James Bloor
Boston University Metropolitan College
MET CS 544
Final Project
Summer 1 2021
BU ID: U45448564

# Top 10 Forbes Highest Paid Athletes 1990-2020

## Dataset Details

Forbes highest paid athletes include earning from their on the field and off the field earnings The data set that we will look at below is from 1990-2020 of the top 10 highest paid athletes of each year. You will see that there is not any data for the year of 2001, this is because Forbes altered their reporting period of the 2002 list from June to June. This dataset was originally from the website topendsports.com.

## Organizing Data

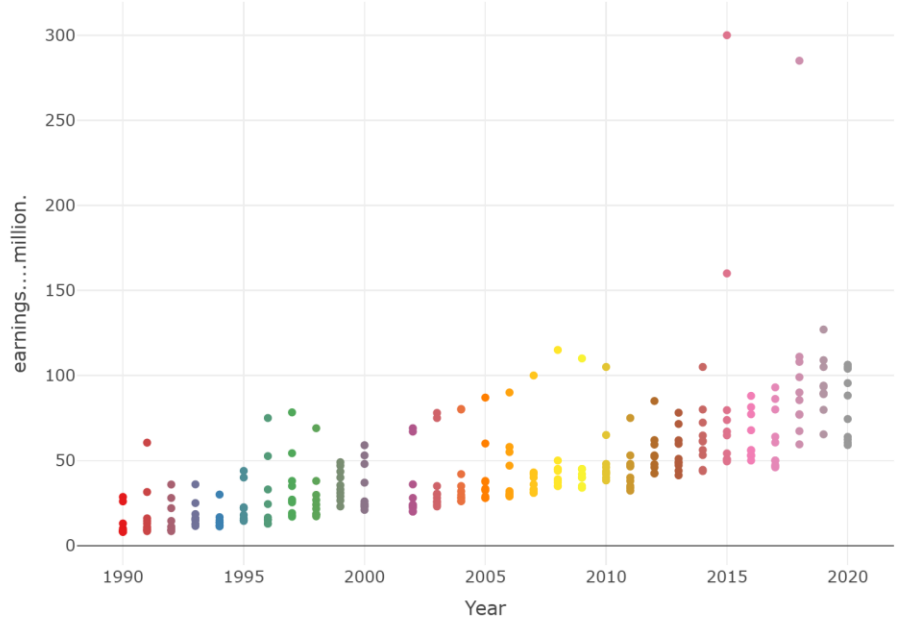
After working with my data set I soon realized that the sport column was becoming a hindrance on my analysis. Looking at the raw data many sports were given different names, i.e. Basketball players could be noted at “NBA”, “Basketball”, and “basketball”. The most names I had to correct were in the Auto Racing group. I joined many of them together to have a more concise data set that would continue to shed light on what I was curious on analyzing in this data.

## Objective

The goal of this analysis is to see how the highest paid athletes in the world are evolving. How diverse are these athletes? What trends can we observe? What athletes stand out above other ones when making the top ten list. I chose this data set because I wondered if athletes in salary capped sports appeared less often then athletes that do not a cap on how much they can earn from their sport team.

