ECON 0150 | Economic Data Analysis

The economist's data analysis skillset.

Part 1.2 | Cross-Sectional (Numerical) Data

Cross-Sectional Numerical Data

Comparing numerical values across entities

> Cross-sectional data: many entities, one point in time

> Numerical variables: values you can do math with (age, income, consumption)

> Key question: How is this variable distributed?

Two Tools for Numerical Distributions

Choose based on sample size and what you want to see

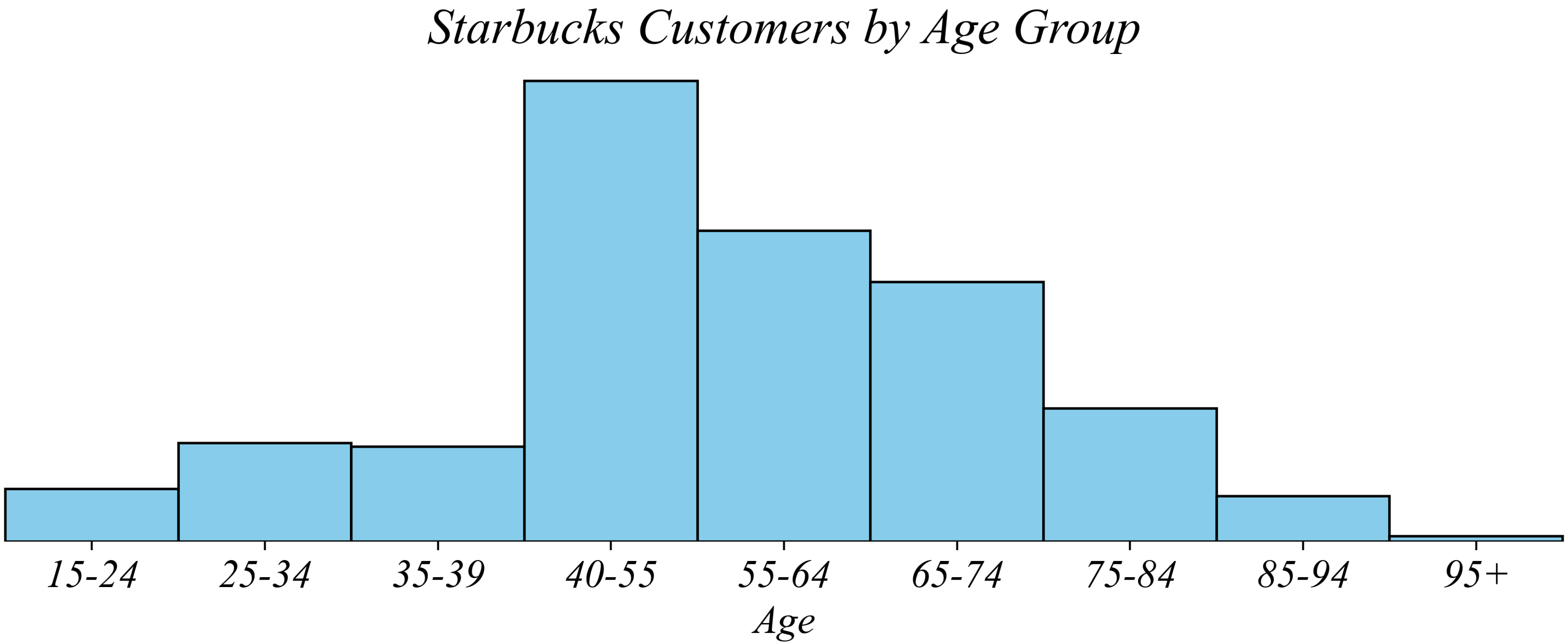
Tool	**Best for**	**Shows**
Histogram	Many observations	Shape of distribution
Boxplot + Stripplot	Fewer observations	Quartiles + individual values

Histograms: Shape of the Distribution

Use when you have many observations

Histograms

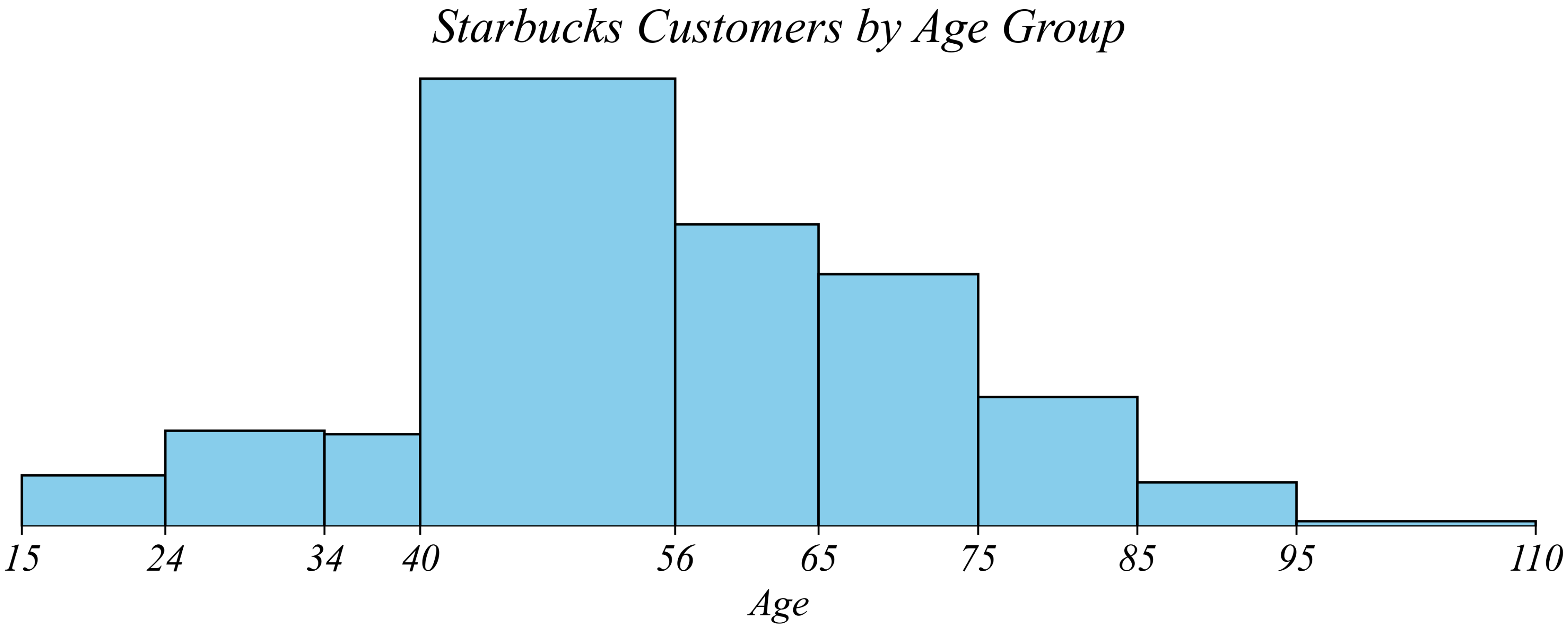
Q. Which age group has the most Starbucks customers?



> the bin sizes aren't even, making it hard to interpret

Numerical Variables: Histograms

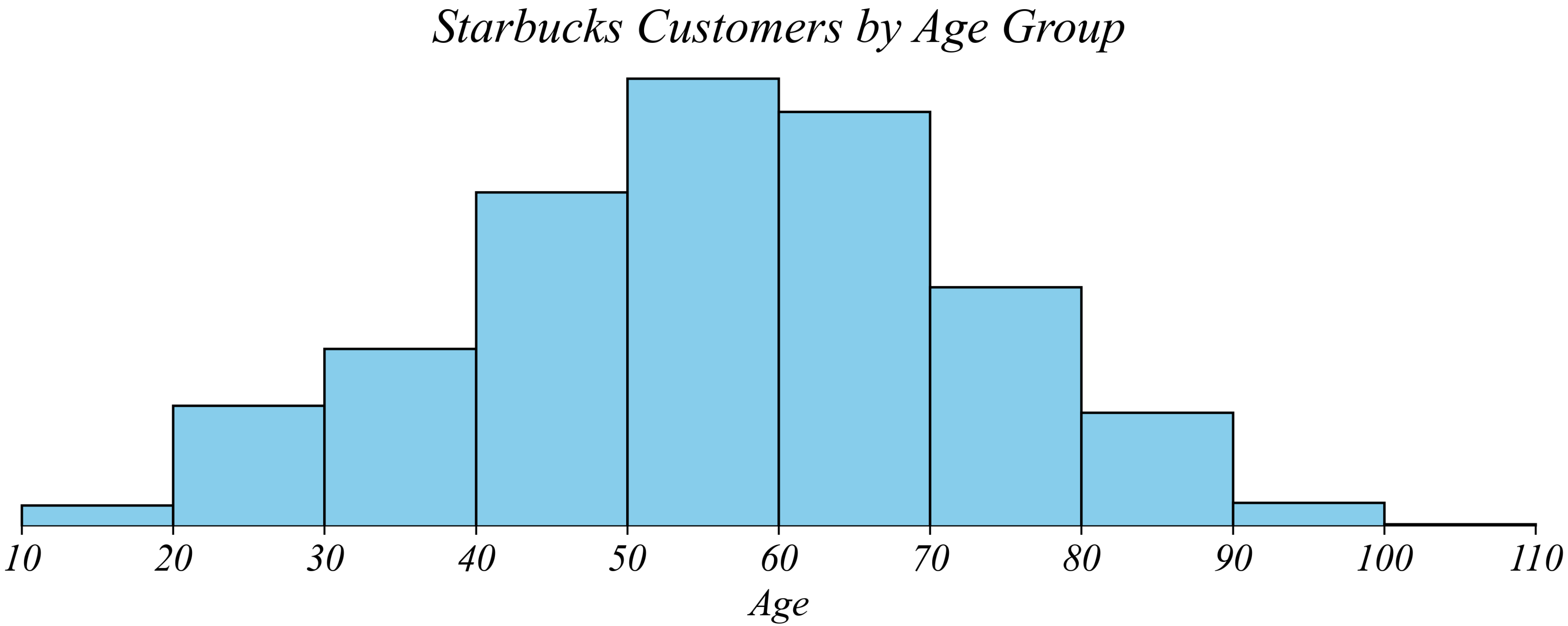
Q. Which age group has the most Starbucks customers?



