Excel 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 file and go to `"Open With" > "Microsoft Excel"`.

Once you have the data in Excel, freeze the header row.

- Go to the `view` tab > `Freeze Top 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 Excel. 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.

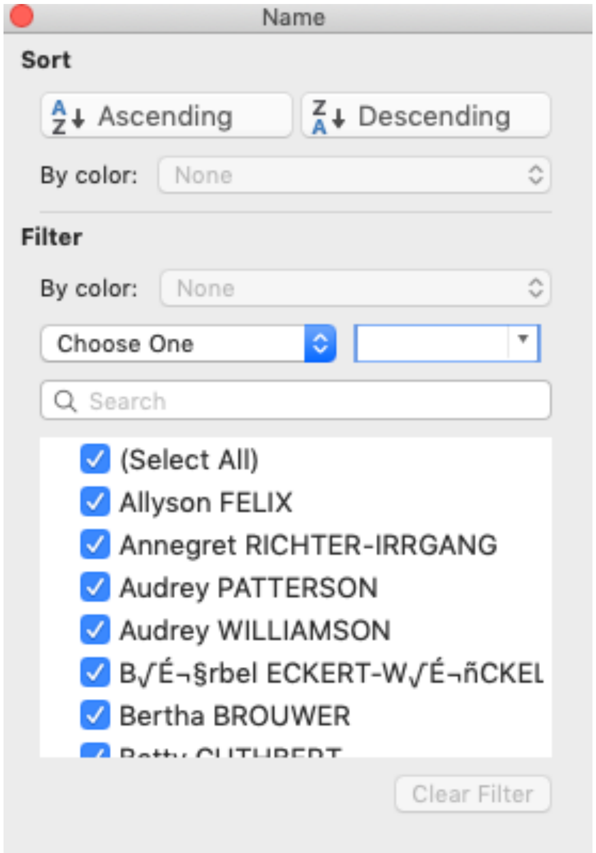
Merlene

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.

1. Highlight the `Name` column and on the `Home` tab click `Sort & Filter`
2. Click the arrow in column Name ▾ to get the following window:



3. Search for `Merlene`. You should see the sheet only displays Merlene's race times, which are race times:

```
22.2
22.09
22.09
22.24
```

Word of caution: when filtering with `Sort & Filter` on a spreadsheet, applying a function on the column applies the function to all rows between the first and last 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 average running time from the average over everyone else.

As mentioned above, we shouldn't apply a function to the rows we get from `Sort & Filter` since it might include hidden rows. Instead to find Merlene's average running time we can use the `AVERAGEIF` function. More information on the function here. The function has syntax `AVERAGEIF(range, criteria, [average_range])`. The way we'll use it is:

- `range` is the range of cells to filter by. In this case it's the `Name` column
- `criteria` is the criteria to filter the `range` cells by. In this case it's `Merlene OTTEY` since we want rows where `Name` is `Merlene OTTEY`.
- `average_range` are the cells to compute the average over. The function will only average the rows where the criteria is true

So, in a cell enter:

```
=AVERAGEIF(E:E, "Merlene OTTEY", G:G)
```

where `E:E` is the `Name` column and `G:G` is the `Result` column. You should get 22.155.

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

```
=AVERAGEIF(E:E, "<>Merlene OTTEY", G:G)
```

In Excel, `<>` is the not equal operator. We're computing the average race time for all medalists who are not Merlene. You should get ~22.77

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

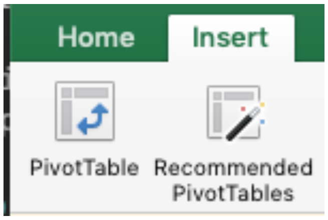
On average Merlene's running time is 0.62s less than the average over everyone else. That's substantial time for olympic sprints.

