SI 618 Project Report II

1. Motivation

The nature of my project is to look at the factors that can affect overall strength and how these factors relate and differ from one another. As a gym fanatic and former powerlifting fanatic, I was curious about strength factors and found a dataset that involves some of the world's finest powerlifters. Using this data allowed me to get a large enough sample size to run regression models, correlations, PCA, and clustering analysis on variables I believe are related to strength. My first question about the dataset was whether or not the weight class distribution amongst these powerlifters were fair for overall standardized judging. Weight classes can differ by over ten pounds at times, so I was curious to see if there were some people at the cusp of a weight class that should in fact be move up in order to make competition fairer for others. My second question was to see whether or not the use of equipment and certain strength variables can accurately represent my dataset. I also did the same assessment for testing for anabolic steroids and strength variables. My last question was to see how strength is related to variables (like equipment usage, steroid testing, body weight and age) and how they differ between sex.

2. Data Sources

Open Powerlifting Dataset: https://openpowerlifting.org/data

The Open Powerlifting Dataset was easy to handle and download. All I had to do was download the zip file from the website provided and import the csv into my markdown file. Luckily the dataset was kept clean and formatted well so I did not have to do too much data manipulation in order to get what I wanted out of the dataset. The dataset has over 1.2 million entries and dates from 1967 all the way to 2018. It contains variables such as bodyweight, mean deadlift, mean bench, mean squat, total kg lifted, Wilks score, sex, equipment, testing for steroids or not, age, and many more. Most of the data is formatted numerically with few categorical variables like weight class, age class, sex, testing, equipment type, and federation for the contest where the data was retrieved. The data also contained variable classifications on a github repository which can be found here: https://github.com/rfordatascience/tidytuesday/tree/master/data/2019/2019-10-08

3. Methods

To begin my analyses, I had to properly prepare my data. I knew that I wanted to work with some categorical variables in my analysis, so I converted some of the values into numerical variables. For example, I looked through the equipment data and saw that there four different types of equipment levels used. With some initial research on equipment and overall effectiveness, I placed equipment levels in a hierarchy using a function where the most helpful equipment of wraps was given a value of 3. The second most useful equipment of double-ply was given a value of 2 and the third most useful equipment of single-ply was given a value of 1. No equipment which was originally labeled as "raw" was given a value of 0. Also, I used a binary conversion for whether or not the data is from a contest that tested for anabolic steroids or not. If it was tested, then it received a value of 1 and if it was not tested then it received a value of 0.

With over a million entries in my dataset, I quickly saw that some of the entries were missing data. I also saw that there were columns that were unnecessary for my analysis like "best4BenchKg", "best4squatKg", "best4deadliftKg", "birthyearclass", "country" and "tested". I only removed tested after I made a new column that converted it into a numeric value. I also removed all data that had any empty cells in any of the remaining columns because I felt that it would cause errors in my analyses. After this initial clean of my data, my dataset went from over a million entries to about 156,000 entries.

Analysis I: For my first question of whether or not some people did truly belong in their weight class for competition, I ran into a big problem after my first test run. I found that there were too many weight class classifications for the different powerlifting federations. To combat this, I looked for the federation with the most data and found that to be USAPL. I then filtered my data for this analysis to only use the USAPL data to correctly bin each entry in its respective weight class by USAPL standards. Also, I had to change the weight class from a categorical classification to numeric. In order to do that I used a regular expression to get rid of any data that had + tacked to the end of them. This was appropriate for my analysis because I knew it would not have an effect of wrong classification on my entries. For example, 120+ became 120 and it did not have effect on my analysis because each entry was uniform in entry for classification and therefore someone being 130 kg would still be marked as 120+ rather than 150+.

Analysis II: For my second question of looking to see how related strength factors were to equipment usage and steroid testing, I had to determine which

variables to use to assess strength. This was a challenge because there is no true measure of strength that can accurately capture all elements of strength because they are all inter-related. I ultimately decided to use average bench, average squat, and average deadlift scores from the dataset because the average top score in each lift is typically what is used to measure a Wilks score, Glossenbrenner score and McCulloch score. I also had to normalize my chosen variables by applying a function that divided the mean of that column by the standard deviation of that column. This was in order to properly scale my data and find a correct variance in my PCA analysis. For this analysis, I did not remove data other than the initial clean I conducted. I used all data entries from all federations and allowed entries with the same person because each entry was from a different contest and resulted in variable values. Therefore, the data was treated as if it were a new person.

Analysis III: For my third question of seeing how age and body weight are related to strength and how it differs between sex, I had to revert my dataset to the initial clean. After analysis II, my strength variables were normalized so I reloaded the dataset and applied my functions to remove noisy entry and create my numerical values for certain factors. I did not run into any more trouble for this analysis in my data preparation after this step.

