Analyzing one variable

This data for this exercise can be found at:

http://bit.ly/1P8ti6T

How do I know I am using one variable?

- Summarizing one variable
  - Average
  - Median
  - Sum
  - Maximum
  - Minimum
- Looking for an outlier in a distribution

How do I know which summary statistic(s) to use?

Well, it depends on the shape of the distribution.

But first some vocab:

Mean/Average - This is the sum of all values divided by the number of observations.

Median - If we rank everyone in the data by value, this is the value associated with the person (or people) in the middle.

Mode - This is the value that occurs most frequently in the data.

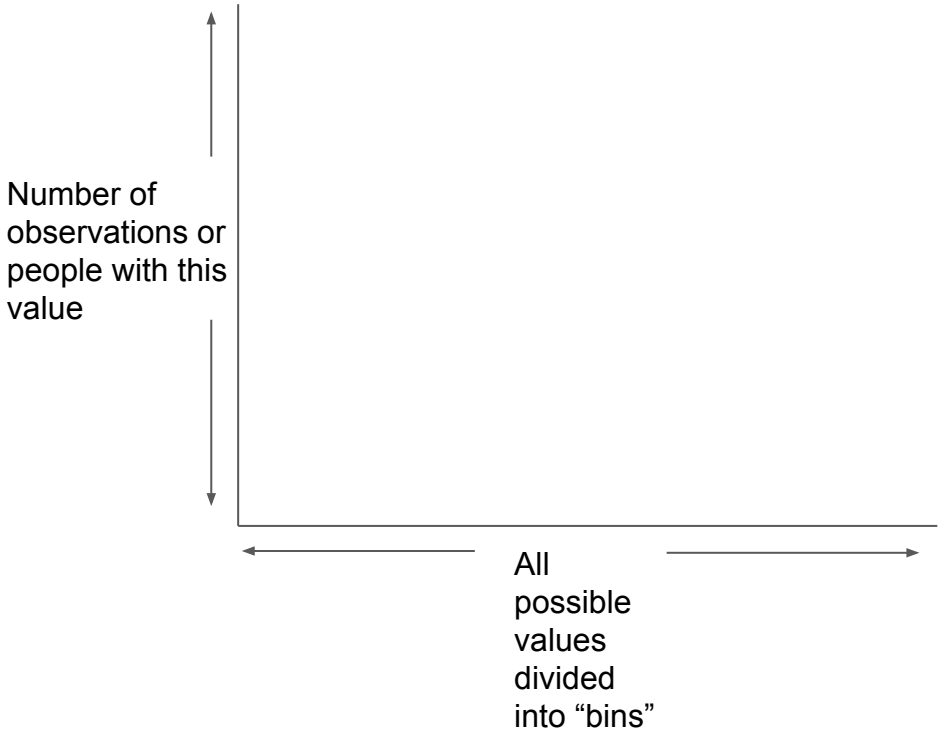
What is a distribution?

In general terms, it reflects how observations are spread out across the range of our data.

The range is all the values between the minimum and maximum values in our column.

A good way to see a distribution is to make a histogram.

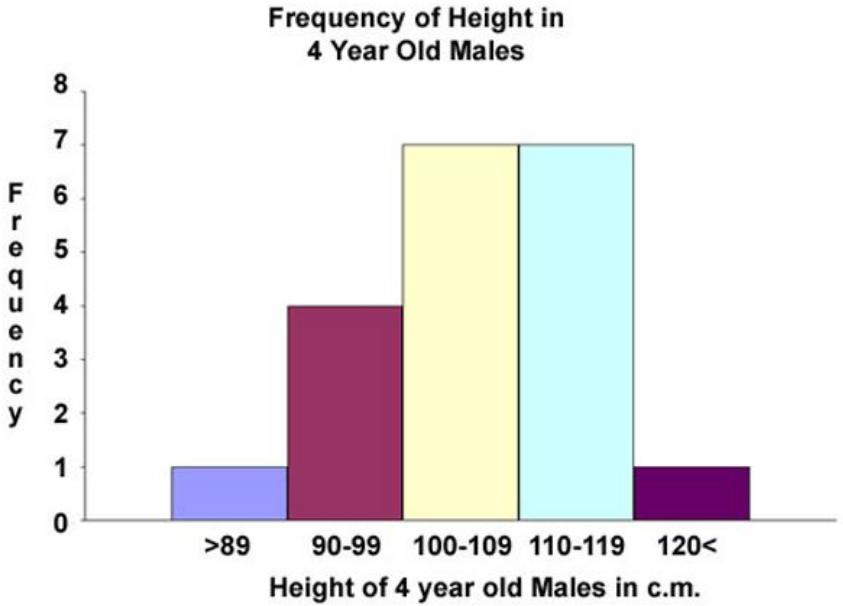
What is a histogram?



