

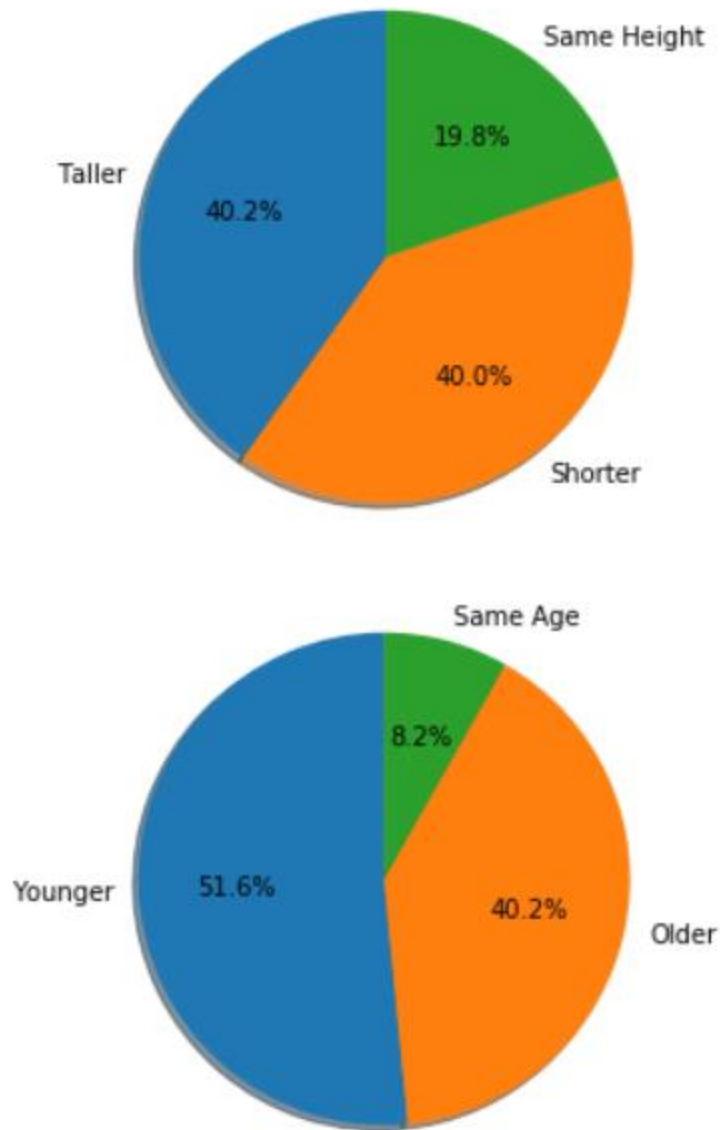
For my final project, I found a data set of UFC (mixed martial arts) match data on Kaggle. This is interesting to me because I watch this sport sometimes, and it seems extremely unpredictable, more so than another sport like baseball or football. In baseball or football or another popular sports the game is a set time interval and will end at a set time. Also, past performance is usually a good predictor of future performance in these sports since they are team sports. in mixed martial arts these types of matches can end at any time and very unpredictably. Past performance doesn't seem to matter since it's an individual sport. Also "mistakes" by one person can lead to them losing a match quickly if the other person capitalizes on it. So, I want to find out if *any* correlation exists between a fighter's performance or statistics and maybe compare it to future performance. To find this out I decided to calculate and graph the following things:

1. Determine which fighters have won the most by a certain method (judge's decision, KO, or submission).
2. Determine which fighters have the most loses by a certain method
3. Determine the stats around strikes landed information vs KO wins and successful grappling information vs submission wins.
4. Determine if younger age is an advantage to win
5. Determine if taller height is an advantage to win

