- The total length should not exceed 8 pages.

Evaluation

Your report will be evaluated based on:

- (format) adherence to formatting and appearance guidelines;
- (clarity) clarity and thoughtfulness in written voice;
- (accuracy) apparent accuracy of quantitative results and technical information;
- (applied a PSTAT100 technique) successful use of one or more techniques in the course.

Notice that no credit is tied to the nature of the results; you can earn credit equally well with an analysis that says little as with one that says a lot. **Negative, neutral, or ambiguous results -- analyses that do not produce any particular insight -- are more than acceptable.** If your analyses turn in one of these directions, present them as clearly as possible, and consider speculating in your discussion section about the absence of signficicant/interpretable findings.

Analysis on USAPL powerlifters using only raw equipment the United States

Peter Chu, Mac Kul

Author contributions

Peter Chu contributed most of the coding in addition to interpretations of the analysis

Mac Kul contributed most of the writings and coding and interpretation of the correlation matrix

Abstract

Prepare an abstract *after* you've written the entire report. The abstract should be 4-6 sentences summarizing the report contents. Typically:

- the first 1-2 sentences introduce and motivate the topic;
- the next 1-2 sentences state the aims;
- the next 1-2 sentences state the findings.

This project was done using a dataset from openpowerlifting.org on variables of interest. This was done to answer the questions of which state had the highest average Dots score over 450 and whether men or women had a higher average Dots score. From our analysis we concluded that New York had the highest average Dots score over 450 and that on average men had a higher Dots score than women over the past 5 years.

Introduction and Background

Powerlifting is an individual sport where the goal is to lift the most amount of weight possible in squat, bench press, and deadlift in a specific weight and age category. Over the years, powerlifting has grown tremendously across the U.S, and even other countries. Our aim for this project is to study the progression of athlete's totals with contributing factors such as age, time, sex, etc in America.

Our main motivation for choosing this data is because we have also competed in Powerlifting. We believe that having domain knowledge in this specific topic will give us leverage with data interpretation, which then helps us answer questions that we are interested in. Aside from our personal involvement with the sport, we believe that the data itself has a lot of variables to analyze.

Aims

Our project aim was to do an in-depth analysis on powerlifting across the United States. Our two main questions were which state has the highest average Dots score over 450, and how to average male and female Dots score compare over the years 2017 - 2022. Our second question is of high interest as the time-old debate of whether men or womer are stronger is still alive today. To answer these questions we subsetted our data for varialbes of interest before numerically answering our questions

---

Materials and methods

The goal of this section is to describe your dataset(s) and sketch out your analysis.

Datasets

We obtained our data from openpowerlifting.org, the main database for all of the competition records. The data we have has the name, age, weight, weight lifted, and the federation scores (DOTS), country, and state of each atheletes. Since the data was retrieved from the official database, the data was obtained through direct submissions from meet directors right after competitions. This data is generalizable to our targetted population, which are powerlifting atheletes.

Variable name	Description	Type	Units of measurement
`Equipment`	`Type of equipment used for each lift`	`str`	`None`
`Event`	`The type of event done at the powerlifting competition`	`str`	`None`
`Name`	`The name of the person (observation)`	`str`	`None`

Variable name	Description	Type	Units of measurement
Sex	The registered gender / sex of the person	str	None
Date	The date when the data was collected	str	None
Division	The weightclass of the person	str	None
State	The State from which the person was born	str	None
WeightClassKg	The weightclass of the person	str	None
Age	The age of the person at the time the data was recorded	Numeric	Years
AgeClass	The age range in which the person is categorized into	str	Years
Best3SquatKg	The best successfull squat done by the person out of 3 tries	Numeric	Kilograms
Best3BenchKg	The best successfull bench press done by the person out of 3 tries	Numeric	Kilograms
Best3DeadlifeKg	The best successfull deadlift done by the person out of 3 tries	Numeric	Kilograms
TotalKg	The total sum of all 3 best lifts	Numeric	Kilograms
Dots	The score calculated by a formula facotring in the person weight aand total weight lifted	Numeric	Points
Tested	Indicator of whether the person was drug tested at the compeitition or not	str	None
Country	The country from which the person is from	str	None
ParentFederation	The powerlifting federation which hosted the compeitition in which the person participated in	str	None