Most Common Nationalities

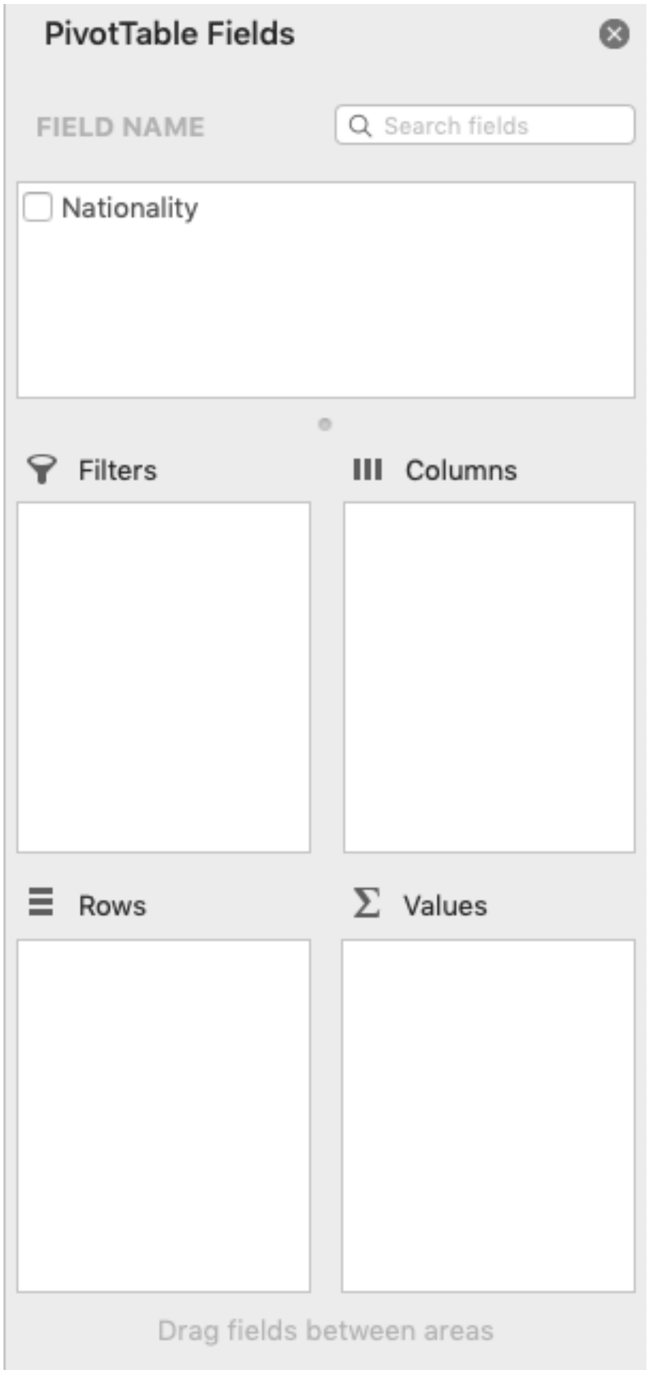
I wonder which nationalities win the most medals? To find out we'll use a pivot table

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.

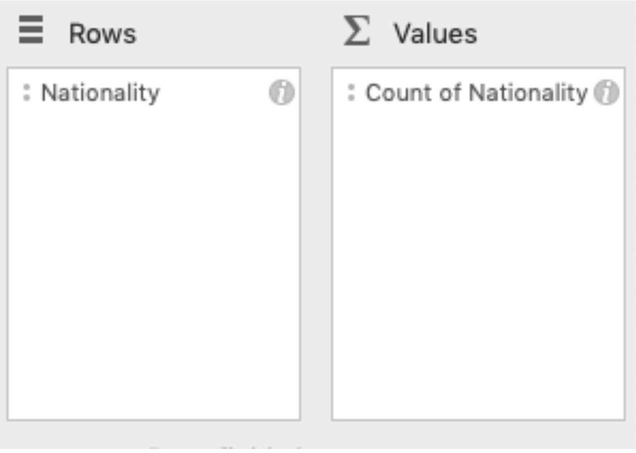
Concretely, highlight the `Nationality` column and go to `Insert` tab > `PivotTable`.



Create the pivot table in a new worksheet. It should open a new sheet. On the right you should see an area to build your pivot table:

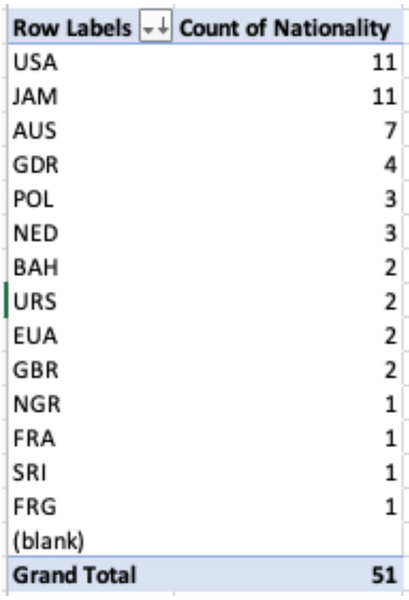


In the `PivotTable Fields` pane, drag `Nationality` to the `Rows` and `Values` sections. You should see this:



Rows	Values
Nationality	Count of Nationality

In the pivot table you should see the count for each nationality. Click on a cell in the `Count of Nationality` column. Then go to the `Data` tab and click ZA↓ to sort in descending order. You'll notice USA and JAM are tied for winning 11 medals.