A histogram tells us how common each range in the data is. This is called the ‘distribution’ of the data.

What is a histogram?

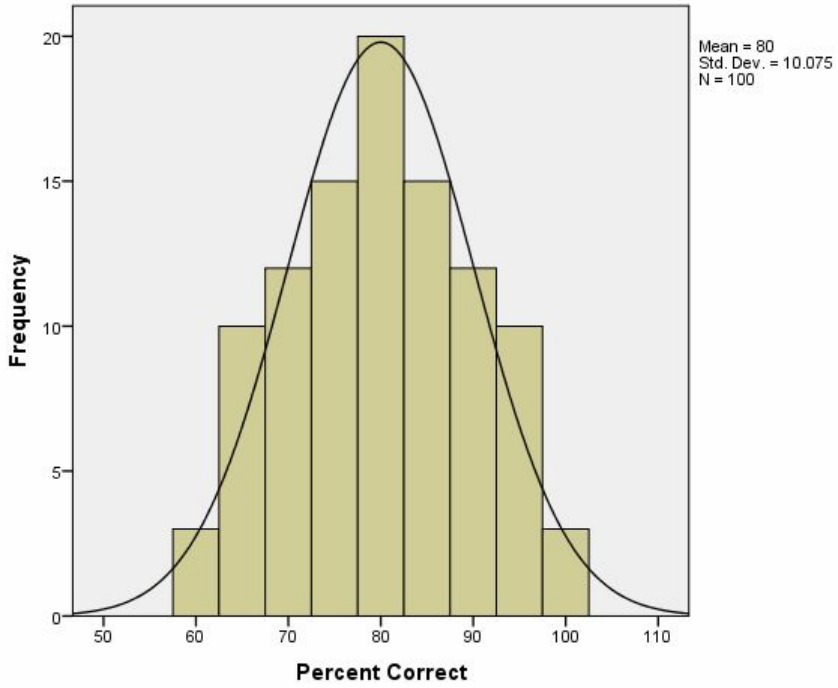
Here’s an example histogram.



Here are some common distribution shapes

Normal

Average is an appropriate summary for this column.



Here are some common distribution shapes

Normal with skew

The long right or left tail can move the mean to a value that isn't typical.

You should consider median or mode here.

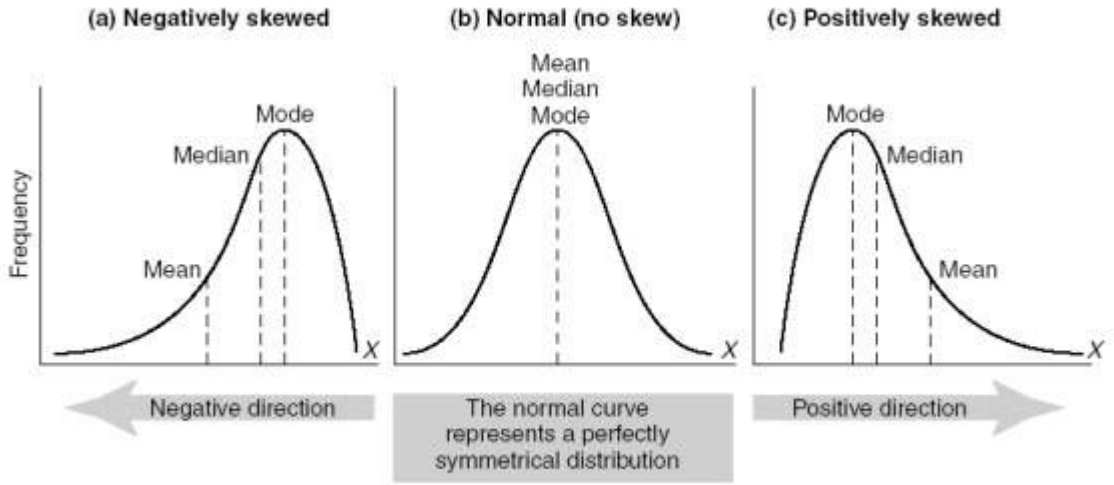
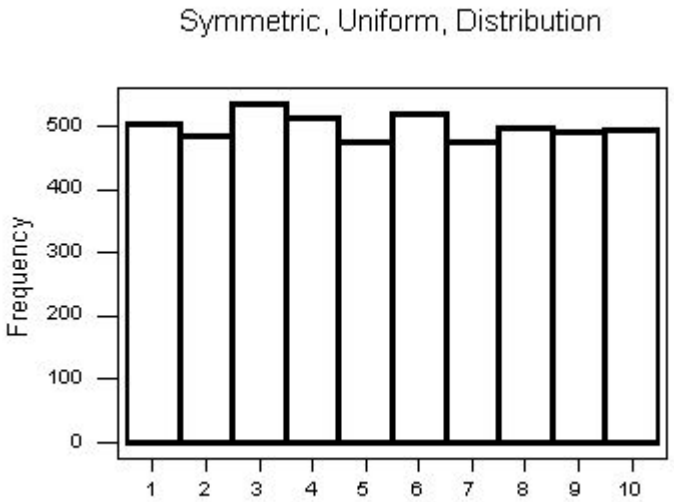


