ECON 0150 | Economic Data Analysis

The economist's data analysis pipeline.

Part 1.4 | Panel Data

Panel Data: Long Format vs Wide Format

Panel data comes in one of two formats.

Panel Data in ***Long Format*** uses lists each entry as a row, using a column (eg. *Shop*) to record the group.

	Hours	Shop
transaction_id		
7	12	Shop A
11	15	Shop A
19	14	Shop A
32	16	Shop A
33	19	Shop A

Panel Data: Long Format vs Wide Format

Panel data comes in one of two formats.

Panel Data in ***Wide Format*** uses lists each group as a row, using a column to record each entry.

	1999	2019
Code		
AUT	8.430589	7.925747
BGR	2.652661	3.638313
HRV	4.480790	5.623266
CYP	3.477888	5.615070
CZE	3.255587	4.739563

Hiring a Barista

Use Coffee_Sales_Receips.csv to help inform where to hire a barista.

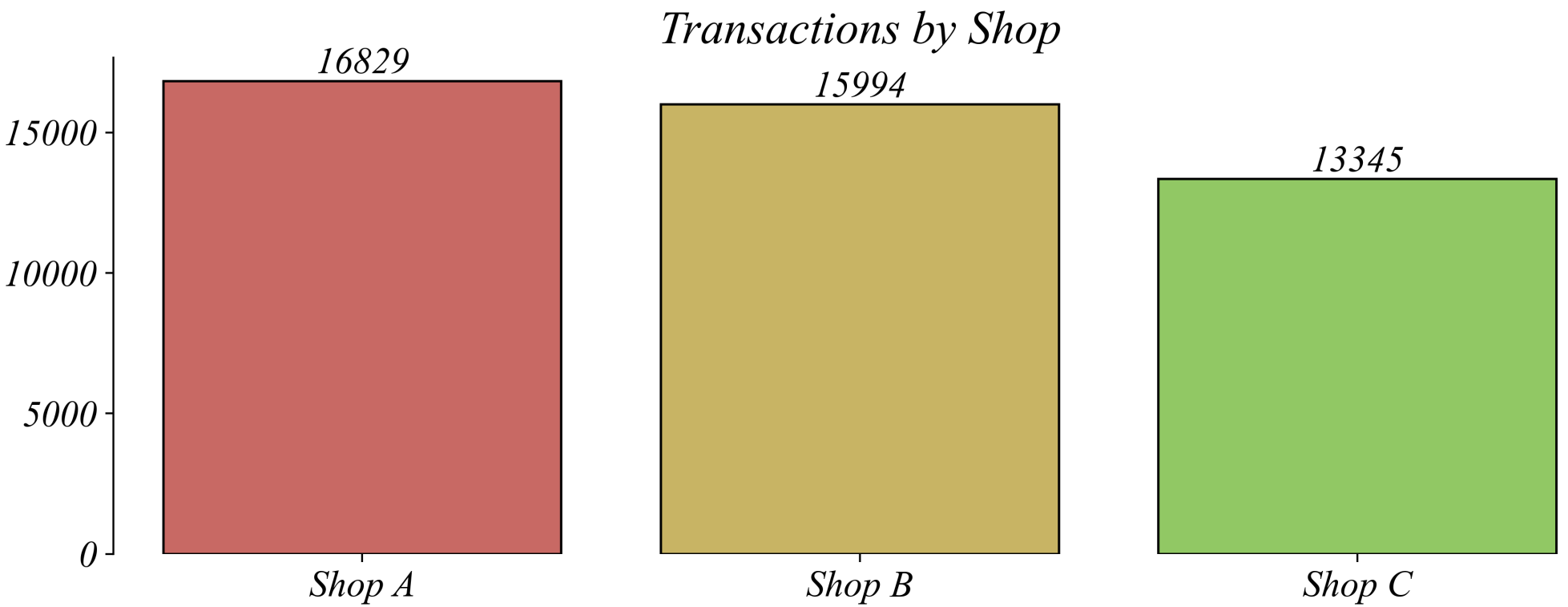
- *You manage three coffee shops and are considering where to hire a new barista.*
- *You have a dataset containing information about the transactions taking place at all three coffee shops throughout the day.*
- *Lets consider how to use this data to inform our decision.*

Hiring a Barista

Q. Which coffee shop is the busiest?

Hiring a Barista: Bar Graphs Compare Shops

Q. Which coffee shop is the busiest?



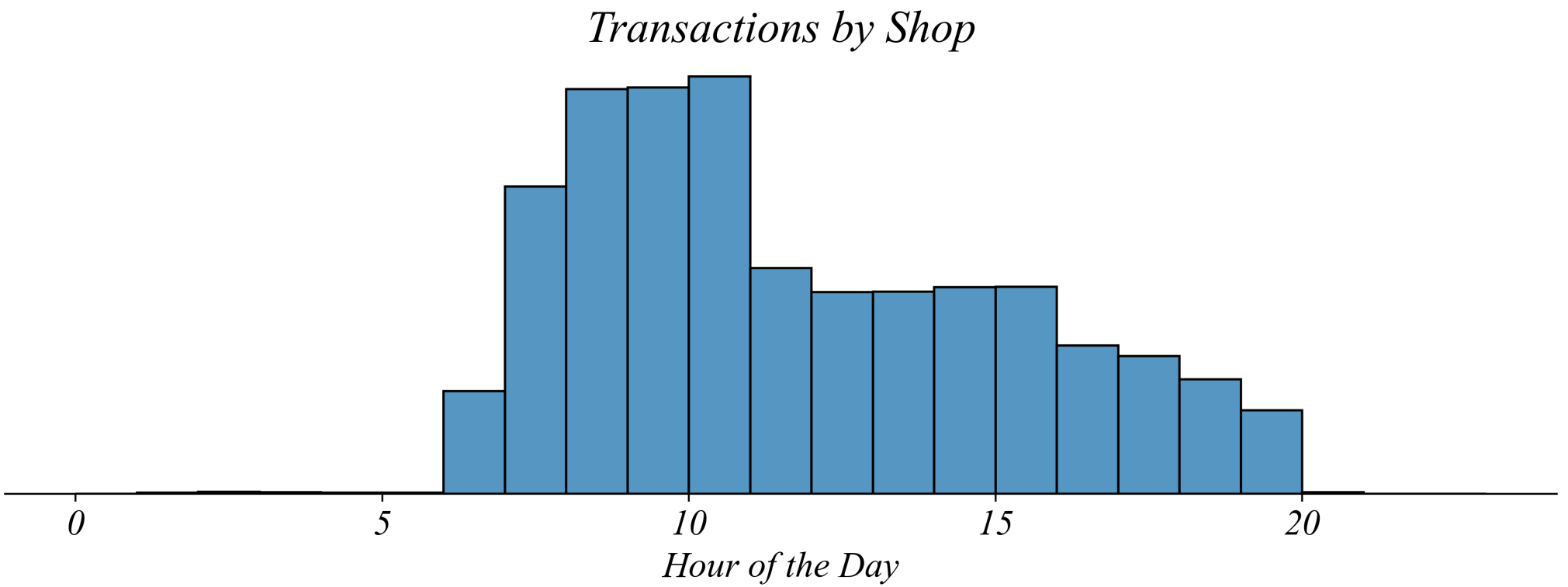
> a bar chart makes it easy to compare shops' busyness

Hiring a Barista

Q. What time of day is the busiest?

Hiring a Barista: Histograms Can Compare Times

Q. What time of day is the busiest?



> a histogram makes it easy to compare transactions by time of day

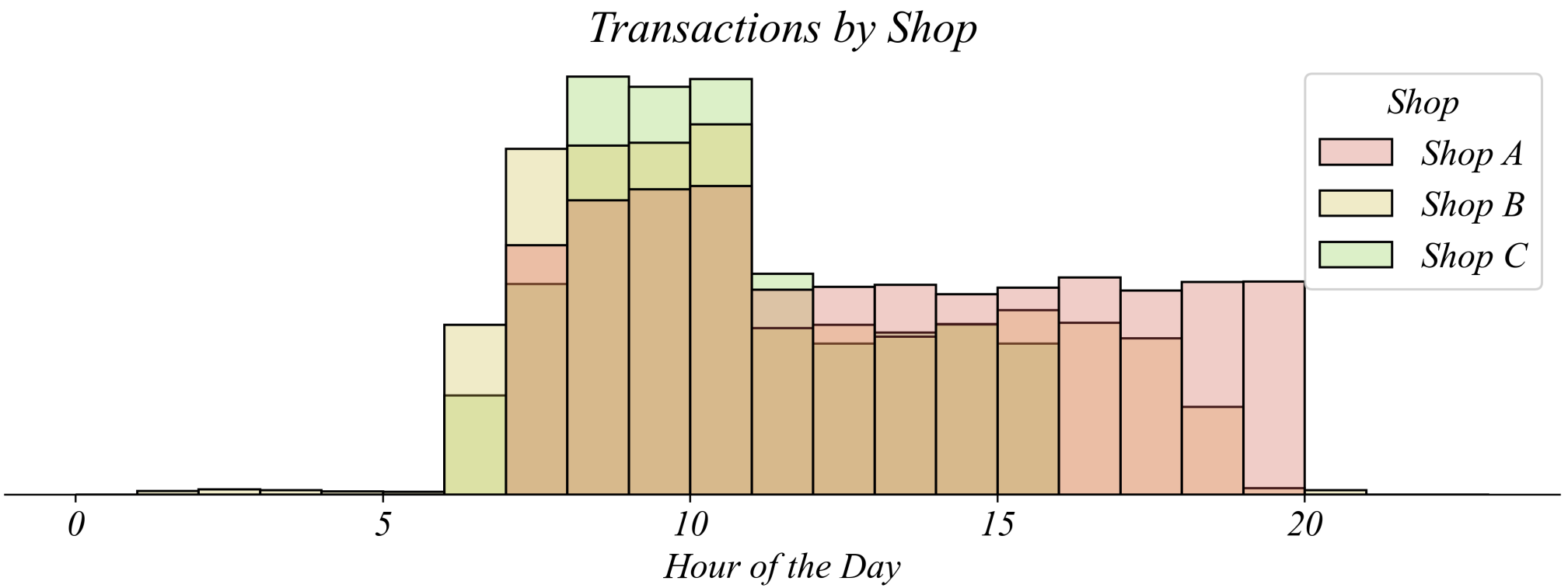
> does this mean the morning shift at Shop A is the busiest?

Hiring a Barista

Q. Which shift is the busiest?

Hiring a Barista: Transactions by Shop

Q. Which shift is the busiest?

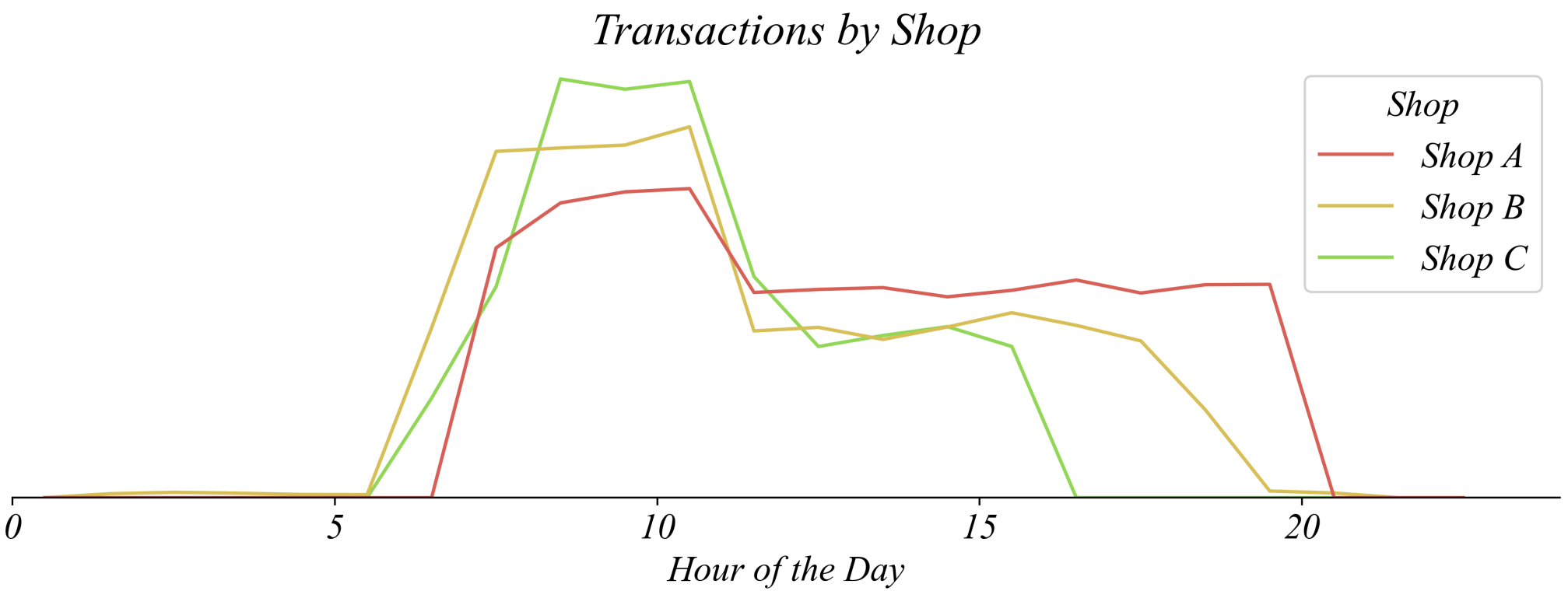


> *an overlaid histogram can show all three groups*

> *does this show the data clearly?*

Hiring a Barista: Transactions by Shop

Q. Which shift is the busiest?



> *instead, lets use a line graph*

Part 1.4 | Panel Data Using Line Graphs

Summary

- *Categorical variables and continuous variables can give us different views of the same data.*
- *We can visualize both views one the same graph.*
- *Line graphs help simplify the visualization of multiple categories.*

Exercise 1.4 | Coffee Shop Transactions

Use Coffee_Sales_Receips.csv to help inform where to hire a barista.

```
# Load Dataset
sales = pd.read_csv(file_path + 'Coffee_Sales_Reciepts.csv')
sales.head()
```

	Hours	Shop
0	12	Shop A
1	15	Shop A
2	14	Shop A
3	16	Shop A
4	19	Shop A

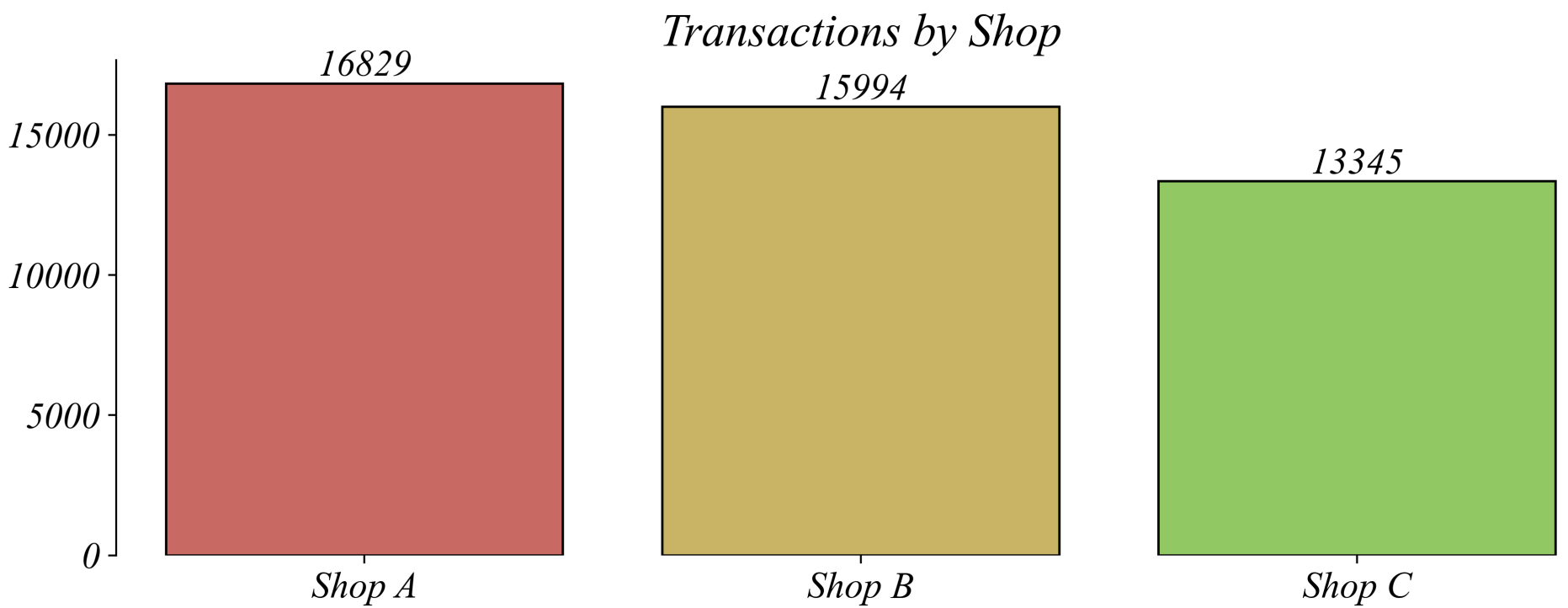
> *you'll see a few more columns in your dataset*

> *this is **Long-Format Panel Data**: transactions are all in the same column*

Exercise 1.4 | Bar Chart

Use Coffee_Sales_Receips.csv to help inform where to hire a barista.