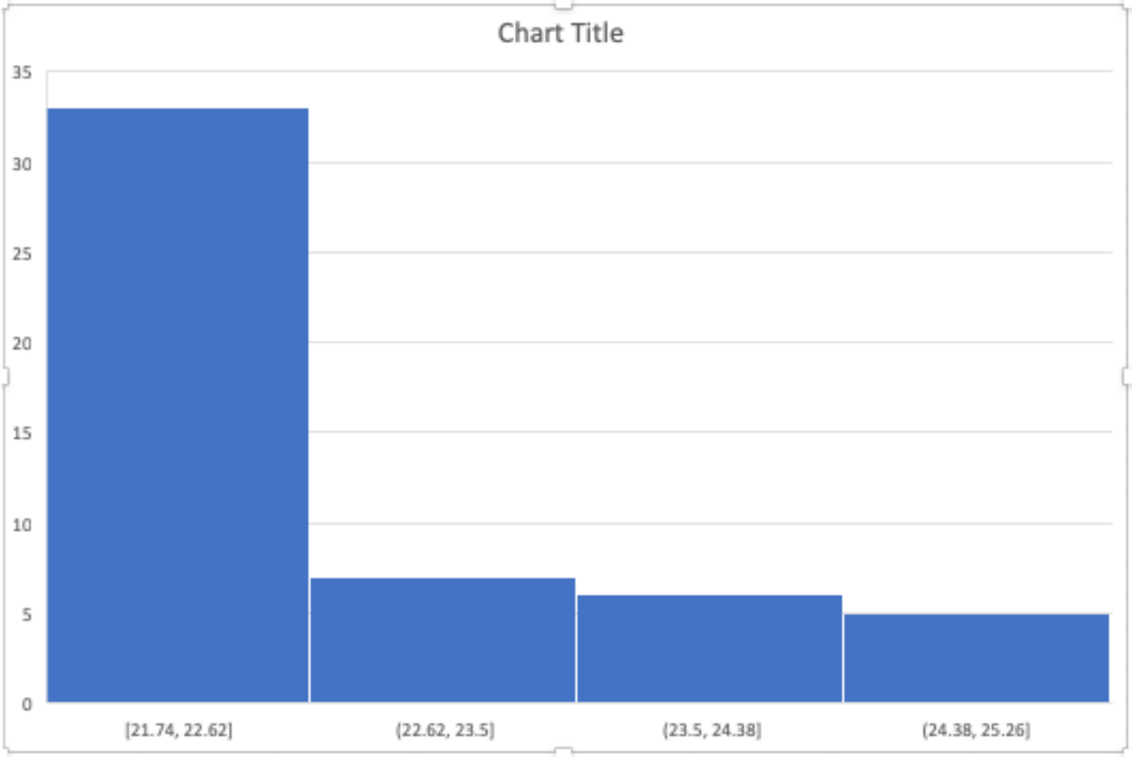
Row Labels	Count of Nationality
USA	11
JAM	11
AUS	7
GDR	4
POL	3
NED	3
BAH	2
URS	2
EUA	2
GBR	2
NGR	1
FRA	1
SRI	1
FRG	1
(blank)	
Grand Total	**51**

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 to this link to learn how to create a histogram. Create a histogram on the `Result` column

The histogram chart should look something like:



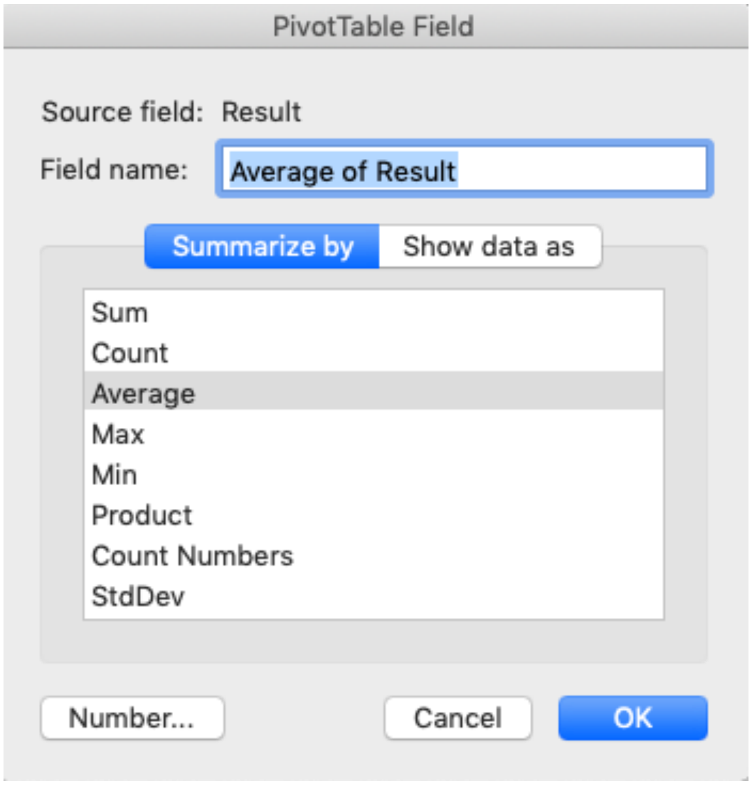
That's interesting, it looks like most running times are in the 22-23s range but in some years medalists have times up to 25s! Maybe we've gotten better at running over the years.

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 in a new sheet. We'll use the pivot table to find the mean race time for each year. In the `PivotTable Fields` Pane:

- Drag `Year` to the `Rows` section
- Drag `Result` to the `Values` section.
- Click on the `i` next to `Result` and choose to Summarize By Average



