*

I Know It When I See It

- or -

Looking for Patterns with Scatter Plots

ISC1057

Janet Peterson and John Burkardt

Computational Thinking

Fall Semester 2016

Scientists look for patterns in nature.

To discover such patterns, they identify interesting and measurable features, and make many observations of how these quantities change.

For example, they may measure the length of the hands and feet of many individuals, and notice that the two measurements seem to be related.

Or they may record the mileage on a car, and measure the amount of wear on its tires.

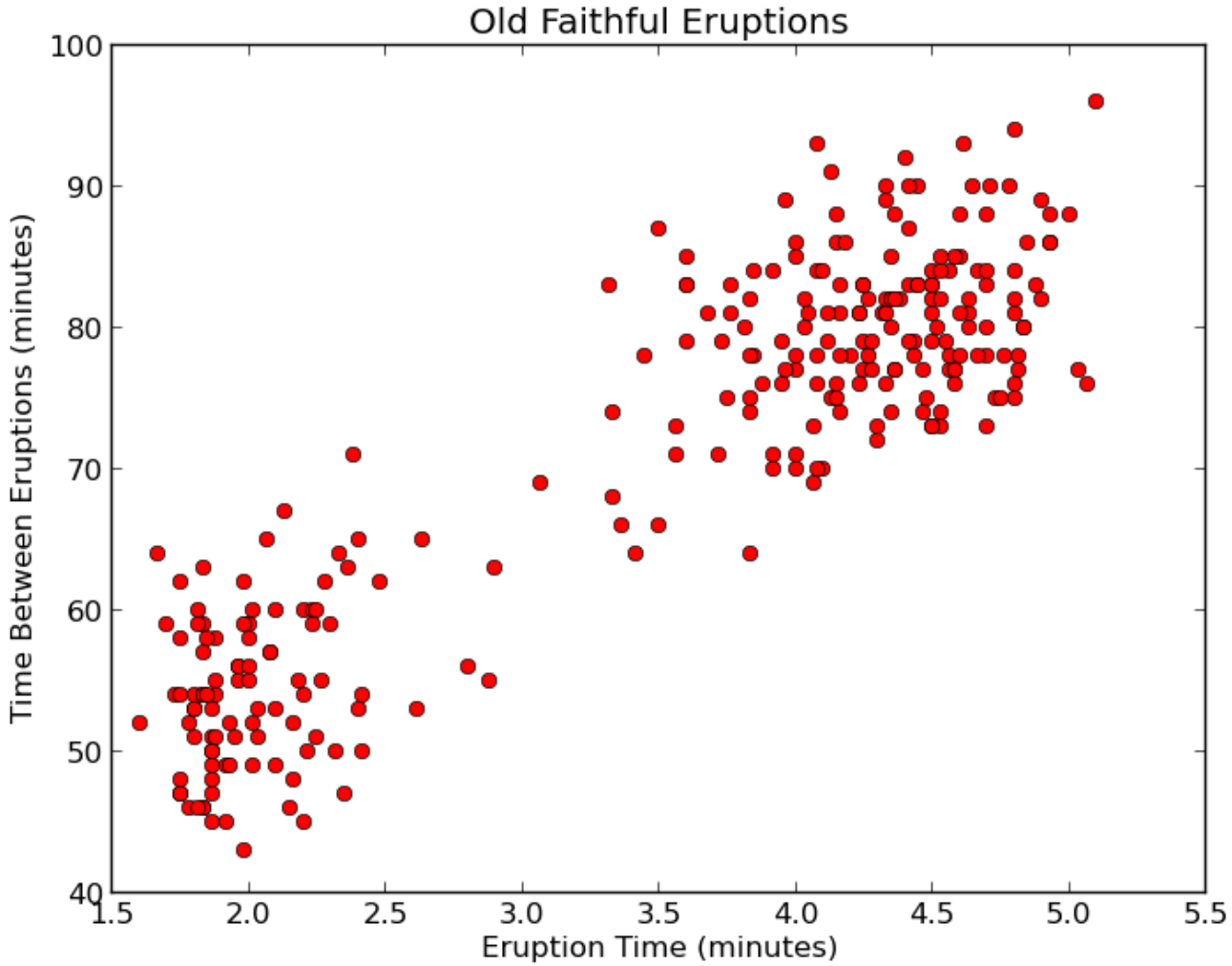
But even if we gather many measurements of such cases, it's much easier to see a pattern in a plot, than in a table of values.

A scatterplot is the simplest way to investigate the relationship between a pair of measured quantities.

For example, a very patient scientist watched the Old Faithful geyser, recording how long the geyser eruption lasted, and then how long it "rested" afterwards. This resulted in 272 pairs of measurements:

Index	Duration	Rest
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

While these numbers tell the whole story, we can't see it. But if we make a scatterplot of the data, some interesting patterns are visible.


Old Faithful Eruptions
Time Between Eruptions (minutes)
Eruption Time (minutes)
100
90
80
70
60
50
40
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5

Once a scatterplot has been made, we can make a judgment about whether our data is trying to tell us something.

For instance, our data might be:

- completely disorganized;
- organized into a blob;
- stretched out into a linear pattern;
- broken up into two or more clusters.

A scatterplot is a quick way to detect whether such patterns exist, and to guide a guess as to what that pattern might be.

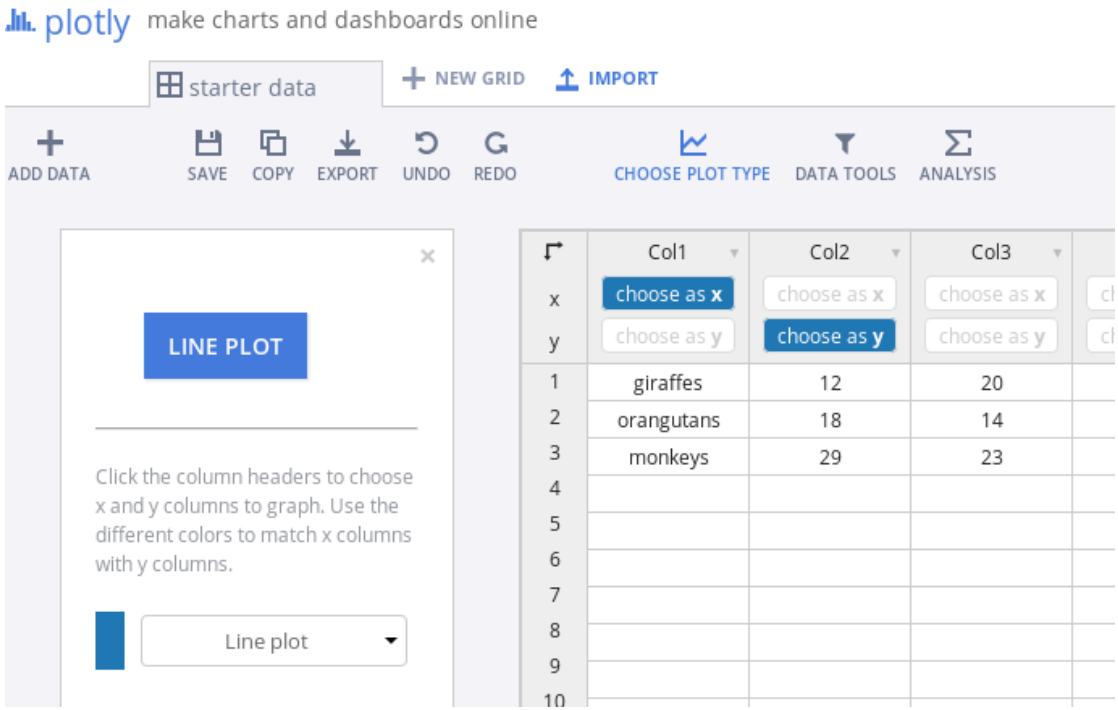
Example #1: Tallahassee Cab Fares

A scientist has taken several cab rides in Tallahassee, and has recorded the mileage and the fare. The data looks like this:

Miles	Fare ($)
4.5	18.00
26.7	73.75
6.7	23.00
16.4	56.00
32.7	83.25
5.7	17.50
77.0	190.50
8.3	19.65

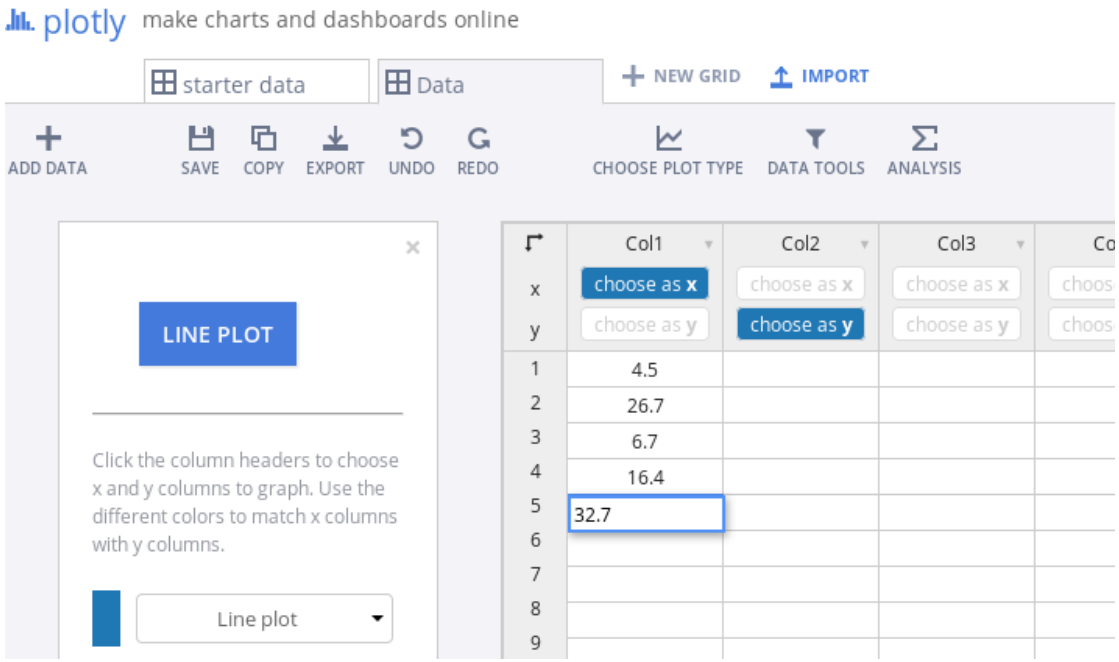
We'll search for a pattern assuming the fare depends on the mileage, so we can try a scatter plot using mileage as the x or independent variable and the fare as the y or dependent variable.

To make a scatter plot, we can begin by logging into Plotly at **https://plot.ly/plot**. There is already some sample data entered in the Plotly table.



Option 1; Type in the Data:

Clear out the sample data by selecting $+$ **New Grid**. Then we can start typing in our data:



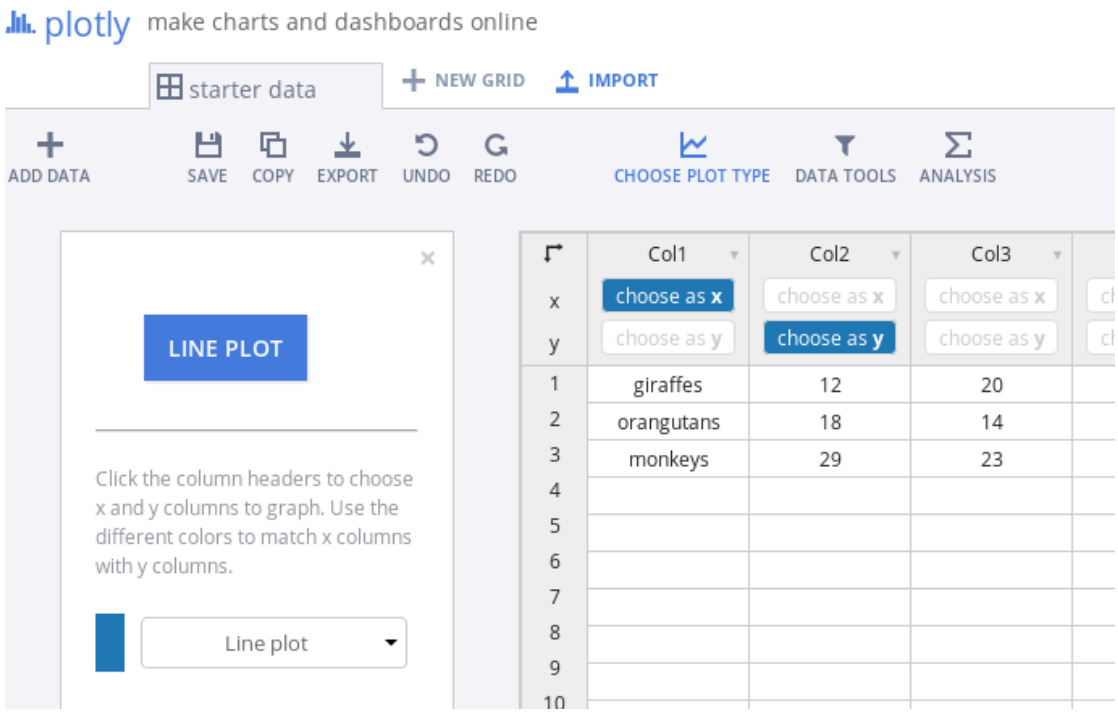
Option 2: Prepare your data in a CSV file.

A Comma Separated Value or "CSV" file is a text file in which each line contains a data observation. The first line of the file can include labels for the data, as long as these are enclosed in quotes.

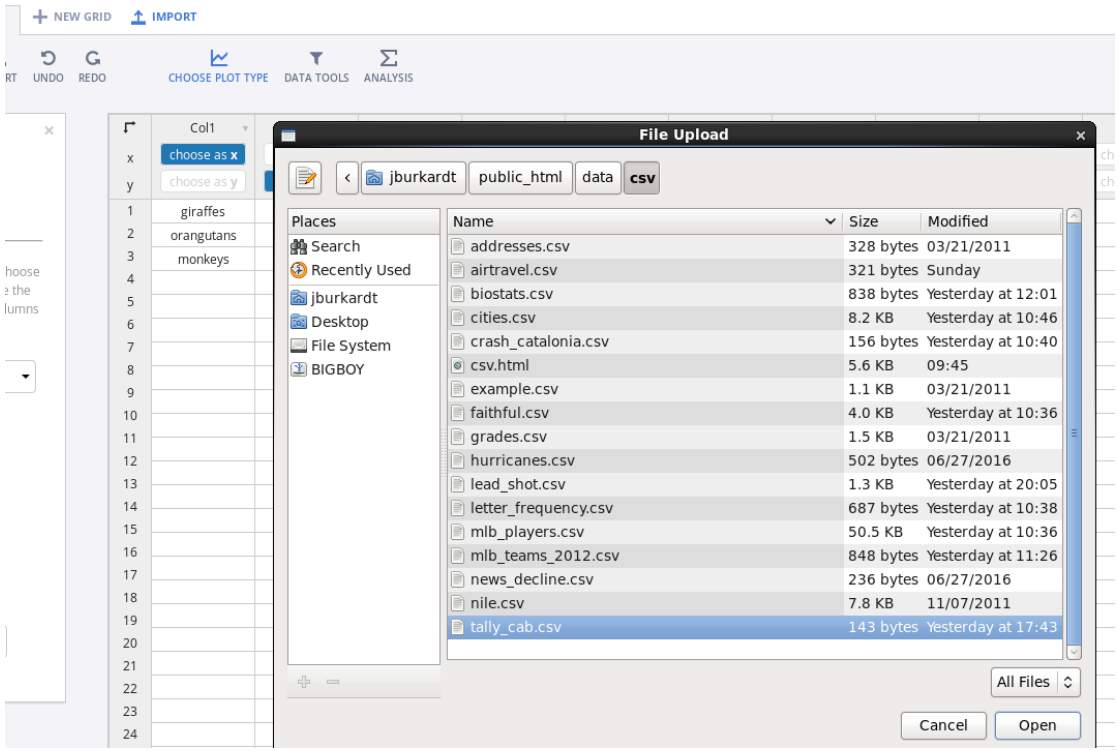
We might prepare a file "taxi.csv" looking like this:

"Miles",	"Fare ($)"
4.5,	18.00
26.7,	73.75
6.7,	23.00
16.4,	56.00
32.7,	83.25
5.7,	17.50
77.0,	190.50
8.3,	19.65

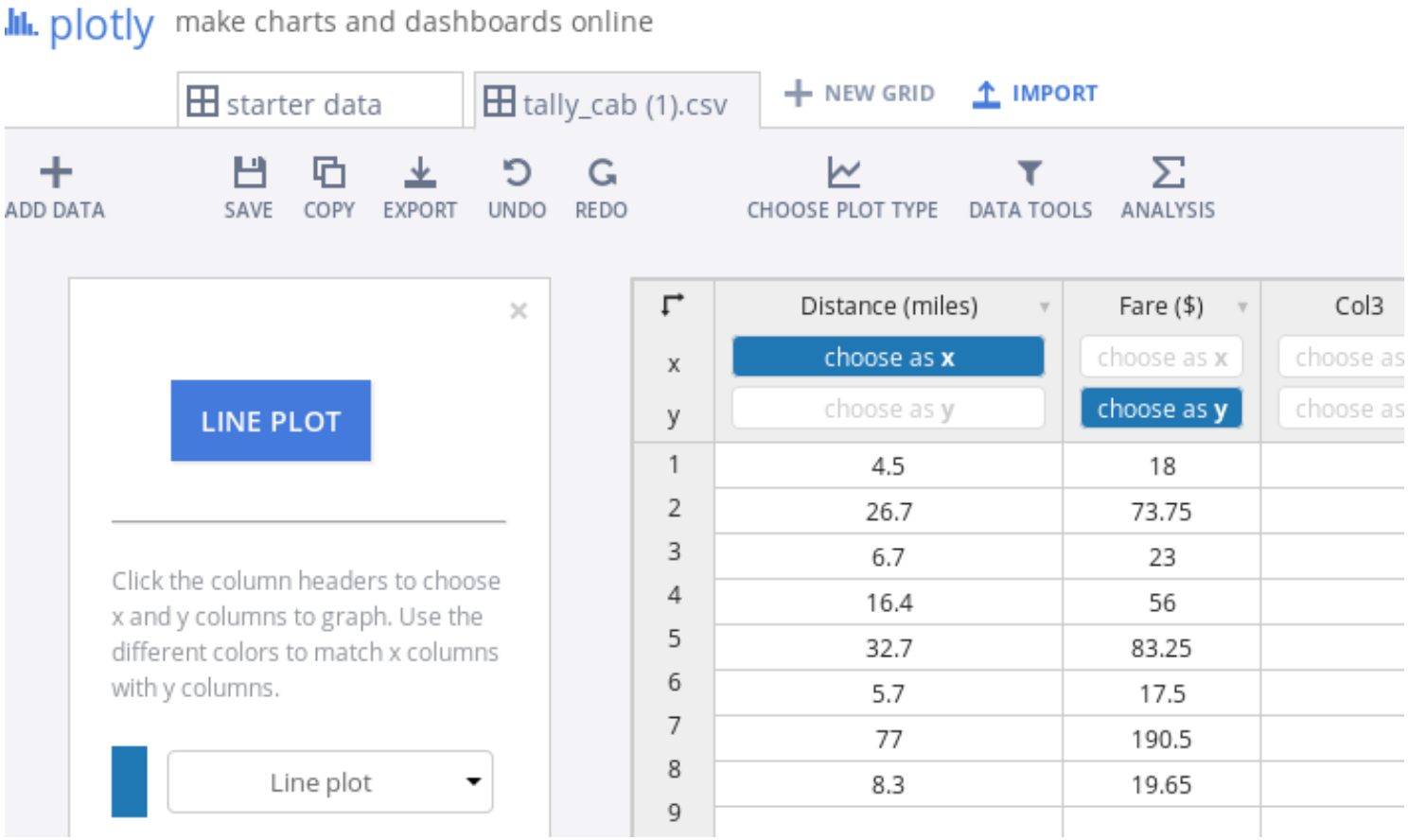
After we have prepared our data in a CSV file, we log into Plotly, as usual. There is already some sample data entered in the Plotly table, but we want to replace this with our file of taxi data. We do this by going to the "import" menu item.