> the bin sizes aren't even, making it hard to interpret

Histograms: Use equal sized bins

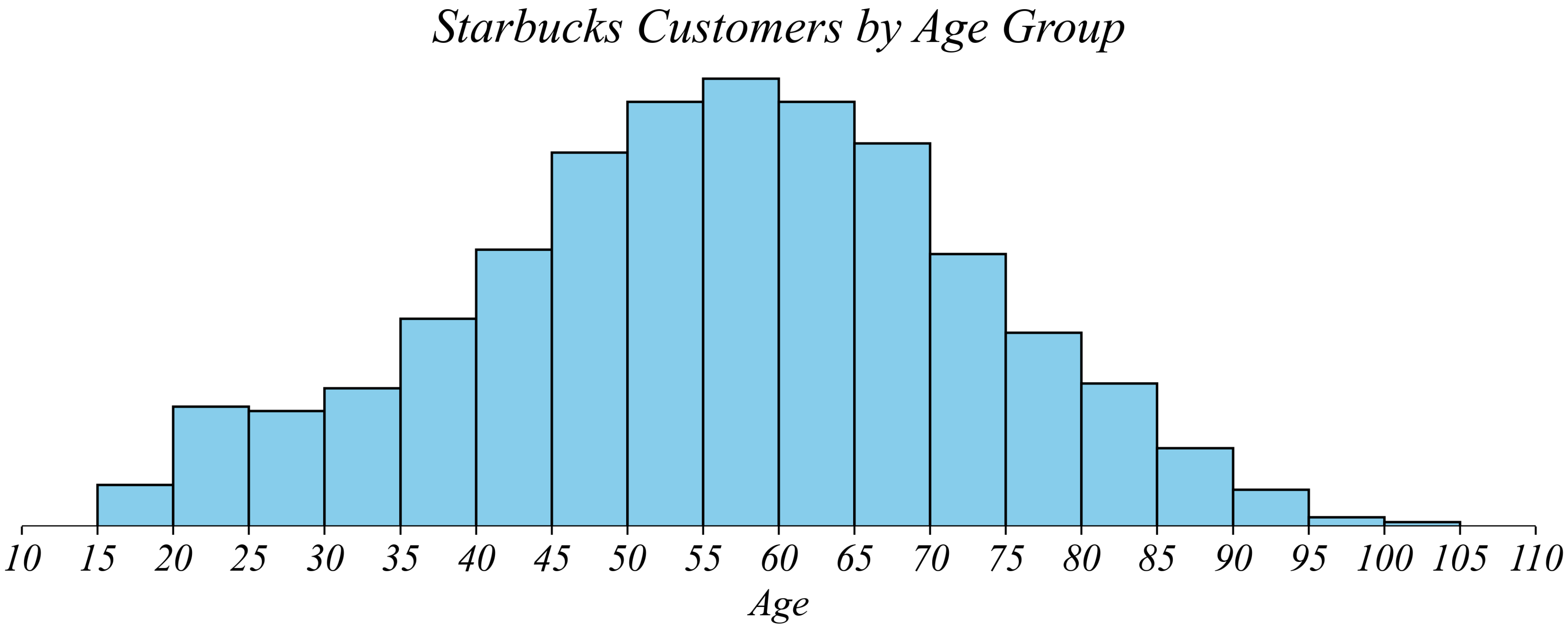
Q. Which age group has the most Starbucks customers?



> but what if we want to distinguish between a 55 year old and a 60 year old?

Histograms: Use narrow enough bins

Q. Which age group has the most Starbucks customers?

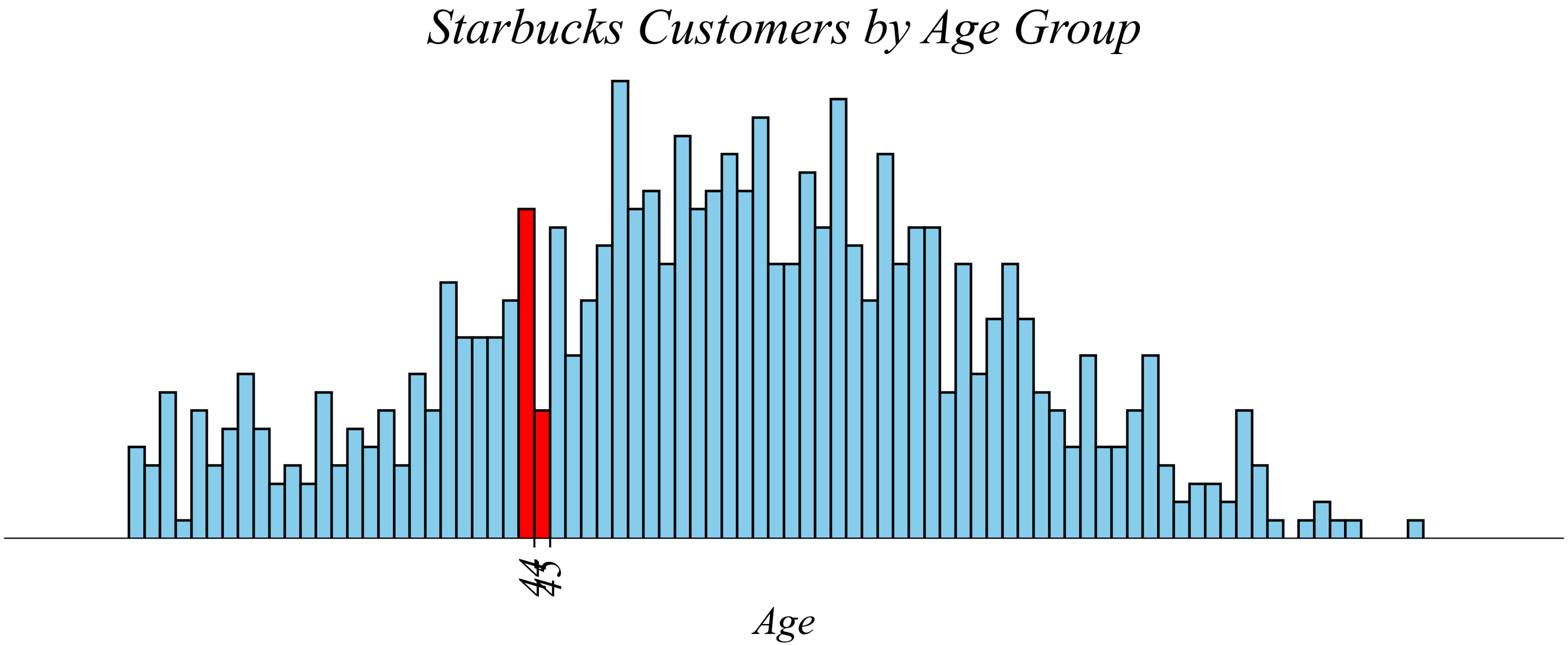


> what if we take this even further?

> what if we compare 44 year olds to 45 year olds?

Histograms: Avoid visualizing noise

Q. Do 44 or 45 year olds spend more at Starbucks?

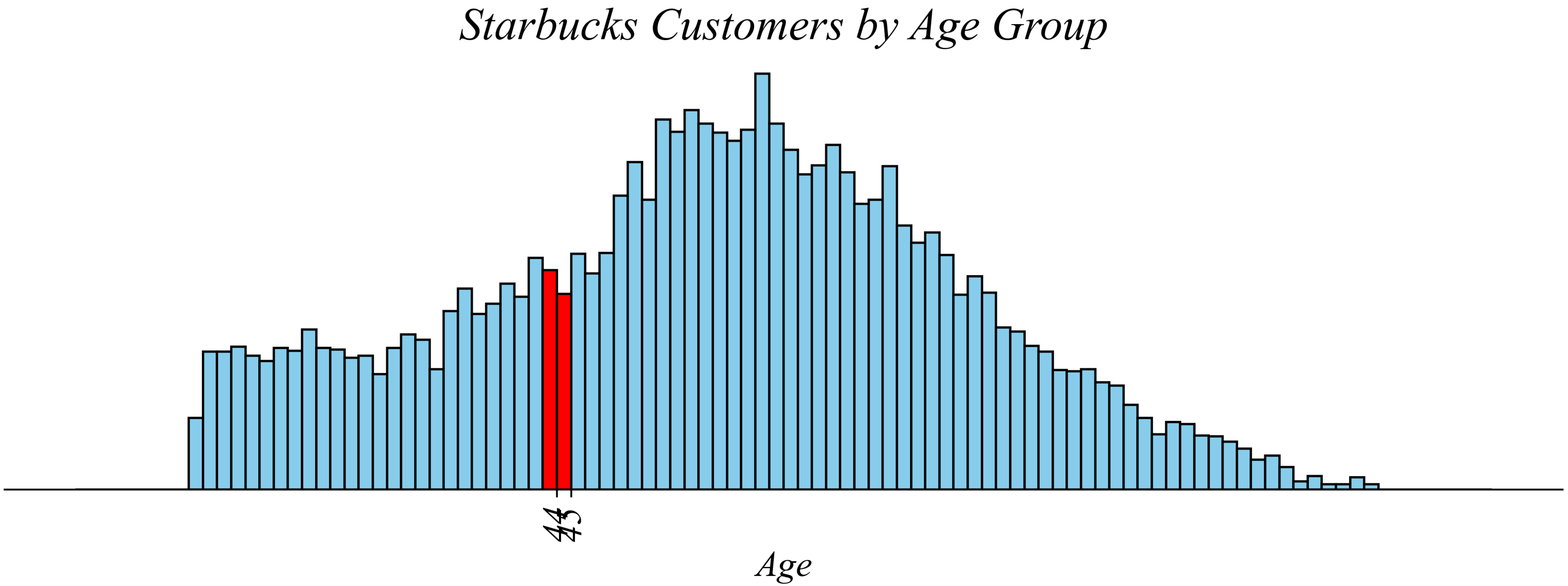


> we can go too far, introducing statistical noise. how do we fix the problem?

> increase the sample size or the bin width!

Histograms: Balance resolution vs noise

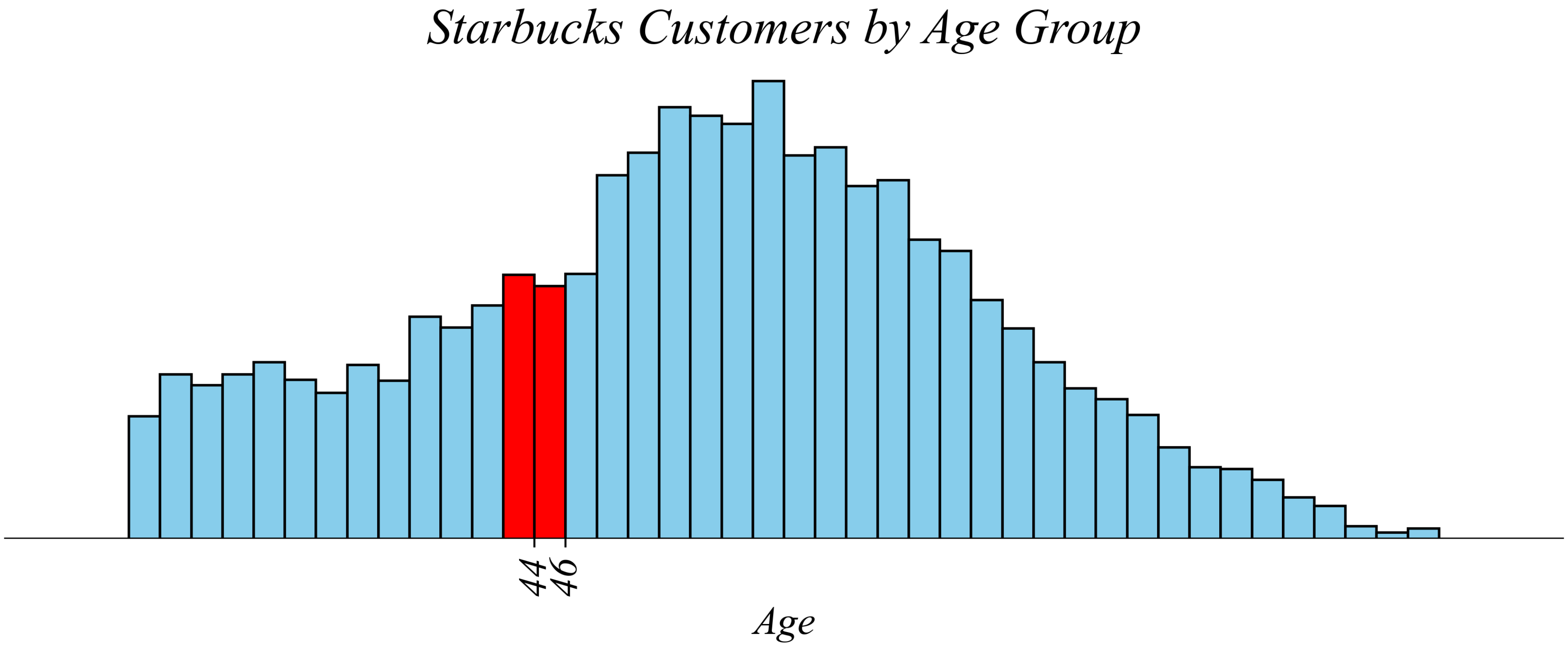
Q. Which age group has the most Starbucks customers?



