Final Project Description

Background and Introduction

The final project for the course will require you to complete some exploratory analyses for a womens clothing firm. The data for this project come from the Kaggle website at the following link:

https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews/data

I have attached the dataset to the Project entry on myCourses so that you do not have to register for an account. The dataset contains information on customer product ratings for an anonymous womens clothing e-commerce firm. For this project, you only need to focus on the following fields:

Variable	Description
Review_ID	Row ID for the review
Clothing_ID	Product ID for particular article of clothing
Age	Age of the reviewer
Rating	Reviewer rating of article on 0-5 scale
Recommended	Whether the reviewer positively recommends the product (1) or not (0)
Department_Name	The department which is responsible for that article of clothing

Objectives and evaluation

The project requires you to complete three tasks, detailed below. You should prepare a report for the e-commerce firm answering their questions for each task. They would also like you to include the code for each task in your report, for reproducibility purposes. You may include the code as code chunks where the analyses are taking place or, if you prefer, you may include it at the end (although the code should be clearly commented so that it is clear which task each block of code corresponds to).

The completion of each task is worth 25 points. The quality of presentation will also be worth 25 points, i.e. clarity of explanation, plots, tables, and code.

The length of the projects will vary, depending on the number and formatting of figures and tables and the conciseness of the writing. Rather than focusing on the number of pages, I encourage students to focus on completing each task (and subtask) below to the best of their ability in the clearest and most efficient manner.

Tasks to complete

Task 1: Exploratory single variable analyses

The first task is to provide some exploratory data analyses and describe the distributions of age, product rating, recommendations, and article departments amongst the respondents. For each of the four variables, first produce appropriate plots and summary tables and then, in words, describe the distribution of each variable individually. Note if there are any missing values and then remove them from the data set for the remainder of the tasks.

Task 2: Exploring associations

The firm is also interested in answering two questions about associations between some of the variables. Please address each question by using both graphical and numerical summaries and describe the nature of the associations described by the questions below.

Question 1: The firm would like to know whether the distribution of age of reviewers varies across product **departments**.

Question 2: For marketing purposes, they would also like to divide respondent age into five demographic categories: 25 and under, 26 - 35, 36-45, 46-64, and 65 and over and compare the distribution of product **ratings** amongst each of the five age groups to see which groups are most enthusiastic about their company's products.

Task 3:

For the final task, the company would like to compile a list of their ten most popular products based on recommendations (with each product indicated by ID number). However, they read an article which indicated that comparing products based just on the average review or the proportion positively recommended is dangerous, as some products have many fewer reviews than others, e.g. one could have 100% of reviews with positive recommendations, but only 3 total reviews.

The company feels that a measure of popularity should balance both the number of reviews with the proportion recommended. One such measure for binary values is known as the Wilson's lower confidence limit approximation for proportions.

Let $\hat{p}_i$ be the proportion of respondents who positively recommended a certain product and n_i be the number of respondents who rated that product (positively or negatively). Then Wilson's lower confidence limit can be computed via:

$$
\begin{aligned}
a_i &= \frac{1.96^2}{2n_i} \\
b_i &= \frac{\hat{p}_i(1-\hat{p}_i)}{n_i} \\
c_i &= \frac{a_i}{2n_i} \\
WLCL(\hat{p}_i) &= \frac{\hat{p}_i + a_i - 1.96 \times \sqrt{b_i + c_i}}{1 + 2a_i}
\end{aligned}
$$

This lower confidence limit is the value for which we are 97.5% confident that the true value of the proportion in the population who positively recommend product i lies above this quantity. For two products with the same proportion of positive recommendations, $\hat{p}_i$, the one with the larger number of reviews will have the higher lower confidence limit. It is possible for a product with a larger number of reviews to have a smaller proportion of positive recommendations than a second product, but a larger lower confidence limit if the second product has many fewer total reviews.

For this task, compile three different lists in the form of tables:

a) the 10 product ID's with the highest average ratings;
b) the 10 product ID's with the highest proportion of positive recommendations; and
c) the 10 product ID's with the highest Wilson lower confidence limits for positive recommendations as described above

In each table, include the product ID, the number of reviews for that product, the average rating, the proportion of positive recommendations, and the department.

Which list do you think best represents the products which are the most popular? Explain your answer clearly to the firm in your report.