# P1 POWERLIFTING

Stefan Leif Brzost-Andersen

22/12/2021

AALBORG UNIVERSITY

STUDENT REPORT

**Title:**

OpenPowerlifting Datasæt

**Project:**

P1

**Project Period:**

08/12/2021 - 22/12/2021

**Project Group:**

B264a (SOLO)

**Group Members:**

Stefan Leif Brzost-Andersen

**Supervisor:**

Christophe Biscio

**Number of Pages:**

15



**The Department of Computer Science**
Data Science
Selma Lagerløfs Vej 300
9220 Aalborg Øst
https://www.cs.aau.dk/

**Abstract:**

Early-concepts of dataexploration, wrangling analysis of OpenPowerlifting's Dataset.

# Indholdsfortegnelse

# Underskrift

Stefan Leif Brzost-Andersen

# Reader's Guide

# 1 Introduktion

Powerlifting er en styrkesport, hvor atleter kan stille op uden udstyr (Typisk refereret som 'RAW / Classic' dvs. uden udstyr) eller med udstyr. Udstyr defineres typisk i form af en mekanisk-støttende vest (Bench, deadlift, squat- suit) og 'wraps', hvis formål er at støtte ens banekurve igennem løftet og sørger for, at atleten kan klare tungere vægte, som i nogle tilfælde er over 100kg end hvad de ellers ville kunne have løftet i RAW. Alt dette er ulovligt i RAW / Classic, imens 'straps', 'bælte', 'squat shoes' og andet udstyr, hvis formål er til at minimere skaderisiko er lovligt.

Atleten kan vælge mellem at stille op i både squat, bænkpres og dødløft. Men også stille op til solo-konkurrencer, hvor der kun stilles op i en enkelt disiplin bl.a. DM i Bænkpres. Ligesom i Olympisk Vægtløftning vil atleten modtage tre løft - uanset disiplin, hvor kun det stærkeste og teknisk-godkendte løft bliver registreret.

Konkurrencerne findes i de fleste steder i verden og er også en af de store paralympiske sportsgrene (bænkpres ONLY). Der findes internationale, interkontinental, nationale, lokale versioner, hvor præmien typisk er æren, fremfor en større økonomisk gevinst, som der ses i andre sportsgrene. Herefter følger der forskellige foreninger med til konkurrencerne, hvor nogle har banlyst andre, da reglerne er forskellige fra forening til forening - både ift. doping-politik, men også teknik, krav for udstyr, valg af strømper, sko, trøje, etc.

Powerlifting har fået en del popularitet de seneste år og vi vil derfor kigge på, hvor hurtigt PL har vokset, hvorfor PL har vokset og om der er noget betydeligt grundlag for større foreninger sanktionere mindre foreninger, såfremt deres reglement ikke støtter deres eget. En af de større problematikker vi vil fokusere på er om der ses en betydelig forskel mellem non-doping foreninger og doping foreninger. Til sidst vil vi se på nogle af de største faktorer til powerlifting, heraf WILKS og om en ny, mere retfærdig og non-bias løsning kan findes til bestemmelse af atleternes styrke-index.

# 2 Exploratory Data Analysis

OpenPowerlifting Datasæt indholder data fra 1962 til 2021 og samtlige konkurrencer, events, meets er registreret. Der er 41 variabler og ca. 2.5 millioner samples i datasættet. Udover dette vil en del variabler blive ekskluderet pågrund af mangel på tid - dermed er der blevet valgt 16 ud af de 41 variabler. Udvalgte variabler vil blive analyseret og beskrevet i de kommende afsnit.

| **Ekskluderet** | | | |
|---|---|---|---|
| Name | BirthYearClass | Division | Place |
| Dots | Glossbrenner | Goodlift | State |
| MeetState | MeetTown | Meetname | Event |
| Squat1Kg | Squat2Kg | Squat3Kg | Squat4Kg |
| Bench1Kg | Bench2Kg | Bench3Kg | bench4Kg |
| Deadlift1Kg | Deadlift2Kg | Deadlift3Kg | Deadlift4Kg |
| Dato | MeetCountry | Udstyr | Deadlift4Kg |

## 2.1 Data Typer

Datasættet indholder forskellige datatyper og de vil derfor opdeles til to kategorier: *Numeriske data* og *Kategoriske data*.