4. Analysis and Results

Analysis I: For my cluster analysis to find outliers in weight classes for individuals based on their weight class and bodyweight, I performed a kmeans test. Because I did not know the exact number of centers was appropriate for my analysis, I kept altering the centers and checking the kmeans results for the total percentage effectiveness of my clusters. Once I maximized this number, I found the number of centers that optimizes my analysis to be around 15 centers.

```
Within cluster sum of squares by cluster:
 [1] 268377.770  49934.816 130108.416   4599.646   4668.852  50126.738 131745.122
 [8] 344768.258  29385.283  39557.165  17946.267  47523.695 267524.854   6741.187
[15]  24309.275
 (between_SS / total_SS =  97.9 %)
```

Figure 1. Percentage of between sum of squares by cluster to total sum of squares by cluster

I then created a distances value by grabbing the rowsum squared of my two variables subtracted by the centers. I then square rooted that function and ordered my outliers to be up to 20 in each cluster in descending order. I plotted the distribution of each variable with its outliers in different colors for distinction.

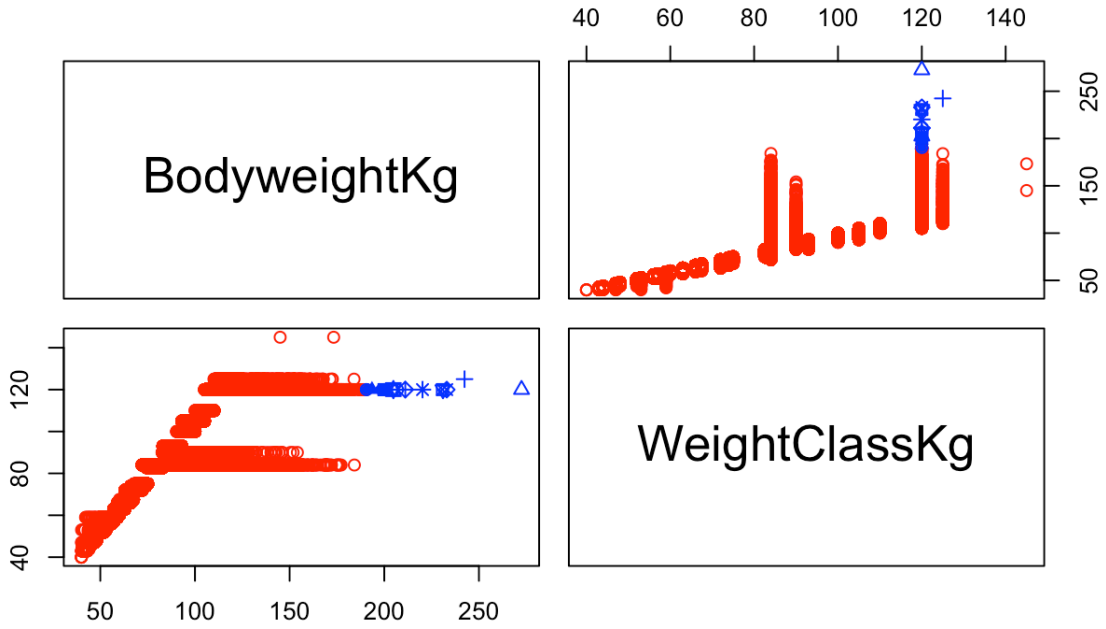

Figure 2. Cluster plot with outliers shown if present in clusters for body weight and weight class