> larger sample has less noise!

Histograms: Balance resolution vs noise

Q. Which age group has the most Starbucks customers?



> larger bins also has less noise!

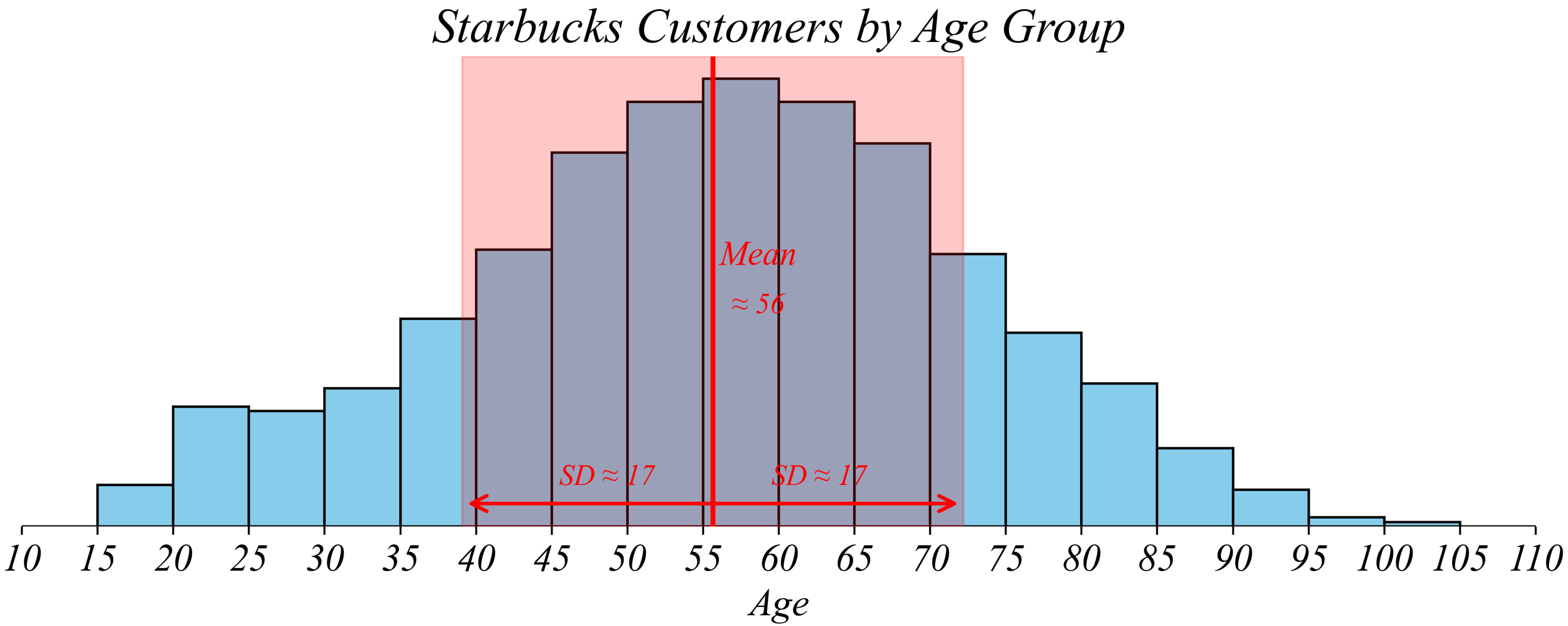
Describing the Distribution: Center and Spread

Two numbers that summarize a histogram

- ***Mean*** *— the average value (center)*
- ***Standard Deviation (SD)*** *— typical distance from the mean (spread)*

Mean and Standard Deviation

Q. What is the average age of Starbucks customers?



> Mean ≈ 56 years; SD ≈ 17 years

> "The average customer is about 56; ages typically vary by about 17 years from that average"

Histograms: Summary

... use the right summary tool for the variable type

- *Use histograms to visualize continuous variables.*
- *Make histograms with equally sized bins.*
- *Histograms with bins that are too narrow increase statistical noise, which can obscure underlying relationships.*

S-T-E for Histograms

What we just did

Step	**Action**
SELECT	All Starbucks customers
TRANSFORM	Count customers within each age bin
ENCODE	Bin → x-position; Count → bar height

> *TRANSFORM for histograms = count within bins*

Exercise 1.2 | Histograms

Q. Which age group among those making $40k or less has the most Starbucks customers?

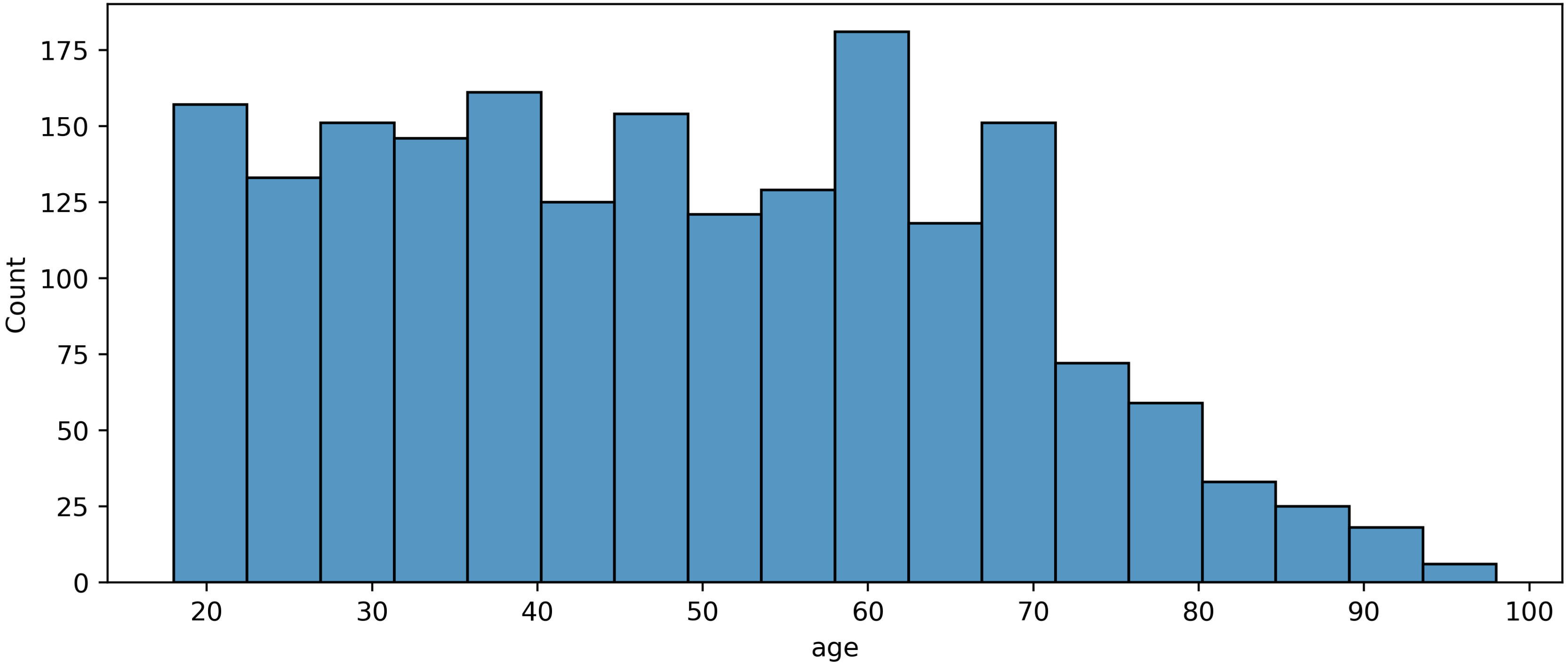
Lets use the data to examine whether customers between 45 - 55 years old spend the most among customers making less than $40k.

Data: `Starbucks_Customer_Profiles_40k.csv`

Exercise 1.2 | Histograms

Q. Which age group among those making $40k or less has the most Starbucks customers?

```python
# Histogram with 5 year bins
sns.histplot(customers, x='age', bins=range(20,100,5))
```



```python
# Save Figure
plt.savefig('exercise_1_2_histogram.png')
```

Exercise 1.2 | Mean and Standard Deviation

Summarize the distribution with two numbers

```python
# Calculate the mean
customers['age'].mean()
```

```python
# Calculate the standard deviation
customers['age'].std()
```

> *Mean tells us the center; SD tells us the spread*

“The average customer is about 48 years old; ages typically vary by about 18 years from that average.”

What if we have fewer observations?

