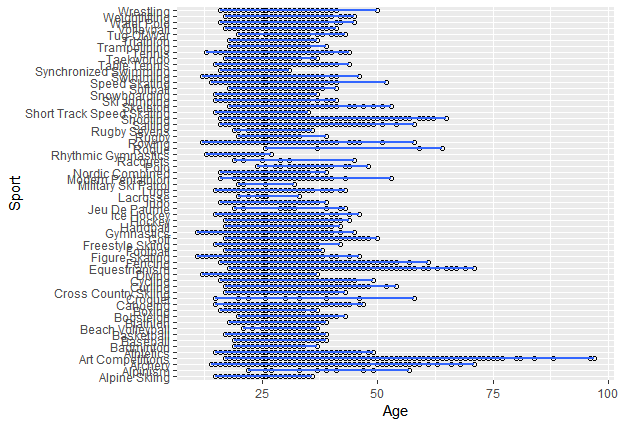
Project: Capstone summary on statistical analysis

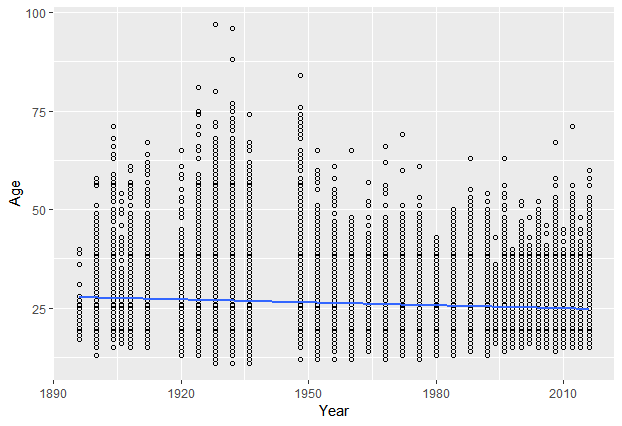
I began my analysis by realizing I needed to better clean my data. As started to run plots and review it was just too much information. The first thing after running the plots, I realized that having it be over 35 was too much so I have adjusted to analyzing over 30 years old. Also after talking with my mentor, I decided to narrow it down to the top 5 Teams based on participation. However, in reality I narrowed it down to the top 7 teams. I combined West Germany, East Germany all into Germany and combined Soviet Union with Russia as both of these countries had changes in team names in the last 30 years. Lastly I did include Japan just to provide a reference on the far east country as otherwise the data is very influenced by Europe and United States.

In my data cleaning I also added two columns, the first to have a simple yes or no to being over 30. The second is to associate a number value to the medal earned, gold =3, silver = 2, bronze =1, no medal = 0. My hope is that while I continue to do statistical analysis on this, I can put some weight to the medal earned.

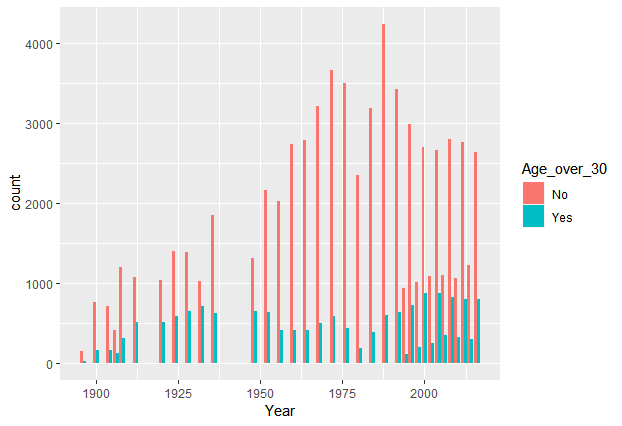
Looking at various plot types, the scatterplots do not provide as much information as the data is very fixed. I mean that Olympics are held every 4 years so the years are very focused. So the scatterplots are visual lines not random.



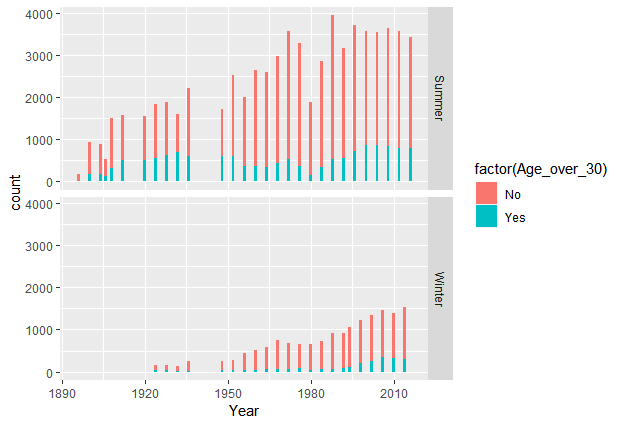
I started to look at by Year and Age, then blended into looking at Sports, Teams, Medals, Sex. Sports are still very broad in scope. Some of the information did provide some insight. Below you can see a slight decline over the years in age of participants.



Even as I looked into the time series plots, I needed to leave the scatterplot and focus on bar charts and histograms. Bar charts showed some interesting information.



And into histograms looking into it based on season showing more variation in summer than winter Olympics which when pulling out summer shows it has waivered over the years.



Based on this review, I now want to focus on the probability of an athlete over 30 earning a medal. And in this I want to consider if any of the top 7 countries has a higher probability or if any sport has a better probability. As I began to review I found that the participation of athletes of those 7 countries has decrease around 1%.(19.6 vs 18.6) if I compare entire years vs since 1980. Plus if I compare medal earnings of those over 30, it decreased 0.02(0.4 vs 0.038). And finally if I compare the Sports of all sports vs a selection of 12 sports, the medals earned by those over 30 have decreased less than 0.005%(0.047 vs 0.042).

I find that these are marginal so going forward I will just consider the years starting with 1980 to 2016 and narrowing the review of sports to the 12 selected.