FIGURE 15.6 Examples of normal and skewed distributions

Here are some common distribution shapes

Uniform

The mean here is 5. Does that accurately reflect reality?

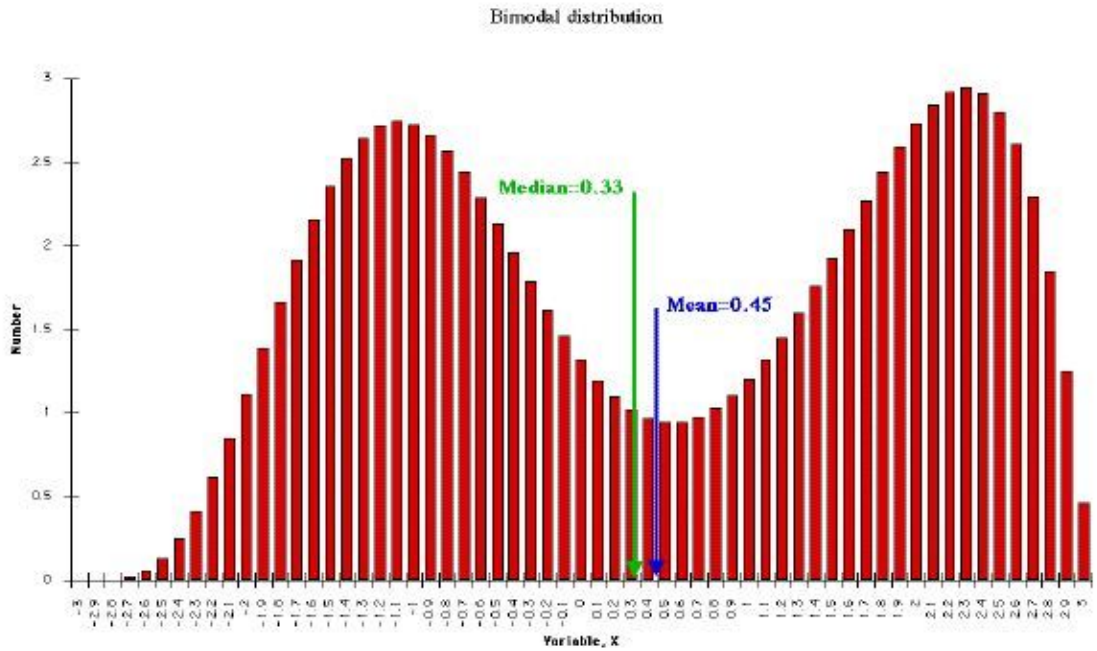


Here are some common distribution shapes

Bimodal

We see this often when we are talking about poverty or race.

Do the mean and median here accurately reflect reality?



