*

Filling the Bins
- or -
Turning Numerical Data into Histograms

ISC1057
Janet Peterson and John Burkardt
Computational Thinking
Fall Semester 2016

A histogram is a kind of chart which shows patterns in sets of numeric data.

Strictly speaking, a histogram is different from a bar chart, because a bar chart is based on categories: the number of times each baseball team has won a World Series, for instance. The categories are given, and we just have to assign a number to each one.

For a histogram, we might have measurements of the circumference of 1,000 trees on campus, that is, just series of 1,000 numbers. These numbers might all be different. If we made a bar chart of this data, then we would end up with 1,000 little bars, one for each measurement.

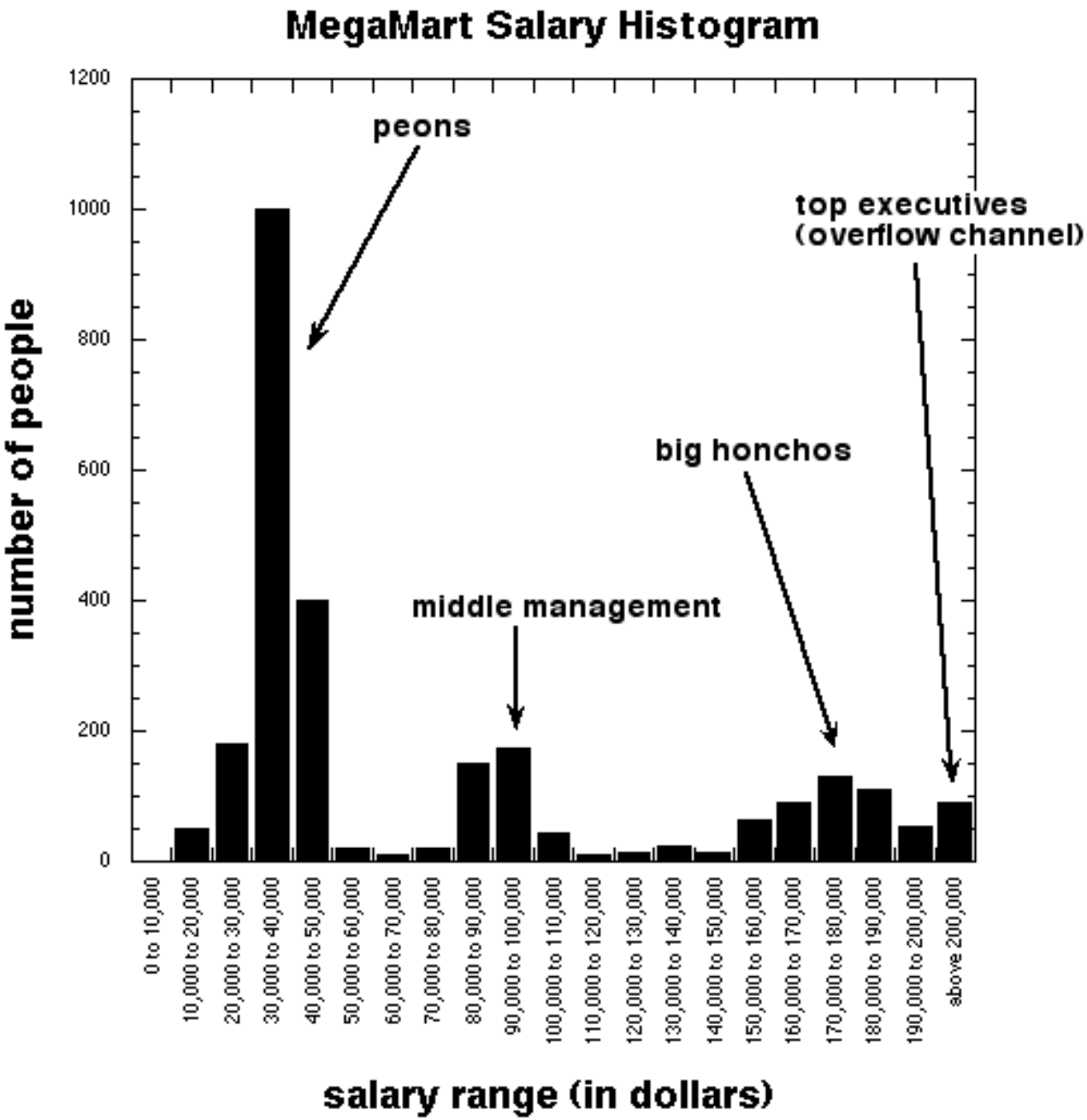
But if the numbers are measurements, then it might make sense to group them together, separating small, medium and large numbers, or, more carefully, dividing the values into ranges, or bins.

The size and number of bins we use will have a significant effect on whether we can see patterns in the data. We don't want every number in a separate bin, or all the numbers in a single bin. We are looking for natural patterns that the numbers arrange themsselve into when we give them a "reasonable" number of bins to choose from.

Sample data:

Employee Number	Position	Annual Salary
1	"Clerk"	32,000
2	"Stocker"	17,125
3	"Senior Manager"	143,000
...	more data	—
2108	"Vice President for Sales"	245,000
2109	"Cleaner"	15,125
2110	"Data Analyst"	74,000

It's possible that no two people in this company earn the same salary. What it we consider salary ranges of $10,000, and count the number of employees in each range?


MegaMart Salary Histogram
number of people
1200
1000
800
600
400
200
0
peons
top executives
(overflow channel)
big honchos
middle management
0 to 10,000
10,000 to 20,000
20,000 to 30,000
30,000 to 40,000
40,000 to 50,000
50,000 to 60,000
60,000 to 70,000
70,000 to 80,000
80,000 to 90,000
90,000 to 100,000
100,000 to 110,000
110,000 to 120,000
120,000 to 130,000
130,000 to 140,000
140,000 to 150,000
150,000 to 160,000
160,000 to 170,000
170,000 to 180,000
180,000 to 190,000
190,000 to 200,000
above 200,000
salary range (in dollars)

In a histogram, it's usual to make the bin widths or ranges the same. That means we still have to decide how many bins to use, and so sometimes a little experimentation is necessary.

However, just as with bar charts, if the data is properly plotted, we may find that it is:

- pretty much flat or even;
- arranged like a staircase (mostly up, or mostly down);
- arranged like a hill (up, then down);
- arranged like a valley (down, then up).
- having several peak values separated by valleys *(as in our salary data on the previous slide)*;
- seemingly random (up and down with no pattern);

A histogram may summarize or illustrate our beliefs, or it may be evidence that asks us to think of an explanation.

Example #1: 200 Height and Weight Measurements

It's a common practice to collect height and weight measurements.

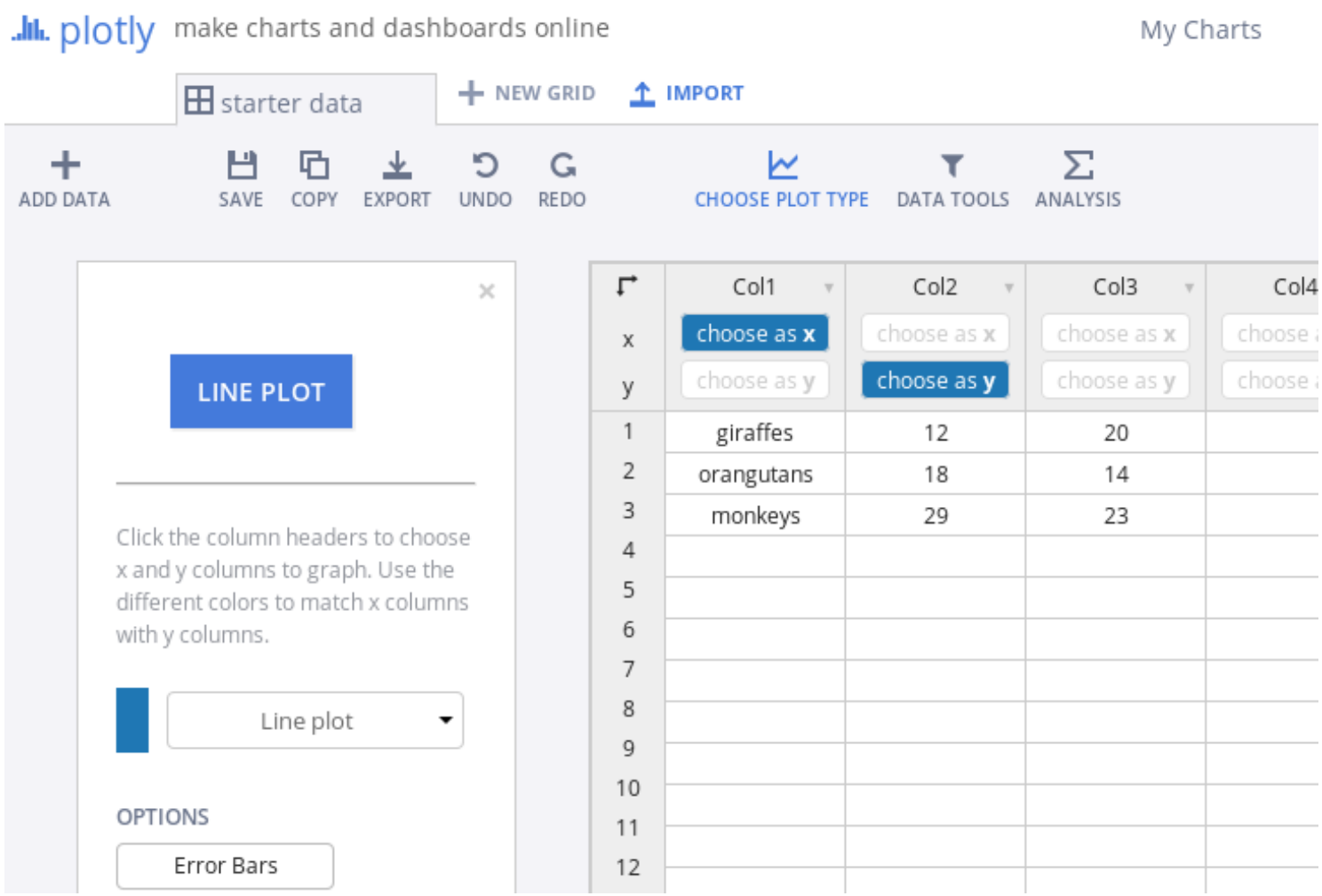
We assume there is an average height, with some range of normal variation. But are there more short people than tall ones?

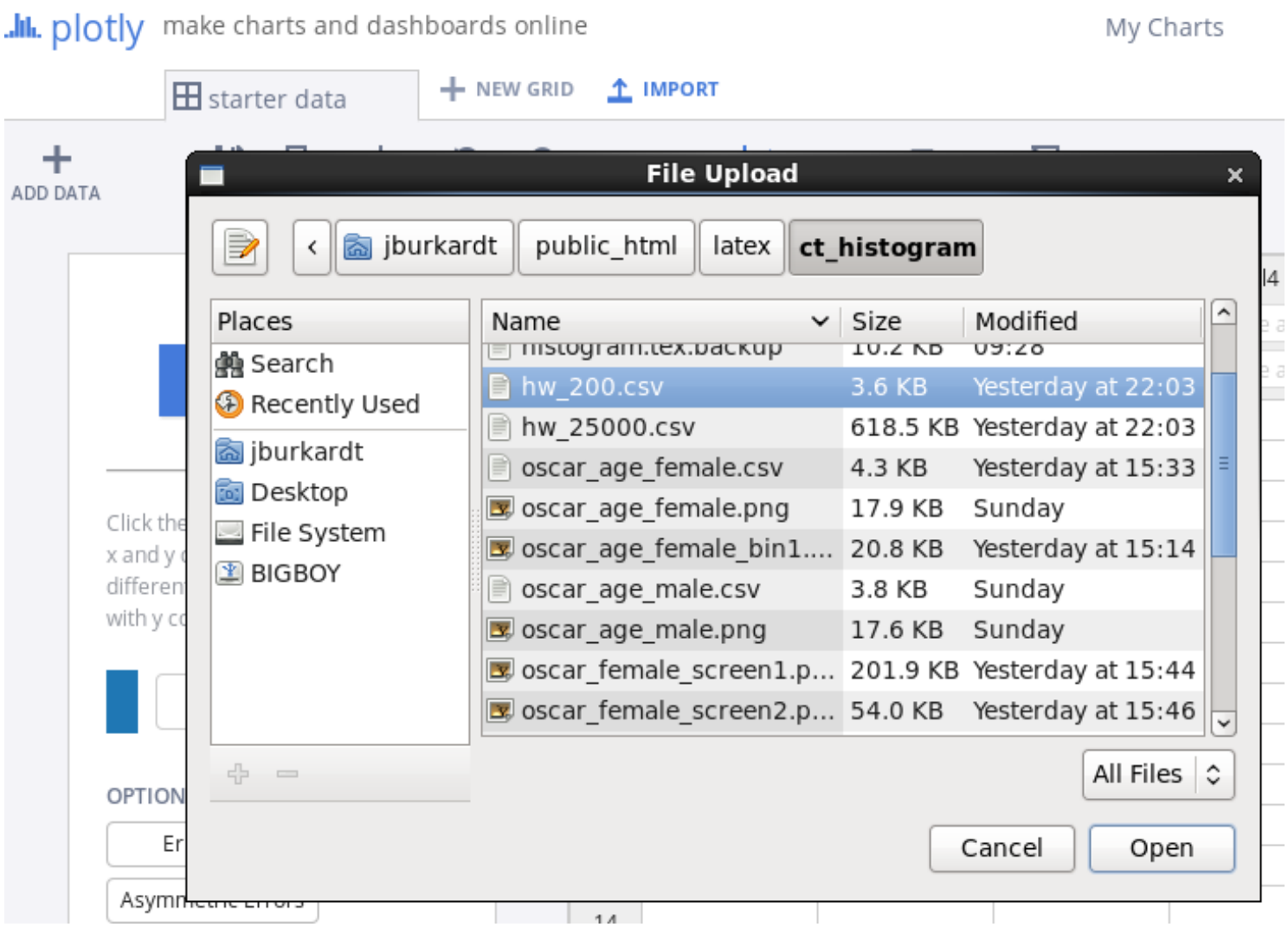
It's not so easy to guess someone's weight, and the range of variation is probably much greater than for height.

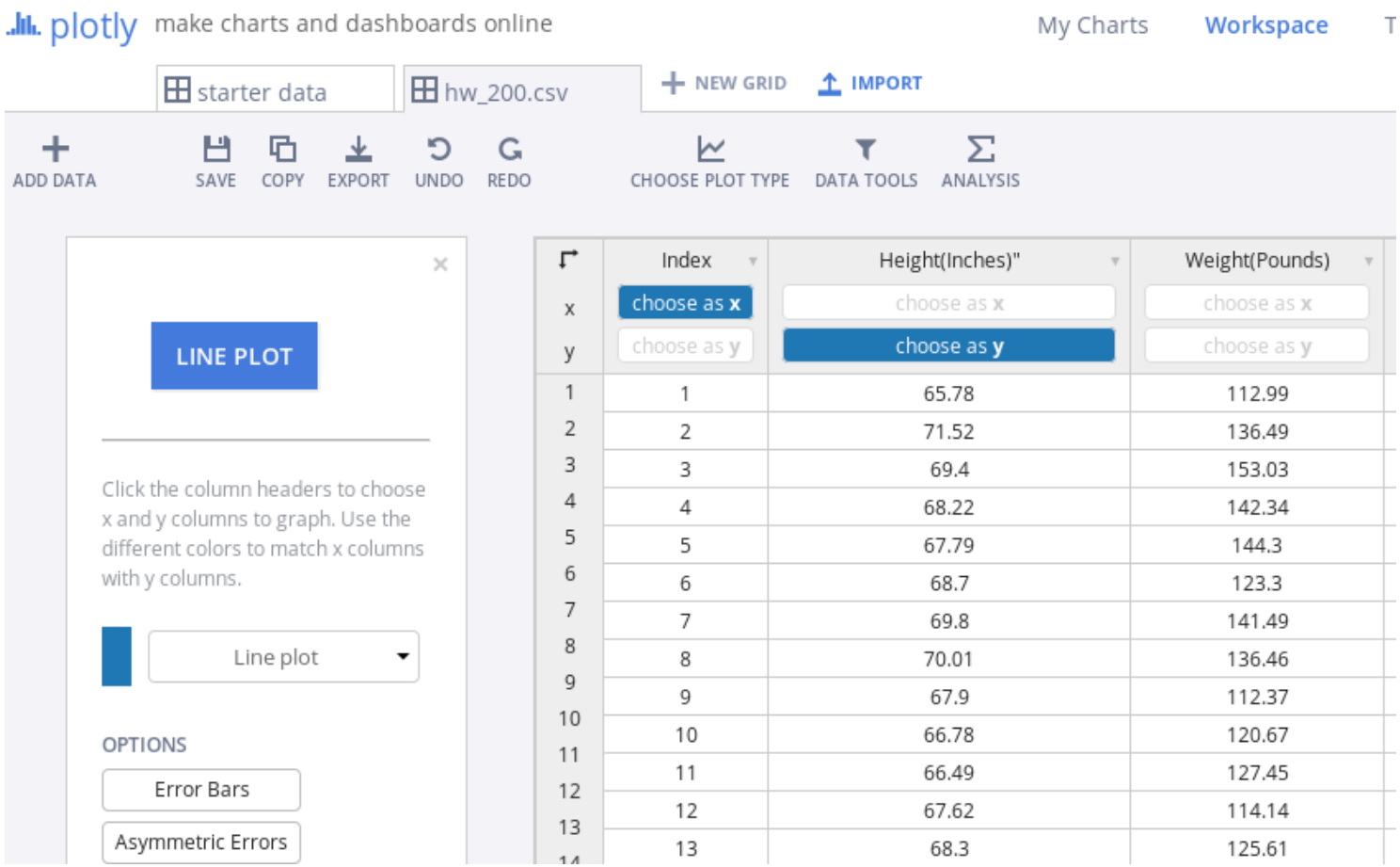
Because both measurements are numbers that are scattered over a numeric range, a histogram can be a good tool to find patterns.

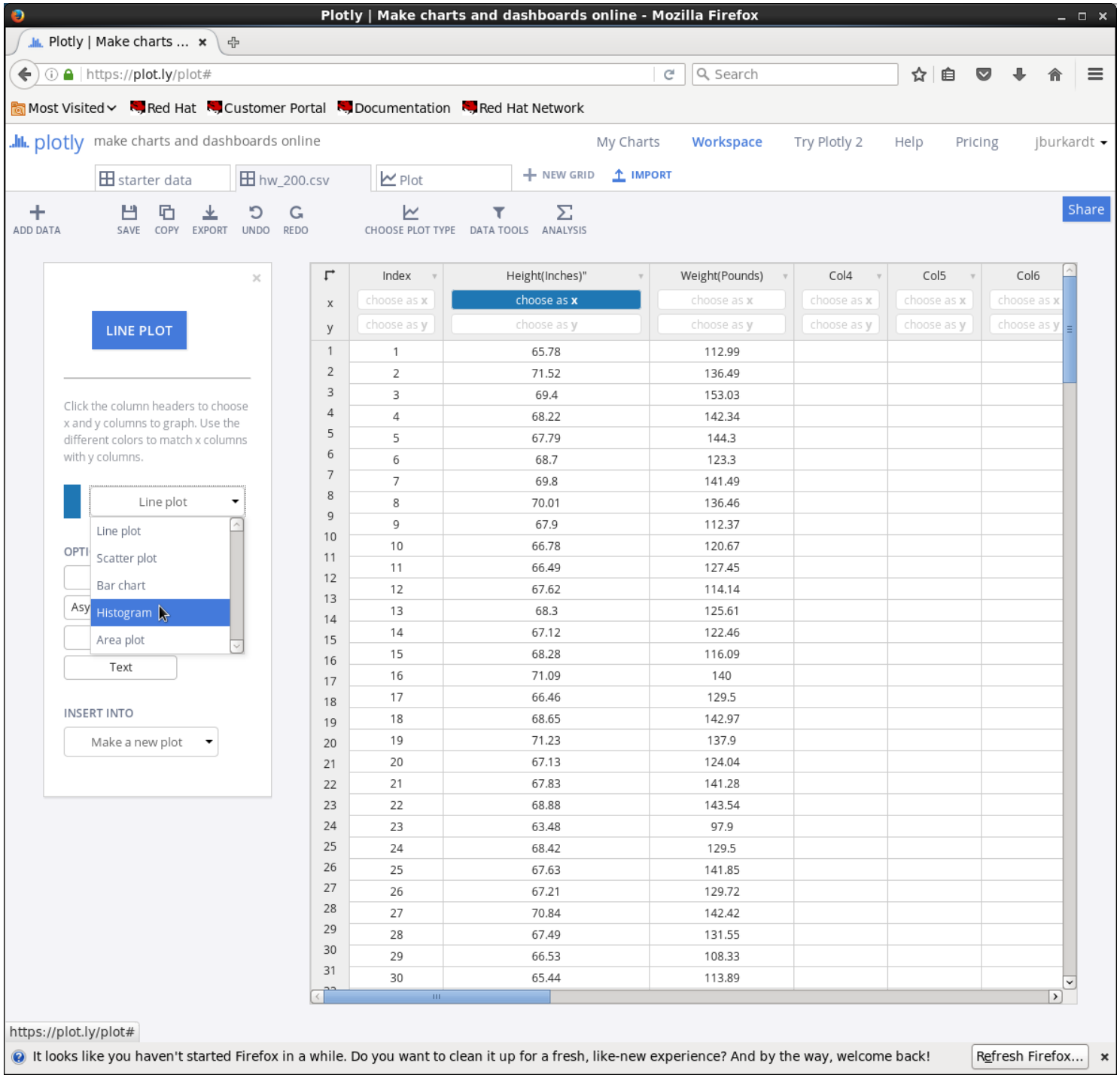
We have 200 height and weight measurements in the file **hw_200.csv**.