## Year over Year Analysis

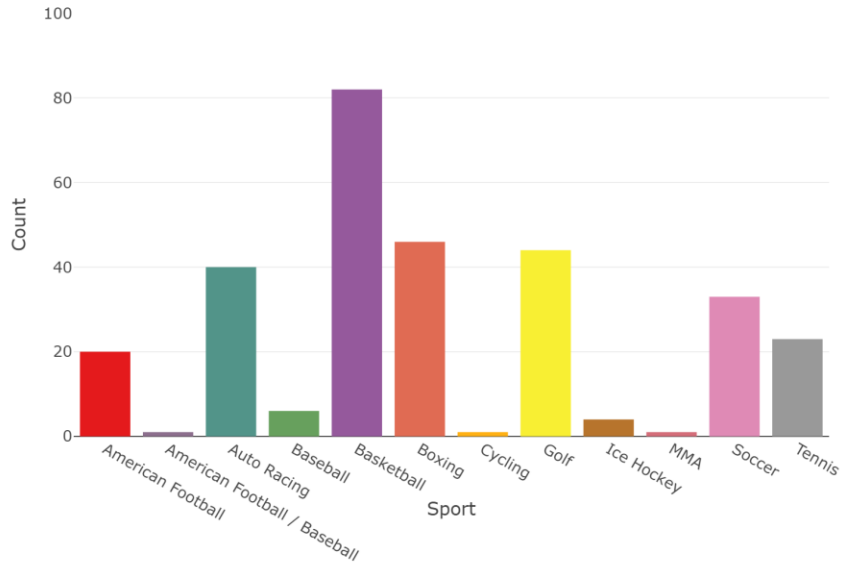
Lets first look at a scatter plot of the top 10 highest paid athletes for our entire data set and see what we will be working with for the remainder of my project. Along the x axis are each year, from 1990-2020, and the y axis shows us the amount of many made by each athlete for a given year in millions.




A trend we can see immediately is the steady increase each year of dots joined in together. As the last three decades have gone along, we have seen sports and athletes become more and more popular in the public eyes. Michael Jordan dominated the 1990’s with the six straight NBA titles. The 2000’s were Tiger’s, Woods blazed a path of excellence in a sport that had never seen anyone with his stature and excellence. In today’s word, Lebron James not only dominates on the Laker floor he plays on but has built his own enterprise to create movies, shows, documentaries, and so much more. As these players become more popular, more and more doors begin to open. With this comes sponsorship, record breaking signings with teams and clubs, and that goes without mentioning what Floyd Mayweather has been able to accomplish with his perfect record (more on him later and how he fairs with his companions on this list).

# Count of Athlete by Sport

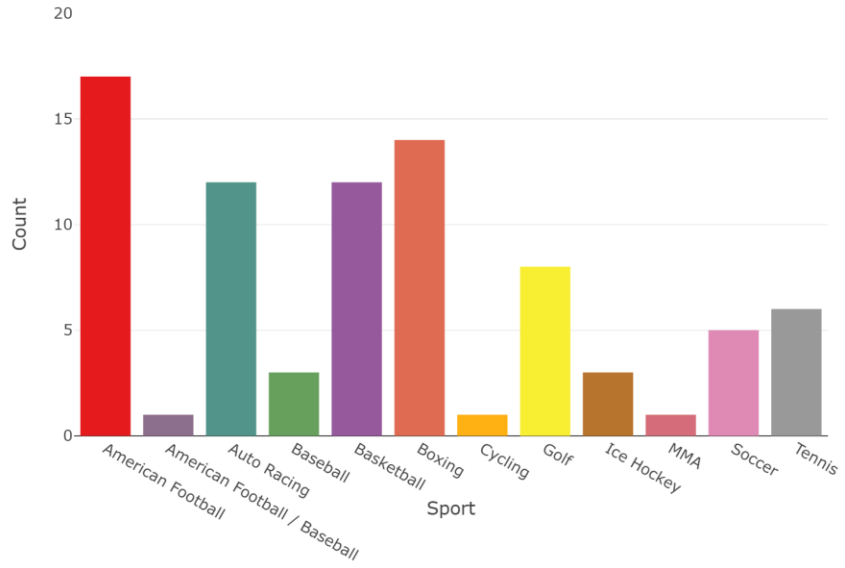
What sport has the most representation in this data? In the plot above, we see who has placed where on a given year in earning but what sport appears the most over the past 30 years?




Above, you can see that Basketball dominates this list, almost double the count of any other sport. Boxing and golf are 2nd and third respectively. This still does not inform us how many unique athletes have appeared in this top 10. Many of these athletes may appear year in and out so looking at a unique list will help is determine the turn over that we see from sport to sport in this list.