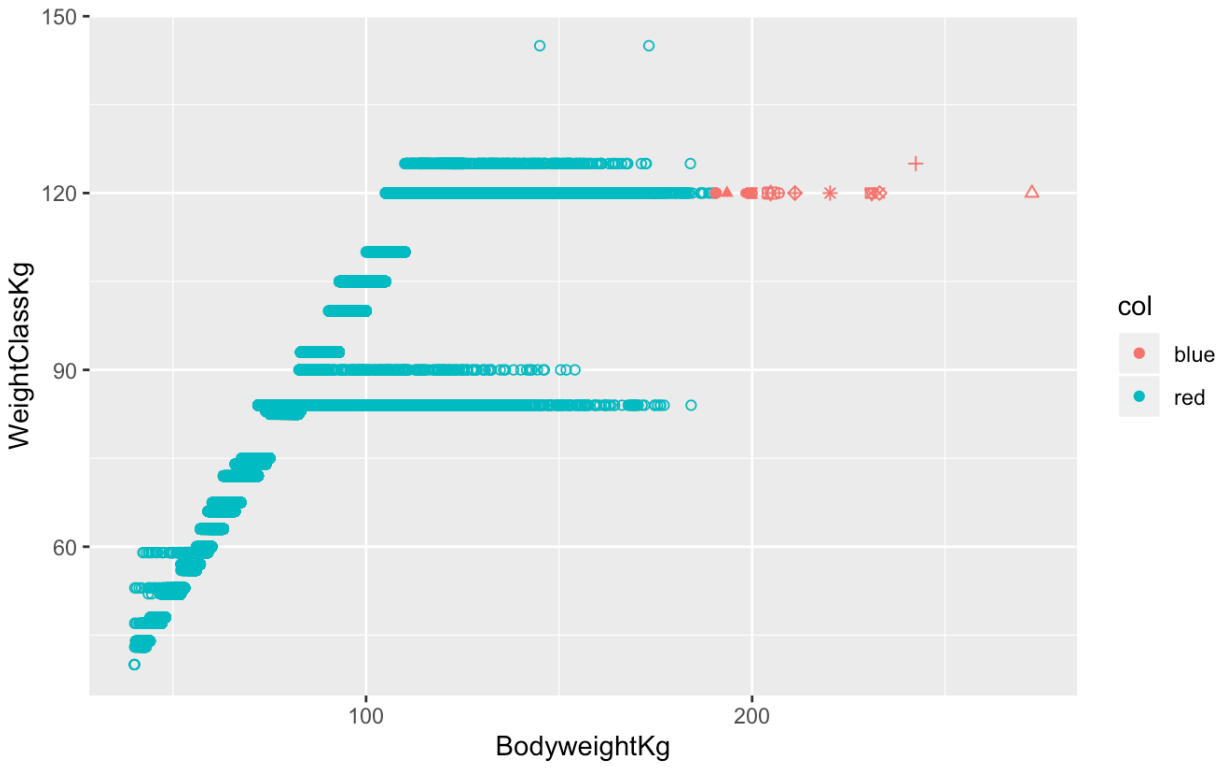

Figure 3. Cluster plot with outliers shown in clusters for body weight and weight class graphed against one another

Looking at the visualizations that this analysis presented, I found that weight classes for lower body weights are accurate in for competition typically because most of the competition are competing at bodyweights similar to each other. However, the higher the weight class and body weight, there tends to be more outliers that belong in a class above their current one especially looking at the 120+ weight class. This makes sense because like other sports, powerlifting loosely defines weight classes after certain weights and the bin classification for weights typically become larger. This will create a larger discrepancy in competitors in the same weight class causing some of them to outliers. A conclusion to make would be create more specific weight classes after 120 kg because the competition within it indicates some people's bodyweights to be higher and these people can have the potential to outscore their competition in these powerlifting meets.

Analysis II: For my dimensionality reduction, I used PCA to see if certain strength variables, equipment usage and steroid testing was an appropriate representation of my entire dataset. I used my scaled data to use the prcomp function in order to help find the variance and the proportion of variance explained by the model. I then plotted this out using an elbow graph to show the difference in proportion PC1-PC4 explains in variance between my variables.