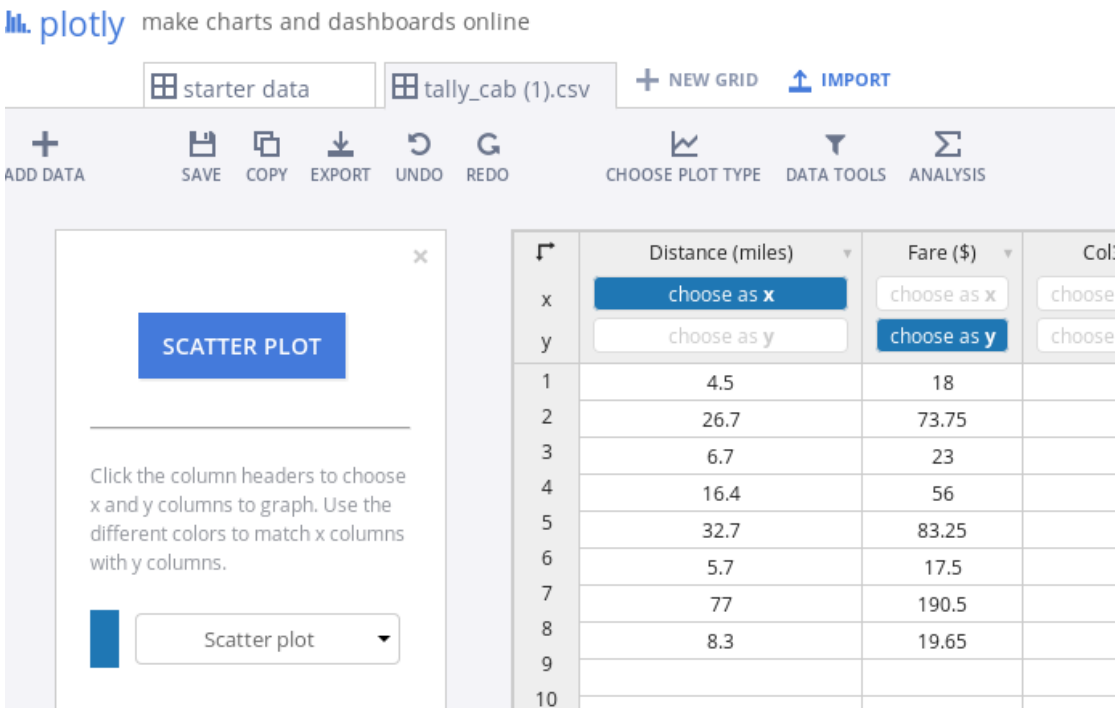
Plotly will provide a menu to help us locate the data file on our computer. Once we have found it and highlighted it, we press **Open** to bring the data into Plotly.



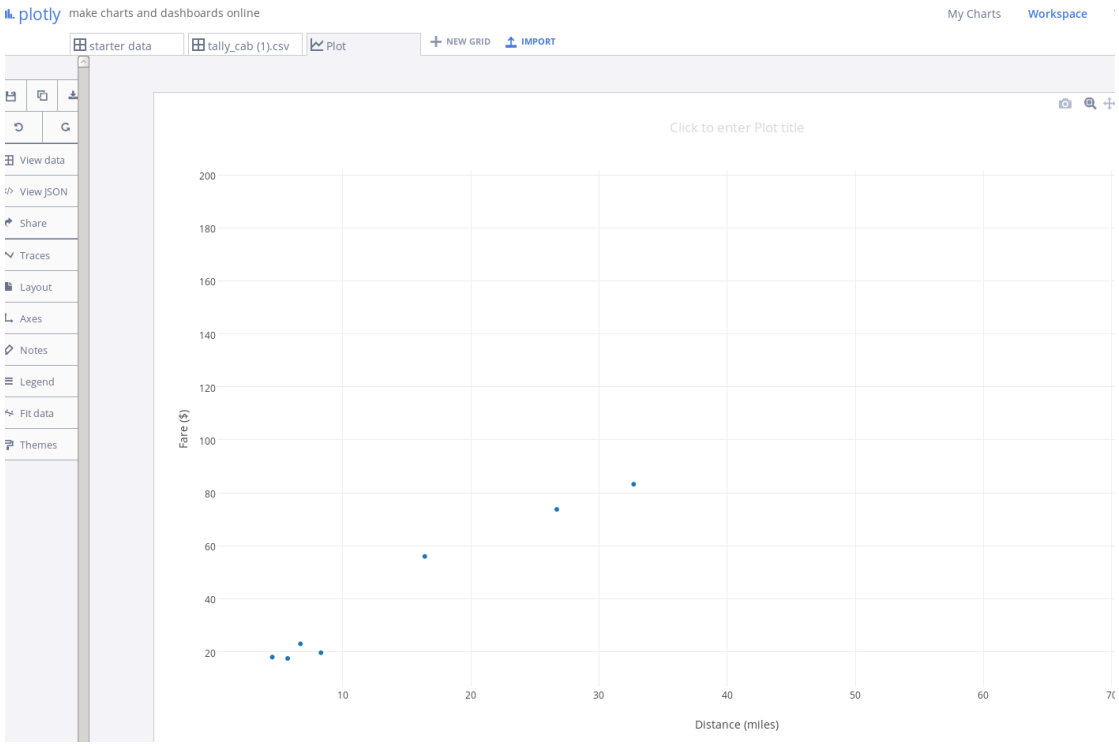
Now the data will show up in the Plotly grid. If the CSV file included an initial header line that labeled the data, then these labels will appear at the tops of the columns in the Plotly grid.



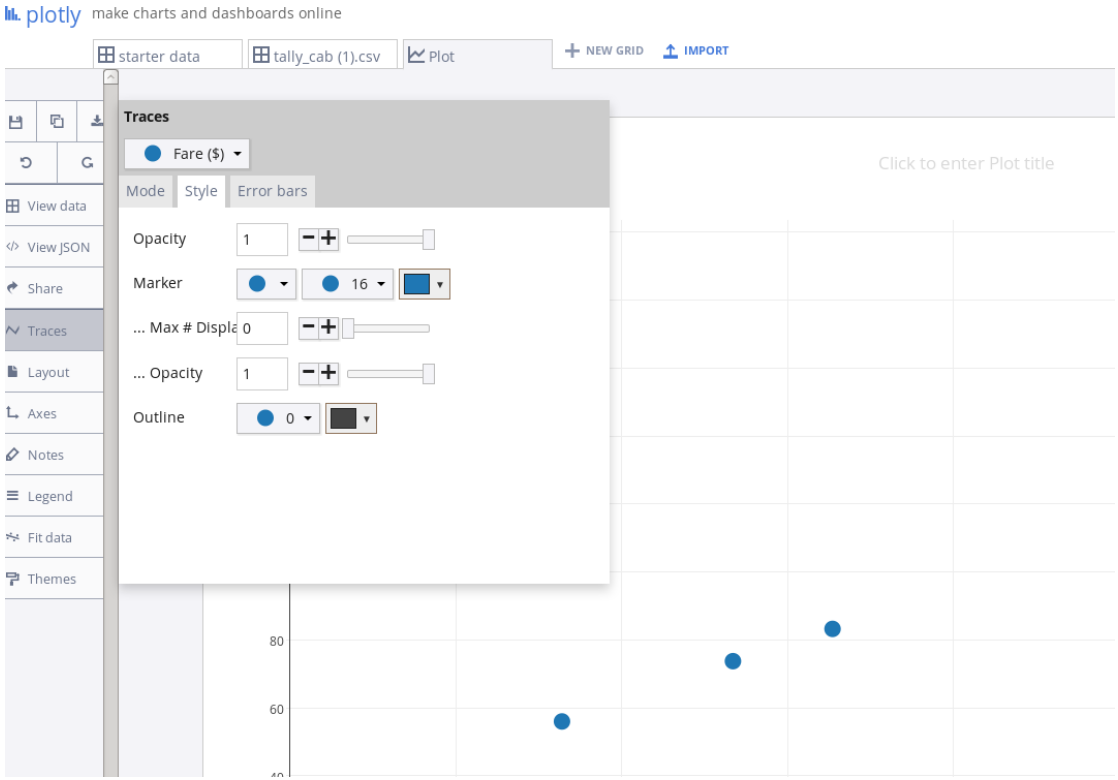
Plotly always starts by assuming we want a line plot. To get a scatter plot, click on the white box that says **Line plot**, and it will display other optons. Choose **scatter plot**. The big blue box will automatically also change to **Scatter plot**. Now press the big blue box to make your plot.



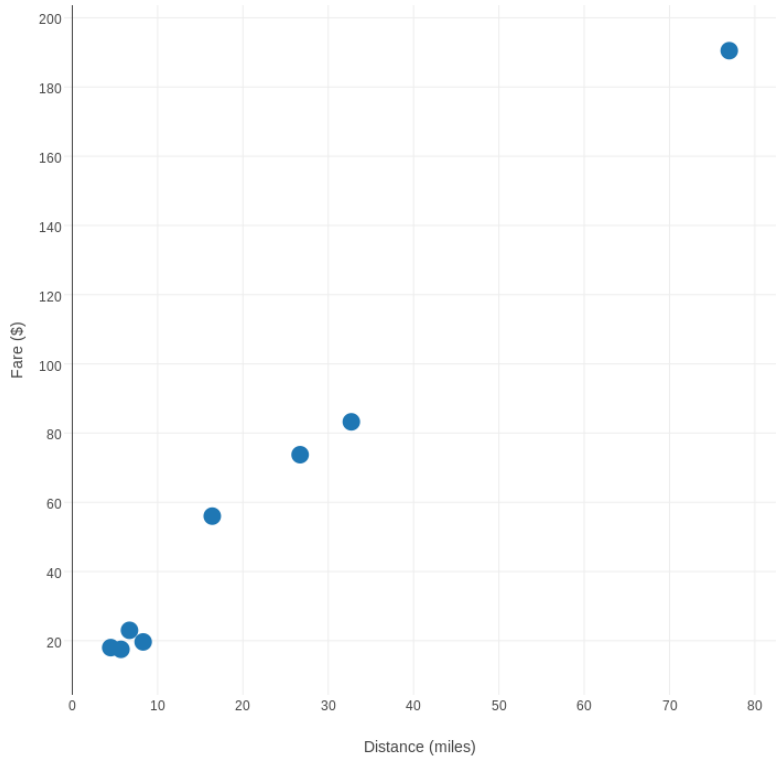
Now we see the plot of our data. We don't have many data points, and Plotly likes to use small dots, so we can't see much. Luckily, we can change the style of the plot.



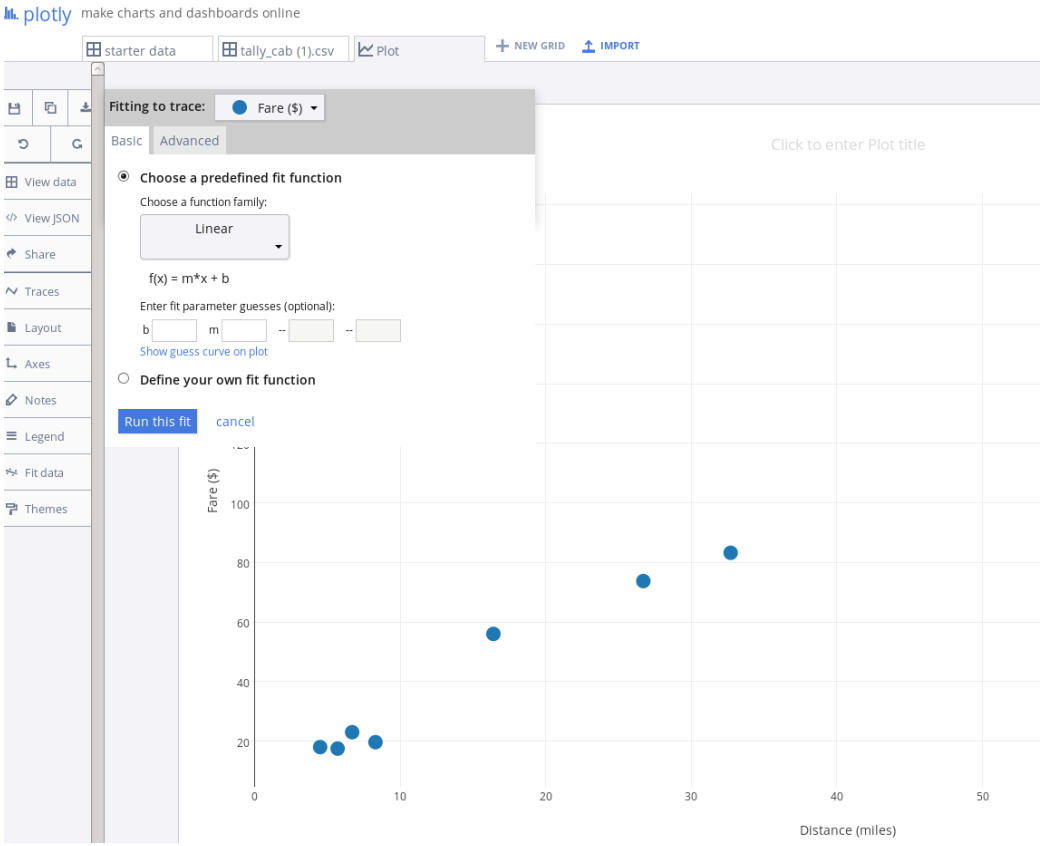
There is a menu strip on the left of the plot. If you select Traces, you will see a box. The tab we want is **Style**. The style option we want is **Marker** option, which controls the color, size, and shape of the dots. The size is usually 6, but let's increase it to 16.



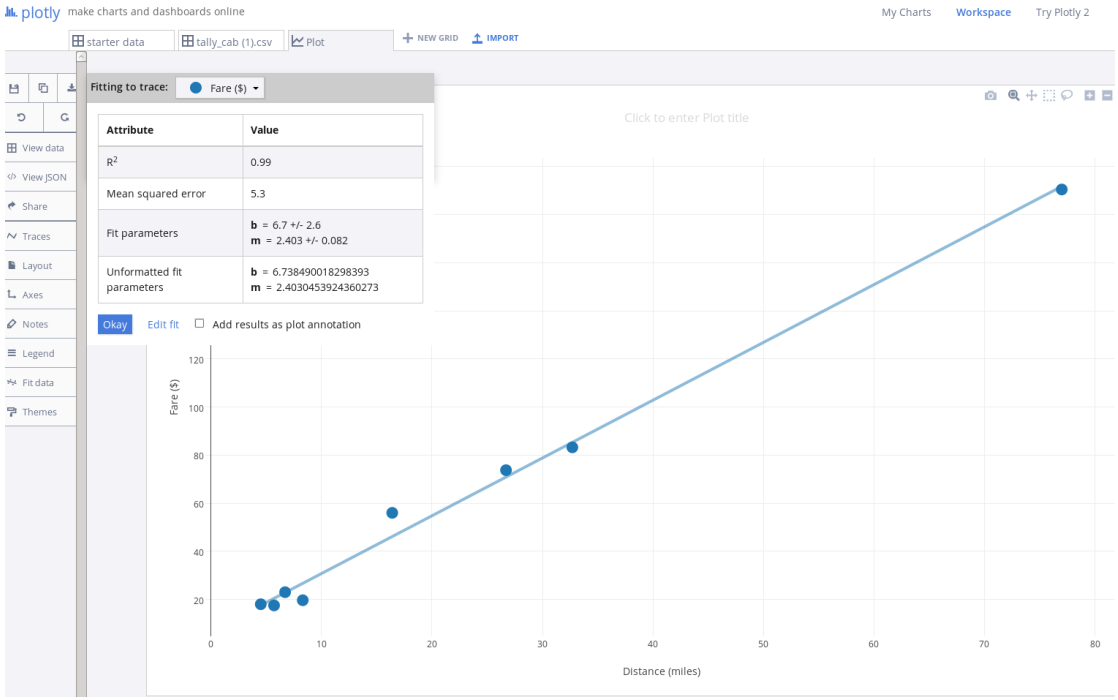
Here's our scatter plot. The shape of the data certainly suggests a simple relationship between mileage and fare.



Select the item **Fit Data** on the menu strip on the left. By default, this option will find a straight line suggested by the shape of the data. Select **Run this fit** to add the line to the plot.



Here's our final scatter plot, including the line suggested by the data.