# Unique Count of Athlete by Sport

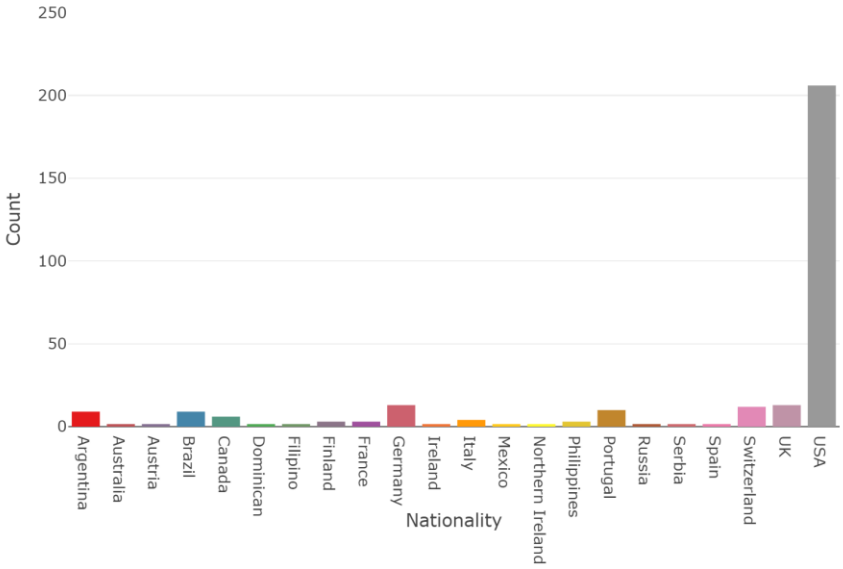
How many unique athletes represent their given sport? In the graph above, many athletes cound have been double counted in representing their sport. Now we can make every athlete count only once and see what sport has the most unique athletes representing them from 1990-2020.




Looking at this unique count, we can see that the American Football has one of the largest turn overs, with only 18 players representing them on this unique graph but a count of 20 on the previous graph. Out of the 82 spots that Basketball has taken in the past three decades, only 12 athletes have stood there. This shows us that player like Jordan and James have appeared on this list year over year and kept their names in the race for the highest paid athlete for decades at a time.

# Count of Athlete by Country

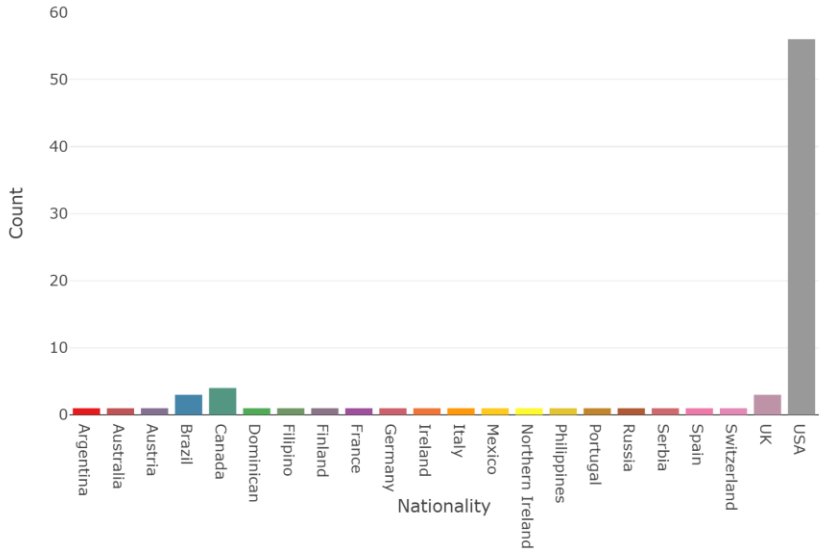
What nationalities are represented in our data? A major interest I had when first looking into this data was the ability to see how diverse these athletes were and which counties they represented. As I mentioned above, sports is a huge enterprise in the States but how do athletes fair overseas and how often to they make the Forbes list of top 10 paid athletes of a given year?