| **Inkluderet** | |
|---|---|
| ***Numerisk Data*** *Dtype = Float64* | **Kategorisk Data** *Dtype = Object* |
| Alder | Køn |
| Kropsvægt | Land |
| Maksimal Squat | Kategori |
| Maksimal Bænkpres | Vægtklasse |
| Maksimal Dødløft | Doping |
| Total Load | Parent Federation |
| Wilks | Federation |

## 2.2 Data Cleaning

OPL-Datasæt er betydeligt større end det almindelige datasæt med sine 2.5 millioner samples. Der er blevet kortlagt en række conditions til datacleaning, som minimere fejlagtig data og sørger for det er gennemskueligt til analyse. Da udgangspunktet i analysen er de inkluderet variabler vil de resterende variabler ikke renses og der opdeles to seperate datasæt. Det rensede datasæt kører igennem fejlsøgning i form af loops og derefter er det klart til videre analyse.

| **Datacleaning** | |
|---|---|
| **Condition** | **Inkluderet** |
| Alder | Ja |
| Kropsvægt | Ja |
| Doping/Non-Doping | Ja |
| WILKS | Ja |
| RAW / Uden Udstyr | Ja |

Da der blev tjekket for manglende værdier er den overaskende lav, da der er minimal tomme felter i OPL-Datasæt. Der blev dog fundet en betydelig del N/A-værdier og besynderlige værdier, såsom kropsvægt på 0.5kg og en alder på 105 år. Dette bringer vores total-mængde ned på 462.965 samples og 16 features.

| **Manglende Værdier (Procent)** | | |
|---|---|---|
| **Data** | **Før** | **Efter** |
| Sex | 0.00 | 0.00 |
| Event | 0.00 | 0.00 |
| Equipment | 0.00 | 0.00 |
| Age | 37.91 | 0.00 |
| BodyweightKg | 1.30 | 0.00 |
| WeightClassKg | 1.15 | 0.20 |
| Best3SquatKg | 34.15 | 0.00 |
| Best3BenchKg | 12.09 | 0.00 |
| Best3DeadliftKg | 28.74 | 0.00 |
| TotalKg | 7.01 | 0.00 |
| Wilks | 7.88 | 0.00 |
| Tested | 26.22 | 0.00 |
| Country | 44.95 | 0.00 |
| Federation | 0.00 | 0.00 |
| ParentFederation | 35.47 | 11.11 |

Fordelingen er dog ikke uniform blandt alle features og af denne grund vil der blive tilfældigt trukket 100.000(n) tilfældige samples. Da der kan være betydelig difference mellem data fra non-doping og doping federationer vil grafer bruge data fra federationer med godkendt doping-politik - uden andet er beskrevet.

## 2.3 Numeriske Data

### 2.3.1 Alder

Alder angiver hvor gammel den givne atlet er på konkurrences tidspunkt. Denne variable bruges til at indstille dem til deres pågældende kategori (beskrives i 2.3.3 Kategori).

