DSCC/CSC 462 Final Project

Due: Thursday, December 15, 2022 at 11:59 p.m. EST

This is the final group project for DSCC/CSC 462. You are expected to complete this project in groups of 3-4 people. You are only allowed to discuss the final project with your assigned group members. You are not permitted to talk with anyone outside of your group aside from the professor and course teaching assistants. You are permitted to use any approved course material for this project. You are NOT allowed to use paid online resources outside of what is available through Blackboard (i.e., Chegg or similar sites).

Your group needs to submit a single file for all of you (i.e., you do not all upload the document; one group member does it, and that's sufficient). Only one person will upload it to their Blackboard at the assignment submission link, but please ensure that all group members' names are listed on the uploaded submission. All group members are expected to equally contribute to the final project, and you are responsible for knowing everything going on in the project (i.e., one person can't just be responsible for one question and not know what happened in the other questions; if I were to email you and ask you to explain to me what your group did for any questions, you should be very comfortable explaining the answer).

The project must be completed in RMarkdown. You will upload both your knitted PDF document and RMD file to Blackboard to be graded. Ensure that all group member's names are on the knitted PDF, and **please put your group number on the document as well**. Please take the time to ensure your interpretations are thorough and thoughtful. Use the spellcheck feature to catch typos.

The dataset your group should use is available on Blackboard next to your group assignment. The file will be of form "ads#.csv", where # is your group number (i.e., group 1 should use file "ads1.csv"). Make sure you are using the correct file listed with your group!

Please keep an eye out for Blackboard announcements I may post relating to the project and procedures for submission.

Important: Please label all graphs (title, axes, and legend, if applicable), and for any hypothesis tests, clearly state the hypotheses, test statistic, and p-value (or rejection region, if you choose to use it). You may use all `.test()` functions, i.e., no need to do any tests "by-hand." We also highly recommend that you use ggplot for graphics, although you may also use base R graphics.

Introduction

Social media is a remarkably effective tool for connecting people to one another. It is also a very powerful medium for advertising. In this final project, you will be examining ad data from a company that is interested in learning how to more efficiently advertise its products on social media. In the company's data, each line corresponds to an ad. For each ad, the company records which social media platform led to the sale (`socialmedia`), the age of the customer (`age`), whether the customer was on mobile or a computer (`mobile`), the season of the year when the sale took place (`season`), whether or not the customer was a new customer (`newcustomer`), the cost of the ad (`adcost`), and the revenue generated by that customer (`adrevenue`).

Your goal is to (1) help the company understand the data they have collected, and (2) provide recommendations about what the company should do moving forward. The following are guidelines for the data analyses you should run, but many have been purposefully left open-ended so you can be creative in how you approach the problem.

Project Description

The CEO (chief executive officer) and CFO (chief financial officer) have come up with the following list of questions. In order to answer them, you will use tools for both descriptive statistics and inferential statistics that we learned during class. Please describe your thought process behind how you approach each question and clearly explain any visualizations (e.g., histograms, side-by-side boxplots, scatterplots, frequency tables, and/or barplots) and tests (e.g., hypotheses, test statistics, p-values, and/or critical values) used.