Let’s go to the data

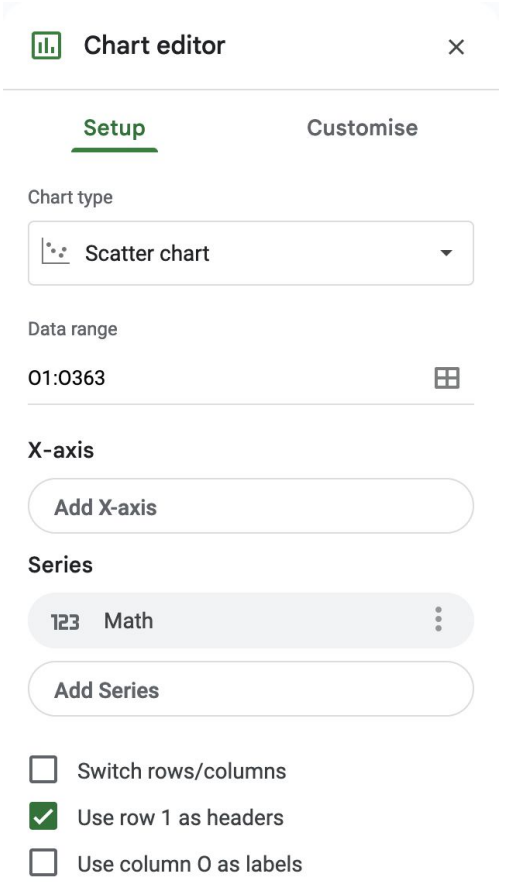
The data for this exercise is school-level. We have average test scores for each school, as well as some characteristics of the school.

Take a look at the data and make sure you understand what’s going on.

Making a histogram

Fortunately google sheets makes it pretty easy to make a histogram.

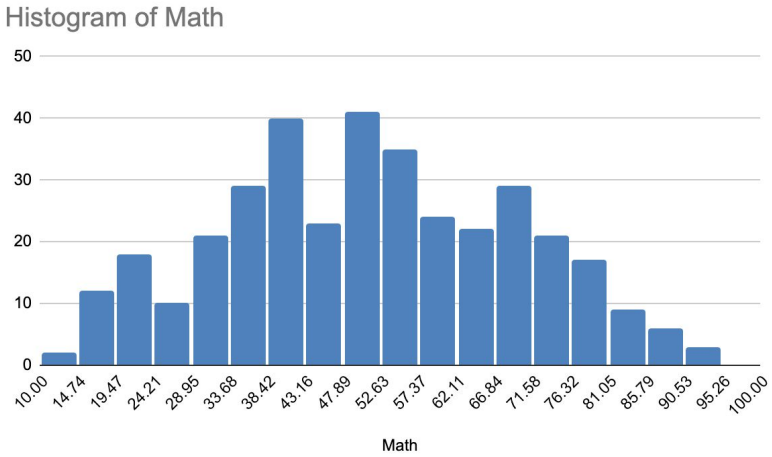
- First let's click on 'Column O' to select the column.
- Then Insert -> Chart
- You should see something like this:

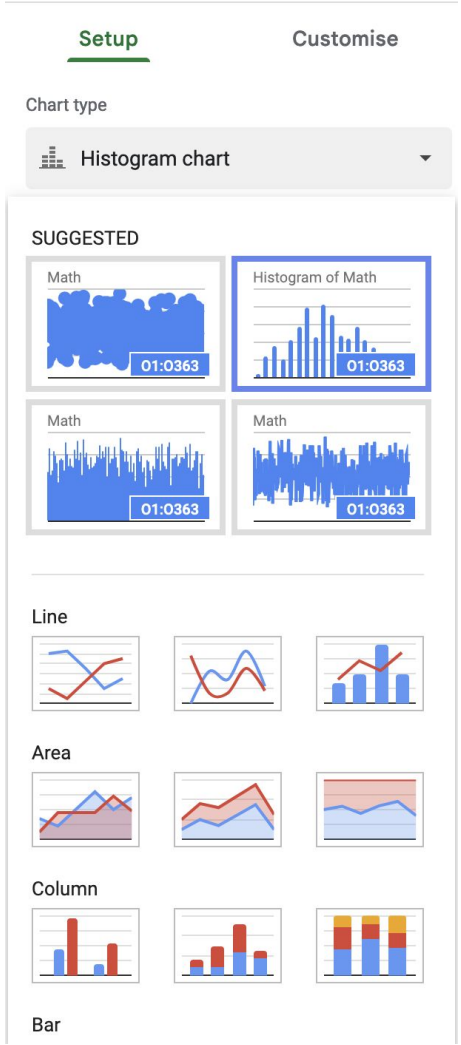


Making a histogram

Choose the histogram option, and you should get this:

Roughly, what distribution type does this approximate?