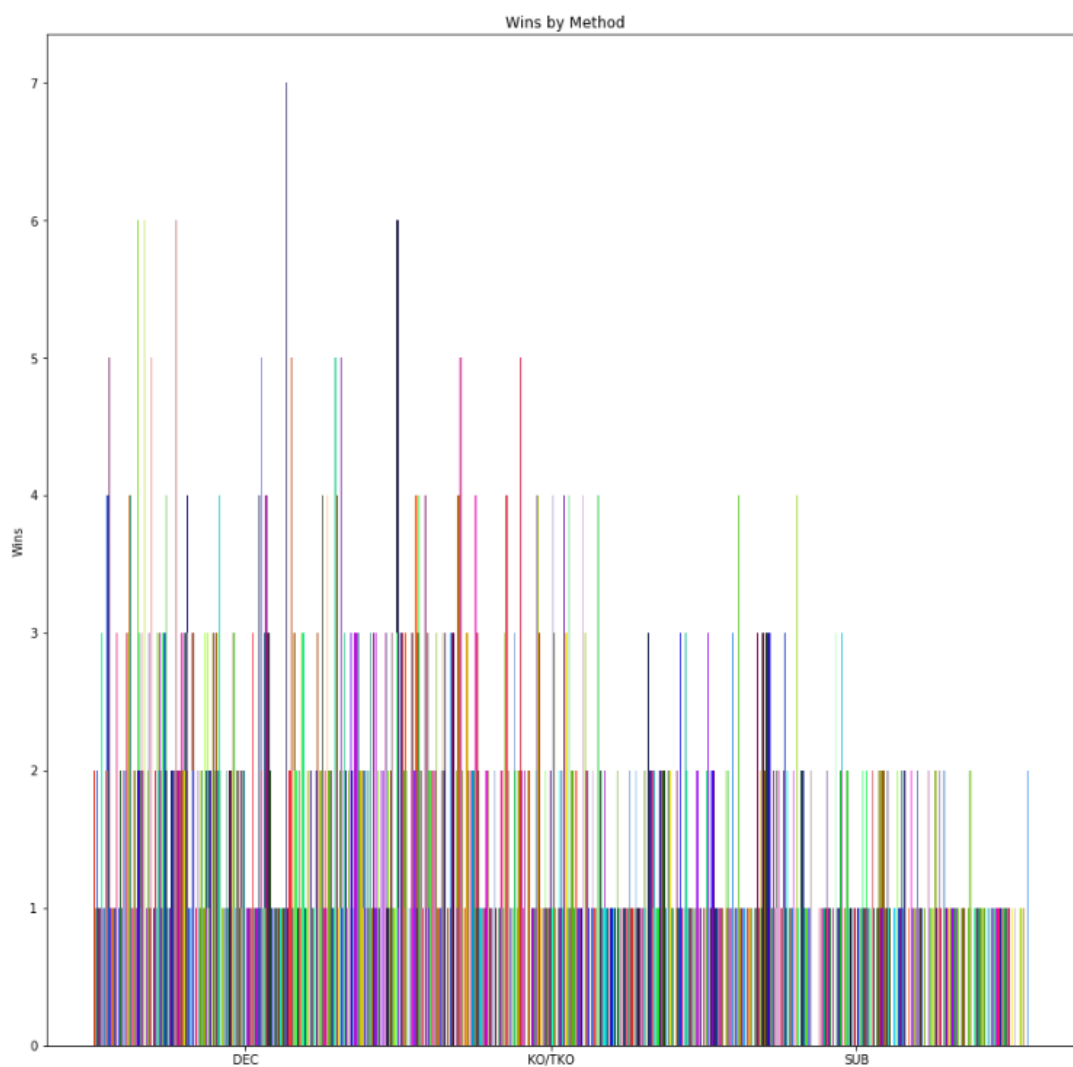
After doing this I would like to compare this data using a made-up scoring system and award points to two randomly selected fighters to hypothetically predict who would win in a matchup. This will not be a perfect system; however, it would be the best solution given the limited data and absence of machine learning (or my own statistics knowledge). **One other important note:** the data set doesn't include gender, so two randomly selected fighters may be male and female which would be unrealistic. It also doesn't consider weight classes. Fighters do move up and down in weight classes, so I decided not to worry about that.