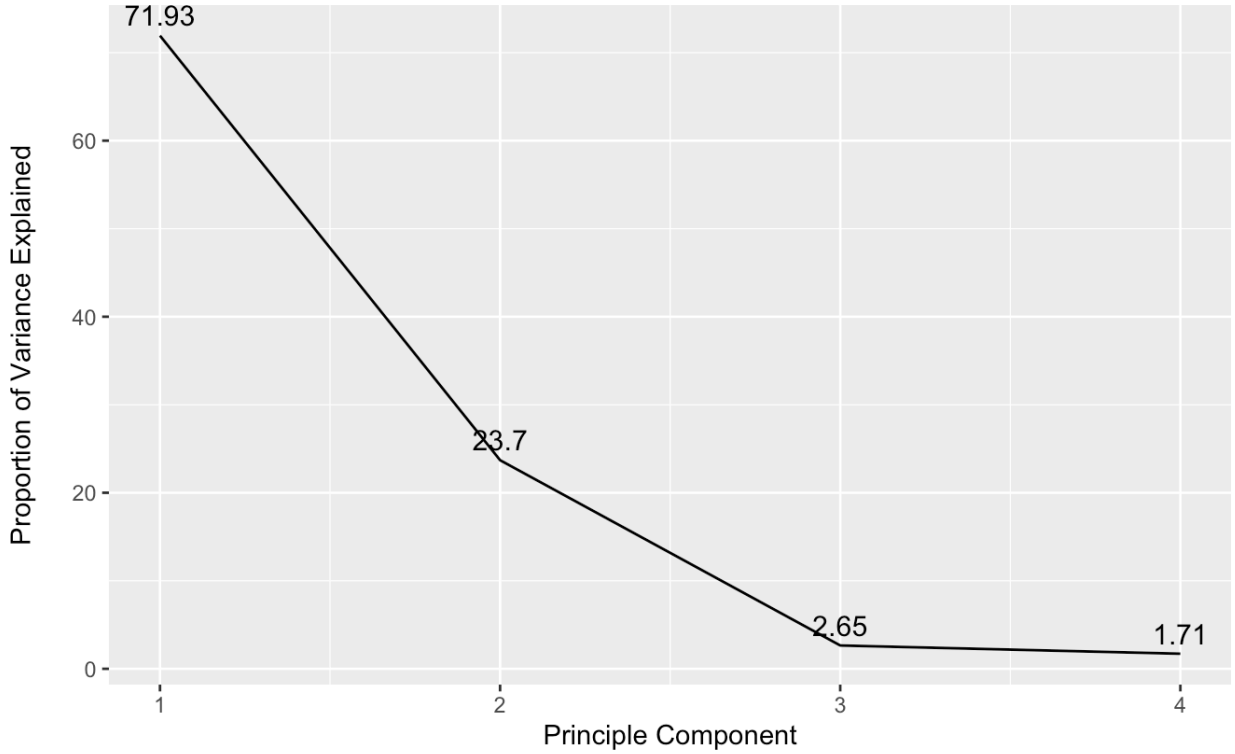

Figure 4. Elbow graph of PC-PC4 variance for factors of bench, squat, deadlift and equipment

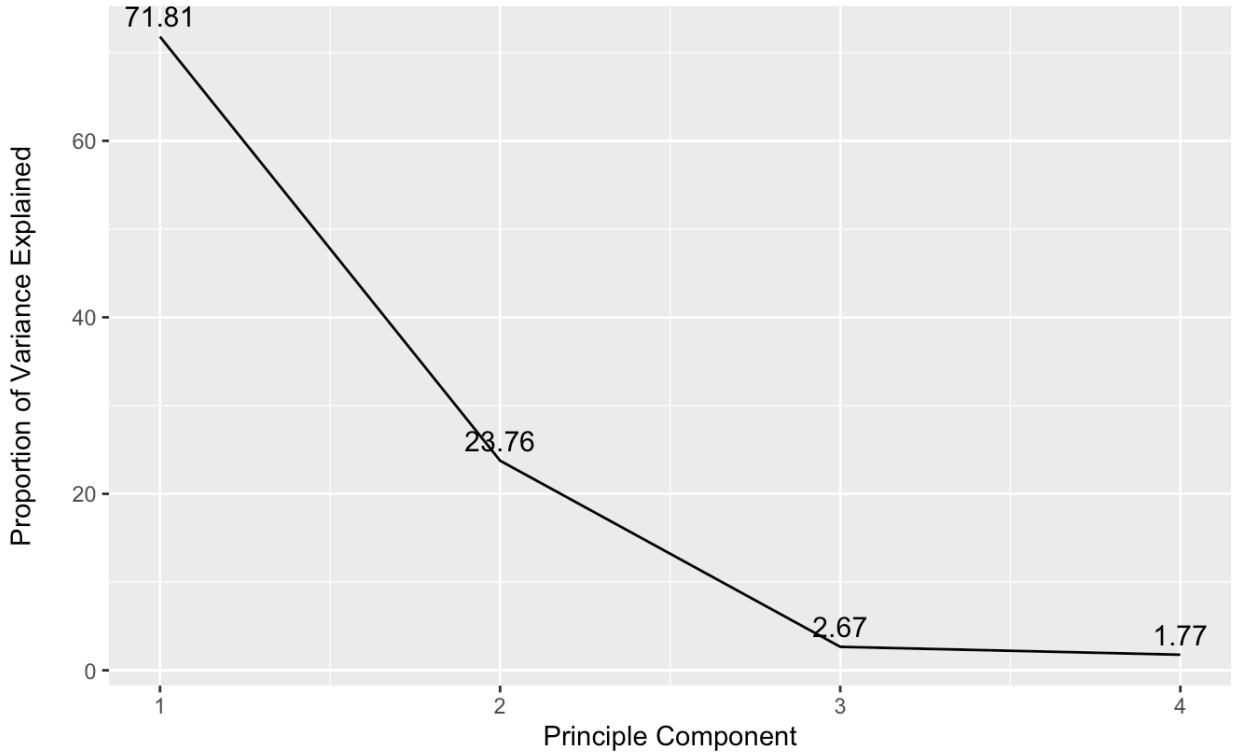

Figure 5. Elbow graph showing PC1-PC4 variance for factors of bench, squat, deadlift and steroid testing

I then put this data into a data frame and added the categorical variable of age classes in order to show different age groupings and different age groups are affected when looking at the difference between PC1 and PC2. I graphed this difference and used color to distinguish age classes for both equipment and testing analyses.