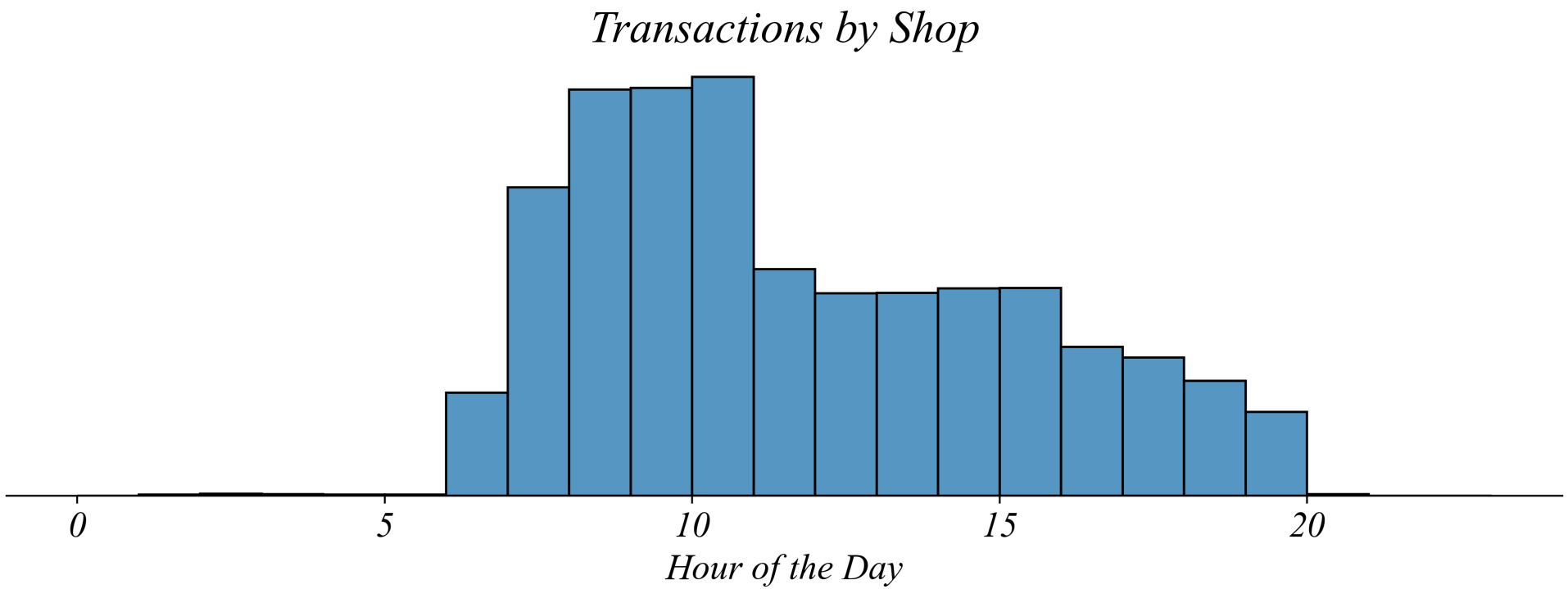
```
# Bar graph
sns.countplot(sales, x='Shop', hue='Shop')
```



Exercise 1.4 | Histogram

Use Coffee_Sales_Receips.csv to help inform where to hire a barista.

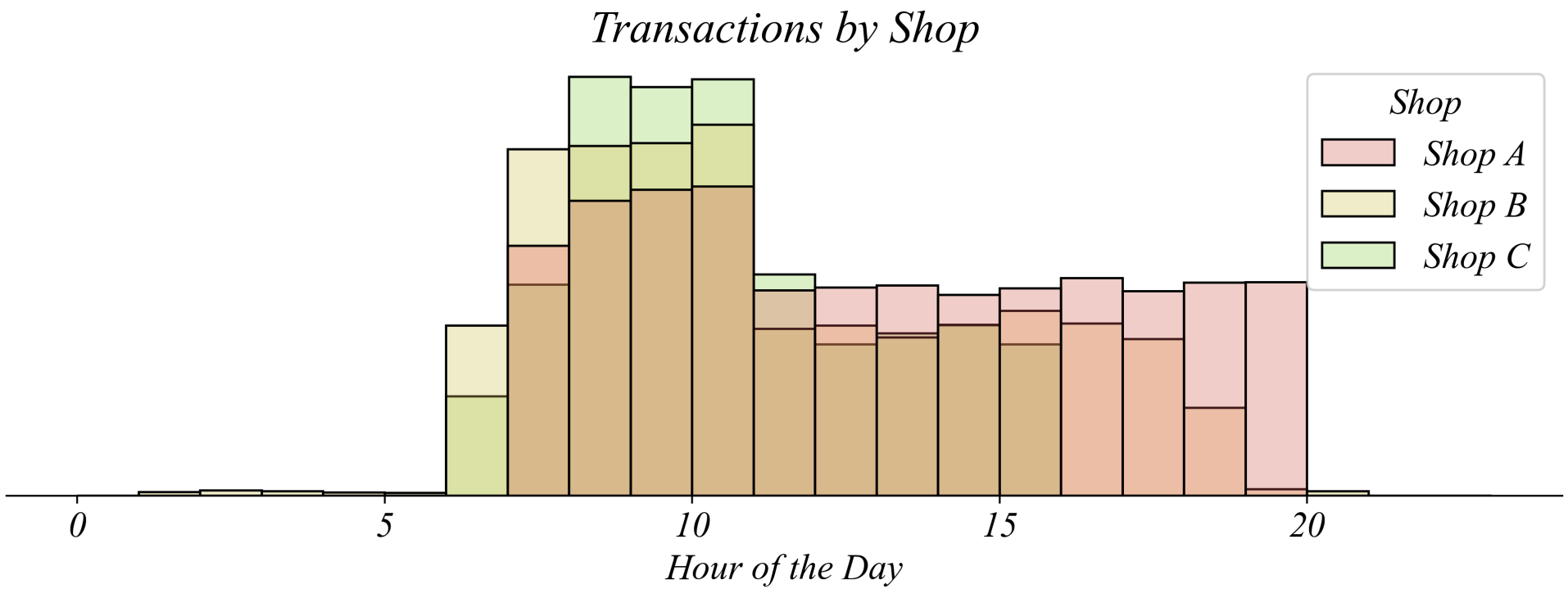
```
# Create a histogram
sns.histplot(sales, x='Hours', bins=range(0,24,1))
```



Exercise 1.4 | Multi-Histogram

Use Coffee_Sales_Receips.csv to help inform where to hire a barista.

```
1 # Create a multi-histogram
2 sns.histplot(sales, x='Hours', hue='Shop', bins=range(0,24,1))
```



Exercise 1.4 | Count Hourly by Shop

Q. Which shift is the busiest?

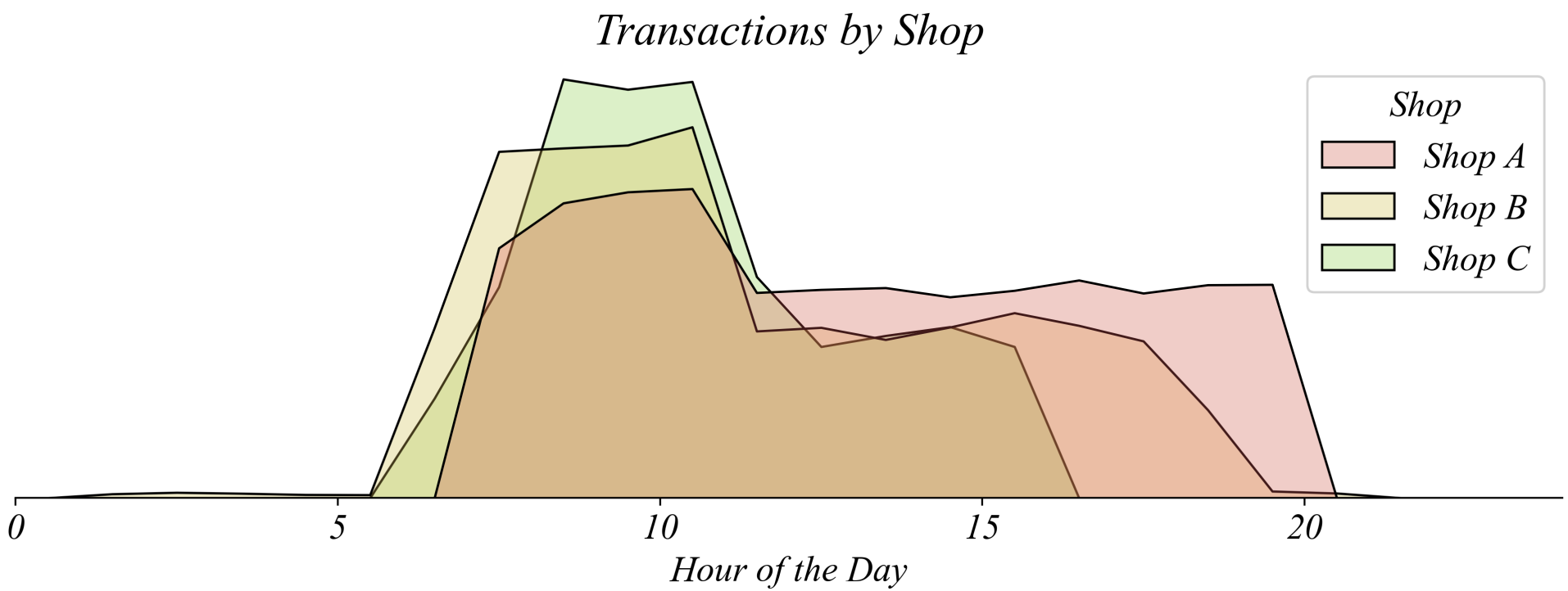
```
# Create hourly counts by shop
hourly_counts = sales.groupby(['Shop', 'Hours']).size().reset_index(name='Count')
```

	Shop	Hours	Count
0	Shop A	7	1383
1	Shop A	8	1632
2	Shop A	9	1693
3	Shop A	10	1711
4	Shop A	11	1136

Exercise 1.4 | Multiple Line Graph

Q. Which shift is the busiest?

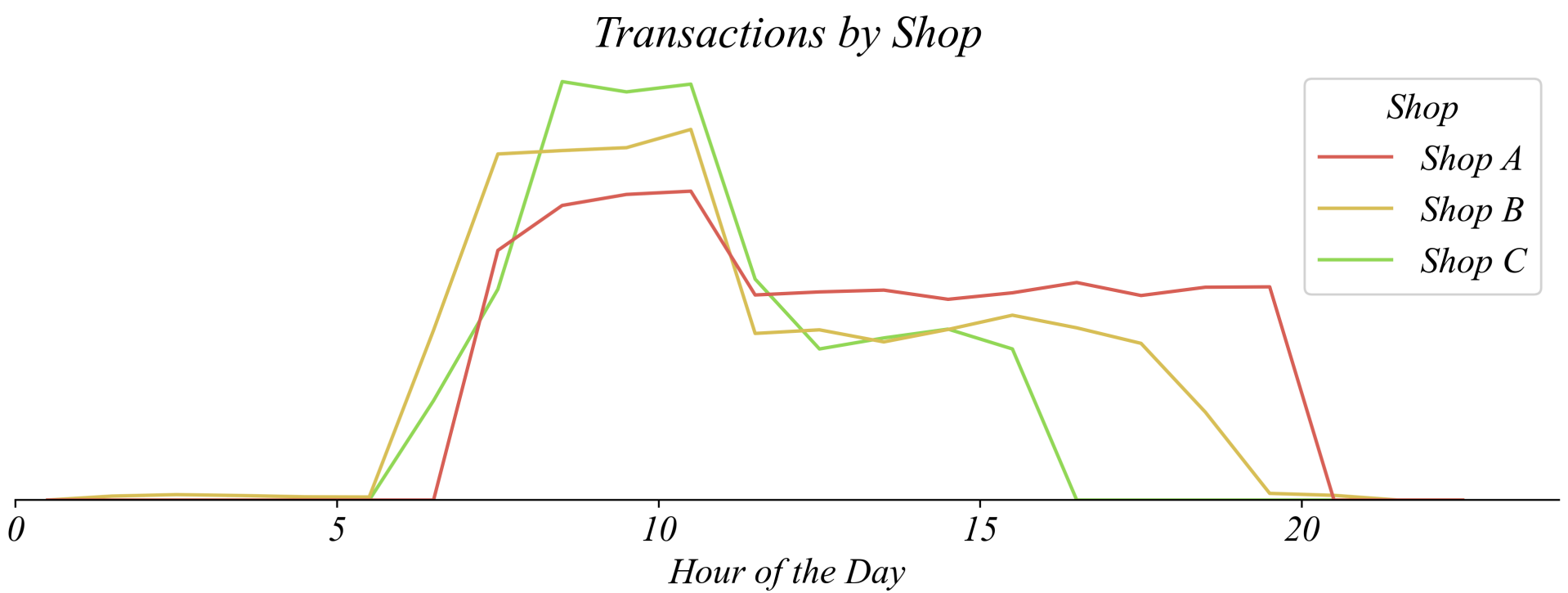
```
# Multiple-Line Graph
sns.histplot(sales, x='Hours', hue='Shop', bins=range(0,24,1), element='poly')
```



Exercise 1.4 | Multiple Line Graph

Q. Which shift is the busiest?

```
# Multiple-Line Graph
sns.histplot(sales, x='Hours', hue='Shop', bins=range(0,24,1), element='poly', fill=False)
```



