MSIN0094 Second Assignment

Due 10 am, 26 Nov

1 Background

The senior marketing manager of Amazon UK, Tom, had been thinking about purchasing a new television for the family so that he could finish watching Squid Game series to chill after work. He really needed some more rest as he had gone through lots of pressure recently: First, he had to leave PineApple Inc due to his miscalculation of the payback period for the influencer marketing program; second, due to his frequent absence from the marketing analytics module, he struggled a lot while computing the customer lifetime value for his bubble tea business; as a consequence, he miscalculated the CLV to CAC ratio, ended up acquiring non-profitable customers, and failed the business. Fortunately, with the prestigious MSc Business Analytics degree he obtained from UCL, he managed to find a senior marketing manager position at Amazon UK, which definitely had nothing to do with the fact that he is the brother-in-law of Jeff Bezos.

"Hey Dad," Tom's train of thoughts was interrupted by his son, Tomson, "The latest Nintendo Switch console has just been back in stock on Amazon. I need a new Sony OLED TV to play 4K games. Also, I want a large screen size; the larger, the better!"

"Son, our flat is only 400 ft^2, so we have no space for a large-size TV; plus, Samsung makes the best TVs; let's buy from Samsung!"

Tom and Tomson argued for the whole evening without reaching any mutual agreement.

In the end, Tom came up with a good idea: "Son, your dream is to join UCL's MSc BA program and become a successful business analyst like me, right? Let me get the past 12 months' TV sales data from Amazon, and let the data speak who is right or wrong!" "Deal Dad!"

On the second day, Tom asked the database team for the TV sales data for the four major brands, including Samsung, LG, Sony, and Philips, for the past 12 months. The database team came back with three datasets (*italic* words are variable names).

The first dataset is named `data_product`, which includes the specification information on each TV model. That is, for each *product_id*, the *data_product.csv* includes the *brand*, *technology* (LCD, QLED, or OLED), *resolution* (4k, 1080p, or 720p), *energy_class* (A+, A, B, or C), *support_HDR* (1 means that the TV model has HDR support and 0 means no HDR support), *refresh_rates* (up to 60 Hz or above 60 Hz), and *screensize* (up to 29 inches, 30-39 inches, 40-49 inches, 50-59 inches, or 60 inches and above). The `data_product`

dataset also includes the information on each model's *Recommended Retail Price (RRP)*, which is the price at which the manufacturer recommends that the retailer sell the product; the intention of RRP was to help standardize prices among retailers.

The second dataset is named `data_sales`, which includes the *product_id*, *week_id*, the *weekly sales* (how many units sold), and *price_discount* (the percentage off the retail price). The *actual transaction prices* during each week would be RRP * (1 – price_discount).

The last dataset is named `data_marketing`, which includes the information on each brand's *weekly marketing expenses* on Amazon for search advertising and featuring their products. Note that, the marketing expense is at the brand level rather than product level.

With the three datasets at hand, Tom was confident that he would be able to use the power of data manipulation tools to convince his son that Samsung makes the best TV and that a larger screen is not always preferred by customers.

2 Dataset Manipluation (30pts)

Note that, the datasets are on different granularity levels: `data_product` is at the product-level because the product attributes are time-invariant; `data_sales` is at the product-week level; `data_marketing` is at the brand-week level. Therefore, the first step should be to correctly join the datasets together based on the common IDs.

1. Use the `dplyr` package's join function to merge `data_product` with `data_sales`, and assign the merged dataset into a new dataset named `data_full`. (**10pts** in total for a, b, and c questions)

    a. Complete the code of joining the two datasets; throughout the assignment, please explicitly use the 'by =' argument to specify the common IDs for your joining operations (**2pts** for using correct common IDs for joining; **2pts** for correct remaining codes)

    b. Generate a variable *final_price* in `data_full` as below (**2pts** for correct code of generating the variable): *final_price* = *RRP* * (1 – *discount*)

    c. Use R's built-in functions to compute and report the mean and standard deviation for *final_price* (correct mean **2pts**; correct standard deviation **2pts**)

2. Merge `data_marketing` with `data_full` based on appropriate IDs. (**8pts** in total for a, b questions)

    a. Complete the code of joining the two datasets (**2pts** for using correct common IDs for joining; **2pts** for correct remaining codes)

    b. Use R's built-in functions to compute and report the summary statistics for *marketing_expense* as in `data_full`. (correct code **2pts**; correct results **2pts**)

3. Please use R codes to answer the following questions (**12pts** in total for a, b, and c questions)

    a. What is the correlation coefficient between *marketing expense* and *RRP* in `data_full`? (correct code **2pts**; correct answer **2pts**)

    b. Discuss whether the correlation is statistically significant? (correct R code to conduct the test, **2pts**; correct discussion **2pts**)

    c. Normally, RRP has been set before the manufacturer releases the inventory to retailers. That is, in terms of the time of occurrence, RRP should go before marketing expense. From the correlation test above, discuss why there is a correlation between RRP and marketing expense. (correct discussion **4pts**)

3 Preliminary Customer Analysis (28 pts)

Hereinafter, please

- Use `dplyr` package to answer the preliminary customer analysis questions
- Use `%>%`, i.e., the pipe operator, to chain your dataset operations in all questions in this section, such that you can get the result **in one step**.