Example #2: Age, Height and Weight for our Office Staff

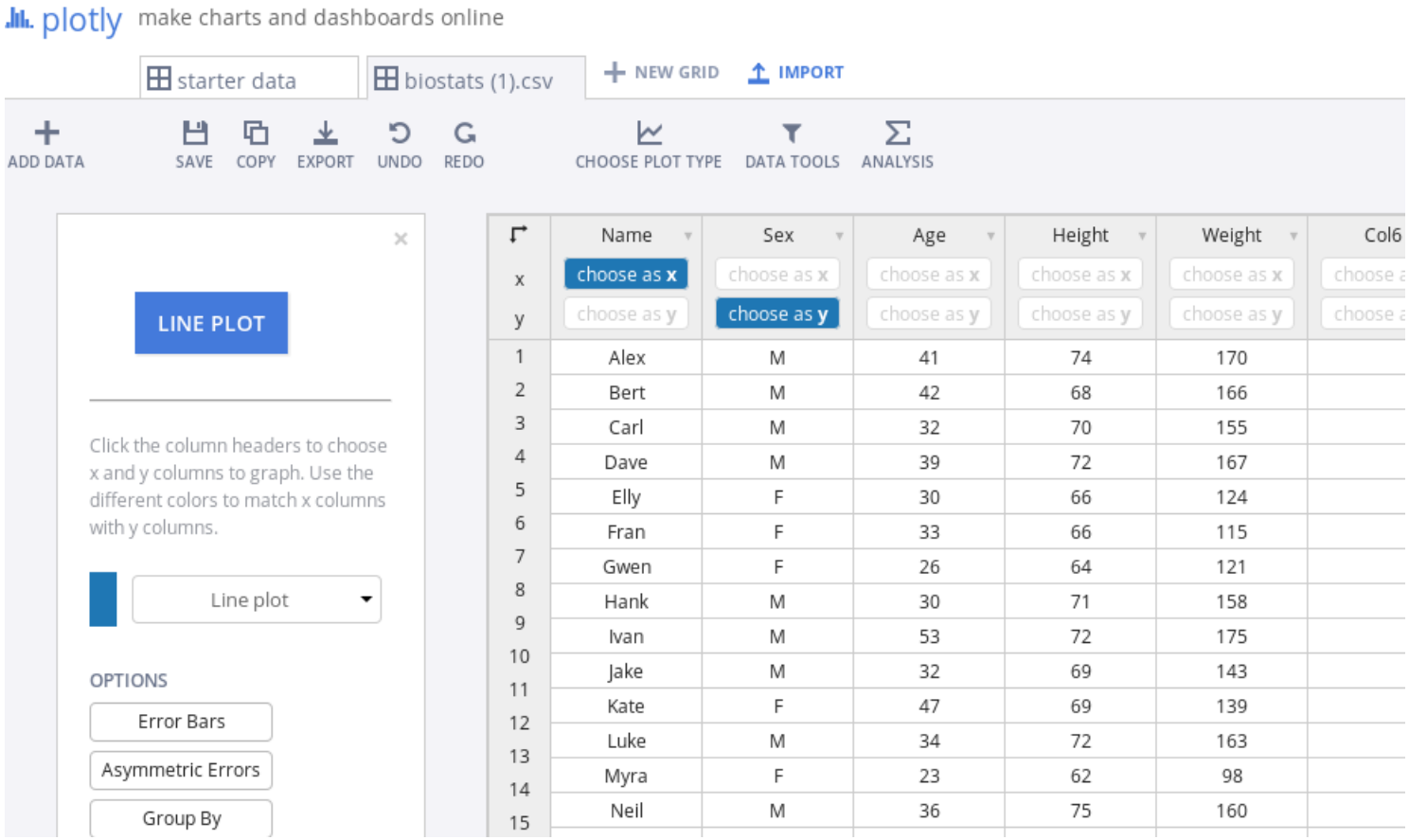
In order to get health insurance, all 18 people in our office must have a medical exam which records, among other things, their age, height and weight.

It's natural to expect some relationship between these quantities; tall people are likely to weight more, for instance.

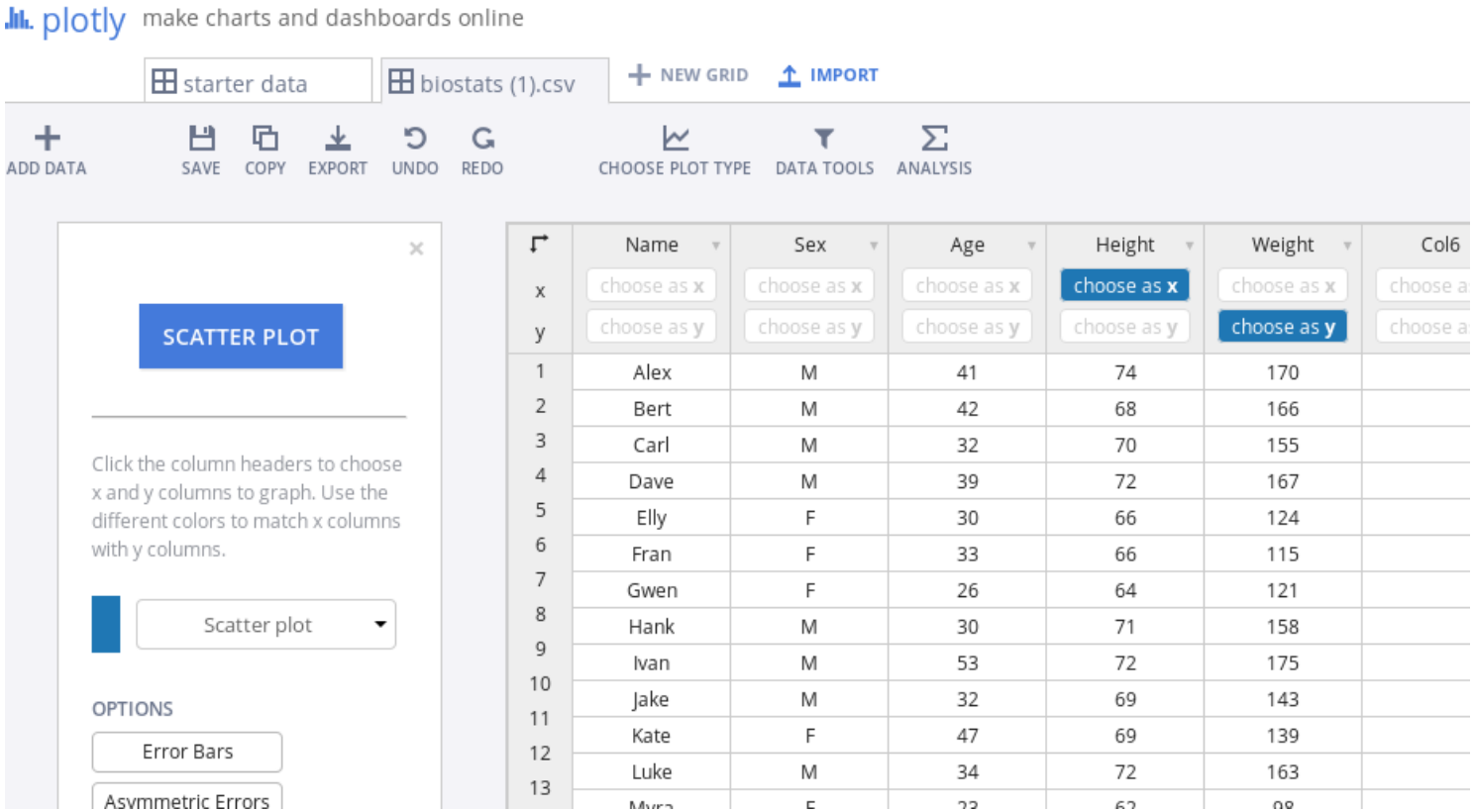
Scatterplots can help us to look into this matter. We have three measurements to work with so there are several plots we can make from this data.

“Name”	“Sex”	“Age”	“Height”	“Weight”
“Alex”	M	41	74	170
“Bert”	M	42	68	166
“Carl”	M	32	70	155
“Dave”	M	39	72	167
“Elly”	F	30	66	124
“Fran”	F	33	66	115
“Gwen”	F	26	64	121
“Hank”	M	30	71	158
“Ivan”	M	53	72	175
“Jake”	M	32	69	143
“Kate”	F	47	69	139
“Luke”	M	34	72	163
“Myra”	F	23	62	98
“Neil”	M	36	75	160
“Omar”	M	38	70	145
“Page”	F	31	67	135
“Quin”	M	29	71	176
“Ruth”	F	28	65	131

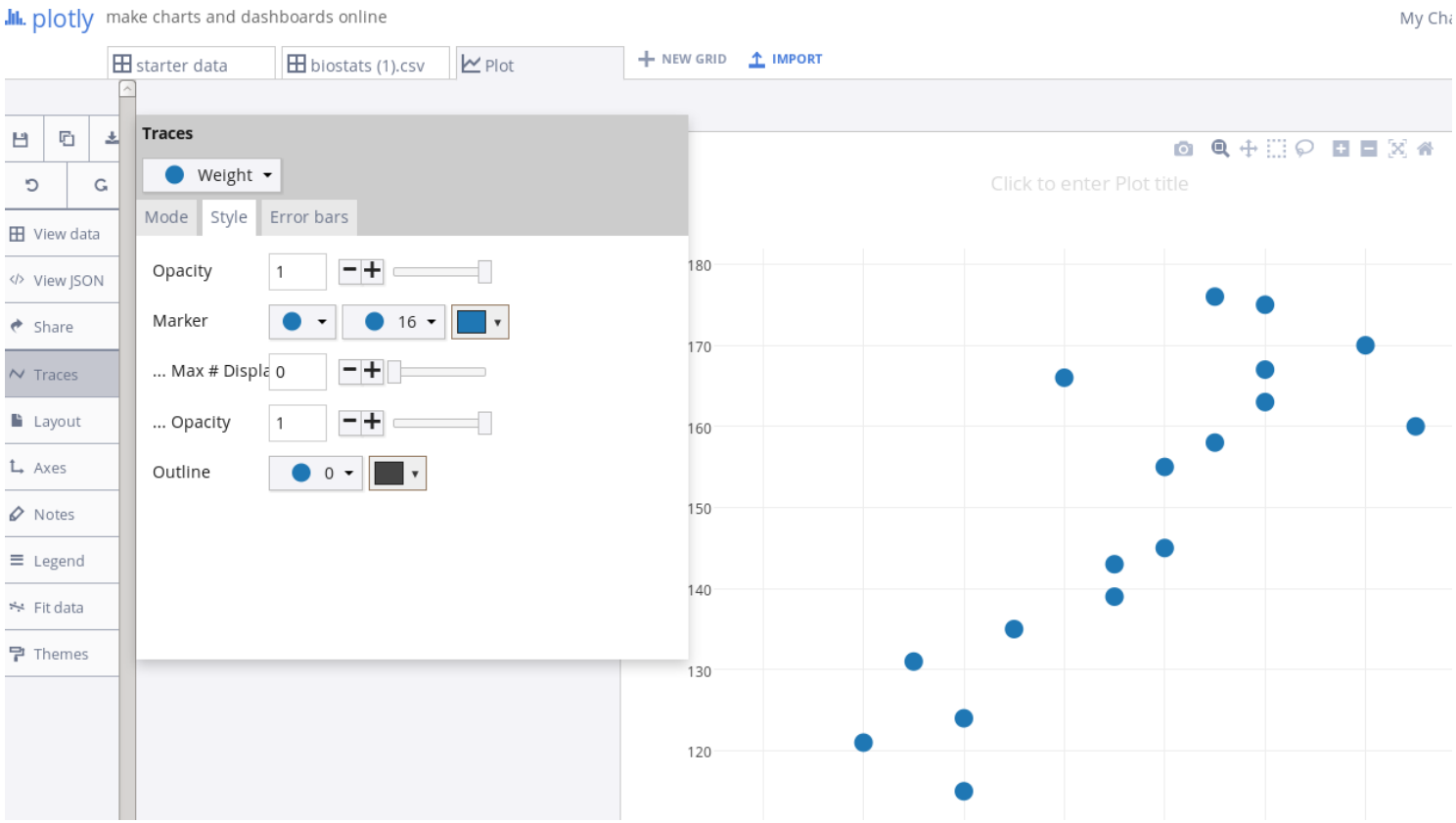
We can put this data into a csv file called **biostats.csv** and import it, or else wait til we log into Plotly, and type it in directly. In either case, we'll get a 5 column Plotly grid something like this:



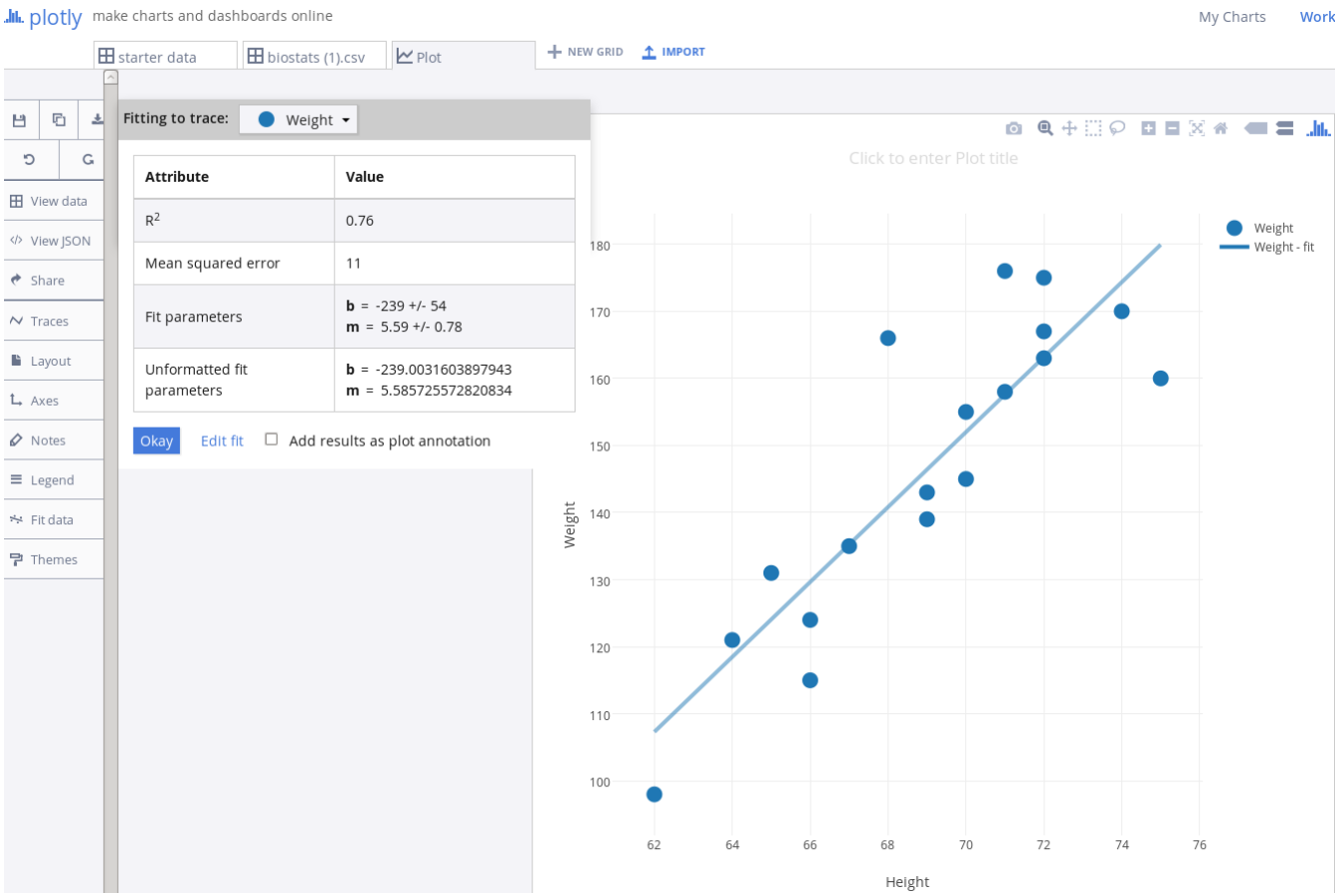
Plotly wants to make a line plot column 1 $(choose\ as\ x)$ versus column 2 $(choose\ as\ y)$ but instead we want a scatter plot, of height versus weight. We need to adjust the Plotly grid so it looks like this:



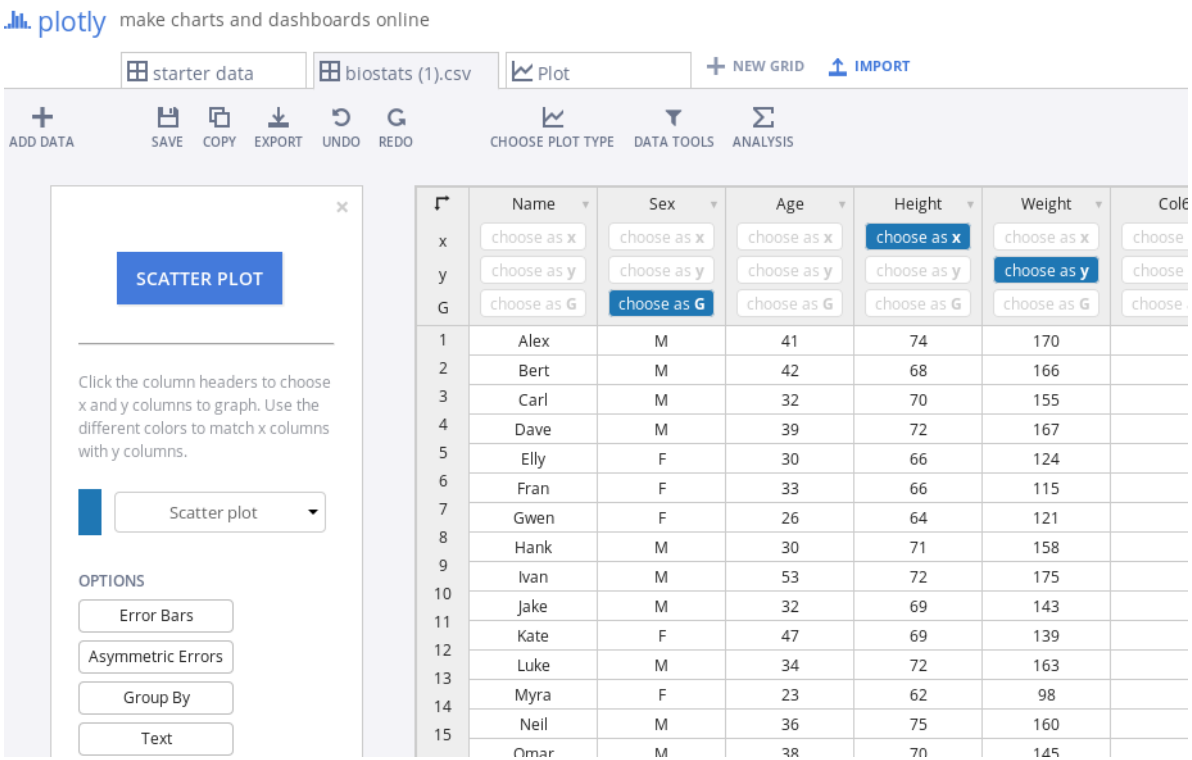
When we press Scatterplot, the dots are too small again, so we go to the left hand Traces option, choose Style and then Marker and increase the dot size to 16:



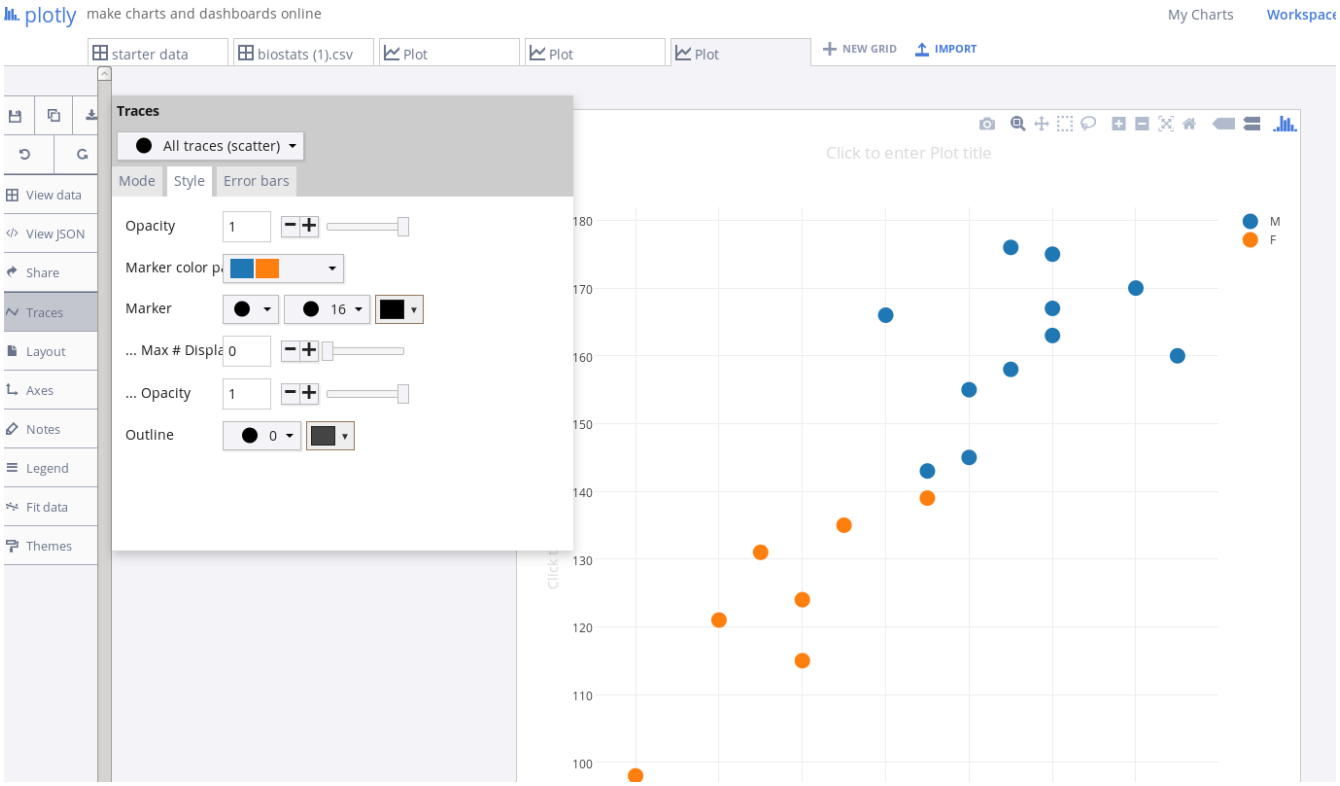
Because the data seems to suggest a line, we can choose the left hand **Fit data** menu option, **Add fit to this trace**, and **Run this fit** so that a line is drawn suggesting the pattern.