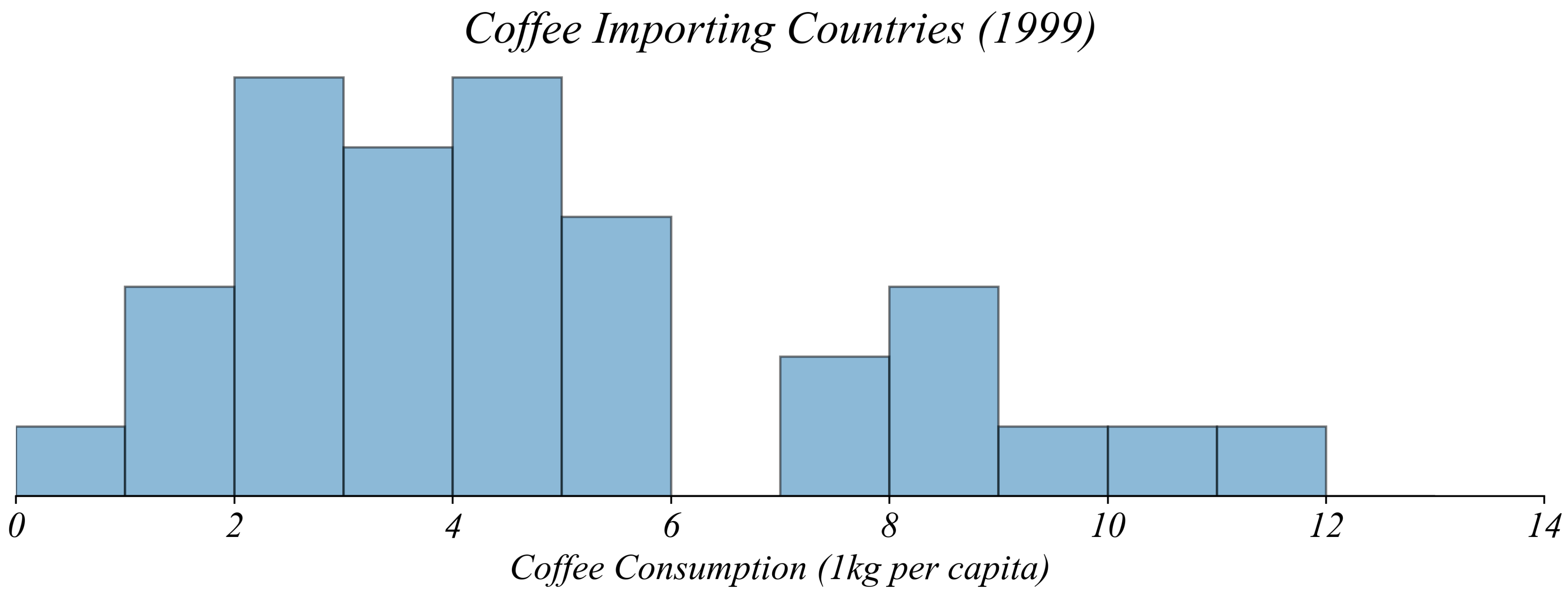
Histograms need many data points to show shape

> *With few observations, histogram bins become noisy or empty*

> *We need a different tool: boxplots + stripplots*

Histograms vs Boxplots

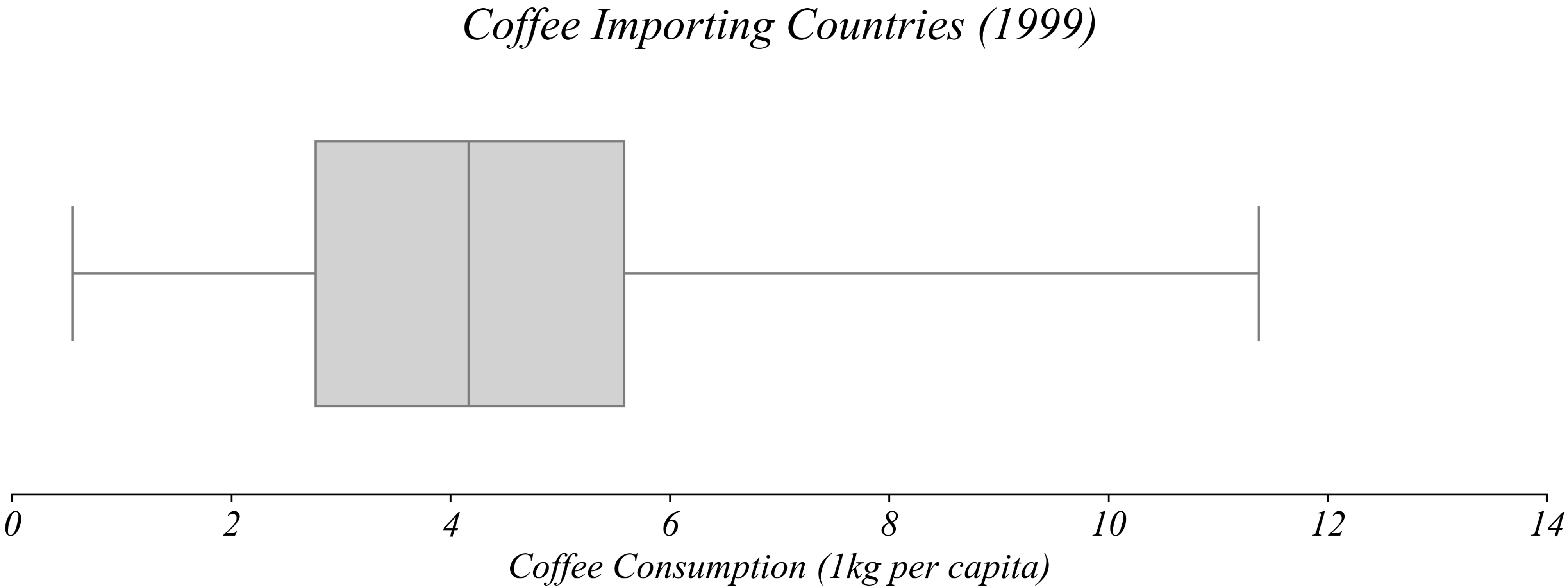
Q. Which countries drank an average amount of coffee?