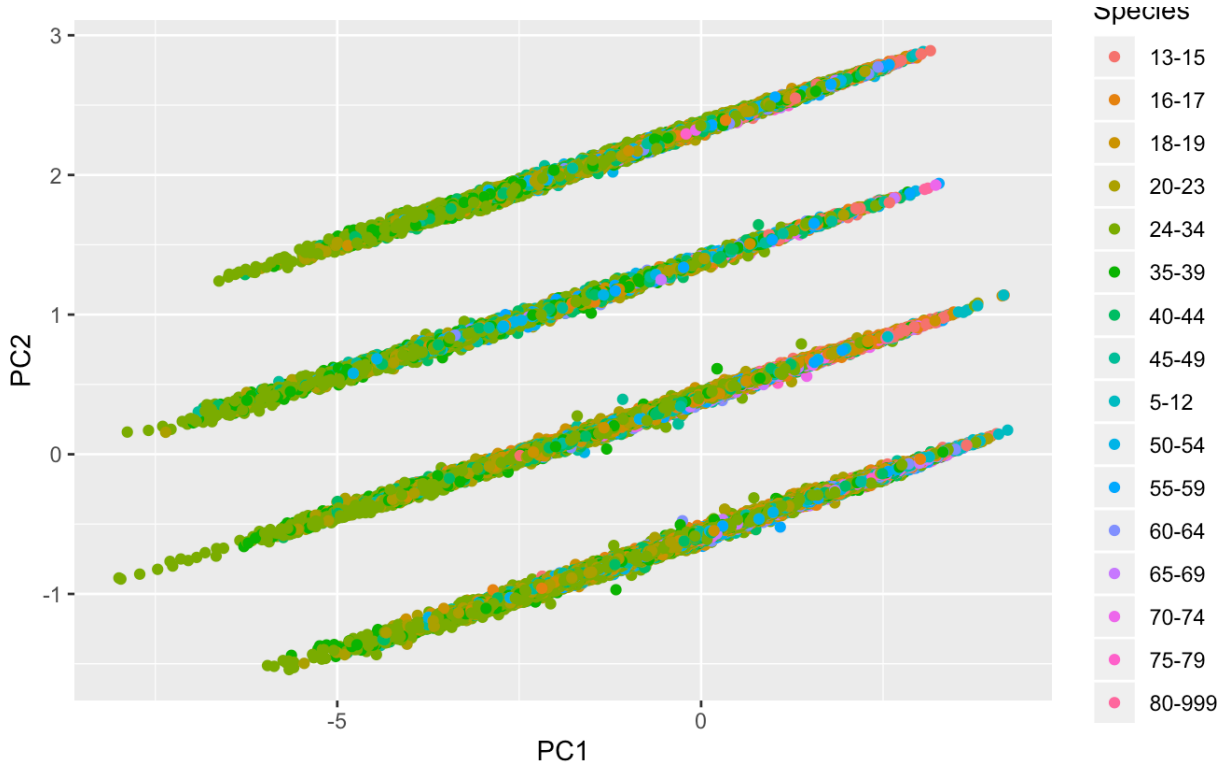

Figure 6. Scatter showing clusters of data for all age groups when plotted for PC1 to PC2 for equipment, bench, squat and deadlift factors