Methods

During the exploratory analysis phase, a correlation matrix was created to see if there were any clear correlations between DOTs scores and other variables. From the coefficients, it was expected that squat, bench, and deadlift numbers are positively correlated to DOTs, because they directly go into the DOTs score calculation formula. Squat and Deadlift share correlation coefficients of 0.77, while bench has a correlation coeffient of 0.67. This intuitively makes sense because the leg and the posterior chain muscles are generally more powerful than the upper body that the bench press movement utilize. In addition, deadlift is also the most correlated to the overall total amount lifter because lifters can usually lift the highest amount with this movement compared to the other two. In addition, the deadlift is the last movement of the competition, meaning that the the lifter had most likely completed the first two movements without failing at least their first attempt. We also used the method of simple linear regression

on Age to see if it could predict Dots scores over 450. This was done following the taught methods by finding residuals, fitted values, and plotting them. The intercept was calculated to be 313.21287349 and our slope coefficient estimate to be 40.96208018

Results

This section should show your results. You can include sub-header structure as suits your aims (or not).

The easiest way to compose this section is to structure it around your figures and tables. Prepare and input your figures and tables in the order you'd like them to appear, and then draft the text. Move through the figures and tables in sequence, and for each one:

- introduce the figure/table;
- describe what it shows;
- and then describe what you see (its significance). Keep the latter brief; you'll have an opportunity to offer more nuanced/extended commentary in the discussion section.

