RMarkdown STAT 2218 Project 2

Jack Ambery and Kevin Wilson

Due December 8, 2023

Work in a team of 1, 2 or 3 students. You will hand in 1 typed report per group.

You are hired by a company sells two main types of products remotely by connecting to and talking to customers by phone. They provide a data set that has the following information. Each row includes information from one employee’s shift randomly selected across the different teams, regions and products.

Data: salesdata.csv

* Team: This represents the team that the employee works on (Team A to J). The company is organized into 10 teams.
* Product: Which product was being sold during shift (Product 1 or 2)
* Region: Which state the employee works in
* Shift: The number of hours the employee worked on that shift
* TotalTalk: The total number of hours during the shift the employee was talking to customers on the phone
* TotalHold: The total number of hours during the shift the employee had customers on hold
* TotalNumSales: The total number of items sold during the shift
* TotalRevenue: The total amount of revenue (in $) employee made during the shift
* NumCalls: The total number of calls the employee took during the shift
* TotalNumFeedback: The total number of customers that submitted feedback on the employee during the shift
* PercentSatisfied: The percent of customer surveys that had positive reviews (Excellent or Good ratings)
* NumPos: The number of customer surveys that were positive reviews (Excellent or Good ratings)
* NumNeg: The number of customer surveys that were negative reviews (Fair or Poor ratings)

For this project, the company wants you to look at the factors related to total revenue. In particular:

* Is there a significant difference in average total revenue based region and product, or an interaction of the two factors?
* Does total revenue differ by team?

Your team will write a report based on your analyses, assuming a 5% significance level, that includes your R code, relevant output and analysis for each of following concepts.

# Input data file below.

Import data into RStudio using the base option, and then cut and paste YOUR file directory in the read.csv command below. It may be different from what I have listed below since your files are located in a different directory/folder. Then you can attach the file.

salesdata <- read.csv("C:/Users/jacka/OneDrive - fairfield.edu/Fall 2023/Statistics II/Project2/salesdata.csv")  
attach(salesdata)

# Question 1: Exploratory Analysis (10 pts)

* Does the total revenue seem to differ based on the main effect of region? Does the total revenue seem to differ based on the main effect of product? Explain why or why not and provide (and interpret) graphical and numerical evidence to support your claim.
* Does there appear to be an interaction effect between region and product on total revenue? Explain why or why not and provide (and interpret) graphical and numerical evidence to support your claim.

## Question 1 Exploratory Analysis R Code

tapply(TotalRevenue, Region, length)

## Alabama California Colorado Connecticut Florida Hawaii   
## 21 20 20 20 19 20

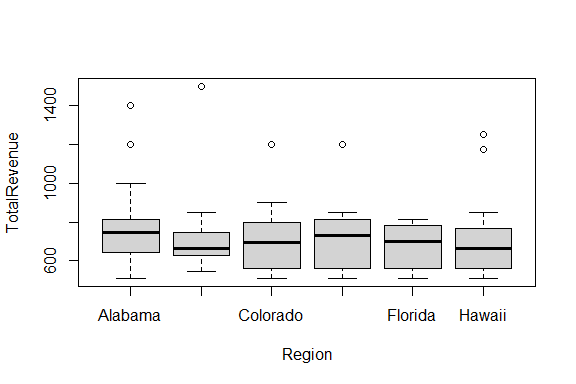
tapply(TotalRevenue, Product, length)

## Product 1 Product 2   
## 60 60

tapply(TotalRevenue, Region, mean)

## Alabama California Colorado Connecticut Florida Hawaii   
## 777.8095 719.7000 709.3000 743.6500 676.8421 700.3500

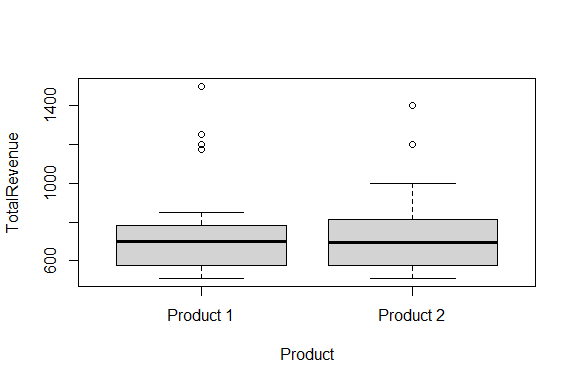
boxplot(TotalRevenue~Region)



tapply(TotalRevenue, Product, mean)