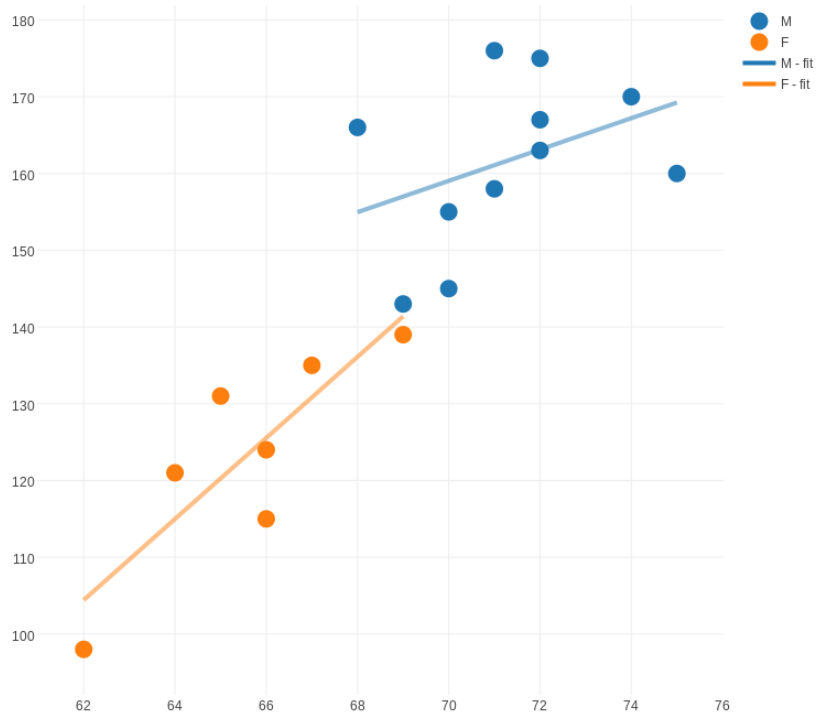
Our data includes men and women, so we wonder if the plot can illustrate any sex differences. Since column 2 of our data is an $\mathbf{M}$ or $\mathbf{F}$, we can select the $\mathbf{Group\ by}$ item on the left, and under column 2, highlight Use as G, to ask that the two sets of data be distinguished.



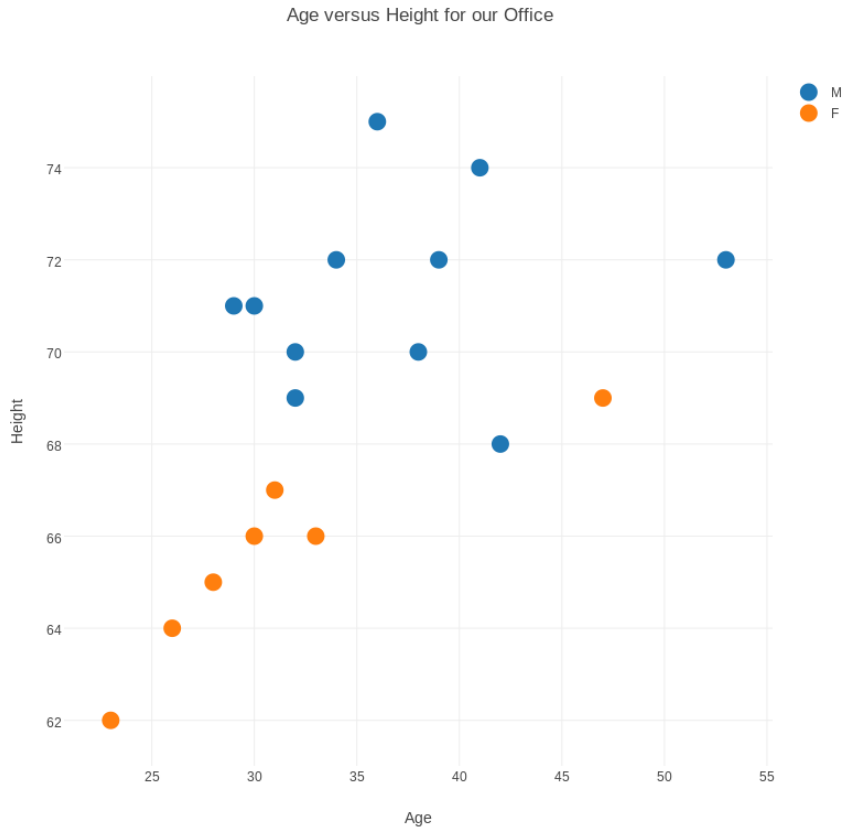
Our scatterplot will again have small dots which we want to increase. And we do indeed see an interesting pattern.



We can again choose the option **Fit data**, but this time, we will have to do it twice, for the female and male data separately. Surprisingly, the two linear relationships have very different slopes now.



Here is another scatter plot, in which we look at age versus height, and again group the data by sex. We can draw several conclusions from this plot.



Example #3: Duration and Pause for Old Faithful

Old Faithful is the nickname for a geyser in Yellowstone National Park.

It got its name because of the regularity of its eruptions. At one time, it was commonly said that it erupted every hour, like a clock. In fact, records show that the next eruption may occur in as little as 35 minutes, or as much as two hours after the previous one.

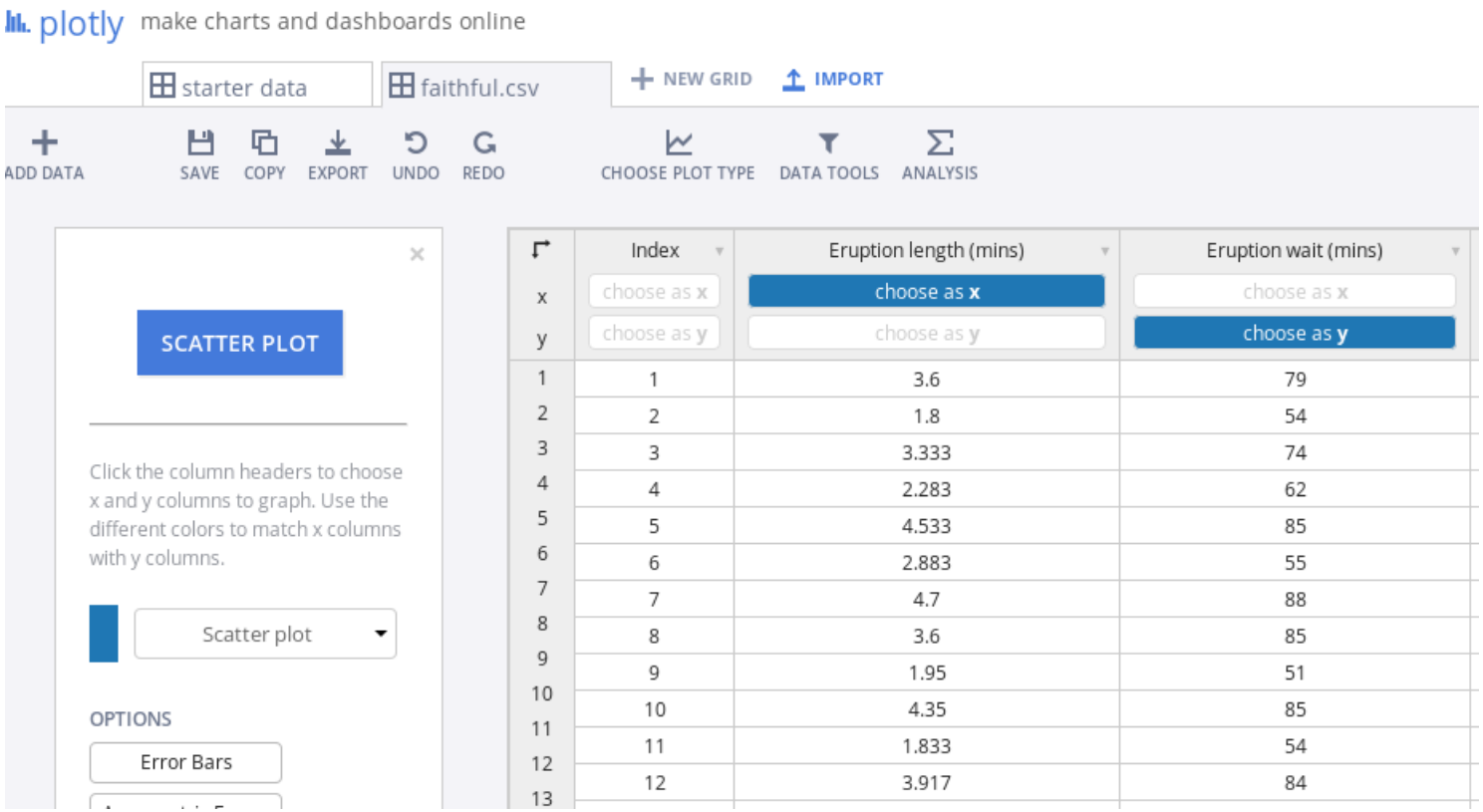
Moreover, the duration of the eruption varies, that is, the length of time during which the geyser sends water shooting up into the air.

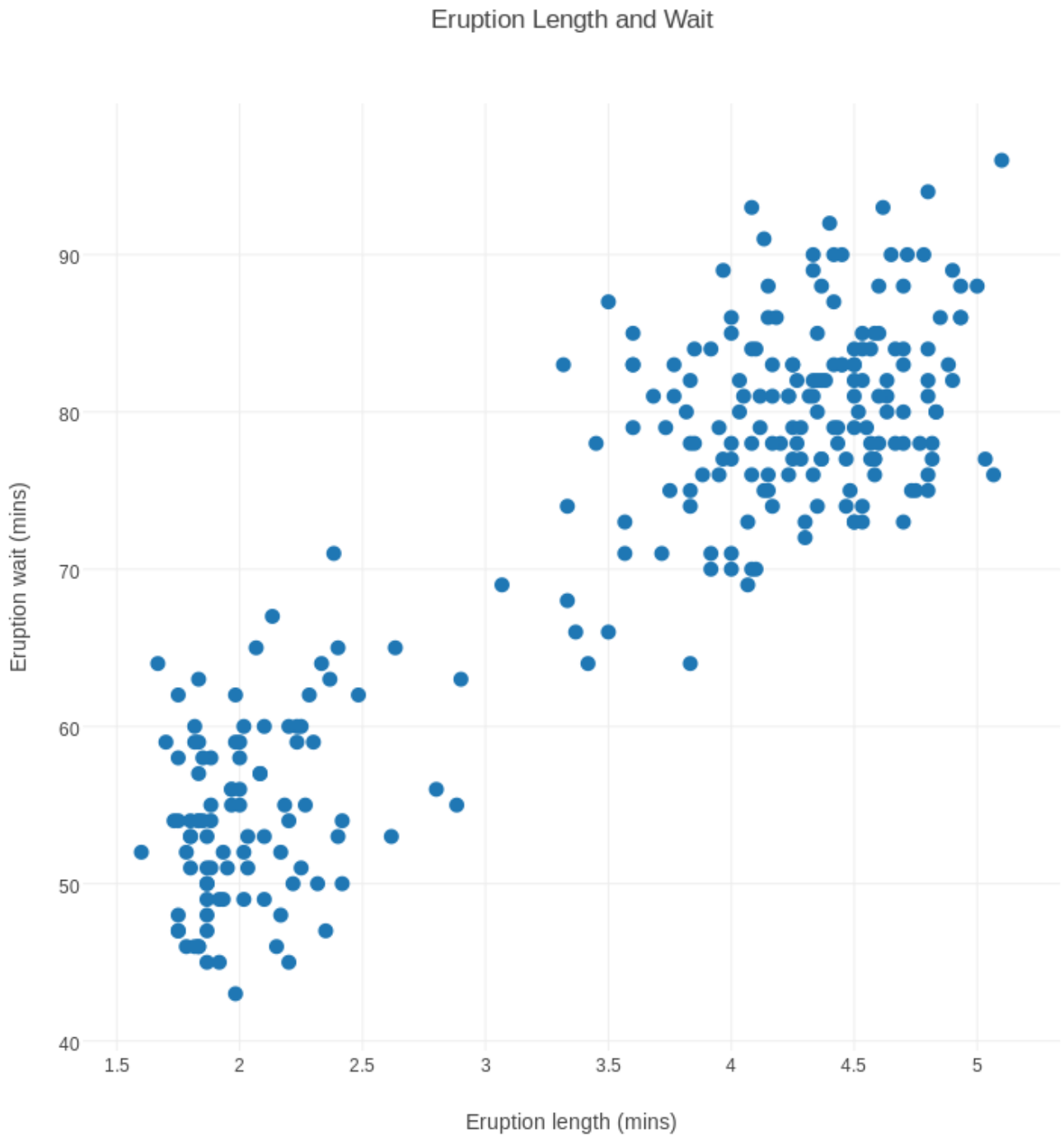
We have 272 records, including the duration of an eruption, and the wait til the next one, in minutes, and we would like to see if there is a pattern between these variables.

Our data is stored in **faithful.csv**. The first line of the file contains labels, followed by 272 lines of data.

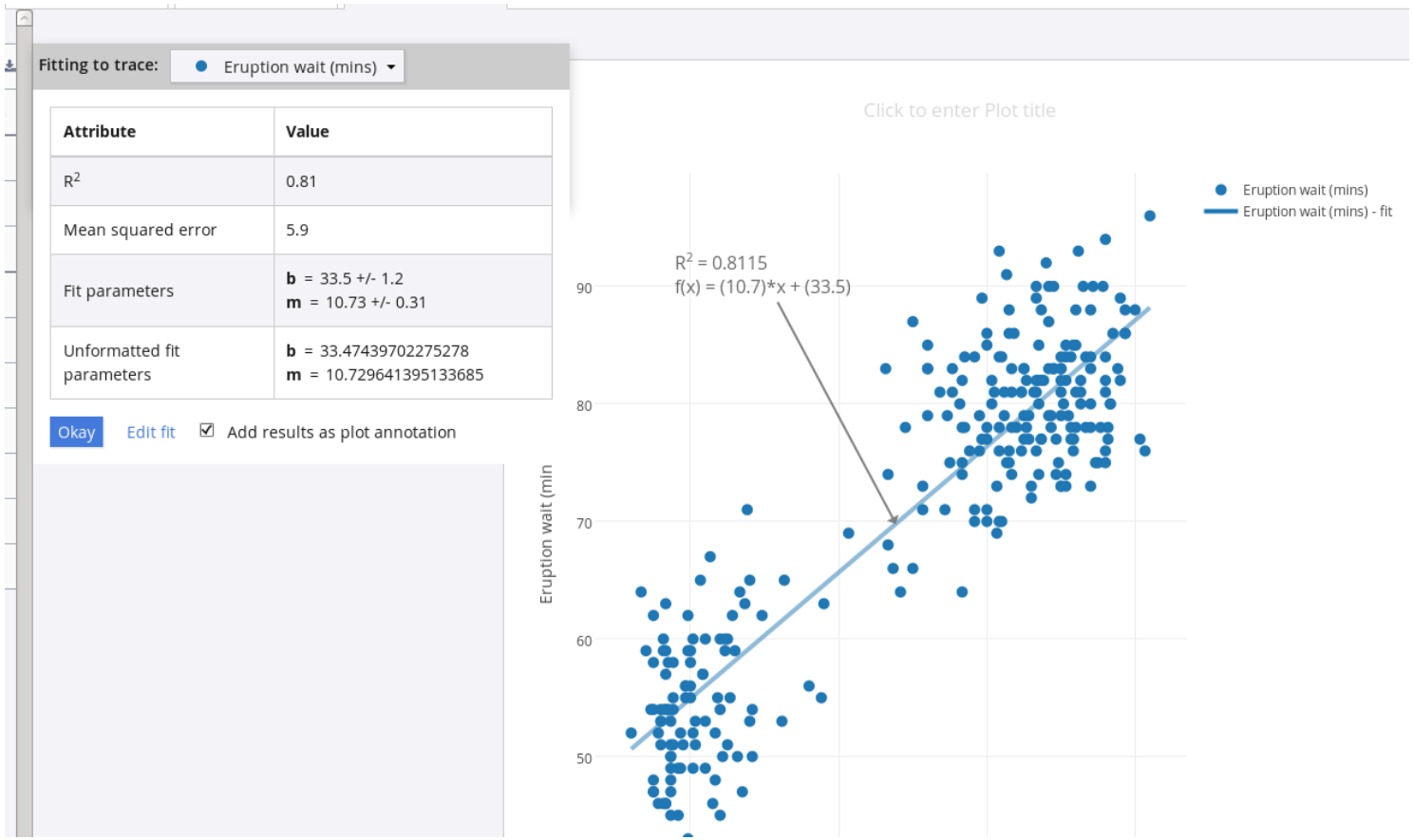
"Index",	"Duration",	"Wait"
1,	3.600,	79
2,	1.800,	54
3,	3.333,	74
—	*more data*	—
270,	4.417,	90
271,	1.817,	46
272,	4.467,	74

We start up Plotlly, import our data file, request that column 2 be used as x, column 3 for y, and specify a scatterplot:




Eruption Length and Wait
Eruption wait (mins)
90
80
70
60
50
40
1.5
2
2.5
3
3.5
4
4.5
5
Eruption length (mins)

We can ask for a linear fit to this data:



The linear fit produces two numbers $\mathbf{b}$ and $\mathbf{m}$ that set up a prediction formula $y = m \times x + b$ from the data.