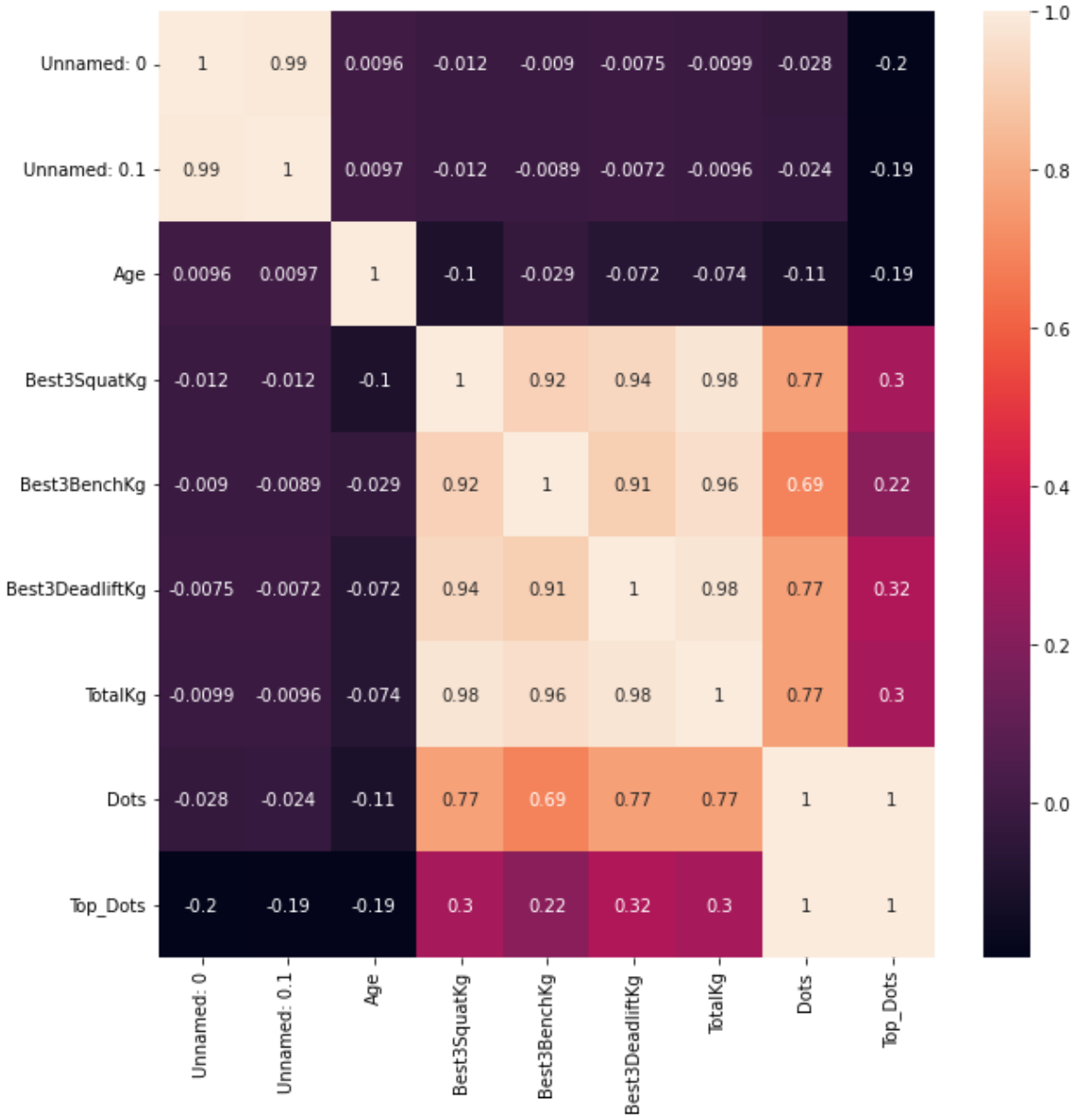
Discussion

This section should conclude your report in 1-2 paragraphs that reiterate the findings and offer any commentary. 'Commentary' could include:

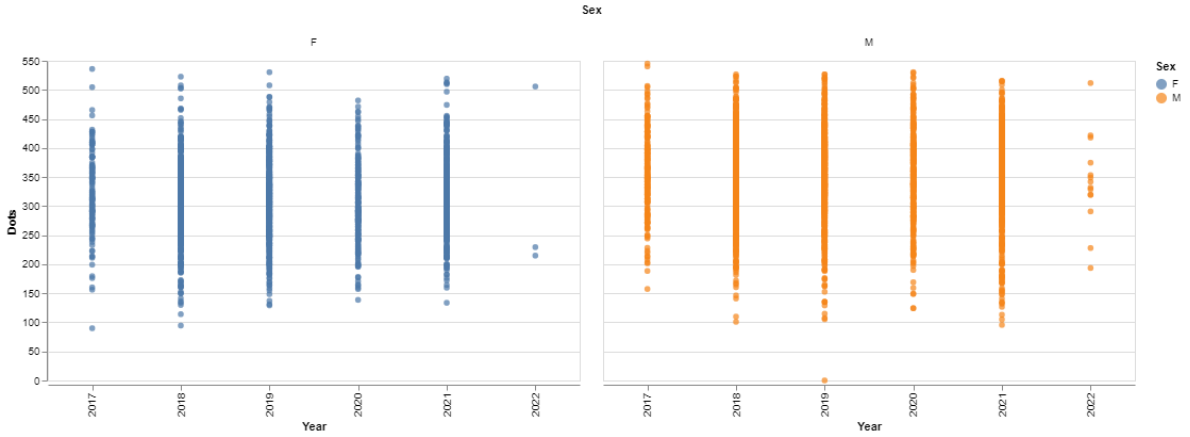
- speculation about the cause of certain findings;
- caveats about interpretation;
- refining of questions or aims;
- further topics you would have liked to explore.

Our project looked at a multitude of variables from powerlifting competitions over the years 2017 to 2022. The analysis focusses on whether men or women are stronger at the top level according the Dots metric. Since Dots takes into account gender, it is not a hiderance to our analysis as it has been standardized. Thus we can not argue that women will automatically have a higher Dots score because of their gender.