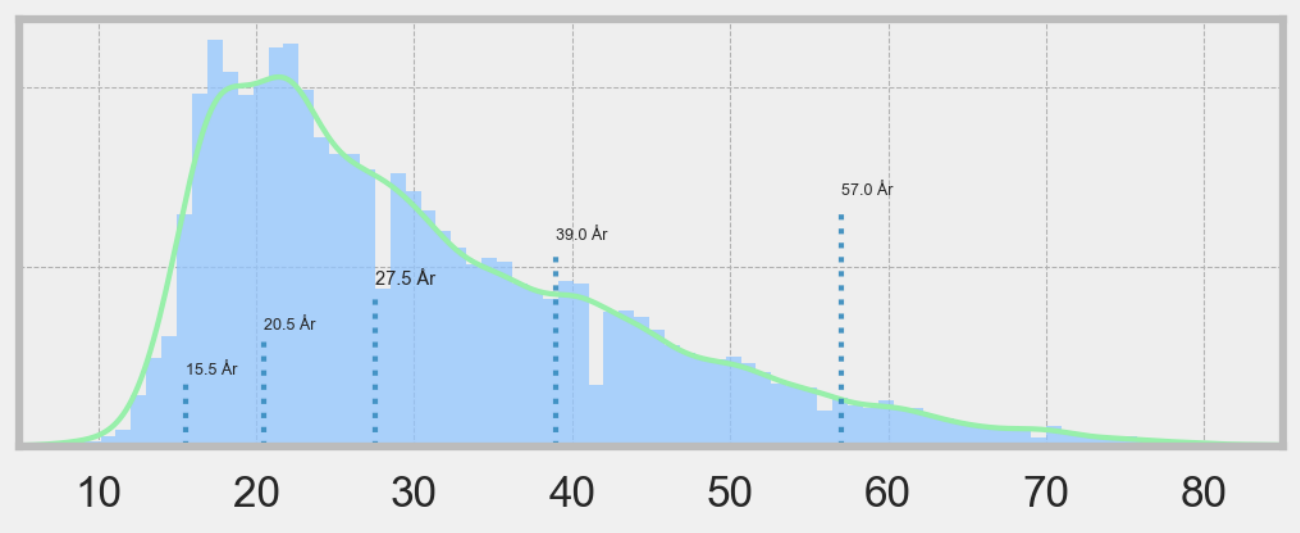



***Figure 2.3.1:*** Graf viser fordeling pr. Alder og hhv. 5, 25, 50, 75, 95 kvantil.

Ved nærmere overview af OPL-Datasæt ser vi, at den ældste atlet er 95 år gammel, imens den yngste atlet 7 år gammel. Fordelingen af alder hos atleterne kan ses i overstående histogram m. Kvartilsæt og KDE. Den gennemsnitlige alder ligger på 27.5 år og i de øvre kvartiller ses der hhv. 39 og 57 år. Sporten er derfor mest populær hos folk på maksimalt 28 år og under, imens det kun er 5 procent, der stiller op efter 57 år. Dette tænkes som en tendens med PL's stigende popularitet i verden.

### 2.3.2 Kropsvægt

Atletens kropsvægt er en af de afgørende værdier indenfor PL, da ens absolut styrke stiger med ens total masse. Vi vil senere kigge på sammenhængen mellem kropsvægt og køn i kombination med absolut styrke.

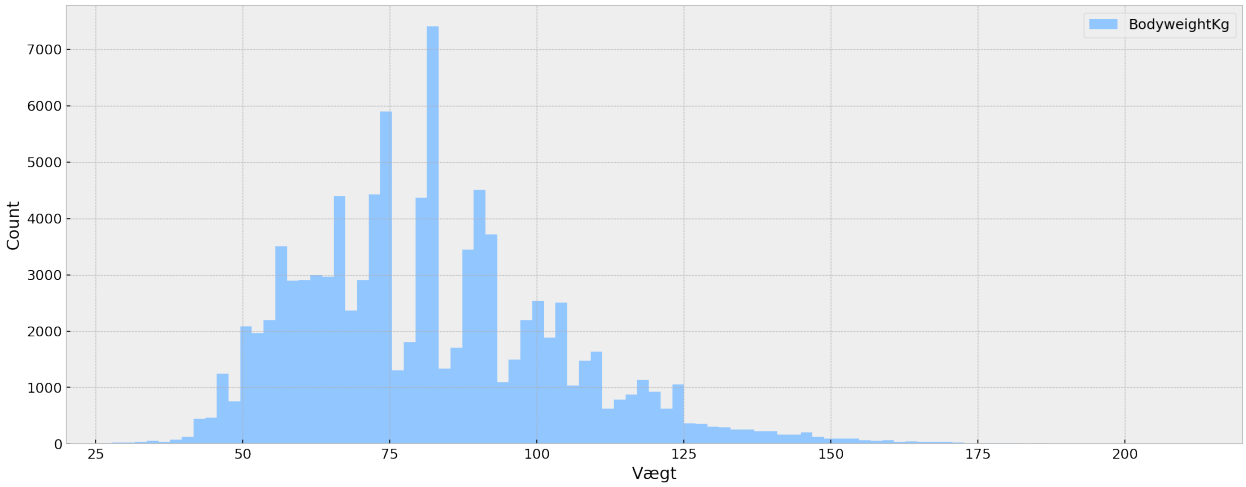



***Figure 2.3.2:*** Graf viser fordeling i atleternes kropsvægt.

| | Count | Mean | Std | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| **BodyweightKg** | n = 100.000 | 81.87 | 22.17 | 21.7 | 65.5 | 80.5 | 94.4 | 220.3 |

Den gennemsnitlige atlet vejer 81.87 kilogram og afvigelsen fra mean er på 22.17 kilogram. Så størstedelen af atleterne vejer mellem ca. 60kg til 104kg og det må automatisk være de mest populære vægtklasser. Den mindste atlet vejer 21.7kg og vores yngste atlet er 7 år gammel, dette hænger sammen med den forventede vægt for et 7-årigt barn.