Based on our data, the linear fit says that:

$$\text{estimated wait} \approx 10.73 \times \text{duration} + 33.5$$

So the formula suggests a 2 minute duration is followed by a 55 minute wait, while a 4 minute duration suggests about a 76 minute wait, which matches the behavior of the line in our plot.

Example #4: Height, Weight, Age of MLB Players

We have seen, when we use the "Fit data" option in Plotly, that along with the formula for the fitting line, there is also a number reported as R^2.

For example, for the first fitting line to our office biostat data, R^2 was about 0.76, and for our Old Faithful fitting line, R^2 was about 0.81.

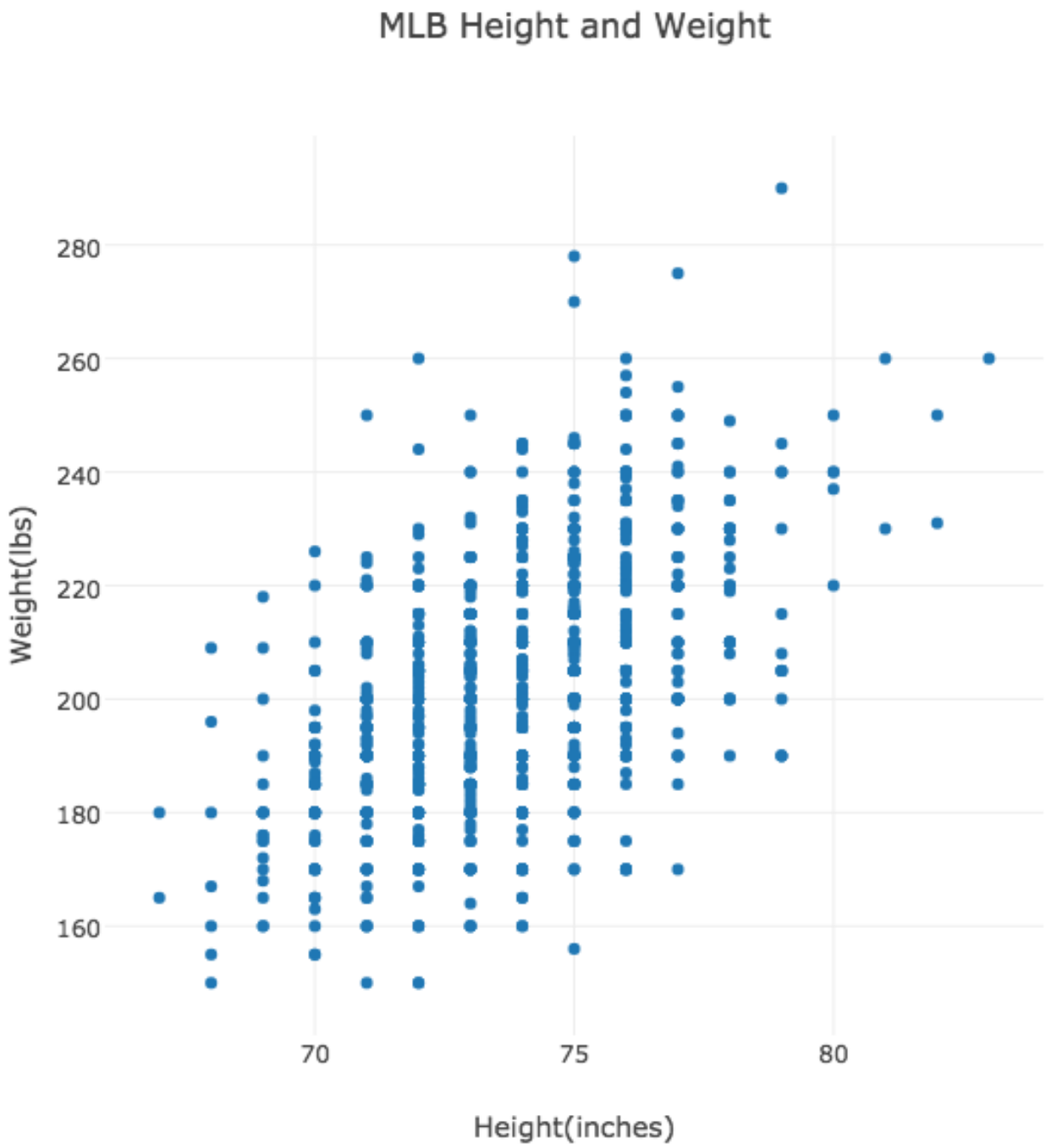
When we make a scatterplot, we expect the data to be scattered, that is, not to lie neatly on a line. When we ask for a linear fit to data, we are asking for the line that comes closest to matching our messy data. The value of R^2 tells us how well we did...in other words, it tells us whether the scattered data is actually close to a linear shape, or is a terrible chaotic mess.

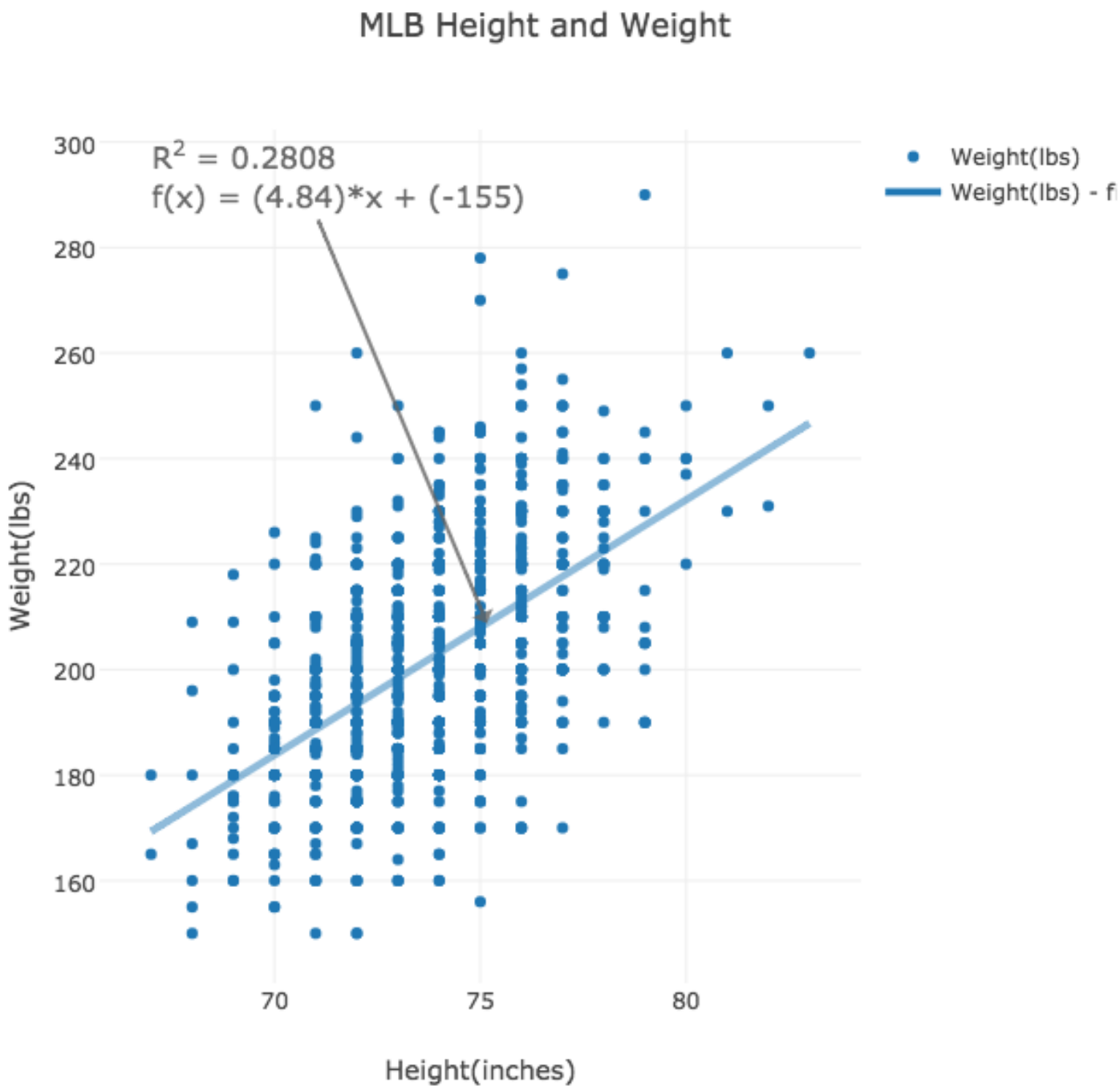
Whenever we ask to fit the data, Plotly will always draw us a straight line. But the value of R^2 is a warning that indicates whether that line is a good explanation of the data, or does not explain much at all.

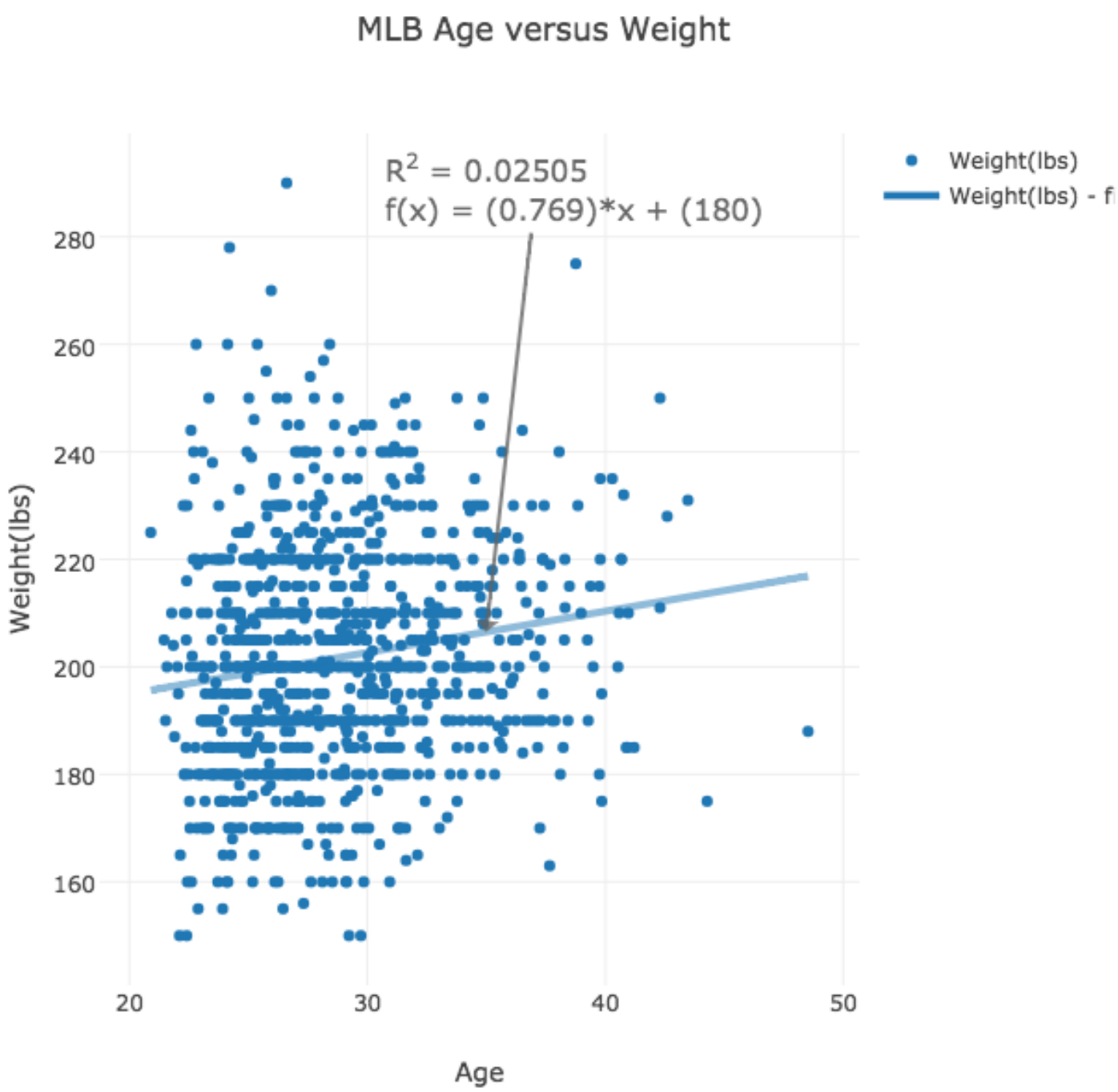
In this example, we will study a set of data about Major League Baseball (MLB) players, including height, weight, and age. We will make some scatter plots, and ask Plotly to show us the fitting lines. With our eyes, we can make some judgment about whether the data has a simple form, but as we examine the R^2 values, we will see that it gives us a clear indication of how orderly or disorderly the data is.

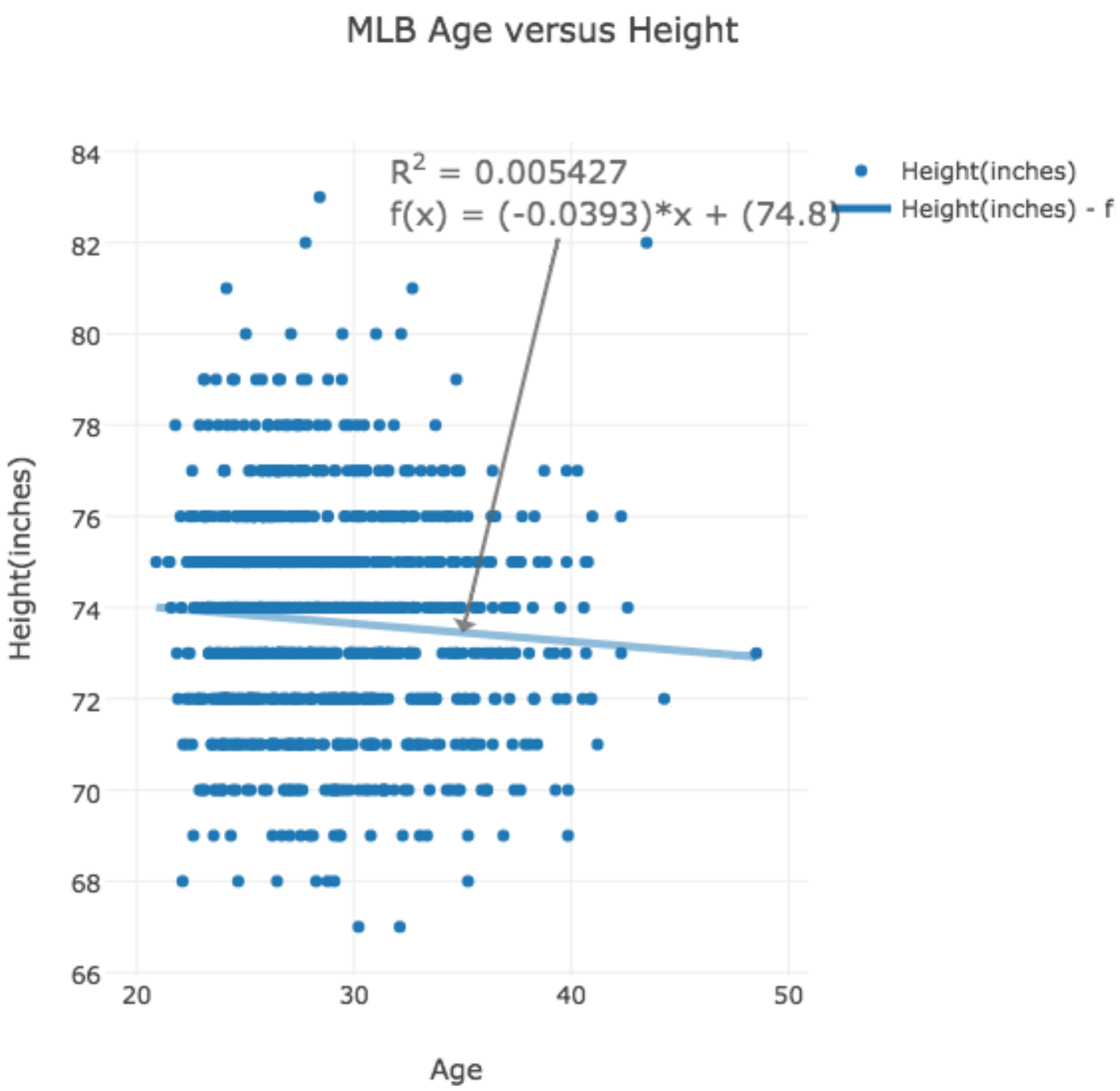
The data file **mlb.csv** describes 1034 baseball players:

"Name"	"Team"	"Position"	"Height(in)"	"Weight(lbs)"	"Age"
"Adam Donachie"	"BAL"	"Catcher"	74	180	22.99
"Paul Bako"	"BAL"	"Catcher"	74	215	34.69
"Ramon Hernandez"	"BAL"	"Catcher"	72	210	30.78
—	—	*more data*	—	—	—
"Chris Narveson"	"STL"	"Relief Pitcher"	75	205	25.19
"Randy Keisler"	"STL"	"Relief Pitcher"	75	190	31.01
"Josh Kinney"	"STL"	"Relief Pitcher"	73	195	27.92


MLB Height and Weight
Weight(lbs)
280
260
240
220
200
180
160
70
75
80
Height(inches)


MLB Height and Weight
R^2 = 0.2808
f(x) = (4.84)*x + (-155)
Weight(lbs)
Weight(lbs) - f
300
280
260
240
220
200
180
160
Weight(lbs)
70
75
80
Height(inches)


MLB Age versus Weight
R^2 = 0.02505
f(x) = (0.769)*x + (180)
Weight(lbs)
Weight(lbs) - f
280
260
240
220
200
180
160
Weight(lbs)
20
30
40
50
Age


MLB Age versus Height
R^2 = 0.005427
f(x) = (-0.0393)*x + (74.8)
Height(inches)
Height(inches) - f
84
82
80
78
76
74
72
70
68
66
Height(inches)
20
30
40
50
Age

The MLB data never really looked as nicely patterned as many of the other datasets have been.

