

## Part 2.1 | Relationships Between Numerical Variables

### *Why Visualize Relationships?*

Before diving into scatter plots, let's see why visualizing relationships matters. Consider four datasets. If we calculate summary statistics for each — the mean of X, the mean of Y, the standard deviation of X, the standard deviation of Y, and even the correlation between X and Y — we get nearly identical numbers.

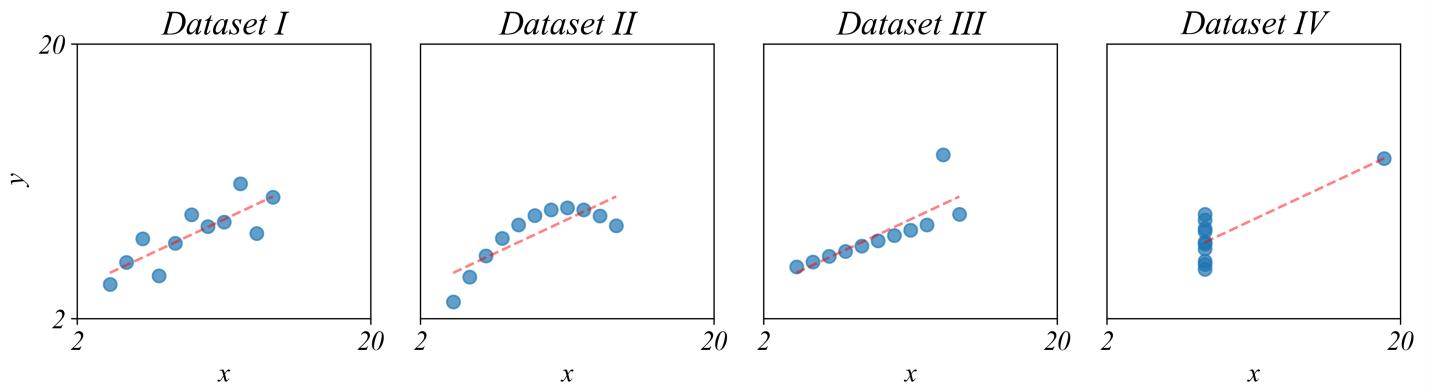
#### *Summary Statistics: Four Datasets*

Dataset	Mean( $x$ )	Mean( $y$ )	Std( $x$ )	Std( $y$ )	Corr( $x,y$ )
I	9.0	7.5	3.32	2.03	0.82
II	9.0	7.5	3.32	2.03	0.82
III	9.0	7.5	3.32	2.03	0.82
IV	9.0	7.5	3.32	2.03	0.82

Based on these statistics alone, you might conclude the four datasets are essentially the same. But are they?

Let's plot each dataset as a scatter plot.

#### *Anscombe's Quartet: Same Statistics, Different Patterns*



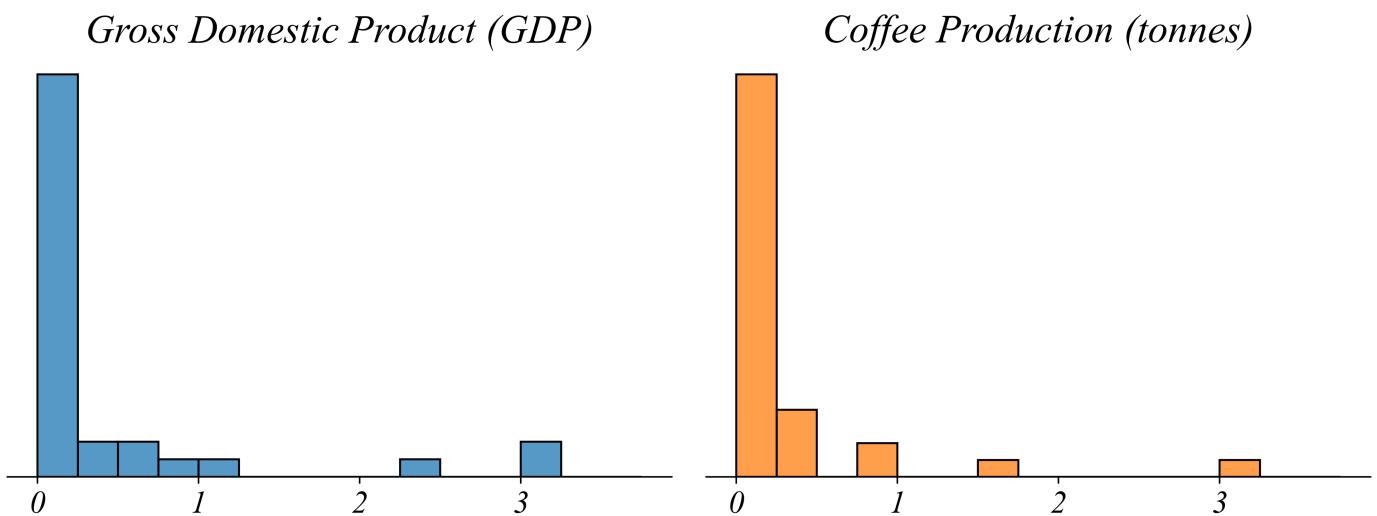
The four datasets look completely different! The first shows a linear relationship. The second shows a curved relationship. The third has one outlier pulling the statistics. The fourth has all points stacked vertically except for one outlier.

This is called **Anscombe's Quartet**, and it teaches us an important lesson: summary statistics can hide important patterns. To truly understand data, we need to visualize the *relationships* between variables, not just summarize each variable separately.