1. kvartil viser at 25 procent og under vejer 65.5kg, hvilket muligvis kan passe med kønsfordelingen mellem mænd og kvinder. Ved 2. Kvartil ser vi 80.5kg og ved 3. Kvartil ser vi 94.4 kg - det observeres derfor, at størstedelen af atleterne deltager i vægtklassen under 103kg. Den største registreret atlet lå på 220.3 kilogram og og betegnes som en outlier, som observeres på grafen.

### 2.3.3 Maksimal Squat, Dødløft, Bænkpress

Best3Squat, Best3Bench, Best3Deadlift er det stærkeste og pæneste løft ud af alle atletens forsøg til en konkurrence. Hvori TotalKg er atletens samlede total løft, hvor man beregner summen af alle tre løft. Den stærkeste atlet er derfor den med det højeste total løft i alle disipliner.

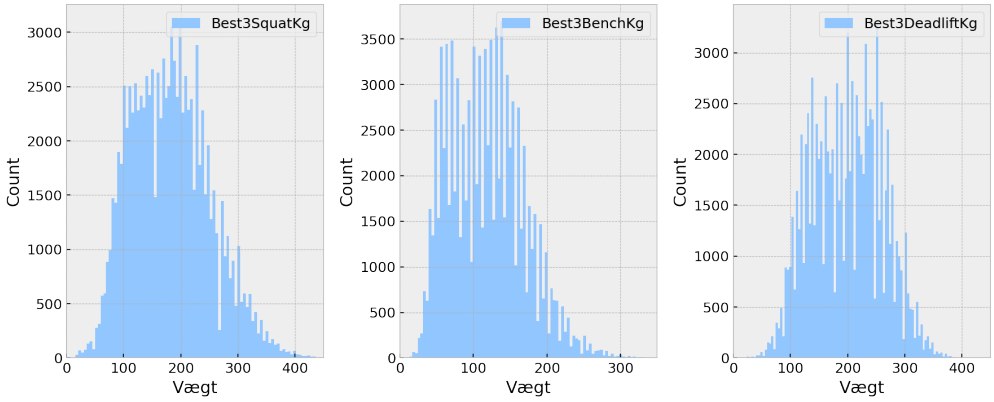



***Figure 2.3.3:*** Graf viser fordeling i hhv. Squat og Bænkpress.

| | Best3SquatKg | Best3BenchKg | Best3DeadliftKg |
|---|---|---|---|
| **Count** | n = 100.000 | n = 100.000 | n = 100.000 |
| **Mean** | 182.1 | 118.1 | 198.23 |
| **Std** | 68.82 | 50.5 | 61.5 |
| **Min** | 1 | 9 | 15 |
| **25%** | 128 | 75 | 147.5 |
| **50%** | 180 | 115 | 200 |
| **75%** | 230 | 150 | 245 |
| **Max** | 490 | 385 | 405 |

Vi kan observere udefra deres mean, at den gennemsnitlige atlet generelt er stærkere i dødløft end i squat og bænkpres. Hvor man generelt er svagest i bænkpres og squat hænger tæt op af dødløft. Derimod er der mindst afvigelse fra mean ved bænkpres, men højest i squat. Der ses også en mindste-værdi på 1 i squat, hvor det var besværligt at finde fejl i dette. Da alle 0.5 værdier er fjernet efter datacleaning og 1kg ikke er et uværdigt løft for en 7-årig til en konkurrence. Derfor vil et teoretisk løft for en 7-årig med 1kg squat, 9kg bænkpres og 15kg dødløft ikke var usædvanligt. Det er dog som forventet, at atleternes mindste værdi er højest dødløft.