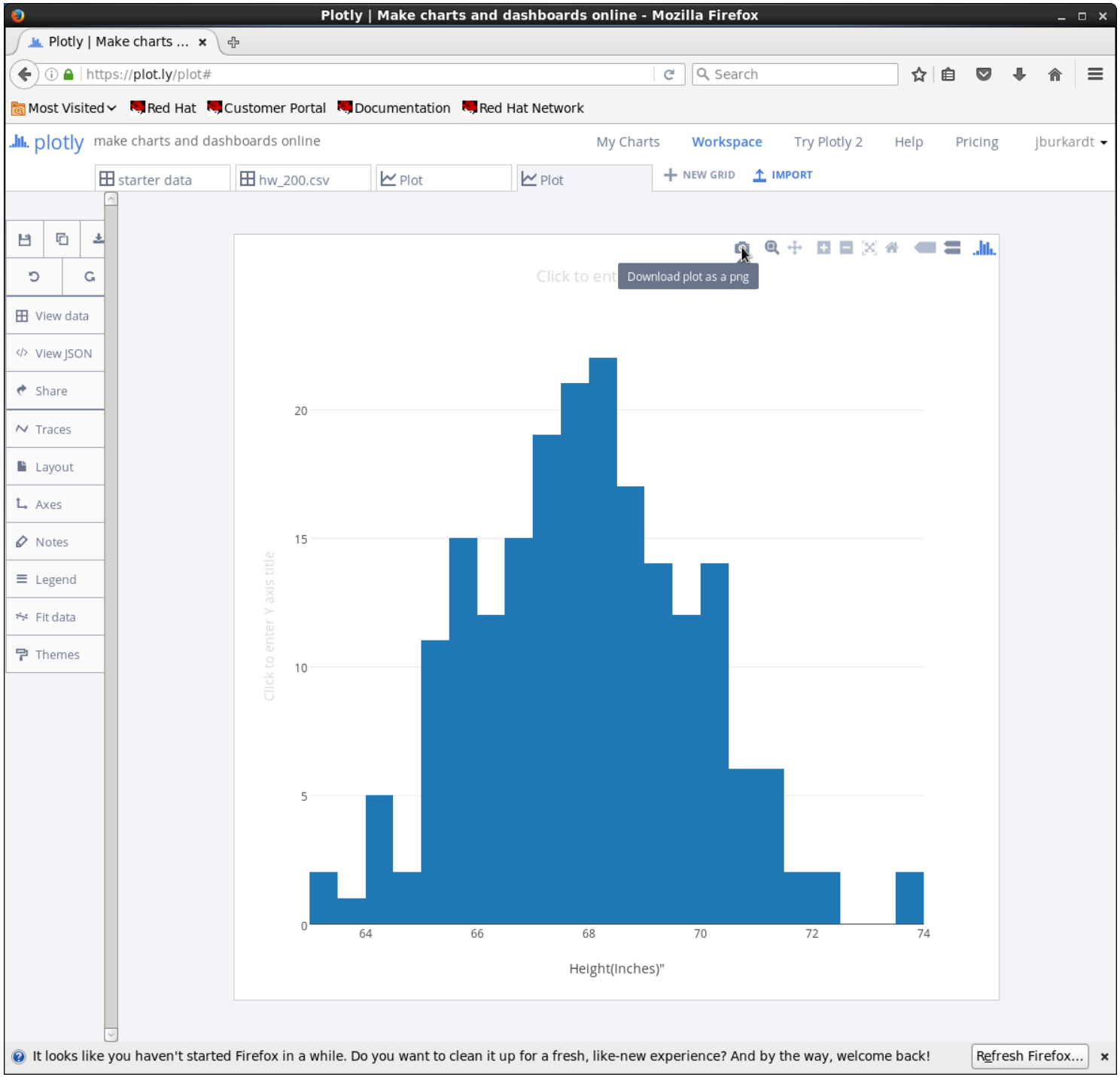
"Index",	"Height (inches)",	"Weight (pounds)"
1,	65.78,	112.99
2,	71.52,	136.49
3,	69.40,	153.03
—	*more data*	—
198,	68.24,	128.30
199,	68.02,	127.47
200,	71.39,	127.88

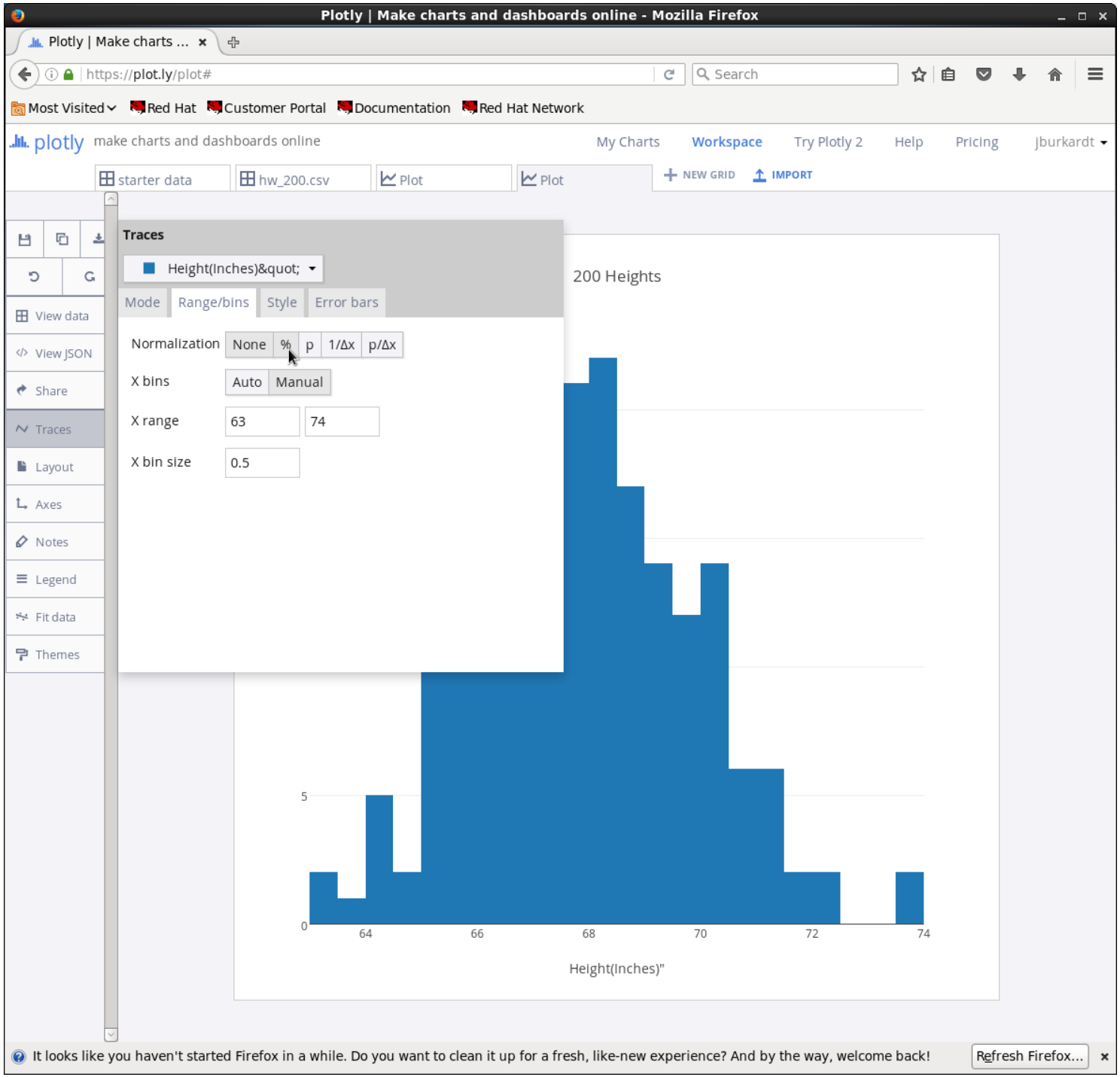

plotly
make charts and dashboards online
My Charts
starter data
NEW GRID
IMPORT
ADD DATA
SAVE
COPY
EXPORT
UNDO
REDO
CHOOSE PLOT TYPE
DATA TOOLS
ANALYSIS
LINE PLOT
Click the column headers to choose x and y columns to graph. Use the different colors to match x columns with y columns.
Line plot
OPTIONS
Error Bars
Col1
Col2
Col3
Col4
x
y
choose as x
choose as y
1 giraffes 12 20
2 orangutans 18 14
3 monkeys 29 23
4
5
6
7
8
9
10
11
12

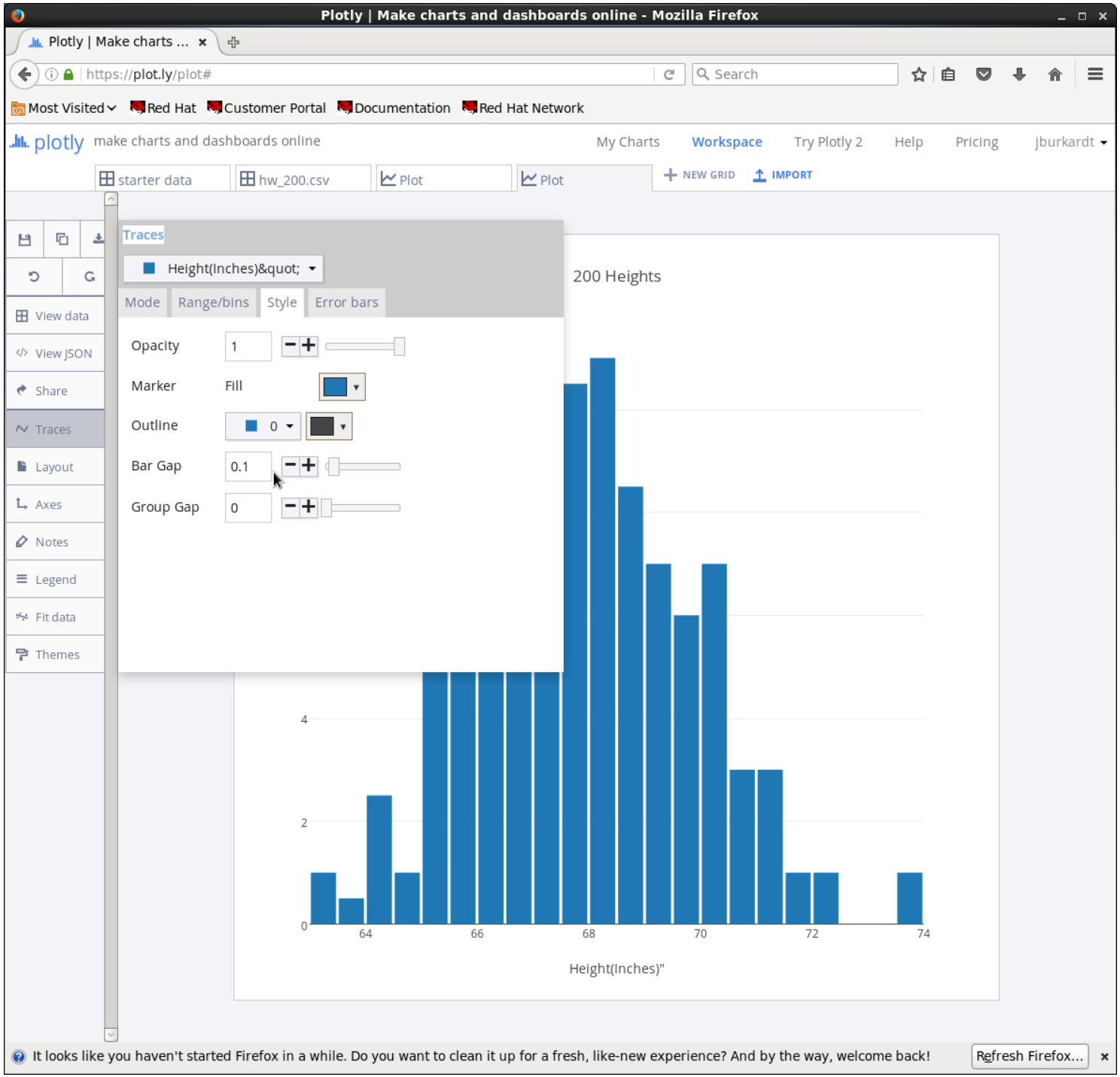

plotly
make charts and dashboards online
My Charts
starter data
NEW GRID
IMPORT
ADD DATA
File Upload
jburkardt
public_html
latex
ct_histogram
Places
Search
Recently Used
jburkardt
Desktop
File System
BIGBOY
Name
Size
Modified
hw_200.csv 3.6 KB Yesterday at 22:03
hw_25000.csv 618.5 KB Yesterday at 22:03
oscar_age_female.csv 4.3 KB Yesterday at 15:33
oscar_age_female.png 17.9 KB Sunday
oscar_age_female_bin1.... 20.8 KB Yesterday at 15:14
oscar_age_male.csv 3.8 KB Sunday
oscar_age_male.png 17.6 KB Sunday
oscar_female_screen1.p... 201.9 KB Yesterday at 15:44
oscar_female_screen2.p... 54.0 KB Yesterday at 15:46
All Files
Cancel
Open

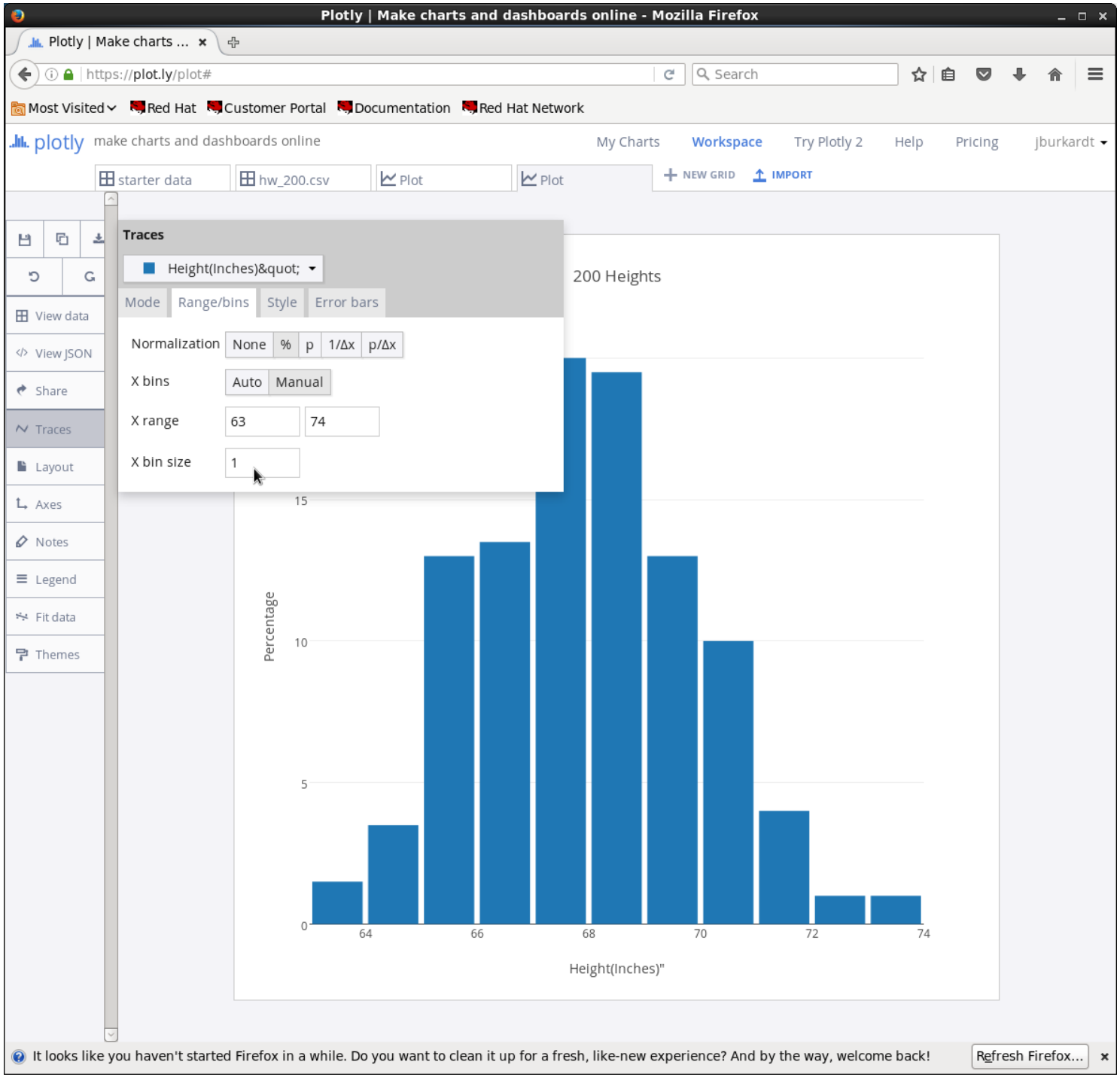

plotly
make charts and dashboards online
My Charts
Workspace
starter data
hw_200.csv
NEW GRID
IMPORT
ADD DATA
SAVE
COPY
EXPORT
UNDO
REDO
CHOOSE PLOT TYPE
DATA TOOLS
ANALYSIS
LINE PLOT
Click the column headers to choose x and y columns to graph. Use the different colors to match x columns with y columns.
Line plot
OPTIONS
Error Bars
Asymmetric Errors
Index
Height(Inches)"
Weight(Pounds)
x
y
choose as x
choose as y
1 1 65.78 112.99
2 2 71.52 136.49
3 3 69.4 153.03
4 4 68.22 142.34
5 5 67.79 144.3
6 6 68.7 123.3
7 7 69.8 141.49
8 8 70.01 136.46
9 9 67.9 112.37
10 10 66.78 120.67
11 11 66.49 127.45
12 12 67.62 114.14
13 13 68.3 125.61

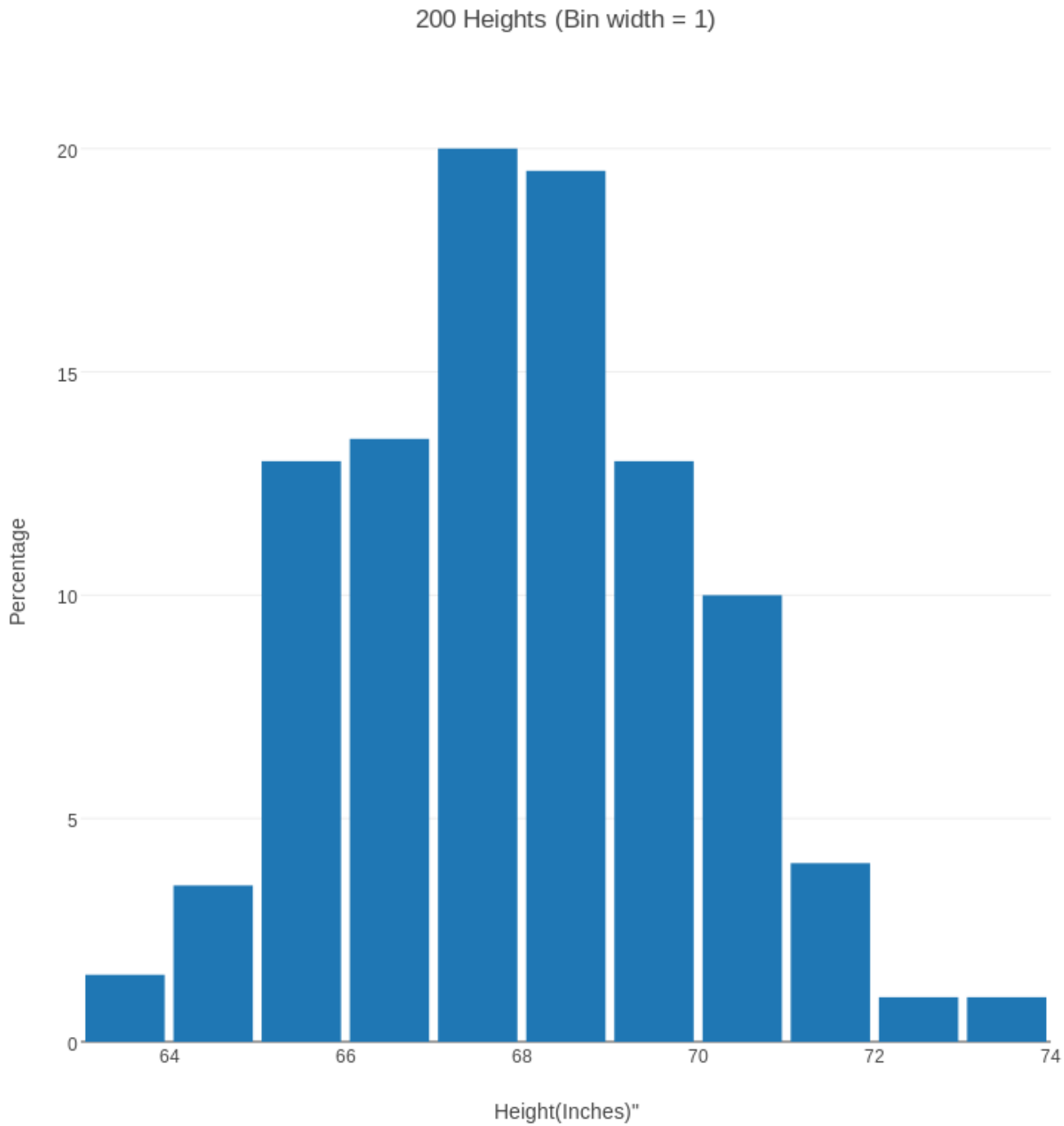

Plotly | Make charts and dashboards online - Mozilla Firefox
Plotly | Make charts ...
https://plot.ly/plot#
Search
Most Visited
Red Hat
Customer Portal
Documentation
Red Hat Network
plotly
make charts and dashboards online
My Charts
Workspace
Try Plotly 2
Help
Pricing
jburkardt
starter data
hw_200.csv
Plot
NEW GRID
IMPORT
ADD DATA
SAVE
COPY
EXPORT
UNDO
REDO
CHOOSE PLOT TYPE
DATA TOOLS
ANALYSIS
Share
LINE PLOT
Click the column headers to choose x and y columns to graph. Use the different colors to match x columns with y columns.
Line plot
Line plot
Scatter plot
Bar chart
Histogram
Area plot
Text
INSERT INTO
Make a new plot
Index
Height(Inches)"
Weight(Pounds)
Col4
Col5
Col6
choose as x
choose as y
1 1 65.78 112.99
2 2 71.52 136.49
3 3 69.4 153.03
4 4 68.22 142.34
5 5 67.79 144.3
6 6 68.7 123.3
7 7 69.8 141.49
8 8 70.01 136.46
9 9 67.9 112.37
10 10 66.78 120.67
11 11 66.49 127.45
12 12 67.62 114.14
13 13 68.3 125.61
14 14 67.12 122.46
15 15 68.28 116.09
16 16 71.09 140
17 17 66.46 129.5
18 18 68.65 142.97
19 19 71.23 137.9
20 20 67.13 124.04
21 21 67.83 141.28
22 22 68.88 143.54
23 23 63.48 97.9
24 24 68.42 129.5
25 25 67.63 141.85
26 26 67.21 129.72
27 27 70.84 142.42
28 28 67.49 131.55
29 29 66.53 108.33
30 30 65.44 113.89
https://plot.ly/plot#
It looks like you haven't started Firefox in a while. Do you want to clean it up for a fresh, like-new experience? And by the way, welcome back!
Refresh Firefox...