> histogram bins make it impossible to see exact values or quartiles

Boxplots

Q. Which countries drank the most coffee in 1999?

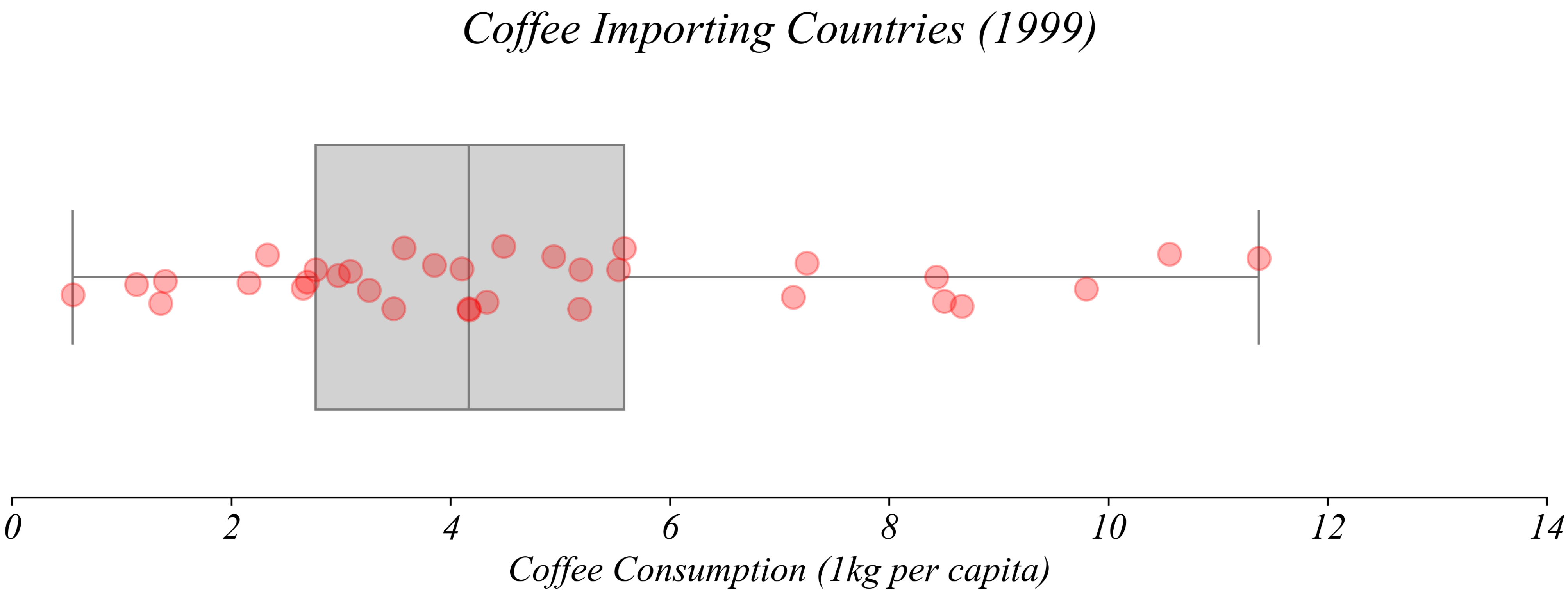


> *as we'll see, boxplots can tell us about quartiles*

> *but boxplots are still pretty unclear for our question*

Boxplots + Stripplots

Q. Which countries drank the most coffee in 1999?

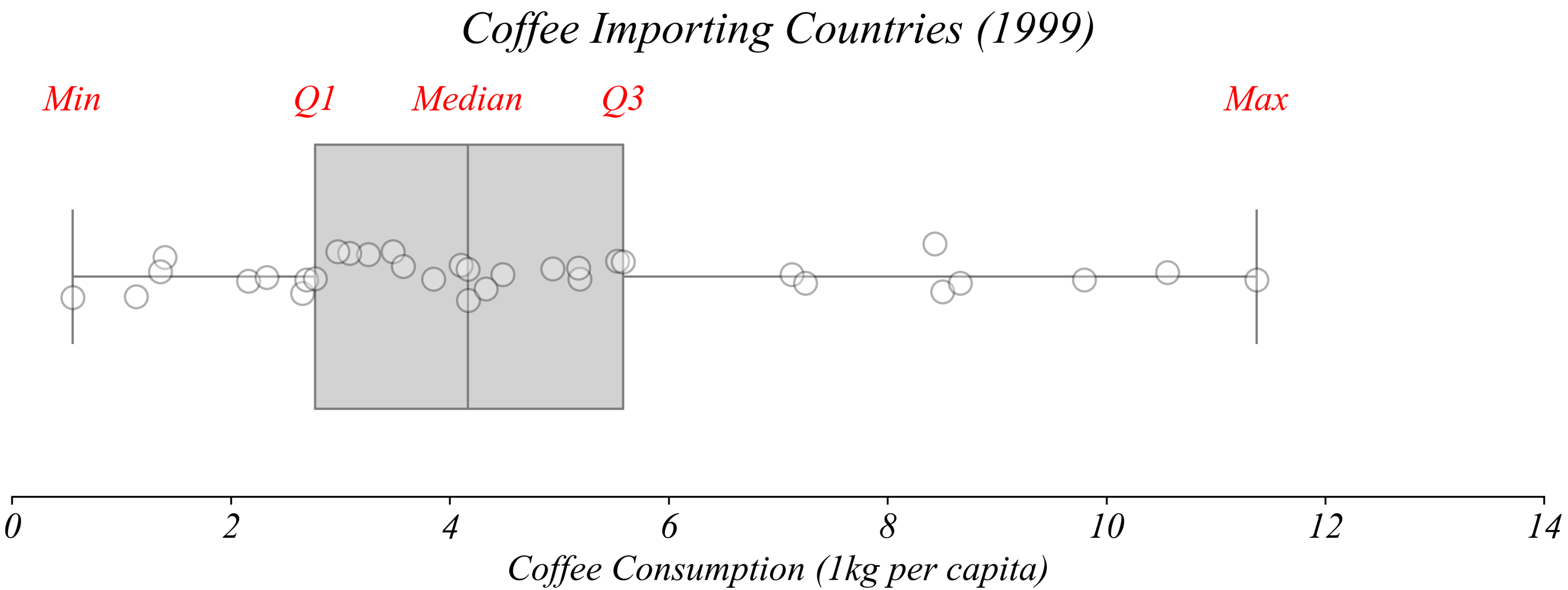


> here we can see the datapoints directly with the boxplot

> each point represents a country's coffee consumption

Boxplots + Stripplots

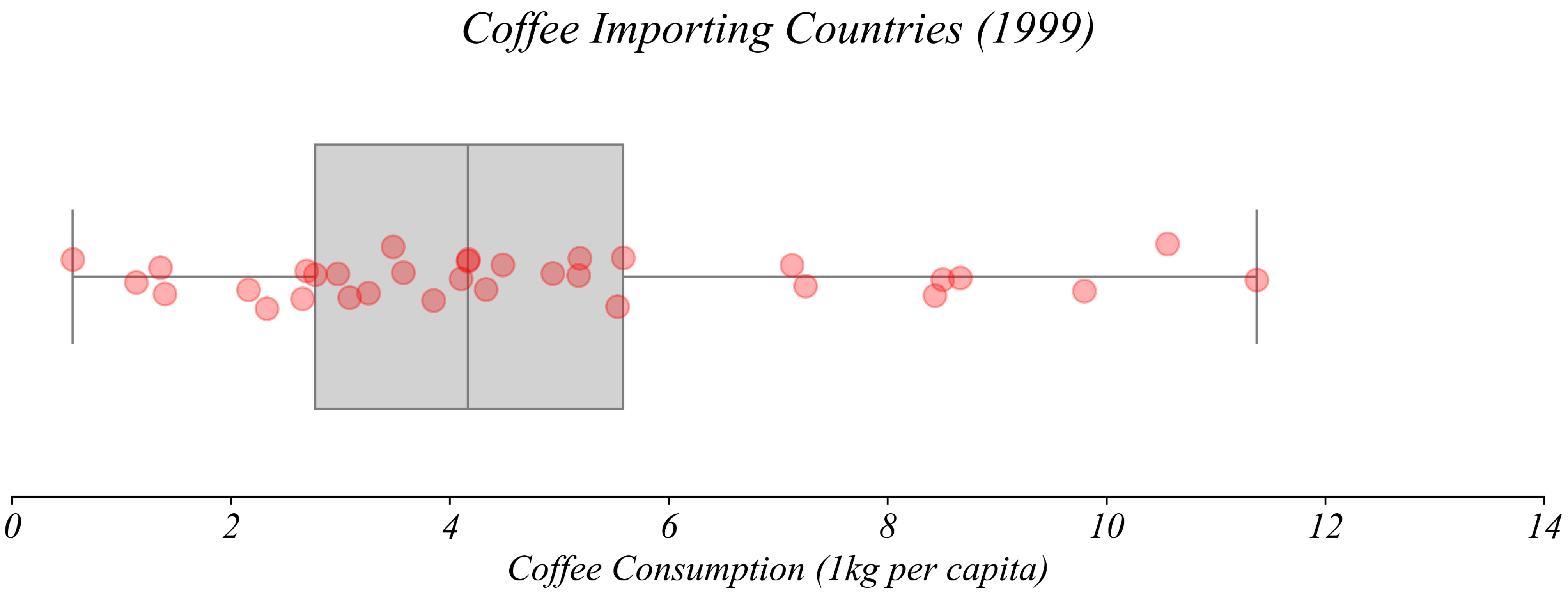
Q. Which countries drank the most coffee in 1999?



> each element of the boxplot represents one of these five quartiles

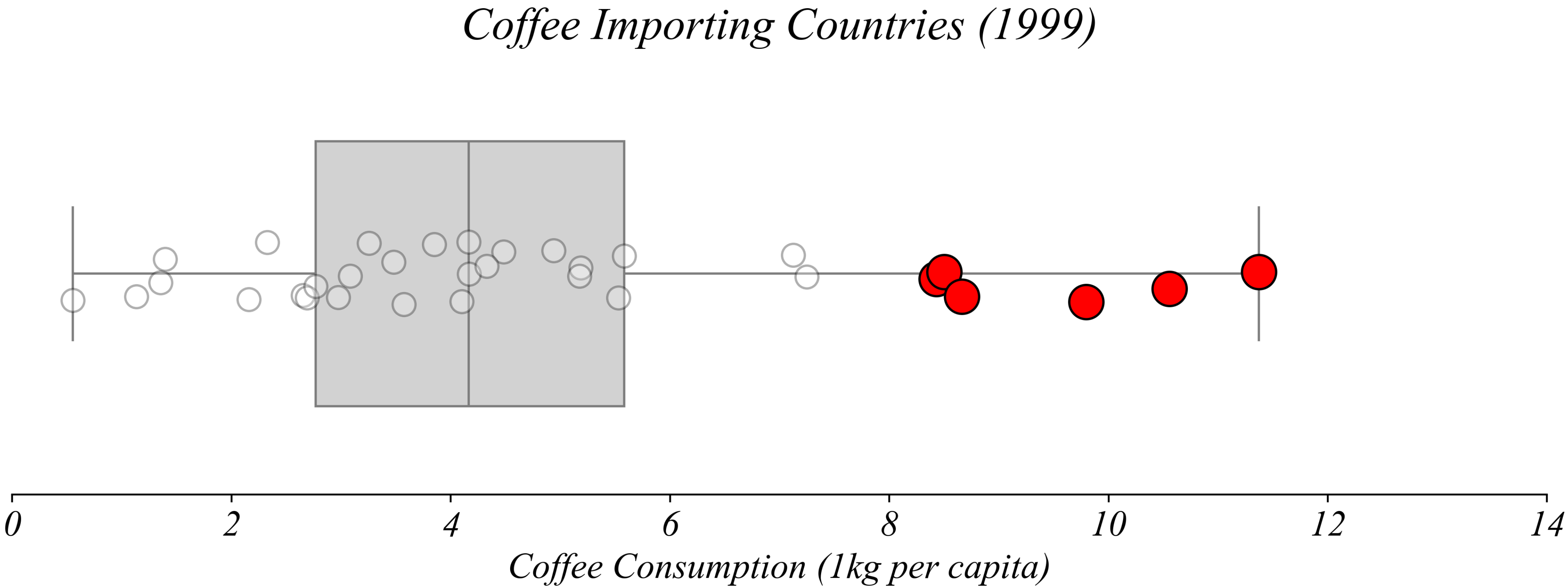
Boxplots + Stripplots

Which countries consumed more than 8 kg per capita?



Boxplots + Stripplots

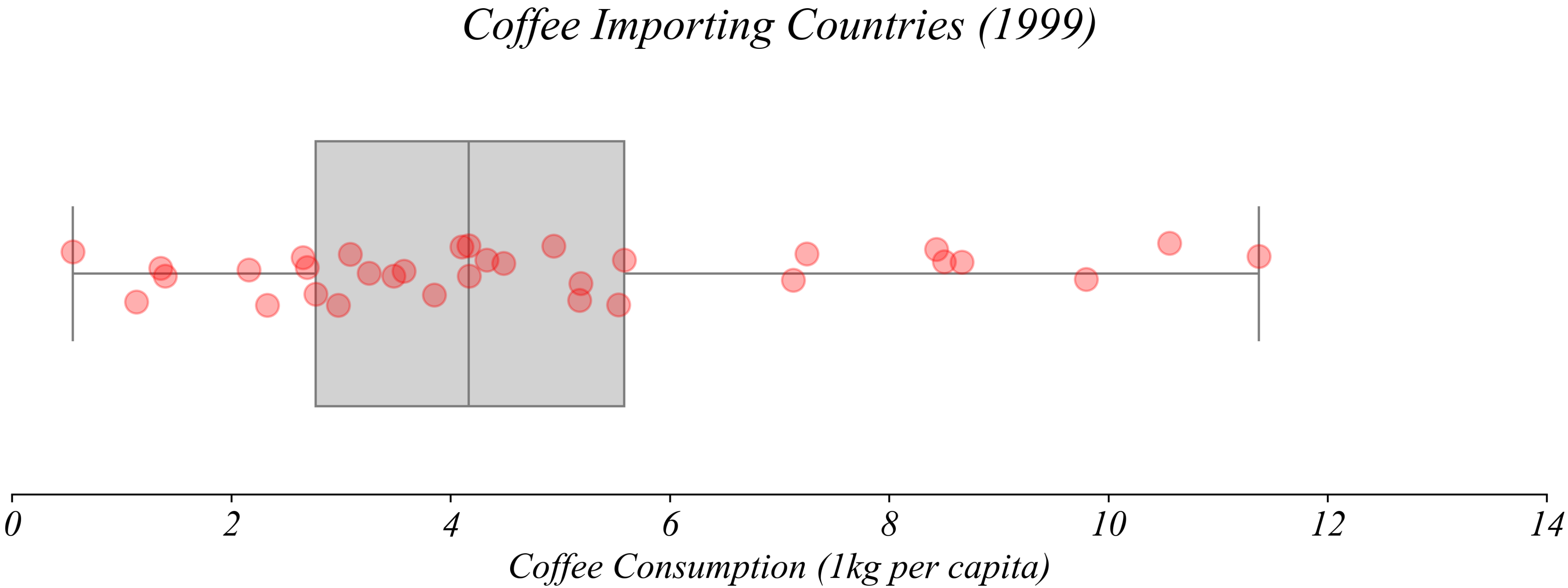
Which countries consumed more than 8 kg per capita?



> we can highlight the relevant subsets of the data

Boxplots + Stripplots

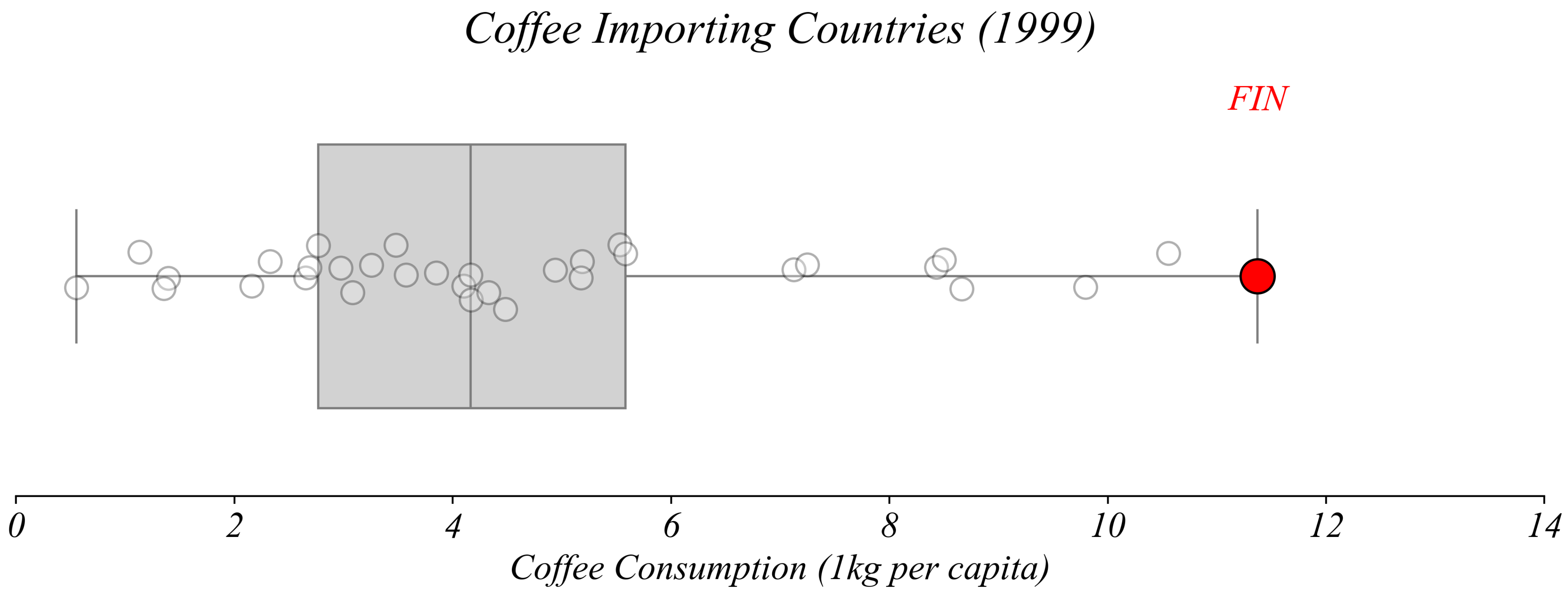
Which country consumed the most coffee per capita?