I hhv. 1, 2 og 3 kvartil ser vi, at 25 procent af atleterne har et maksimal løft på 128kg i squat, 75kg i bænkpres og 147.5kg i dødløft. Derefter er næste niveau på 180kg i squat, 115kg i bænkpres og 200kg i dødløft. Sidste niveau er på 230kg i squat, 150kg i bænkpres og 245kg i dødløft. Det skal dog noteres at der ikke er taget forhold for absolut styrke i disse værdier og de reele tal er markant anderledes, når vægtklasser kommer ind i formlen.

### 2.3.4 Total Load

Total Load er summen af alle tre disipliner. Eksempelvis, hvis en atlet har et SBD-løft på hhv. 200kg, 150kg og 200kg vil det give et total load på 550kg. Total load's mean er på 498.55 kg med en afvigelse på 174.88kg. De fleste registreret atleter vil derfor ligge mellem 323.67kg til 673.43kg i total løft og dette forhold kan sammenlignes med atleternes kropsvægt, da det højeste total load er på 1205 kilogram og der skal være en tilsvarende, proportional kropsvægt for at opnå den total load.

| TotalKg | Count | Mean | Std | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| **Total Load** | n = 100.000 | 498.55 | 174.88 | 38.6 | 355 | 498.95 | 622.5 | 1205 |

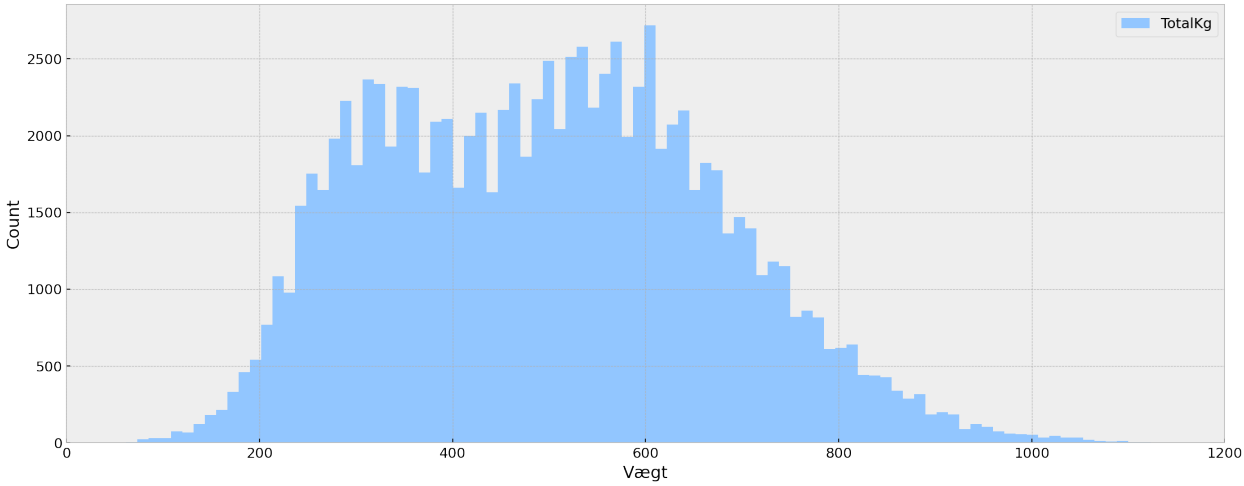
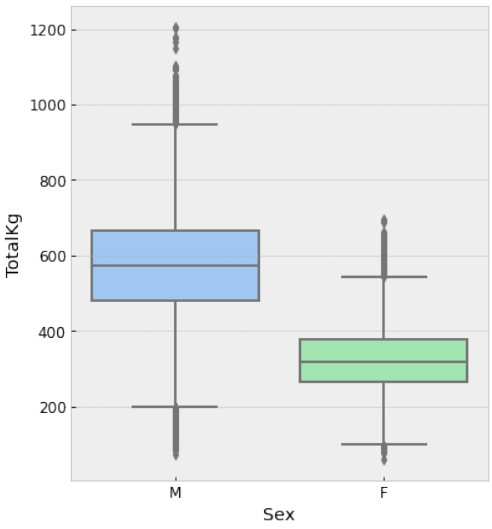



***Figure 2.3.4:*** Graf viser fordeling i Total Load (KG)

I 2.3.5 har vi 'M' for Mand og 'F' for kvinde. Grafen giver en nogenlunde visuel præsentation af styrkeforskellen mellem mænd og kvinder, hvor der ses højest afvigelse fra de øvre outliers, der kan betegnes som de absolut stærkeste kvinder og mænd. Derimod er de svageste mænd og kvinder på nogenlunde samme niveau, dog er der talrige fejl i denne udtalelse og den er kun observeret fra denne graf.