Plotly | Make charts and dashboards online - Mozilla Firefox
Plotly | Make charts ...
https://plot.ly/plot#
Search
Most Visited
Red Hat
Customer Portal
Documentation
Red Hat Network
plotly
make charts and dashboards online
My Charts
Workspace
Try Plotly 2
Help
Pricing
jburkardt
starter data
hw_200.csv
Plot
Plot
NEW GRID
IMPORT
View data
View JSON
Share
Traces
Layout
Axes
Notes
Legend
Fit data
Themes
Download plot as a png
Click to enter Y axis title
0
5
10
15
20
64
66
68
70
72
74
Height(Inches)"
It looks like you haven't started Firefox in a while. Do you want to clean it up for a fresh, like-new experience? And by the way, welcome back!
Refresh Firefox...


Plotly | Make charts and dashboards online - Mozilla Firefox
Plotly | Make charts ...
https://plot.ly/plot#
Search
Most Visited
Red Hat
Customer Portal
Documentation
Red Hat Network
plotly
make charts and dashboards online
My Charts
Workspace
Try Plotly 2
Help
Pricing
jburkardt
starter data
hw_200.csv
Plot
Plot
NEW GRID
IMPORT
Traces
Height(Inches)"
Mode
Range/bins
Style
Error bars
Normalization
None
%
p
1/Δx
p/Δx
X bins
Auto
Manual
X range
63
74
X bin size
0.5
View data
View JSON
Share
Traces
Layout
Axes
Notes
Legend
Fit data
Themes
200 Heights
5
0
64
66
68
70
72
74
Height(Inches)"
It looks like you haven't started Firefox in a while. Do you want to clean it up for a fresh, like-new experience? And by the way, welcome back!
Refresh Firefox...


Plotly | Make charts and dashboards online - Mozilla Firefox
Plotly | Make charts ...
https://plot.ly/plot#
Search
Most Visited
Red Hat
Customer Portal
Documentation
Red Hat Network
plotly
make charts and dashboards online
My Charts
Workspace
Try Plotly 2
Help
Pricing
jburkardt
starter data
hw_200.csv
Plot
Plot
NEW GRID
IMPORT
Traces
Height(Inches)"
Mode
Range/bins
Style
Error bars
Opacity
1
Marker
Fill
Outline
0
Bar Gap
0.1
Group Gap
0
View data
View JSON
Share
Traces
Layout
Axes
Notes
Legend
Fit data
Themes
200 Heights
4
2
0
64
66
68
70
72
74
Height(Inches)"
It looks like you haven't started Firefox in a while. Do you want to clean it up for a fresh, like-new experience? And by the way, welcome back!
Refresh Firefox...


Plotly | Make charts and dashboards online - Mozilla Firefox
Plotly | Make charts ...
https://plot.ly/plot#
Search
Most Visited
Red Hat
Customer Portal
Documentation
Red Hat Network
plotly
make charts and dashboards online
My Charts
Workspace
Try Plotly 2
Help
Pricing
jburkardt
starter data
hw_200.csv
Plot
Plot
NEW GRID
IMPORT
Traces
Height(Inches)"
Mode
Range/bins
Style
Error bars
Normalization
None
%
p
1/Δx
p/Δx
X bins
Auto
Manual
X range
63
74
X bin size
1
View data
View JSON
Share
Traces
Layout
Axes
Notes
Legend
Fit data
Themes
200 Heights
Percentage
15
10
5
0
64
66
68
70
72
74
Height(Inches)"
It looks like you haven't started Firefox in a while. Do you want to clean it up for a fresh, like-new experience? And by the way, welcome back!
Refresh Firefox...

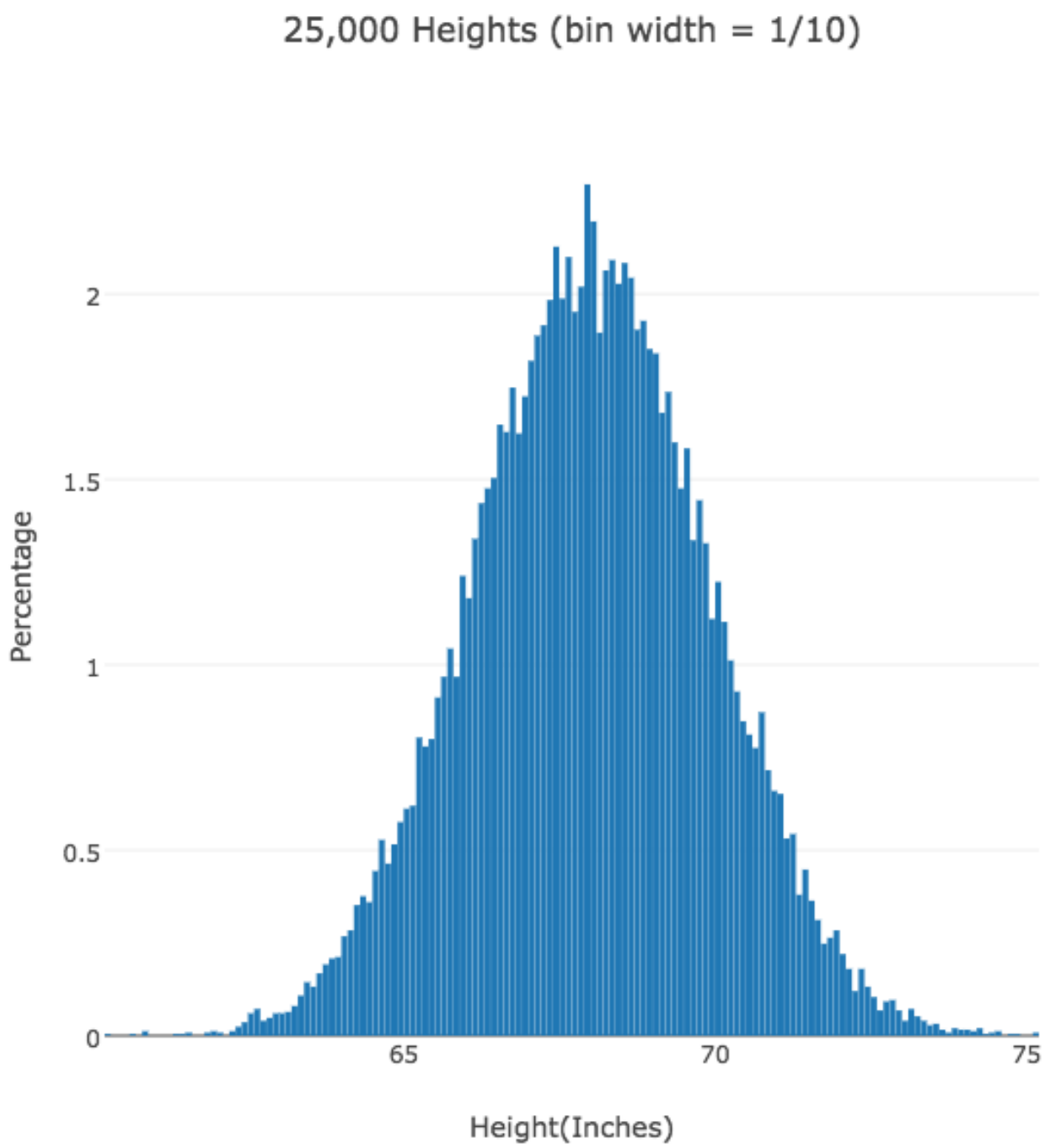

200 Heights (Bin width = 1)
Percentage
20
15
10
5
0
64
66
68
70
72
74
Height(Inches)"

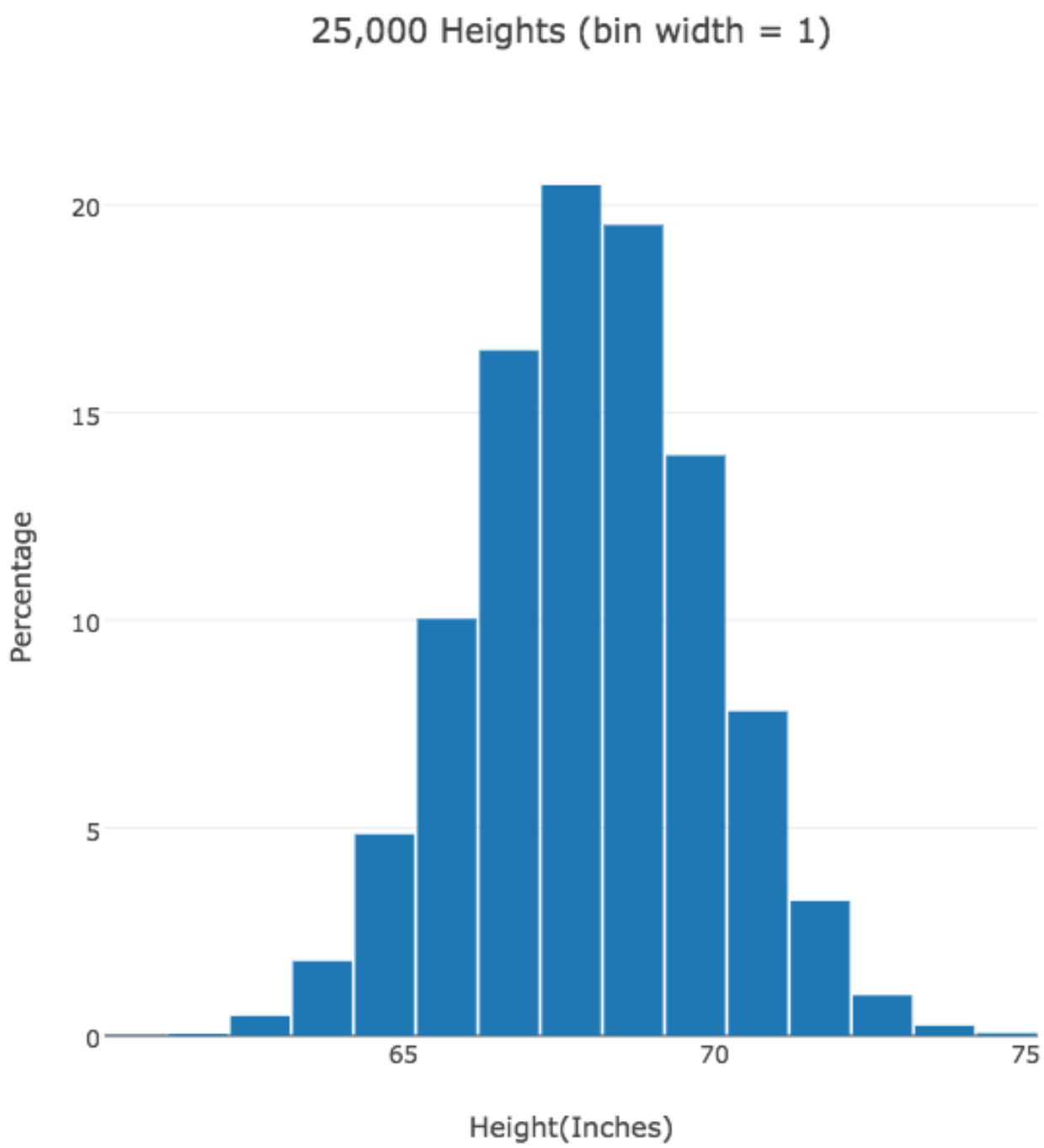
Example #1.5: (DO OVER!) 25,000 Height and Weight Measurements

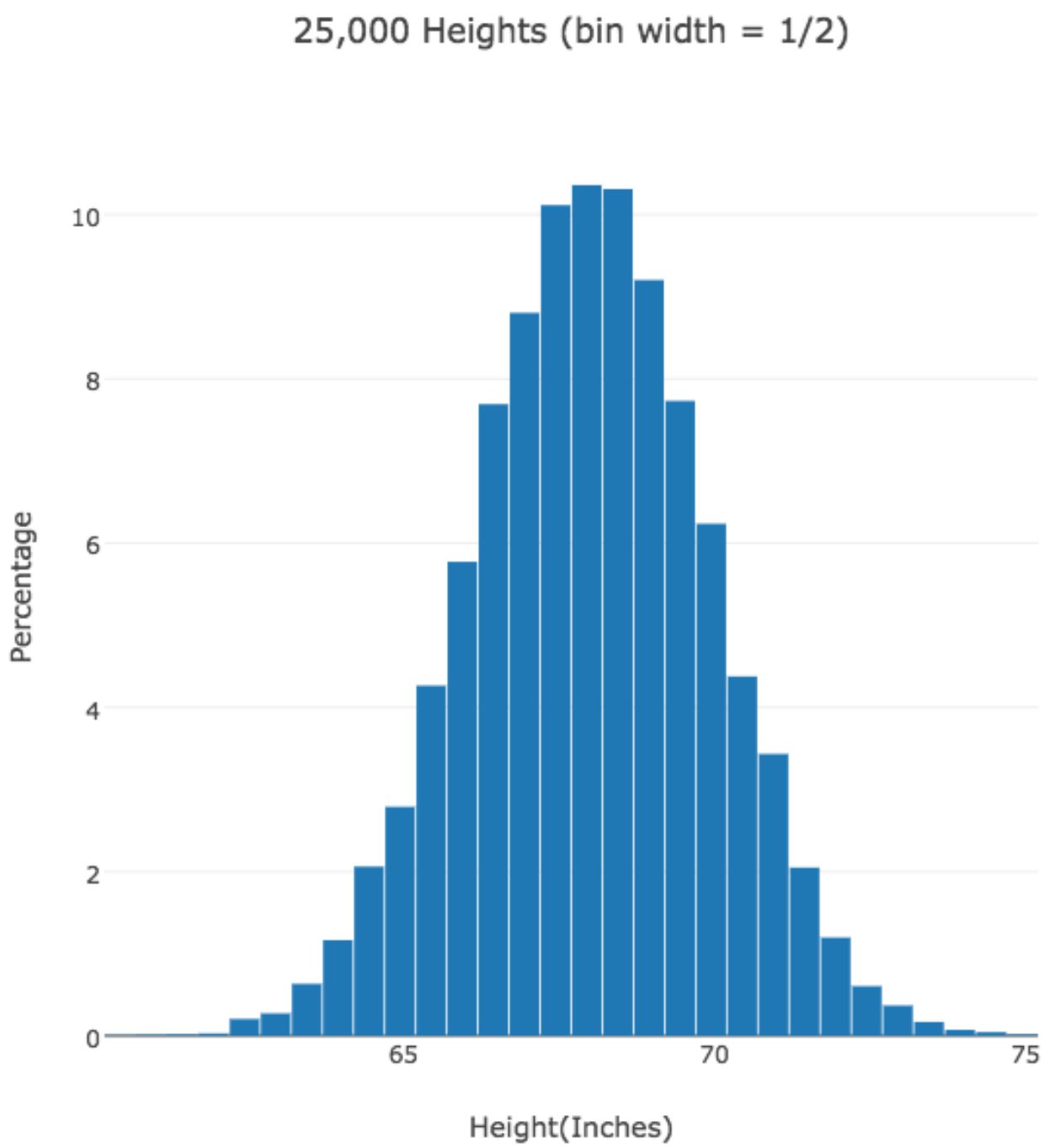
Our histogram of the height measurements was somewhat disappointing. There was a pattern visible, with the typical height seeming to 67.5 inches or $5'7\frac{1}{2}"$, and other heights tending to stay close to that value.

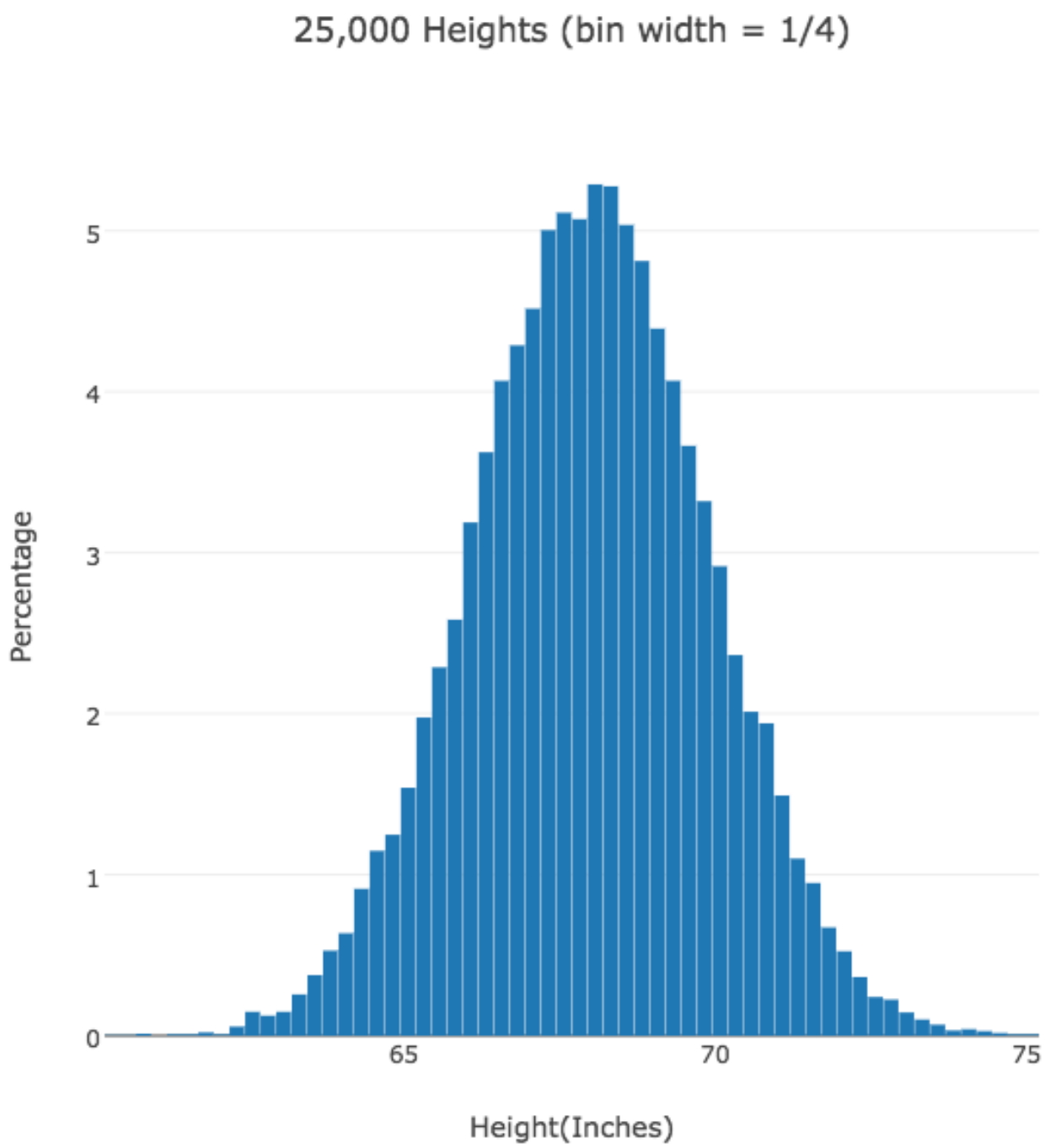
However, the histogram pattern looked very "blocky" and irregular, especially when we had a bin width of $\frac{1}{2}$. We would expect that a law of nature would have a smoother appearance.