> we can find the exact values according to quartiles

Boxplots + Stripplots

Which country consumed the most coffee per capita?

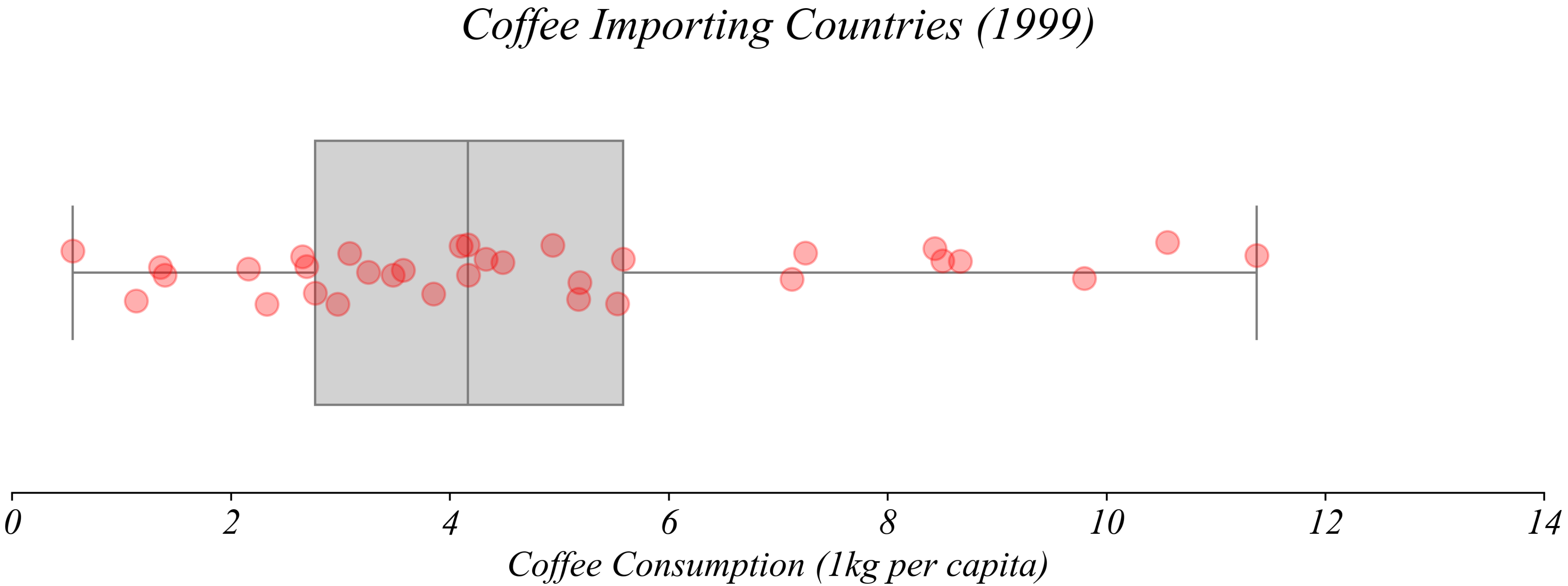


> we can find the exact values according to quartiles

> Finland consumed the most coffee per capita in 1999

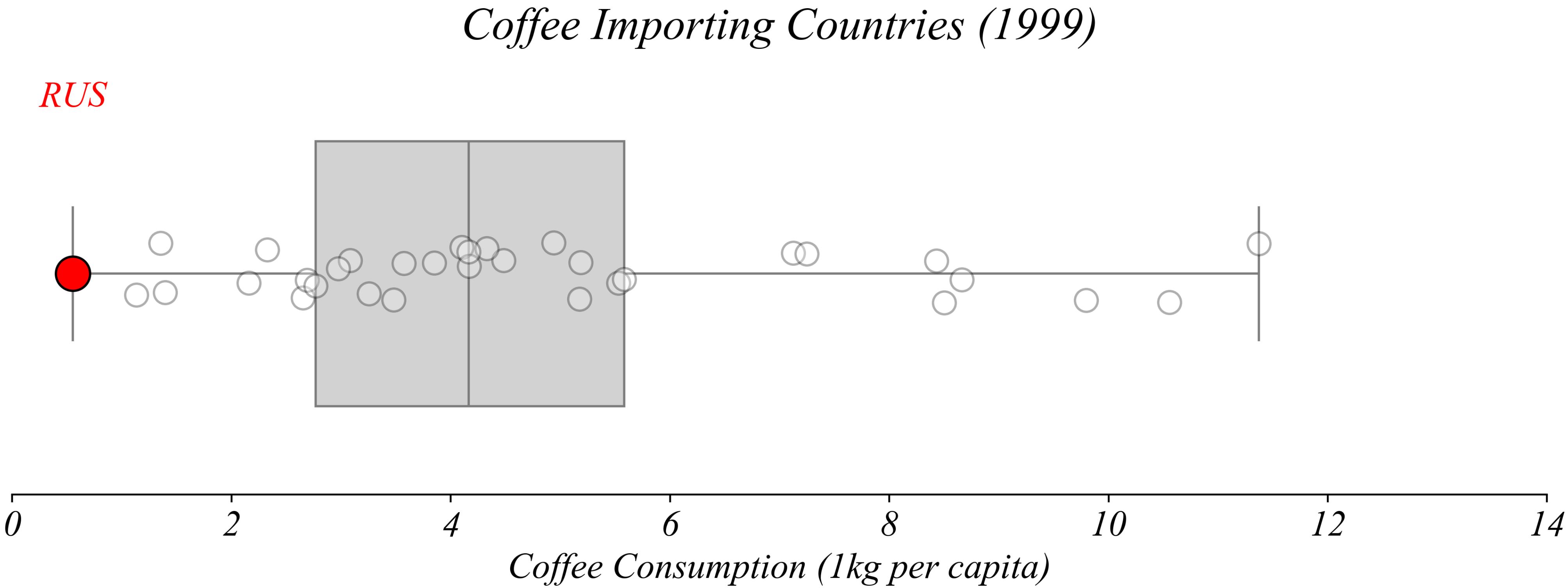
Boxplots + Stripplots

Which country consumed the least coffee per capita?



Boxplots + Stripplots

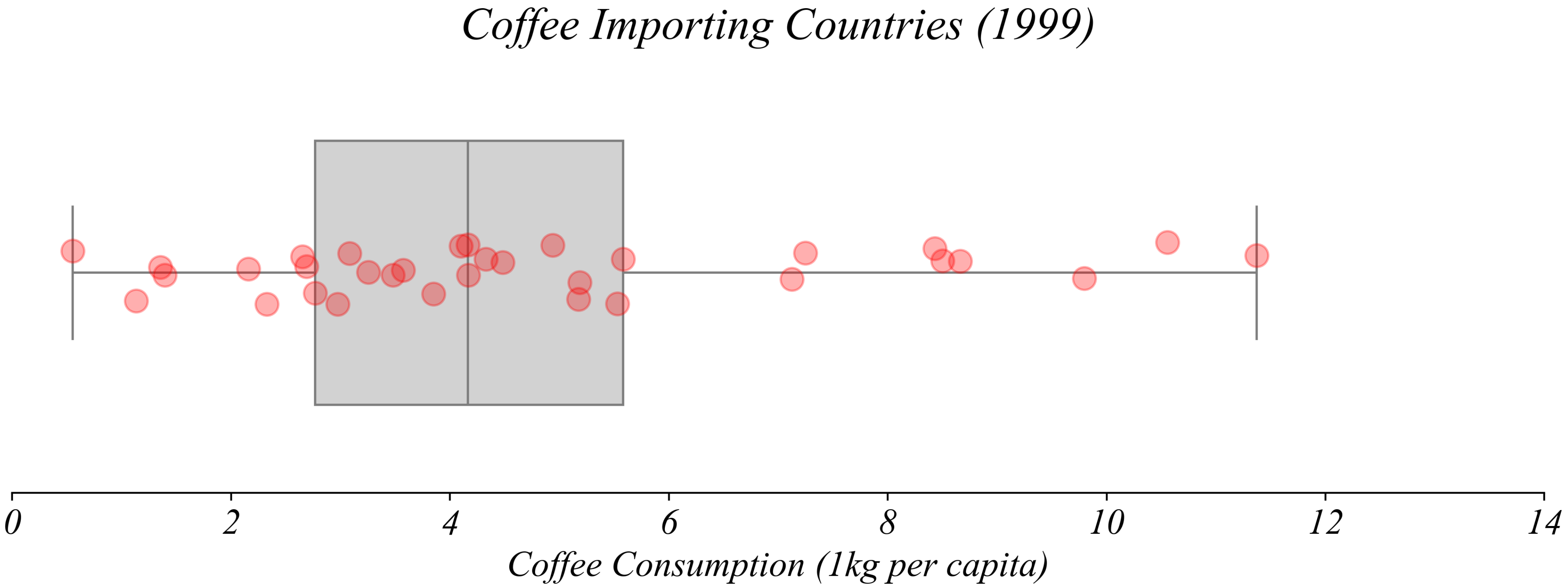
Which country consumed the least coffee per capita?