***Figure 2.3.5:*** x = Sex, y = TotalKg



### 2.3.5 Wilks

'*Mass Moves Mass*'-citatet kan bruges til forklaring af PL's vægtklasse system. Der har været gentagne matematiske modeller, som er designet til et enkelt formål; Målingen af relativ styrke mellem de forskellige vægtklasser - en model der skal neutralisere det naturlige hierarki, der ses ved måling af mekanisk styrke, hvor absolut styrke placeres i den absolutte top.

| | Mænd | Kvinder |
|---|---|---|
| **a** | 47.46178854 | -125.4255398 |
| **b** | 8.472061379 | 13.71219419 |
| **c** | 0.07369410346 | -0.03307250631 |
| **d** | -0.001395833811 | -0.001050400051 |
| **e** | $7.07665973070743 \times 10^6$ | $9.38773881462799 \times 10^6$ |
| **f** | -1.20804336482315 x $10^8$ | $-2.3334613884954 \times 10^8$ |

***Table 1:*** Oversigt over tal.

En af disse matematiske modeller er Wilks og er på nuværende tidspunkt en af de mest brugte metoder til, at måle relativ styrke hos styrkeatleter. På denne måde vil man se hvem der er stærkest, når man fjerner '*Mass Moves Mass*' fra den naturlige formel. Problemet i modellen er, at den bestemt ikke er perfekt og der findes optimeret formler, som ville passe bedre til formålet. Wilks har i tidligere modeller haft en betydelige køns-bias og det ses typisk at mænd sætter højere op end kvinder. Imens der også ses en alders-bias i en enkelt vægtklasse (Open Class).

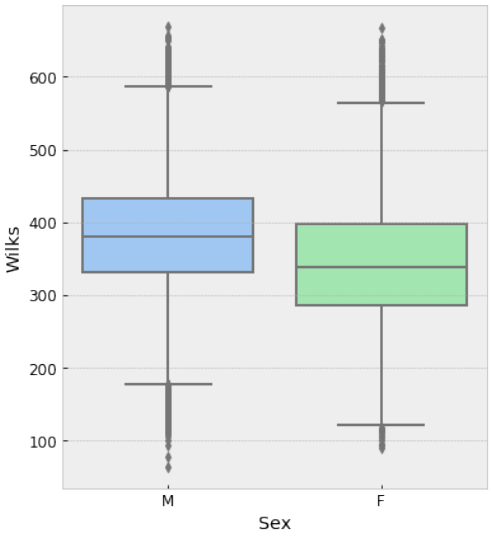



***Table 2:*** Outlierboxplot

Der observeres en markant udligning af køns-bias mellem tidligere modeller og dette forhold kan ikke isoleres til reelt køns-bias ved kort omfang. Det kan diskuteres om en højere kønsdeltagelse ville neutralisere afvigelsen mellem mænd og kvinder ved sammenligning af WILKS.

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

***Table 3:*** 2020-Version: Wilks Formula

## 2.4 Kategorisk Data

### 2.4.1 Land

Da Powerlifting er en international sport vil atleter fra forskellige lande tit kæmpe mod hinanden.