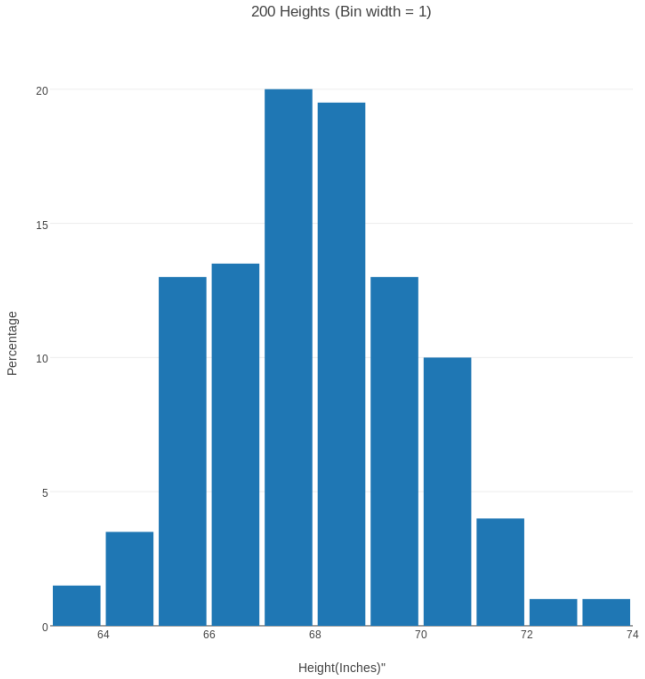
However, it's possible that we just haven't looked at enough data for the pattern to become clear. We can test this idea, because we have another dataset of 25,000 height and weight measurements.

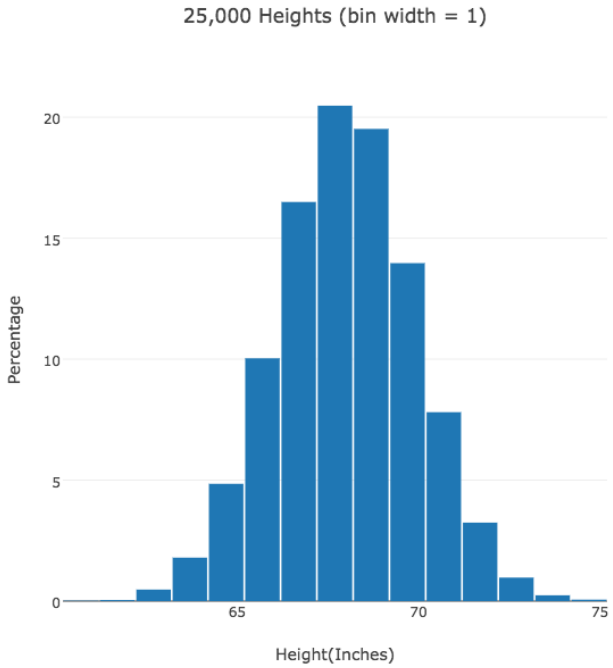

25,000 Heights (bin width = 1/10)
Percentage
2
1.5
1
0.5
0
65
70
75
Height(Inches)


25,000 Heights (bin width = 1)
Percentage
20
15
10
5
0
65
70
75
Height(Inches)


25,000 Heights (bin width = 1/2)
Percentage
10
8
6
4
2
0
65
70
75
Height(Inches)


25,000 Heights (bin width = 1/4)
Percentage
0
1
2
3
4
5
65
70
75
Height(Inches)


200 Heights (Bin width = 1)
Percentage
20
15
10
5
0
64
66
68
70
72
74
Height(Inches)"


25,000 Heights (bin width = 1)
Percentage
20
15
10
5
0
65
70
75
Height(Inches)

Example #2: Female Oscar Winners

The Academy awards or "Oscars" for best Actor and Actress have been given every year since 1928, a total of 89 years.

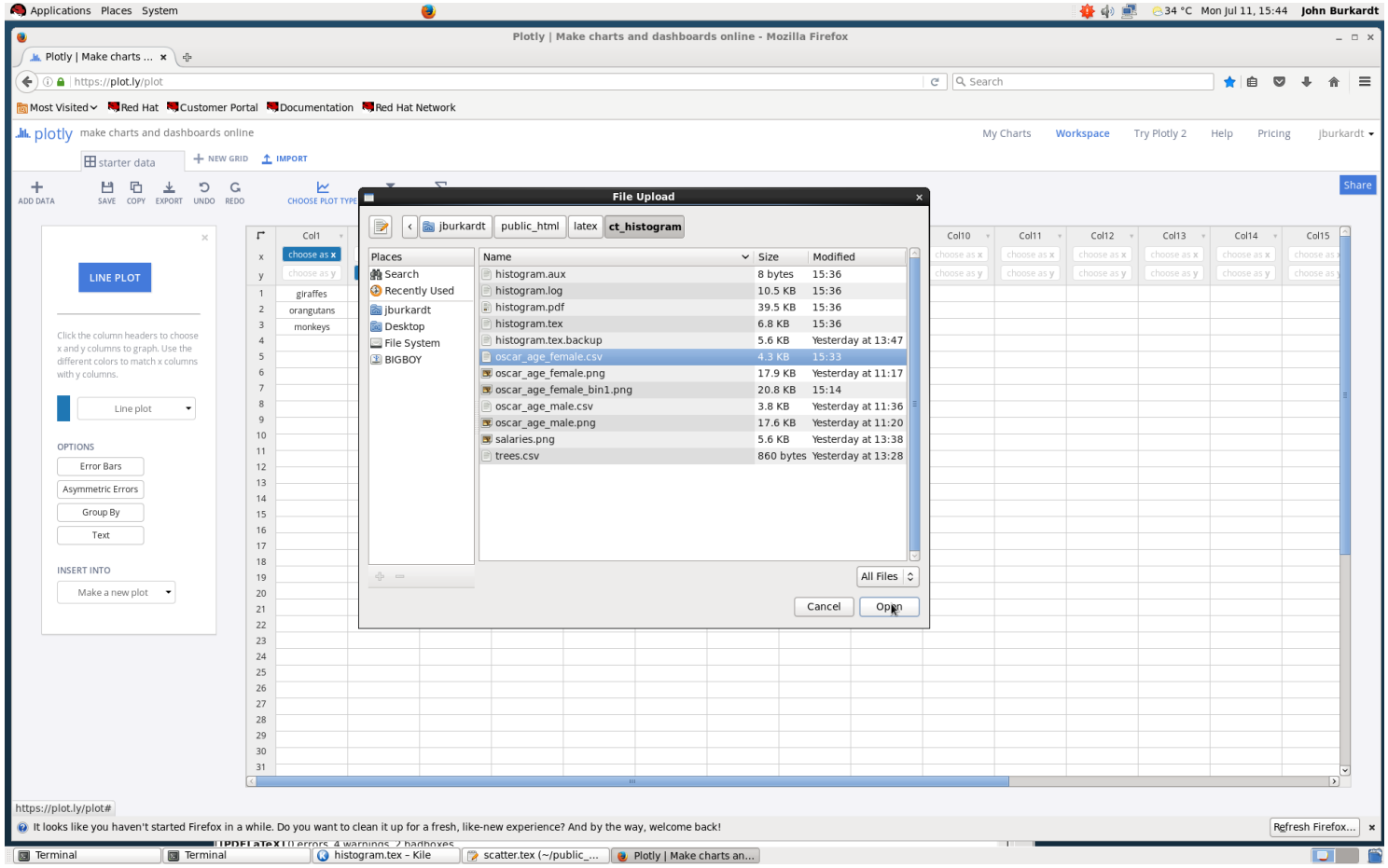
It has often been said that Hollywood treats men and women stars very differently: male actors have long careers, and can have leading roles into their 60's, while actresses are often allowed only a few hit roles before being dropped in favor of younger stars.

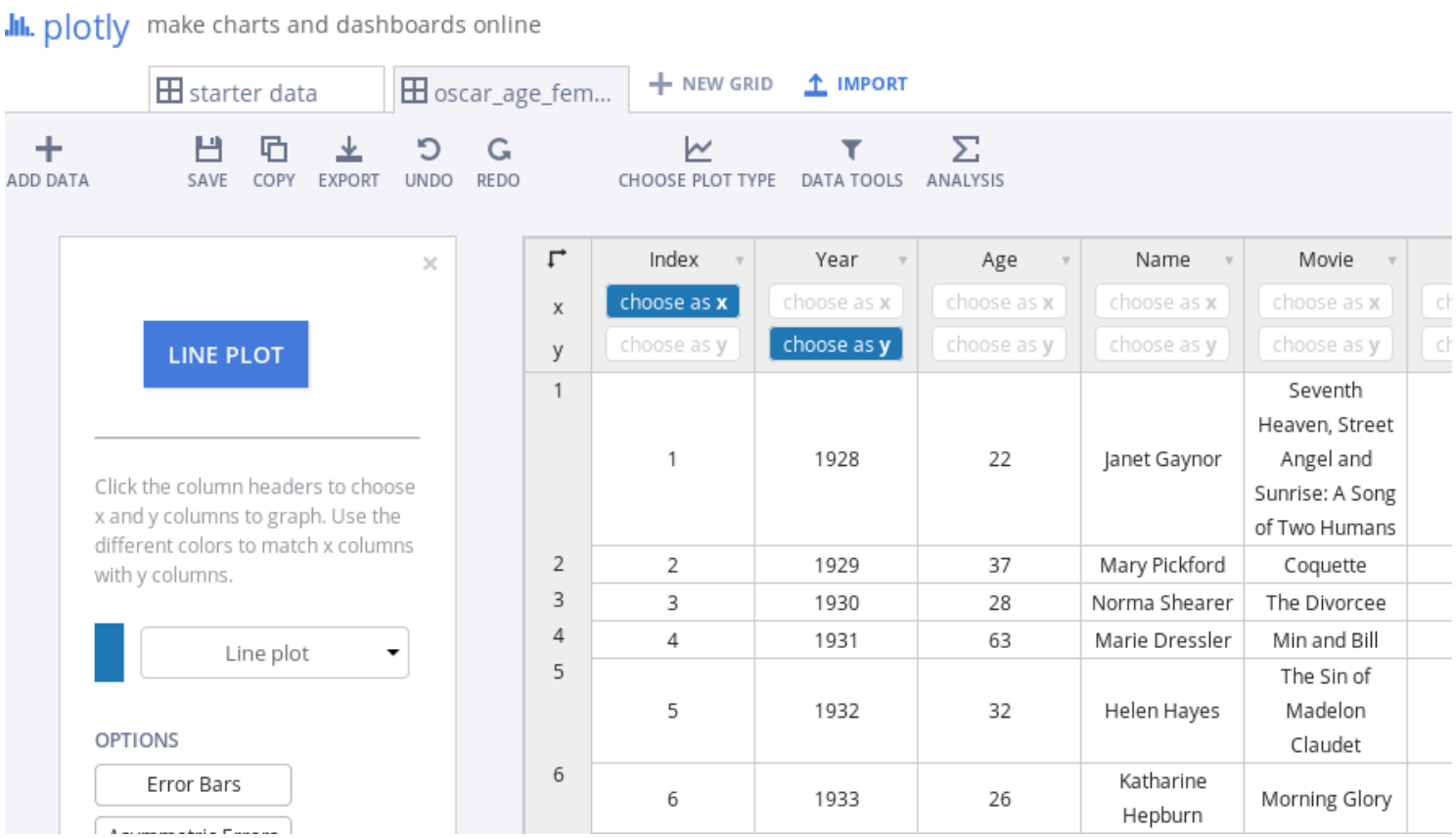
As a small test of this statement, we can plot the ages of female Oscar winners, and see if the plot has any story to tell.

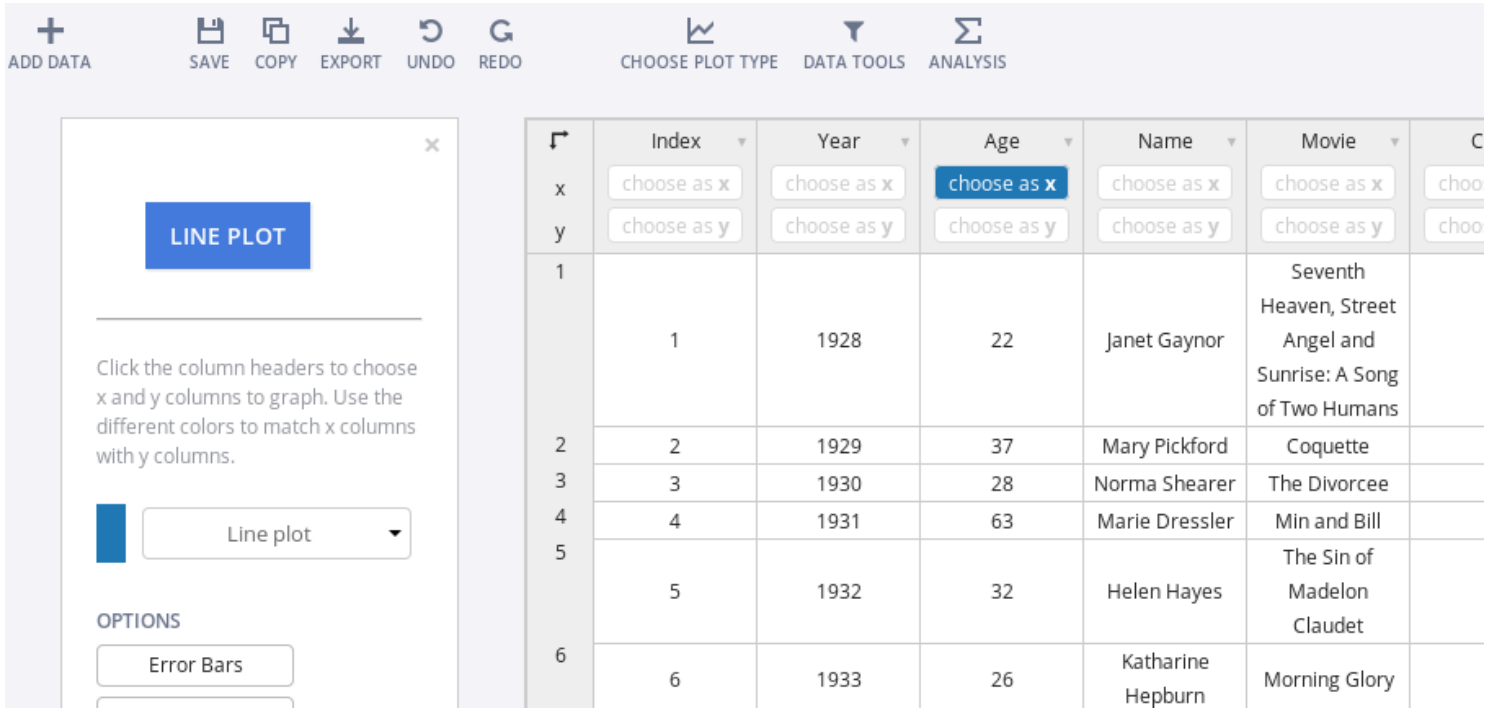
Our data is stored in the file **oscar_female.csv**.

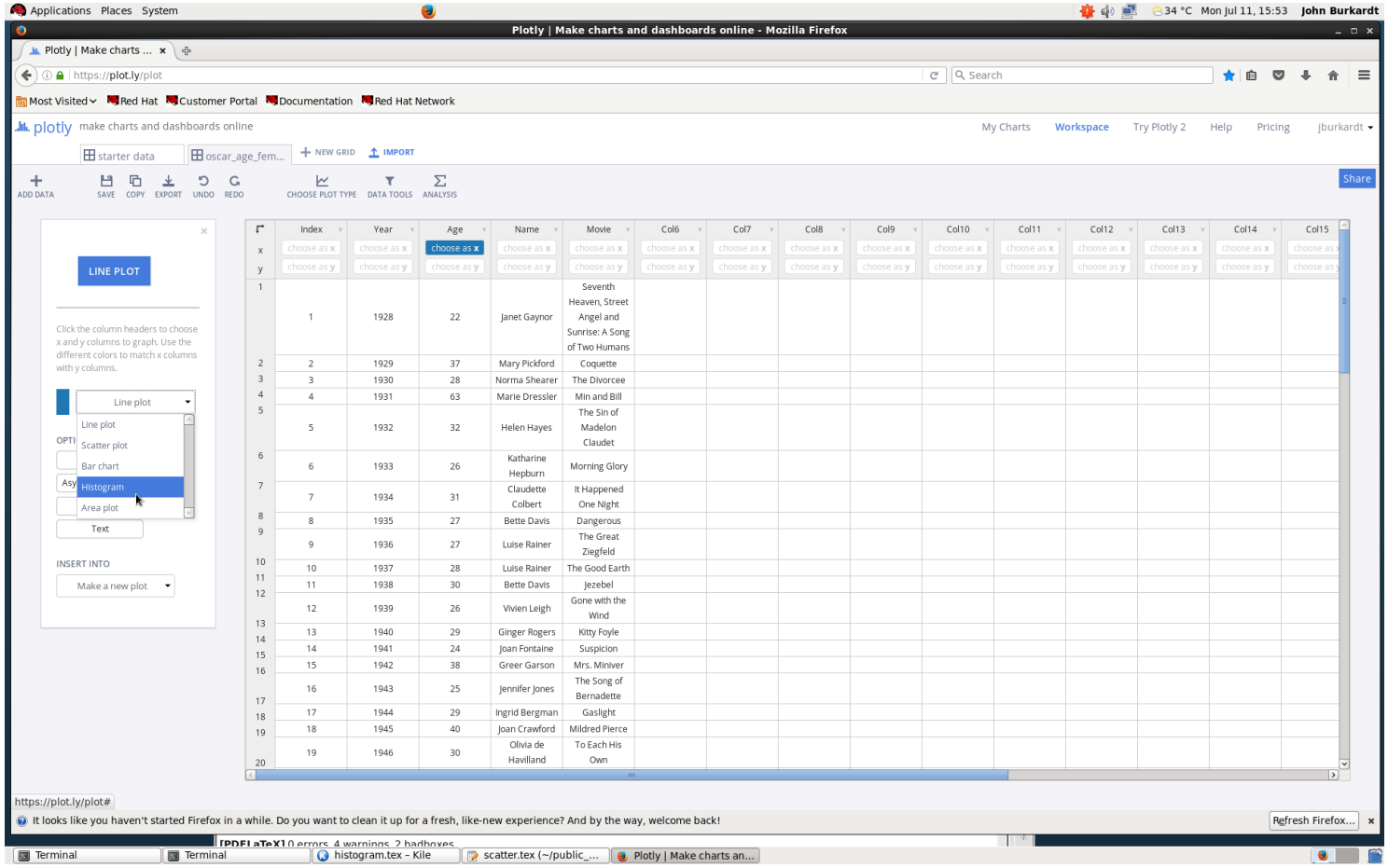
"Index",	"Year",	"Age",	"Name",	"Movie"
1,	1928,	22,	"Janet Gaynor",	"Seventh Heaven"
2,	1929,	37,	"Mary Pickford",	"Coquette"
3,	1930,	28,	"Norma Shearer",	"The Divorcee"
—	—	*more data*	—	—
87,	2014,	44,	"Cate Blanchett",	"Blue Jasmine"
88,	2015,	54,	"Julianne Moore",	"Still Alice"
89,	2016,	26,	"Brie Larson",	"Room"