Panel Data: Coffee Consumption Per Capita

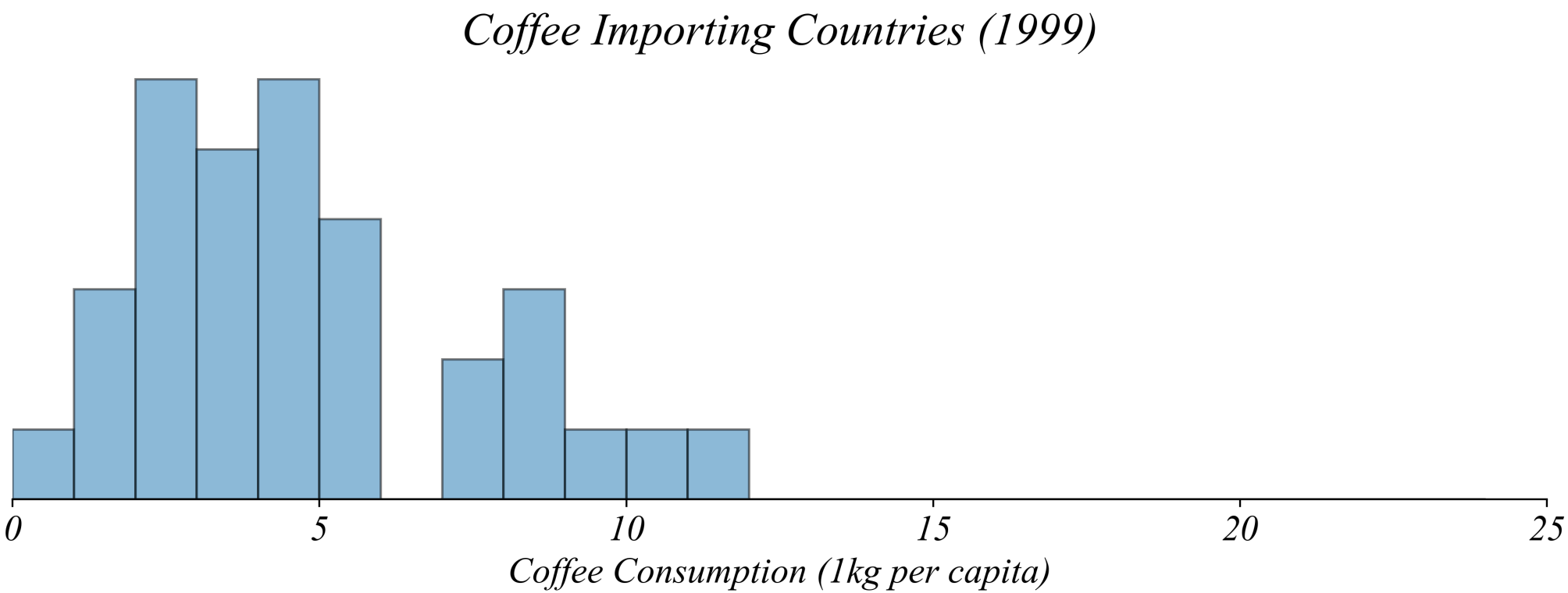
Is the world drinking more coffee?

Lets examine whether the world is drinking more coffee today than in the 1990s.

- ***Data:*** `Coffee_Per_Cap.csv`

Panel Data: Coffee Consumption Per Capita

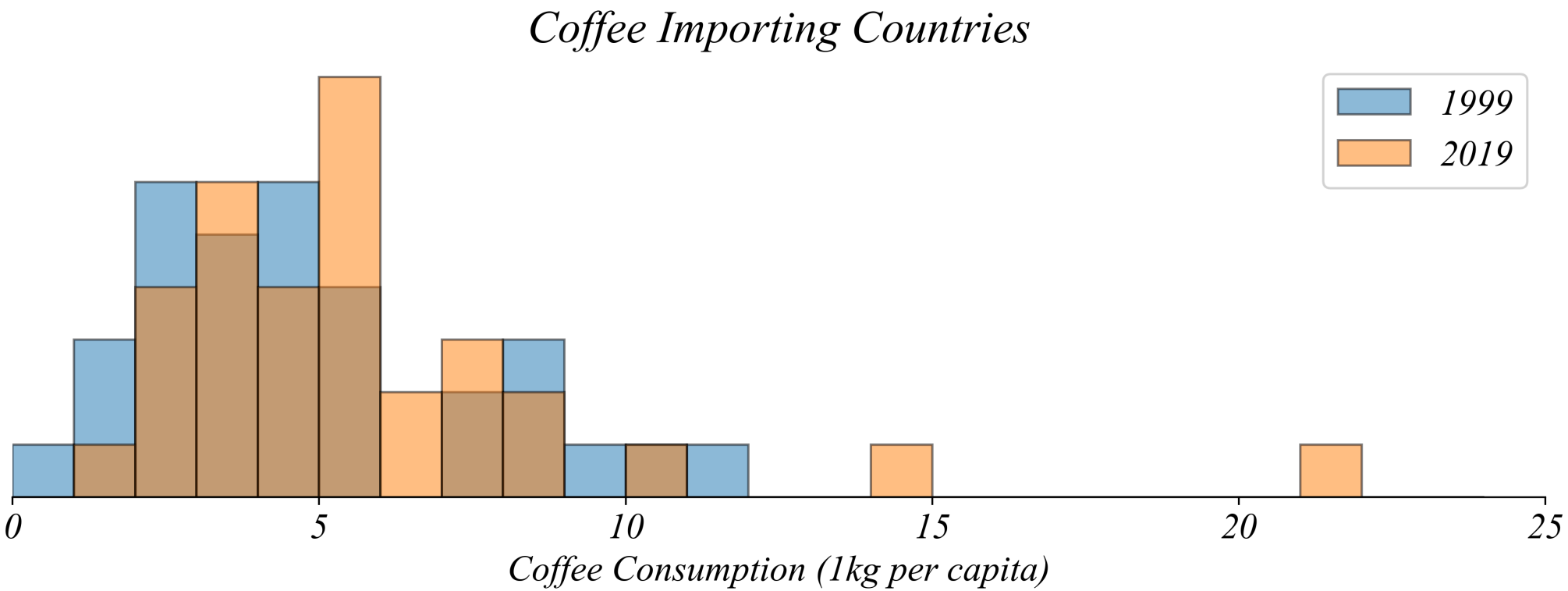
Is the world drinking more coffee?



> compared to what…?

Panel Data: Coffee Consumption Per Capita

Is the world drinking more coffee?

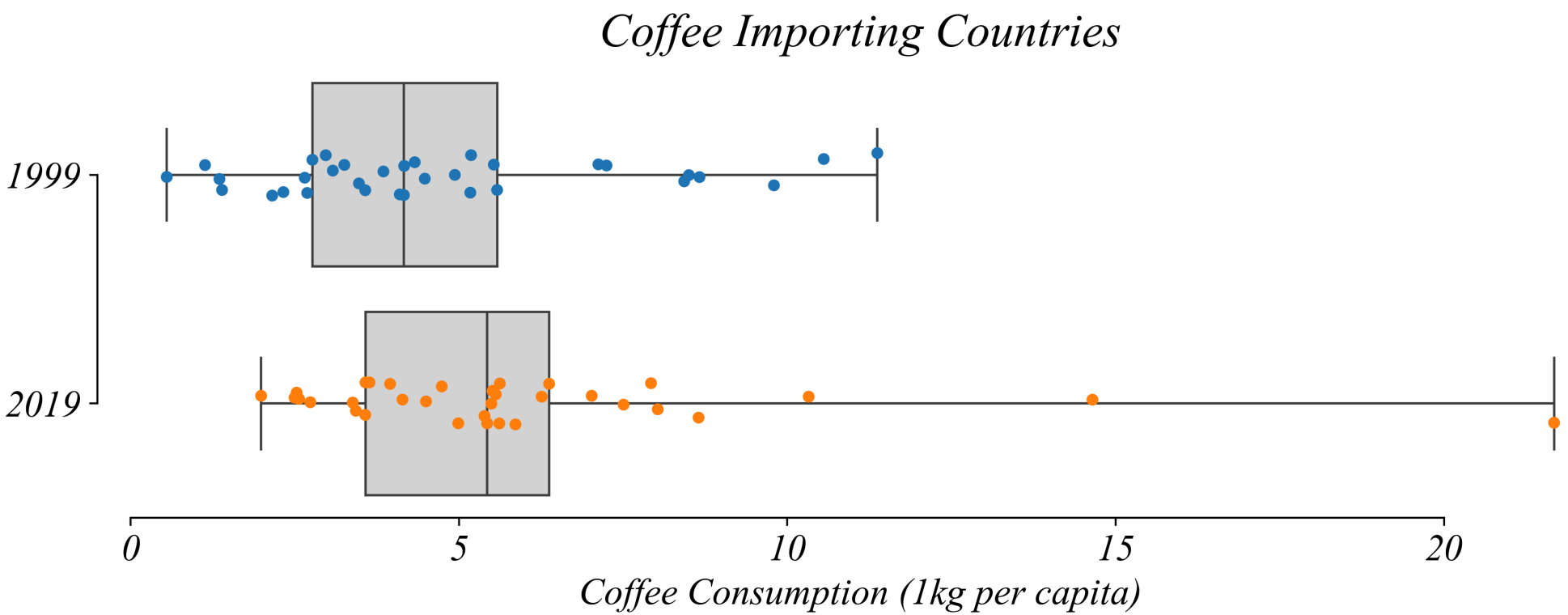


> *this is still pretty unclear: histograms aren't great for comparison*

> *lets use a multi-boxplot*

Panel Data: Multi-Boxplots

Is the world drinking more coffee?

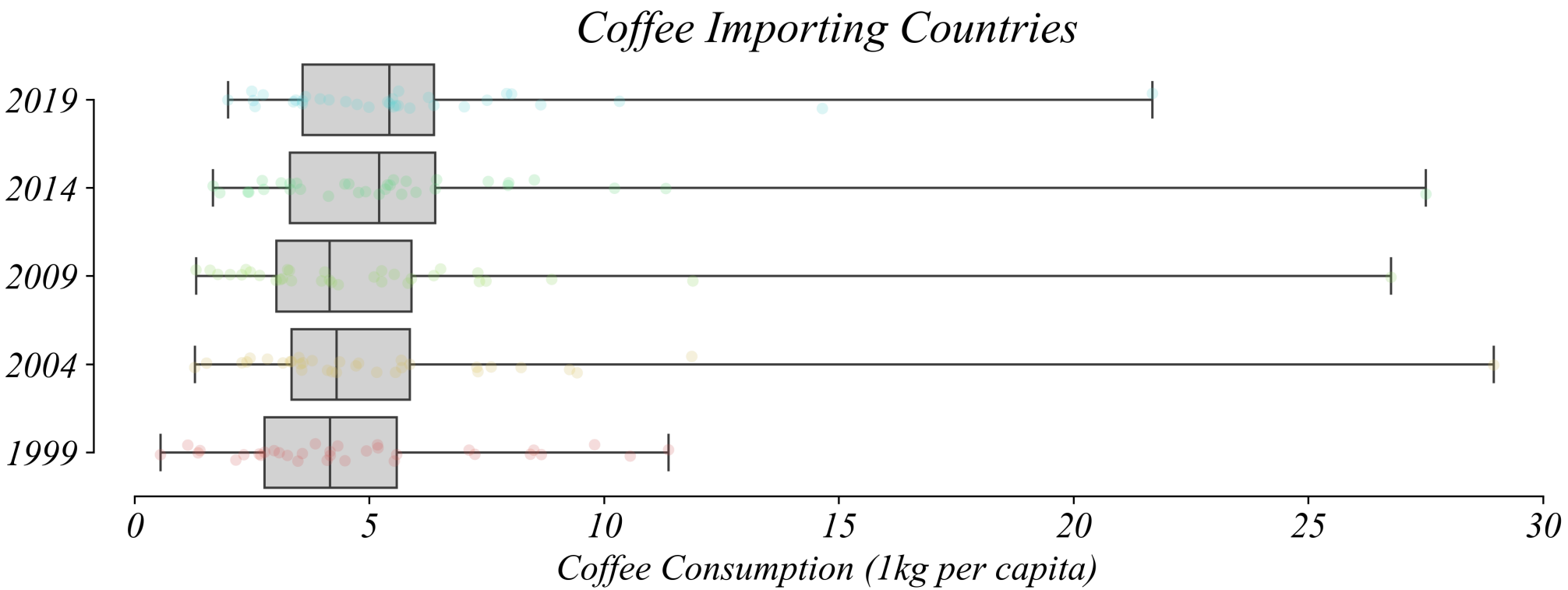


> *this is better: it looks like the distribution is shifted higher!*

> *lets examine the years in between to see how the distribution evolved*

Panel Data: Multi-Boxplots

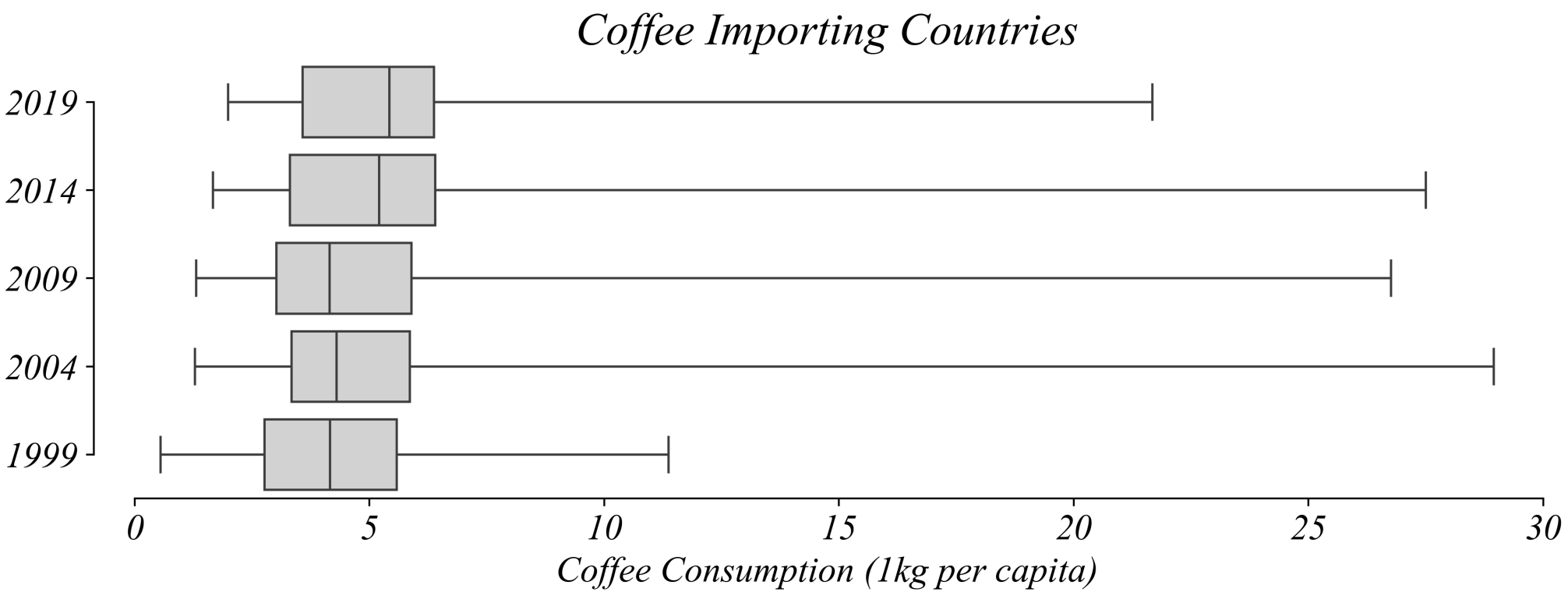
Is the world drinking more coffee?