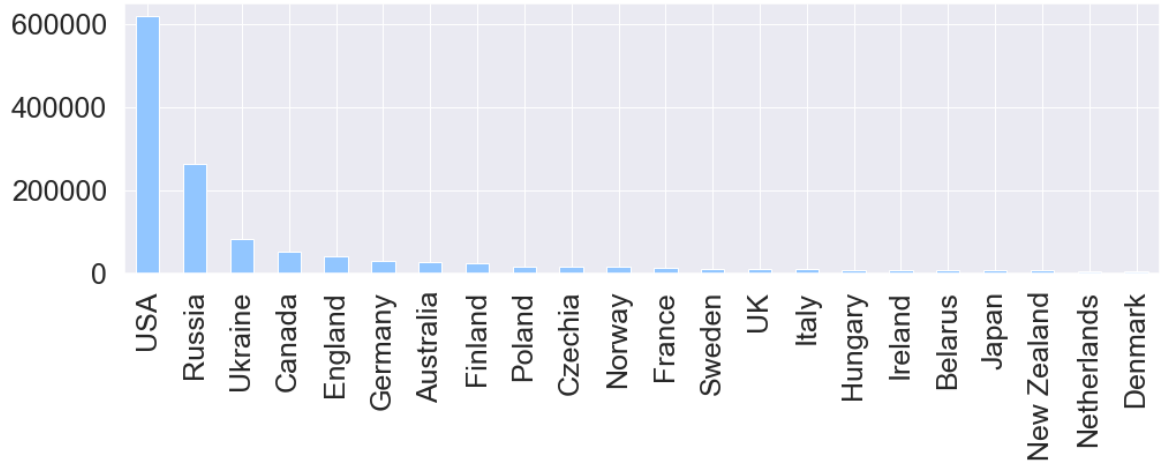



***Figure 2.4.1:*** Deltagende Lande

Det ses at USA er det mest deltagende land, imens Rusland er lige efter. Ukraine er på tredje pladsen og deres version af sporten, samt træningsmetoder blev tit brugt af berømte trænere rundt i verden. Danmark ligger på en nr. 23 plads og indirekte sidstepladsen, da der ikke være brugbart data for de resterende lande.

### 2.4.2 Tested (Non-Doping / Doping)

Variablen hentyder til om federationen håndterer kontrol af deres atleter. Typisk er det op til organisationen, dog er det en relativ stor økonomisk belastning, at kræve ens atleter bliver testet. I Danmark sørger staten for det i gennem Anti-Doping Danmark (ADD). Sammen med Norge og Sverige konkurrere disse tre lande betydeligt mod hinanden ift. andre lande, da kontrollen i andre lande er kan sammenlignes med den nordiske model. Denne Nordiske Alliance holder derfor flest konkurrencer mellem hinanden, da de ser det som mest fair for deres atleter.

### 2.4.3 Federation

Her ses de mest populære federationer i OPL-Datasæt. THSPA, USAPL og FPR tager hhv. 1, 2 og 3-pladsen. Dette ses som forventet, hvis man tager total mængde atleter pr. land i sammenligning. De mere strikse federationer, såsom IPF har i de seneste år indhentet en del, da atleter foretrækker klare regler og minimering af snyd.

