Module Answer Key

Abby Hustler

2/3/2023

library(readr)  
library(here)

library(mosaic)

library(dplyr)  
library(readr)  
library(here)  
library(mosaic)  
library(dplyr)  
library(tidyverse)

theme\_set(theme\_bw())  
data2022=read.csv("https://moneypuck.com/moneypuck/playerData/seasonSummary/2022/regular/teams.csv", header = FALSE)  
data2021=read.csv("https://moneypuck.com/moneypuck/playerData/seasonSummary/2021/regular/teams.csv")  
names(data2022)=names(data2021)

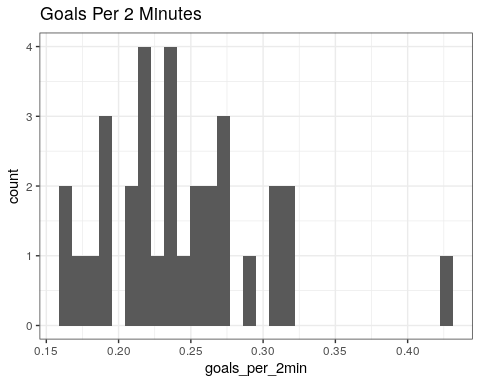
powerplay\_data<- data2022|> filter(situation=="5on4")|>  
 mutate(IceTime= iceTime/60)|>  
 mutate(SOG\_per\_2min= (shotsOnGoalFor/IceTime)\*2)|>  
 mutate(hits\_per\_2min= (hitsFor/IceTime)\*2)|>  
 mutate(play\_continued\_inzone\_per2min= (playContinuedInZoneFor/IceTime)\*2)|>  
 mutate(takeaways\_for\_per2min= (takeawaysFor/IceTime)\*2)|>  
 mutate(faceoffs\_won\_per2min= (faceOffsWonFor/IceTime)\*2)|>  
 mutate(goal\_differential= goalsFor-goalsAgainst)|>  
 mutate(goals\_per\_2min= (goalsFor/IceTime)\*2)|>  
 select(c(1, 108:115))

vec\_eastern <-c("BOS","TBL", "NYR", "TOR", "NJD", "CAR", "PIT",   
 "NYI", "WSH", "FLA", "OTT", "BUF", "DET", "PHI", "MTL","CBJ")  
  
powerplay\_data<- powerplay\_data|> mutate(League = if\_else(team %in% vec\_eastern,true = "Eastern", false = "Western"))

## 

1. Create a histogram that will show us the distribution of goals per 2 minutes.

ggplot(data = powerplay\_data, aes(x=goals\_per\_2min))+  
 geom\_histogram()+  
 labs(title = "Goals Per 2 Minutes")



1. Look over the data set. Give one example each of a categorical and quantitative variable.

Conference variable (east vs west) is a categorical while goals\_per\_2mins is a quantitative value.

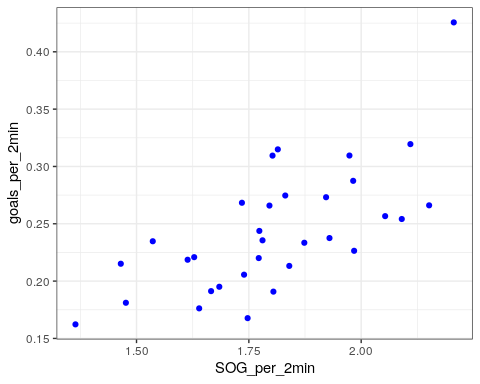
1. Find the mean for the goals\_per\_2\_min variable.

mean(powerplay\_data$goals\_per\_2min)

## [1] 0.2435952

1. Create a scatterplot that shows us the shots on goal compared to the goals per 2 minutes Does the graph you created appear to have a strong correlation? First make an initial guess of the correlation’s value. Now compute it using the cor function and compare it to your estimate?

ggplot(data= powerplay\_data, aes(y= goals\_per\_2min, x= SOG\_per\_2min))+  
 geom\_point(color="blue")



cor(powerplay\_data$goals\_per\_2min, powerplay\_data$SOG\_per\_2min)

## [1] 0.6853592

1. Create a hypothesis test to see if there is a correlation between goals per 2 minutes and shots on goal per 2 minutes. Find the T value which will help you make your conclusion. Ho: p=0 Ha: p≠0

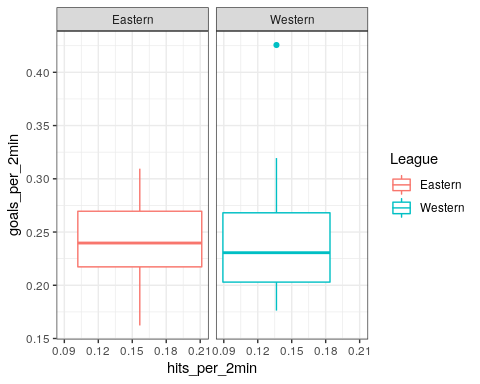
cor.test(powerplay\_data$goals\_per\_2min, powerplay\_data$SOG\_per\_2min)

##   
## Pearson's product-moment correlation  
##   
## data: x and y  
## t = 5.155, df = 30, p-value = 1.504e-05  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.4423871 0.8345999  
## sample estimates:  
## cor   
## 0.6853592

t = 4.5197 so pvalue is zero which means we can reject the null. We have significant evidence that the correlation is different from zero.

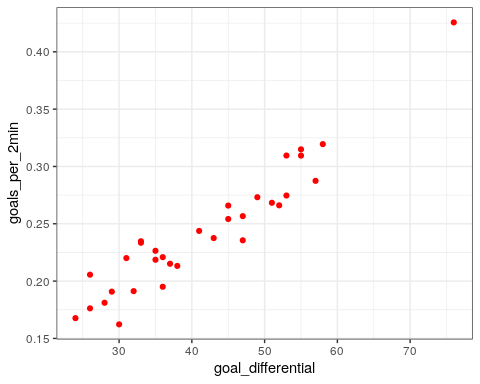
1. Create a boxplot that shows us goals\_per\_2min vs hits\_per\_2min for each conference (hint: add facet\_wrap(~League) to your ggplot to separate the east vs west). Once you have created the plot tell us about the differences between leagues. Are there any appearant outliers?

ggplot(data= powerplay\_data, aes(y= goals\_per\_2min, x= hits\_per\_2min, colour= League))+  
 geom\_boxplot()+  
 facet\_wrap(~League)



1. Create a scatterplot that will compare goals\_per\_2min with goal differential. Do you see any outliers? If so, do it follow the linear pattern of the other data points?

ggplot(data=powerplay\_data, aes(x=goal\_differential, y=goals\_per\_2min))+  
 geom\_point(colour="red")



1. Look at the output below. This is the model you can use to predict goals\_per\_2min. What is the best individual predictor out of the model? Is the predictor significant?

mod1= lm(goals\_per\_2min ~ SOG\_per\_2min + faceoffs\_won\_per2min + hits\_per\_2min, data = powerplay\_data)  
summary(mod1)  
## Call:  
## lm(formula = goals\_per\_2min ~ SOG\_per\_2min + faceoffs\_won\_per2min +   
## hits\_per\_2min, data = powerplay\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.056680 -0.030386 -0.007734 0.026543 0.100337   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.11358 0.07815 -1.453 0.157   
## SOG\_per\_2min 0.20729 0.04534 4.572 8.92e-05 \*\*\*  
## faceoffs\_won\_per2min -0.04387 0.06235 -0.704 0.488   
## hits\_per\_2min 0.23182 0.21507 1.078 0.290   
## ---  
## Residual standard error: 0.04079 on 28 degrees of freedom  
## Multiple R-squared: 0.4961, Adjusted R-squared: 0.4421   
## F-statistic: 9.188 on 3 and 28 DF, p-value: 0.0002147

The best predictor is SOG\_per\_2min considering its high tvalue which results in the lowest pvalue.

1. Based on the same output, does the entire model appear to be significant? If so, explain how you know this.

Yes the entire model is significant, and we know this by solely loooking at the pvalue at the bottom of the output.