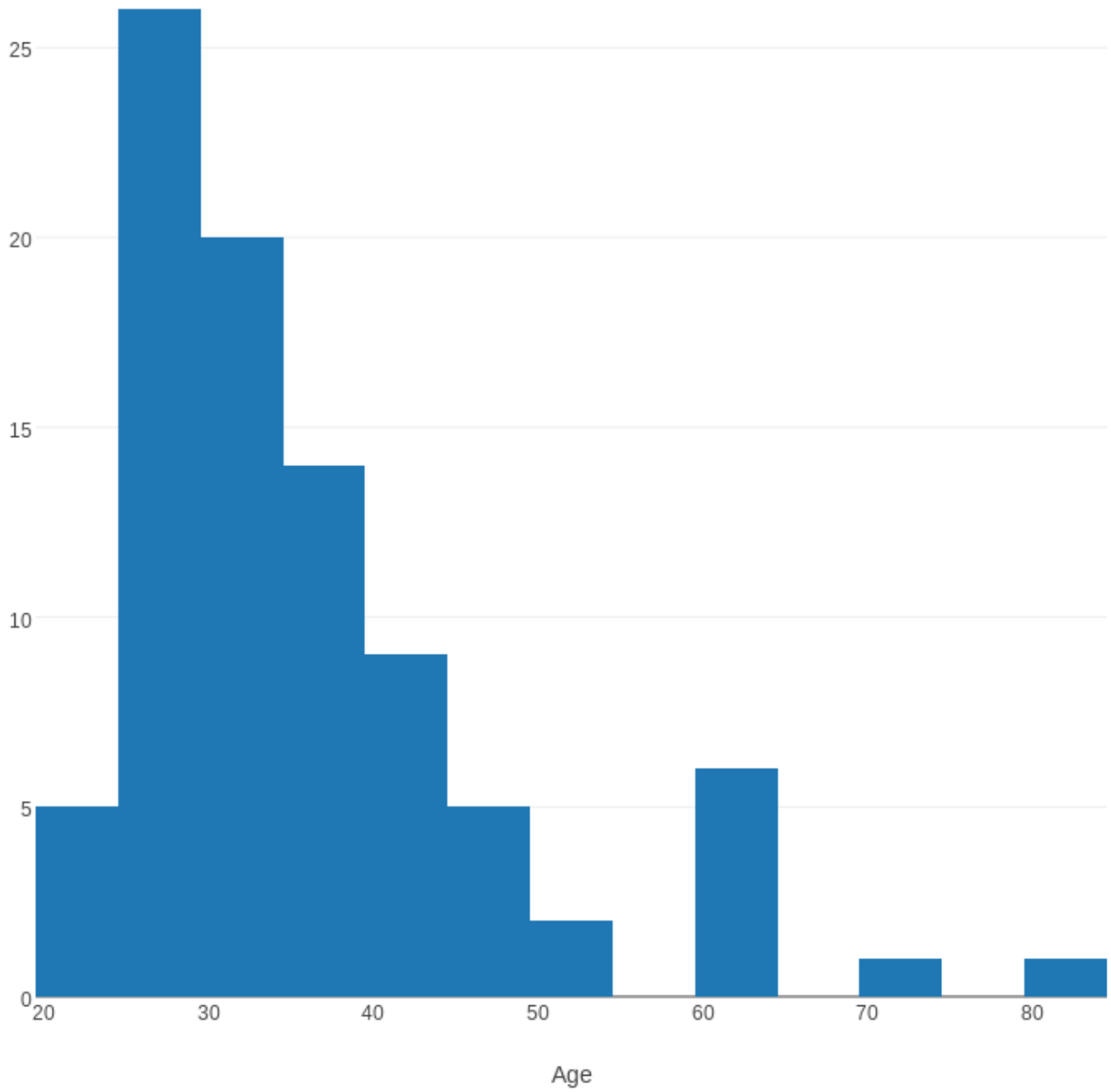
We begin our histogram creation by logging into Plotly. There is already some sample data entered in the Plotly grid, but we want to replace this with the data from our file. We do this by going to the **IMPORT** menu item, choosing **Upload file**, then using the browse menu to find our file and clicking on it.

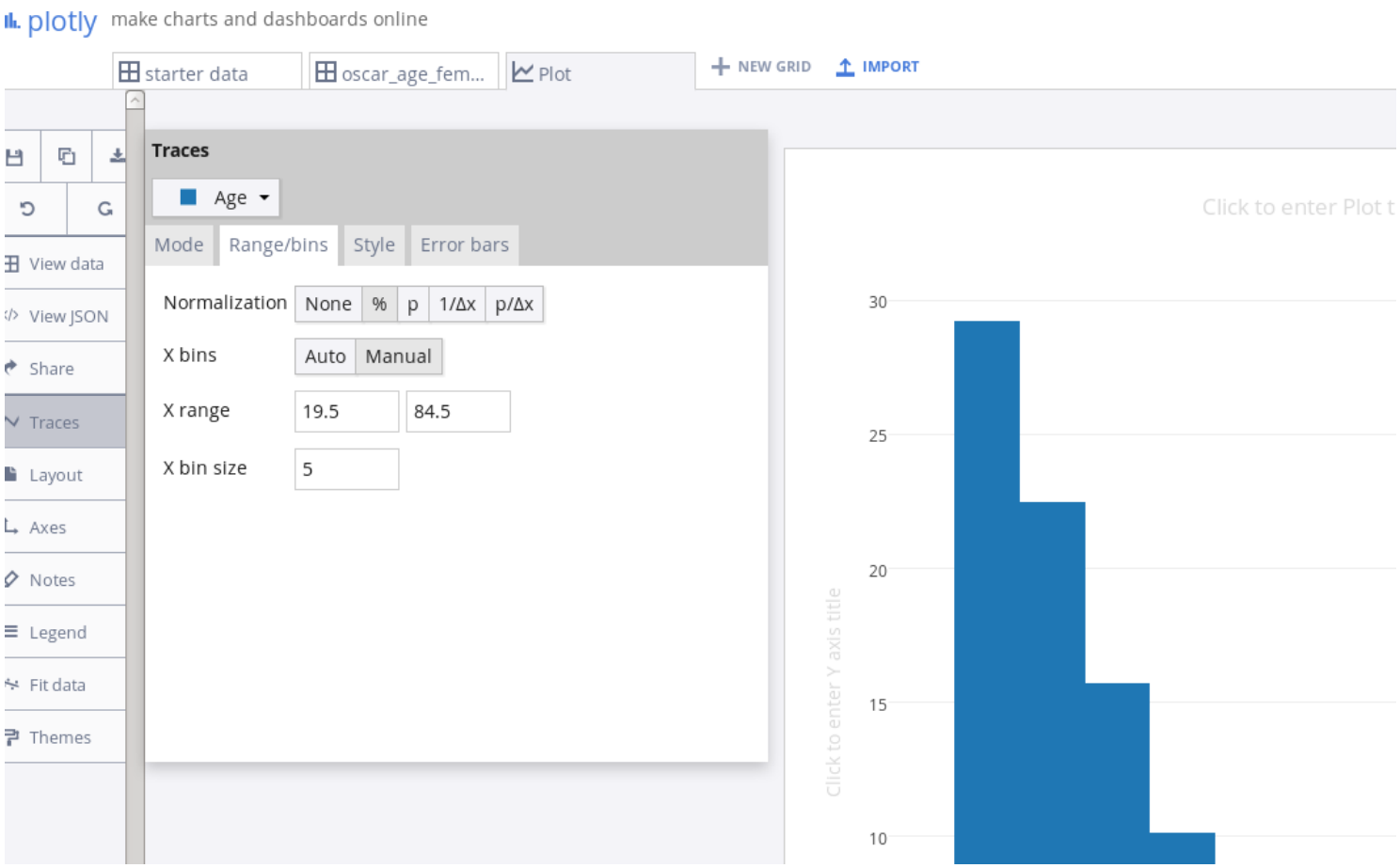

Applications Places System
34 °C Mon Jul 11, 15:44 John Burkardt
Plotly | Make charts and dashboards online - Mozilla Firefox
Plotly | Make charts ...
https://plot.ly/plot
Search
Most Visited Red Hat Customer Portal Documentation Red Hat Network
plotly make charts and dashboards online
My Charts Workspace Try Plotly 2 Help Pricing jburkardt
starter data
NEW GRID
IMPORT
ADD DATA
SAVE
COPY
EXPORT
UNDO
REDO
CHOOSE PLOT TYPE
Share
LINE PLOT
Click the column headers to choose x and y columns to graph. Use the different colors to match x columns with y columns.
Line plot
OPTIONS
Error Bars
Asymmetric Errors
Group By
Text
INSERT INTO
Make a new plot
File Upload
jburkardt
public_html
latex
ct_histogram
Places
Search
Recently Used
jburkardt
Desktop
File System
BIGBOY
Name
Size
Modified
histogram.aux 8 bytes 15:36
histogram.log 10.5 KB 15:36
histogram.pdf 39.5 KB 15:36
histogram.tex 6.8 KB 15:36
histogram.tex.backup 5.6 KB Yesterday at 13:47
oscar_age_female.csv 4.3 KB 15:33
oscar_age_female.png 17.9 KB Yesterday at 11:17
oscar_age_female_bin1.png 20.8 KB 15:14
oscar_age_male.csv 3.8 KB Yesterday at 11:36
oscar_age_male.png 17.6 KB Yesterday at 11:20
salaries.png 5.6 KB Yesterday at 13:38
trees.csv 860 bytes Yesterday at 13:28
All Files
Cancel
Open
giraffes
orangutans
monkeys
https://plot.ly/plot#
It looks like you haven't started Firefox in a while. Do you want to clean it up for a fresh, like-new experience? And by the way, welcome back!
Refresh Firefox...
Terminal
Terminal
histogram.tex – Kile
scatter.tex (~/public_...
Plotly | Make charts an...

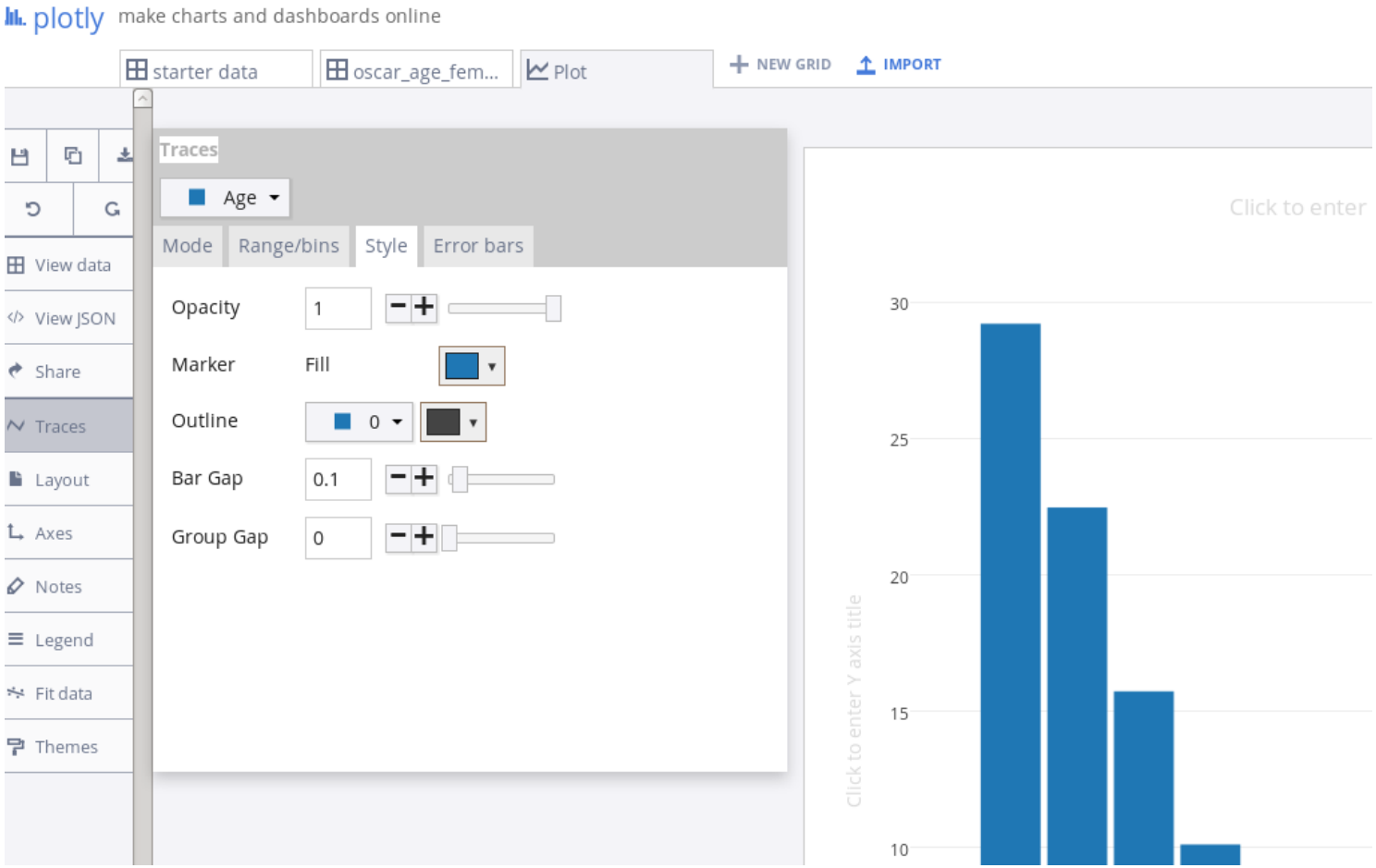

plotly make charts and dashboards online
starter data
oscar_age_fem...
NEW GRID
IMPORT
ADD DATA
SAVE
COPY
EXPORT
UNDO
REDO
CHOOSE PLOT TYPE
DATA TOOLS
ANALYSIS
LINE PLOT
Click the column headers to choose x and y columns to graph. Use the different colors to match x columns with y columns.
Line plot
OPTIONS
Error Bars
Index
Year
Age
Name
Movie
x
y
choose as x
choose as y
1 1 1928 22 Janet Gaynor Seventh Heaven, Street Angel and Sunrise: A Song of Two Humans
2 2 1929 37 Mary Pickford Coquette
3 3 1930 28 Norma Shearer The Divorcee
4 4 1931 63 Marie Dressler Min and Bill
5 5 1932 32 Helen Hayes The Sin of Madelon Claudet
6 6 1933 26 Katharine Hepburn Morning Glory

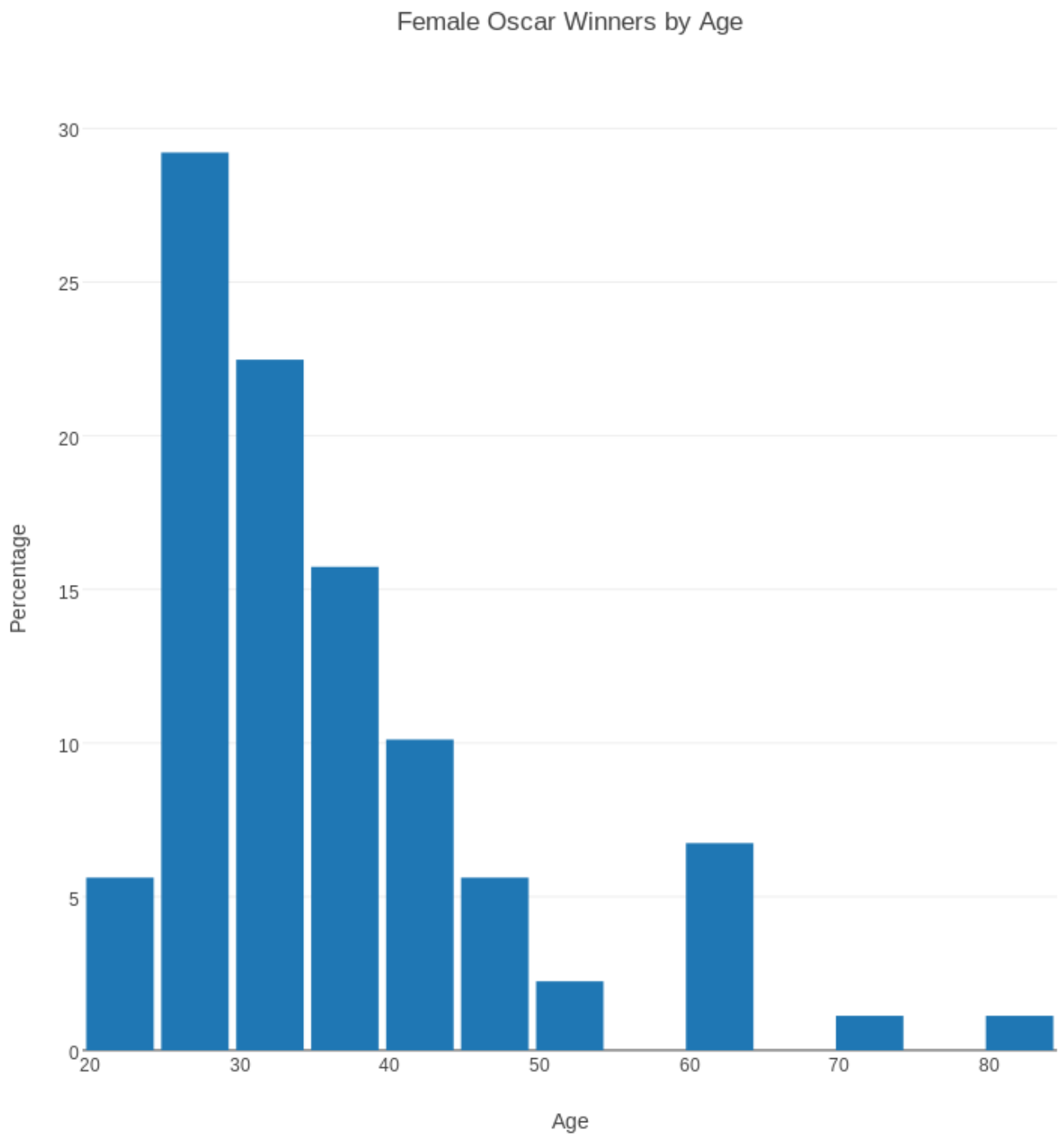

ADD DATA
SAVE
COPY
EXPORT
UNDO
REDO
CHOOSE PLOT TYPE
DATA TOOLS
ANALYSIS
LINE PLOT
Click the column headers to choose x and y columns to graph. Use the different colors to match x columns with y columns.
Line plot
OPTIONS
Error Bars
Index
Year
Age
Name
Movie
x
y
choose as x
choose as y
1
1
1928
22
Janet Gaynor
Seventh Heaven, Street Angel and Sunrise: A Song of Two Humans
2
2
1929
37
Mary Pickford
Coquette
3
3
1930
28
Norma Shearer
The Divorcee
4
4
1931
63
Marie Dressler
Min and Bill
5
5
1932
32
Helen Hayes
The Sin of Madelon Claudet
6
6
1933
26
Katharine Hepburn
Morning Glory