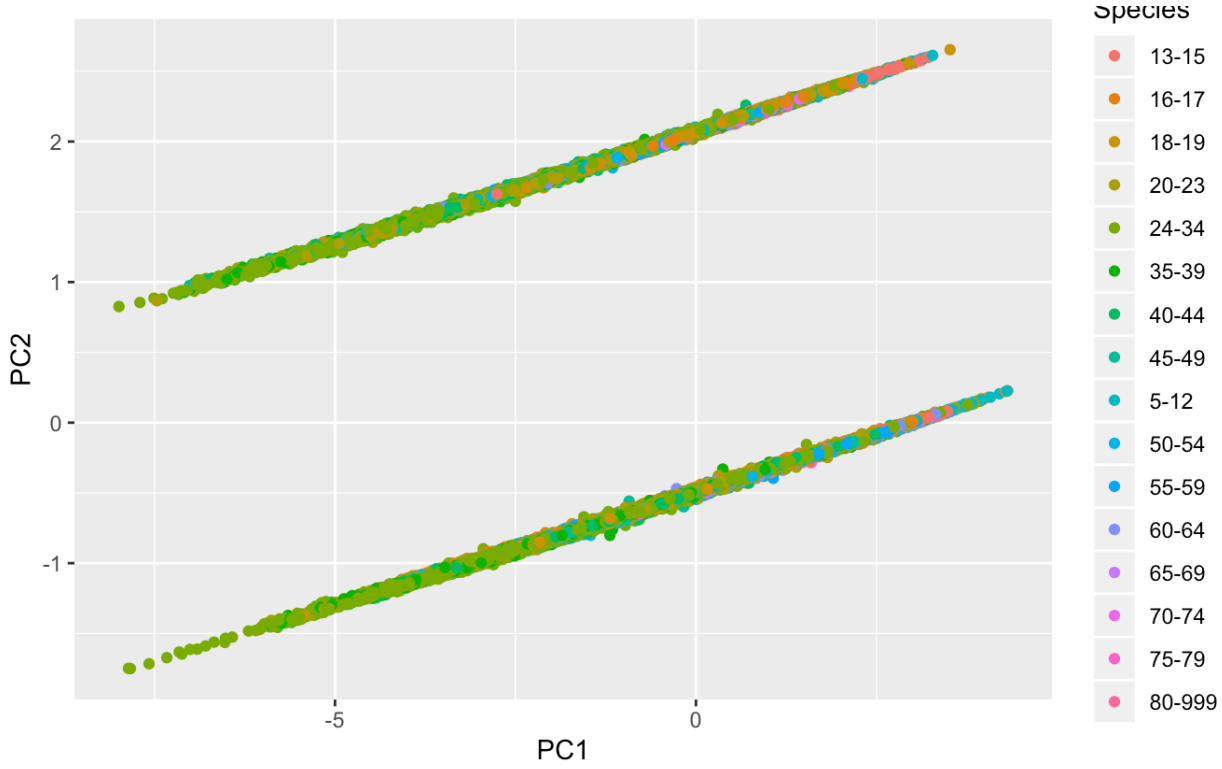

Figure 7. Scatter showing clusters of data for all age groups when plotted for PC1 to PC2 for steroid testing, bench, squat and deadlifting factors

I found that based on my PCA analysis that about 72% of the variance for my dataset can be represented by PC1 alone for both my equipment and steroid testing analysis. This allows me to assume that using these four variables I am able to capture nearly three quarters of my data and my analysis gives a somewhat accurate result. Also, I found that with my PCA analysis that my data has multiple clusters that may not be seen by plotting the regular data. Most of the data points have a strong linear correlation to PC1 and further strengthen my conclusion that my analyses are accurate enough to represent the entirety of the data.

Analysis III: My last analysis was to look at how strength differs for each sex when accounting for different factors at powerlifting meets. My first plot was a bar plot with age classes to show the entire distribution for equipment usage to total kg for both male and females. I also made a graph for a similar distribution except instead of equipment I used steroid testing.

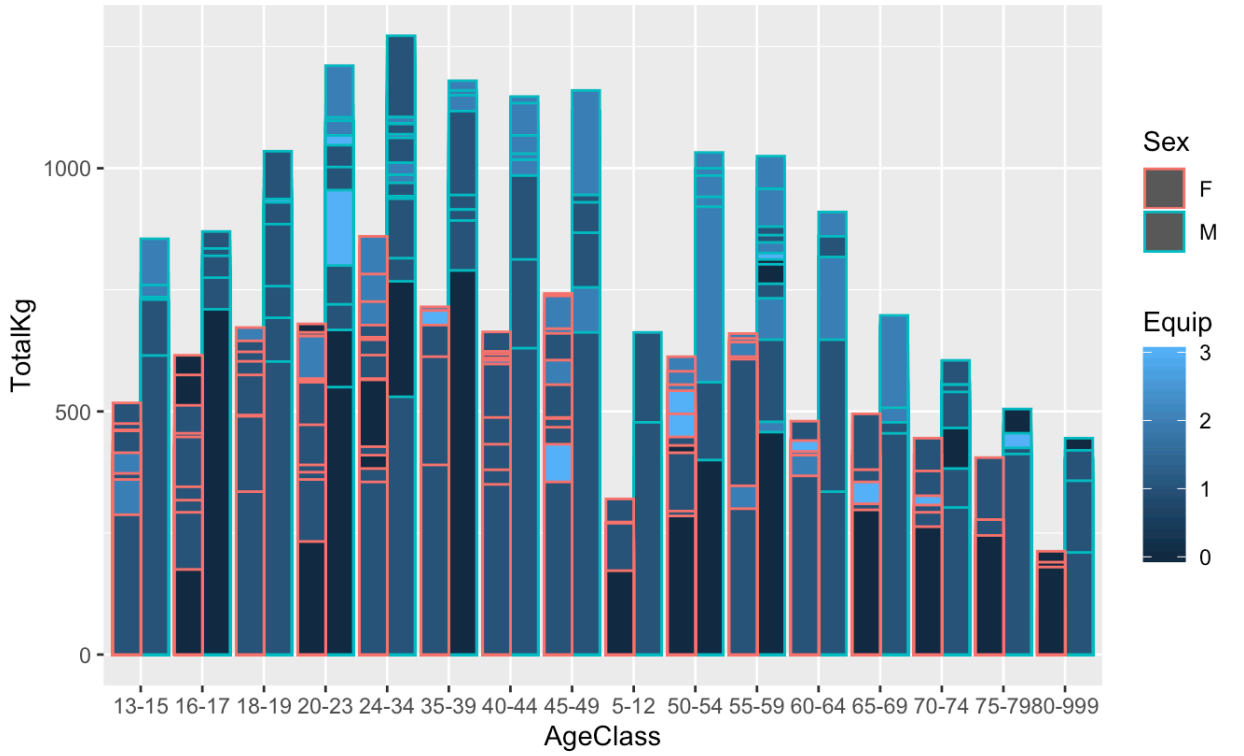

Figure 8. Bar graph showing total kgs per age group with sex accounted for and equipment usage