## Product 1 Product 2   
## 724.7333 719.5000

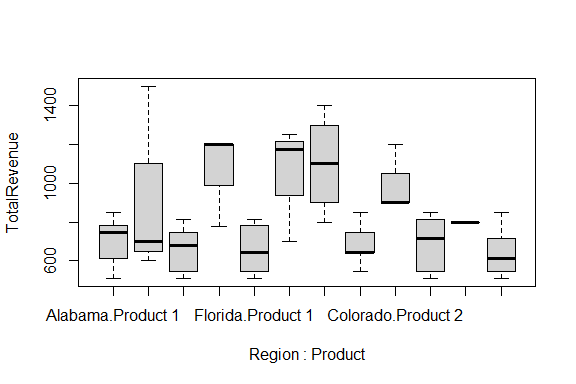
boxplot(TotalRevenue~Product)



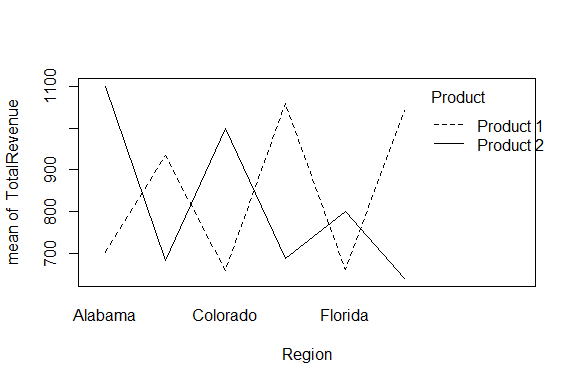
tapply(TotalRevenue, list(Region, Product), mean)

## Product 1 Product 2  
## Alabama 702.0000 1100  
## California 933.3333 682  
## Colorado 658.0000 1000  
## Connecticut 1059.0000 688  
## Florida 662.3529 800  
## Hawaii 1042.3333 640

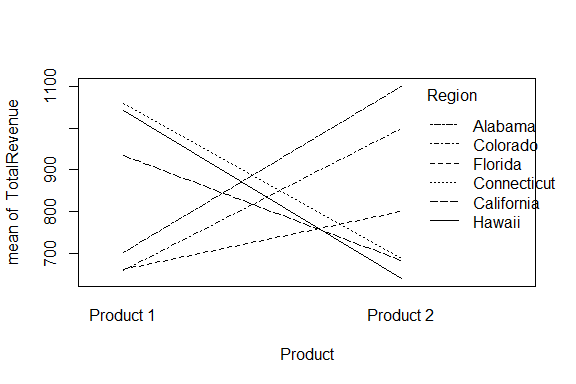
boxplot(TotalRevenue~Region\*Product)



interaction.plot(Region, Product, TotalRevenue)



interaction.plot(Product, Region, TotalRevenue)



## Question 1 Discussion of Exploratory Analysis

While there is a range of about 100 for the average total revenue by region, the boxplots of the total revenue by region seem to be pretty similar, with most regions having far outliers. Thus, it is difficult to tell if the average total revenue differs based on the main effect of region, but a hypothesis test would help clarify this.

The mean total revenue for product 1 is 724.7333 and the mean total revenue for product 2 is 719.5000. These means are similar and the boxplots of total revenue for product 1 and product 2 also appear to be very similar. Thus, upon initial exploratory analysis, it does not appear that the average total revenue differs based on the main effect of product.

The pairwise means of total revenue for product and region seem very different. For example, the mean total revenue for product 1 sold in Colorado is 658, whereas the mean total revenue for product 2 sold in Alabama is 1100. The pairwise boxplots are also very different. Also, the interaction plots display many points of intersection, which also indicates a significant difference in pairwise means of total revenue for product and region. Therefore, there appears very likely that there is a significant interaction effect between region and product on the average of total revenue.

