More Statistical Analyses

Open RStudio and create a new project under your Module 6 folder and call it **Mod6Assignment2.** For this assignment, you will be creating an R Markdown document that will include topics including regression anlaysis, t-tests, ANOVAs, etc.

Create the R Markdown Document

1.) In RStudio, select *File -> New File -> Text File*. This will create a blank text file in the same area that scripts were created in previous assignments (upper left panel). Save this file to your project as **Mod6Assign2Answer.rmd** (it is important to save with the .rmd extension as this saves the text file as an R Markdown file).

2.) Create a Header 1 with the title: **Module 6 - Assignment 2**

3.) Create a Header 2 with the title: **Last Name, First Name** (replace with your name)

4.) Create a Header 3 with the title: **Statistical Analyses**

5.) Click on the dropdown arrow next to the Knit icon Knit at the top of the R Markdown Pane in RStudio and select Knit to Word.

6.) Notice that you now have a document in your files for the project named **Mod6Assign2Answer.docx**. This is the file you will be uploading later to Canvas.

7.) For this assignment, you will need to download all of the datasets listed on the assignment from Canvas. Save these in the same folder that you created the project in.

8.) Create a new chunk of R code that will load both the tidyverse.

9.) We will also be using some new packages that need to be loaded. From the console (not R Markdown document), run the install.packages() command for the following:

a. corrplot

Part 1: Importing the dataset

10.) Within R Studio, click on the files you just added to the project folder and select Import Dataset…

11.) Instead of importing using the Import Text Data screen, copy the code in the code preview and save it in the chunk of code you created to load the libraries needed for this exercise. Note: delete the view(dataset) command within the chunk of code.

12.) Do this for each of the files you have downloaded to the project folder which include:

  a. RespiratoryExchangeSample.xlsx
  b. Advertising.csv
  c. insurance.csv
  d. Perceptions.xlsx

Part 2 – Regression and Correlation

13.) Create a new title labeled: Regression and Correlation

14.) After the title, add the following text:

Regression analysis is a statistical method that allows you to examine the relationship between two or more variables of interest. Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables (e.g. height and weight). This particular type of analysis is useful when a researcher wants to establish if there are possible connections between variables.

15.) Create a new title: Insurance Costs

16.) Add the following text after the title:

We would like to determine if we can accurately predict insurance costs based upon the factors included in the data. We would also like to know if there are any connections between variables (for example, is age connected or correlated to charges).

17.) This dataset includes a number of factors that may impact medical insurance costs. The following columns of data are included within this dataset:

  a. age: age of primary beneficiary
  b. sex: insurance contractor gender, female, male
  c. bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
  d. children: Number of children covered by health insurance / Number of dependents
  e. smoker: Smoking
  f. region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest
  g. charges: Individual medical costs billed by health insurance

18.) For this part, you will be using the Insurance.csv dataset you imported.

19.)Add a new title: Correlations of bmi, age, children and cost

20.)Add a new chunk of code. We would first like to look at correlations among the variables in our dataset. Using previous tidyverse packages, create a new dataset called Insurance2 that will include only the age, bmi, children and charges. Next, add the code that will run a correlation of all the variables in Insurance2.

21.)We would also like to visualize the correlation matrix using the package (corrplot) you installed at the beginning of this exercise. Add an additional line of code that will save your correlation matrix to a variable (e.g., Corr_matrix <- *the correlation code goes here*). Finally, add the following code to the current chunk:

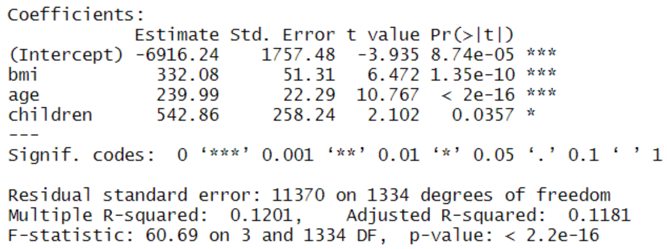
```
library(corrplot)
library(RColorBrewer)

corrplot(Corr_matrix, type="upper", order="hclust",
     col=brewer.pal(n=8, name="RdYlBu"))
```

