

Google Sheets tutorial

In this tutorial we'll explore the dataset `200M_women.csv` in the same folder as this file. The file contains olympic race times for the track & field event "200M Women". Each row contains data related to the race time of a medalist, and there are race times since 1948.

Read the data

The pre-requisite to any data analysis is being able to read the data in, so let's do that! One way is to right click the CSV and go to `"Open with" > "Google Sheets"`.

Another way is:

- create a new Google sheets
- use `file > import...` to import the CSV

Once you have the data in Google Sheets, freeze the header row.

- highlight the first row, then go to `view > Freeze > 1 row`

Here are what the columns in the data mean:

- Event - the name of the Olympic event
- Location - the city the Olympics were held at
- Year - the year of the Olympics
- Medal - the medal the runner won
- Name - the name of the runner
- Nationality - the nationality of the runner
- Result - the runner's race time in seconds

Explore the data

Average time

Great, we now have the data in Sheets. Let's explore a bit. `Result` contains runners' race times in seconds. What is the mean running time over all rows?

In any cell use the `AVERAGE` function to get the average over all values in the `Result` column. In a cell

you might do:

```
=AVERAGE ( G:G )
```

You should get an average time of ~22.7 seconds.

Merlene

This dataset contains runner Merlene Ottey. According to Wikipedia

Ottey has won the most career Olympic medals (9 medals: 3 silver and 6 bronze) in women's track and field


Wowzers! Let's filter down to her running times.

Use the [FILTER](#) command to get the filtered race times in another column. In a cell you might enter

```
=FILTER(G:G, E:E = "Merlene Ottey")
```

This filters down the **Result** column according to the criteria that the row in the **Name** column equals "Merlene Ottey". You should see four race times:

```
22.2  
22.09  
22.09  
22.24
```

Sidenote: another way to filter data is to use a filter view. They can be created with this icon:  Word of caution: if you're using a filtered view on a spreadsheet, applying a function on the column applies the function to all rows, even ones not showing.

Merlene vs world

Merlene is fast. How much faster is she than the average time of all other medalists? We'll subtract Merlene's mean running time from the average over everyone else.

FILTER comes in handy again because it can be the input to another function. That is we can do

```
=AVG(FILTER(<parameters>))
```

 to first filter based on some condition and compute the mean on those filtered rows. In a cell enter:

```
=AVERAGE(FILTER(G:G, E:E = "Merlene Ottey"))
```

Now we need the average time for all other medalists. In another cell below the previous one enter:

```
=AVERAGE(FILTER(G:G, E:E <> "Merlene Ottey"))
```

In google sheets, `<>` is the not equal operator. We're computing the average race time for all medalists who are not Merlene.

Finally in a third cell subtract Merlene's average time from the average time of everyone else. You should get ~0.62

On average Merlene's running time is 0.62s less than the average over everyone else. That's substantial time for olympic sprints.

Most Common Nationalities

I wonder which nationalities win the most medals? To find out we'll use a [pivot table](#)

A pivot table allows you to summarize your data by groups. If you're familiar with SQL it allows `GROUP BY` operations. In our case our groups are nationalities, and the operation we want to perform on those groups is count the number of medalists. That is for each nationality we want to count the number of rows. Pivot tables are great for these group by then perform operation tasks. For example another thing we might want is to group by runner name and find their shortest race time.

Concretely, highlight the `Nationality` column and go to tab `Data > Pivot Table...`. You want to create a pivot table in a new sheet. Click on the table to see the `Pivot table editor` on the right



In the `Pivot table editor`

- Under `Rows` click `Add` and pick `Nationality`. These are the values we'll group by
- Under `Values` click `Add` and pick `Nationality`. It should automatically show you the `COUNTA of Nationality`. `Nationality` is the column we want to aggregate over for each group, and the operation we want to perform on that column for each group is `COUNTA`.

You should see:

Rows

Add

Nationality

Order

Ascending

Sort by

Nationality

☒

Show totals

Columns

Add

Values

Add

Nationality

Summarize by

COUNTA

Show as

Default

In the pivot table you should see the count for each nationality. You'll notice USA and JAM are tied for winning 11 medals.

	A	B
1	Nationality	COUNTA of Nationality
2		0
3	AUS	7
4	BAH	2
5	EUA	2
6	FRA	1
7	FRG	1
8	GBR	2
9	GDR	4
10	JAM	11
11	NED	3
12	NGR	1
13	POL	3
14	SRI	1
15	URS	2
16	USA	11
17	Grand Total	51

Distribution of running times

We've looked at Merlene's time vs everyone else. I want to know how running times are distributed over all rows. To do so we can plot a histogram of running times. A histogram buckets running times and counts the number of values in each bucket.