# Question 1 Hypothesis Testing and Models (20 pts)

* Test if the average total revenue depends on an interaction between team and product.
* If not, test the main effects.
* For each test you run use a significance level of 5%. Be sure to write Ho and Ha, give the R code and relevant output, and write your conclusion specifically for this problem.
* If necessary, run a post-hoc test and interpret the findings specifically for this problem. If a post-hoc test is not needed, explain why.
* Assess your final model by determining if each of the assumption conditions is met or not. Interpret results and include evidence to support your claims.
* Is there anything else you would do or recommend?

## R Code for Hypothesis Tests

results = aov(TotalRevenue ~ Region \* Product)  
  
library(car)

## Loading required package: carData

Anova(results, type="III")

## Anova Table (Type III tests)  
##   
## Response: TotalRevenue  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 8377668 1 409.3962 < 2.2e-16 \*\*\*  
## Region 918999 5 8.9818 3.676e-07 \*\*\*  
## Product 512927 1 25.0655 2.170e-06 \*\*\*  
## Region:Product 1767149 5 17.2712 1.573e-12 \*\*\*  
## Residuals 2210055 108   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

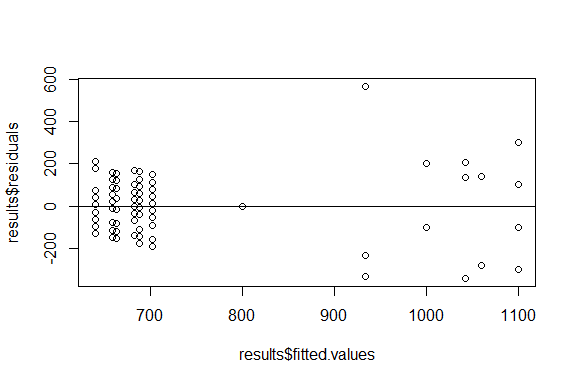
# omitted for printing  
# TukeyHSD(results)  
  
# Assumptions  
  
# Assumption 1 - the residuals are normal  
shapiro.test(results$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: results$residuals  
## W = 0.9671, p-value = 0.004912

# Assumption 2 - equal variance  
leveneTest(TotalRevenue~factor(Product)\*factor(Region))

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 11 1.9657 0.03882 \*  
## 108   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Assumption 3 - constant variance of errors  
plot(results$residuals~results$fitted.values)  
abline(0,0)



# Assumption 1 - the residuals are normal

## Discussion of Hypothesis Tests

The hypotheses for testing the interaction are below. Our null hypothesis (H) is that the differences in mean response are not dependent on an interaction between region and product. Our alternative hypothesis (H) is that the differences in mean response are dependent on an interaction between region and product. We made sure to use a Type III two-way anova test since our categorical variables were unbalanced.

H: There is not an interaction between product and region on total revenue.

H: There is an interaction between product and region on total revenue.

We found a p-value of 1.573e-12 for the interaction, which is less than 0.05, so we reject H. Thus, the interaction term is included in the full model and the differences in mean response of total revenue are dependent on an interaction between region and product.

A post-hoc test is necessary because our p-value is significant, less than 0.05. We used the Tukey test as our post-hoc test. The Tukey test returns only p-values greater than 0.05 for the main effect of Region, which means the pairwise means of total revenue based on region are not different. The Tukey test also returns strictly p-values greater than 0.05 for the main effect of Product, which indicates the means of total revenue based on region are not different.

The Tukey test returns some p-values less than 0.05 so some significant p-values. This means some of the pairwise means of total revenue based on region and product are different. For example, Connecticut product 1 and Alabama Product 1 have a p-value of 0.0065608, which is less than 0.05. This means that the pairwise means for the total revenue of product 1 sold in Connecticut and the total revenue of product 1 sold in Alabama are significantly different.

Looking at the assumptions:

Assumption 1 - Normality of the Residuals

Since the Shapiro test run on the residuals yields a p-value of 0.004, there is evidence that this distribution is not normal.

This assumption is not met.

