Sarah Patten

Capstone 1: Milestone Report

Problem:

Through the last few years, there has been a large shift in people focusing on a solely strength based exercise regimen, with more competitions available for those athletes to compete.

- As the sport grows, how is the demographic changing?
- Is overall strength increasing?
- How much of an impact does sex have on powerlifting?

These outlined questions above need to be explored to allow new powerlifters see how they fit in, as well as allow current powerlifters identify where they can improve to stand out.

The Data:

The powerlifting database contains two tables.

The first table is filled with the competition meet data. This includes the MeetID, the date, location, and name. A preview of this is displayed below.

	MeetID	MeetPath	Federation	Date	MeetCountry	MeetState	MeetTown	MeetName
0	0	365strong/1601	365Strong	2016-10-29	USA	NC	Charlotte	2016 Junior & Senior National Powerlifting Cha...
1	1	365strong/1602	365Strong	2016-11-19	USA	MO	Ozark	Thanksgiving Powerlifting Classic
2	2	365strong/1603	365Strong	2016-07-09	USA	NC	Charlotte	Charlotte Europa Games
3	3	365strong/1604	365Strong	2016-06-11	USA	SC	Rock Hill	Carolina Cup Push Pull Challenge
4	4	365strong/1605	365Strong	2016-04-10	USA	SC	Rock Hill	Eastern USA Challenge

The other database includes each of the competitors information for each meet. It includes information like meetID, name, age, weight, meet division, best lift amounts for bench, squat, and deadlift, as well as the competitors place, and Wilks score. The meetID in each table allows for combining these tables for more information about when the lifts took place.

A preview of this dataFrame prior to any adjustments is displayed below; because there are so many columns it is split on 2 lines.

	MeetID	Name	Sex	Equipment	Age	Division	BodyweightKg	WeightClassKg	Squat4Kg	BestSquatKg	Bench4Kg	BestBenchKg	Deadlift4Kg
0	0	Angie Belk Terry	F	Wraps	47.0	Mst 45-49	59.60	60	NaN	47.63	NaN	20.41	NaN
1	0	Dawn Bogart	F	Single-ply	42.0	Mst 40-44	58.51	60	NaN	142.88	NaN	95.25	NaN
2	0	Dawn Bogart	F	Single-ply	42.0	Open Senior	58.51	60	NaN	142.88	NaN	95.25	NaN
3	0	Dawn Bogart	F	Raw	42.0	Open Senior	58.51	60	NaN	NaN	NaN	95.25	NaN
4	0	Destiny Dula	F	Raw	18.0	Teen 18-19	63.68	67.5	NaN	NaN	NaN	31.75	NaN

BestDeadliftKg	TotalKg	Place	Wilks
70.31	138.35	1	155.05
163.29	401.42	1	456.38
163.29	401.42	1	456.38
NaN	95.25	1	108.29
90.72	122.47	1	130.47

The first table was pretty clean and didn't need much manipulation. The date column was converted to date type object to allow for the creation of year and month columns. Pandas to_datetime function allowed for this conversion, after using it, lambda functions were used to create the two new columns.

The second table needed more work. The Squat4Kg, Bench4Kg, and Deadlift4Kg columns had very few values, so they were removed. After exploring that, the method describe was used to explore the maximum and minimum values to see if anything seemed off. It turned out a small portion of the table had negative values for the squat, deadlift, or bench values. This was due to an unsuccessful attempt at the weight. Since there was only a small percentage of records that contained negative values (<1%) the records were removed from the table.

The Equipment column originally had seven columns that had mislabeled "wraps" as "straps"; these columns were corrected to having the right label. This column, as well as the Sex column, was changed to a category.

A large range of ages participate in these competitions (from 5-95), and since we couldn't easily reclassify the weight classes, and the "division" column has too many options, an ageGroup column was created. The USAPL age classifications were used. There were some rows that were missing an age; these were assigned the "Open" ageGroup. The first few rows of the updated dataFrame are shown below.