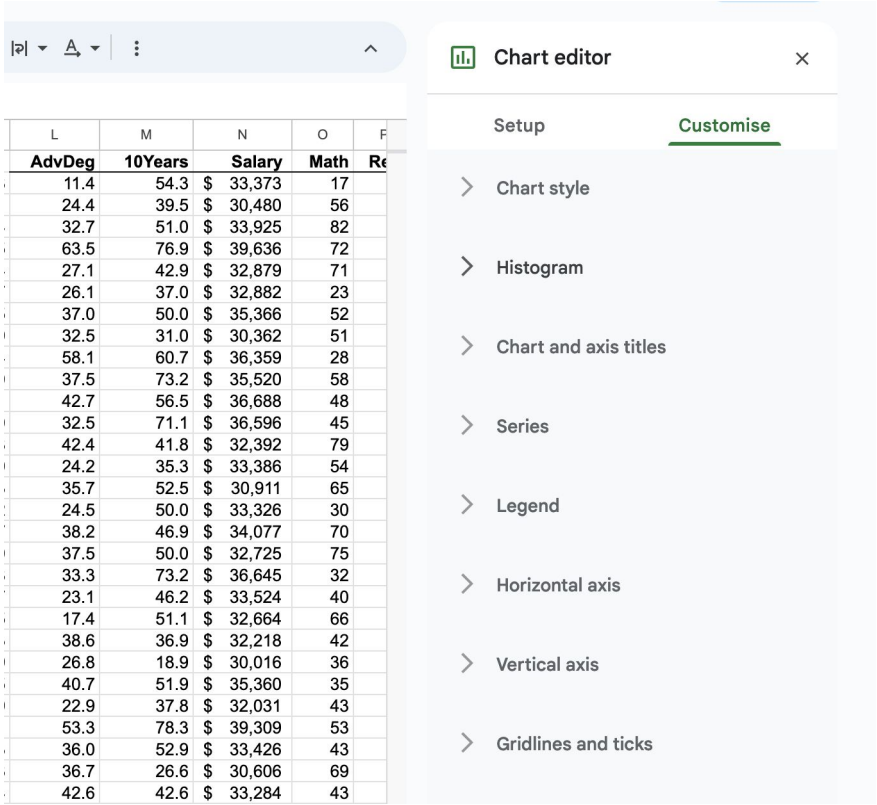




Making a histogram

This is looking okay, but there are a few things we can do to make it look a bit better.

- In the Chart Editor box, let's click the tab that says "Customize" (you can also click the three small dots in the upper right corner to bring this box up in the future)