Assumption 2 - Equal Variance

Then we used Levene’s test to see if the variance of total revenue across every pairwise group was equal. Since the test yielded a p-value of 0.03, this means at the 5% level, we have evidence to suggest that the variances of each group may not be equal.

Therefore this assumption is also not met.

Assumption 3 - errors have constant variance

Looking at our plot of residuals vs. fitted values, it is quite clear that there is not constant variance. There does appear to be in the left side of the plot, but the right side of the plot is vastly different. This condition is not met.

The assumptions of random assignment of subjects to treatment and independence of subjects and groups is met by the nature of the problem.

Since most assumptions were not met for this two-way anova test, we would not suggest using this model or making any conclusions from it.

# Question 2 (20 pts):

Does the total revenue significantly differ for the different teams? Include a discussion that includes exploratory analysis, hypothesis testing, checking all of the assumptions for the test and running post-hoc tests, if needed. You will be graded on the completeness, and correctness, of your analysis.

##Question 2 R Code

# This is a one-way anova test  
# unbalanced does not matter for one-way anova  
  
tapply(TotalRevenue, Team, summary)

## $`Team A`  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 510.0 646.0 731.0 766.8 816.0 1400.0   
##   
## $`Team B`  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 510.0 637.5 748.0 738.2 816.0 1000.0   
##   
## $`Team C`  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 510.0 544.0 680.0 705.0 786.5 1200.0   
##   
## $`Team D`  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 510.0 544.0 646.0 729.8 816.0 1200.0   
##   
## $`Team E`  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 544.0 669.5 782.0 742.5 824.5 900.0   
##   
## $`Team F`  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 510.0 637.5 680.0 722.9 778.2 1200.0   
##   
## $`Team G`  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 510.0 671.5 707.0 725.9 748.0 1177.0   
##   
## $`Team H`  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 510.0 569.5 680.0 706.7 782.0 1250.0   
##   
## $`Team I`  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 510.0 544.0 612.0 717.2 790.5 1500.0   
##   
## $`Team J`  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 510.0 578.0 690.0 666.2 748.0 816.0

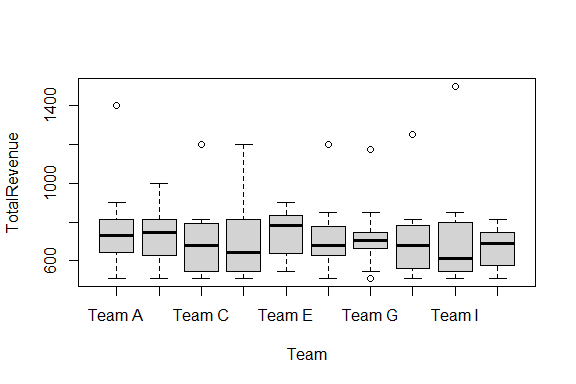
tapply(TotalRevenue, Team, sd)

## Team A Team B Team C Team D Team E Team F Team G Team H   
## 229.4665 132.5813 192.8457 241.0801 118.0759 179.8906 168.1776 200.4165   
## Team I Team J   
## 273.2185 106.4449

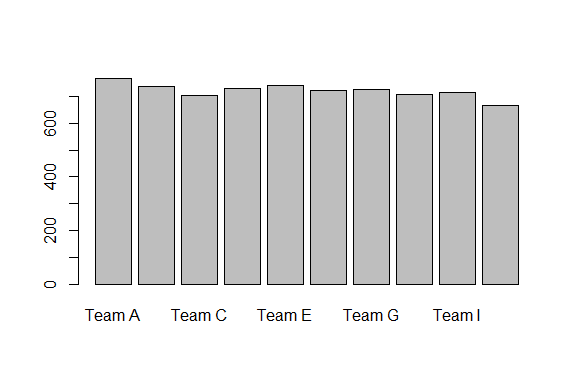
tapply(TotalRevenue, Team, length)

## Team A Team B Team C Team D Team E Team F Team G Team H Team I Team J   
## 12 12 12 12 12 12 12 12 12 12

boxplot(TotalRevenue ~ Team)



barplot(tapply(TotalRevenue, Team, mean))