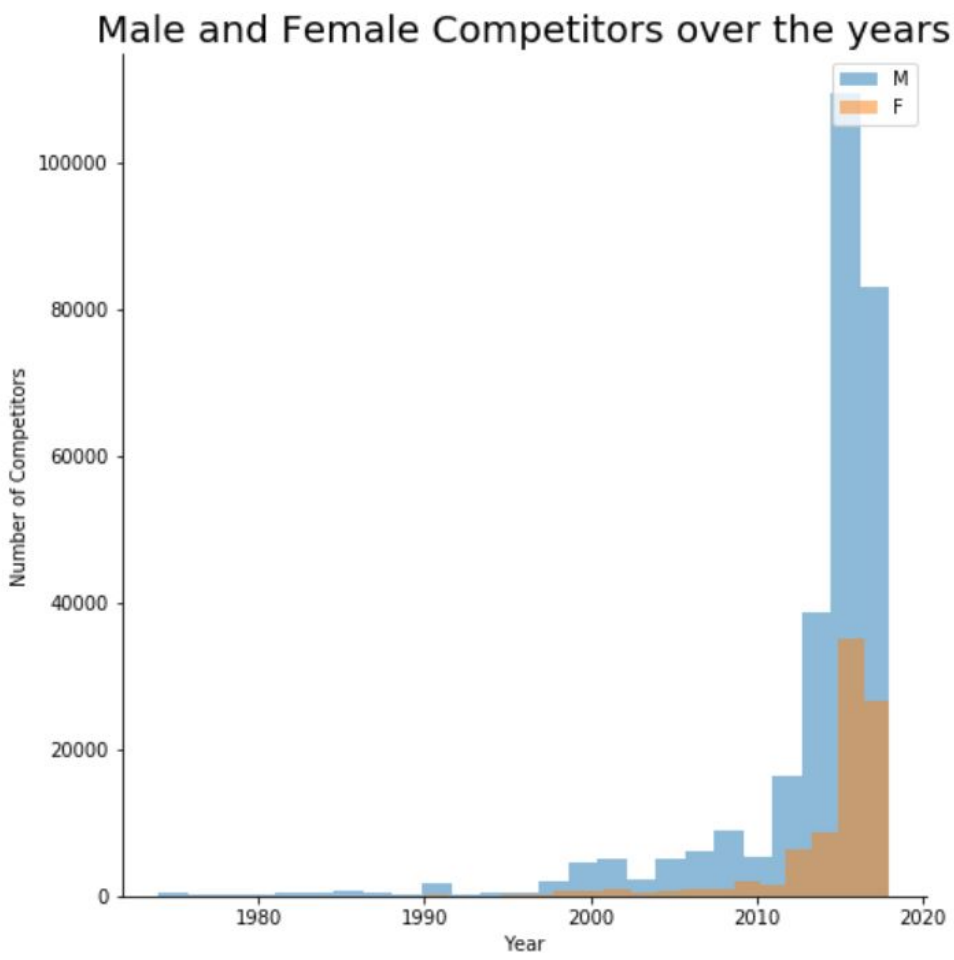
	MeetID	Name	Sex	Equipment	Age	Division	BodyweightKg	WeightClassKg	BestSquatKg	BestBenchKg	BestDeadliftKg	TotalKg	Place	Wilks	AgeGroup
23	0	Kevin Gingerich	M	Raw	32.0	Open Junior	71.94	75	154.22	115.67	183.70	453.59	2	333.01	OPEN
24	0	Juan Bollo	M	Raw	20.0	Open Junior	70.67	75	163.29	111.13	204.12	478.54	1	356.03	JUNIOR
25	0	James McManus	M	Wraps	36.0	Open Junior	74.93	75	NaN	NaN	NaN	NaN	DQ	NaN	OPEN
26	0	James McManus	M	Raw	36.0	Open Junior	74.93	75	NaN	115.67	156.49	272.16	1	194.06	OPEN
27	0	Scott Faircloth	M	Wraps	27.0	Open M/P/F	71.30	75	181.44	99.79	188.24	469.47	1	346.96	OPEN

Exploratory Data Analysis:

Initial findings from exploratory analysis (get this from your data story and inferential statistics reports)

1. Summary of findings
2. Visuals and statistics to support findings

To get an idea of how demographics were changing, a bar graph was created to show the number of male vs female competitors there have been over the years since the sport began being recorded. The figure below shows that it is gaining a lot of popularity in the more recent years (the most recent date in the dataset is January 2018), though males are still competing far more than females.