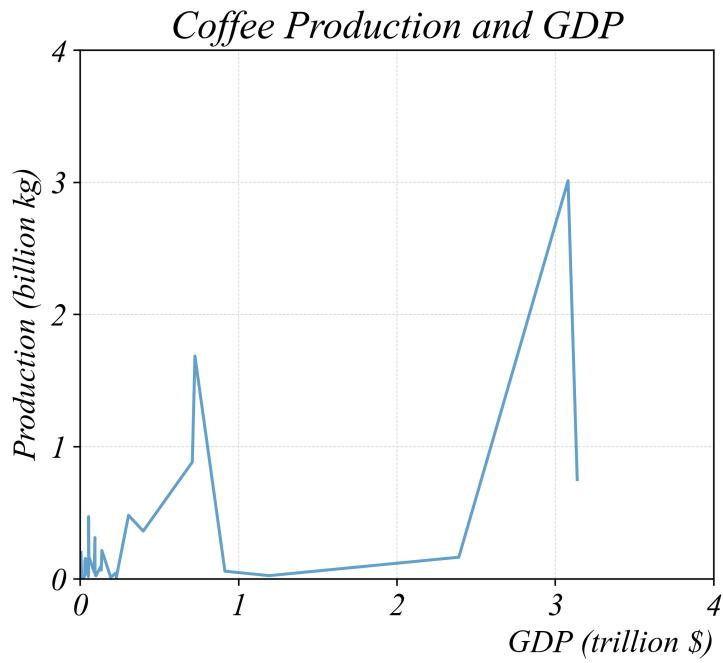
That's what Part 2 is all about — exploring relationships between variables.

## *Scatterplots in Cross-Section*

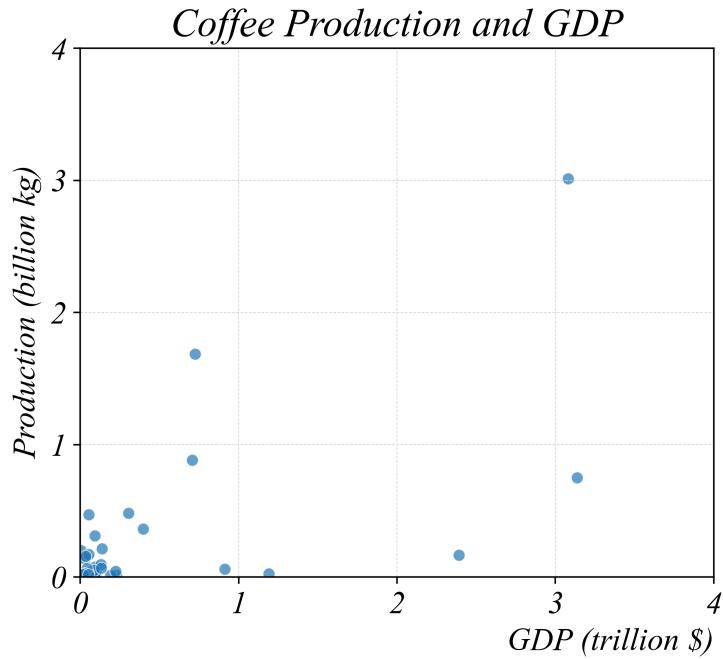
Coffee is consumed globally, but grown in just a few countries. Is there any pattern in the relationship between coffee production and gross domestic product (GDP)? **Gross domestic product (GDP)** measures the value of goods and services produced in a country over a defined period. This dataset has coffee production (in billion kilograms) and GDP (in trillion US dollars). I've dropped very large GDP countries (US and China and India) to focus on smaller countries with higher coffee production. Lets find a useful way to visualize the data. The first thing we might do is create histograms of the two variables.



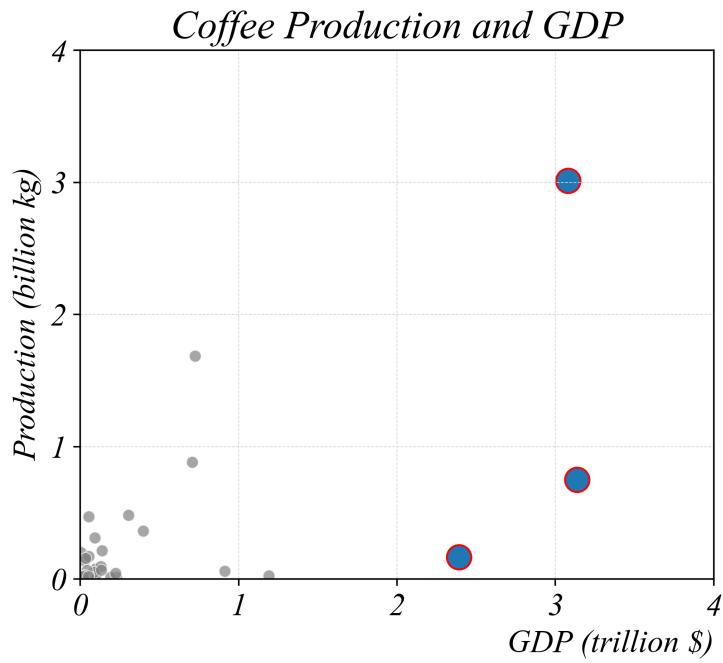
But maybe we want to understand the relationships between the variables. A line graph can plot one variable against the other but isn't going to work in this case. Line graphs are used to measure relationships between individual points, which is not what we're trying to do here.