22.)**Based on the matrix and visuals, explain the results from your correlation matrix in a paragraph after the chunk of code. Are any of the variables highly correlated?**

23.) Add a new title: Regression Analysis

24.)Add a new chunk of code that will include a multiple linear regression of the data within the Insurance2 dataset you created during the correlation analysis. Specifically, we want to see the impact of age, bmi and children on charges. For this code, follow the guidelines from Cookbook for R reading this week, specifically I want to see both the coefficients and summary of output (see below).



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6916.24    1757.48  -3.935 8.74e-05 ***
bmi           332.08      51.31   6.472 1.35e-10 ***
age           239.99      22.29  10.767  < 2e-16 ***
children      542.86     258.24   2.102   0.0357 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11370 on 1334 degrees of freedom
Multiple R-squared:  0.1201,    Adjusted R-squared:  0.1181 
F-statistic: 60.69 on 3 and 1334 DF,  p-value: < 2.2e-16
```

25.)The results show both the coefficient estimates and p-values. The coefficient will tell the impact of each independent variable (e.g., age) on the dependent variable (i.e., charges) in our model. The p-value will tell you if the variable is significant or not.

Write a paragraph after the code. Based on the results, which variables were significant and what particular significant variable had the largest impact on charges? Provide a summary after the chunk of code.

26.) Add a new chunk of code. If you noticed, we can only run numeric values for a regression but we are still interested if sex (i.e., gender) and smoker impacts charges. To do this, we will need to create a new variable for each that will be coded as 1 or 0. Add the following code to change sex (or gender) to be 1 or 0:

Insurance <- mutate(Insurance, gender=ifelse(sex=="female",1,0))

This code will create a new variable called gender and uses the *ifelse* command. The *ifelse* examines the sex field to see if female is listed. If it is, a 1 will be placed in gender while if it is male, a 0 will be placed into gender. The code is structured such as:

ifelse (test,value_if_true, value_if_false)

Using code similar to that above, create a new variable called smoker2 that will assign a value of 1 if they are a smoker and a value of 0 if they are not.

27.) Finally, run another regression including all the values from the previous regression (i.e., age, bmi and children) with the addition of gender and smoker2. **After the code, provide an explanation of the new results. Does gender and smoking have an impact on cost?**

<u>Part 3 – Group Comparisons</u>

28.) Create a new title: Group Comparisons with t-tests

29.) After the new title, add the following text explaining what a t-test is:

The t-test is used to compare the values of the means from two samples and test whether it is likely that the samples are from populations having different mean values. This is often used to compare 2 groups to see if there are any significant differences between these groups.

30.) Create a new title: Caffeine Impacts on Respiratory Exchange Ratio

31.) After this title, add the following text explaining what this dataset is and what the study is trying to find out:

A study of the effect of caffeine on muscle metabolism used volunteers who each underwent arm exercise tests. Half the participants were randomly selected to take a capsule containing pure caffeine one hour before the test. The other participants received a placebo capsule. During each exercise the subject's respiratory exchange ratio (RER) was measured. (RER is the ratio of CO_2 produced to O_2 consumed and is an indicator of whether energy is being obtained from carbohydrates or fats).

The question you are trying to answer is whether caffeine impacts RER during exercise.

32.) For this part, you will be using the RespiratoryExchangeSample.xlsx dataset you imported.

33.) Add a new chunk of code. First, add code that will give you summary statistics about the data. Then, run a t-test between the 2 groups (Caffeine and Placebo) to see if there was a difference between participants taking caffeine compared to those with the placebo. You should use the default Welch technique for this t-test.

(**Hint: Since the data is already in the "wide-date" format, use the t-test code for two separate vectors (i.e., wide data) approach from the Cookbook for R reading this week**)

34.) **Finally, interpret your results in a paragraph following the code. You can examine the p-value to see if there was a significant difference between the groups**.