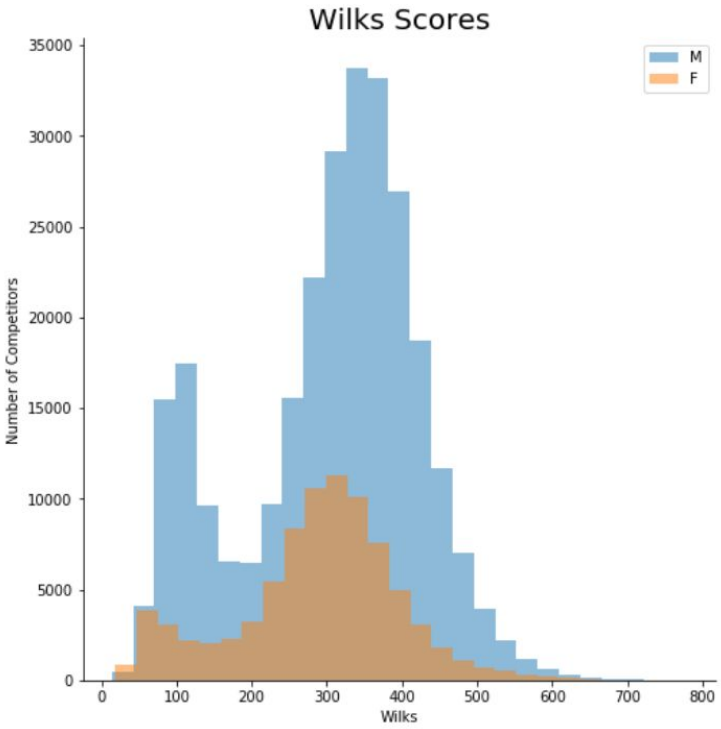
The Wilks score is a number used to assign a strength value for each individual. It is computed with a formula to normalize lifts across weight classes. The constant values on the bottom differ between Males and Females, to help bridge a gap between them.

$$Coeff = \frac{500}{a + bx + cx^2 + dx^3 + ex^4 + fx^5}$$

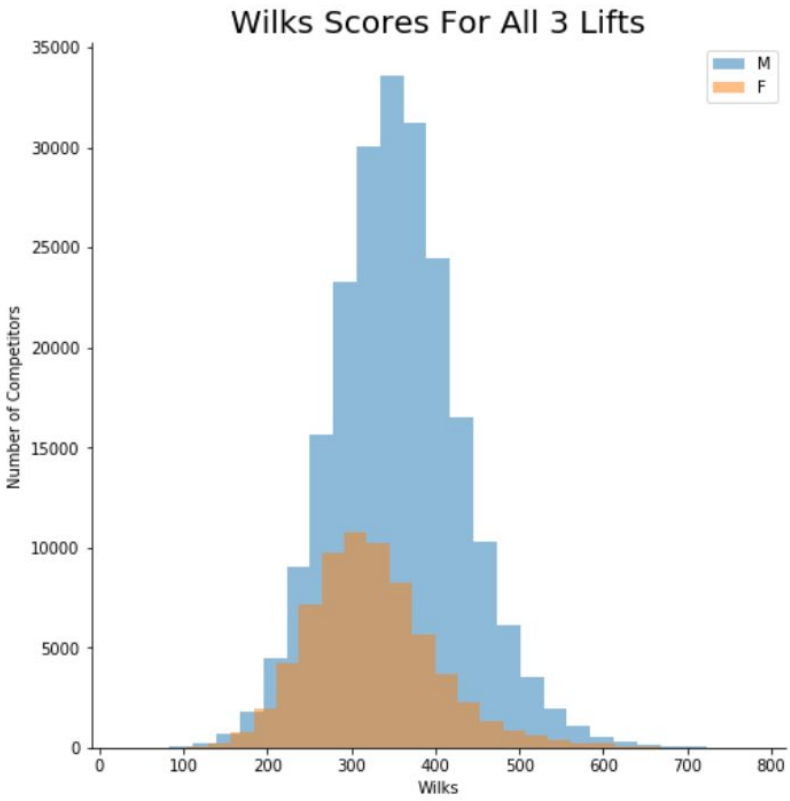
This value was used to help see how strength is changing over time. The graph below shows how the max Wilks scores for males and females have been changing. It was interesting to see that the max Wilks scores for both males and females were so close; the max for females is 776.17, while the max for males is 779.38. It was interesting to see that women's Wilks scores are increasing at a faster rate than males; this could be because women got into lifting later than males, and there are less competitors so stronger individuals stand out more.