Go back to the main sheet with the raw data. Highlight the **Result** column. Go to the tab **Insert** > **Chart** . Change the **Chart type** in **Chart Editor** to **Histogram Chart** . You should see:



Chart editor



Setup

Customize

Chart type



Histogram chart



Data range

G1:G1000



X-AXIS

Add X-axis



SERIES

123

Result



Add Series



Switch rows / columns



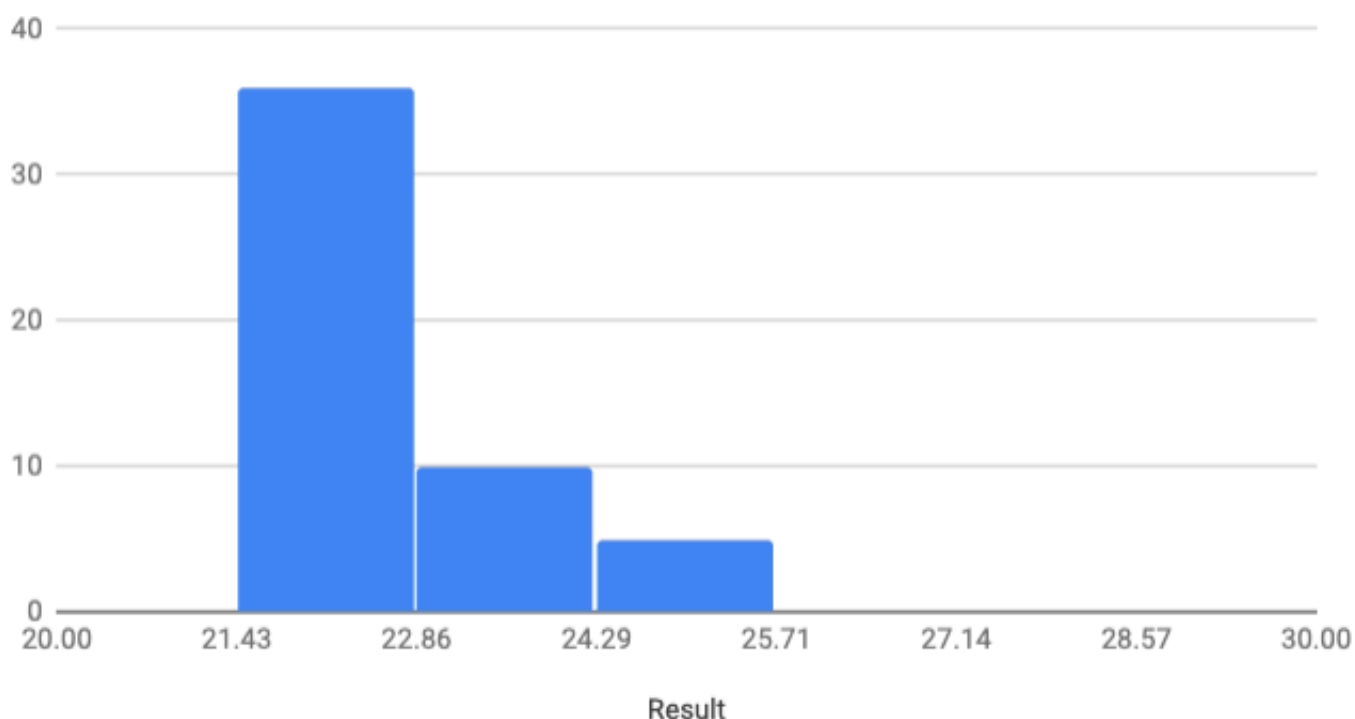
Use row 1 as headers



Use column G as labels

The histogram chart should look something like:

Histogram of Result



That's interesting, it looks like most running times are in the 22-23s range but in some years medalists have times up to 25s! Maybe we've gotten better at running over the years.

Average time vs year

Can we plot the mean running time for each year in a scatterplot?

Highlight the entire sheet and create a pivot table. We'll use the pivot table to find the mean race time for each year. In the `Pivot table editor`:

- Under `Rows` click `Add` and pick `Year`. Order in `Ascending` order.
- Under `Values` click `Add` and pick `Result`. Under `Summarize by` pick `Average`.

