# Module 6 - Assignment 2

## Black, Tyler

### Statistical Analyses

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(readxl)  
Advertising <- read\_csv("Advertising.csv")

## Rows: 1000 Columns: 3  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (3): ID, Rating, Group  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Insurance <- read\_csv("Insurance.csv")

## Rows: 1338 Columns: 7  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (3): sex, smoker, region  
## dbl (4): age, bmi, children, charges  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Perceptions <- read\_excel("Perceptions.xlsx")  
RespiratoryExchangeSample <- read\_excel("RespiratoryExchangeSample.xlsx")

### Regression and Correlation

*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.*

### Insurance Costs

*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).*

### Correlations of bmi, age, children and cost

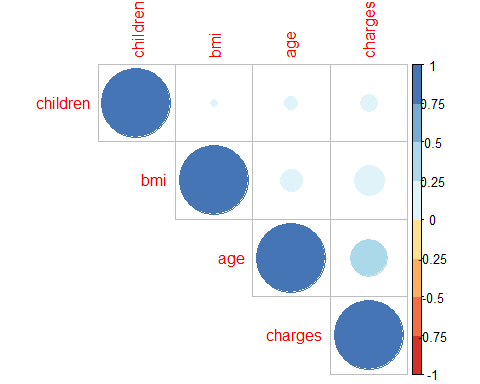
Insurance2 <- select(Insurance, age, bmi, children, charges)  
cor(Insurance2, use = 'everything', method = 'pearson')

## age bmi children charges  
## age 1.0000000 0.1092719 0.04246900 0.29900819  
## bmi 0.1092719 1.0000000 0.01275890 0.19834097  
## children 0.0424690 0.0127589 1.00000000 0.06799823  
## charges 0.2990082 0.1983410 0.06799823 1.00000000

library(corrplot)

## corrplot 0.92 loaded

library(RColorBrewer)  
Corr\_matrix <- cor(Insurance2, use = 'everything', method = 'pearson')  
corrplot(Corr\_matrix, type="upper", order="hclust",  
col=brewer.pal(n=8, name="RdYlBu"))



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?  
**No, none of the variables are highly correlated to one another. Age and charges are the ‘the most’ correlated with a coefficient of .299, but this is not highly correlated**

### Regression Analysis

charges\_fit <- lm(charges ~ age + bmi + children, data = Insurance2)  
summary(charges\_fit)

##   
## Call:  
## lm(formula = charges ~ age + bmi + children, data = Insurance2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13884 -6994 -5092 7125 48627   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6916.24 1757.48 -3.935 8.74e-05 \*\*\*  
## age 239.99 22.29 10.767 < 2e-16 \*\*\*  
## bmi 332.08 51.31 6.472 1.35e-10 \*\*\*  
## 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

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.  
**All the variables are significant in terms of their p-value, but children have the biggest impact on charges per its coefficient**

Insurance <- mutate(Insurance, gender=ifelse(sex=="female",1,0))  
Insurance <- mutate(Insurance, smoker2=ifelse(smoker=="yes",1,0))  
Insurance3 <- select(Insurance,age,bmi,children,charges,gender,smoker2)  
charges\_fit2 <- lm(charges ~ age + bmi + children + gender + smoker2, data = Insurance3)  
summary(charges\_fit2)

##   
## Call:  
## lm(formula = charges ~ age + bmi + children + gender + smoker2,   
## data = Insurance3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11837.2 -2916.7 -994.2 1375.3 29565.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12181.10 963.90 -12.637 < 2e-16 \*\*\*  
## age 257.73 11.90 21.651 < 2e-16 \*\*\*  
## bmi 322.36 27.42 11.757 < 2e-16 \*\*\*  
## children 474.41 137.86 3.441 0.000597 \*\*\*  
## gender 128.64 333.36 0.386 0.699641   
## smoker2 23823.39 412.52 57.750 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6070 on 1332 degrees of freedom  
## Multiple R-squared: 0.7497, Adjusted R-squared: 0.7488   
## F-statistic: 798 on 5 and 1332 DF, p-value: < 2.2e-16