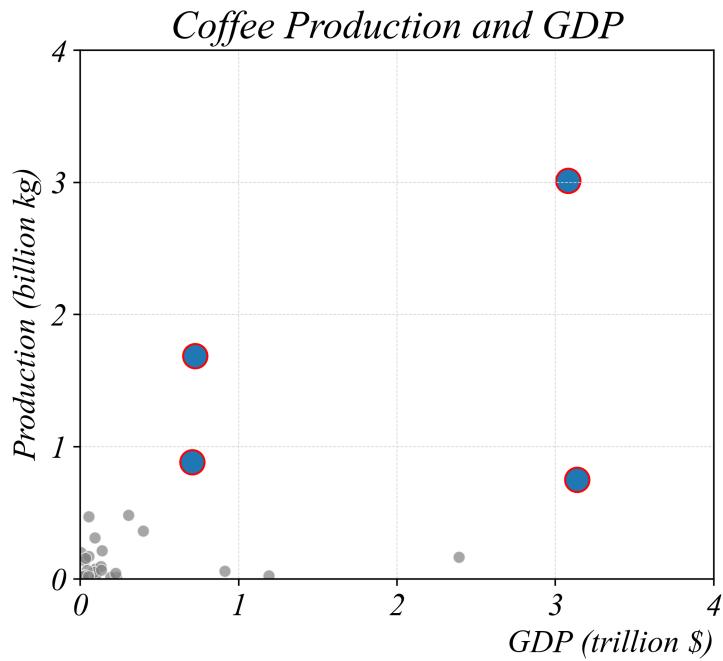
We want to measure a systematic relationship between the two variables not the individual points. A scatter plot allows us to explore the systematic relationship between two variables. Here we can use coffee production and GDP with a scatter plot to show the relationship between the two variables.



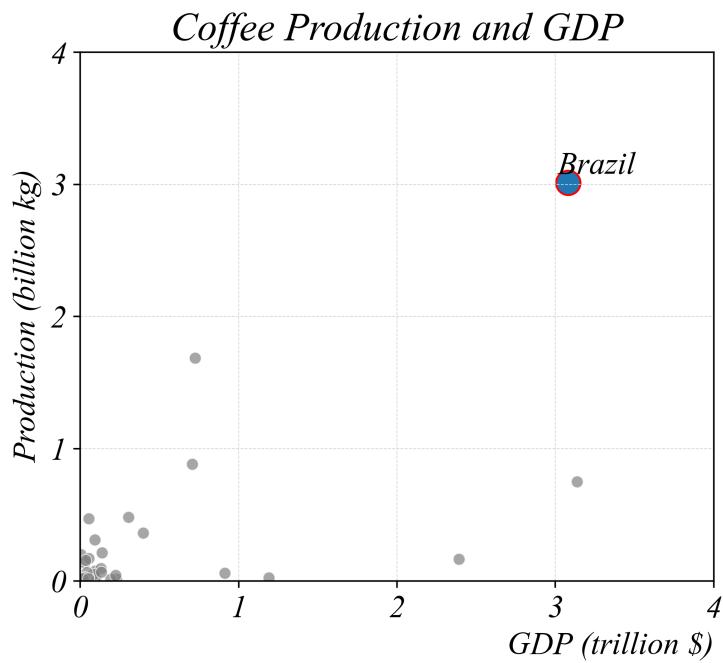
First, let's see what it can tell us about each of the variables individually. Which countries have a GDP above 2 trillion USD? The four highlighted countries — India, Brazil, Mexico, and Indonesia — have a GDP larger than 2.5 trillion USD. This can be seen from their position along the GDP axis.



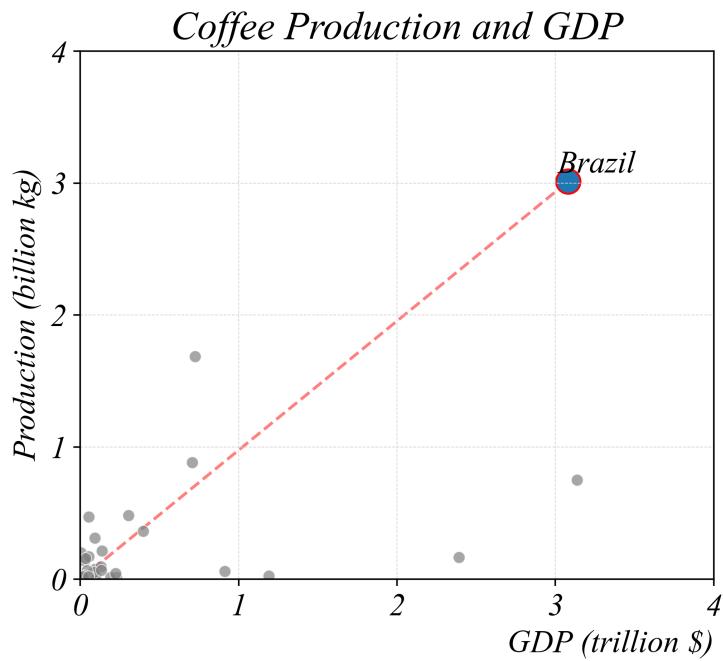
In our scatter plot, the more to the right a country is, the larger its GDP, and the more to the top, the larger its coffee production. Which points represent an production greater than 0.5 Billion kg per year? Brazil, Colombia, Indonesia, and Vietnam. These are not the same countries as those with high GDP. The vertical axis tells us different information than the horizontal axis, which is the main point of scatter plots.



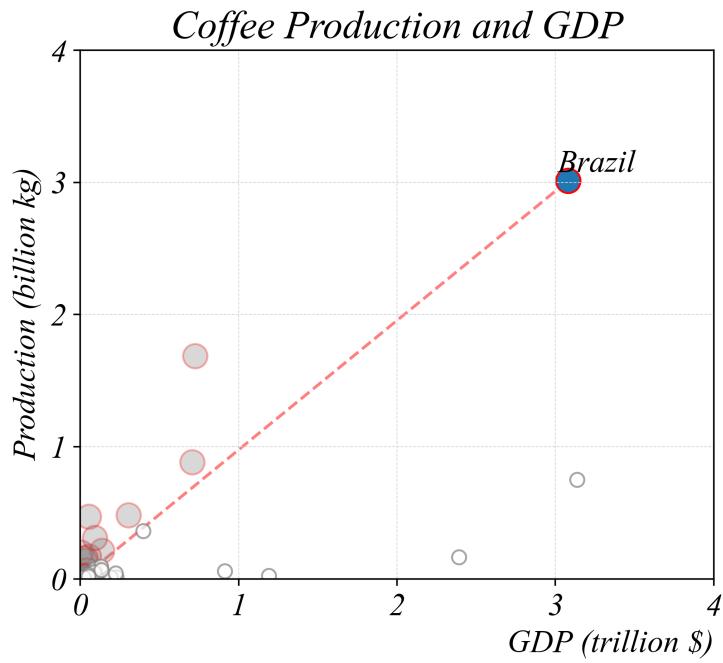
With this view we can examine relationships *between* the two variables. Let's consider a subset of coffee producers. I've drawn a line between the origin and Brazil's point. Any country *on* the red line would produce exactly as much coffee per dollar of GDP as the largest coffee producer — Brazil. Which countries produce **less** coffee per dollar of GDP than Brazil?



Countries below the red line produce less coffee per dollar of GDP than Brazil, and countries above the red line produce more coffee per dollar of GDP than Brazil.

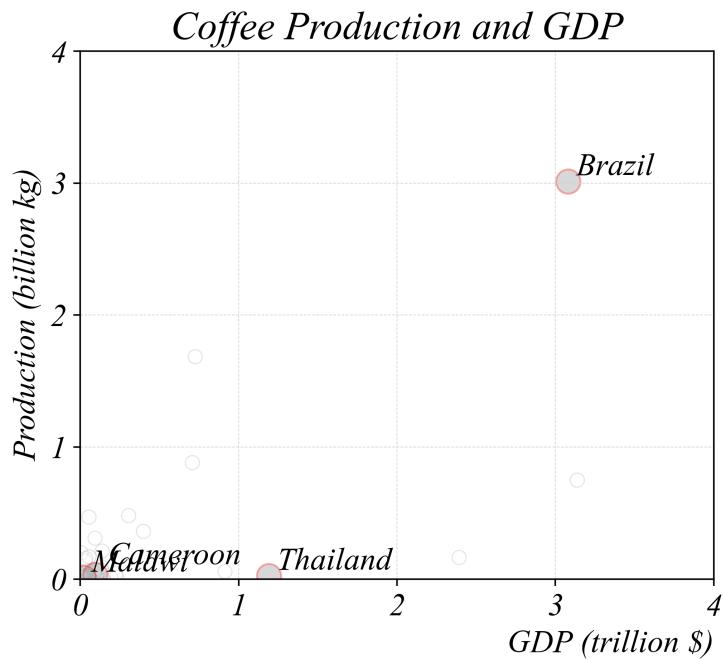


What generalization can you make from this chart? It seems that larger GDP and coffee production tend to go hand in hand, but the plot isn't clear. Lets transform the data to get a clearer picture.



### *Adjusting The Axes*

Our scatter plot provided some insights into the GDP — coffee production relationship for a handful of countries, but we're still seeking a clear overall pattern. Let's look at the GDPs of four countries: Brazil, Malawi, Cameroon, and Thailand.

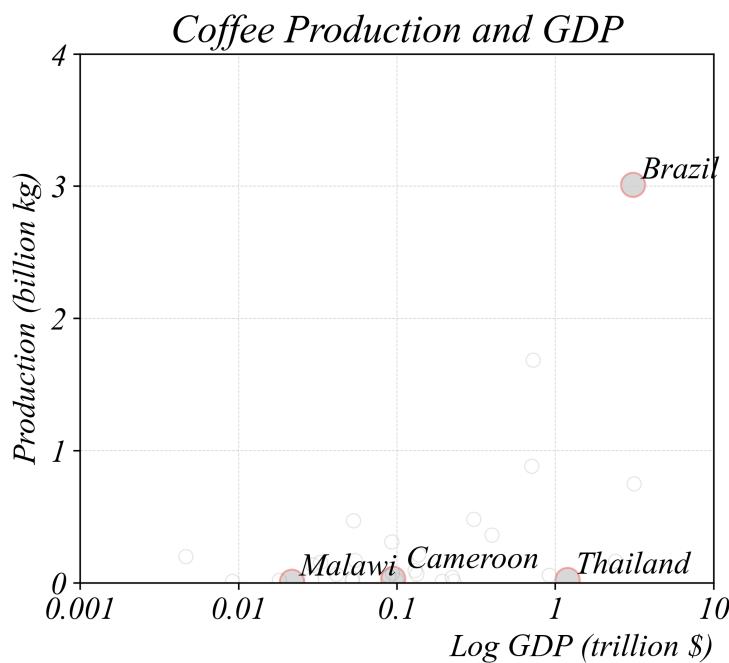


The GDPs of which pair differ by a larger amount?

- Brazil and Thailand
- Cameroon and Malawi

By comparing their position on the  $x$ -axis, we can see that the GDPs of Brazil and Thailand differ by almost 2 trillion dollars. The difference between Malawi and Cameroon is very small on this plot. So, the first difference is much bigger. Now consider the ratios of GDPs: Brazil's to Thailand's, and Cameroon's to Malawi's. Which ratio is larger?

In our scatter plot, each unit on the  $x$ -axis represents the same difference between GDPs. This means that Malawi and Cameroon are placed almost on top of each other, while there's a large empty space between Thailand and Brazil. To make the plot less crowded, we'll transform it so that each unit on the  $x$ -axis represents the same **ratio** between GDPs, using a **logarithmic scale**.

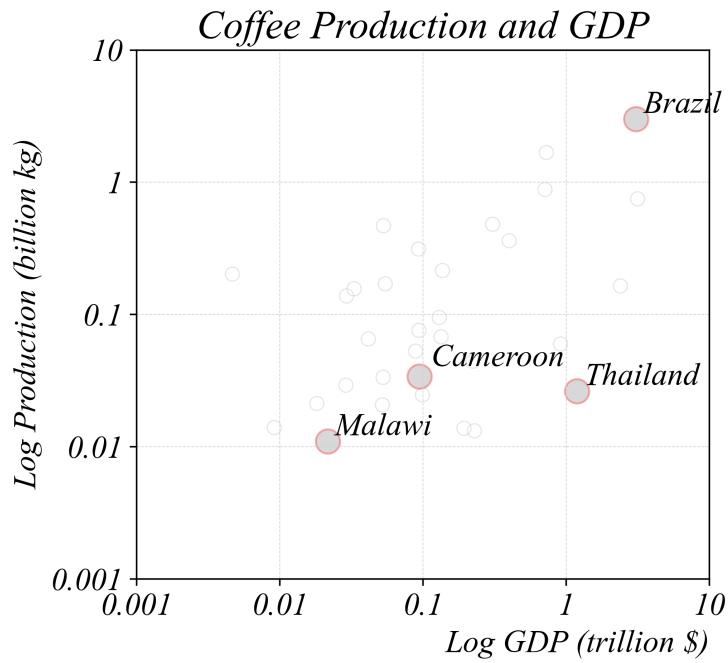


On a **linear scale**, each unit corresponds to adding the same amount. On a **logarithmic scale**, each unit corresponds to multiplying by the same amount.

A base-10 **logarithm** tells us how many times we need to multiply 10 to get another number. For example, it takes 3 multiplications to get 1000 from 10, so  $\log(1000) = 3$ . A base-10 logarithm gives us roughly the number of digits in a number, or its order of magnitude.

Thanks to the logarithmic  $x$ -axis, we can clearly see the GDPs of all four countries. Excluding Brazil, which of these countries produced the largest amount of coffee?

All three countries are very close to the same horizontal line, so it's difficult to tell visually. To answer this question, we need to adjust the plot. We can perform this same ratio-scaling operation, adjusting the vertical axis from linear to logarithmic.

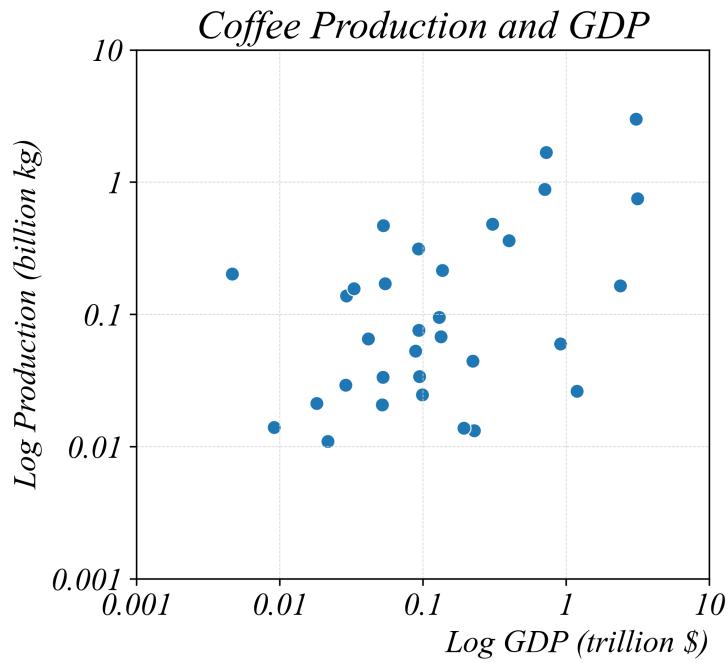


Which of these three countries produced the largest amount of coffee? By changing the  $y$ -axis from linear to logarithmic it's much easier to see that Cameroon produced more coffee than both Malawi and Thailand.

Now we're ready to explore the relationship between coffee production and GDP.

### *Coffee Production vs. GDP*

The new tool — logarithmic scales — will help us describe the relationship between coffee production and GDP. The original data presented in a **log-log scale** makes it much easier to see general relationships between coffee production and the GDP.

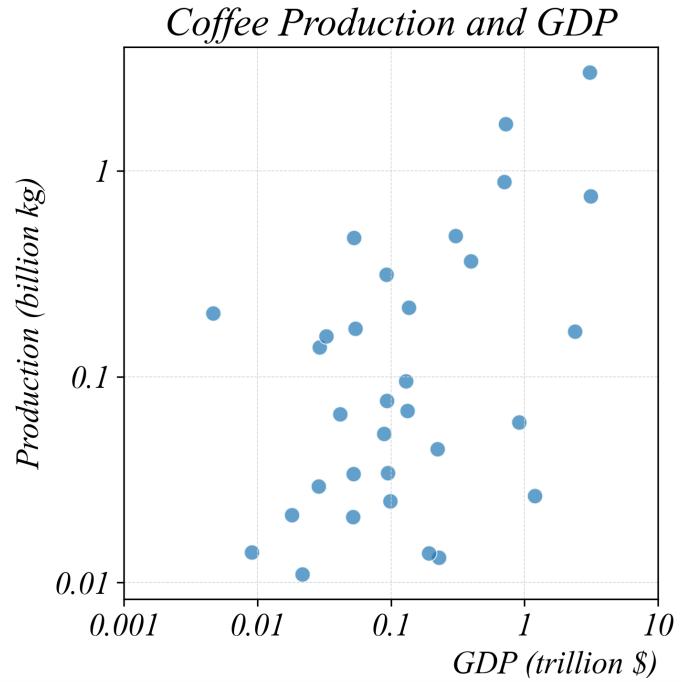


In the log-log plot we can see that in general, countries with larger GDP tend to produce more coffee, and vice versa. On the whole, larger GDP indicates larger coffee production. But is it always true? We'll explore filtering by region in Part 2.3.

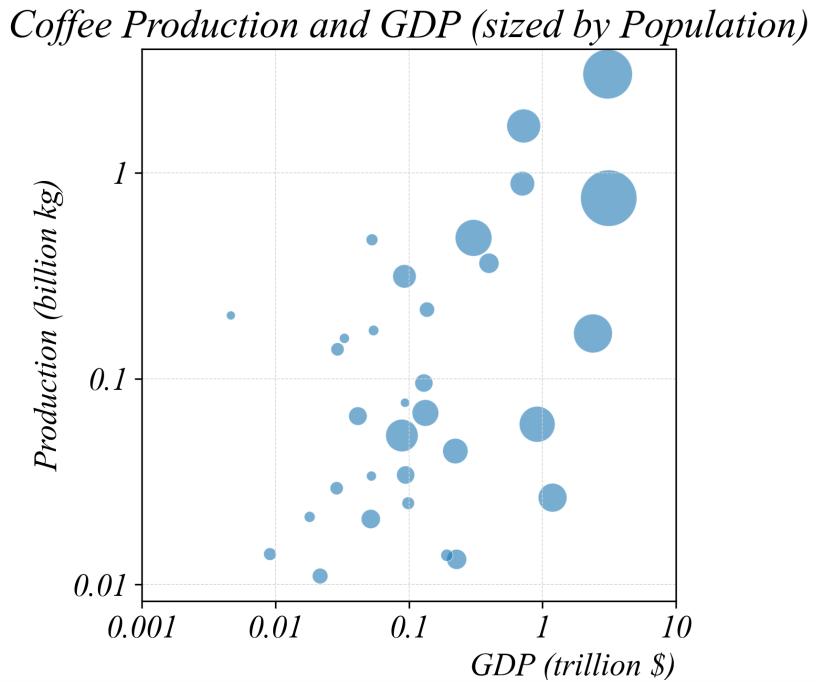
## *Adding More Variables: Size Encoding*

So far we've used position to encode two variables — GDP on the x-axis and coffee production on the y-axis. But what if we want to show a third variable? We can use the **size** of each point.

Here's our standard scatter plot:



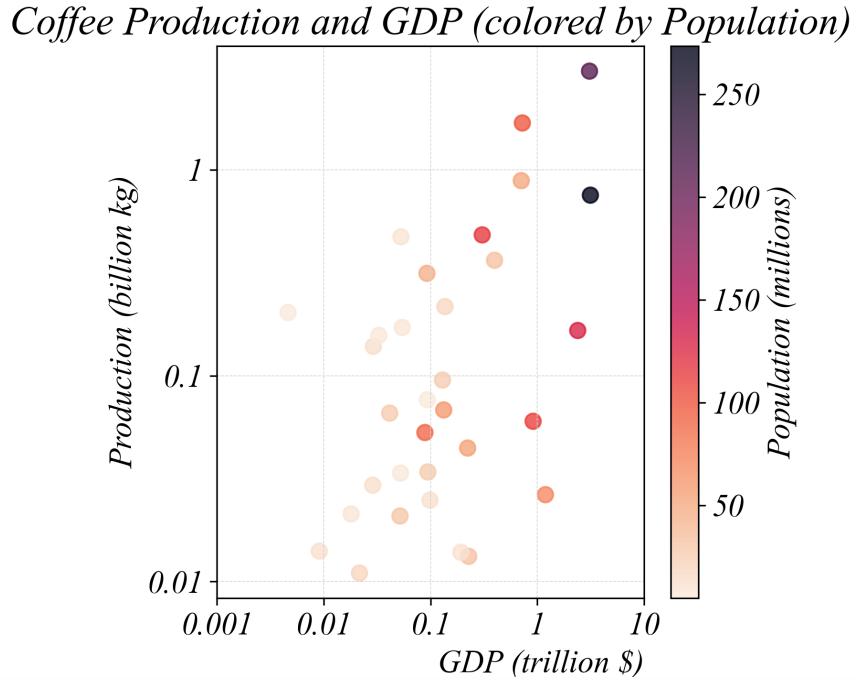
Now let's make each point's size represent that country's population. Larger bubbles mean larger populations.



This type of visualization is called a **bubble chart**. We can now see three variables at once: GDP (horizontal position), coffee production (vertical position), and population (bubble size). Larger countries stand out giving us richer insight than a simple two-variable scatter plot, even if only marginally.

## *Adding More Variables: Color Encoding*

We can also use **color** to encode a third numerical variable. A **continuous color scale** maps values to shades — lighter colors indicate lower values, darker colors indicate higher values.



Here, each point's color represents that country's population. Darker points have larger populations. Similar to bubble charts, color scales can add a small amount of visual information but can also make the main visualized relationship less clear.

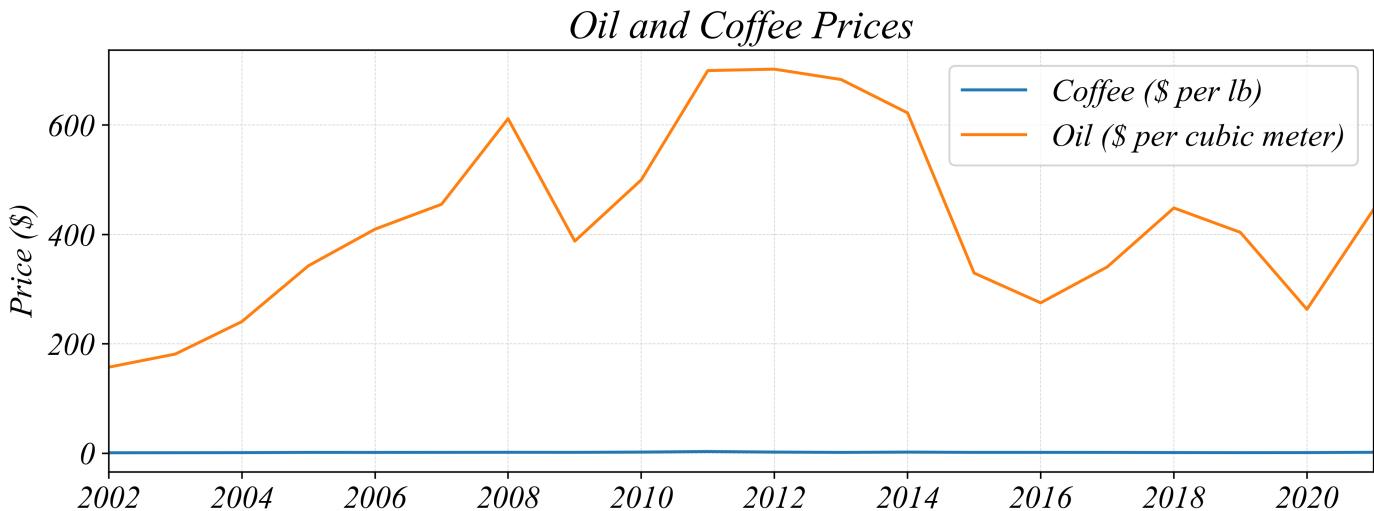
## *Scatterplots and Timeseries Relationships*

Next, we'll compare coffee prices with the prices of another commodity: oil. A **commodity** is a product that can be traded or sold, such as coffee or oil.

Coffee and oil prices depend on many factors, from transportation costs to geopolitical events to economic conditions. Let's see how the prices of these two commodities relate to each other.

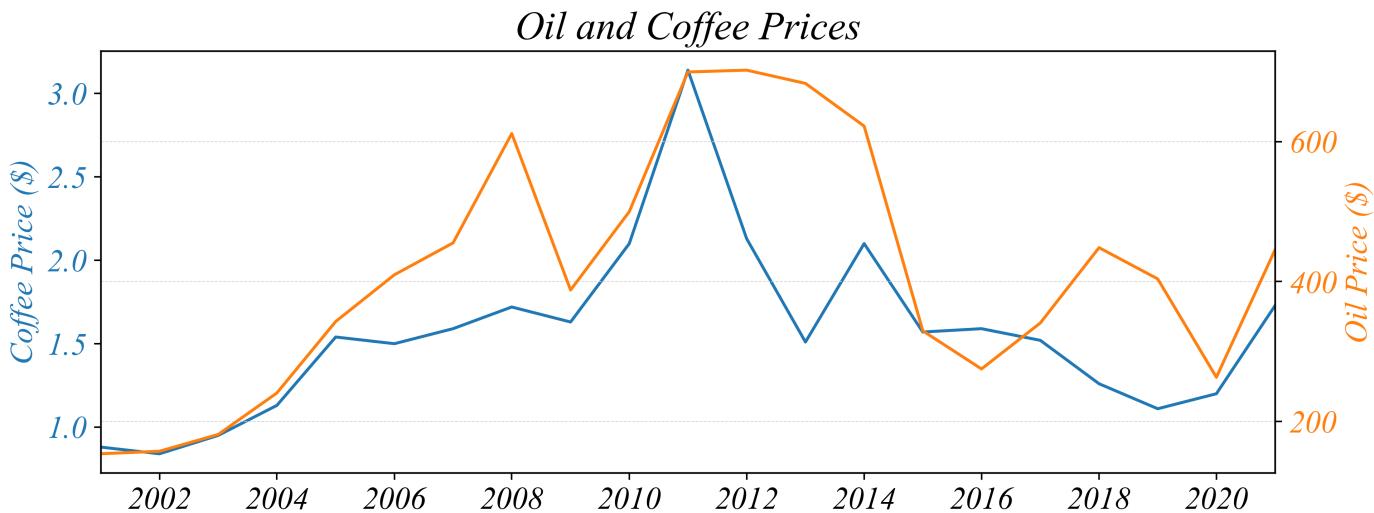
To focus on the bigger picture, we'll consider the average yearly prices of coffee (in US dollars per pound) and crude oil (in US dollars per cubic meter).

We can see that oil is much more expensive in these units. We can see this because the crude oil line is consistently higher than the coffee line, indicating a higher unit price.