Above we can see the strangle hold that the USA has over every other country that appears on this list. Germany and the United Kingdom are ties for 2nd. Lets do the same exercise that we did above with unique athletes per sport here as well.

# Unique Athletes by Country

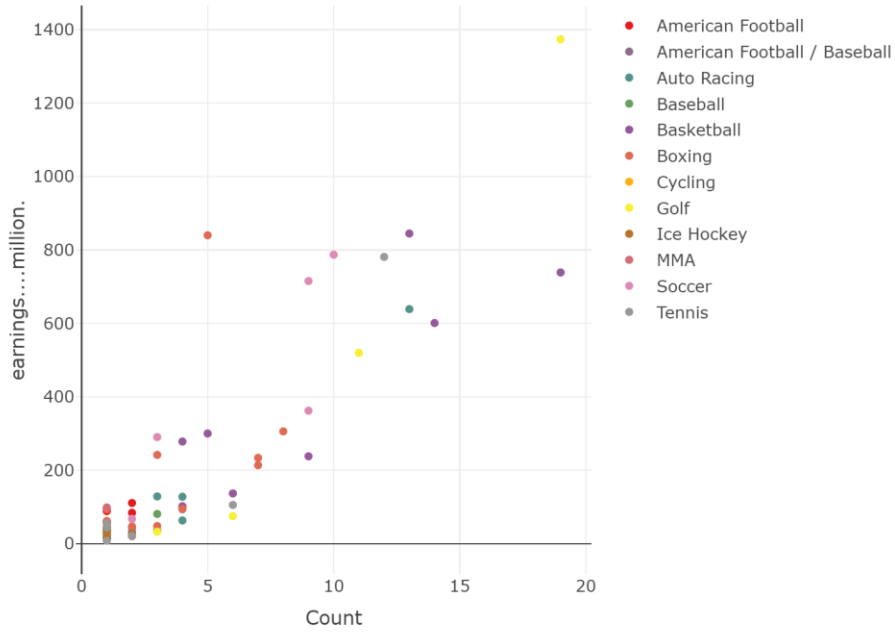
Now we are looking at unique athletes that have appeared on the Forbes list over the past 3 decades. There is no doubt that the USA will still have a strong hold on the lead but lets see what else we can recognize from the data we will see in this next bar graph.




While American Football, Basketball, and Baseball are huge American Sports, Soccer and Tennis are dominated by non-American athletes nowadays. This is where we are seeing countries like Portugal, Argentina, Switzerland show up on our country list because they have a one player that is a dominate force in a specific sport. Another country that stands out to me here is Germany, from a count of 13 on the previous bar graph, it now only shows a count of one on this unique athletes graph. The German that has 13 sport on our data set is Michael Schumacher, and F1 driver that dominated his sport in the 2000's.