> lets ask some smaller more focussed questions

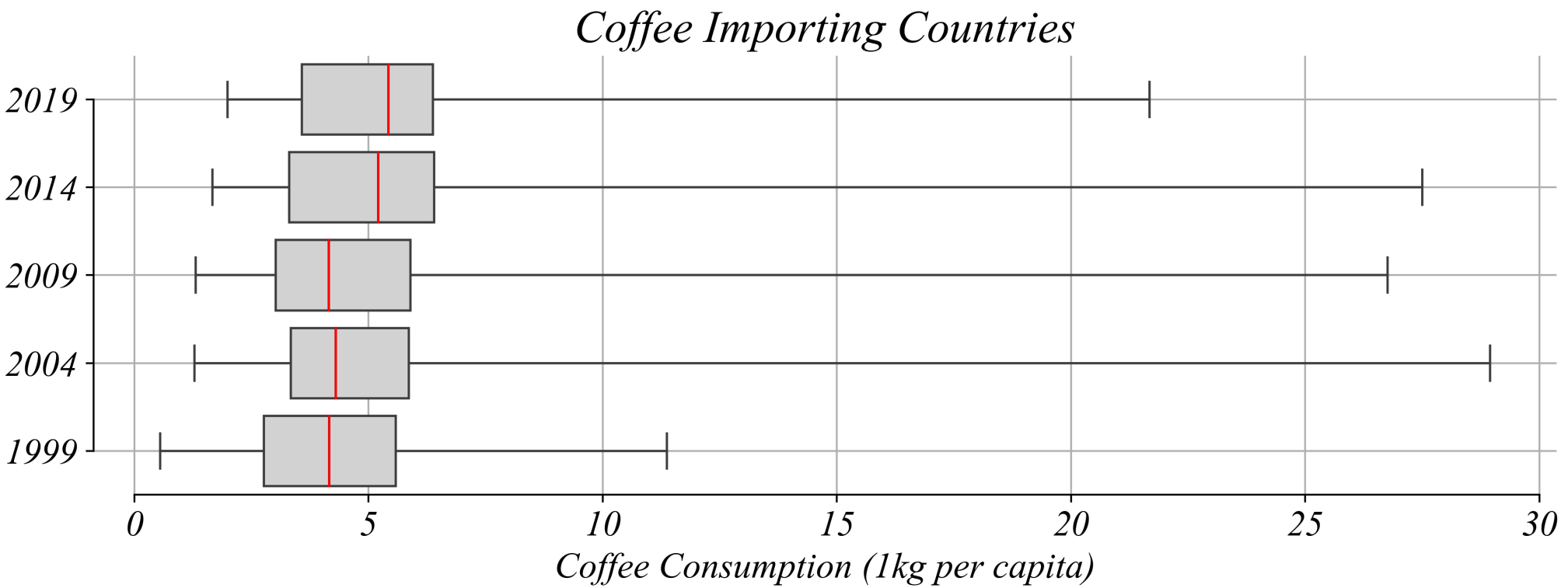
Panel Data: Multi-Boxplots

Which years show at least half consuming less than 5 kg per cap?



Panel Data: Multi-Boxplots

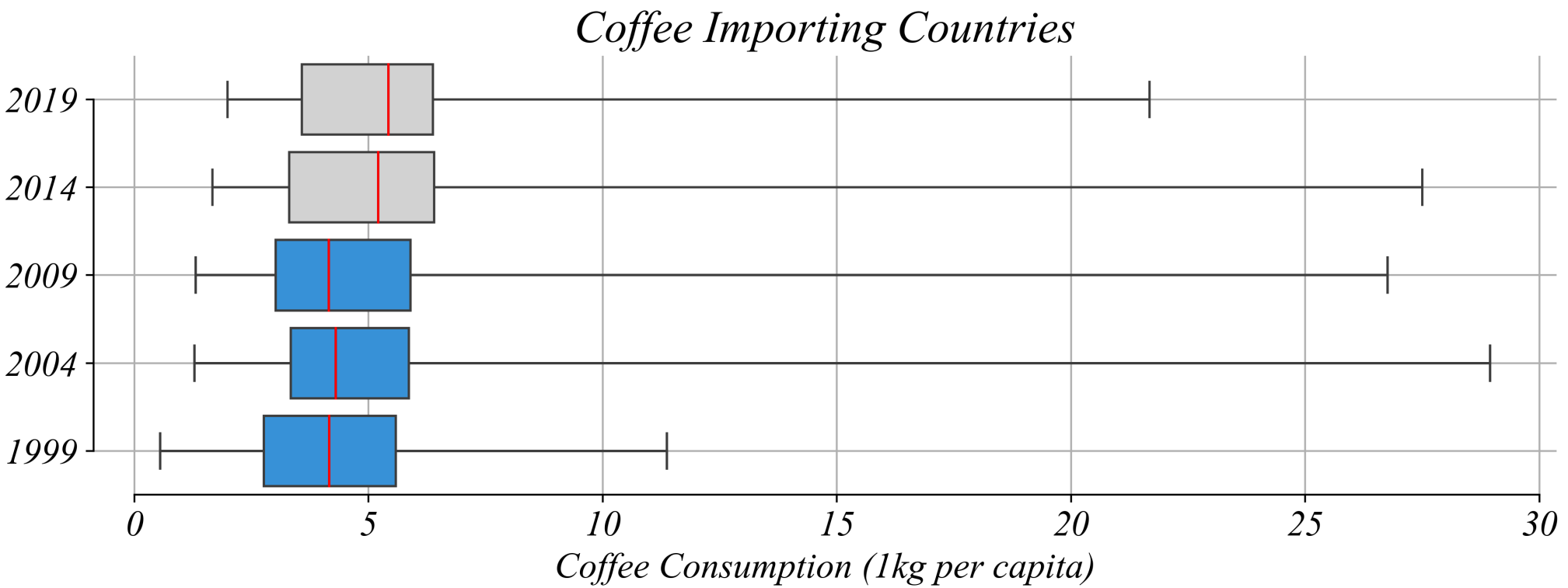
Which years show at least half consuming less than 5 kg per cap?



> *focus on the medians*

Panel Data: Multi-Boxplots

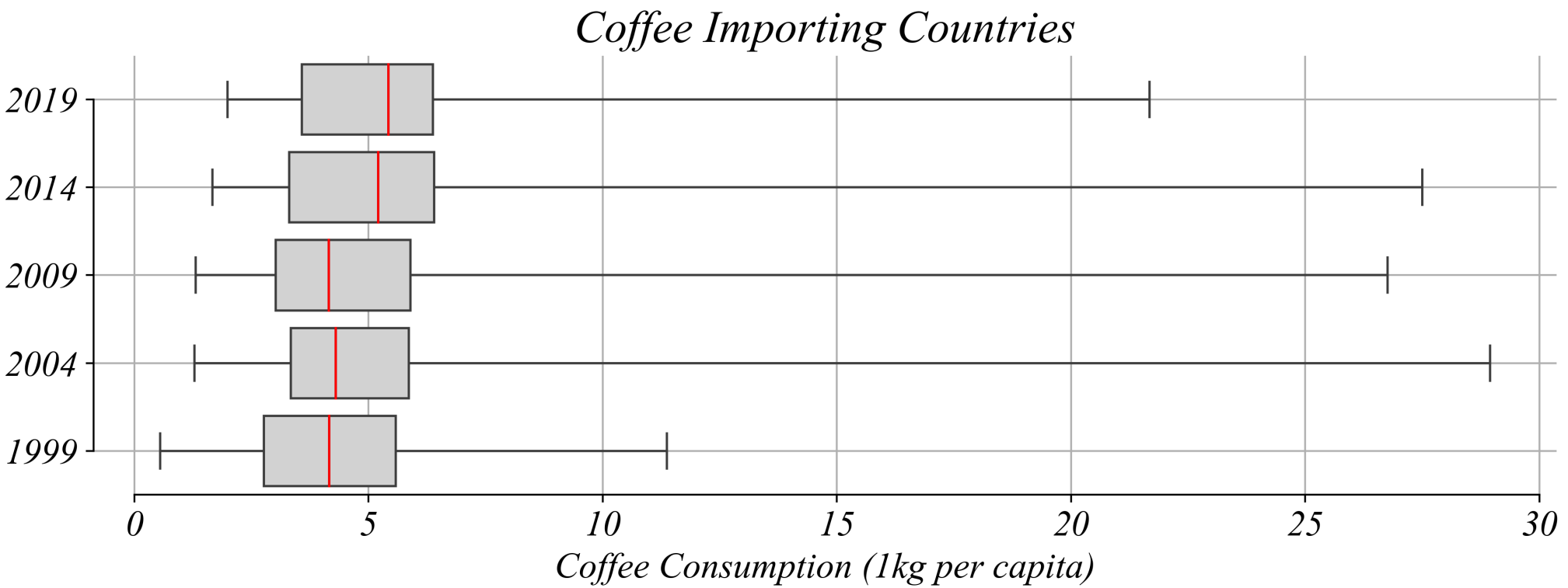
Which years show at least half consuming less than 5 kg per cap?



> *... when the median is above 5 kg per cap*

Panel Data: Multi-Boxplots

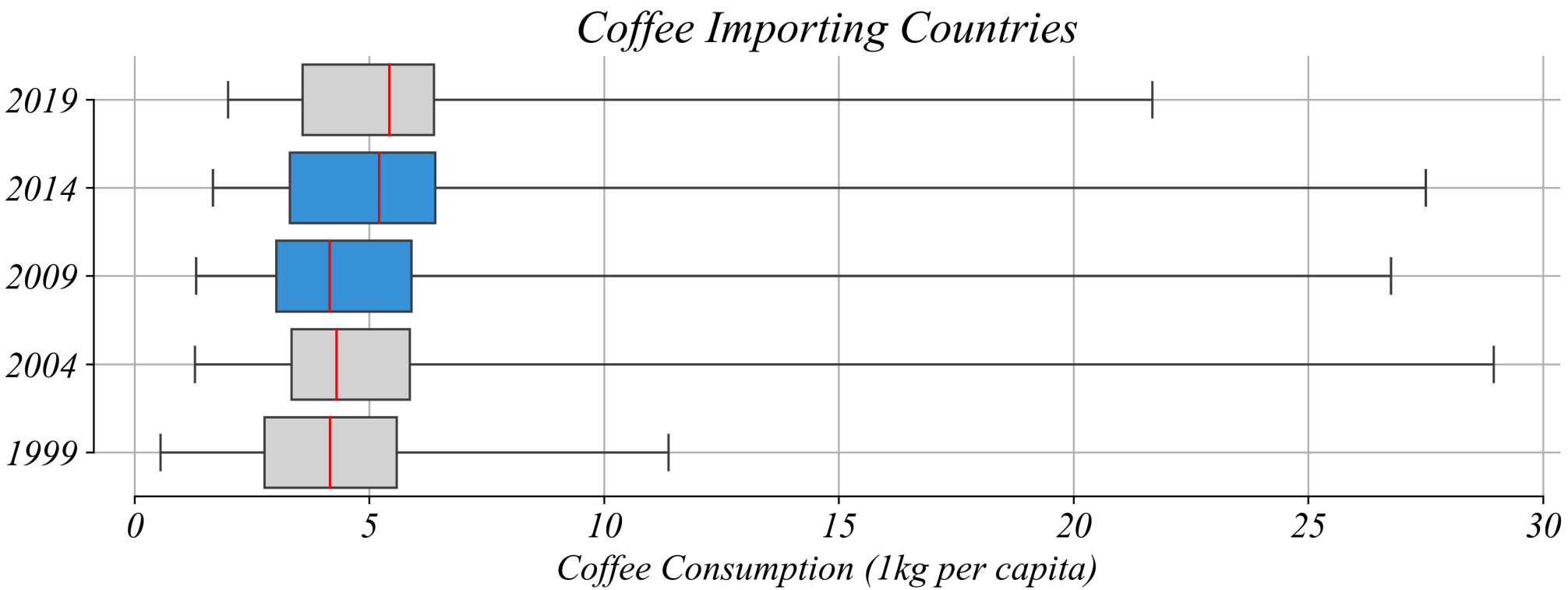
Which years saw the largest jump in the median?



> *... a little difficult to see*

Panel Data: Multi-Boxplots

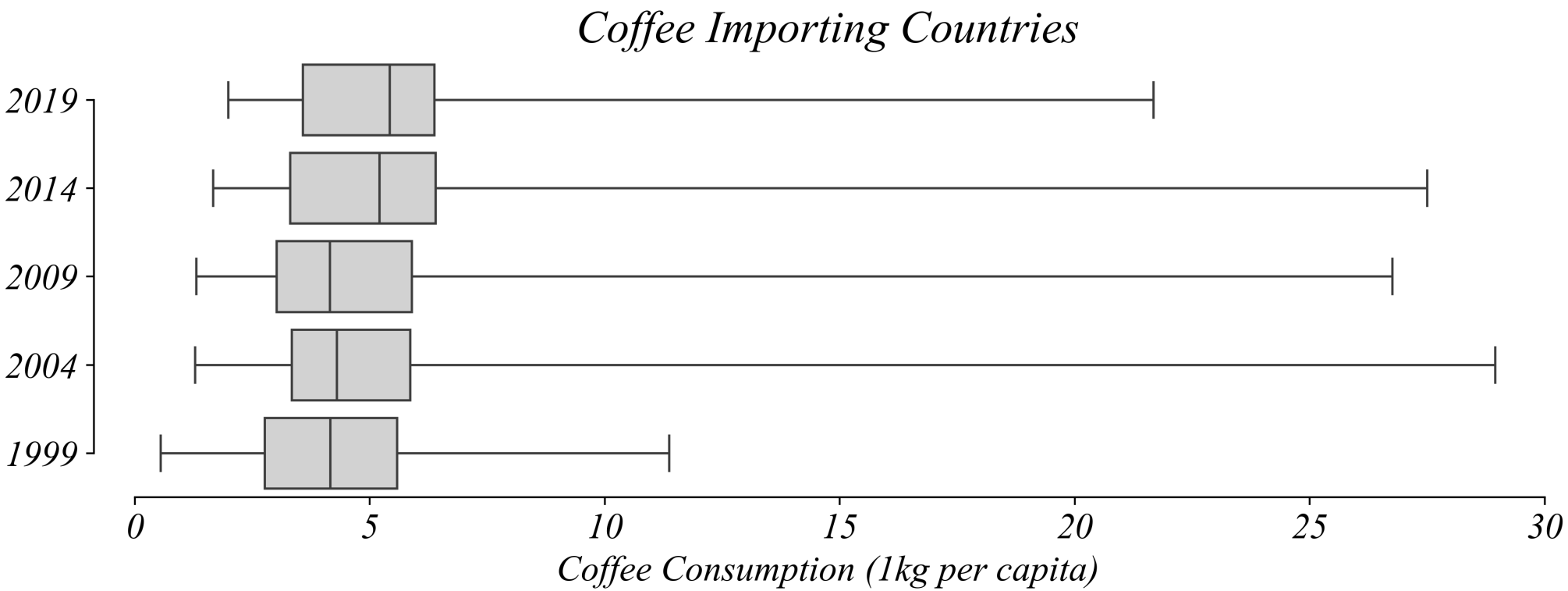
Which years saw the largest jump in the median?