The fact that the data really looks scattered, and that R^2 is so low, suggests that:

- we can only roughly estimate MLB player weight from height;
- we have little accuracy in estimating weight or height from age.

There is a weak general trend from weight to increase and height to decrease with increasing age.

So...just because we can draw a fitting line doesn't mean we've discovered a strong relationship between two items of data!

Example #5: Making Sense of House Prices

One of the biggest tasks for a realtor is deciding how to determine the asking price for a house. The computation is based on the intrinsic value of the house, the asking price for other houses in the neighborhood, the current state of the local housing market, interest rates and so on.

The final judgment comes with the selling price, what someone is actually willing to pay for the house, given all these factors.

We would like to use scatterplots to study how much certain features influence the size of the final selling price for a house.

The data file **homes.csv** describes a list of home sales:

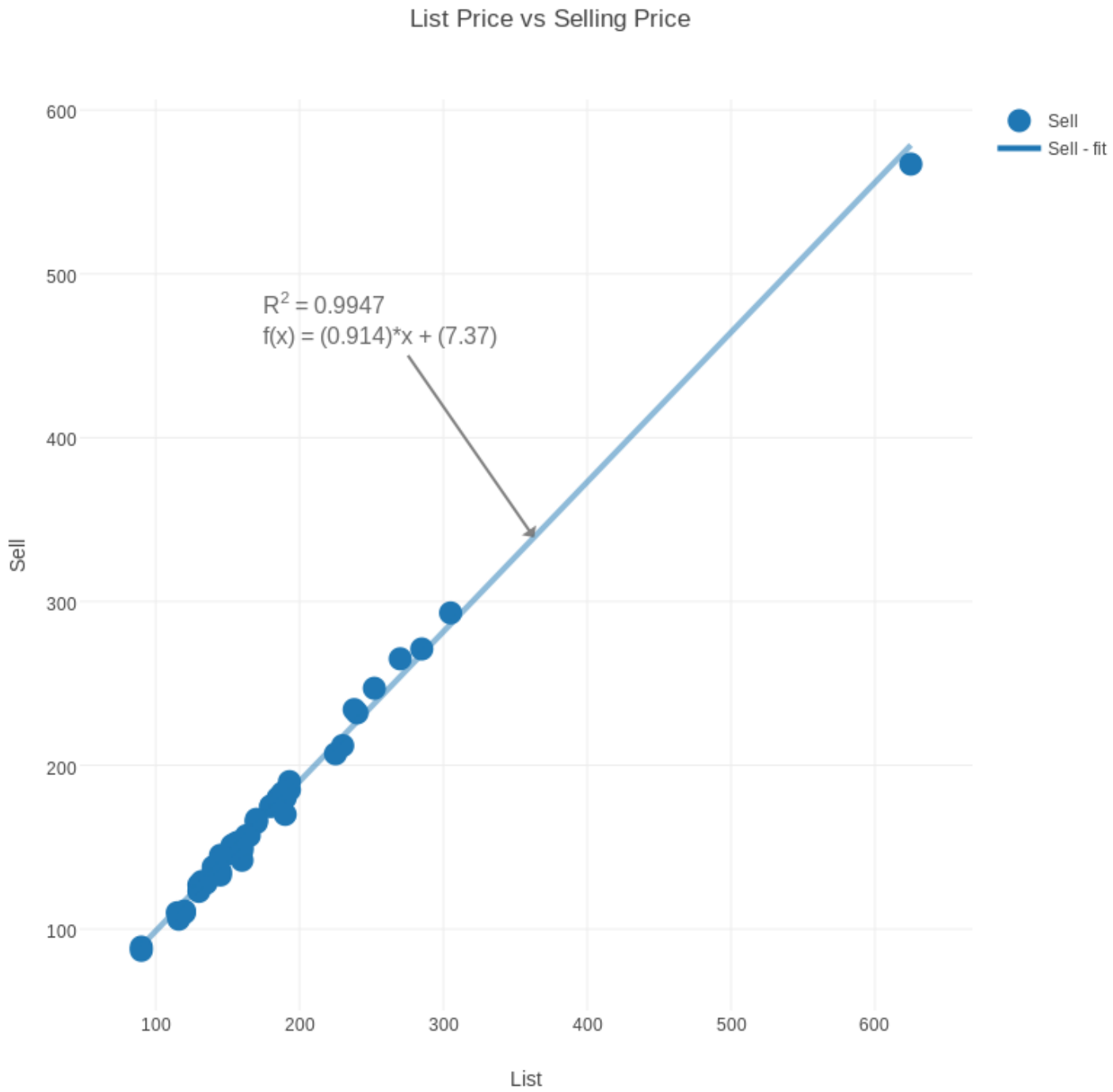
"Sell"	"List"	"Living"	"Rooms"	"Beds"	"Baths"	"Age"	"Acres"	"Taxes"
142	160	28	10	5	3	60	0.28	3167
175	180	18	8	4	1	12	0.43	4033
129	132	13	6	3	1	41	0.33	1471
—	—	—	—	*more*	—	—	—	—
247	252	29	9	4	2	4	1.25	4626
111	120	15	8	3	1	97	1.11	3205
133	145	26	7	3	1	42	0.36	3059

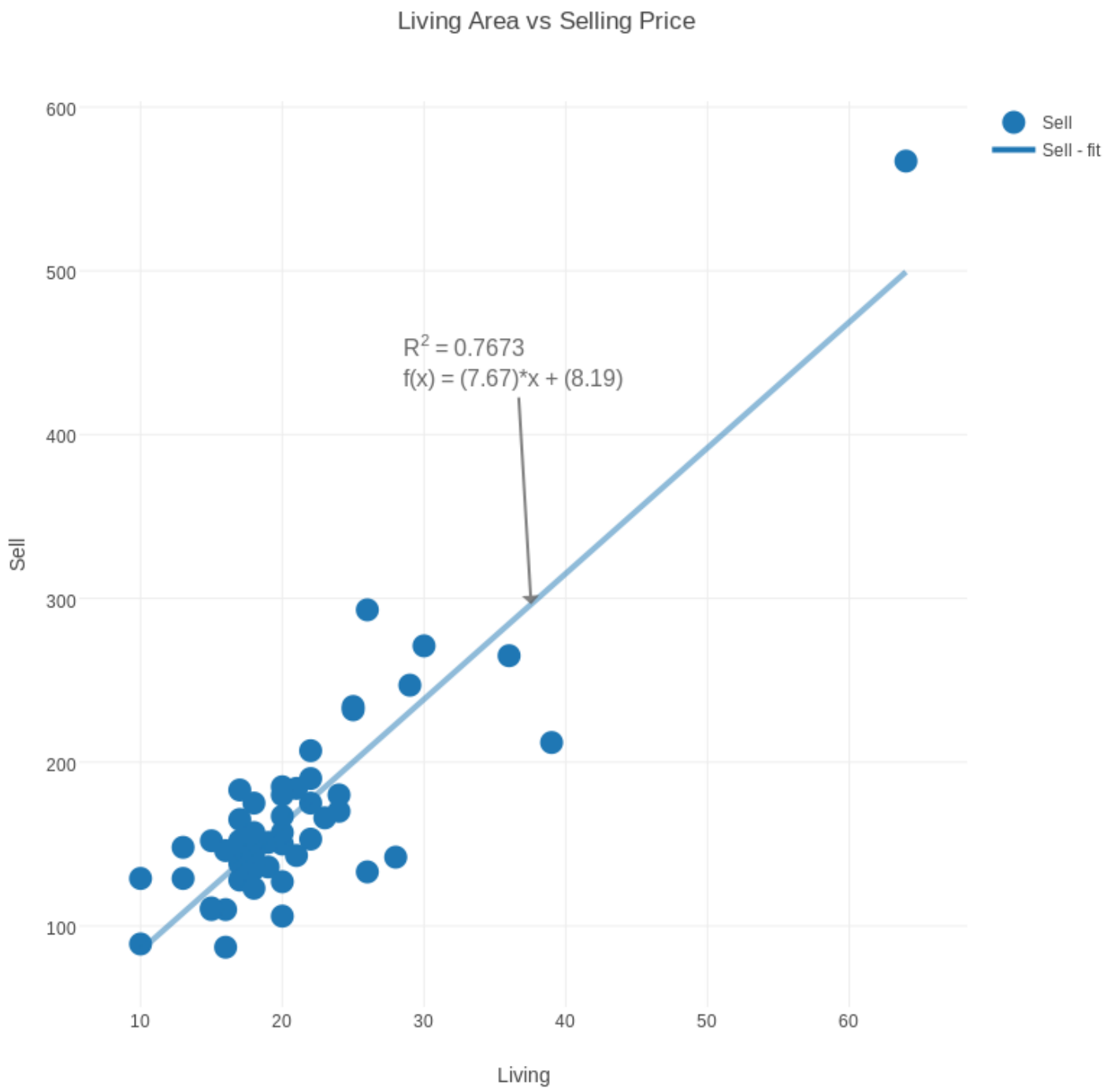
We will investigate by plotting the selling price as the dependent variable (the $\mathbf{y}$ value) and choosing each of the remaining quantities as the $\mathbf{x}$ value.

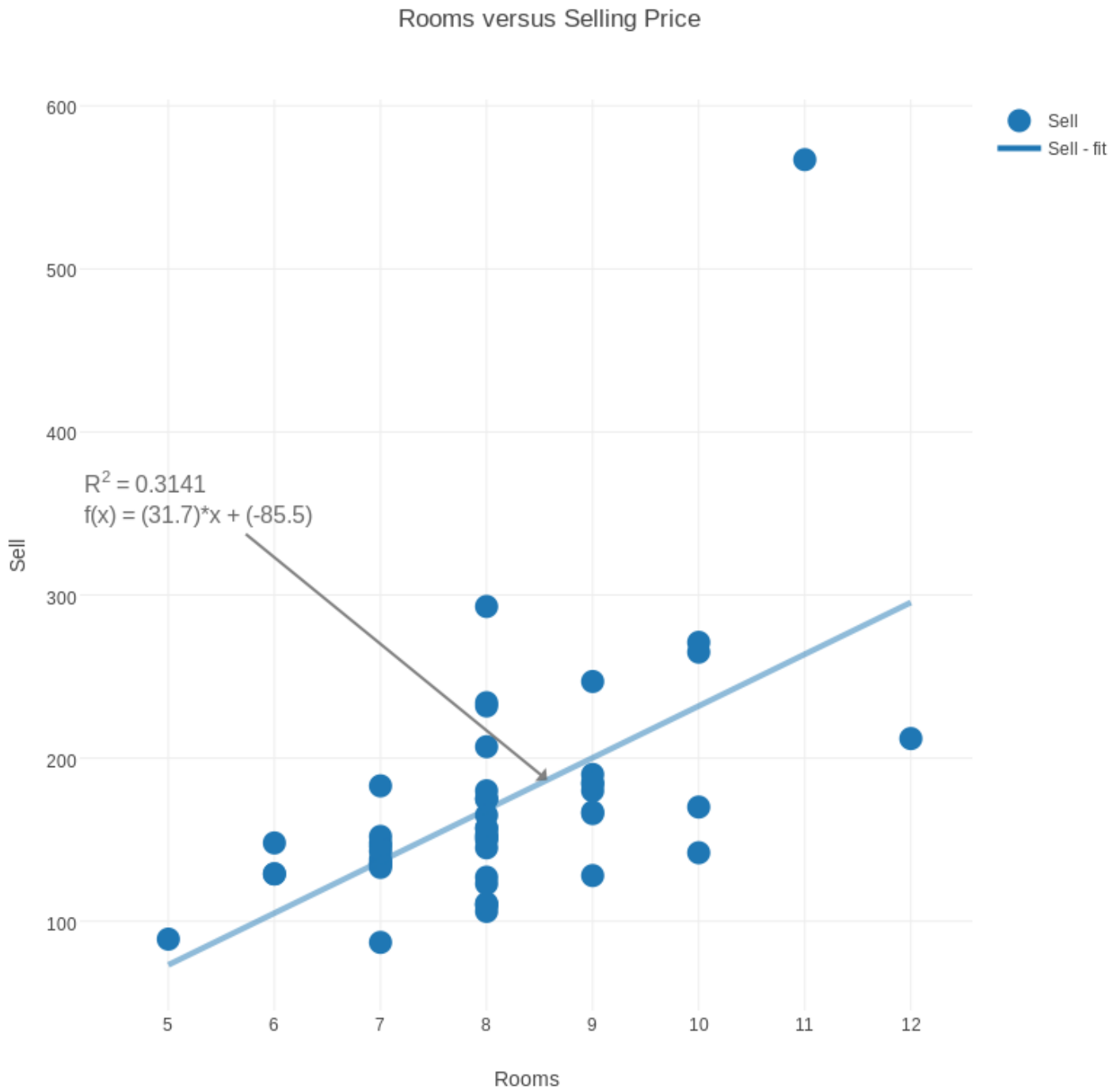
Then we can ask Plotly to fit the data, that is, to come up with the best formula that predicts the value of $\mathbf{y}$ given the value $\mathbf{x}$.

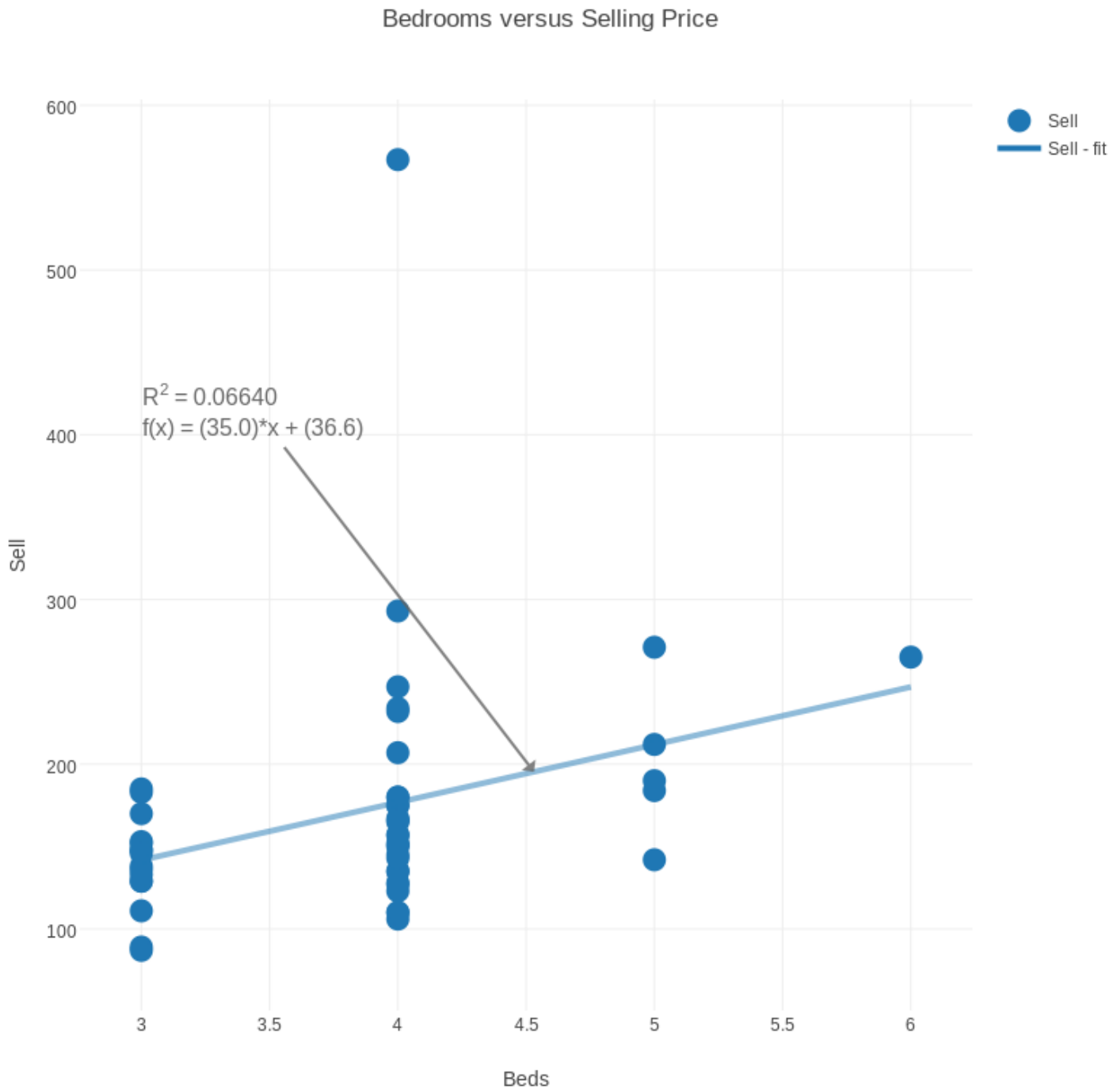
Plotly will also report the value of R^2, which will tell us whether $\mathbf{x}$ is actually a useful clue for the value of $\mathbf{y}$.