# Athlete Earnings Over Time

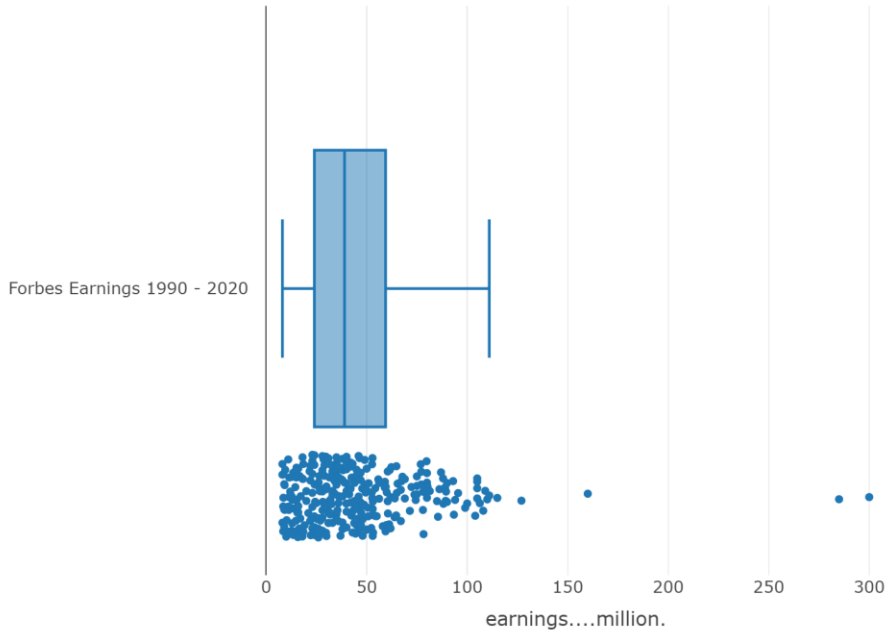
After seeing the different sports and nationalities represented, I wanted to see what put athletes on this list, their earnings. What athlete, that appears on the top 10 of Forbes Highest Earnings has earned the most from 1990-2020 and compare this total to the number of times that an athlete has appeared on the top 10 of this Forbes lists? Below we will use the x axis to show how many times an athlete has appeared in our data set and the y axis as total payout of all these top 10 finishes.




At first look we can see that there are many athletes that never topped the 200 million dollars mark with less than five appearances in this data set. In the middle of the graph we see our usual suspects that are global super stars, Messi, Ronaldo, James, Jordan. There is one athlete that is in the top right with over a billion in earning and 19 appearances, Tiger Woods. Compared to the Basketball players, and other salary capped sports that are in this list, it is not too much of a surprise to see Tiger in a league of his own. The only other athlete I think has a chance of catching Tiger is James. With 13 appearances and 844 million dollars made thus far, I can see James using his Hollywood portfolio, blockbuster movies, HBO projects, and his manager agency, in the next half decade to excel himself near Tiger. The other athlete that stands out to me on this scatter plot is Mayweather, with five appearances and 840 million dollars, if he was a younger man, I do not think it would be a question that he would dominate this list.

# Athlete Earnings Per Year

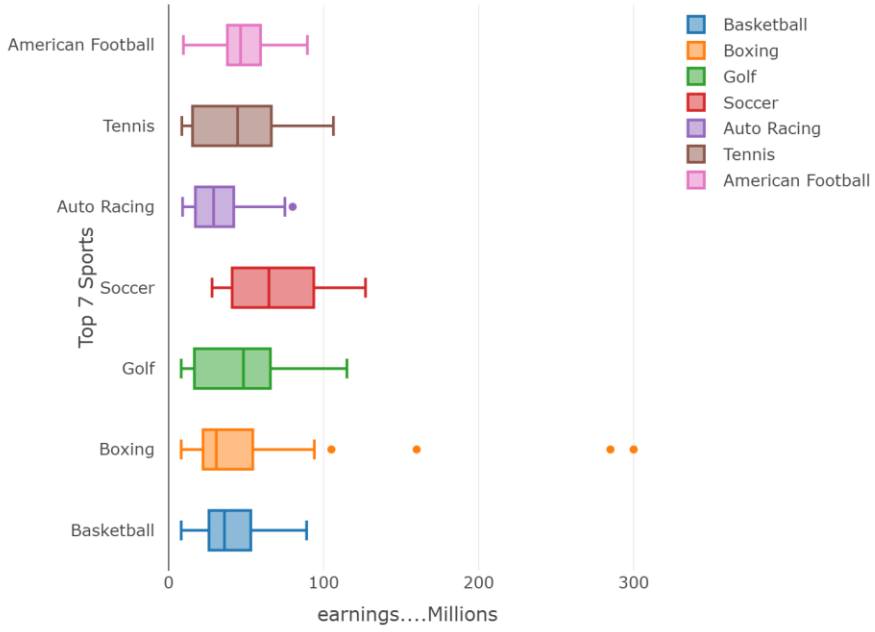
Now let us begin to break down the distribution of the data set. From looking at our first scatter plot, it looks like athletes are making more and more money each year; however, I believe that by looking at a box plot we can see if more athletes in recent years are stretching the earning possibilities.