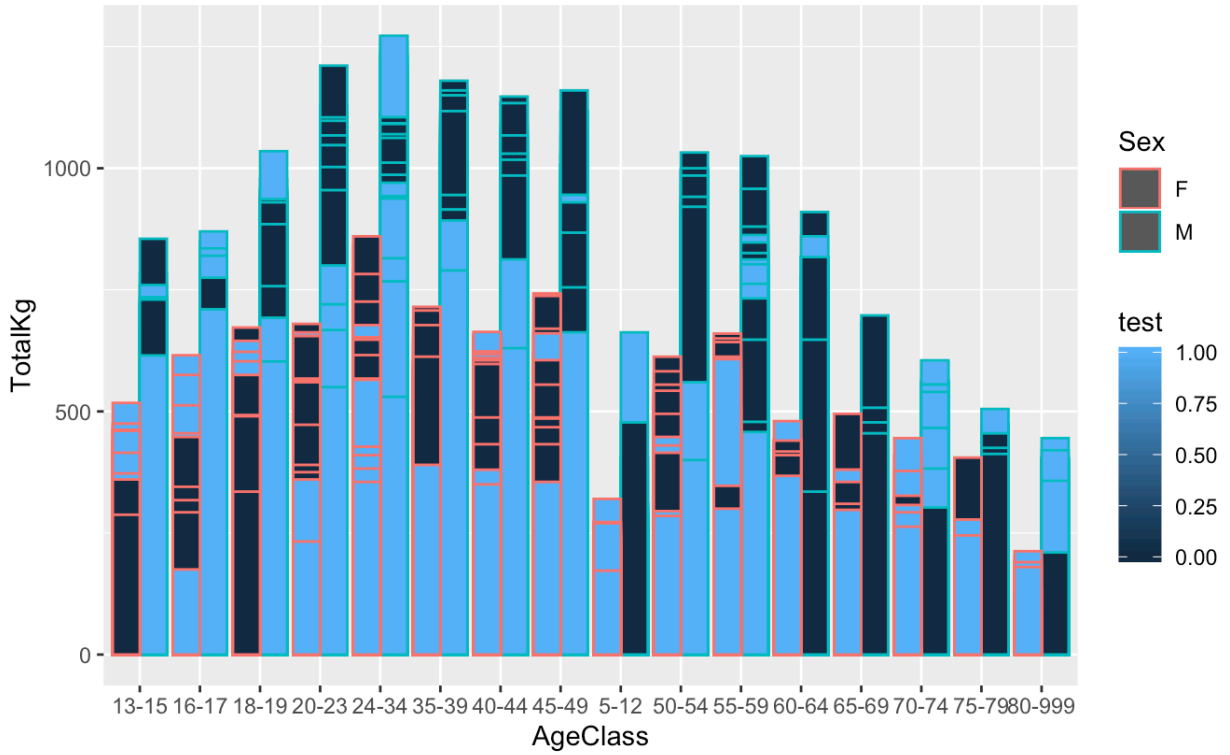

Figure 9. Bar graph showing total kgs per age group with sex accounted for and steroid testing

From the initial visualization for my distribution, I can see that people that men competing in non-testing competitions are lifting more than men in the same age group in tested leagues. The same seems to be the case for women. However, it appears that more people in both sexes are competing in tested competitions rather than non-tested ones. As for equipment, both men and women seem to some sort of equipment for all age classes. The more equipment that these people use does not seem to represent lifting more in their meets either.