This time we're grouping by year and for each year we're finding the average of the `Result` column. You should see:

Rows

Add

Year ✕

Order

Ascending ▾

Sort by

Year ▾

☒ Show totals

Columns

Add

Values

Add

Result ✕

Summarize by

AVERAGE ▾

Show as

Default ▾

Highlight the years and winning times and create a chart (**Insert** > **Chart**). Click on the chart to open the **Chart Editor**

- Go to the **Setup** tab. Change the **Chart Type** to **Scatter chart**



Chart editor



Setup

Customize

Chart type



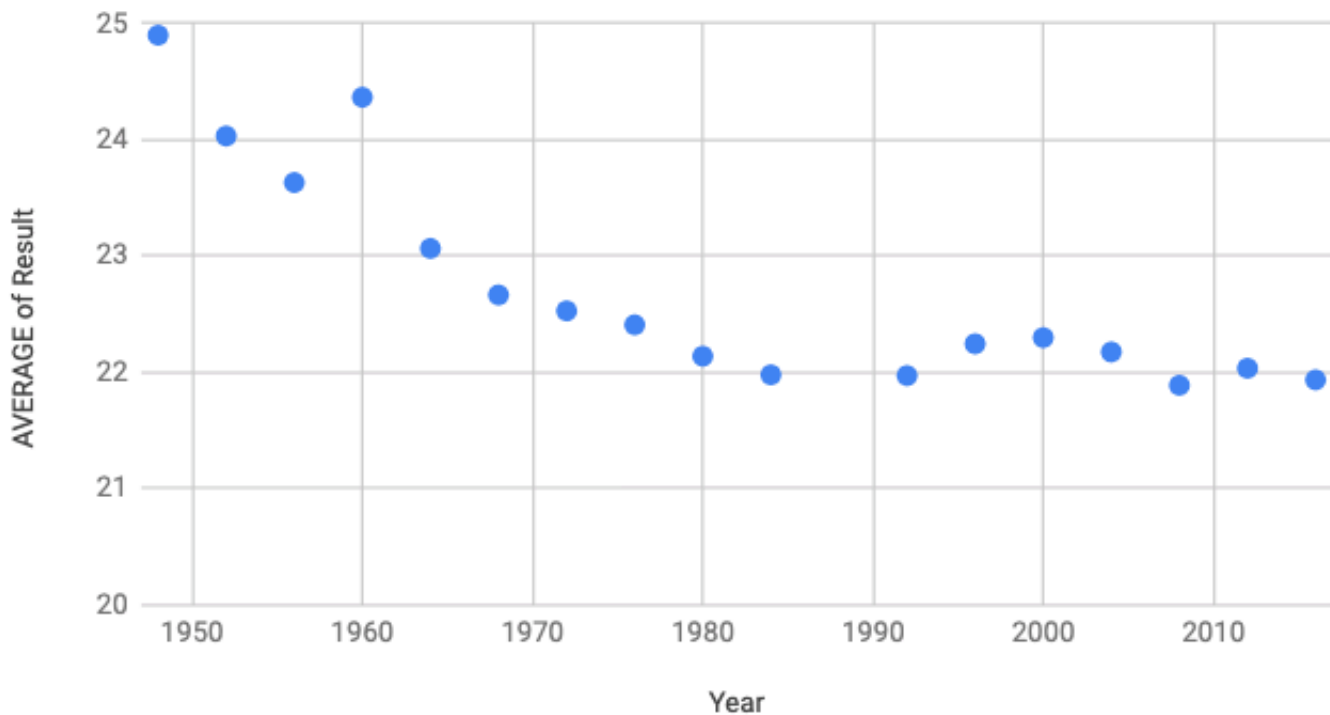
Scatter chart



- In **Customize** open the **Vertical axis** drop down. Add a minimum value of 20 to make the points easier to see.

Now the chart should look something like this:

AVERAGE of Result vs. Year



People have definitely gotten better at running over the years. Now it makes more sense why we see the slower times in the histogram

Conclusion

That's it! You've gone through a simple tutorial of Google Sheets. Thanks for taking the time to get more familiar with this data analysis environment. It'll make you much more productive for the workshop.

If you're interested in additional reading:

- [More on pivot tables](#)

Feel free to explore the dataset further! You might ask some other questions like:

- How does the winning race time change over the years?
- Which runner has been a medalist the most times? What about just gold medalists?
- How does representation of runners from different nationalities change over the years?
- On average how many olympics does each person compete in?