> Russia consumed the least coffee per capita in 1999

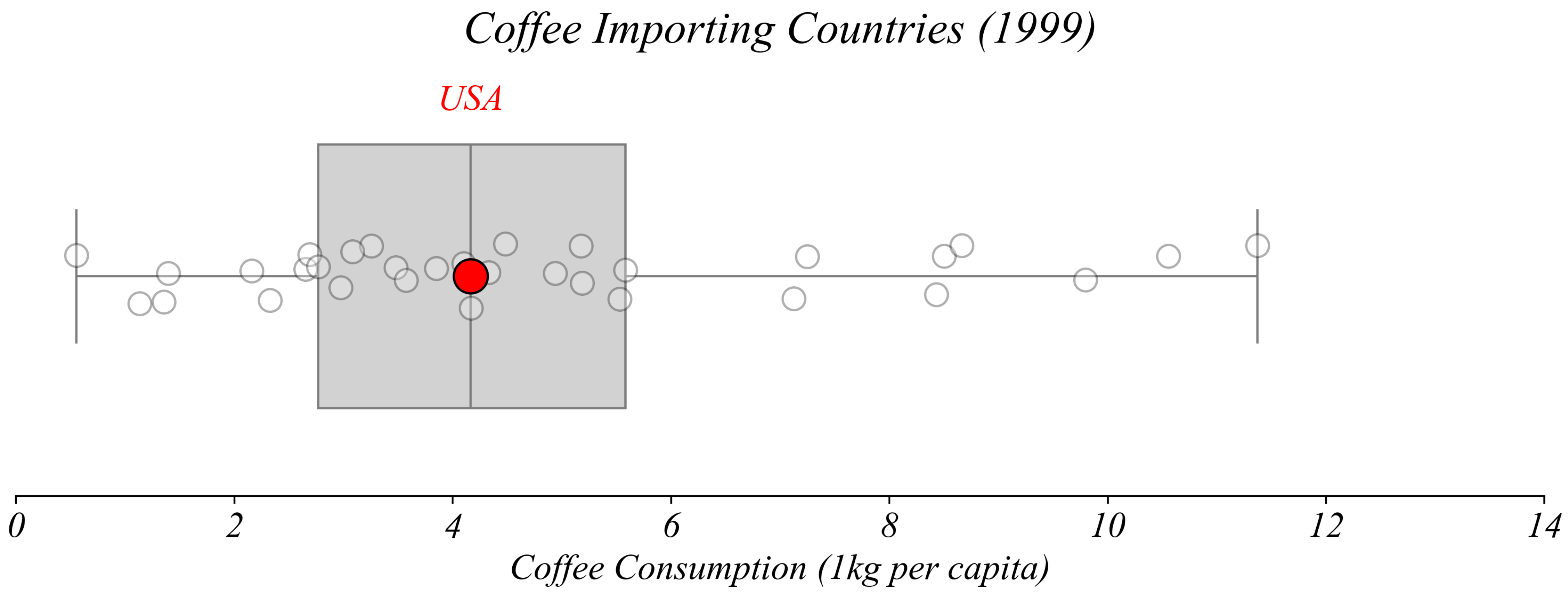
Boxplots + Stripplots

How about the median?



Boxplots + Stripplots

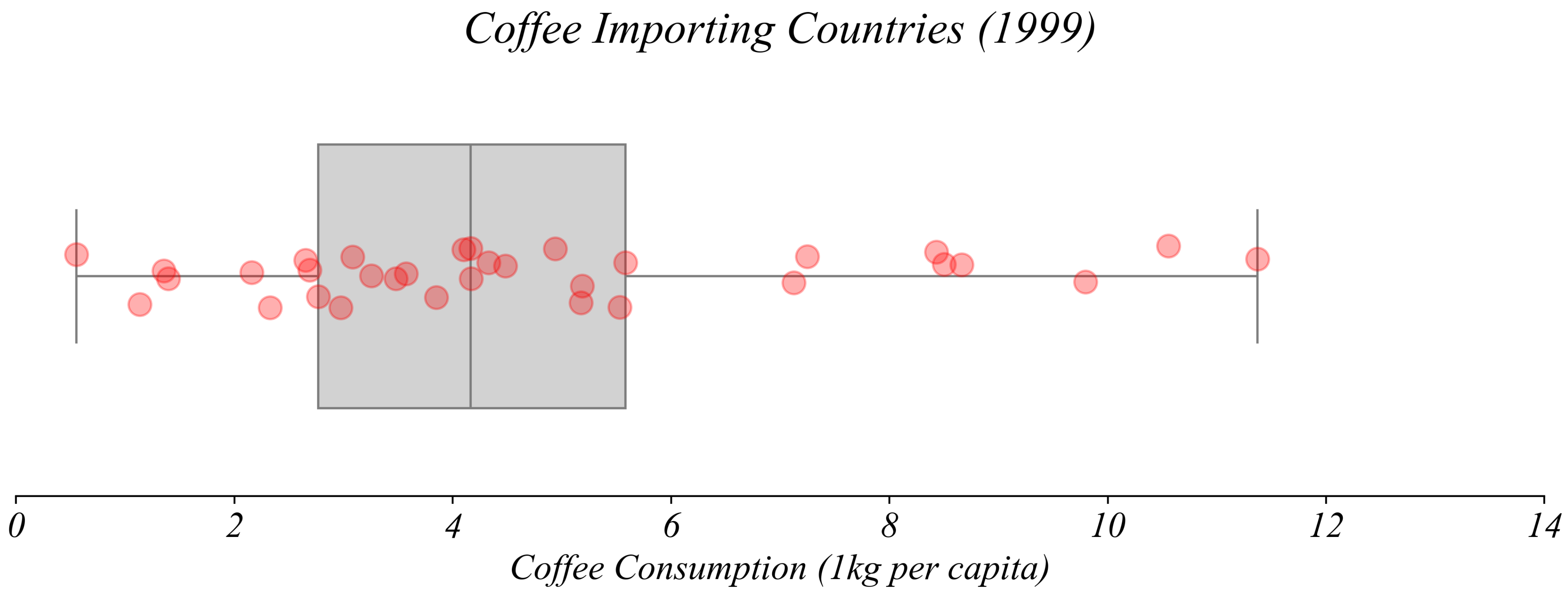
How about the median?



> the US!

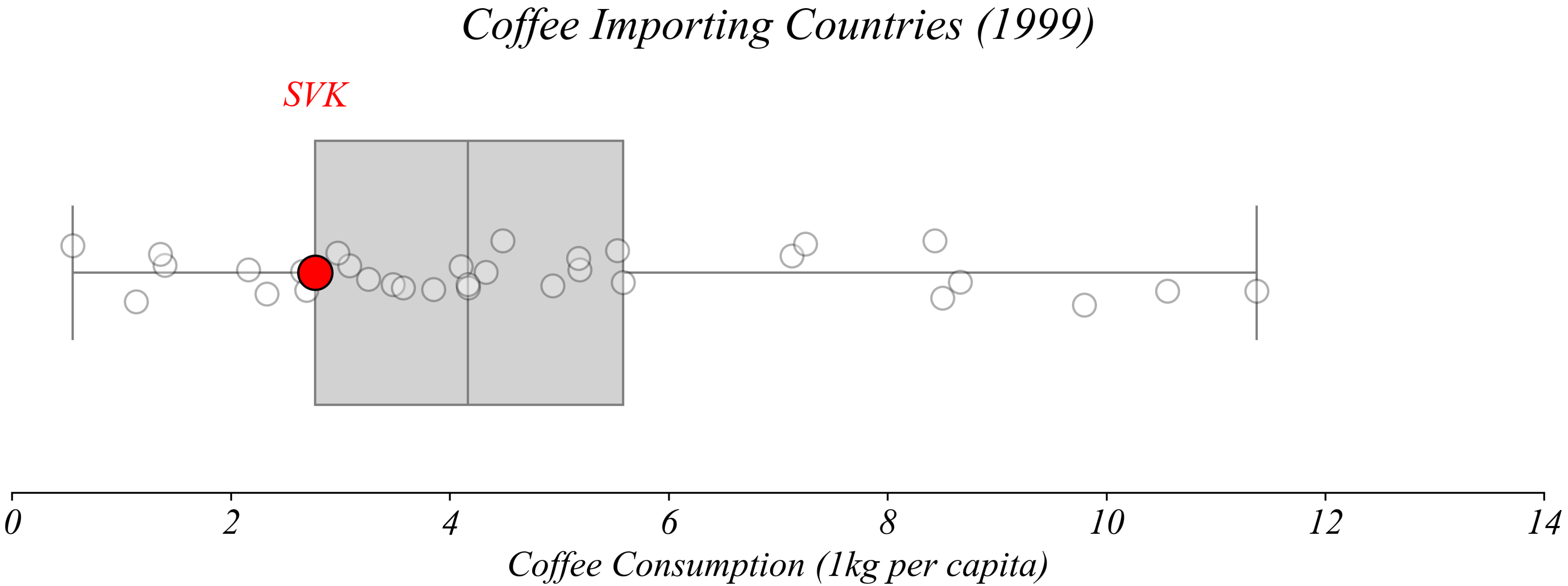
Boxplots + Stripplots

Which country consumes more than exactly 25% of countries?



Boxplots + Stripplots

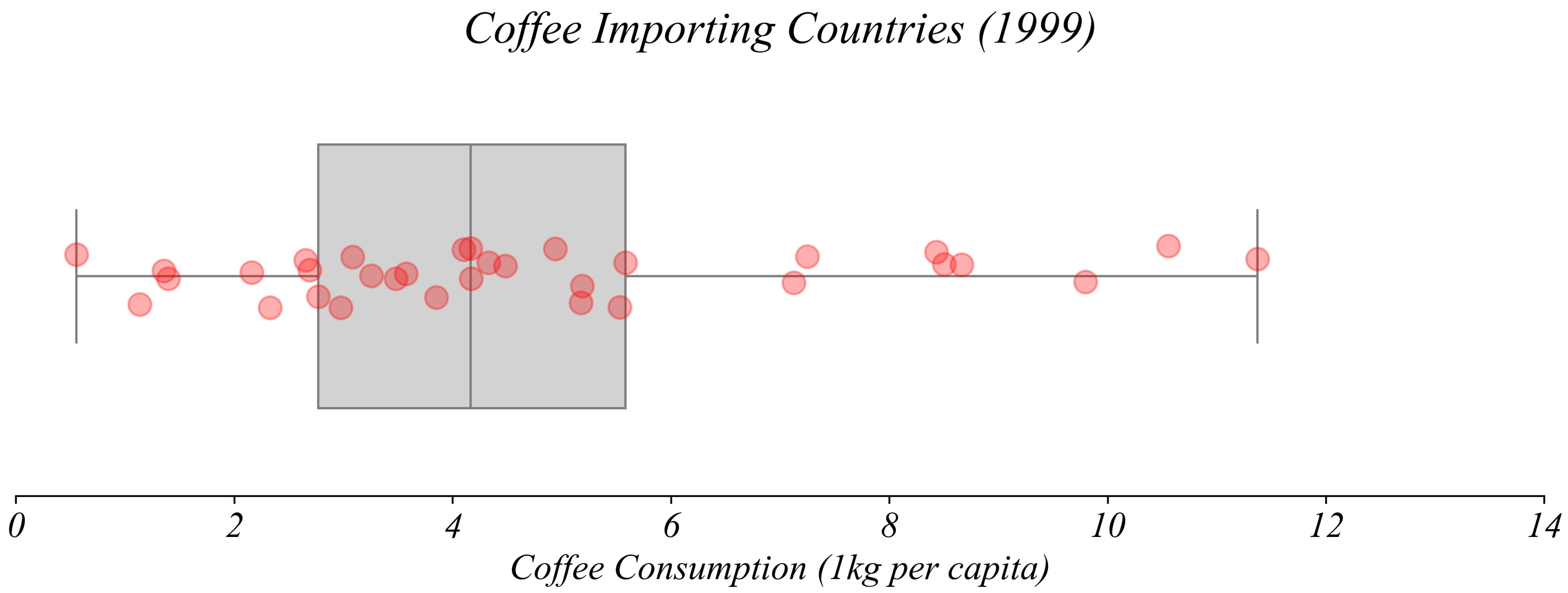
Which country consumes more than exactly 25% of countries?



> Slovakia!