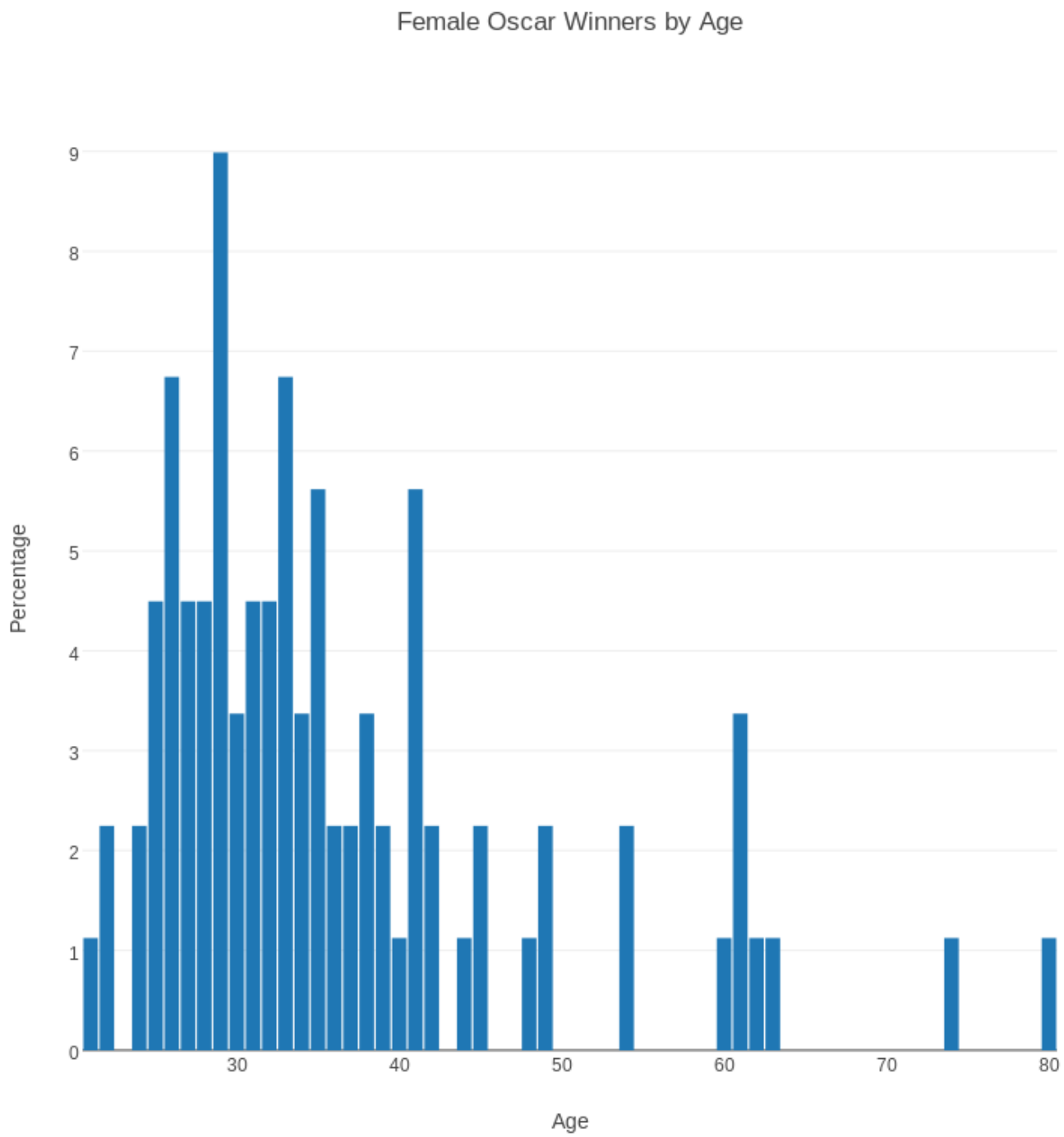

Applications Places System
34 °C Mon Jul 11, 15:53 John Burkardt
Plotly | Make charts and dashboards online - Mozilla Firefox
Plotly | Make charts ...
https://plot.ly/plot
Search
Most Visited Red Hat Customer Portal Documentation Red Hat Network
plotly make charts and dashboards online
My Charts Workspace Try Plotly 2 Help Pricing jburkardt
starter data
oscar_age_fem...
NEW GRID
IMPORT
ADD DATA
SAVE
COPY
EXPORT
UNDO
REDO
CHOOSE PLOT TYPE
DATA TOOLS
ANALYSIS
Share
LINE PLOT
Click the column headers to choose x and y columns to graph. Use the different colors to match x columns with y columns.
Line plot
Line plot
Scatter plot
Bar chart
Histogram
Area plot
Text
INSERT INTO
Make a new plot
Index Year Age Name Movie Col6 Col7 Col8 Col9 Col10 Col11 Col12 Col13 Col14 Col15
x
y
choose as x
choose as y
1 1928 22 Janet Gaynor Seventh Heaven, Street Angel and Sunrise: A Song of Two Humans
2 1929 37 Mary Pickford Coquette
3 1930 28 Norma Shearer The Divorcee
4 1931 63 Marie Dressler Min and Bill
5 1932 32 Helen Hayes The Sin of Madelon Claudet
6 1933 26 Katharine Hepburn Morning Glory
7 1934 31 Claudette Colbert It Happened One Night
8 1935 27 Bette Davis Dangerous
9 1936 27 Luise Rainer The Great Ziegfeld
10 1937 28 Luise Rainer The Good Earth
11 1938 30 Bette Davis Jezebel
12 1939 26 Vivien Leigh Gone with the Wind
13 1940 29 Ginger Rogers Kitty Foyle
14 1941 24 Joan Fontaine Suspicion
15 1942 38 Greer Garson Mrs. Miniver
16 1943 25 Jennifer Jones The Song of Bernadette
17 1944 29 Ingrid Bergman Gaslight
18 1945 40 Joan Crawford Mildred Pierce
19 1946 30 Olivia de Havilland To Each His Own
20
https://plot.ly/plot#
It looks like you haven't started Firefox in a while. Do you want to clean it up for a fresh, like-new experience? And by the way, welcome back!
Refresh Firefox...
Terminal
Terminal
histogram.tex – Kile
scatter.tex (~/public_...
Plotly | Make charts an...


25
20
15
10
5
0
20
30
40
50
60
70
80
Age


plotly
make charts and dashboards online
starter data
oscar_age_fem...
Plot
NEW GRID
IMPORT
View data
View JSON
Share
Traces
Layout
Axes
Notes
Legend
Fit data
Themes
Traces
Age
Mode
Range/bins
Style
Error bars
Normalization
None
%
p
1/Δx
p/Δx
X bins
Auto
Manual
X range
19.5
84.5
X bin size
5
Click to enter Plot t
Click to enter Y axis title
30
25
20
15
10


plotly
make charts and dashboards online
starter data
oscar_age_fem...
Plot
NEW GRID
IMPORT
View data
View JSON
Share
Traces
Layout
Axes
Notes
Legend
Fit data
Themes
Traces
Age
Mode
Range/bins
Style
Error bars
Opacity
1
Marker
Fill
Outline
0
Bar Gap
0.1
Group Gap
0
Click to enter
Click to enter Y axis title
30
25
20
15
10


Female Oscar Winners by Age
Percentage
30
25
20
15
10
5
0
20
30
40
50
60
70
80
Age


Female Oscar Winners by Age
Percentage
0
1
2
3
4
5
6
7
8
9
30
40
50
60
70
80
Age

Example #3: Male Oscar Winners

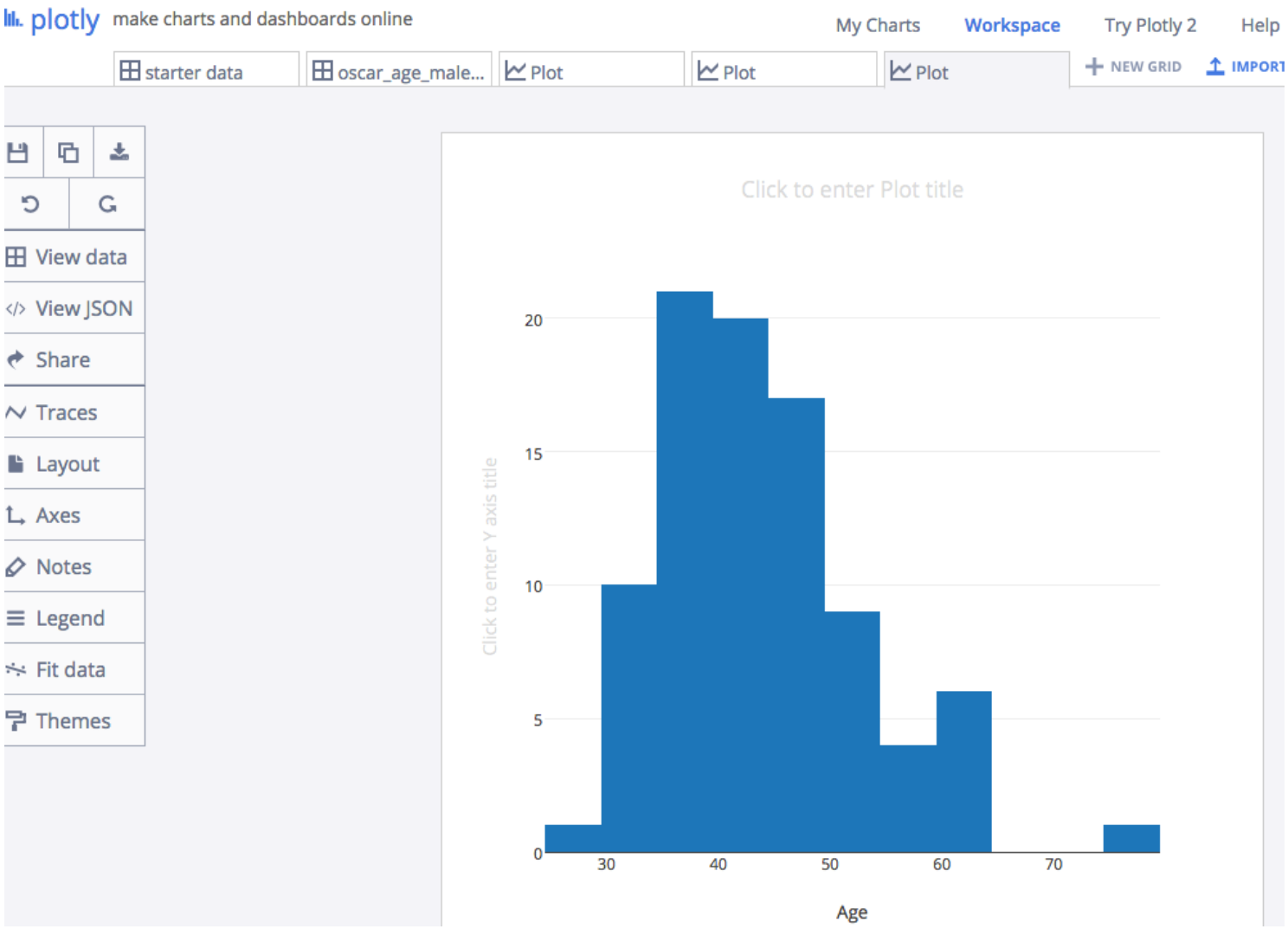
The female Academy Award winners represent a small and untypical sample of the acting careers of women in Hollywood, but our histogram does suggest that women have the greatest success at a very early age, and seem to decrease in visibility each time we move 5 years further in age range, although there are some "blips" popping up in the later end of the plot.

If our data is correctly showing us the "story", then it's natural to wonder about what we would see if we examined the same data for male Oscar winners.

First we need to grab the corresponding data file!

Our data is stored in the file **oscar_male.csv**.

"Index",	"Year",	"Age",	"Name",	"Movie"
1,	1928,	44,	"Emil Jannings",	"The Last Command"
2,	1929,	41,	"Warner Baxter",	"In Old Arizona"
3,	1930,	62,	"George Arliss",	"Disraeli"
—	—	*more data*	—	—
87,	2014,	44,	"Matthew McConaughey",	"Dallas Buyers Club"
88,	2015,	33,	"Eddie Redmayne",	"The Theory of Everything"
89,	2016,	41,	"Leonardo DiCaprio",	"The Revenant"


plotly
make charts and dashboards online
My Charts
Workspace
Try Plotly 2
Help
starter data
oscar_age_male...
Plot
Plot
Plot
NEW GRID
IMPORT
View data
View JSON
Share
Traces
Layout
Axes
Notes
Legend
Fit data
Themes
Click to enter Plot title
Click to enter Y axis title
20
15
10
5
0
30
40
50
60
70
Age

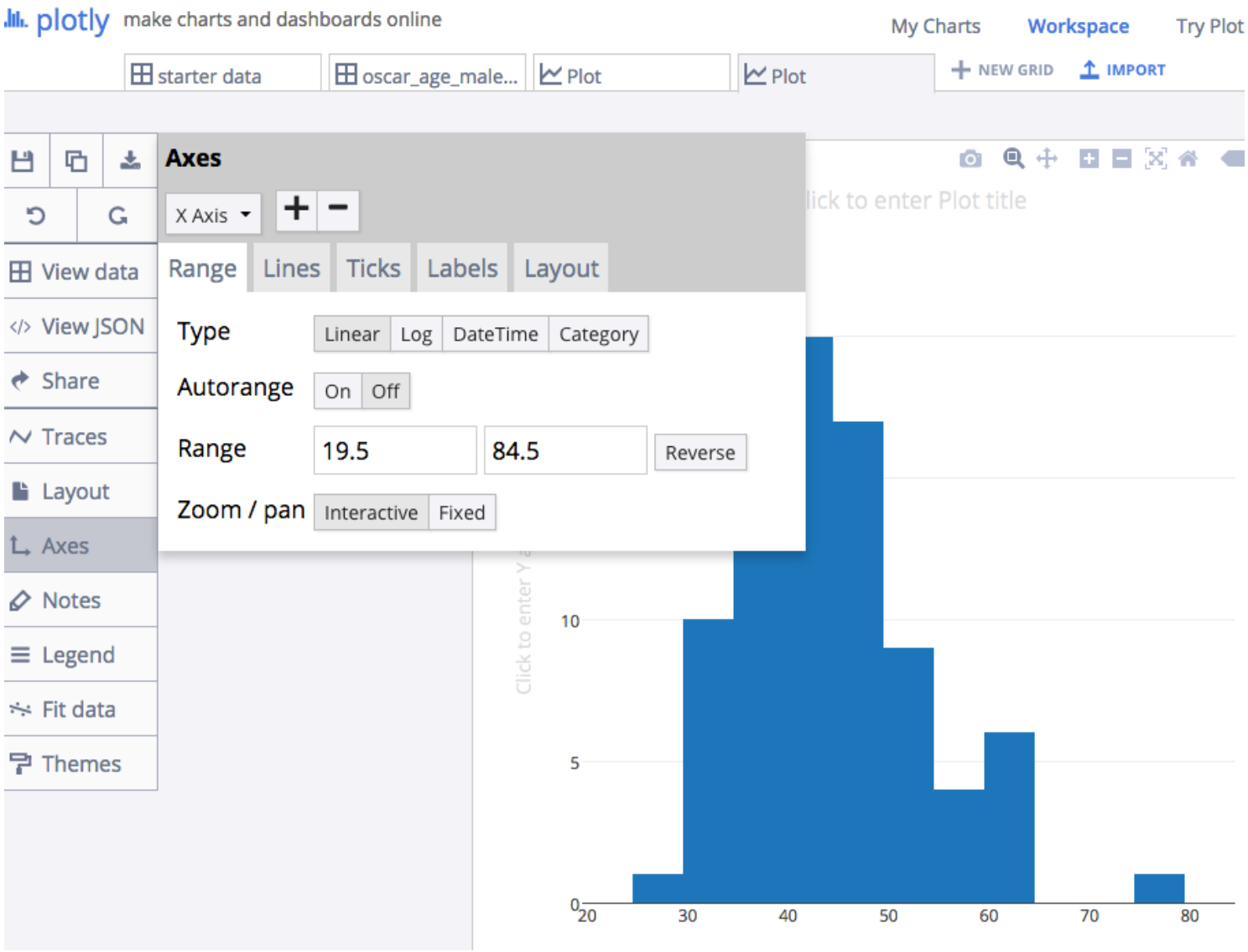
Our first plot has an interesting shape, but if we are going to compare it to the same plot for the female Oscar winners, we need to make some modifications. In particular, note that the horizontal data (the "Age" values) doesn't seem to have the same extent as in the plot of the female data.

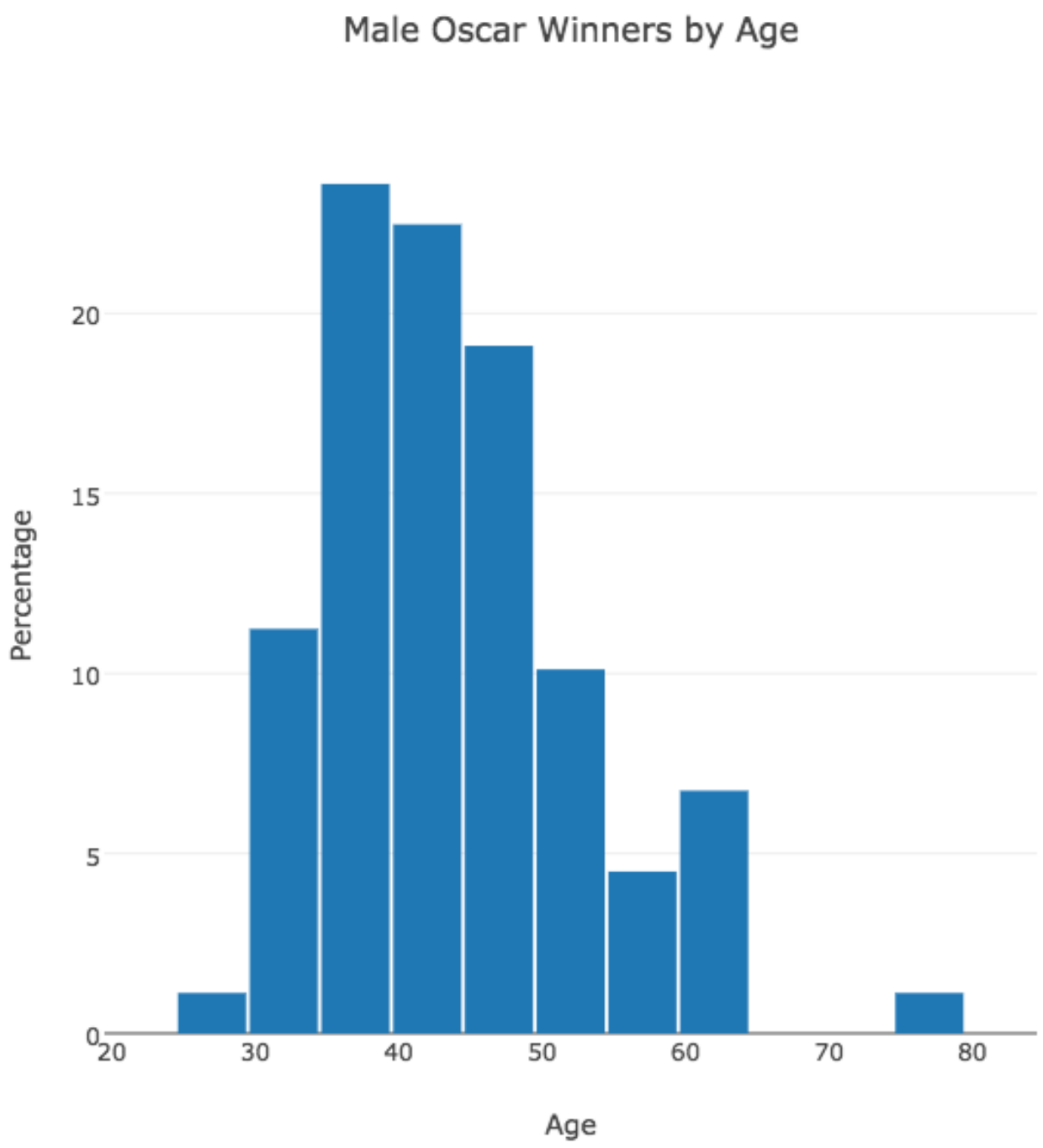
Unless we say so, Plotly will make the graph just wide enough for the data it sees. But we can go into the Axes menu item, and pick the **Range** item and then change the range from the values

```
Range 24.5  79.5
```

to match the range on the female plot:

```
Range   19.5  84.5
```

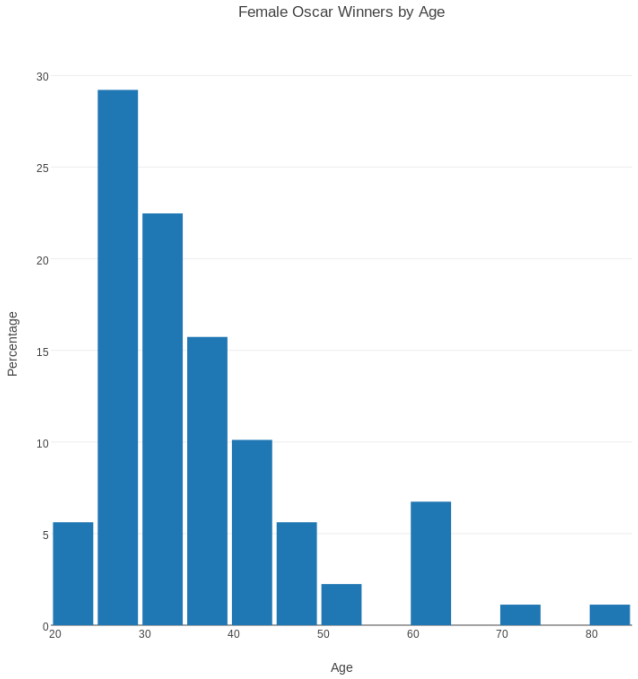

plotly
make charts and dashboards online
My Charts
Workspace
Try Plot
starter data
oscar_age_male...
Plot
Plot
NEW GRID
IMPORT
Axes
X Axis
Range
Lines
Ticks
Labels
Layout
Type
Linear
Log
DateTime
Category
Autorange
On
Off
Range
19.5
84.5
Reverse
Zoom / pan
Interactive
Fixed
View data
View JSON
Share
Traces
Layout
Axes
Notes
Legend
Fit data
Themes
Click to enter Plot title
10
5
0
20
30
40
50
60
70
80

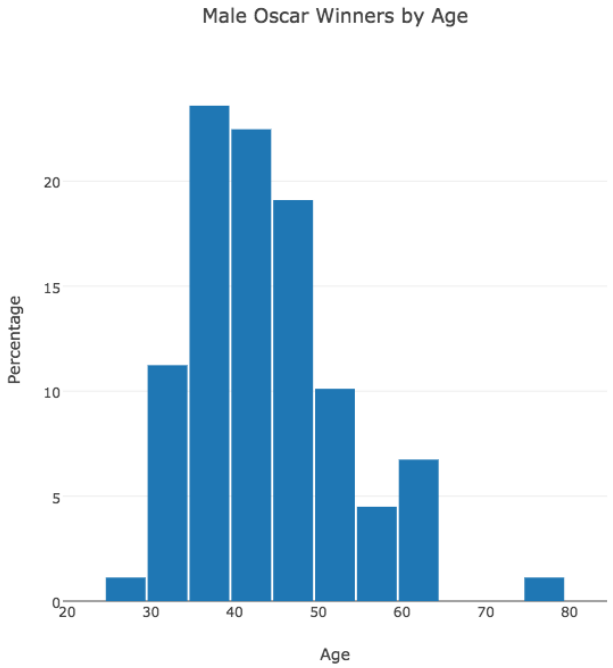

Male Oscar Winners by Age
Percentage
0
5
10
15
20
20
30
40
50
60
70
80
Age

Now we have both sets of data plotted with the same age range matched up.