> *... a little difficult to see*

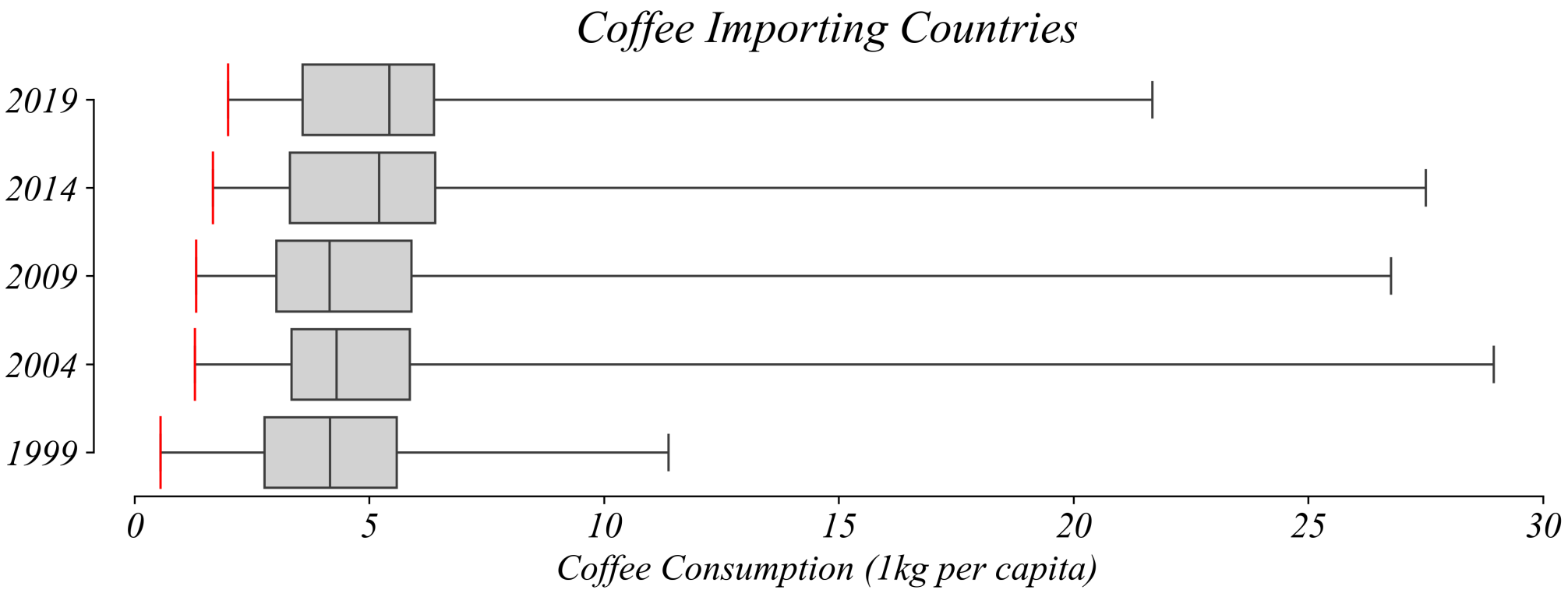
Panel Data: Multi-Boxplots

Is the country with the lowest consumption consuming more today?



Panel Data: Multi-Boxplots

Is the country with the lowest consumption consuming more today?

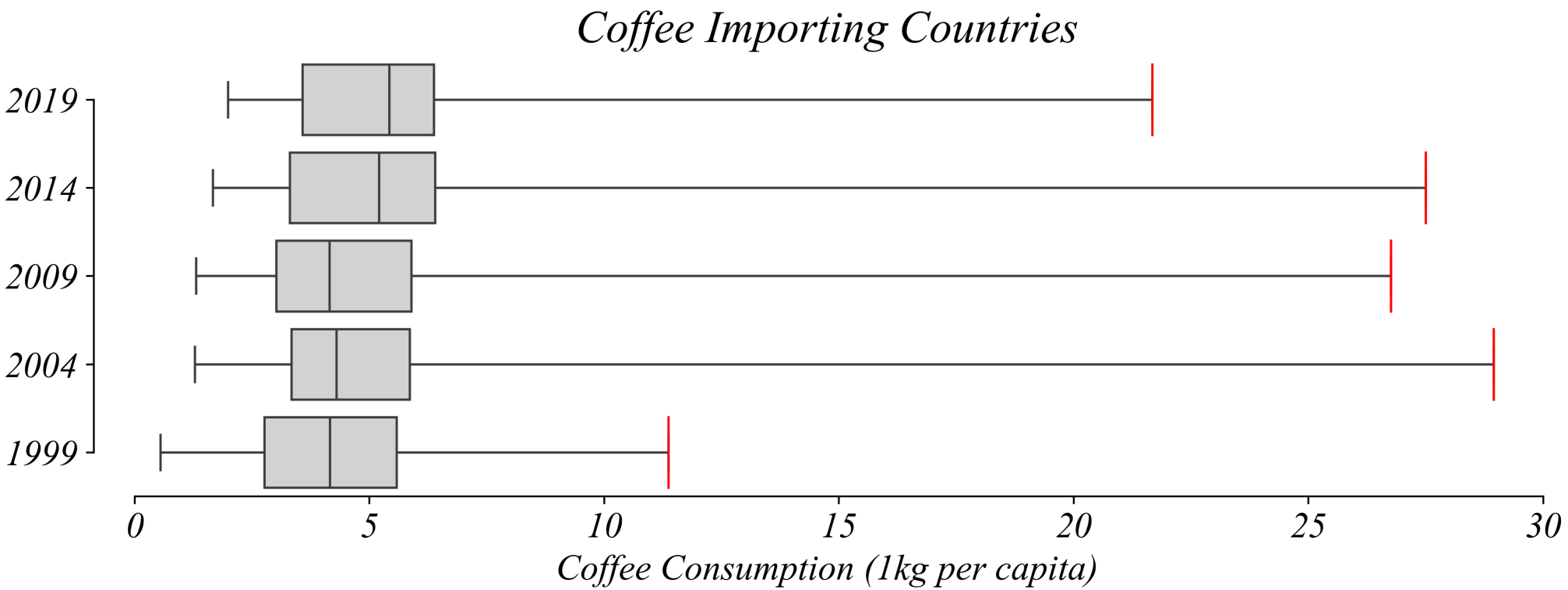


> *focus on the minimums*

> *yes!*

Panel Data: Multi-Boxplots

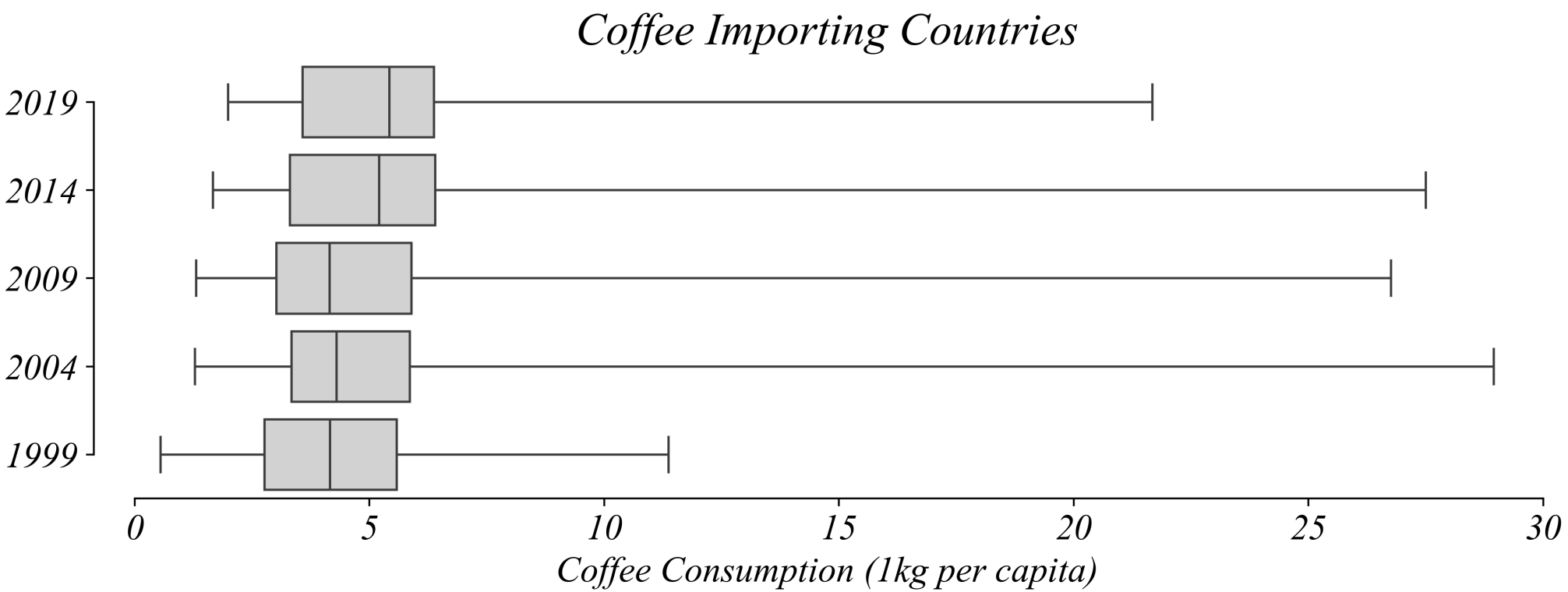
What patterns do we observe about the maximums?



> *same with the maximums*

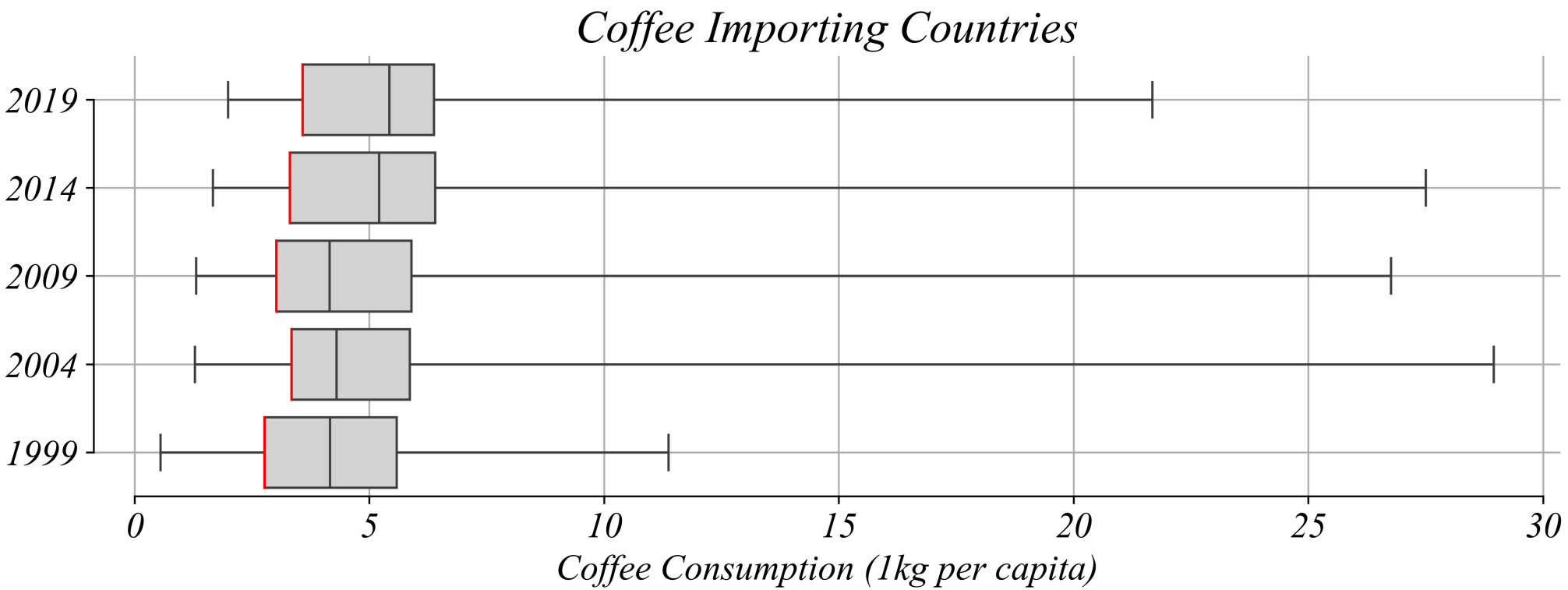
Panel Data: Multi-Boxplots

Which years did more than 25% consume less than 5 kg?



Panel Data: Multi-Boxplots

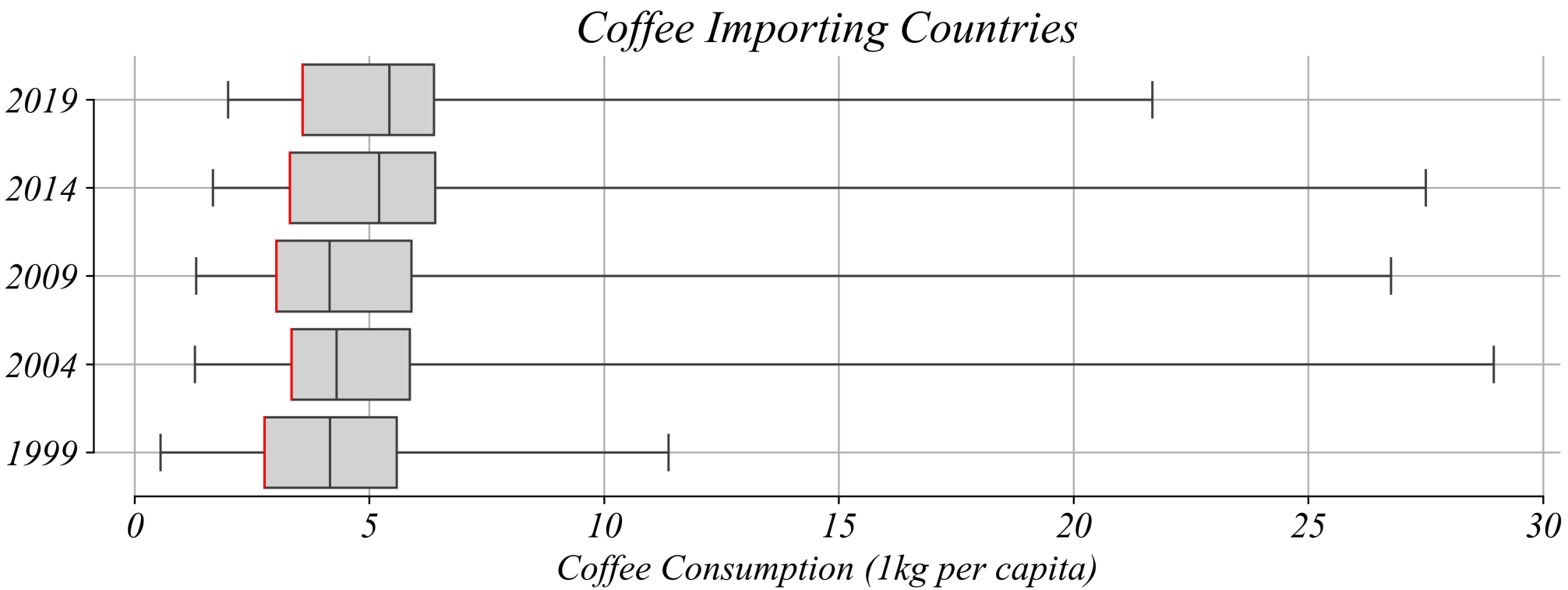
Which years did more than 25% consume less than 5 kg?