As we can see in the plot, there is a very concentrated amount of earnings in between the lower and upper fence, only five instances that are above the upper fence, Tiger, Messi, Pacquiao, and Mayweather twice. With what we saw in the last scatter plot, Mayweather is not a surprise. He would have to be an outlier on this bar plot to have made as much money as he has in the past decade and has made half a billion more dollars than the next athlete with five appearances in our data. From the Median to the upper fence, we see a larger spread than we do from the lower fence to the median. As athletes are paid above the median, there is less of a precedent for this pay but it is becoming more frequent.

# Athlete Earnings by Sport

What would the above data look like if we filtered it down by sport? Let us now look at the same box plot above but split it out by sport. As we have seen in our data there are a few sports that do not have any data points. I will filter down the Forbes data to only hold sports with more than 10 data points. This will allow the box plot to have enough data to create a manageable box plot that we can make inferences on.




Now we can see how sports compare against each other and how their earning spread is across our data. Out of our data we have seven sports, boxing of course show us Mayweather's high totals but also has the lowest median out of all the sports we are looking at. This tells us that while Mayweather makes a ton of money in Boxing, others have not done the same and stay closer to the Median of our overall data. Soccer does have the highest upper fence and while at first this surprised me, I can infer that this is possible because Soccer is also an uncapped sport overseas. This allows athletes to make as much money as their teams want to pay them and keep the sporting world a free market. Capping a sporting team's spending on salary is silly to me but that is an argument for another time.

# Central Limit Theorem

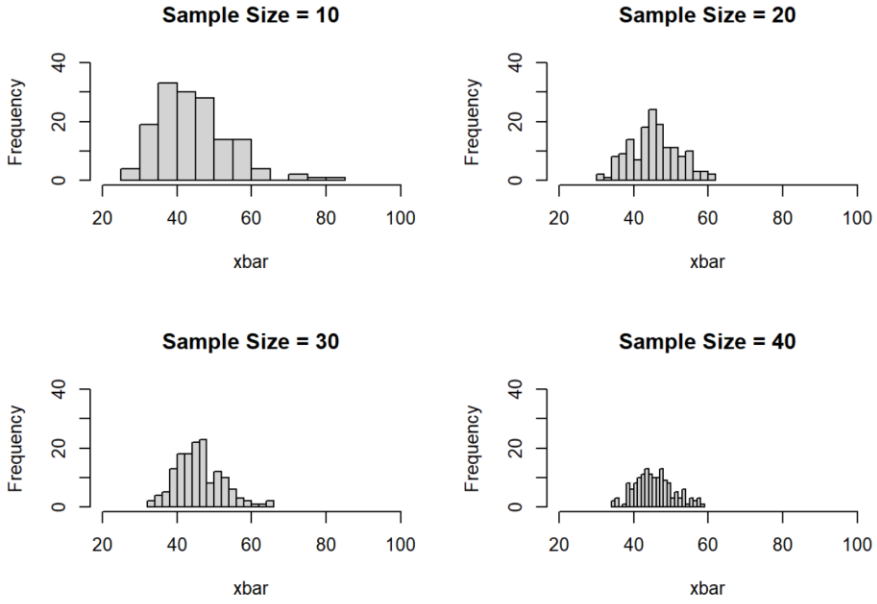
“The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement , then the distribution of the sample means will be approximately normally distributed.” Below you will see that I have taken 4 sample sizes to see what the distribution looks like in all cases. I have also placed the mean and standard deviation for each of these 4 samples that we take.

```
## Sample Size =  10  Mean =  44.38567  SD =  9.811974
```

```
## Sample Size =  20  Mean =  45.5379  SD =  6.435861
```

```
## Sample Size =  30  Mean =  45.88638  SD =  6.190601
```

