Final Project

Due: Monday, December 14, 2020 at 2 p.m. EST

This is the final group project for DSCC/CSC/STAT 262/462. You are assigned a group to work with (group assignments are on Blackboard). 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 online resources outside of what is available through Blackboard (i.e. no Chegg, stack exchange/overflow type sites, Google, etc.). If it is not official course material or part of the R help files, you cannot use it.

Your group needs to submit a single file for all three 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 will receive the same group for their group's 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 you group did for any questions, you should be very comfortable with the submission).

The project must be completed in RMarkdown. You will upload your knitted PDF document to Blackboard to be graded (if you like to be safe and want to also upload the RMD file in your same submission, that is fine, but make sure you have the PDF uploaded). 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 "hawaii#.csv", where # is your group number (i.e. group 1 would have file "hawaii1.csv"). Make sure you are using the correct file listed with your group!

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

Students enrolled in 462 are required to use ggplot for their graphics (aside from the boxcox plot that defaults from the boxcox() function). Students in 262 can choose whether they use base R graphics or ggplot.

We all could use a vacation right now to escape the craziness of the pandemic and life generally. Since vacations are not too realistic, we instead are going to pretend we work for a Hawaiian Islands travel agency and are analyzing trends among visitors.Suppose the travel agency has given you data on visitors that have recently traveled to Hawaii. Information collected on each visitor is the island visited (`island`), traveler's age (`age`), reason for travel (`reason`), length of stay (`length`), season when they visited (`season`), whether or not it was their first time visiting a Hawaiian island (`firstTime`) and total cost of trip in USD (`cost`).

1. Begin by reading the data file in R.
2. Examine the variable `island`. There are four main islands that visitors stay on: Big Island, Kauai, Maui, or Oahu. Assume that each visitor has only traveled to one island.
    a. Create a relative frequency table of `island`.
    b. Create a barplot of the absolute frequency of `island`, making each of the four bars a different color. Include a title of "Barplot of Island Visited".
    c. Conduct an appropriate hypothesis test at the $\alpha = 0.05$ significance level to see if the proportion of visitors to each island are equal to one another. Report the test statistic and p-value, and interpret the results within the context of the problem.
3. Examine the age of visitors (`age`).
    a. Calculate the five-number summary of age.
    b. Create a modified boxplot of age, labeling the y-axis as "Visitors' Ages".
    c. Create a histogram of age, and color the bars yellow. Add a red vertical line on the histogram at the median of age, and report the calculated value of the median in red next to the line using the form $\tilde{x}$. Briefly comment on the distribution.
    d. Calculate the standard deviation of age.
    e. Calculate the skewness of age.
    f. Construct a one-sided upper-bound 90% confidence interval for the mean age. Interpret the interval within the context of the problem.
    g. Evaluate how well the empirical rule applies to age. In particular, calculate the proportion of ages that falls within one, two, and three standard deviations of the mean. Compare these values to the theoretical percentages as stated by the empirical rule. Overall, does the empirical rule do a good job at describing age? Briefly justify your response.
    h. Construct a normal qq plot for age (include the 1:1 line). Comment on the plot.
    i. Use a Box-Cox power transformation to determine an appropriate transformation of age. In particular, use the boxcox() function in the MASS library to determine the optimal transformation. Report the optimal transformation, but do not apply it to the data.
    j. Apply a log transformation to age of visitors. Create a histogram of the transformed data. Comment on the plot.
4. Examine the island visited (`island`) by reason for travel (`reason`).
    a. Create a two-way table with marginal totals of the joint distribution of `island` and `reason`.
    b. Construct a stacked barplot with a bar for each island. Each bar should be broken into two pieces: one for those who are traveling for business and one for those who are traveling for pleasure.. Color the business pieces blue and color the pleasure pieces green. Label the x-axis to say "Island". Add a legend (it is fine if the legend covers part of the plot itself, as long as the plot can still be reasonably read).
    c. Conduct an appropriate test at the $\alpha = 0.05$ significance level to determine if `island` and `reason` are associated with one another. Make sure to state the hypotheses, report the test statistic and p-value, and interpret the results in the context of the problem.

5. Examine length of stay (`length`) as a function of season (`season`).

   a. Create side-by-side boxplots of length of stay for each of the four different seasons.

   b. Does the mean length of stay differ by season? To test this, construct an ANOVA table for testing $H_0 : \mu_f = \mu_w = \mu_{su} = \mu_{sp}$ at the $\alpha = 0.05$ significance level. State the value of the test statistic and the p-value, and interpret the results within the context of the problem.

   c. Further explore the results from question 5b using a Bonferroni multiple comparison procedure with an overall familywise error rate of $\alpha_{FWE} = 0.05$. Which means are significantly different from each other?

6. Examine `age` as a function of `firstTime`.

   a. At the $\alpha = 0.01$ significance level, test whether the variance of the age is different for those visiting Hawaii for the first time and those making a return visit (i.e. $H_0 : \sigma_F^2 = \sigma_R^2$ vs. $H_1 : \sigma_F^2 \neq \sigma_R^2$). Report the test statistic and p-value, and interpret the results within the context of the problem.

   b. Conduct an appropriate two-sample t-test at the $\alpha = 0.05$ significance level to determine if the mean age of first-time visitors is less than the mean age of return visitors ($H_0 : \mu_F - \mu_R \geq 0$ vs. $H_1 : \mu_F - \mu_R < 0$. Use your results from question 6a to determine whether or not to assume equal variances. Report the test statistic and p-value, and interpret the results within the context of the problem.

7. Examine cost of trip (`cost`) as a function of other variables.

   a. Construct a scatterplot of cost of trip over length of stay (`length`), plotting points as $\times$ instead of the default $\circ$. Briefly comment on the plot and whether linear regression seems appropriate.

   b. Calculate the Spearman correlation coefficient between length of stay and cost of trip.

   c. Conduct a test at the $\alpha = 0.05$ significance level to determine if the Pearson correlation between length of stay and cost of trip is different from 0 (i.e. $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$). Report the test statistic and p-value, and interpret the results within the context of the problem.

   d. Fit a linear regression to model cost of trip as a function of length of stay. Call this model `lm1`. Report the regression equation.

   e. Construct a two-sided 90% confidence interval for the coefficient of length of stay from `lm1`. Interpret the interval.

   f. What is the value of the coefficient of determination for `lm1`? Interpret this value within the context of the question.

   g. Construct diagnostic plots for `lm1`, and briefly comment on the fit of the model with respect to linearity and normality assumptions.

   h. Fit a linear regression to model the cost of trip as a function of both length of stay and island visited (i.e. `length and island`). Report the regression equation.

   i. At the $\alpha = 0.05$ significance level, conduct an F-test to determine whether island visited significantly predicts the cost of trip once we have accounted for length of stay. Report the test statistic and p-value, and interpret the results within the context of the problem.