> *look at the 25%*

Panel Data: Multi-Boxplots

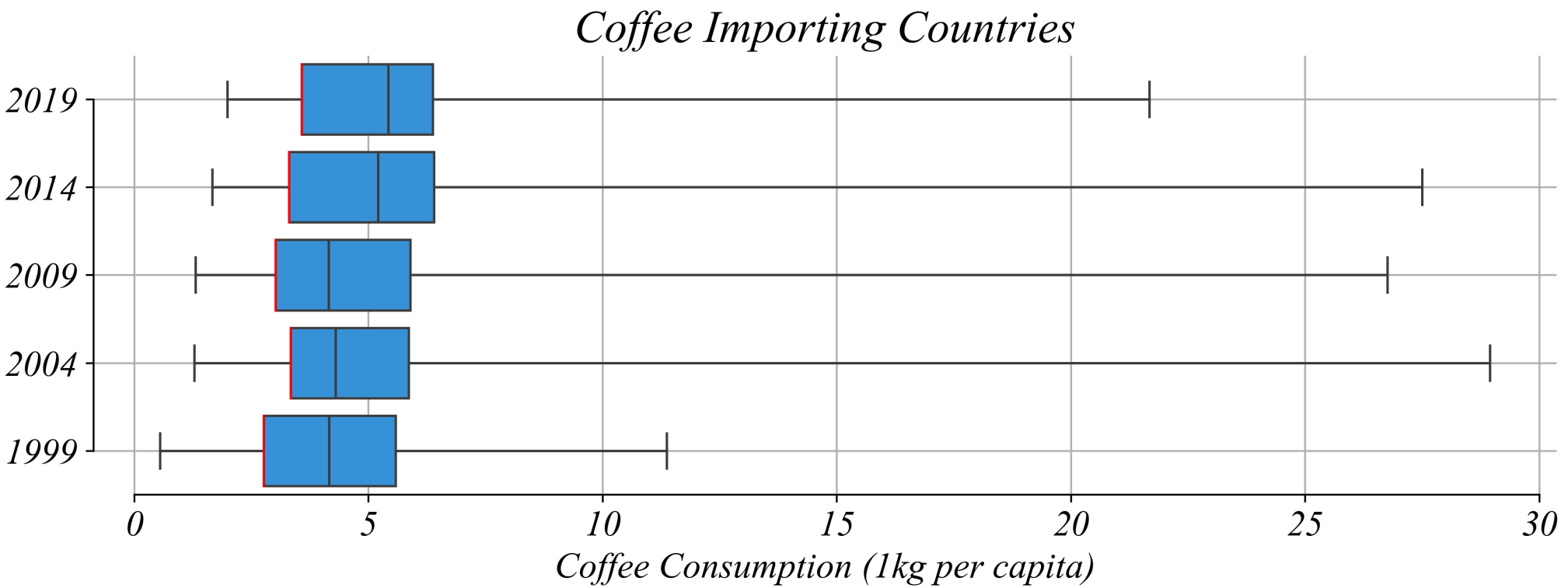
Which years did more than 25% consume less than 5 kg?



> look at the 25% and compare it to 5 kg per cap

Panel Data: Multi-Boxplots

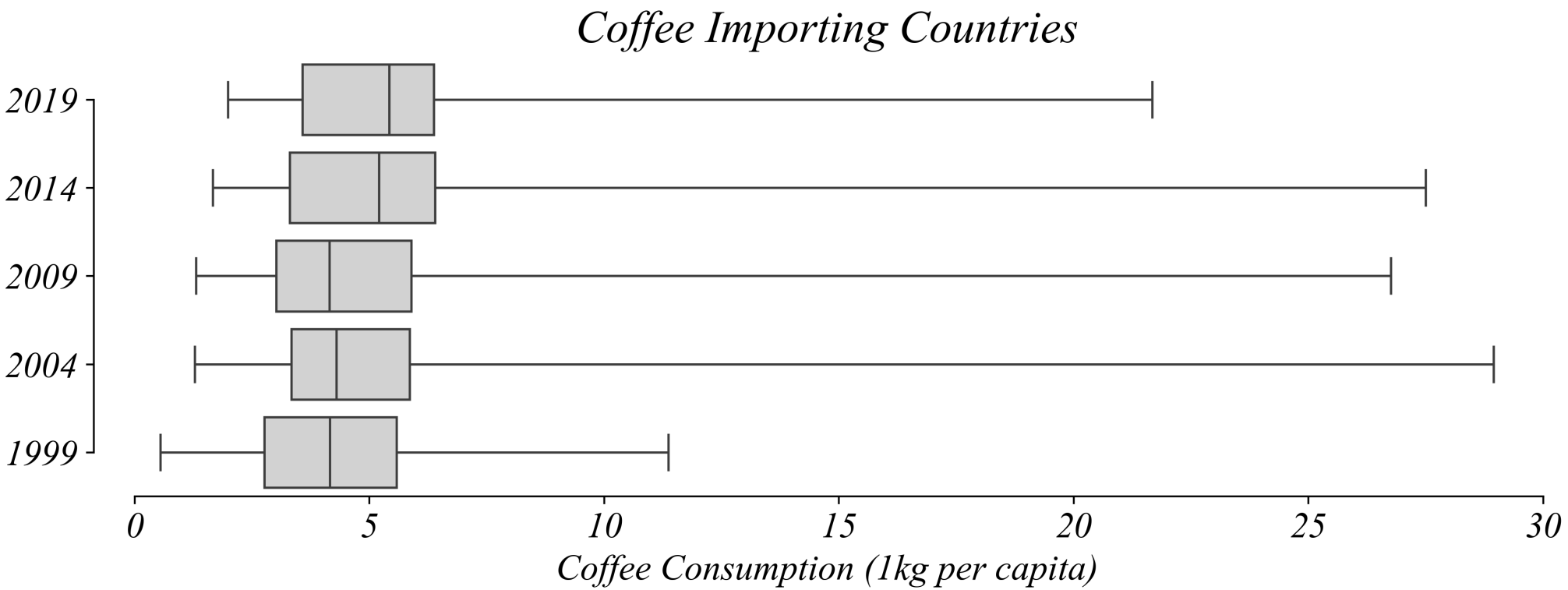
Which years did more than 25% consume less than 5 kg?



> all of them

Panel Data: Multi-Boxplots

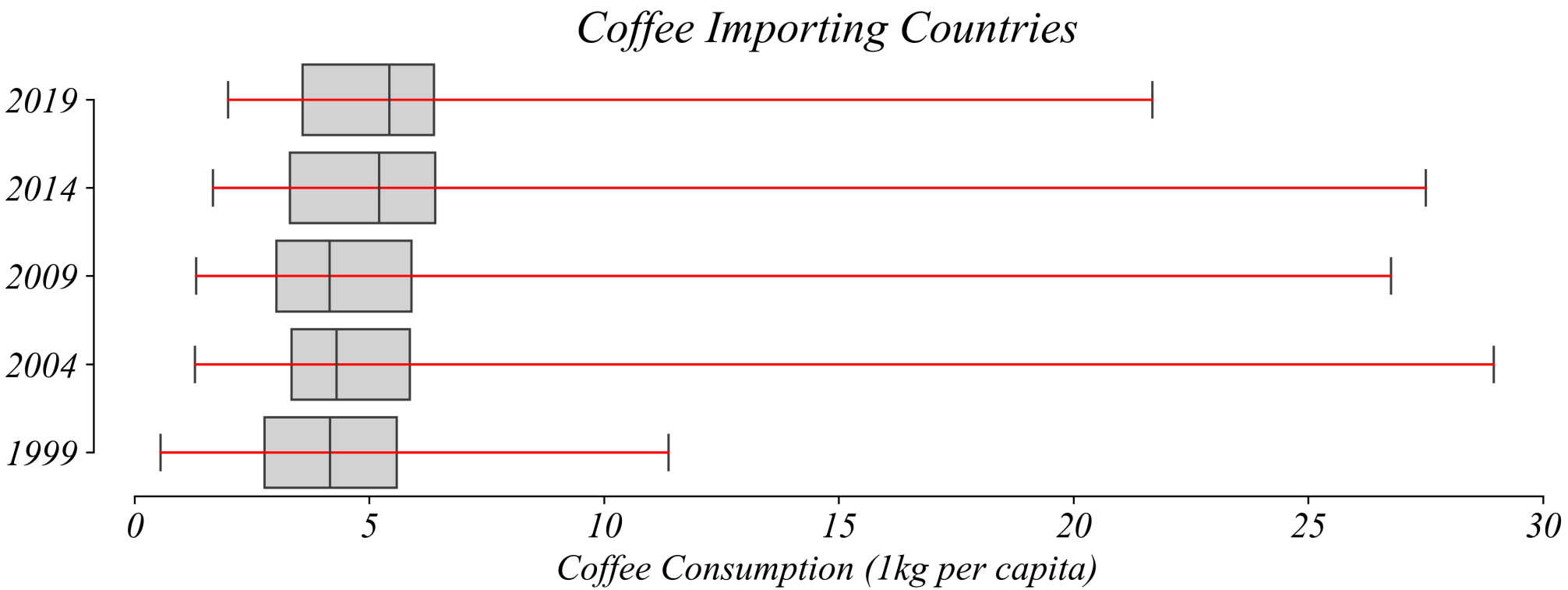
Which year saw the greatest difference between any two countries?



> look at the range

Panel Data: Multi-Boxplots

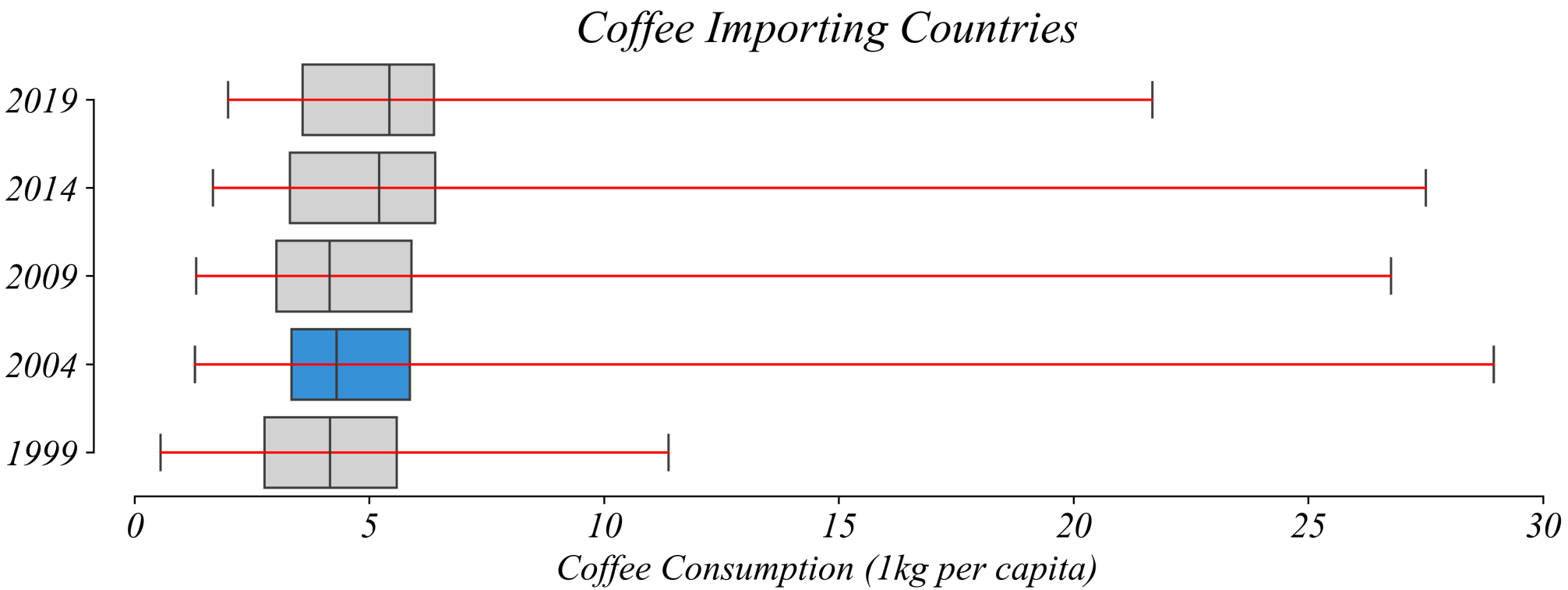
Which year saw the greatest difference between any two countries?



> *look at the range*

Panel Data: Multi-Boxplots

Which year saw the greatest difference between any two countries?



> *look at the range and select the largest*

Exercise 1.4 | Multi-Boxplots

Is the world drinking more coffee?

We're going to use a set of boxplots to visually compare across years the distributions of coffee consumption per capital among coffee importing countries.

- ***Data:*** `Coffee_Per_Cap.csv`

	Code	1999	2009	2019
0	AUT	8.430589	6.371562	7.925747
2	BGR	2.652661	3.296419	3.638313
3	HRV	4.480790	5.100831	5.623266
4	CYP	3.477888	4.050500	5.615070
5	CZE	3.255587	3.016104	4.739563

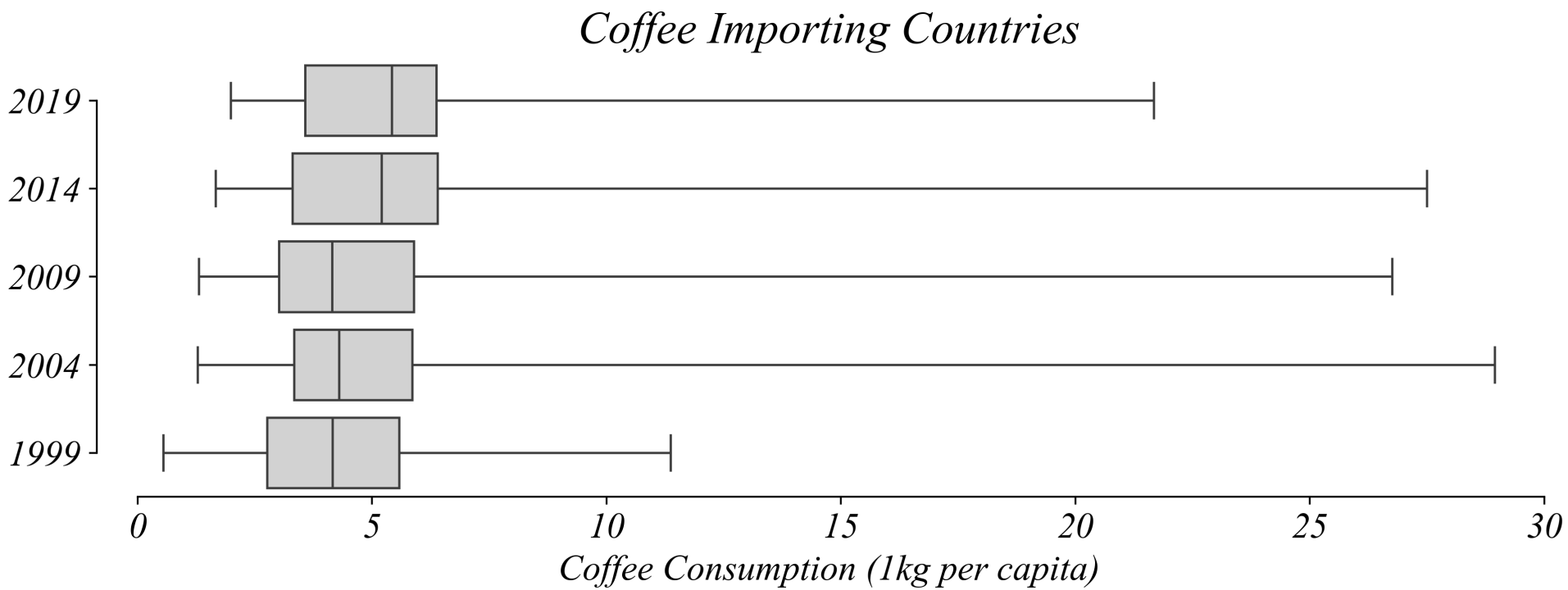
> *this is **Wide-Format Panel Data**: each year is in a separate column*

Exercise 1.4 | Multi-Boxplots

Is the world drinking more coffee?