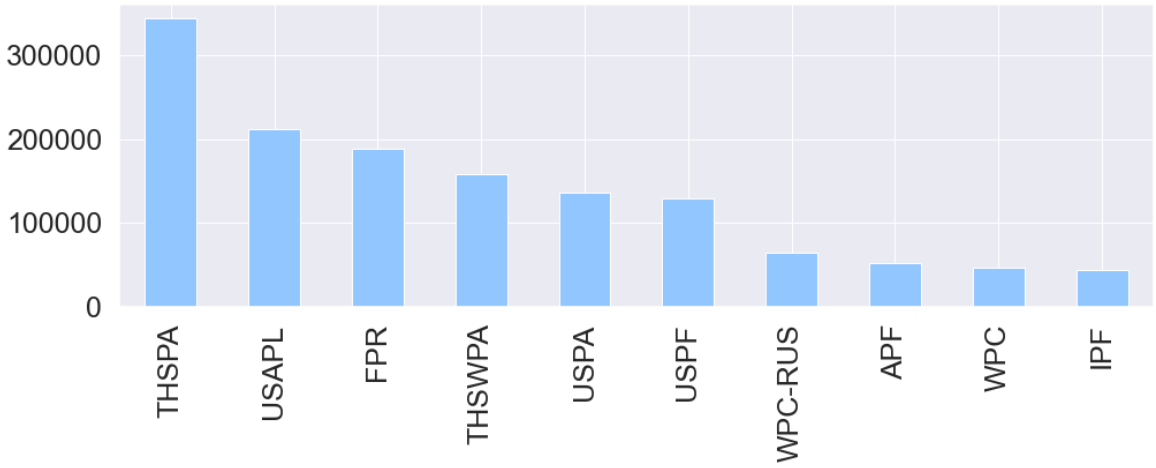



***Figure 2.4.2:*** Mest populære federationer blandt OPL

### 2.4.4 Køn

'Sex' fastsætter hvilken kategori atleten skal deltage i. Der er tre kategorier i OPL-Datasæt: M, F og Mx.
'Mx' vil ikke bruges til videre analyse, da den primære interesse er for Mand eller kvinde. Hvor en tidligere mandlig atlet, som der er skiftet til intetkøn vil have betydelig fordel ift. en kvinde, der skifter til intetkøn. Dette vil gøre en sammenligning besværlig, da man sammenligner miksede resultater.

### 2.4.5 Vægtklasse

'WeightClassKg' er vægtklassen atleten har stillet op til. Den kan enten specificeres som et maksimum eller minimum - eksempelvis kan der stå '90', hvilket betyder atletens maksimale vægt er til og med 90 kilogram. Hvis der stod '90+' ville atleten stiller op med en minimums kropsvægt på 91 kilogram, da skal betegnes som: 91...N Kilogram.

### 2.4.6 Dato

Dato henviser til en bestemt konkurrences tidspunkt. Da Powerlifting er en voksende sportsgren kan nedstående popularitetskurve give et overblik på dens vækst. Det ses at det først var ved 2008 til 2009 sporten fik den eksplosive vækst vi ser i dag. Hvor at der var et stærkt fald i slutningen af 2019 og hen af starten af 2020, som der muligvis kan relateres til Corona og nedlukninger rundt i verden.

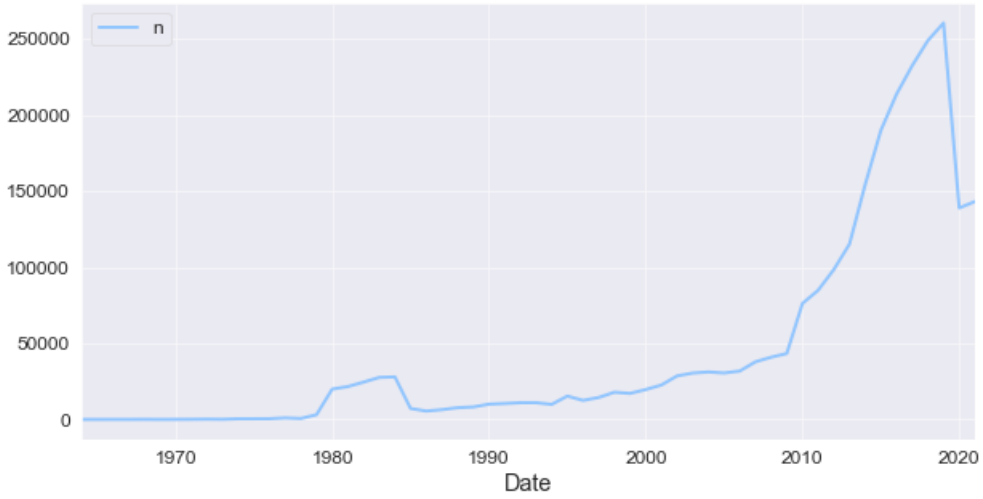



***Figure 2.4.3:*** Popularitetskurve

## 2.5 Afrundning af databeskrivelse

I databeskrivelsen blev datasættet drejet rundt og der blev fundet relevant tendens til videre analyse. Det originale aspekt som opgaven skulle handle om er dog ikke muligt indenfor den begrænsede tid, og en forsnævring af opgaven er uundgåeligt. Dette førte virkede til to primære spørgsmål, som der er interessant at arbejde videre med og passer til solo-omfanget i opgaven.

- Kan man forudse atletens vægt ved brug af deres maksimale løft i hhv. squat, bænkpres og dødløft?
- Er der hold i federationernes erklæring og behandling af andre federationer, som ikke følger samme regler som dem med fokus på doping-politik?

Det vil derfor også være interessant med et korrelations heatmap for tjek af variablerne, samt en videre datanalyse af mænd og kvinders fysiske træk i Powerlifting.

# 3 Problemanalyse

# 4 Data Exploration

# 5 Data analysis

# Appendices