35.) Create a new title: Impact of Advertising

36.) After this title, add the following text explaining what this dataset is and what the study is trying to find out:

You are a marketing researcher conducting a study to understand the impact of a new marketing campaign. To test the new advertisements, you conduct a study to understand how consumers will respond based on see the new ad compared to the previous campaign. One group will see the new ad and one group will see the older ads. They will then rate the ad on a scale of 0 to 100 as a percentage of purchase likelihood based on the ad.

The question you are trying to answer is whether to roll out the new campaign or stick with the current campaign.

37.) For this part, you will be using the Advertising.csv dataset you imported.

38.) Add a new chunk of code. First, add code that will give you summary statistics about the data. Then, run a t-test between the 2 groups (Group 1 is new ad, Group2 is old ad) to see if there was a difference between participants seeing the new advertising campaign compared to those not seeing the new ads. You should use the Student's t-test technique instead of the default Welch.

(**Hint: use the code for Student t-test approach from the Cookbook for R reading this week**)

39.) Finally, interpret your results in a paragraph following the code. You can examine the p-value to see if there was a significant difference between the groups. You are trying to answer the question if the new advertising campaign should move forward.

<u>Part 4 – Analysis of Variance</u>

40.) Create a new title: ANOVA

41.) After the new title, add the following text explaining what a t-test is:

An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis. Basically, you're testing groups to see if there's a difference between them. Examples of when you might want to test different groups:

- *A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others.*
- *A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.*
- *Students from different colleges take the same exam. You want to see if one college outperforms the other.*

42.)Create a new title: Perceptions of Social Media Profiles

43.)After this title, add the following text explaining what this dataset is and what the study is trying to find out:

This study examines how certain information presented on a social media site might influence perceptions of trust, connectedness and knowledge of the profile owner. Specifically, participants were shown weak, average and strong arguments that would influence their perceptions of the above variables. Using the dataset provided, the following code runs an ANOVA with post-hoc analyses to understand argument strength impacts on perceptions.

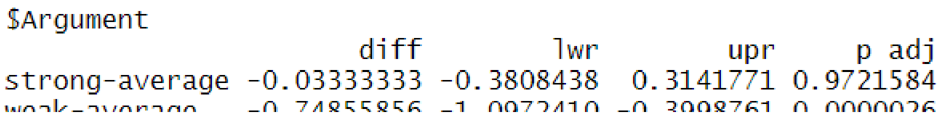
44.)For this part, you will be using the Perceptions.xlsx dataset you imported.

45.)Add a new chunk of code after the paragraph. In this chunk, run three different ANOVAs using the instruction provided in the Cookbook for R (use of aov command) reading this week:

a. ANOVA 1 – examine the difference of Trust across Argument
b. ANOVA 2 – examine the difference of Connectedness across Argument
c. ANOVA 3 – examine the difference of Knowledge across Argument

46.)**Examine the results in a paragraph after the code chunk, specifically looking at the p-value (Pr(>F)) to see which of the ANOVAs above were significant. You should see that two of these ANOVAs are significant.** However, we know the groups are different but we don't know which specific group is different from another. For example, we know there is a difference between weak, average and strong but we don't know the difference of weak to average, weak to strong and average to strong. To do this, run a post-hoc analysis (TukeyHSD) on the significant ANOVA models.

47.)You should now see a table showing the comparisons above (see the example below):



```
$Argument
                      diff        lwr        upr     p adj
strong-average -0.03333333 -0.3808438  0.3141771 0.9721584
weak-average   -0.74855856 -1.0972410 -0.3998761 0.0000026
```

48.)We examine the p adj in this table to see which groups are different. In the example above, we can see that the difference between a strong and average argument is not significant as p adj is 0.97 (to be significant it should be 0.05 or less). **Based on your post-hoc analysis, add some additional text to the paragraph explaining your results.**

49.) Finally, knit your R Markdown to Word. Before uploading the word document, open it and make any changes to the layout to make the presentation of the material better. Next, upload the document to Canvas.