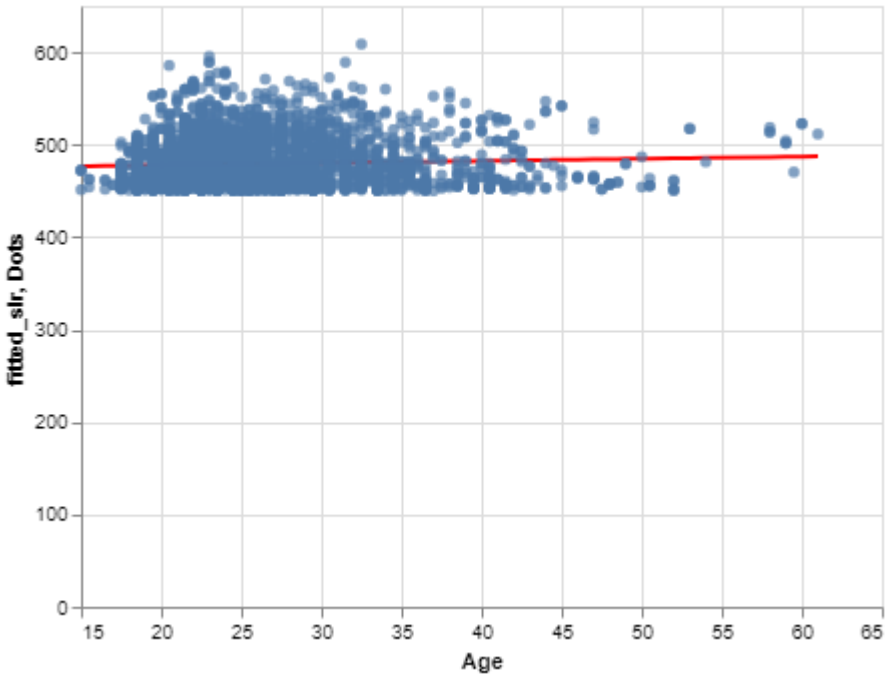
For our first question of aim we looked at which state has the highest average Dots score over 450, which was New York. Some underlying reasons as to why this may be is because as people get stronger they become more recognized. Similar to Los Angeles, New York is a hotspot for celebrities which could explain why so many of the strongest powerlifters reside there. Another reason may be that since New York has the 4th highest state population and may have more gyms catered to their needs. Since they are top-level athletes they may not goto commercial gyms which New York has a lot of. Having more data on relevant aforementioned topics could enable us to do a deeper analysis on why New Yrk has the highest average Dots score over 450.



Our second question of aim looked at whether male or females had a higher average dot score over the years. Our analysis concluded that over the past 5 years, men had a higher average Dots score by around 40 points every year. This may be due to the fact that there is a 2-1 ratio of men to women who compete in powerlifting competitions, thus the average for women is lower. As stated before since Dots takes into account gender, biological reasons do not have a heavy impact as to why this may be. Thus the result of analysis should not be taken as absolute due to the smaller amount of observations for women. Another significant result from our analysis is that over time the Dots score for women tended to go up, while for men the score was stagnant around the 355 mark. More time and more observations for women could result in different conclusions, but we do not have this data for this project.



Our third question of aim was whether or not Age could predict the Dots score over 450. From our analysis, it appeared that Age does not do a very "good" job at predicting this as our regression line was almost horizontal. However, this makes sense as the Dots formula also takes into account age. Similar to gender, this means that despite the fact that people get weaker the older they get, the formula adusts for this. Thus while people do get weaker the older they get in terms of absolute strength, relative strength appears the same for these individuals our analysi was done on.



Further topics we would've liked to explore were what results our analysis would yielded for observations on different countries and federations. However, there is simply too much data too analyze for this project. Overall while we were limited to only the United States, the results are still practical as we are both active competers in this sport in the United States. Thus analysis on powerlifters in Russia would not apply to us, but would be interesting to do so.