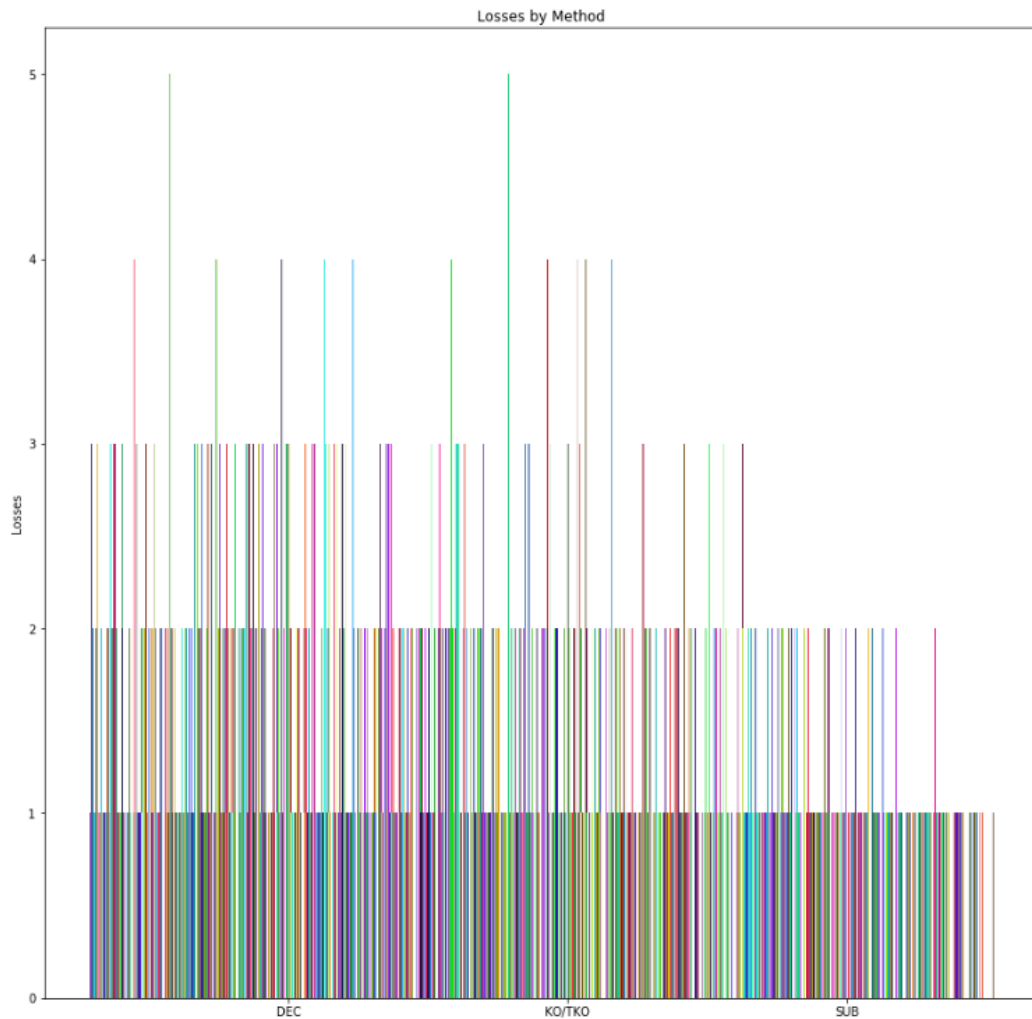
I created three separate classes: One for parsing, calculating and collecting the data, one for graphing the results, and one for comparing the data and predicting a winner. In the data calculation class, UFCDData, I created functions for getting the winners by number of wins of a type and a similar function for losses. These returned key value mappings of a fighter's name and a 3-item tuple of the records. Then I had functionality for getting height and age totals for winners who were taller, winners who were shorter and winners who were same height. I did the same for age with younger, older, and same age. Lastly for this class I created a very complex data structure that held all the winners of each two finish (aka not a judge's decision) type and tallied the strikes landed for ko's for each fighter, and grappling maneuvers landed for each fighter for submission wins. I did the same thing for the losers, so I could compare against the data for the winners.

In the graph class, UFCDDataCharts, I graphed the height and age totals in two separate pie charts. From the pie charts I could conclude that there was a miniscule advantage to being taller to win (40.2% vs 40%, and height made almost no difference. In the age category, younger age had more of advantage, but not by much (51.6% vs 40.2%):