We can concentrate on the "story" that the two sets of histogram bars seem to be telling. One reading of this story is:

- Male actors have their peak popularity in their late 30's, and their careers don't start to drop off until around 50, continuing til 65.
- Female actores have their peak popularity in their late 20's, and their careers immediately begin to drop off substantially every 5 years thereafter til about 55.


Female Oscar Winners by Age
Percentage
Age
0
5
10
15
20
25
30
20
30
40
50
60
70
80


Male Oscar Winners by Age
Percentage
Age
0
5
10
15
20
20
30
40
50
60
70
80

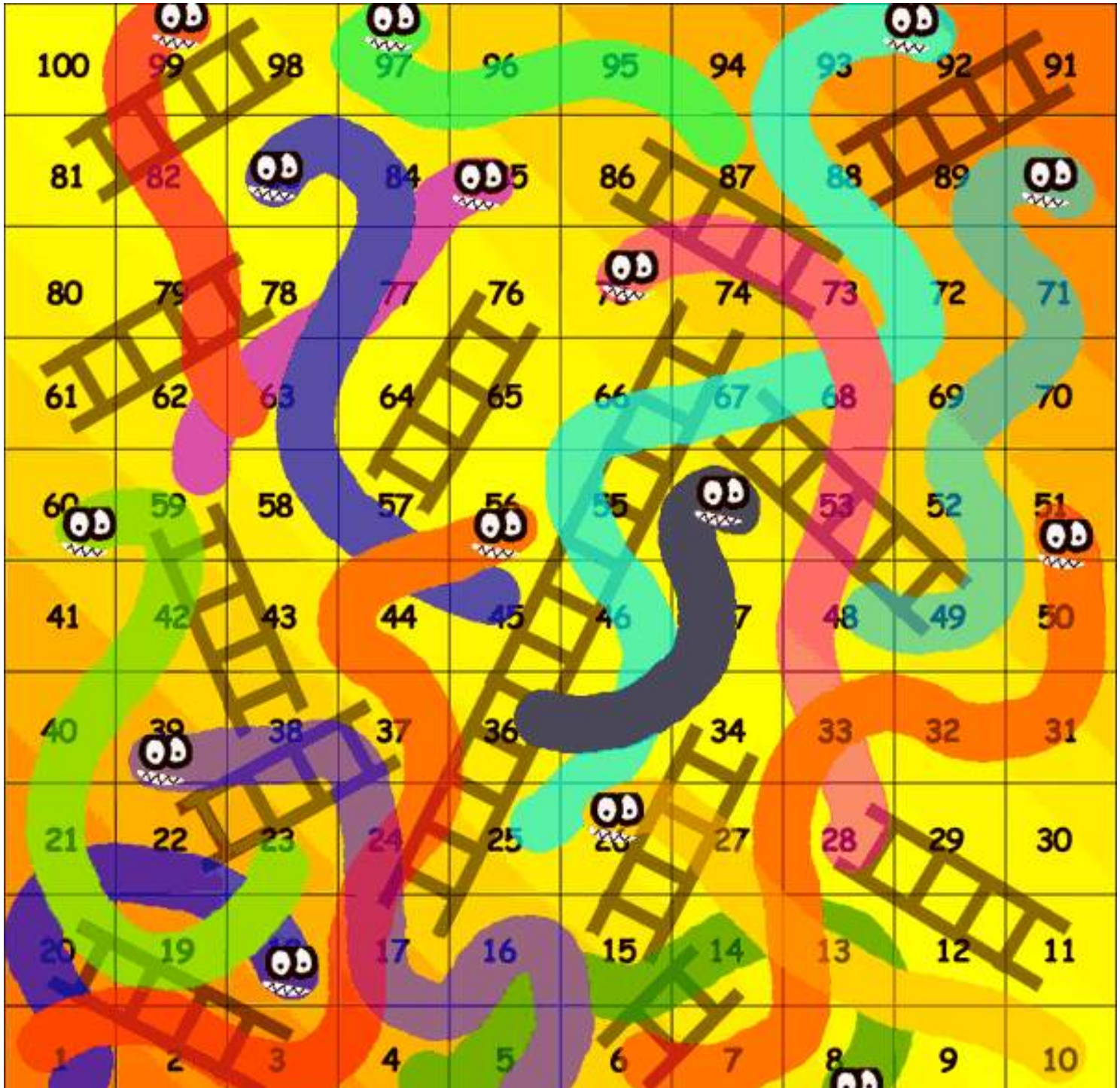
Example #4: Probabilities (Number of Turns in Snakes and Ladders)

Snakes and Ladders is a game played on a board of 100 numbered squares.

From the starting square, players repeatedly roll a die, trying to reach the final square first.

Certain squares have a ladder that moves you ahead; others are snakes that drag you backwards.

A simple question to investigate: *How many turns does it take for one player to reach the final square?*


100 99 98 97 96 95 94 93 92 91
81 82 84 86 87 89
80 79 78 77 76 74 73 72 71
61 62 63 64 65 66 67 68 69 70
60 59 58 57 56 53 52 51
41 42 43 44 45 46 48 49 50
40 38 37 36 34 33 32 31
21 22 23 24 25 27 28 29 30
19 17 16 15 14 13 12 11
1 2 3 4 5 6 7 9 10

Since dice are involved, there's an element of chance, so we have to revise our question and ask, what is the typical or average number of turns required for one player?

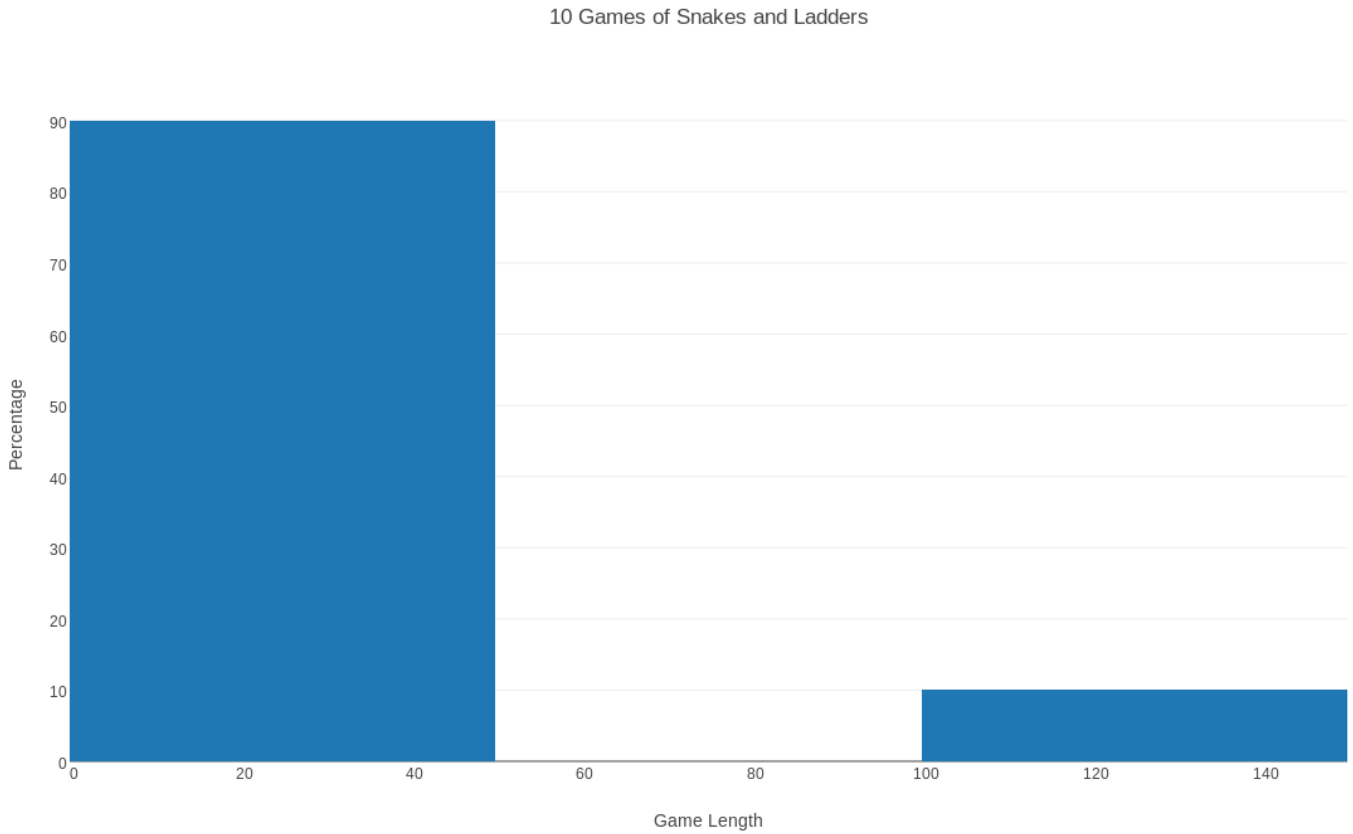
It's not easy to see a way to answer this question, except to play a lot of games and see what happens.

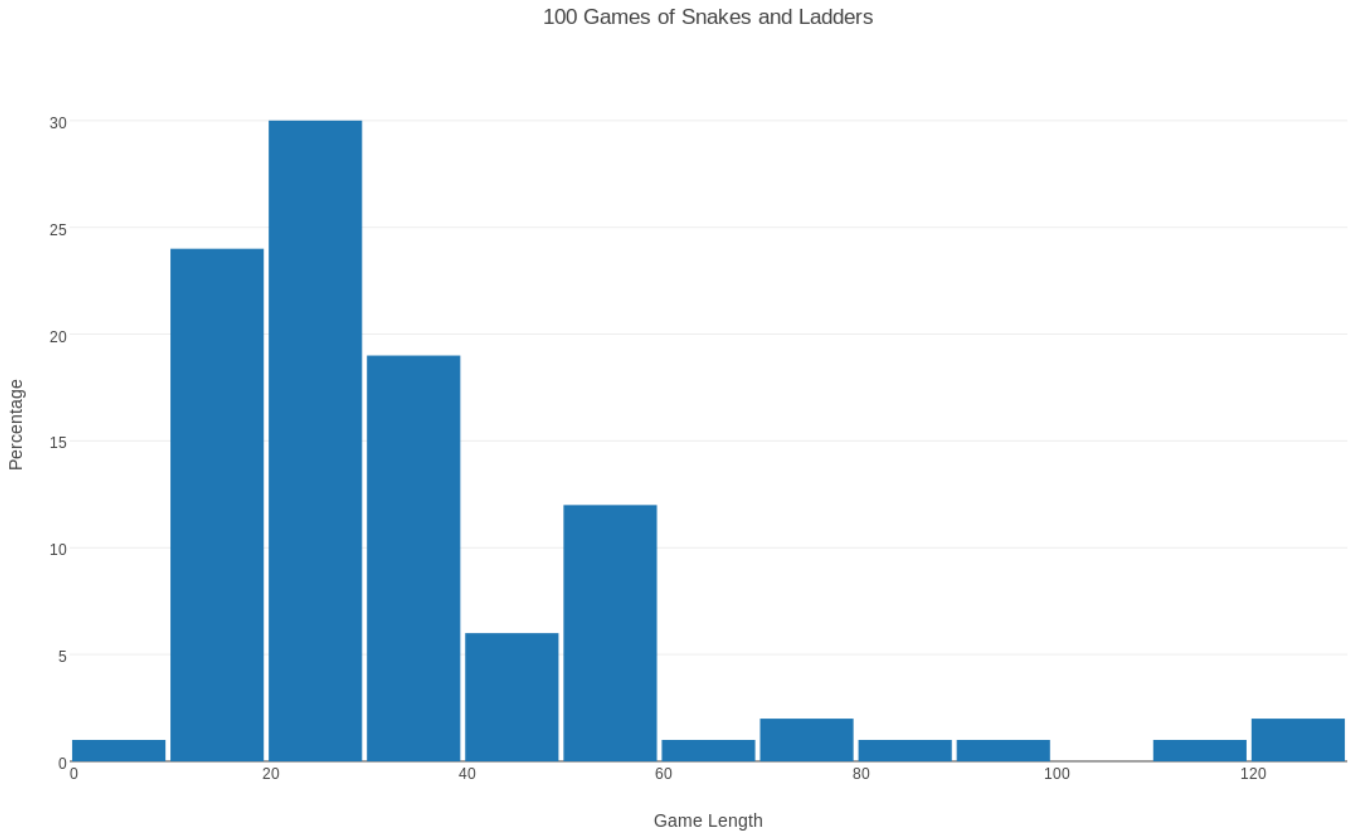
To keep things simple, we imagine a single player, who rolls the die over and over, following the rules, until reaching the end, and writing down the number of turns required.

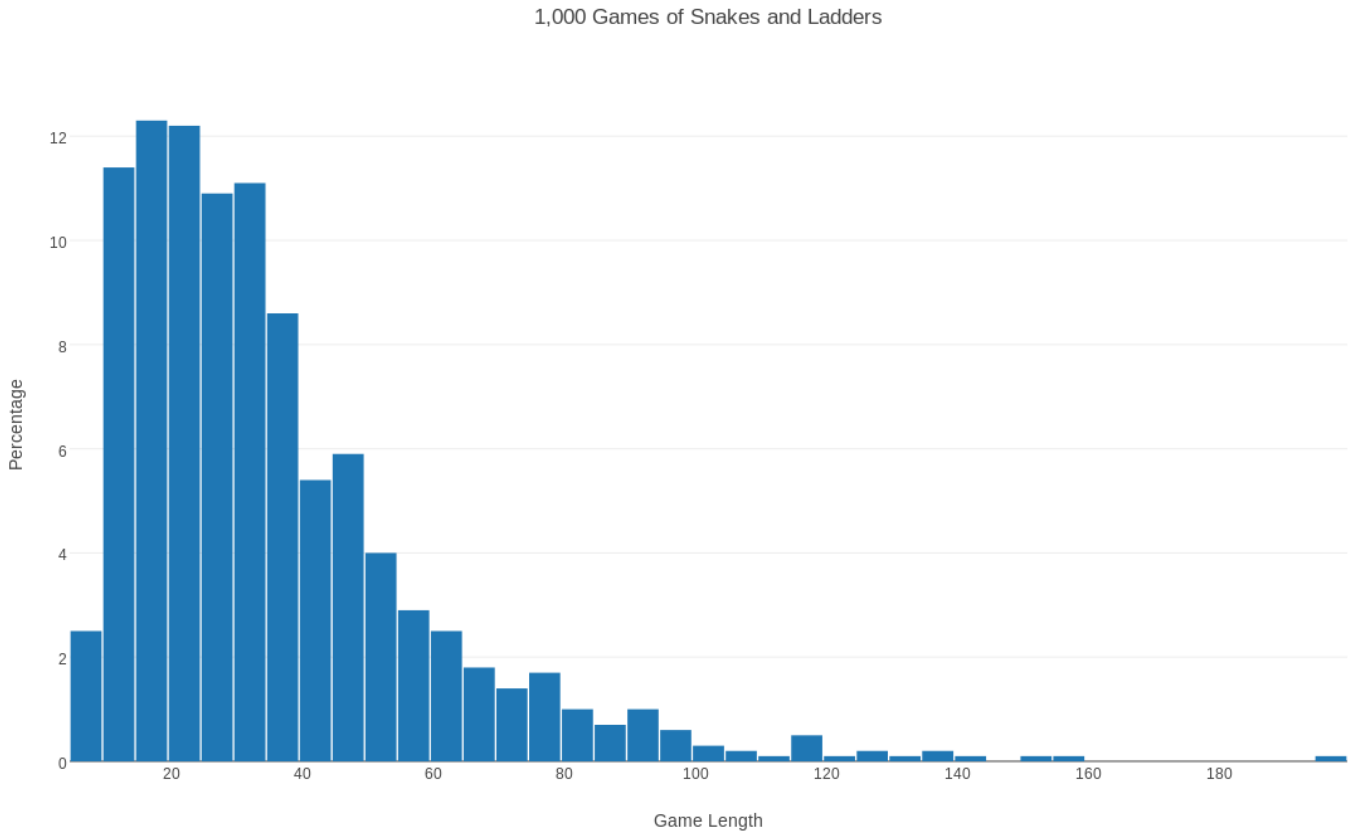
We could store our data in a file called **snakes_10.csv**.

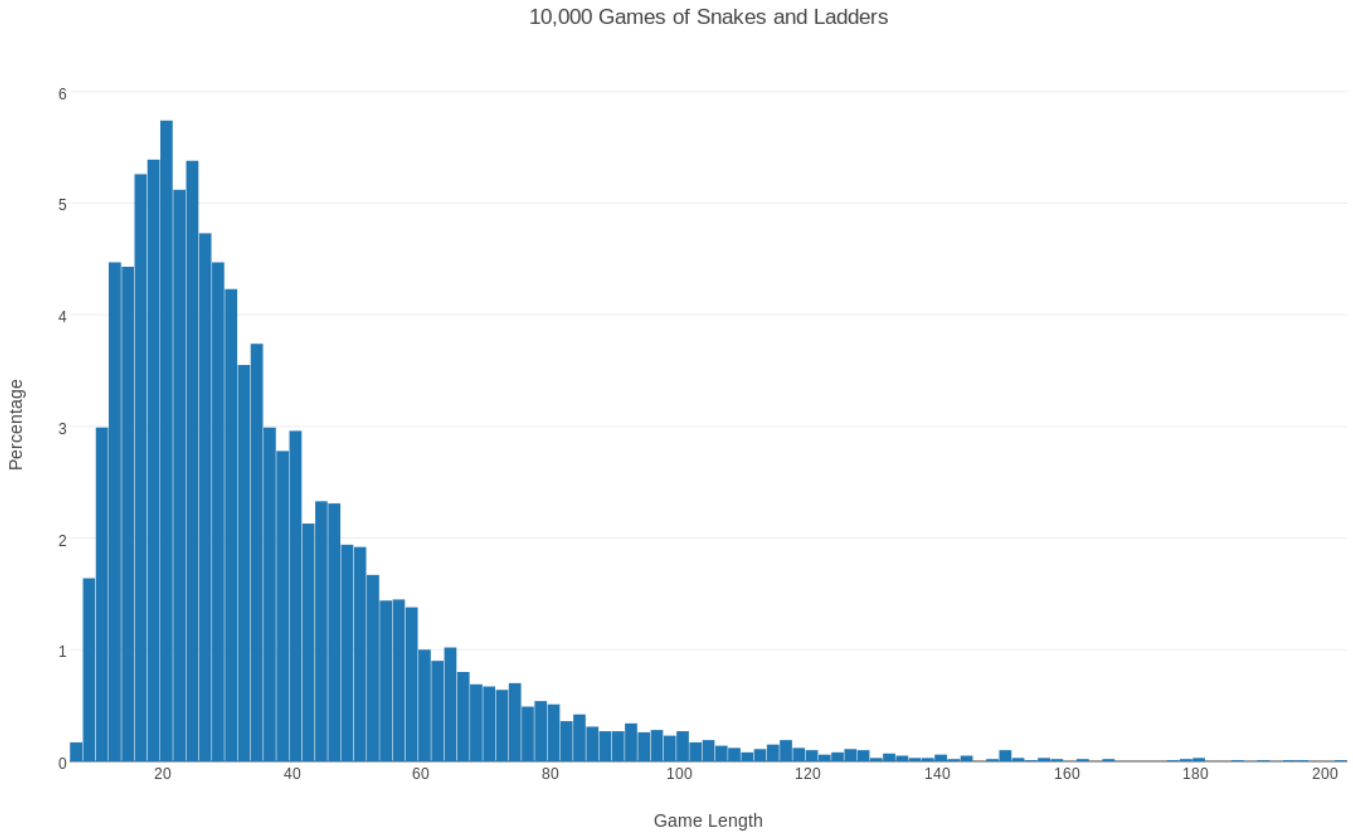
"Game Number",	"Game Length"
1,	30
2,	29
3,	31
4,	16
5,	24
6,	29
7,	28
8,	117
9,	42
10,	23

but that one long 117 turn game suggests we need more data!