Appendix

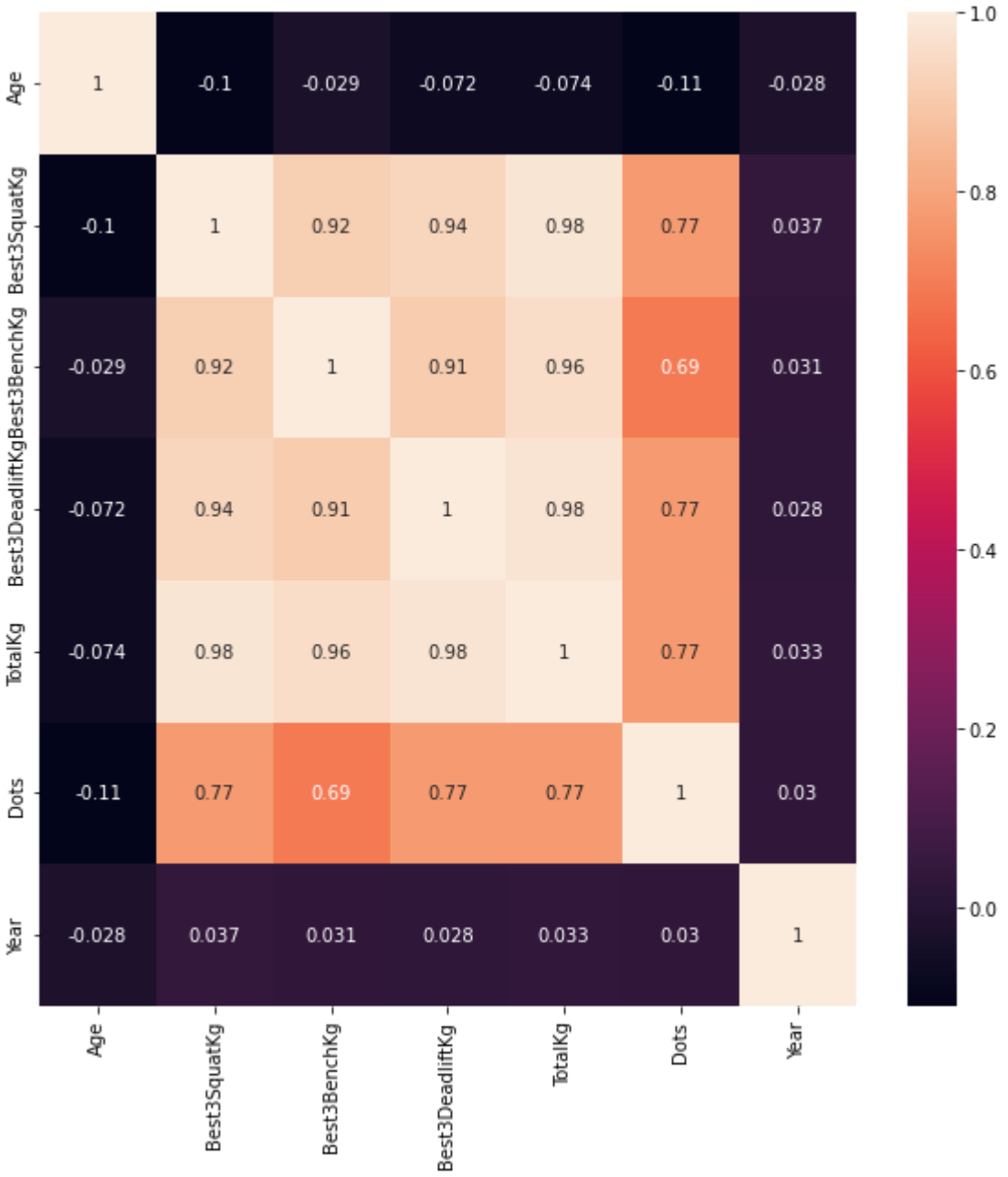
In [1]:
```
#Import necessary stuff
import pandas as pd
import numpy as np
import altair as alt
import seaborn as sns
import random
from random import sample
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
from sklearn.preprocessing import add_dummy_feature
import matplotlib.pyplot as plt

#Read in dataset
data = pd.read_csv('TidiedPLData.csv')
```

In [2]:
```
#Tidy dataset
untidy = data.drop(columns = 'Unnamed: 0')
list1 = list(range(0,69610))
list2 = list(range(0,69610))

data2 = untidy.set_index('Sex').drop('Mx').reset_index()
```

In [3]:
```
#Create new variable of year rather than full date
nsim = 69610

temp = np.zeros(nsim)

for i in range (0,69610):
  temp[i] = int(data2['Date'][i][0:4])

temp1 = temp.tolist()

data2['Year'] = temp1
```

In [4]:
```
#Random seed to get 5000 observations as this is the max number of plottable points ir
random.seed(117) #117, 51, is GOOD
rand_samp = random.sample(list2,5000)

#Set new var so we don't have to type out the whole thing everytime
data3 = data2.iloc[rand_samp,:]
```

In [5]:
```
corrMatrix = data2.corr()
fig, ax = plt.subplots(figsize=(10,10))
sns.heatmap(corrMatrix, annot=True)

plt.show()
```