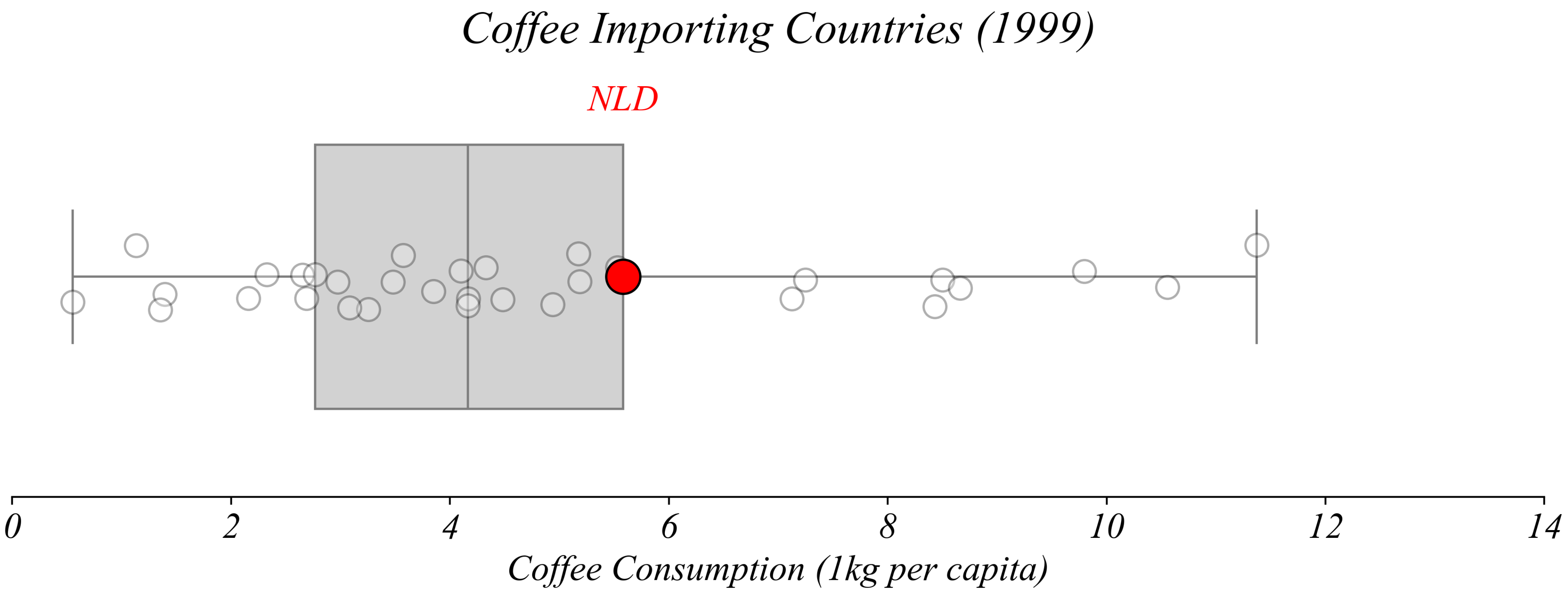
Boxplots + Stripplots

Which country consumes more than exactly 75% of countries?



Boxplots + Stripplots

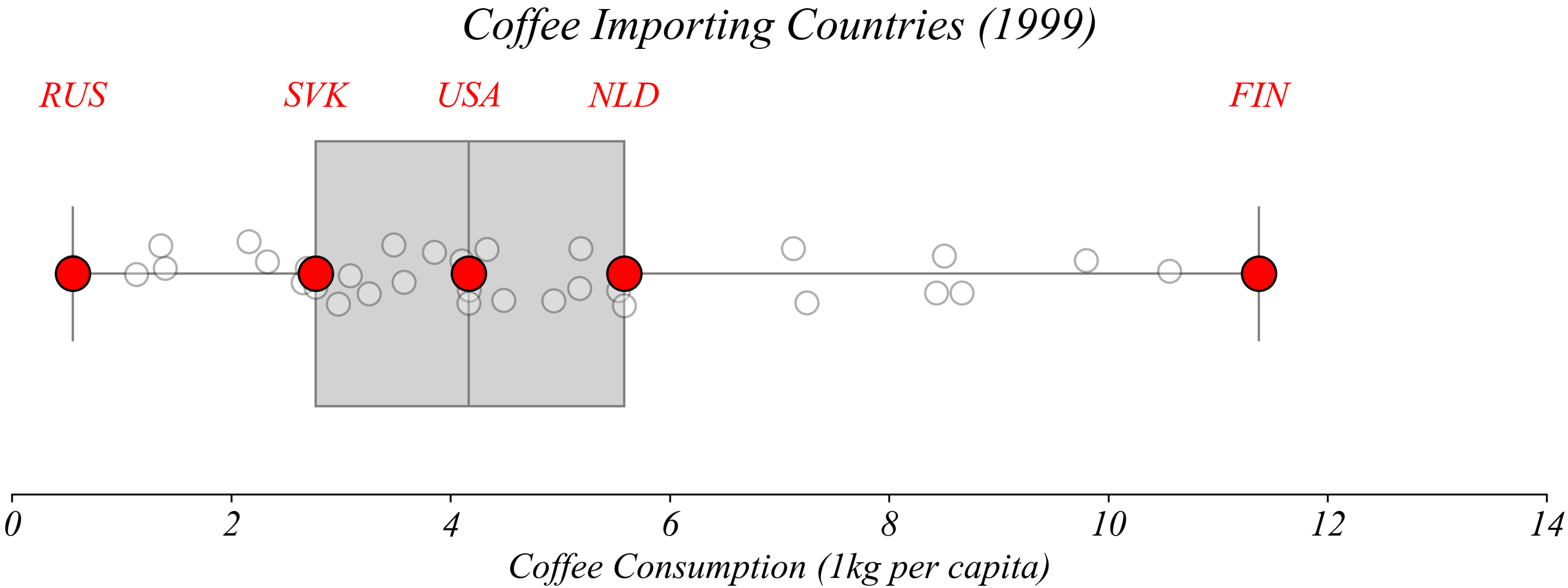
Which country consumes more than exactly 75% of countries?



> Netherlands

Boxplots + Stripplots

Boxplots show quartiles; stripplots show the data.



Boxplots + Stripplots: Summary

Boxplots show quartiles; stripplots show the data.

- *Boxplots make it easy to show the quartiles.*
- *Stripplots can show the distribution of the data.*
- *We can highlight subsets of the data.*

S-T-E for Boxplots + Stripplots

What we just did

Step	**Action**
SELECT	All coffee-importing countries in 1999
TRANSFORM	Calculate quartiles (min, Q1, median, Q3, max)
ENCODE	Quartile → box position; Value → point position

> *TRANSFORM for boxplots = calculate quartiles*

Exercise 1.2 | Boxplots + Stripplots

Show the distribution of coffee consumption per capita in 2019.

Lets use a boxplot and stripplot to examine the distribution of coffee consumption per capita among coffee-importing countries in 2019.

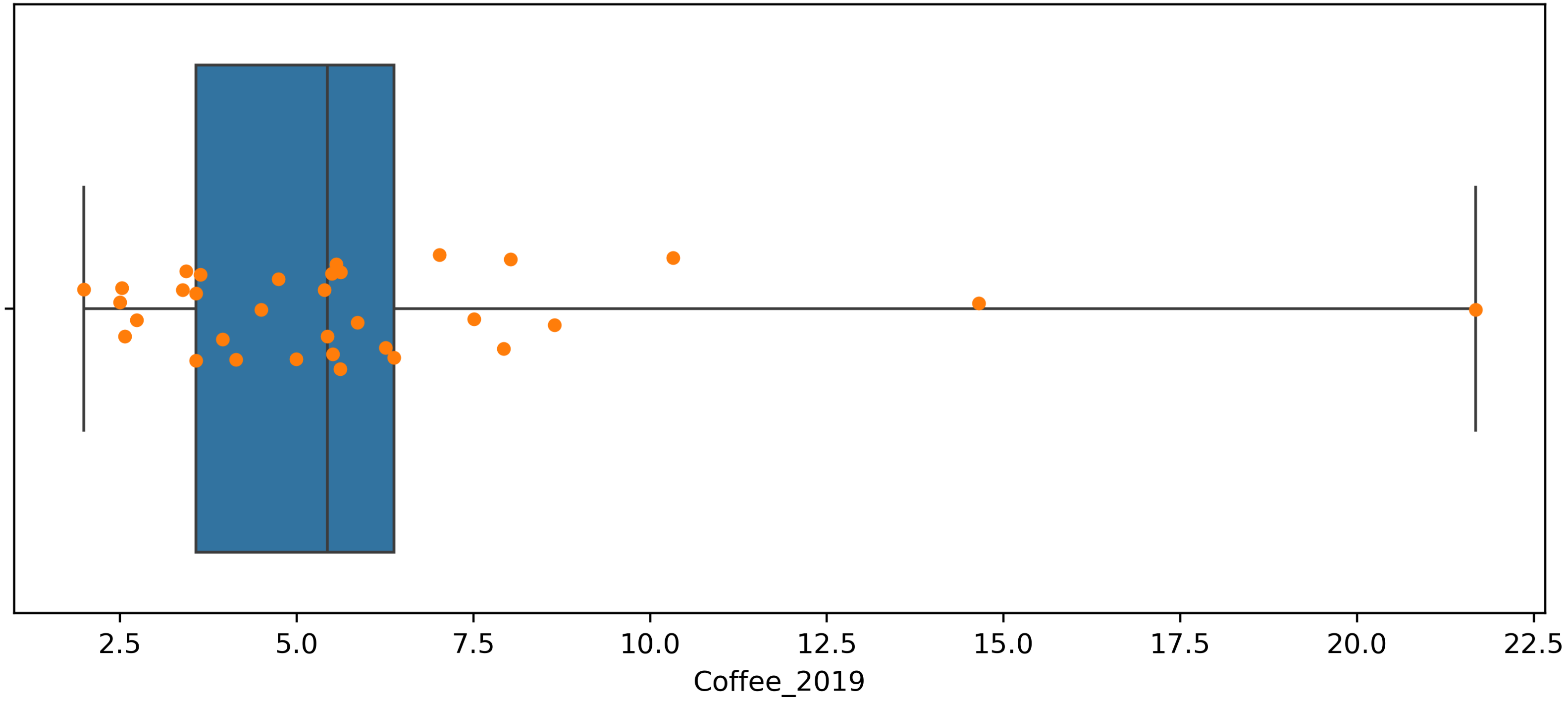
- ***Data:*** `Coffee_Per_Cap_2019.csv`

Exercise 1.2 | Boxplots + Stripplots

Show the distribution of coffee consumption per capita in 2019.

```
# Boxplot with no outliers
sns.boxplot(coffee, x='Coffee_2019', whis=(0,100))
```

```
# Stripplot
sns.stripplot(coffee, x='Coffee_2019')
```



```
# Save Figure
plt.savefig('exercise_1_2_boxplot.png')
```

Exercise 1.2 | Quartiles

Calculate the five-number summary

```python
# Minimum and Maximum
coffee['Coffee_2019'].min()
coffee['Coffee_2019'].max()
```

```python
# Quartiles (Q1, Median, Q3)
coffee['Coffee_2019'].quantile(0.25)
coffee['Coffee_2019'].median()
coffee['Coffee_2019'].quantile(0.75)
```

> These five numbers define the boxplot: min, Q1, median, Q3, max

Building Blocks

What this unit adds to your toolkit

Block	**New in 1.2**
Variables	Numerical
Structures	Cross-section
Operations	Bin, Mean, SD, Quartiles
Visualizations	Histogram, Boxplot, Stripplot

> Next: lets add a time dimension