10 Games of Snakes and Ladders
Percentage
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
120
140
Game Length


100 Games of Snakes and Ladders
Percentage
30
25
20
15
10
5
0
0
20
40
60
80
100
120
Game Length

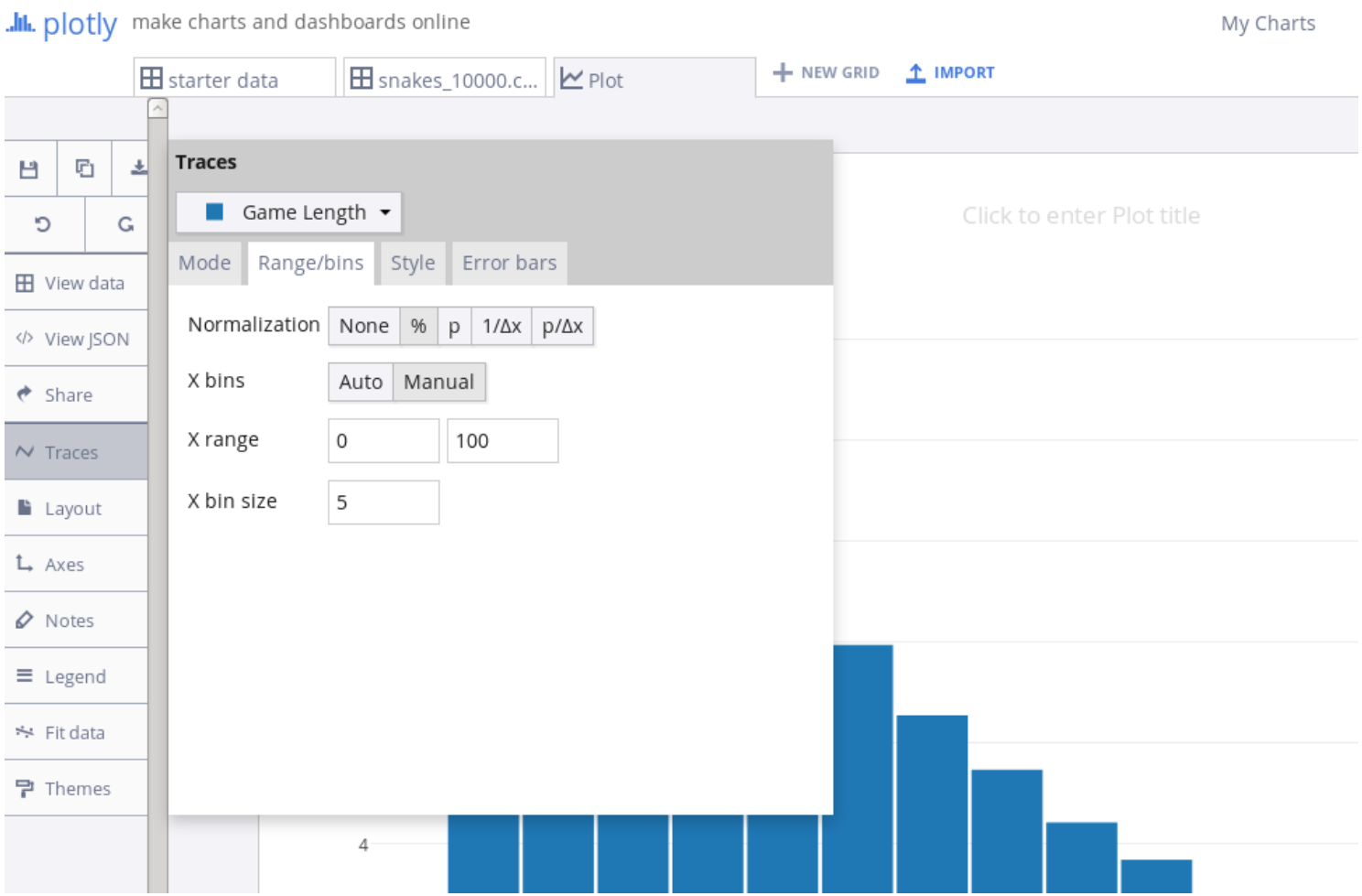

1,000 Games of Snakes and Ladders
Percentage
0
2
4
6
8
10
12
20
40
60
80
100
120
140
160
180
Game Length

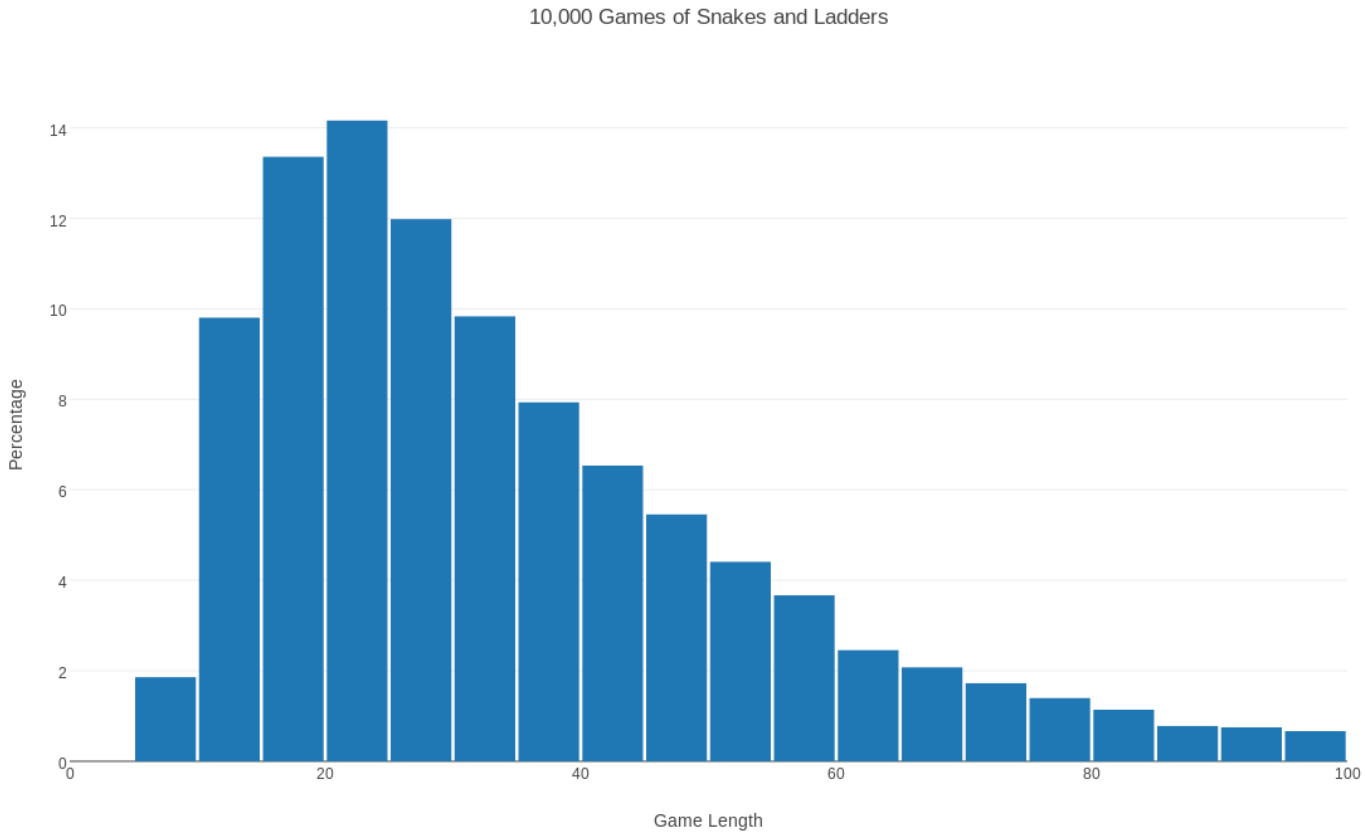

10,000 Games of Snakes and Ladders
Percentage
0
1
2
3
4
5
6
20
40
60
80
100
120
140
160
180
200
Game Length

Using data from 10,000 games, our histogram seems to have a fairly smooth shape, which suggests that the data has "settled down", and that the effects of chance have been averaged out.

However, we still can see a certain amount of saw-tooth or up-and-down variations in the plot. If we look in the **Traces** menu, we can see that our bin width is 2. If we use wider bins, we may be able to smooth out our plot. At the same time, we can reduce the maximum interesting value of Game Length to 100, which will allow us to focus on the most commonly occurring values.

The result is a simplified plot that is smoother and easier to read.


plotly
make charts and dashboards online
My Charts
starter data
snakes_10000.c...
Plot
NEW GRID
IMPORT
Traces
Game Length
Mode
Range/bins
Style
Error bars
Normalization
None
%
p
1/Δx
p/Δx
X bins
Auto
Manual
X range
0
100
X bin size
5
View data
View JSON
Share
Traces
Layout
Axes
Notes
Legend
Fit data
Themes
Click to enter Plot title
4


10,000 Games of Snakes and Ladders
Percentage
0
2
4
6
8
10
12
14
0
20
40
60
80
100
Game Length

Example #5: A Camel with Two Humps (Bimodal data)

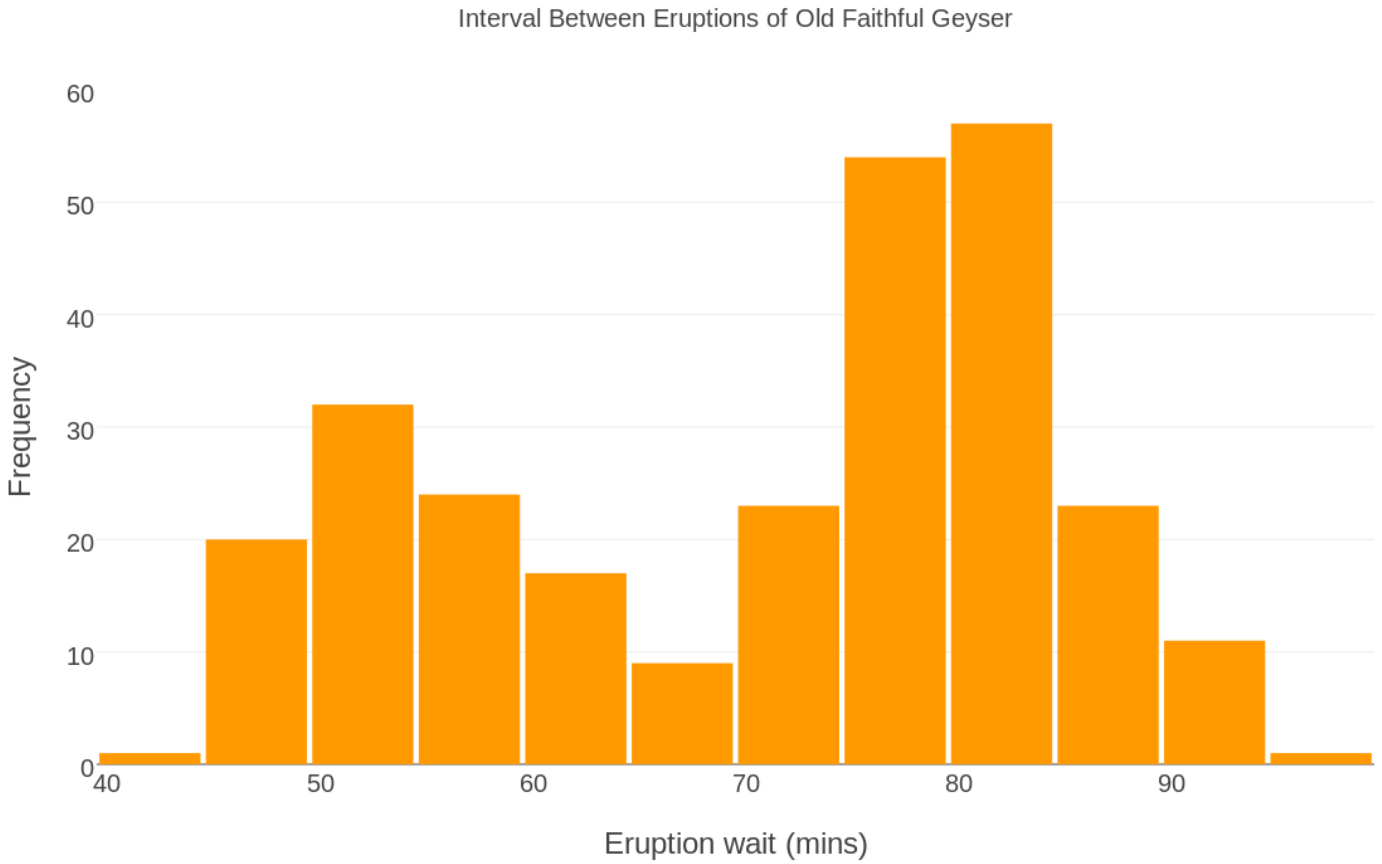
Sometimes a histogram can show an interesting pattern in the data. It may look like a double mountain range, having two peaks.

Often, a peak represents a typical behavior, and if the data has two peaks, then it suggests that sometimes the system wants to do one thing, and sometimes the other.

A peak in a histogram is sometimes called a **mode** and a histogram that has two peaks is called a bimodal histogram.

Our data is stored in the file **faithful.csv**.

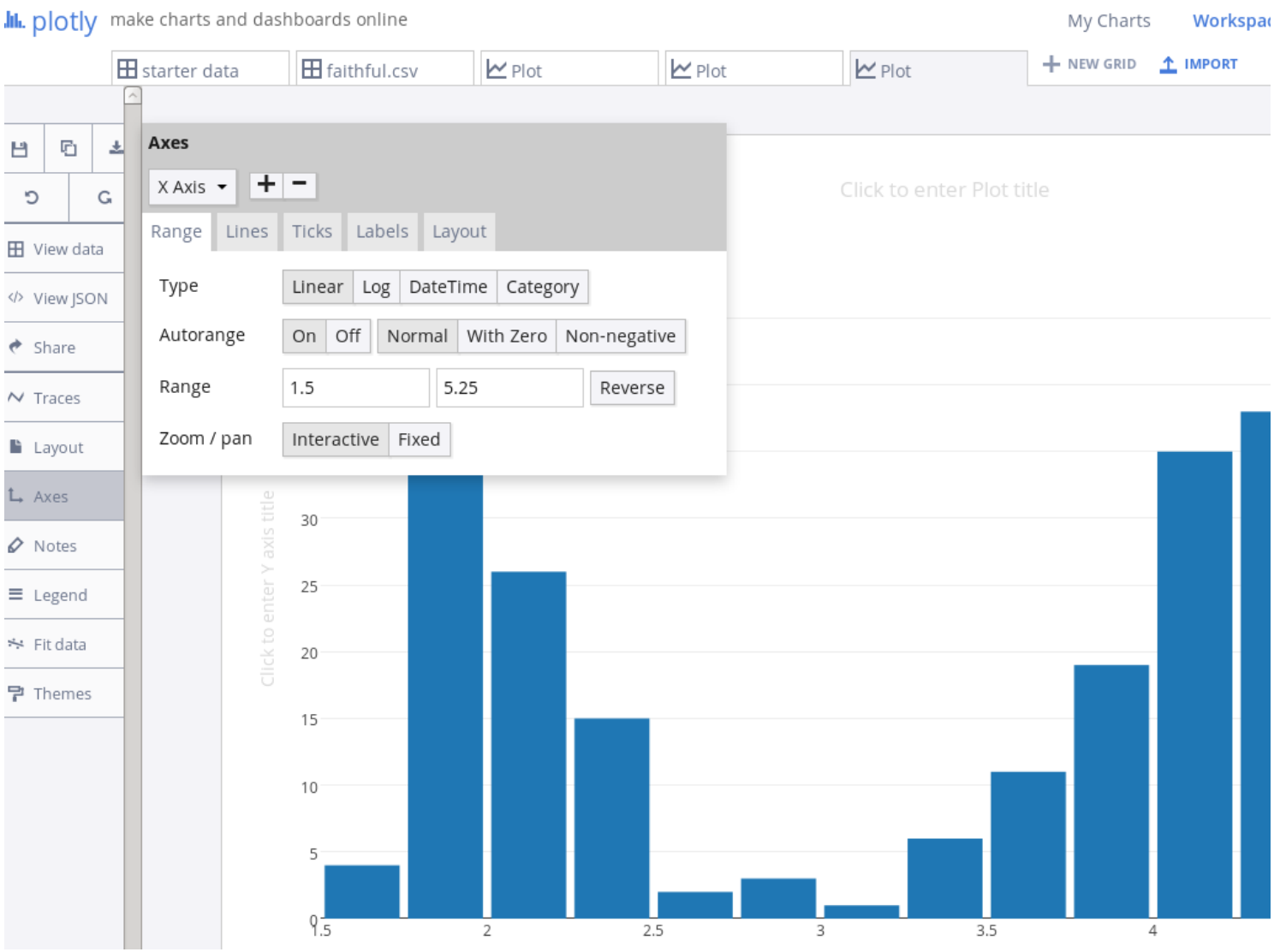
"Index",	"Eruption length (mins)",	"Eruption wait (mins)"
1,	3.600,	79
2,	1.800,	54
3,	3.333,	74
—	*more data*	—
270,	4.417,	90
271,	1.817,	46
272,	4.467,	74

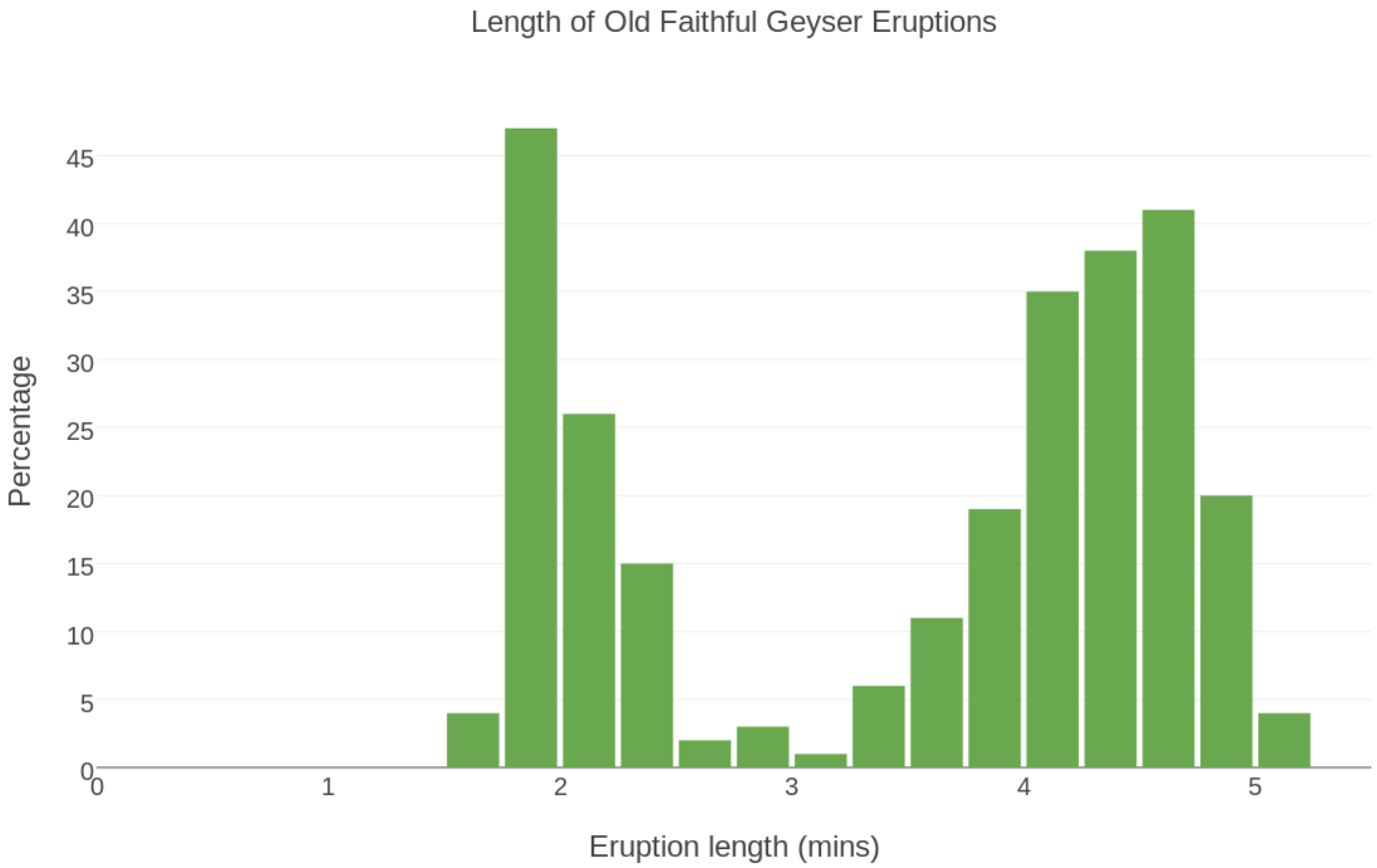

Interval Between Eruptions of Old Faithful Geyser
Frequency
60
50
40
30
20
10
0
40
50
60
70
80
90
Eruption wait (mins)

Now we will do a histogram of the duration of an eruption, that is, after however long we had to wait, how long the geyser actually sprays water upwards in a spectacular show.

Since the geyser has a short and a long waiting period, we might expect that sometimes the water doesn't have so much time to heat up, and other times it does, so that we have a corresponding short and long eruption pattern as well.

By default, the horizontal plot range starts with the smallest observed time. In order to better suggest the time scale, I went to the **colorredAxes left hand menu item and reset the plot range to start at 0.**


plotly
make charts and dashboards online
My Charts
Workspac
starter data
faithful.csv
Plot
Plot
Plot
NEW GRID
IMPORT
Axes
X Axis
Range
Lines
Ticks
Labels
Layout
Type
Linear
Log
DateTime
Category
Autorange
On
Off
Normal
With Zero
Non-negative
Range
1.5
5.25
Reverse
Zoom / pan
Interactive
Fixed
View data
View JSON
Share
Traces
Layout
Axes
Notes
Legend
Fit data
Themes
Click to enter Plot title
Click to enter Y axis title
0
5
10
15
20
25
30
1.5
2
2.5
3
3.5
4


Length of Old Faithful Geyser Eruptions
Percentage
45
40
35
30
25
20
15
10
5
0
0
1
2
3
4
5
Eruption length (mins)