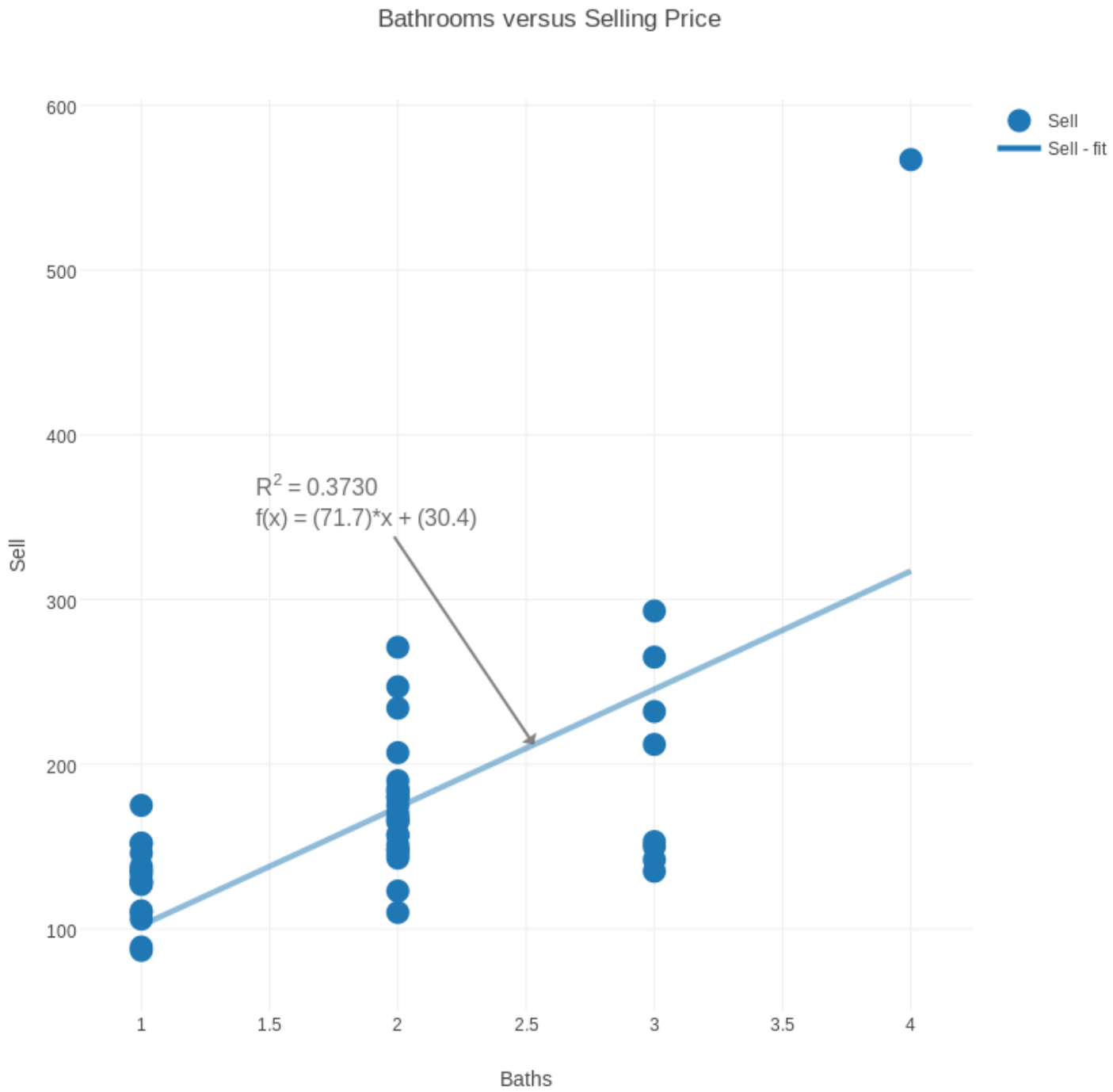
If you were only allowed to know one of the quantities in our table, and had to guess the selling price of the house, what are the best quantities to know? Which ones don't tell you much?

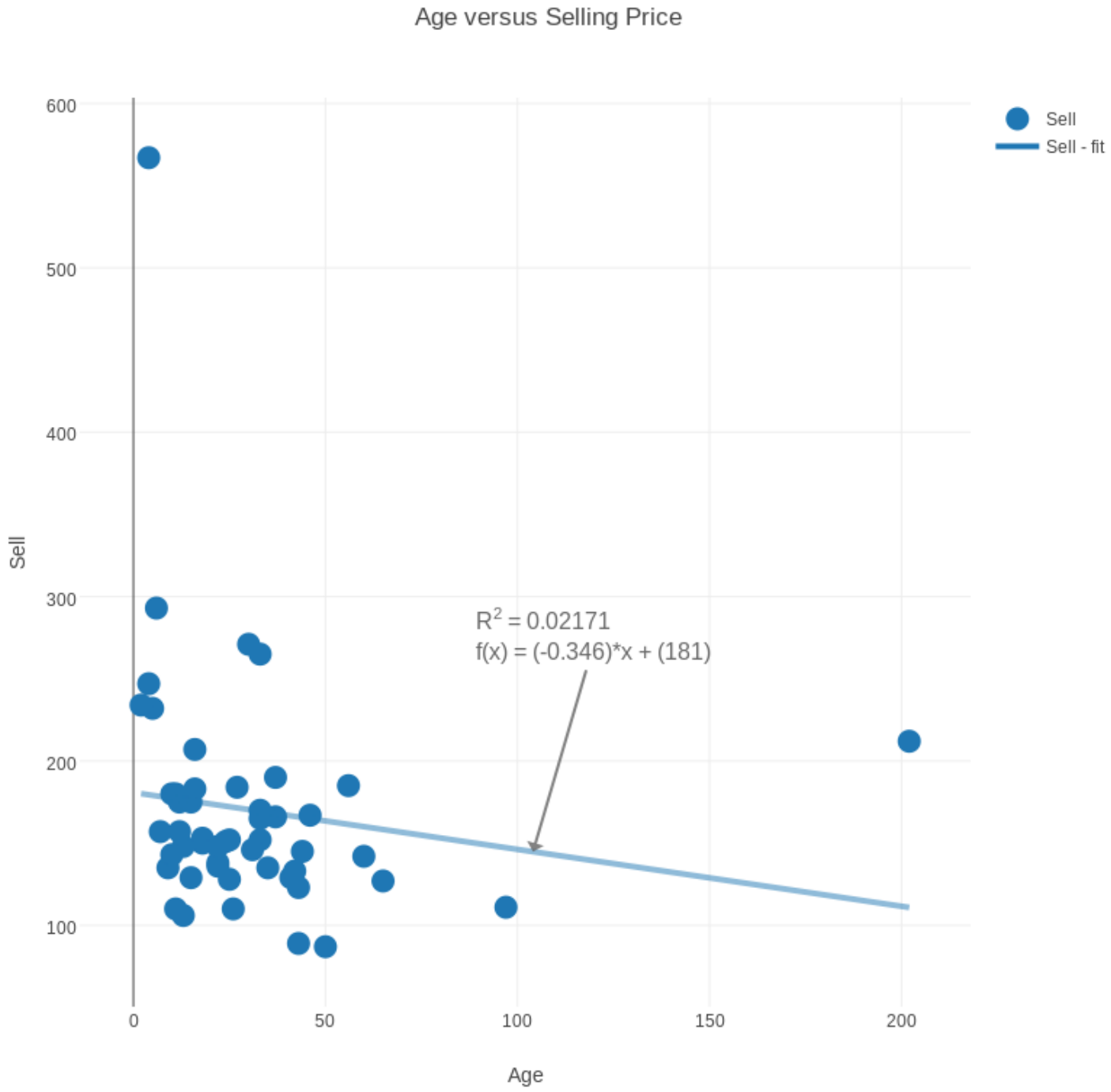

List Price vs Selling Price
Sell
Sell - fit
R^2 = 0.9947
f(x) = (0.914)*x + (7.37)
600
500
400
300
200
100
Sell
100
200
300
400
500
600
List

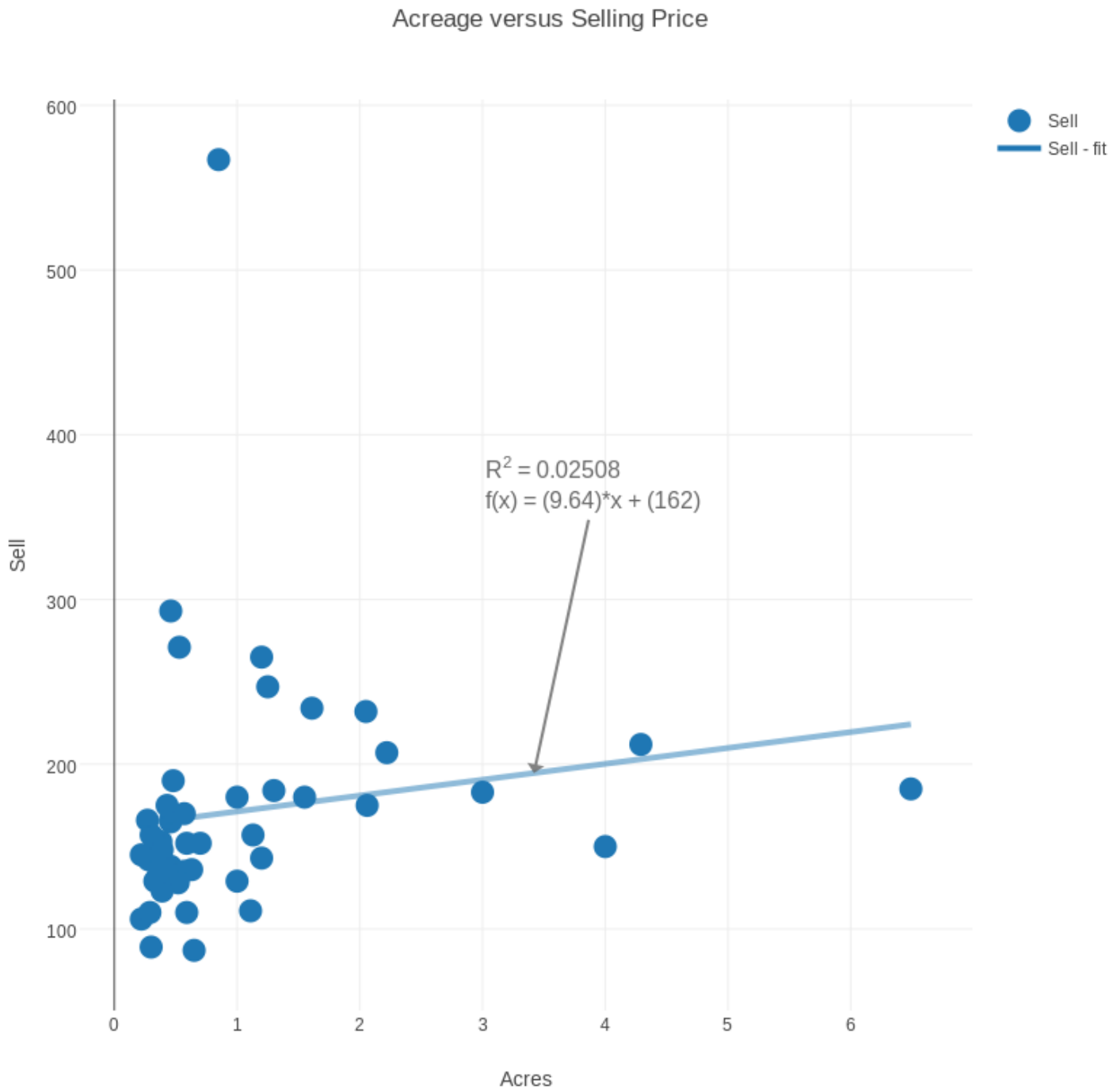

Living Area vs Selling Price
Sell
Sell - fit
$R^2 = 0.7673$
f(x) = (7.67)*x + (8.19)
600
500
400
300
200
100
Sell
10
20
30
40
50
60
Living

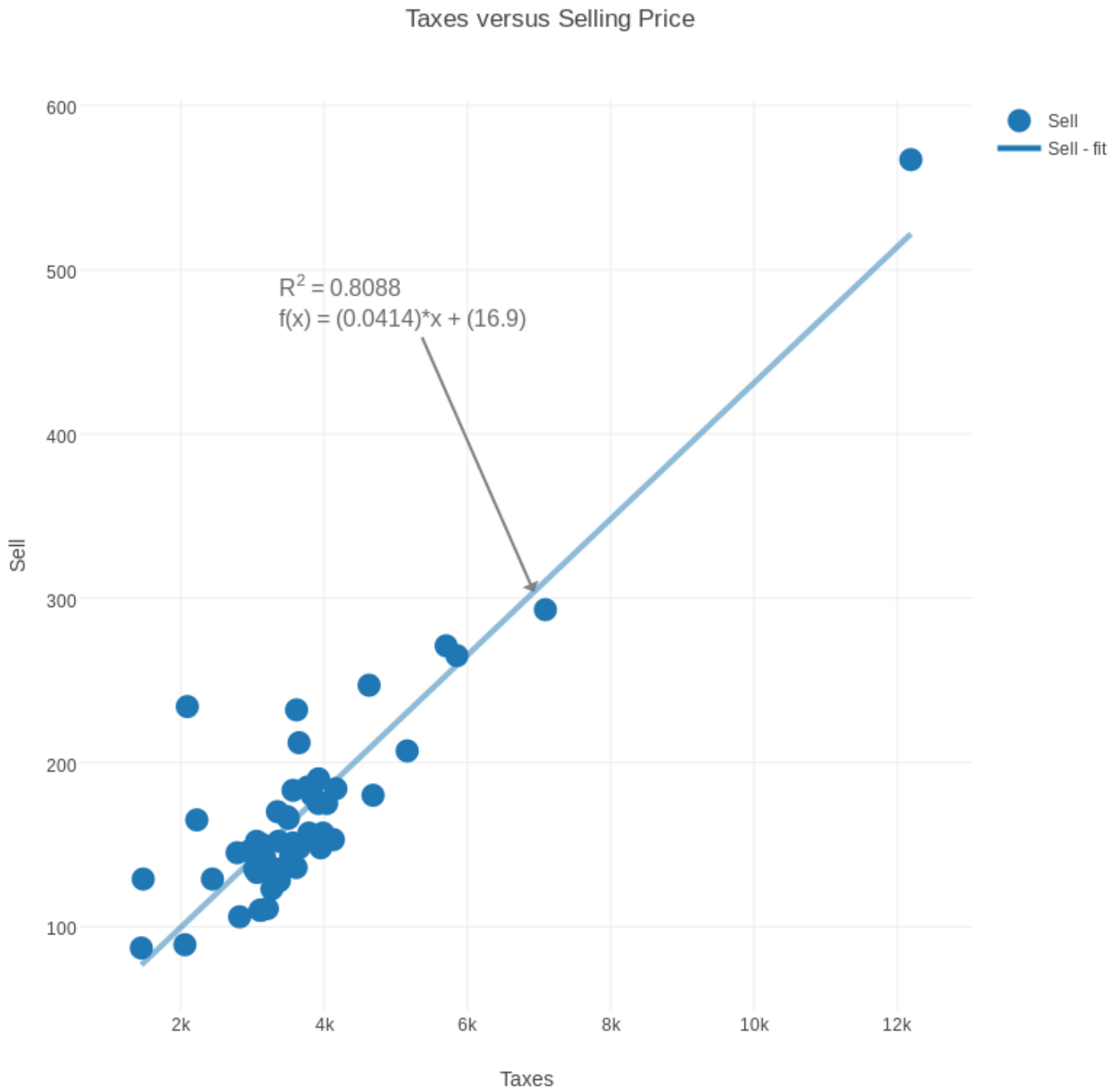

Rooms versus Selling Price
Sell
Sell - fit
600
500
400
300
200
100
Sell
R^2 = 0.3141
f(x) = (31.7)*x + (-85.5)
5
6
7
8
9
10
11
12
Rooms


Bedrooms versus Selling Price
Sell
Sell - fit
R^2 = 0.06640
f(x) = (35.0)*x + (36.6)
600
500
400
300
200
100
Sell
3
3.5
4
4.5
5
5.5
6
Beds


Bathrooms versus Selling Price
Sell
Sell - fit
600
500
400
300
200
100
Sell
$R^2 = 0.3730$
f(x) = (71.7)*x + (30.4)
1
1.5
2
2.5
3
3.5
4
Baths


Age versus Selling Price
Sell
Sell - fit
600
500
400
300
200
100
Sell
R^2 = 0.02171
f(x) = (-0.346)*x + (181)
0
50
100
150
200
Age


Acreage versus Selling Price
Sell
Sell - fit
R^2 = 0.02508
f(x) = (9.64)*x + (162)
Sell
Acres
600
500
400
300
200
100
0
1
2
3
4
5
6


Taxes versus Selling Price
Sell
Sell - fit
$R^2 = 0.8088$
f(x) = (0.0414)*x + (16.9)
600
500
400
300
200
100
Sell
2k
4k
6k
8k
10k
12k
Taxes

Obviously, all the factors we've looked at can affect the selling price.

It's reasonable that the selling price is strongly related to the asking price, and close to the taxes (shouldn't it be even closer?).

After that, if we only had one item to work from, it seems like living area, number of bathrooms, and number of rooms are strong predictors.

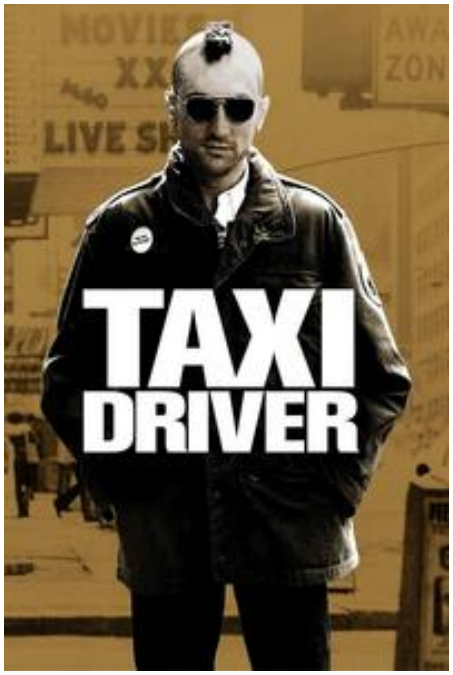
Notice that our data included an outlier, that is, one data item that was very different from the others, a house which sold for half a million dollars, double the typical price. Sometimes, a single outlier can distort the results of an analysis. If that is suspected, the analysis can be redone by removing that outlier and seeing if the results change significantly.

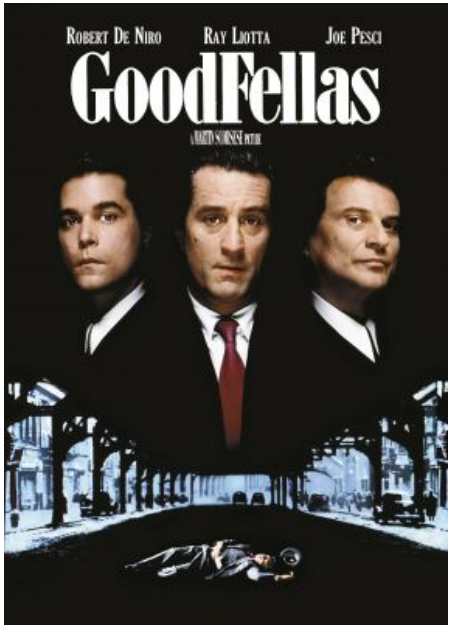
Example #6: Robert DeNiro's Rotten Tomato Scores

Robert DeNiro has a long career in movies. Some of his early movies are critically acclaimed, while many of his recent movies have done poorly at the box office, or were criticized because it felt that DeNiro was just "phoning in" his role, having lost interest in acting.

It's not possible to be scientific about such statements, but, on the other hand, there are a number of movie databases that collect ratings from viewers about movies, and as long as we're dealing with numbers, we have some ideas about how to investigate them.

James Chapman had the idea of looking at the Rotten Tomato ratings for movies starring DeNiro, and making a scatter plot from them.