4. From `data_full`, count how many distinct TV models does LG sell with "1080p" resolution, HDR support, and "LCD" technology? (**2pts** for code; **2pts** for correct result; **4pts** in total)

    - **Tips:** You can search (e.g., on Google) for the R function that can return the unique distinct values of a vector.

5. Select an appropriate dataset, and use dataset aggregation to count the total number of distinct "4k" "OLED" TV models carried by each brand, ranked from lowest to highest.(**2pts** for correct code for aggregation; **2pts** for correct count; **2pts** for correct ranking; **6pts** in total)

6. From `data_product` (**8pts** in total),

    a. use dataset aggregation to compute the mean of *RRP* for "4k" TVs across "OLED" and "LCD" groups (**2pts**).

    b. Discuss if the mean RRP across the above two groups are statistically different by running an appropriate statistical test in R (**4pts**).

    c. Discuss the potential causes of the price difference across OLED and LCD TVs (open question,**2pts**)

7. Aggregate and compute the total sales of each of the 4 TV brands over all weeks (i.e., the total sales per brand), ranked from the highest to the lowest (**2pts** for correct code for aggregation; **2pts** for correct total sales; **2pts** for correct ranking; **6pts** in total)

8. For Samsung, which week has the highest sales of its "QLED" technology TVs (**2pts** for correct code; **2pts** for correct answer; **4pts** in total)

4 Simple Linear Regression (28 pts)

For all regression related questions, please use `summary()` function to print out the regression results so that our TAs would be able to directly examine the results.

9. Run a univariate regression specified as follows (**5 pts** in total for a, b questions)

$$sales = a + b * finalprice + e$$

   a. Report the results from R using `summary()` function. **1pts** for correct codes for regression; **1pts** for correct regression output)

   b. Interpret the estimated coefficient b (including coefficient and statistical significance, **2pts**) and R-squared of the regression (**1pts**)

10. Run a univariate regression specified as follows (**7pts** in total)

    - Tips: As some TV models have sales of zero during certain weeks, taking log will generate NAs in R; a common technique to circumvent this problem is to add a small number to sales within log.

$$\log(sales + 0.01) = a + b * \log(finalprice) + e$$

   a. Report the results from R using `summary()` function. (**1pts** for correct codes for regression; **1pts** for correct regression output)

   b. Interpret the estimated coefficient b (including coefficient and statistical significance; **2pts**) and R-squared of the regression (**1pts**)

   c. Based on R-squared in questions 9 and 10, discuss whether you should take log-transformation? (**2pts**)

11. We can add brand dummy variables in the regression, so that the coefficients of the brand dummies represent the brand equity of each brand relative to the baseline group (i.e., the one group that's dropped out from the regression); that is, the larger the coefficient of a brand dummy, ceteris paribus, the higher the sales, and the "difference" is due to a brand's perceived premium by customers. (**8pts** in total for a, b, and c questions)

    a. From the available datasets, design and run an appropriate regression that can examine the brand effects (**4pts**; please carefully think about what variables to include and not to include in the regression)

    b. Interpret the coefficients of brand dummies (**2pts**)

    c. Discuss which brand has the highest brand equity perceived by customers? (**2pts**)

Tips:

- You can use `factor(brand)` in R to create brand dummies
- When designing the regression, think about the practical tips I have shown you in designing regressions in week 4.

12. Answer the following questions to understand the effect of screen size on sales. (**8pts** in total for a, b, and c questions)

    a. From the available datasets, design and run an appropriate regression that examines the effect of *screensize* (**4pts**; please carefully think about what variables to include and not to include in the regression)

    b. Interpret the coefficient(s) for screen size related variables (**2pts**).

    c. Discuss whether customers have a monotonic preference for larger screens? (i.e., ceteris paribus, the larger the screen, the higher the sales?) (**2pts**)

5 Endogeneity and Instrumental Variables (14pts)

13. According to economic theory, a higher price should normally lead to a lower quantity of sales. **Fully discuss** any possible reasons why the two regressions from the above give counter-intuitive results. (**6pts**)

14. One way to obtain the unbiased estimates for price is to use the instrumental variable method.

    a. Discuss the instrumental variables candidate you would use for the endogenous TV prices in the dataset. (**4pts**)

    b. Denote the instrument you found as Z. Describe the steps of how would you obtain the unbiased estimates for price using two-stage least square method. (**4pts**)

Thank you for solving Tom's problem with his son. Don't forget to ask Tom for consultancy fee next time you run into him!