1. The marketing team first wants to understand how many ads they are running on each social media platform, as well as the demographics of each social media platform's user base.
    a. Create a relative frequency table and a corresponding relative frequency barplot to visualize the fraction of ads on each platform. Make sure to label the plot (title, axes), and comment on trends you observe.
    b. The CFO's ad strategy is supposed to run 10% of all ads on Twitter, 10% on Facebook, 20% on Instagram, 30% on TikTok, and 30% on YouTube. Is the marketing department following this strategy? Run an appropriate statistical test at the $\alpha = 0.05$ significance level and comment on the results.
    c. For each social media platform, calculate the variance, standard deviation, coefficient of variation, and skew of age, and visualize the distribution of age using an appropriate tool from descriptive statistics. Comment on any trends you see.
2. The CEO of the company believes that ads differ in effectiveness (measured in terms of profit, or ad revenue - ad cost) depending on the season. However, is his intuition correct?
    a. Visualize the data using four histograms (one from each season). On each plot, draw and label vertical lines for the mean, median, and 10% trimmed mean. Make sure to label the plots (title, axes, legend), and comment on trends you observe.
    b. In particular, the CEO believes that summer ads yield more profit than winter ads. At the $\alpha = 0.05$ significance level, run an appropriate statistical test (or series of tests) and comment on your results.
    c. What if you wanted to compare all seasons at once at the $\alpha = 0.05$ significance level with a familywise error of $\alpha_{FWE} = 0.05$? Run an appropriate test (or series of tests) and comment on your results.
3. The CFO wants to know if the mean profits (ad revenue - ad cost) are the same on each platform, but he adds the stipulation that the Type I error of the analysis can be at most 5%, and the familywise error of any follow-up tests can also be at most 5%. If mean profits are indeed not equal on each platform, please identify pairs of platforms for which there is a statistically significant difference between mean profits.
    a. Visualize the mean profits for each platform using side-by-side boxplots. Identify any outliers and comment on trends.
    b. Perform an appropriate statistical test (or series of statistical tests) and comment on your findings.

4. The CFO also wants to better understand the relationship between acquiring new customers and net profit.
    a. Visualize the relationship between whether or not someone is a new customer and the net profit off of that customer using an appropriate tool from descriptive statistics. Comment on any trends you observe.
    b. Is advertising on different social media platforms associated with different rates of acquiring new customers? Run an appropriate statistical test at the $\alpha = 0.05$ significance level and comment on the results.
    c. Construct a two-sided 95% confidence interval for the proportion of ads that lead to new customers.
    d. An analyst on another team claims that acquiring new customers is more profitable than trying to sell more products to existing customers. Test their claim at the $\alpha = 0.05$ significance level, and comment on your results.
5. The CEO and CFO disagree about whether being on a mobile phone affects average profits. The CEO thinks that being on your phone or computer doesn't affect overall profits, whereas the CFO thinks there is a difference.
    a. Visualize profits by mobile phone status for each social network platform using an appropriate tool from descriptive statistics. Make sure to label the plot (title, axes, legend), and comment on trends you observe.
    b. At the $\alpha = 0.05$ significance level, examine whether or not being on a mobile phone affects average profits for each social network platform. Discuss your findings.
6. It's time to start understanding how profit depends on other variables, notably ad cost and age.
    a. Visualize the relationship between profit and age with different colored points for each social network. Make sure to label the plot (title, axes, legend), and comment on trends you observe.
    b. Visualize the relationship between profit and ad cost, again with different colored points for each social network. Make sure to label the plot (title, axes, legend), and comment on trends you observe.
    c. Are profit and age correlated? Perform an appropriate statistical test using *Pearson* correlation at the $\alpha = 0.05$ significance level and comment on the results.
    d. Are profit and ad cost correlated? Perform an appropriate statistical test using *Spearman* correlation at the $\alpha = 0.05$ significance level and comment on the results.
    e. Fit a linear regression to model the profit as a function of ad cost. Report the regression equation, a 90% confidence interval for the coefficient of ad cost, and the coefficient of determination.
    f. Fit a linear regression to model the profit as a function of ad cost and age. Comment on the results.
    g. At the $\alpha = 0.05$ significance level, conduct an F-test to determine whether ad cost significantly predicts profit once we have accounted for age. Report the test statistic and p-value, and interpret the results within the context of the problem.
7. Suppose you are given \$100 to spend on advertising for this company. How would you spend it? Explain and interpret any additional analyses you want to do, and provide a detailed description of why you used the analyses you did. This portion should involve significant thought, perhaps partially based on the types of analysis you did earlier. (Hint: It may help to consider each social media platform separately.)