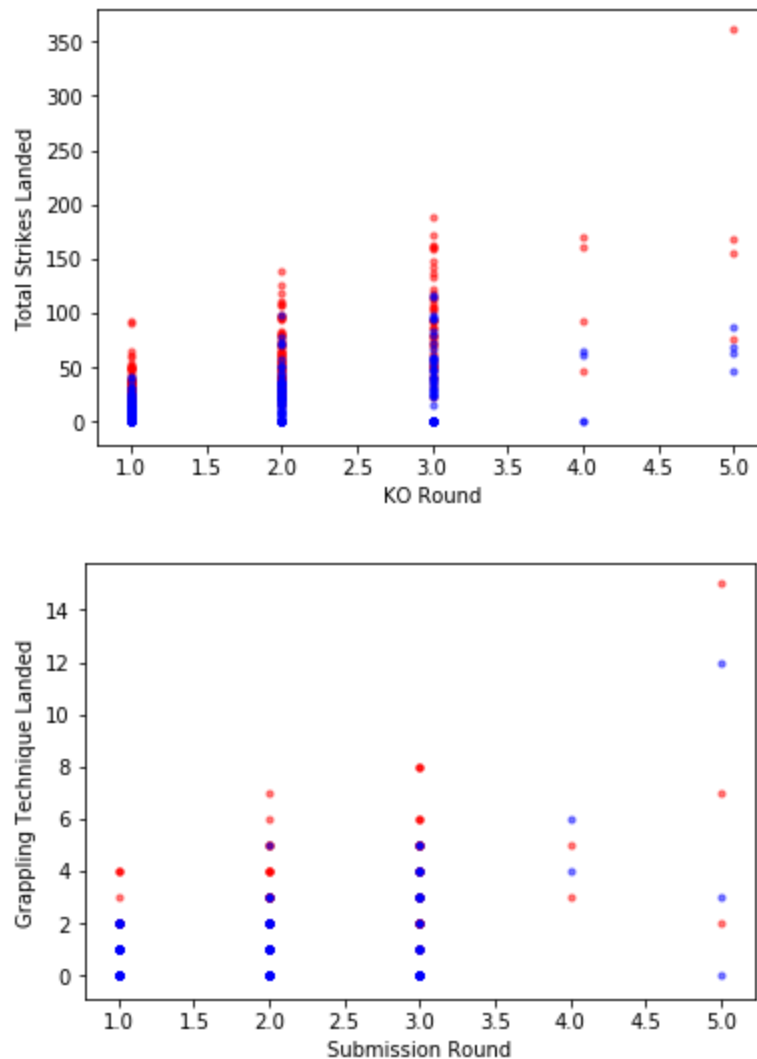


For wins by type and losses by type I plotted two bar charts. Each bar chart was grouped into three categories for the win type, and then each winner for winners and loser for losers was graphed. Unfortunately, this graph is admittedly very hard to read. The legend is huge and uses a lot of different colors. The bar charts are small and hard to compare. The data in this case matters more than the graph since it will weigh heavily into the prediction, but it could be modified to show individual charts for each win type or top 10 in each win type. Graphs with legend removed:





The last graph I plotted was for landed techniques vs a win by that type. This means I calculated landed grappling techniques for a submission win and landed strikes for a ko win. I also did the same for losses. I decided to use a scatter plot for this. I made two separate scatter plots, one for strikes/ko and one for grappling/submission. **Red dots represent the winners, and the blue dots represent the losers.** The scatter plot might seem like an odd choice since the x axis only had two tick intervals but I think it worked well, and gave a good representation of what I was looking for: in the case of ko victories there was a clear correlation between winning and more strikes landed, and in the case of submission victories more grappling techniques landed mattered, but not as much.



The UFCDataPredict gets initialized with all the collected and calculated data. There is a function for randomly grabbing two of the fighters from the data set. These randomly selected entries are passed to the prediction method. In this method I use all the data to compare the two fighters and award points which will be tallied and determine a winner. I created a get stat method as a helper to get data from rows in the NumPy arrays. For taller height I awarded 1 point since that made little difference. For younger age I awarded 3 points. For the rest of the points I first created a ‘multiplier’ where I took the data from landed techniques, but only the winners, and tallied every time a fighter’s name appeared there. Then I iterated through all the win by method data and tallied those numbers, subtracted by the loss by method data and then multiplied the result by each respective multiplier. I added both numbers onto the existing tally to get the fighters final score and compared the two and printed the result.

To improve upon this project, I would break out the bar charts into multiple charts with only the top ten fighters listed. I would also try to make some changes about how the randomly selected fighters were selected to make a more realistic matchup to what you might see in real life. I have no idea where the Kaggle user obtained this data from, but as someone who has watched the sport I would like to factor in other methods. I would possibly need to scrape Wikipedia or another source for this information.