I would like to note that the current two highest Wilks scores noted above were both obtained with the use of wraps, and the current top 5 highest Raw wilks scores are held by females. (This information is available on the [open powerlifting website]).

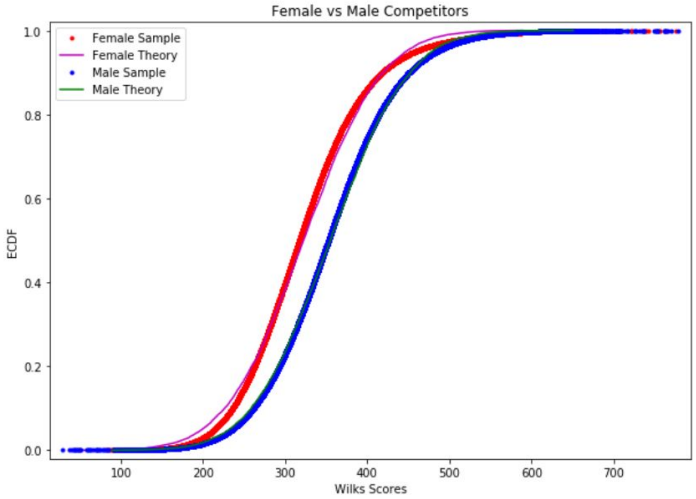
Because of its normalization, the wilks score makes for an interesting data field to apply some analytics on. Originally when looking at this data, and creating the bar graph below, I noticed that there were two peaks in the chart for the scores.



This was due to individuals who didn’t complete all three lifts. A new subset of the dataset was obtained that contained only individuals that had values for bench, squat, and deadlift. The new bar graph of the data resembled more of a normal distribution.



That graph appears to show that there is a different mean Wilks Score for male and female competitors, even though it's normalized. This data was then put into the ECDF function to help determine whether it was a normal distribution. Both male and females tend to closely follow a theoretical curve for the ECDF.



The means displayed above do look different, but still fairly close. The null hypothesis to be tested was that the average wilks score for males was equal to the average wilks score for females.

- H0: mean(females.Wilks) = mean(males.Wilks)
- H1: mean(females.Wilks) ≠ mean(males.Wilks)

To test this, bootstrapping was first used. I adjusted the females wilks scores so that the mean matched that of the males, and drew 10,000 replicates to see how many times we could obtain the number of times our bootstrap sample had a mean less than or equal to our females average wilks score. Out of the 10,000 samples, not a single one came back with a mean as extreme as our data; a P-value of 0.

After this test, a t-test for two datasets, with an assumed identical mean, was performed. This test (in scipy.stats) resulted in a P-value of 0, and a T-statistic value of 94.2.

From this, we can conclude that we should reject the null hypothesis, and offer up our H1 hypothesis that there is a difference in Wilks scores for males and females, even with altered constants used in computing them. Were males and females ever to compute against each other, this Wilks score formula would need to be adjusted to account for the difference in male and female strength. In fact, some powerlifting competitions have started switching away from using the wilks score, and are normalizing lifts differently.

There is still a large gap in the number of male and female lifters, and in the future, this gap in the wilks score could close on its own, as more females begin competing, and more research is done of female strength, but for now it is safe to say male competitors are stronger overall.