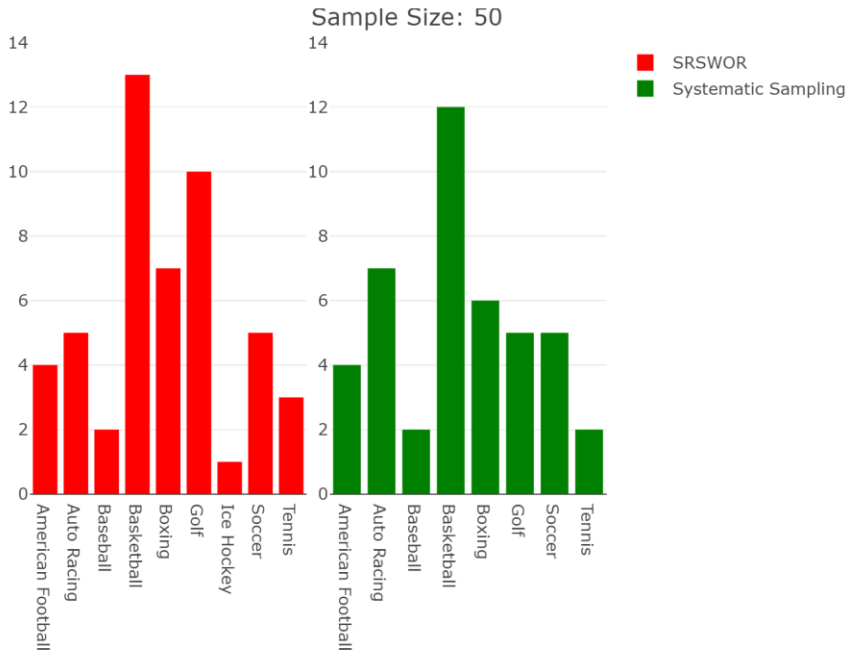
```
## Sample Size =  40  Mean =  45.65528  SD =  5.175468
```




Starting off by looking at the mean and standard deviation, while our mean stays with a range of less than one our standard divination is almost double from our sample size of 40 to 10. This just shows that as we use larger and larger samples, our data becomes tighter and tighter. We can also see our data is close to a normal distribution in each example here.