After the code, provide an explanation of the new results. Does gender and smoking have an impact on cost?  
**All the variables except *gender* have an impact on charges. Smoking has the highest impact**

### Group Comparisons with t-tests

*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.*

### Caffeine Impacts on Respiratory Exchange Ratio

*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 CO2 produced to O2 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.*

summary(RespiratoryExchangeSample)

## Placebo Caffeine   
## Min. : 80.00 Min. :100.0   
## 1st Qu.: 85.00 1st Qu.:106.0   
## Median : 90.00 Median :110.5   
## Mean : 90.11 Mean :110.8   
## 3rd Qu.: 95.25 3rd Qu.:117.0   
## Max. :100.00 Max. :120.0

t.test(RespiratoryExchangeSample$Placebo,RespiratoryExchangeSample$Caffeine)

##   
## Welch Two Sample t-test  
##   
## data: RespiratoryExchangeSample$Placebo and RespiratoryExchangeSample$Caffeine  
## t = -33.742, df = 397.67, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -21.95369 -19.53631  
## sample estimates:  
## mean of x mean of y   
## 90.105 110.850

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.  
**Since the p-value is extremely large, we reject the null hypothesis in favor of the alternative hypothesis that the difference in means is not equal to 0 (the groups have different mean values)**

### Impact of Advertising

*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.*

summary(Advertising)

## ID Rating Group   
## Min. : 1.0 Min. : 0.00 Min. :1.000   
## 1st Qu.: 250.8 1st Qu.: 25.75 1st Qu.:1.000   
## Median : 500.5 Median : 53.00 Median :1.000   
## Mean : 500.5 Mean : 51.06 Mean :1.499   
## 3rd Qu.: 750.2 3rd Qu.: 76.00 3rd Qu.:2.000   
## Max. :1000.0 Max. :100.00 Max. :2.000   
## NA's :184

t.test(Rating ~ Group, Advertising, var.equal = TRUE)

##   
## Two Sample t-test  
##   
## data: Rating by Group  
## t = 1.2509, df = 814, p-value = 0.2113  
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0  
## 95 percent confidence interval:  
## -1.440198 6.501170  
## sample estimates:  
## mean in group 1 mean in group 2   
## 52.33827 49.80779

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.  
**Since the p-value is quite large, we support the null hypothesis that the difference in means is equal to zero, and we discourage the new advertising campaign.**

### ANOVA

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.

### Perceptions of Social Media Profiles

*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.*

anova1 <- aov(Trust ~ Argument, Perceptions)  
summary(anova1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Argument 2 26.59 13.293 16.34 2.4e-07 \*\*\*  
## Residuals 221 179.75 0.813   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova2 <- aov(Connectedness ~ Argument, Perceptions)  
summary(anova2)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Argument 2 29.7 14.859 9.869 7.85e-05 \*\*\*  
## Residuals 221 332.7 1.506   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova3 <- aov(Knowledge ~ Argument,Perceptions)  
summary(anova3)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Argument 2 0.47 0.2333 0.315 0.73  
## Residuals 221 163.67 0.7406

TukeyHSD(anova1)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Trust ~ Argument, data = Perceptions)  
##   
## $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  
## weak-strong -0.71522523 -1.0639077 -0.3665427 0.0000073

TukeyHSD(anova2)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Connectedness ~ Argument, data = Perceptions)  
##   
## $Argument  
## diff lwr upr p adj  
## strong-average -0.2733333 -0.7461312 0.1994645 0.3615643  
## weak-average -0.8736637 -1.3480561 -0.3992712 0.0000628  
## weak-strong -0.6003303 -1.0747228 -0.1259378 0.0087959

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  
**The ANOVAs that are significant are, Trust ~ Argument and Connectedness ~ Argument.**

Based on your post-hoc analysis, add some additional text to the paragraph explaining your results.  
**There are significant difference in Trust means between weak-average arguments and weak-strong arguments. Additionally, there are significant difference in Connectedness means between weak-average arguments and weak-strong arguments**