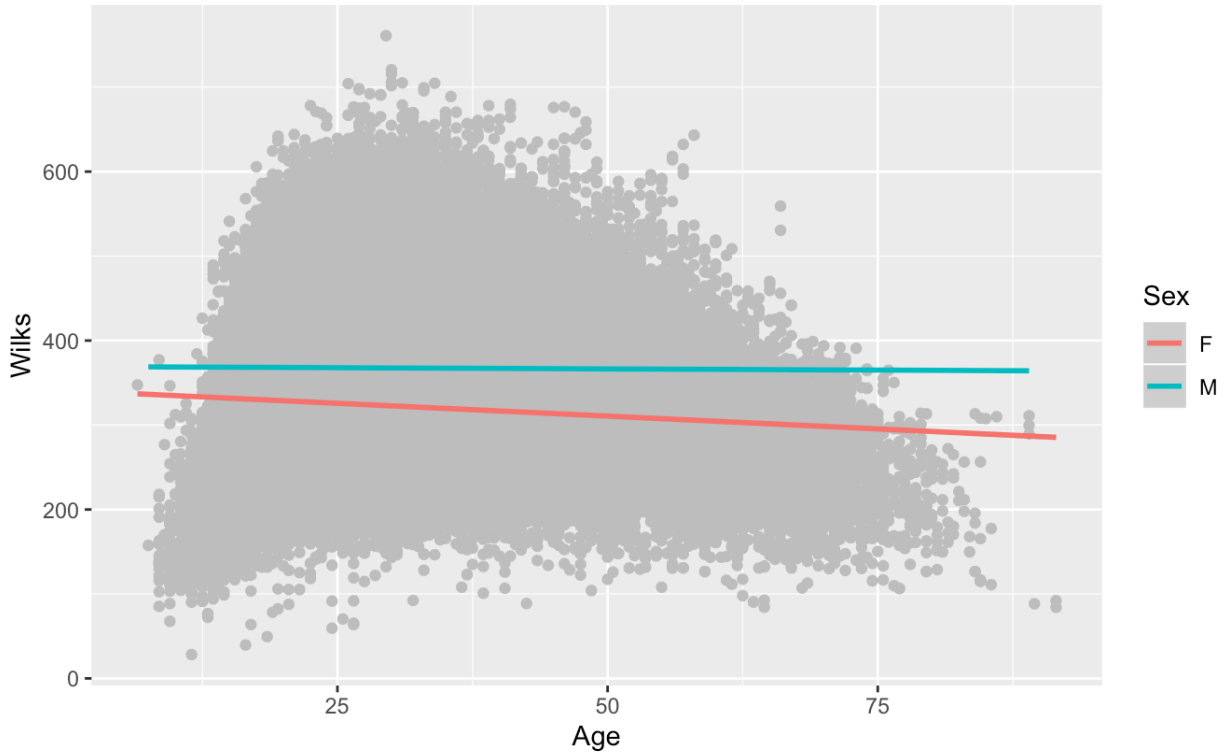

Figure 10. Linear model between wilks and age for both women and men

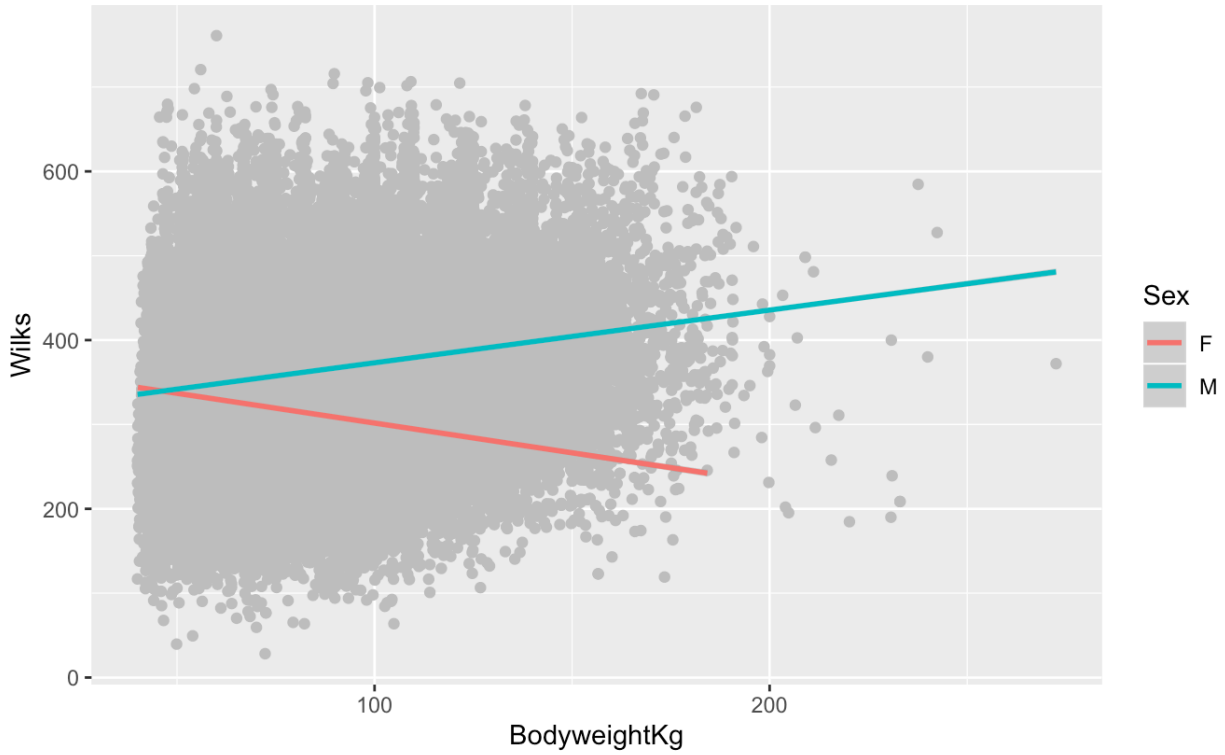

Figure 11. Linear model between wilks and body weight for both women and men

For this specific analysis it is interesting to find that as body weight increases for females, overall strength seems to be going down. However, for males, as body weight increases overall strength seems to increase. Perhaps this may be some kind of difference in overall steroid usage or hormonal changes between genders. When looking at age and overall strength for both females and males, strength decreases as age increases. This makes sense because as our bodies get older the less wear and tear the body can typically take.

Based on my three analyses strength and factors affecting strength, there seems to be