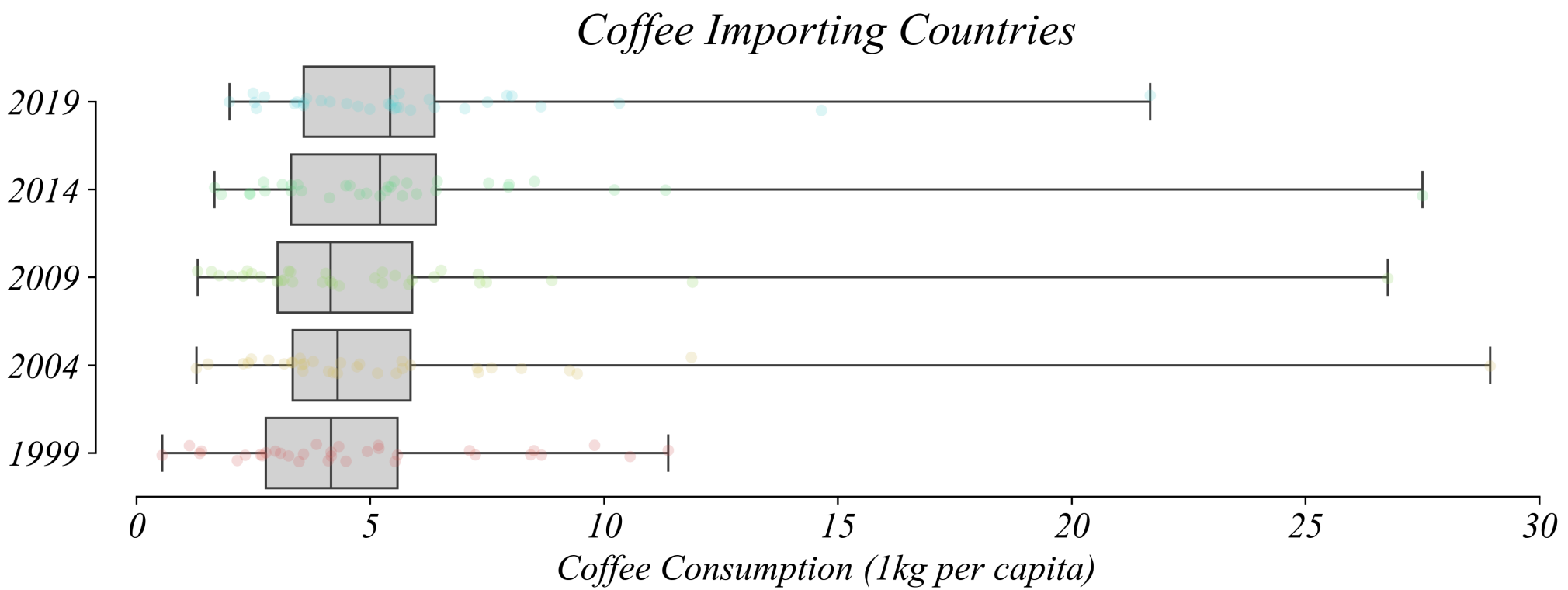
With wide-format panel data seaborn looks a little different.

```
# Wide Format Multi-Boxplot
sns.boxplot(percap[['1999','2004','2009','2014','2019']], orient='h', whis=(0, 100))
```



Panel Data: Multi-Boxplots

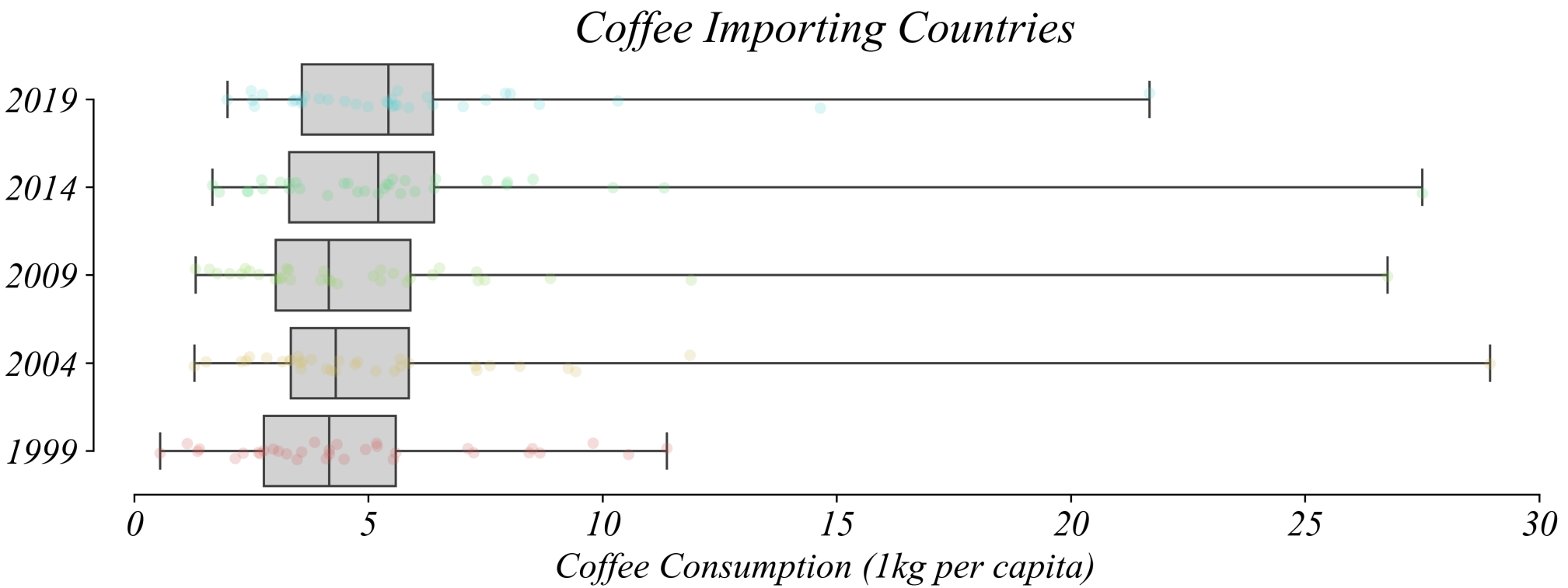
In which year did most countries increase their coffee consumption?



> not visible in the figure!

Panel Data: Relationships Between Years

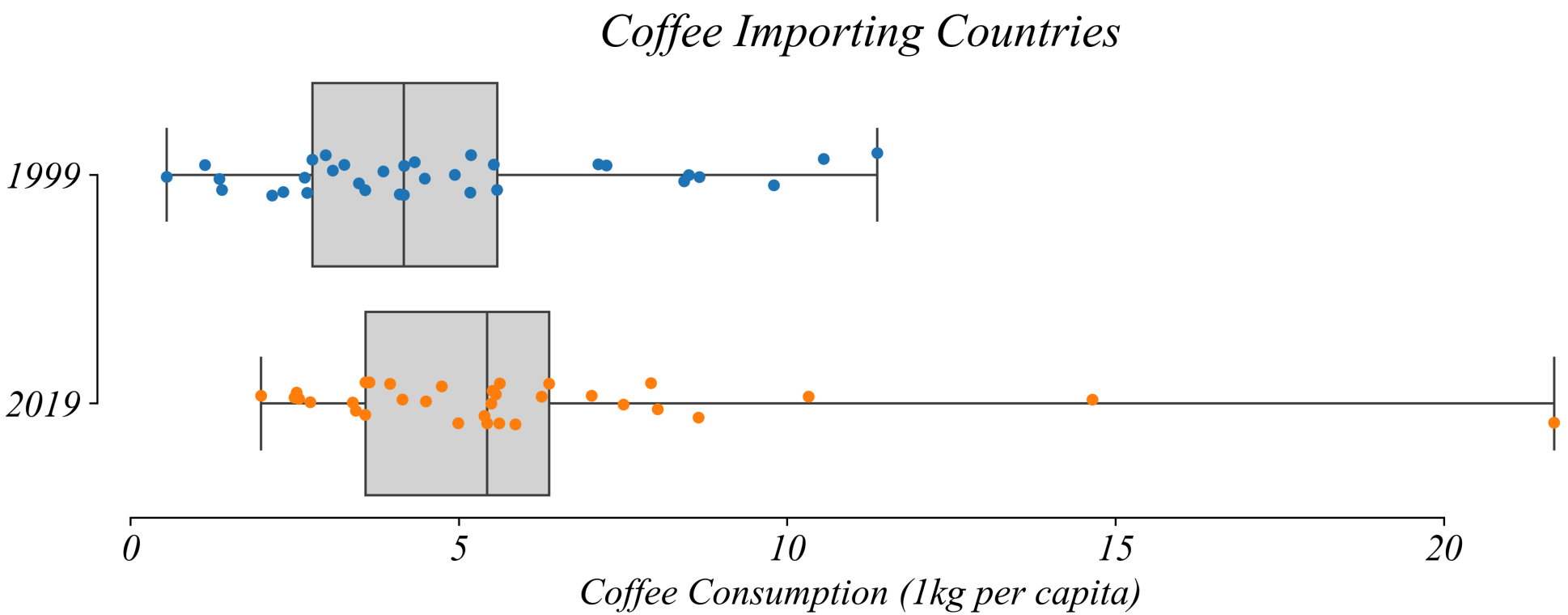
How many countries increased their coffee consumption between 1999 and 2019?



> *also not visible with this figure!*

Panel Data: Relationships Between Years

How many countries increased their coffee consumption between 1999 and 2019?



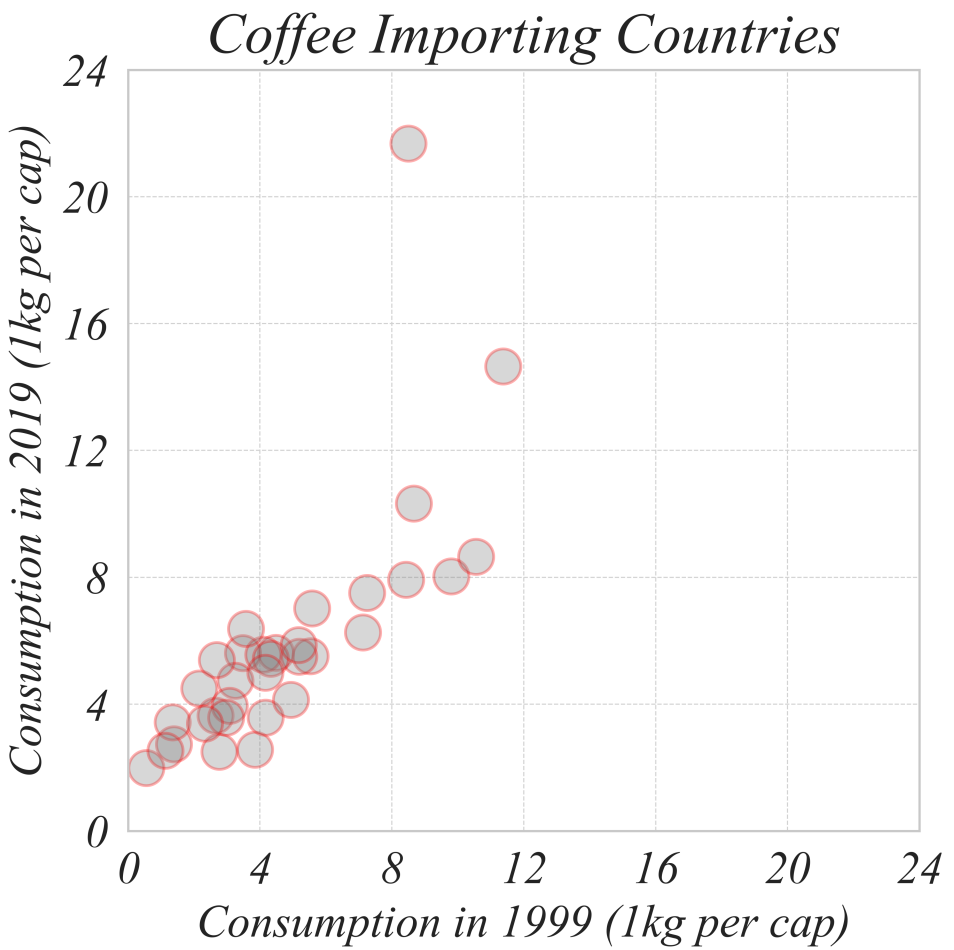
> better, but this figure still doesn't let us keep track of countries between years…

Panel Data: Relationships Between Years

How many countries increased their coffee consumption between 1999 and 2019?

Panel Data: Relationships Between Years

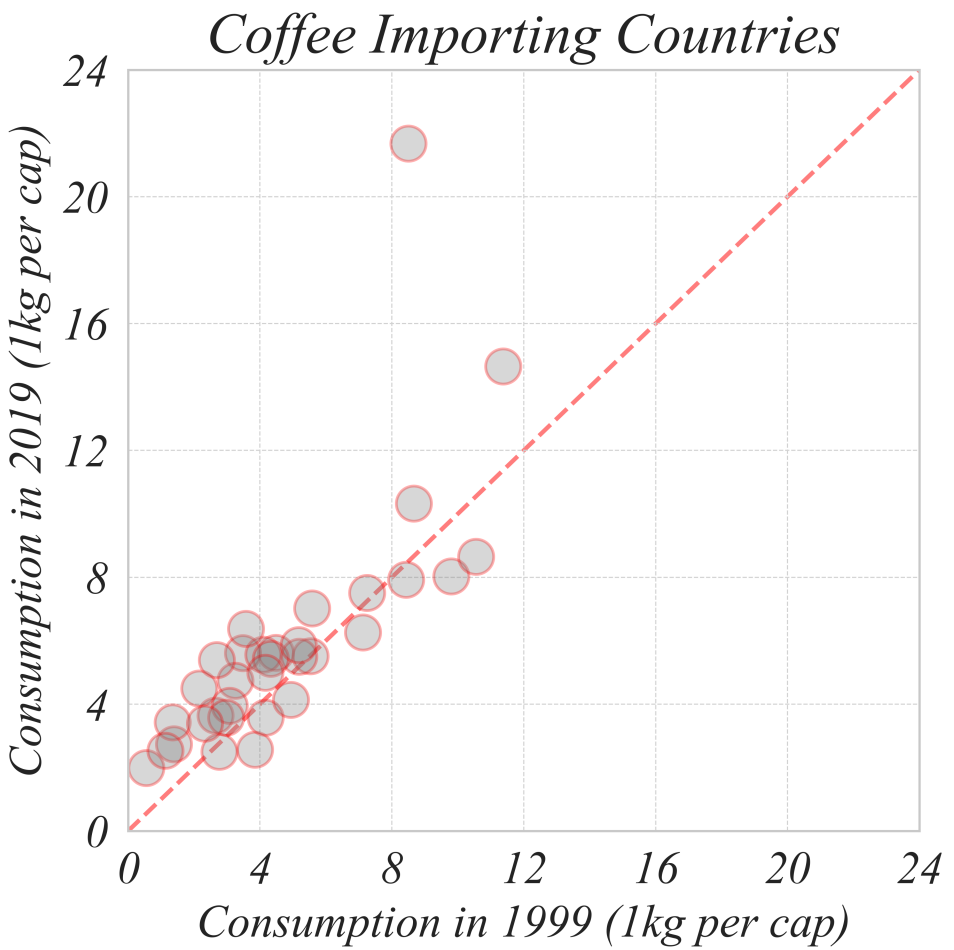
How many countries increased their coffee consumption between 1999 and 2019?