# Sampling

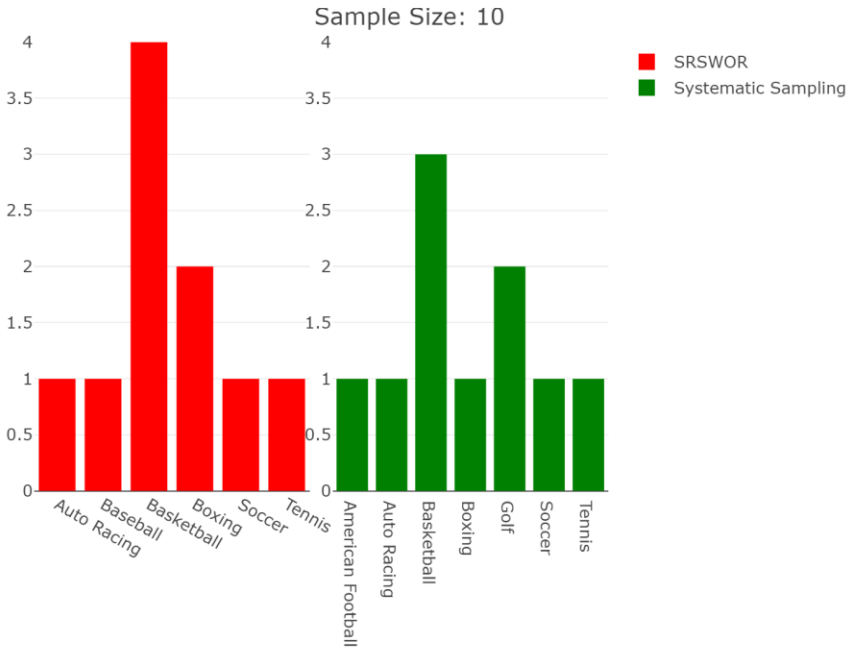
Systematic Sampling & Simple Random Sampling Without Replacement are the two sampling methods I used to analyze representation in data sets. I did not use stratified sampling using using proportional sizes based on the Sport variable because this would force me to filter out some sports that would have a less then 1% chance of making it into the sample. This means that I would have to filter out a sport that had less that 3 data points in my data and doing this would never allow those sports to be seen/samples from in stratifies sampling. Here I use a random sample of 50 athletes in my data set. From these samples I look at simple random sampling without replacement method (srswor) and a systematic method. Simple random sampling is where a specified amount of data points are selected at random from a set of data with every data point having an equal opportunity of being chosen. Systematic sampling is a type of sampling method where data points from a set of data are selected according to a random starting point but with a fixed interval. I have used both methods to see what sports would be selected if we took 50 athletes at ramdom and looked what sports they represented.




As we can see above, Basketball is chosen most in both examples. From there the only significant difference between these two examples is that systematic sampling has one Ice Hockey athlete in the count and simple random sampling without replacement does not.

# Sampling (Cont.)

As I did above, I now take the same steps in my sampling but with a sample size of 10. These examples below can be used a possible predictor on what a year of the top 10 earning athletes might look like.




Again, basketball is atop both counts with various other sports coming into this sample of 10 athletes. Looking at this data this small cound be considered a toss up and having a sample size like 50 would be a better representation on what a ramdom smaple can consist of.

# Conclusion

In conclusion, we have gone over various ways to look at this data and receive insight on where the highest paid athletes in the world come from. While the majority of the data shows athletes from the USA, there are a countries from around the world that have been represented by their most successful athletes. Basketball dominates this list because of names like James and Jordan that are/were on this list every year during their playing time. With basketball players appearing so much, I can infer a few possibilities. The first, even though these athletes are playing in a salary capped sport, they are still able to hold their own on this list because they take so much of their teams spending power. Another inference we can make is that these athletes in basketball have very large endorsement deals. Endorsement deals are sometimes not public so seeing how large these contracts that have with players is difficult but all of the top athletes are known as huge names for the brands that sponsor them. The last item I would like to touch on is the diversity of this list. While many sports are represented, there is not a single female athlete in this data. No female has ever finished in the top ten highest earning athletes in a year, the data set does not even include a variable that denotes the gender for these athletes because every single one is a male. With names like Naomi Osaka, Alex Morgan, Serena Williams dominating in their sports, I look forward to seeing female sports grow their enterprises and for the first female athlete to break the top 10.

# Bibliography

Central Limit Theorem. (2016). Https://Sphweb.Bumc.Bu.Edu/Otlt/Mph-Modules/Bs/Bs704_probability/BS704_Probability12.Html. https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability12.html

Forbes Highest Paid Athletes 1990–2020. (2020, December 19). Kaggle. https://www.kaggle.com/parulpandey/forbes-highest-paid-athletes-19902019

Sievert, C. (2020). Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC. https://plotly-r.com

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Yves Tillé and Alina Matei (2021). sampling: Survey Sampling. R package version 2.9. https://CRAN.R-project.org/package=sampling

G. Jay Kerns (2018). prob: Elementary Probability on Finite Sample Spaces. R package version 1.0-1. https://CRAN.R-project.org/package=prob