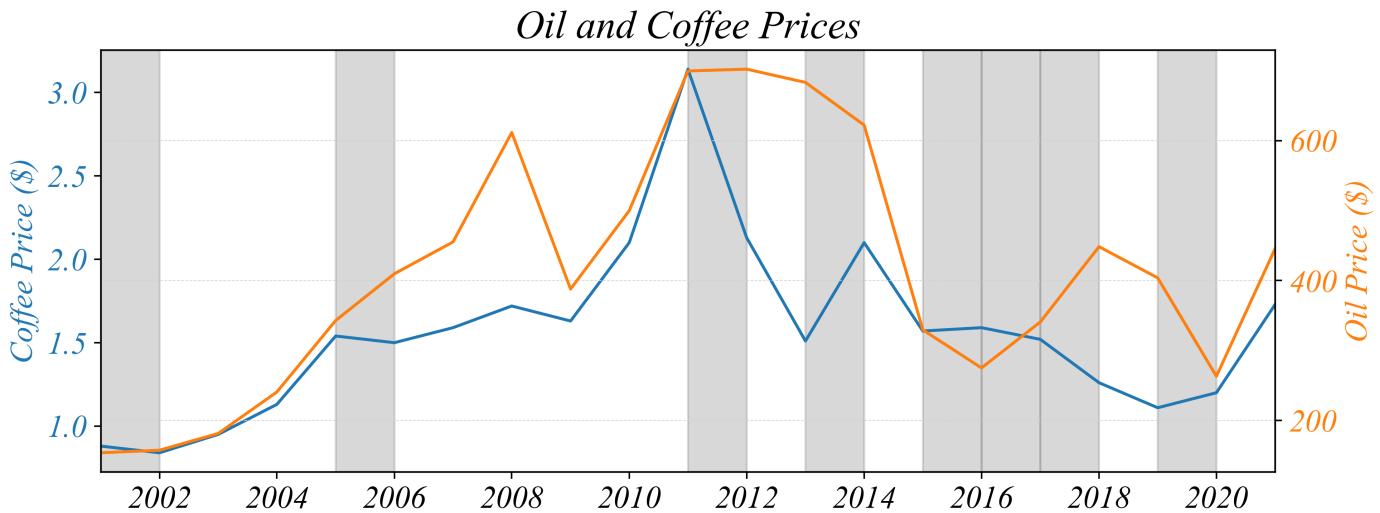


What can be inferred about the two time series? It's difficult to tell. The unit of oil is always much more expensive than the unit of coffee making the variations in coffee prices negligible when compared to oil prices. So, we can't use this graph to analyze the patterns of coffee prices.

Because the values of the two time series differ in orders of magnitude, the previous graph didn't show the comparatively small variations in coffee prices. We can fix this graph by adding a separate *y*-axis.



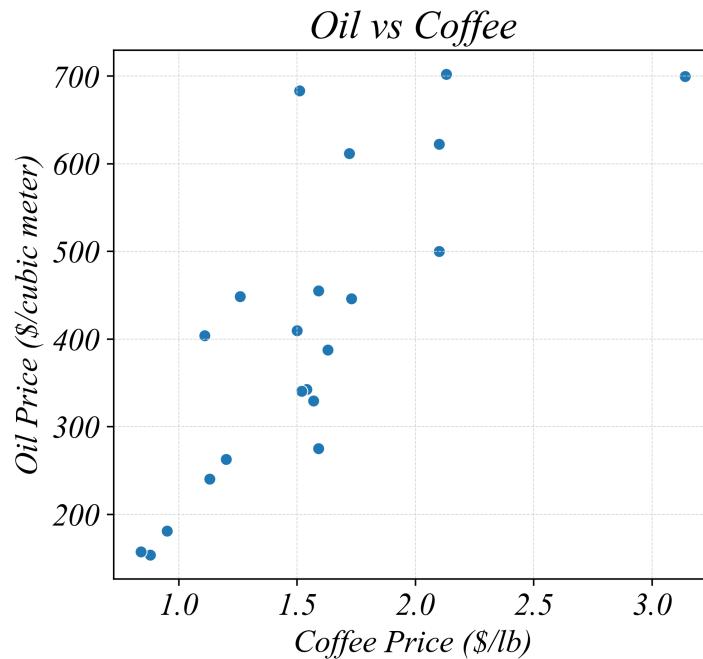
Most of the time, when the price of oil increased, so did the price of coffee, and vice versa. Which years did oil and coffee prices move in **opposite** directions?



In 2001 coffee decreased and oil increased. In 2002, 2003, and 2004, both prices increased. In 2005 oil increased and coffee decreased. ....

This is one way of understanding the relationships between prices. But we've already talked about another one: we can think of each year's price as a separate point in a scatter plot.

The line graph shows that in general, lower coffee prices corresponded to lower oil prices. This is reflected in the scatter plot.



What do you think is true?

- Coffee price determines oil price.
- Oil price determines coffee price.

- Coffee and oil prices are related.

Coffee prices could impact oil prices, oil prices could impact coffee prices, they both could be influenced by a different, unknown process, or their graphs could look similar by chance.

We need to be careful with drawing too many conclusions from the data! We'll talk about how we might disentangle correlation from causation later in the semester. But we always need to be careful.

Next, we'll look closer at seasonal changes in coffee prices.

## *Excel Exercise*

We'll look at 1999 and 2019. The data needs to be adjacent. So to do this we'll create a new sheet, copy over all the day, and delete the data we don't want. We have to be careful with the label. If we accidentally select the title, because it's a number, it will add it as a point. So only select the data.

Then go to insert, insert scatter plot, and it should be good. Add a title. We want to be a little careful with which axis is which. So we can go to "Select Data" and make sure we know which column is the x and which is the y.

That's about it. But we might want to be careful with logs. This one doesn't matter so much, but many data examples will be best with logs. So let's create a log version. We can use the excel function

=LOG(CELL, BASE)

Then we can plot this one with a scatter.