> *a scatter plot can visualize changes between two points in time*

Panel Data: Relationships Between Years

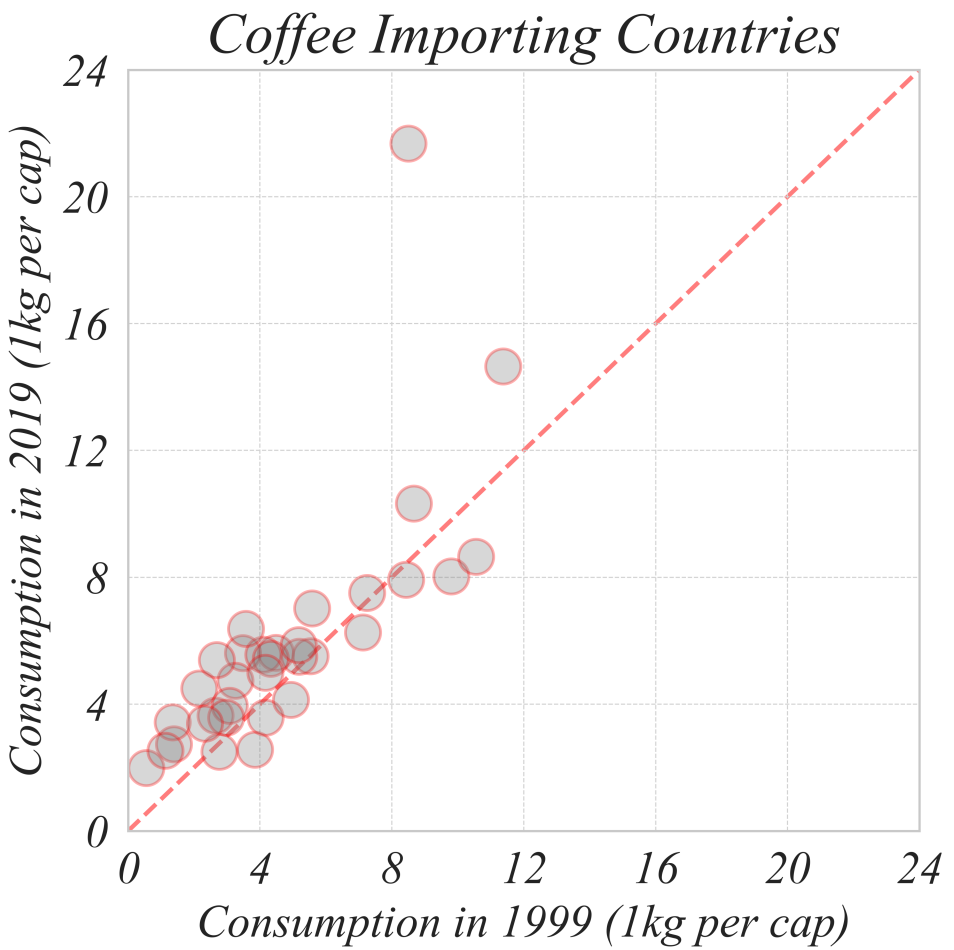
How many countries increased their coffee consumption between 1999 and 2019?



> a 45 degree line shows all the possible points with no change

Panel Data: Relationships Between Years

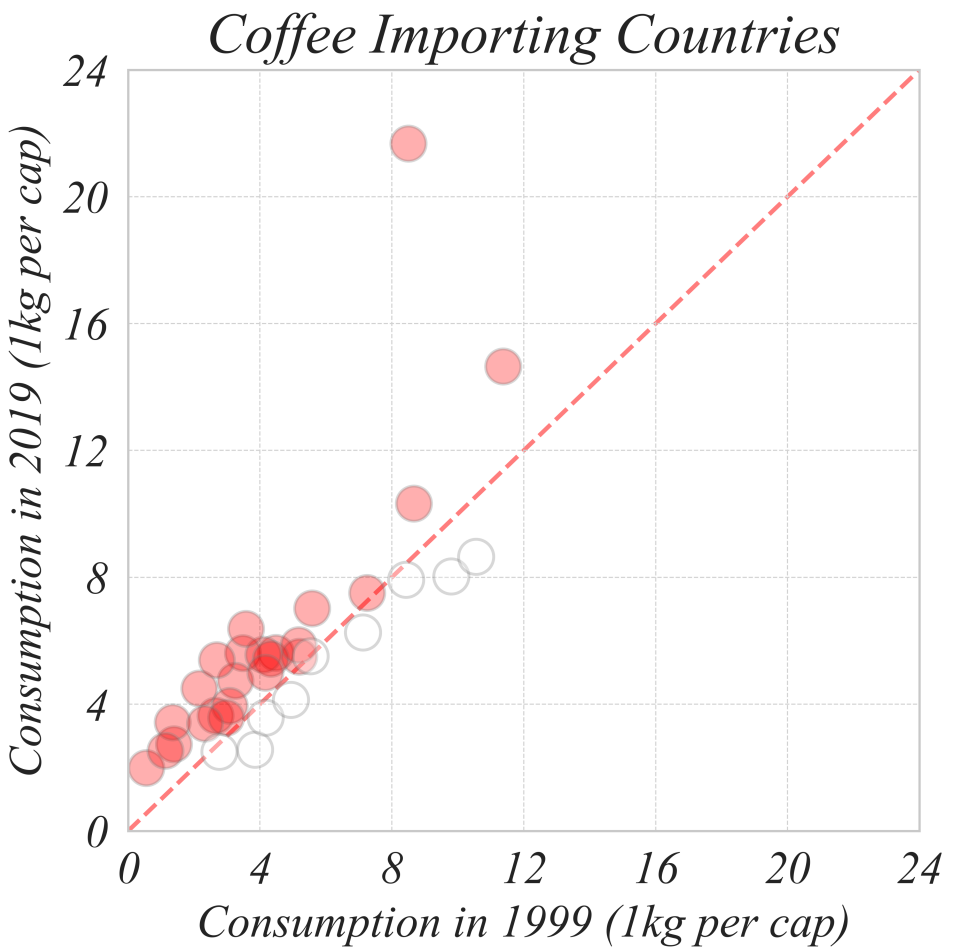
How many countries increased their coffee consumption between 1999 and 2019?



> a 45 degree line shows all the possible points with no change

Panel Data: Relationships Between Years

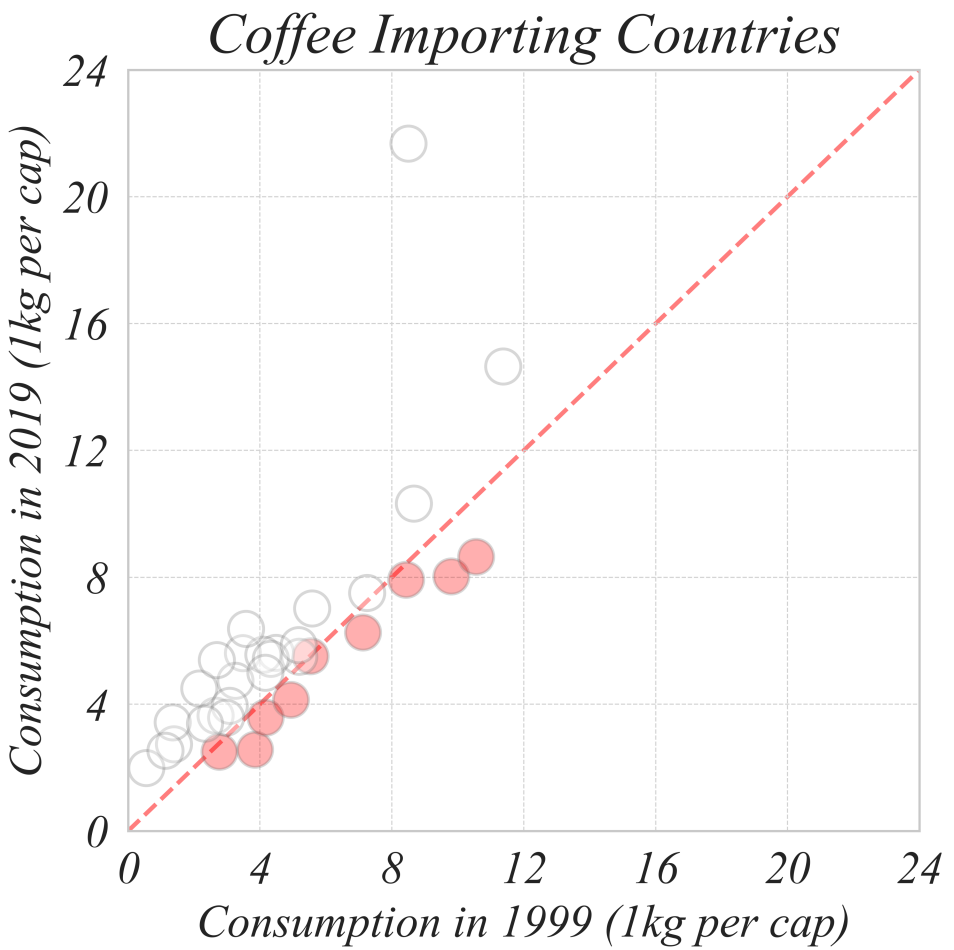
How many countries increased their coffee consumption between 1999 and 2019?



> *points above the line increased*

Panel Data: Relationships Between Years

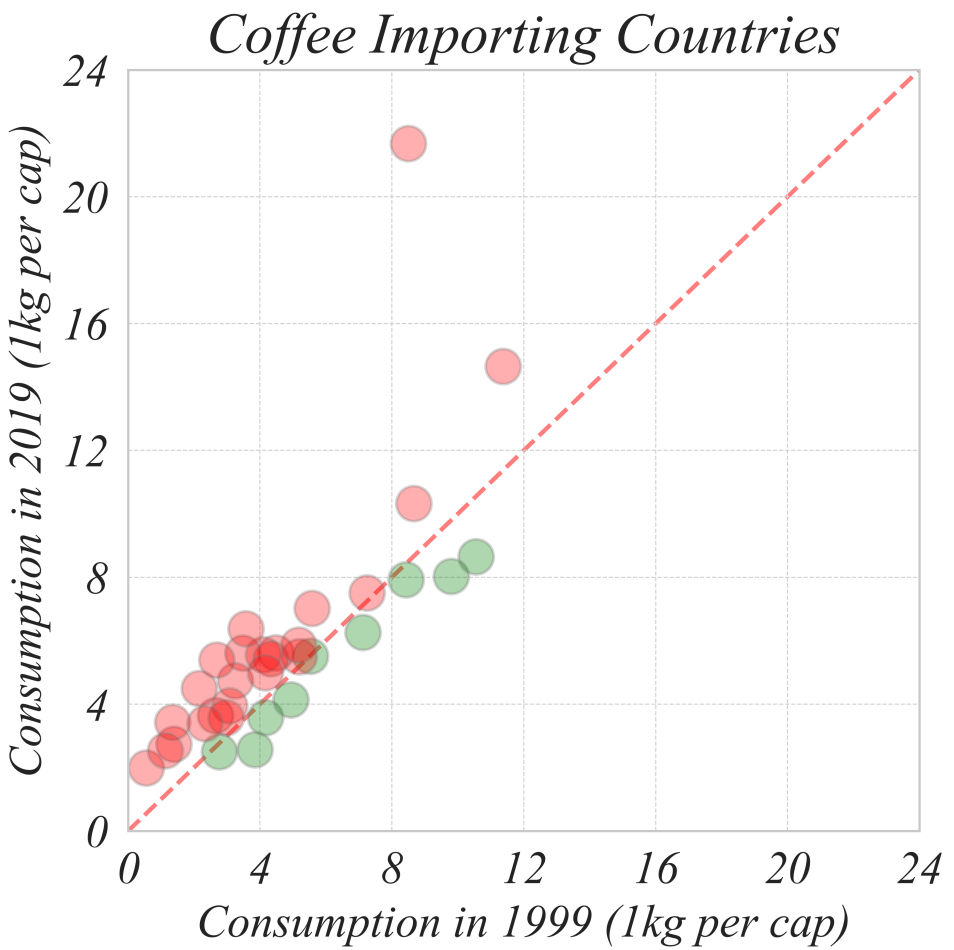
How many countries decreased their coffee consumption between 1999 and 2019?



> points below the line decreased

Panel Data: Relationships Between Years

Does the data confirm that the world is drinking more coffee?



> *we can use colors to visualize both increases and decreases*

Exercise 1.4 | Scatterplots

Is the world drinking more coffee?

We're going to use a scatterplot to visually examine how countries' coffee consumption changed between 1999 and 2019.

- ***Data:*** `Coffee_Per_Cap.csv`

Exercise 1.4 | Scatterplots

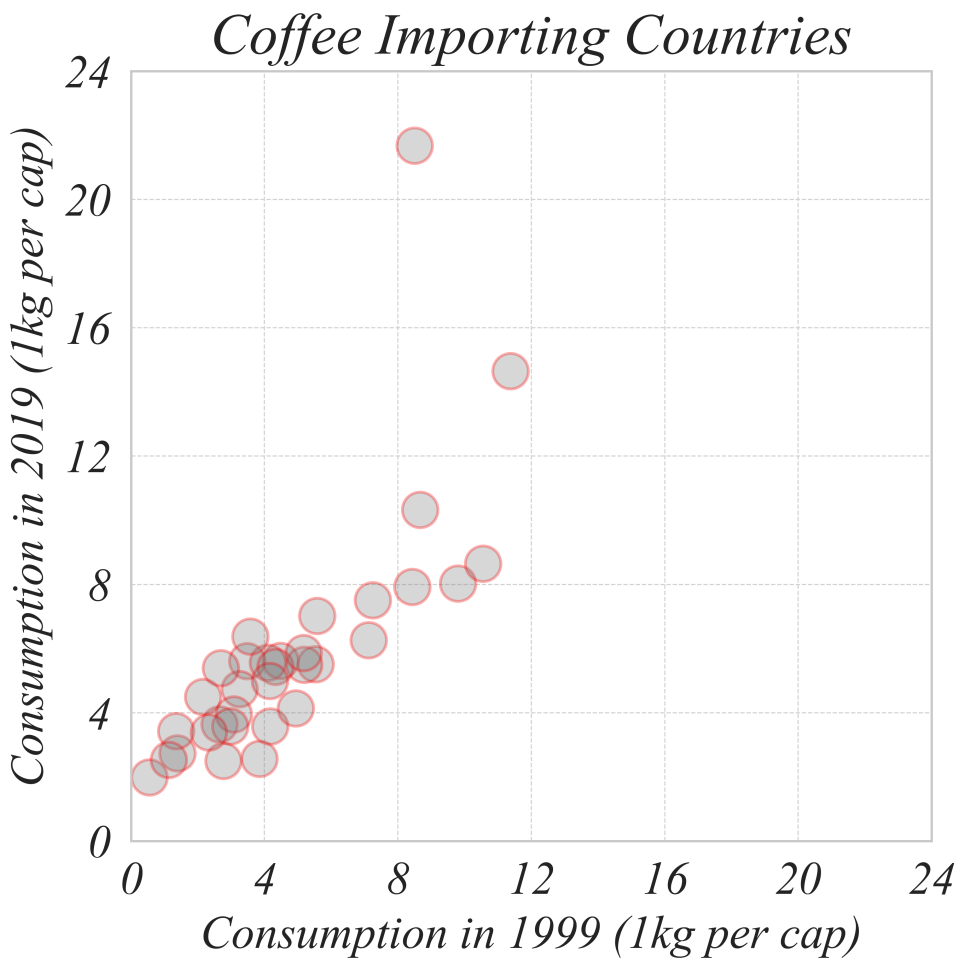
Is the world drinking more coffee?

```
# Wide Format Scatterplot
sns.scatterplot(percap, x='1999', y='2019')
```

Exercise 1.4 | Scatterplots

Is the world drinking more coffee?

```
# Wide Format Scatterplot
sns.scatterplot(percap, x='1999', y='2019')
```



Part 1.4 | Panel Data Using Scatterplots

Summary

- *Multi-Boxplots can help visualize changes in the distribution, but cannot track individual changes.*
- *Scatterplots can show how repeated observations change through time within a single unit.*
- *A 45 degree line and colors can help visually communicate changes.*