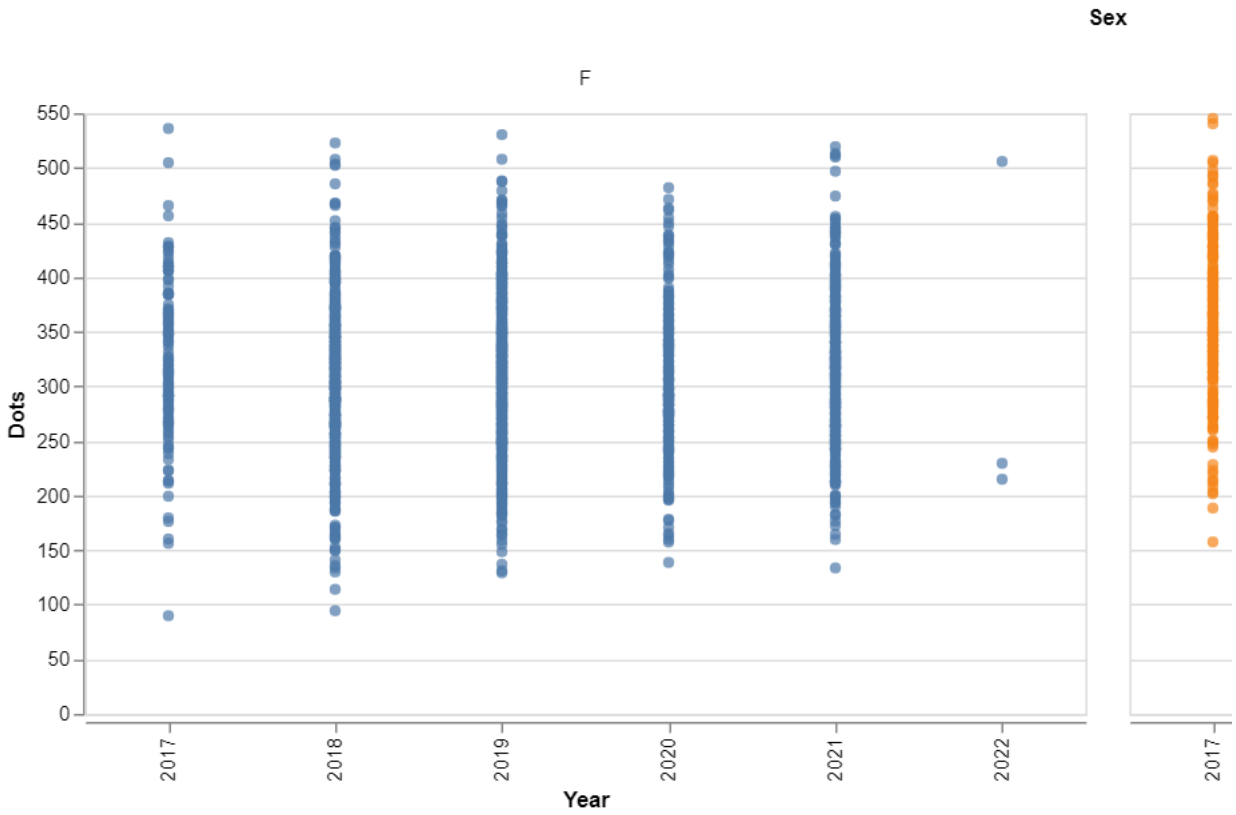
In [6]:
```
grouped_single = data2.groupby(['Sex','Year']).agg({'Dots':['mean']})

grouped_single
```

Out[6]:

		Dots
		mean
Sex	Year	
F	2017.0	317.686932
	2018.0	309.244968
	2019.0	312.555496
	2020.0	315.329202
	2021.0	317.906787
	2022.0	321.553614
M	2017.0	359.371386
	2018.0	351.745458
	2019.0	353.017546
	2020.0	354.652028
	2021.0	357.568808
	2022.0	348.040570

In [7]:
```
#Male vs Female dots by Year plotted next to each other for 5000 random observations

alt.Chart(data3).mark_circle().encode(
    x = alt.X('Year:N'),
    y = alt.Y('Dots'),
    color = 'Sex'
).properties(width = 500).facet(column = 'Sex')
```

Out[7]:



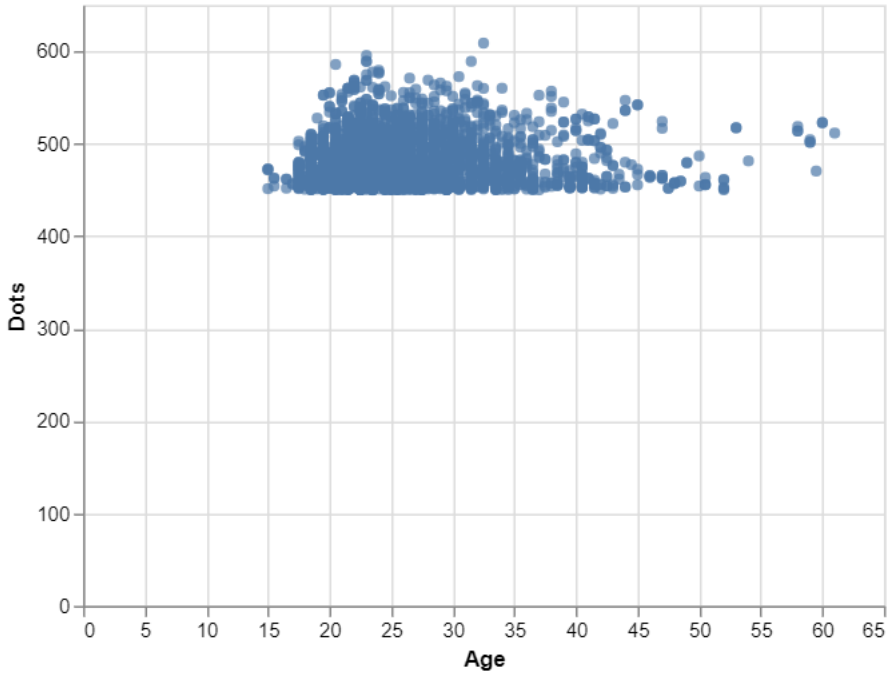
In [8]:
```
#New dataframe
data4 = data2[data2['Dots'] > 450]
data4.head()
```

Out[8]:

	Sex	Name	Country	State	WeightClassKg	Division	Age	AgeClass	ParentFederation	Date
68	M	Owen Ward	USA	FL	105	MR-O	27.0	24-34	IPF	2022-01-29
85	M	Jerry Chavez	USA	AZ	120+	MR-O	25.0	24-34	IPF	2022-05-14
104	F	Heather Connor	USA	NC	47	FR-O	30.0	24-34	IPF	2022-04-01
106	F	Marisa Inda	USA	CA	52	FR-O	45.0	45-49	IPF	2022-04-01
109	F	Meghan Scanlon	USA	MA	63	FR-O	34.0	24-34	IPF	2022-04-01

In [9]:
```
reg_data = data4.loc[:,['Age','Dots']]

simporig = alt.Chart(reg_data).mark_circle().encode(
    x = alt.X('Age'),
    y = alt.Y('Dots')
)

simporig
```

Out[9]:



In [10]:

```
y = reg_data.Dots

x_slr_df = reg_data.loc[:, ['Age']]

# add intercept column
x_slr = add_dummy_feature(x_slr_df, value = 1)

slr = LinearRegression(fit_intercept = False)

# fit model
slr.fit(x_slr, y)

coef = slr.coef_

coef

#BETA 0 = 313.21287349, BETA 1 = 40.96208018

fitted_slr = slr.predict(x_slr)

fitted_slr.shape

resid_slr = y - fitted_slr
reg_data['fitted_slr'] = fitted_slr
reg_data['resid_slr'] = resid_slr

n, p = x_slr.shape

# compute estimate of error variance
sigma2_hat = ((n - 1)/(n - p)) * resid_slr.var()

xtx_slr = x_slr.transpose().dot(x_slr)

# compute matrix of parameter variances/covariances: estimated error variance x (X'X)^
slrcoef_vcov = np.linalg.inv(xtx_slr) * sigma2_hat

# print
```

```
slrcoef_vcov

# take square root of matrix diagonals to get standard errors
slrcoef_se = np.sqrt(slrcoef_vcov.diagonal())

# print
slrcoef_se

slrcoef_table = pd.DataFrame(
    data = {'coefficient estimate': slr.coef_, 'standard error': slrcoef_se},
    index= ['intercept', 'Age']
)

# print
slrcoef_table

r2_score(reg_data.Age, reg_data.fitted_slr)
slr_line = simporig.mark_line(color = 'red').encode(y = 'fitted_slr')

slr_line + simporig
```

Out[10]: