Capstone Project Milestsone report

Sheila Lynch

October 7, 2018

## An introduction to the problem (based on your earlier Capstone submissions)

As I watch the Olympics this past year and listening the the athlete’s stories, most are young adults competing in the variety of sports. It raised the question, how many athletes are over 30 years old. And how many of those over 30 years old earn a medal. And while this information is good for trivia answers, I am validating as manufactures of these sports focus on specific age groups to market product and is there a market for athletes over 30.

## A deeper dive

Once I found a dataset with Olympic data from 1900 - 2016, detailing all Sports and Teams.

load(athlete\_events) Athlete\_event\_whole <- mutate(athlete\_events, Age = ifelse(is.na(Age), mean(Age,na.rm = TRUE), Age)) str(Athlete\_event\_whole)

In order to clean the data, I needed to check the structure and determine if any critical colunms had NA data. I found both Age and Medal columns did include NA or blanks. I first fixed the Medal missing data which consisted of Gold, Silver, Bronze and I filled in the blanks with “No medal earn”.  
-Clean up Medal fields with NA

Top\_6\_Teams <- mutate(athlete\_events, Medal = ifelse(is.na(Medal), “No Medal earned”, Medal)) Secondly to correct the NA in Age, I considered removing those with NA, but as I looked at the age data, I thought filling those in with the mean age avoided losing data rows. Top\_6\_Teams <- mutate(Top\_6\_Teams, Age = ifelse(is.na(Age), mean(Age,na.rm = TRUE), Age))

Next, as countries have divided over the years or come back together, I looked a the top 7 teams by participation and combined West Germany, East Germany and Germany as one team as well as took USSR and Russia data to combine in one team. I realize a few other countries have split off of USSR, but for data volume, it was not impactful.

Clean up major country changes of German and Russia to combine

Top\_6\_Teams <- mutate(Top\_6\_Teams, Team = ifelse(Team == “Soviet Union”, “Russia”, Team)) Top\_6\_Teams <- mutate(Top\_6\_Teams, Team = ifelse(Team == “East Germany”, “Germany”, Team)) Top\_6\_Teams <- mutate(Top\_6\_Teams, Team = ifelse(Team == “West Germany”, “Germany”, Team))

A few other steps of cleaning the data was to create a column to reflect if athlete was under 30(No) or over 30(Yes). This provided better table reviews than true / false - in preference. - create new column to simplify those over / under 30

Top\_7\_TeamAge >= 30,‘Yes’,‘No’)

And I also created a column to assign a value to the medals earned. While I honestly have not finalized if I just compare the change in age based on medals, I wanted this field to explore possibly weighting the level of medal. So that is why I gave Gold 3, Silver 2, Bronze 1 as there is consideration of giving more credit to gold medals earned than bronze. # -build value to medal

Top\_7\_TeamMedal == “Gold”, 3, ifelse(Top\_7\_Team$Medal == "Silver", 2, + ifelse(Top\_7\_Team$Medal == “Bronze”, 1,0)))

## Any preliminary exploration you’ve performed and your initial findings.

I ran some plots to see how the mean age has changed over time. This plot shows the mean age has slightly decreased {r, echo=FALSE} Athlete\_event\_whole <- mutate(athlete\_events, Age = ifelse(is.na(Age), mean(Age,na.rm = TRUE), Age)) ggplot(Athlete\_event\_whole, aes(x=Year, y=Age)) + geom\_point(shape=1) + # Use hollow circles geom\_smooth(method=lm) # Add linear regression line

This plot is reflecting the variance per total involved either over or under 30 years old. {r} ggplot(Top\_7\_Team, aes(x = Year, fill = Age\_over\_30)) + geom\_bar(position = “dodge”)

Then I started to think what possible articles have already been written on this topic. So I went to Google and found 3 articles that help to provide more direction.

The first article discusses Elite Athletes over 40 for the Pyeongchange Winter Olympics: <https://www.bloomberg.com/news/articles/2018-02-14/wisdom-and-drive-older-olympians-becoming-the-norm> . The second article discusses is 30 the new 20 for Olympic Athletes: <https://venngage.com/blog/olympics/> . Both of these articles discuss athletes that were successful and the latter article does analysis comparing Gymnastics to Track & Field. This analysis combined with the third article I read <https://www.verywellfit.com/olympic-sports-youre-not-too-old-for-4075439> which discusses 10 Summer Olmpic Sports Your not too old to quality for. As the 2020 Summer Olympic games are upcoming and the dataset data runs through 2016 Summer Olmpics, I thought I will select 2 of the sports on the list and compare if the average age is actually increasing. Or if they have always been sports that have above typical Olympic athlete mean age.

I selected **Rowing** Rowing

And **Equestrianism** [alt text]<http://images.nbcolympics.com/www.nbcolympics.com/field_image/19August2016/usatsi_9486292-1_0.jpg>

Both sports have a long history in the Olmpics and moderate participation.

The mean age of all sports over time is 25.5569 while the mean age of just Rowing and Equestianism is 28.59793. Now the question is if the mean age is higher will that translate to having more athletes over 30 and is it increasing. this plot shows for both sports it is increasing. {r} tapply(Compare\_Row\_Equest\_allYear, meanfun)

{r} ggplot(Compare\_Row\_Equest\_all, aes(x=Year, y=Age)) + geom\_point(shape=1) +  
geom\_smooth(method=lm)

## Based on these findings, what approach are you going to take?

Ultimately I would like to know if I am looking to earn a medal and over age 30 which sport should I choose. I am now going to compare these two sports, each of which have a higher mean age and discover if this mean has increased over the years. I need to run analysis on each sport and if the mean age has changed. As well as expore how it differs by gender. And does the higher mean age increase or decrease the probablity of those athletes over 30 earning a medal. Is there a Team that has a higher probability in these two sports.

## How has your approach changed from what you initially proposed, if applicable?

Intially I thought I would just review how many athletes over 35 earn a medal over the Olympic years. Then realizing how much data it involved making very messy plots, I decided to narrow it down and focus on two sports. It is still applicapble, just more specific.