# running the one-way anova test  
results = aov(TotalRevenue~Team)  
summary(results)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Team 9 77207 8579 0.234 0.989  
## Residuals 110 4029010 36627

# assumption 1 - Observations from each group are normally distributed  
tapply(TotalRevenue, Team, shapiro.test)

## $`Team A`  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.80528, p-value = 0.01076  
##   
##   
## $`Team B`  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.97214, p-value = 0.9319  
##   
##   
## $`Team C`  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.84758, p-value = 0.03428  
##   
##   
## $`Team D`  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.7867, p-value = 0.00664  
##   
##   
## $`Team E`  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.89611, p-value = 0.1413  
##   
##   
## $`Team F`  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.84858, p-value = 0.03527  
##   
##   
## $`Team G`  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.82035, p-value = 0.01611  
##   
##   
## $`Team H`  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.80501, p-value = 0.01069  
##   
##   
## $`Team I`  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.70335, p-value = 0.0009079  
##   
##   
## $`Team J`  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.93228, p-value = 0.4049

# assumption 2 - Homogeneity of Variances  
library(car)  
leveneTest(TotalRevenue, factor(Team))

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 9 0.4803 0.8852  
## 110

# assumption 3 - Random assignment of subjects to a treatment  
# assumption 4 - Subjects within groups are independent, and subjects in different groups are independent

## Question 2 Discussion and Analysis

Upon initial exploratory analysis, it does seem evident that there will be a significant relationship between team and total revenue.

Right away, we can see in the numerical summary that Team A has an average revenue of 766.8 while Team J has an average revenue of 666.2. The large difference in these means will likely prove to be significant in testing.

The pairwise boxplots also help show the significantly different average total revenue between each Team.

It seems likely that there will be a significant relationship between total revenue and team.

For this question, we utilized a one-way anova test since we only had one categorical variable which had more than 2 categories. Below are our hypotheses:

= Average Total Revenue for Team A.

= Average Total Revenue for Team B.

= Average Total Revenue for Team C.

= Average Total Revenue for Team D.

= Average Total Revenue for Team E.

= Average Total Revenue for Team F.

= Average Total Revenue for Team G.

= Average Total Revenue for Team H.

= Average Total Revenue for Team I.

= Average Total Revenue for Team J.

H: = = = = = = = = = ; The average total revenue for each Team is equal.

H: At least one Team has a different average total revenue compared to the others.

F = 0.235 df = 9, 110 p-value = 0.989

Since the one-way ANOVA test yields a p-value of 0.989, we do not reject H0.

At the 5% level, there is not evidence to suggest that any Team has a different average total revenue compared to the other Teams.

Since we do not have evidence of difference in the average total revenues, there is no need to run any post-hoc tests.

Assumption 1 - Observations from each group are normally distributed

The pairwise shaprio tests show that only 2 of the 10 teams have average total revenues that are normally distributed. This condition is not met.

Assumption 2 - Homogeneity of Variances

= Average Total Revenue for Team A.

= Average Total Revenue for Team B.

= Average Total Revenue for Team C.

= Average Total Revenue for Team D.

= Average Total Revenue for Team E.

= Average Total Revenue for Team F.

= Average Total Revenue for Team G.

= Average Total Revenue for Team H.

= Average Total Revenue for Team I.

= Average Total Revenue for Team J.

H: = = = = = = = = = ; The variance of total revenue for each Team is equal.

H: At least one Team has a different variance of total revenue compared to the others.

F = 0.48 df = 9, 110 p-value = 0.89

Since the Levene test yields a p-value of 0.89, we do not reject H0.

At the 5% level, there is not evidence to suggest that any Team has a different variance in total revenue compared to the other Teams.

Therefore, this condition is met.

Assumption 3 - Random assignment of subjects to a treatment

Assumption 4 - Subjects within groups are independent, and subjects in different groups are independent

Assumption 3 and 4 are both met by the nature of the problem. Each Team is randomly assigned and are independent from each other.

Based on one significantly failed assumption, we would recommend to use caution or transform the data if using a two-way anova test is necessary.