations required) how are you going to determine the price for the summer season.
    c. The back-end team would like to use your latest model in production through a Restful API where the input features are sent via a POST request. In what format do you plan to deliver you model such that it can be seamlessly integrated by the back-end team?

The expected time to complete the task is 4 hours.