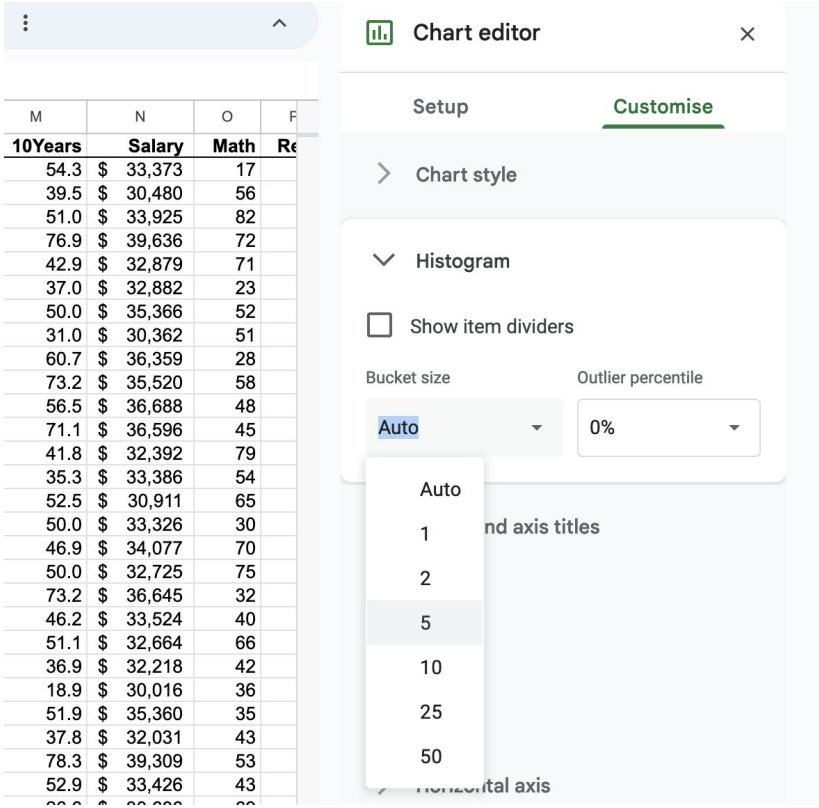


L	M	N	O	P
AdvDeg	10Years	Salary	Math	Re
11.4	54.3	$ 33,373	17	
24.4	39.5	$ 30,480	56	
32.7	51.0	$ 33,925	82	
63.5	76.9	$ 39,636	72	
27.1	42.9	$ 32,879	71	
26.1	37.0	$ 32,882	23	
37.0	50.0	$ 35,366	52	
32.5	31.0	$ 30,362	51	
58.1	60.7	$ 36,359	28	
37.5	73.2	$ 35,520	58	
42.7	56.5	$ 36,688	48	
32.5	71.1	$ 36,596	45	
42.4	41.8	$ 32,392	79	
24.2	35.3	$ 33,386	54	
35.7	52.5	$ 30,911	65	
24.5	50.0	$ 33,326	30	
38.2	46.9	$ 34,077	70	
37.5	50.0	$ 32,725	75	
33.3	73.2	$ 36,645	32	
23.1	46.2	$ 33,524	40	
17.4	51.1	$ 32,664	66	
38.6	36.9	$ 32,218	42	
26.8	18.9	$ 30,016	36	
40.7	51.9	$ 35,360	35	
22.9	37.8	$ 32,031	43	
53.3	78.3	$ 39,309	53	
36.0	52.9	$ 33,426	43	
36.7	26.6	$ 30,606	69	
42.6	42.6	$ 33,284	43	

Making a histogram

What happens if we change things like bucket size and the legend placement?

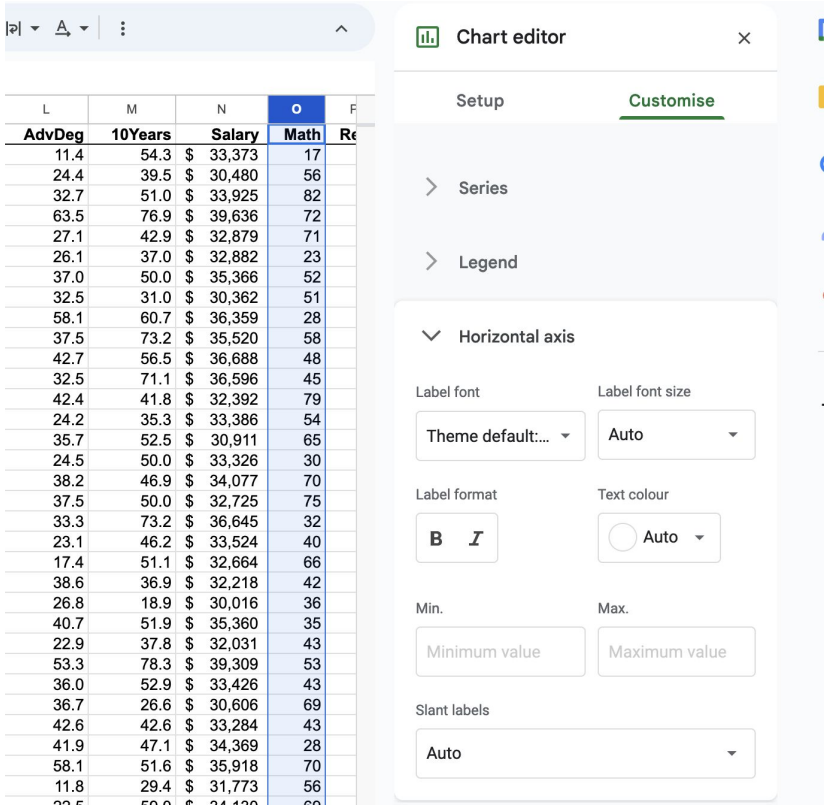
Play with this for a moment.



Making a histogram

We can also adjust qualities of the axis labels and other components of our chart.

- Experiment with the various options in the Chart editor



Making a histogram

As an exercise, let’s make a histogram of the ‘Poverty’ variable. What do you see?