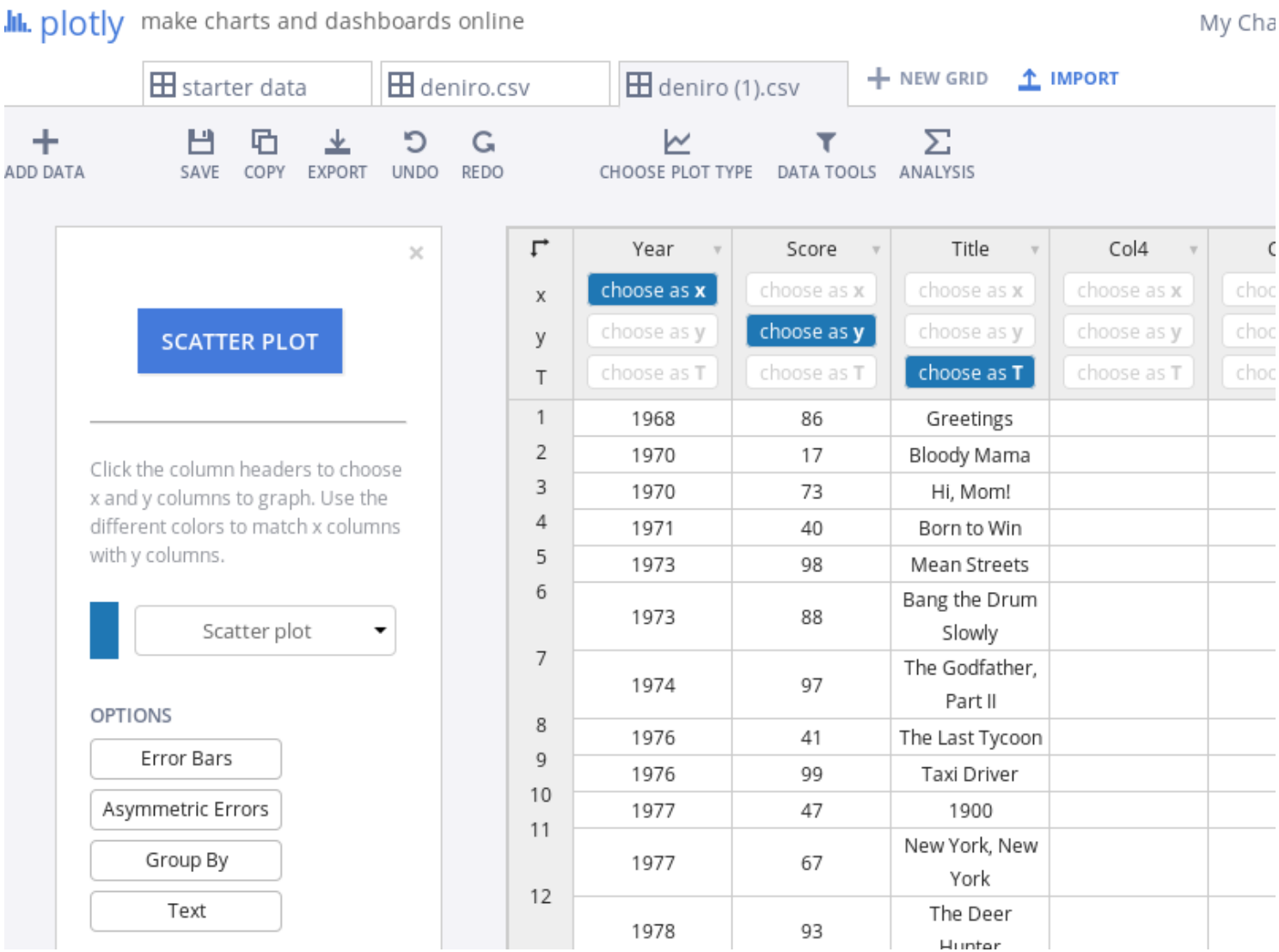

TAXI
DRIVER

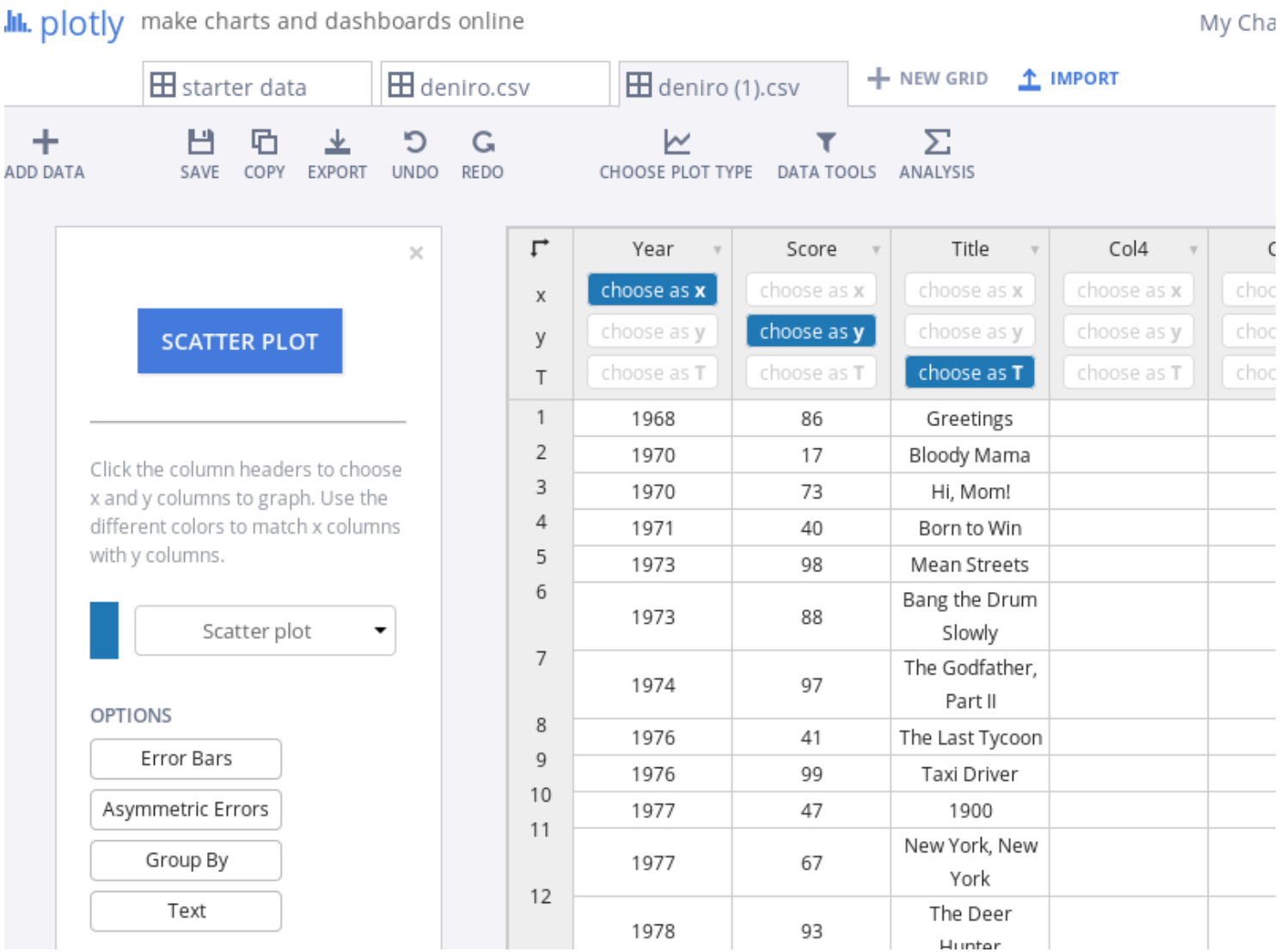

Robert De Niro
Ray Liotta
Joe Pesci
GoodFellas

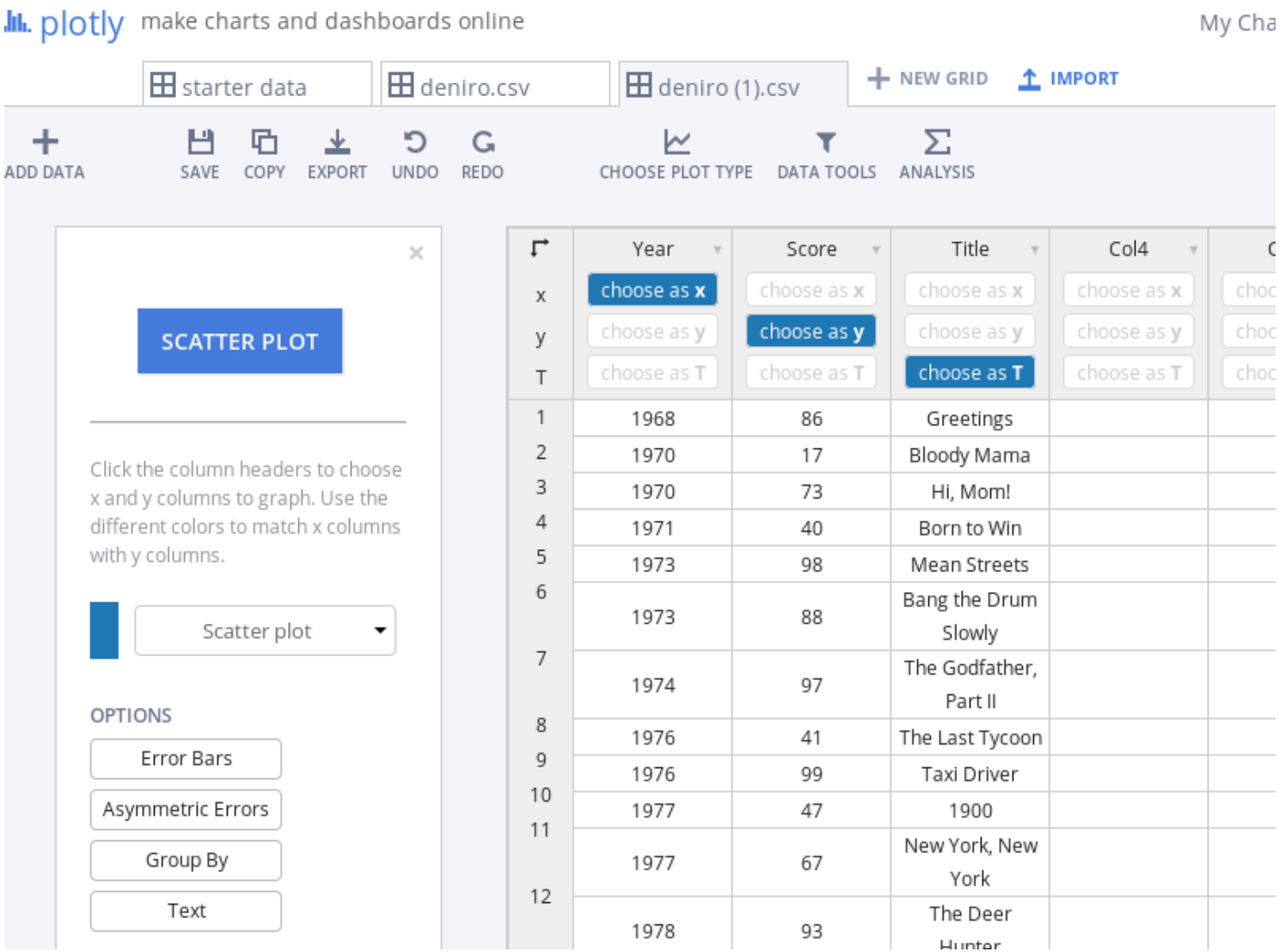
Meet the
Parents

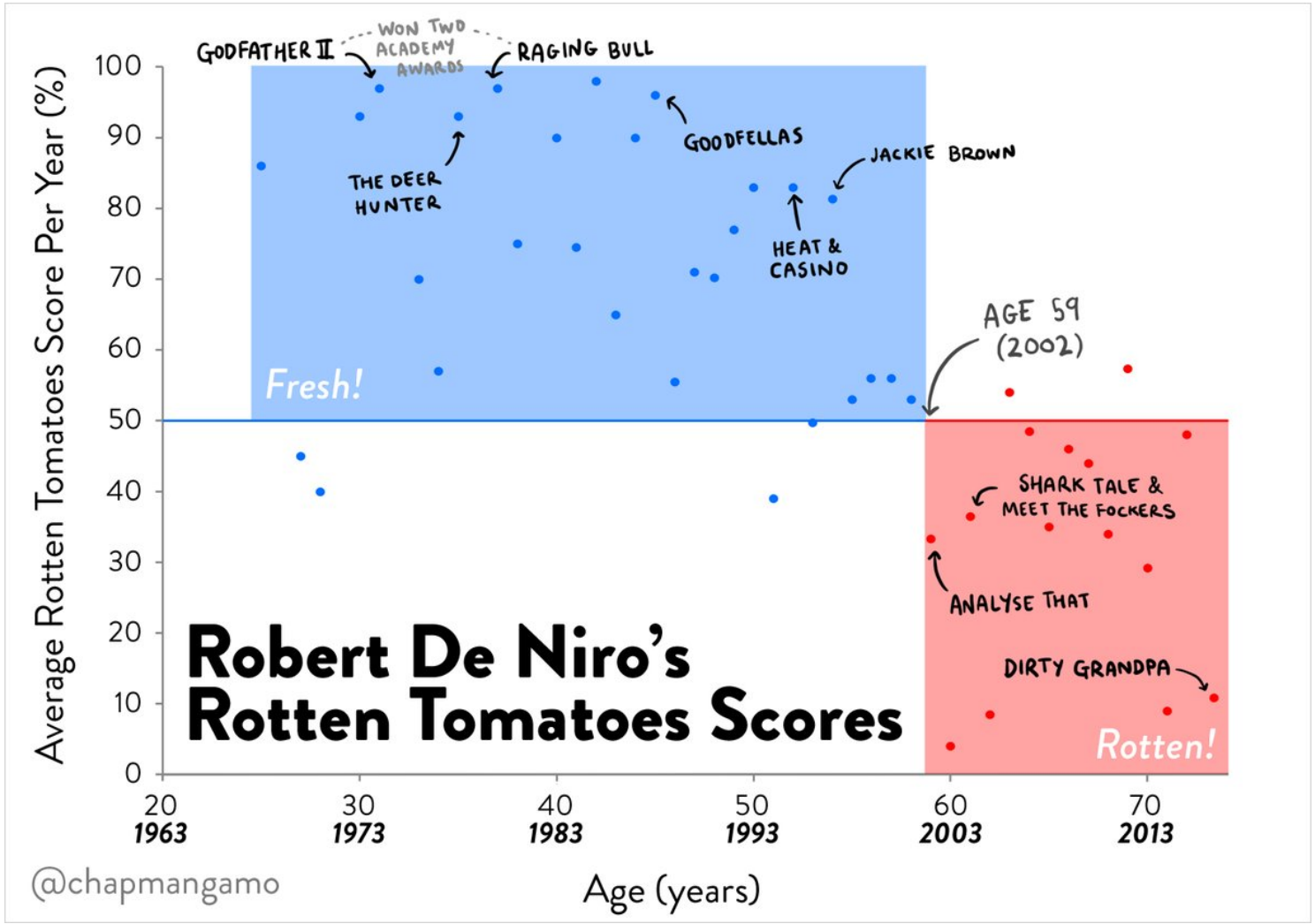
The data file **deniro.csv** describes 88 movies of Robert Deniro:

"Year"	"Score"	"Title"
1968	86	"Greetings"
1970	17	"Bloody Mama"
1970	73	"Hi, Mom!"
—	*more*	—
2015	26	"Heist"
2015	61	"The Intern"
2016	11	"Dirty Grandpa"


plotly
make charts and dashboards online
My Cha
starter data
deniro.csv
deniro (1).csv
NEW GRID
IMPORT
ADD DATA
SAVE
COPY
EXPORT
UNDO
REDO
CHOOSE PLOT TYPE
DATA TOOLS
ANALYSIS
SCATTER PLOT
Click the column headers to choose x and y columns to graph. Use the different colors to match x columns with y columns.
Scatter plot
OPTIONS
Error Bars
Asymmetric Errors
Group By
Text
Year
Score
Title
Col4
x
y
T
choose as x
choose as y
choose as T
1 1968 86 Greetings
2 1970 17 Bloody Mama
3 1970 73 Hi, Mom!
4 1971 40 Born to Win
5 1973 98 Mean Streets
6 1973 88 Bang the Drum Slowly
7 1974 97 The Godfather, Part II
8 1976 41 The Last Tycoon
9 1976 99 Taxi Driver
10 1977 47 1900
11 1977 67 New York, New York
12 1978 93 The Deer Hunter


plotly
make charts and dashboards online
My Cha
starter data
deniro.csv
deniro (1).csv
NEW GRID
IMPORT
ADD DATA
SAVE
COPY
EXPORT
UNDO
REDO
CHOOSE PLOT TYPE
DATA TOOLS
ANALYSIS
SCATTER PLOT
Click the column headers to choose x and y columns to graph. Use the different colors to match x columns with y columns.
Scatter plot
OPTIONS
Error Bars
Asymmetric Errors
Group By
Text
Year
Score
Title
Col4
x
y
T
choose as x
choose as y
choose as T
1 1968 86 Greetings
2 1970 17 Bloody Mama
3 1970 73 Hi, Mom!
4 1971 40 Born to Win
5 1973 98 Mean Streets
6 1973 88 Bang the Drum Slowly
7 1974 97 The Godfather, Part II
8 1976 41 The Last Tycoon
9 1976 99 Taxi Driver
10 1977 47 1900
11 1977 67 New York, New York
12 1978 93 The Deer Hunter


plotly
make charts and dashboards online
My Cha
starter data
deniro.csv
deniro (1).csv
NEW GRID
IMPORT
ADD DATA
SAVE
COPY
EXPORT
UNDO
REDO
CHOOSE PLOT TYPE
DATA TOOLS
ANALYSIS
SCATTER PLOT
Click the column headers to choose x and y columns to graph. Use the different colors to match x columns with y columns.
Scatter plot
OPTIONS
Error Bars
Asymmetric Errors
Group By
Text
Year
Score
Title
Col4
x
y
T
choose as x
choose as y
choose as T
1 1968 86 Greetings
2 1970 17 Bloody Mama
3 1970 73 Hi, Mom!
4 1971 40 Born to Win
5 1973 98 Mean Streets
6 1973 88 Bang the Drum Slowly
7 1974 97 The Godfather, Part II
8 1976 41 The Last Tycoon
9 1976 99 Taxi Driver
10 1977 47 1900
11 1977 67 New York, New York
12 1978 93 The Deer Hunter


GODFATHER II
WON TWO ACADEMY AWARDS
RAGING BULL
GOODFELLAS
JACKIE BROWN
THE DEER HUNTER
HEAT & CASINO
AGE 59 (2002)
Fresh!
SHARK TALE & MEET THE FOCKERS
ANALYSE THAT
DIRTY GRANDPA
Rotten!
Robert De Niro's Rotten Tomatoes Scores
Average Rotten Tomatoes Score Per Year (%)
100
90
80
70
60
50
40
30
20
10
0
20 1963
30 1973
40 1983
50 1993
60 2003
70 2013
Age (years)
@chapmangamo