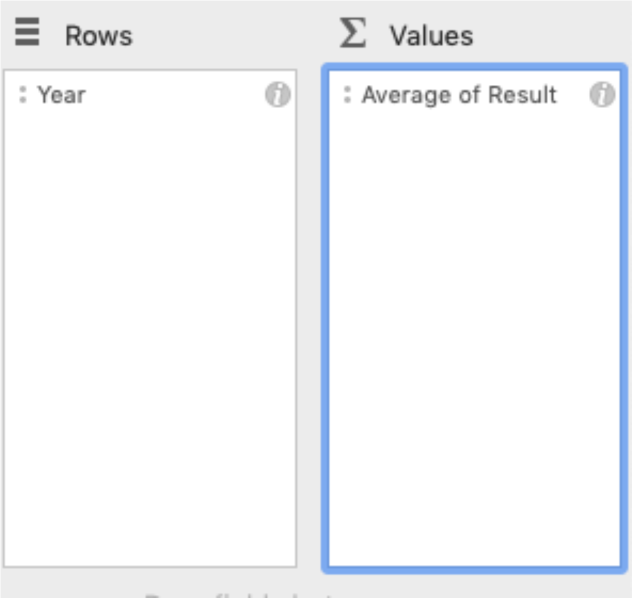
PivotTable Field

Source field: Result

Field name: Average of Result

Summarize by: Average

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

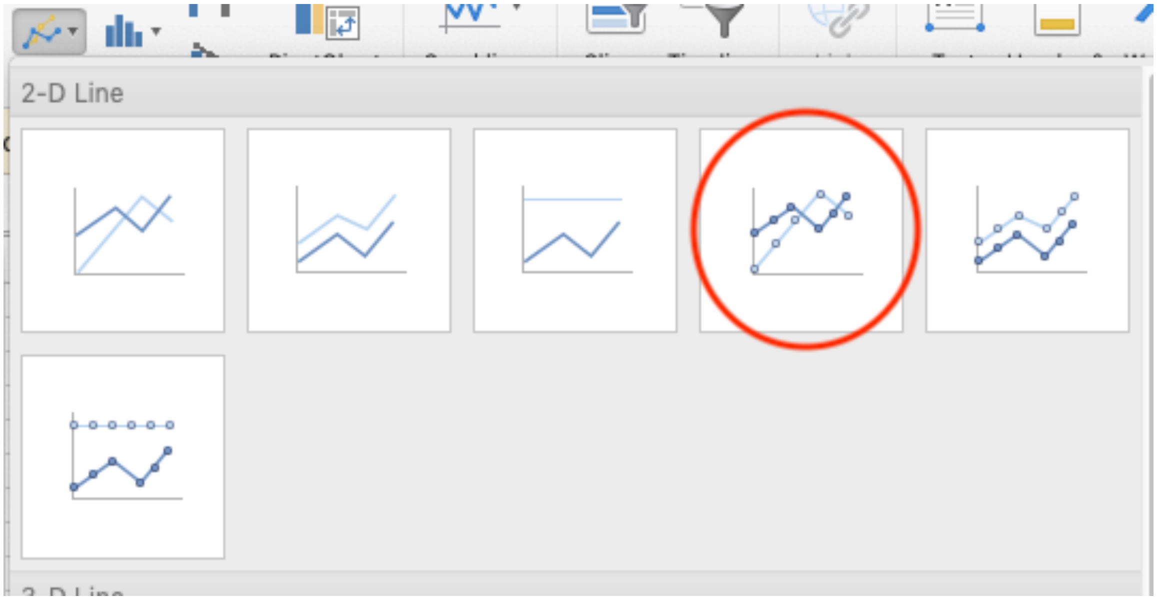


Rows: Year

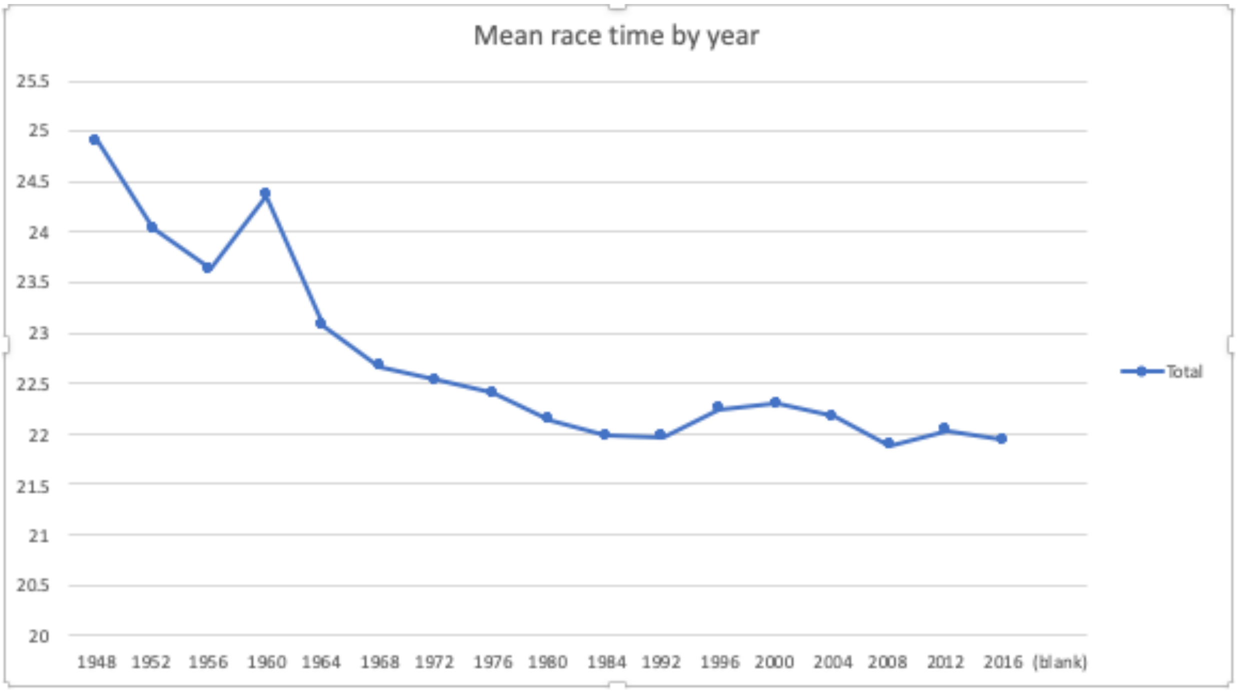
Values: Average of Result

In the pivot table you should see the average time for each year.

Highlight the years and average times and create a chart. Go to the `Insert` tab and click the line chart icon:



Now the chart should look something like this:



I changed the title to be more descriptive. 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 Excel. Thanks for taking the time to get more familiar with this data analysis environment. It'll make you much more productive for the CTF.

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?