

ECON 0150 | Economic Data Analysis

The economist's data analysis pipeline.

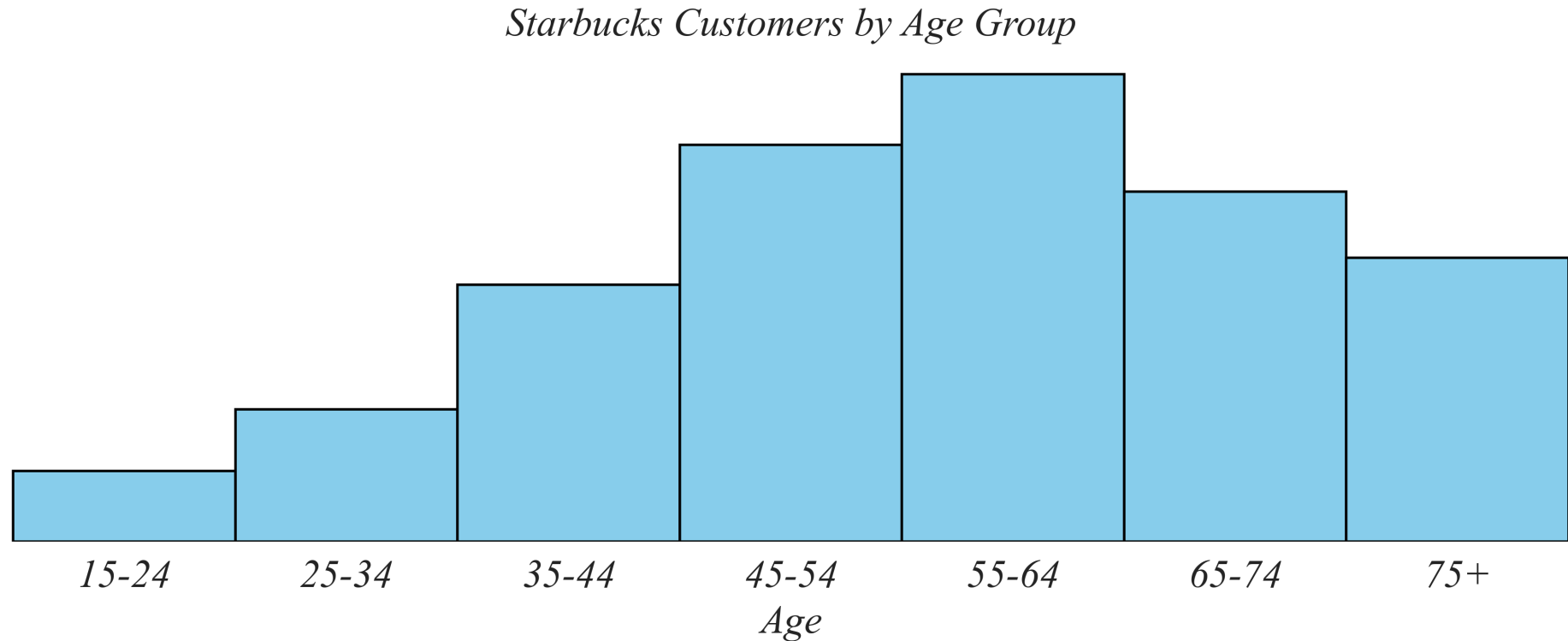
Part 1.2 | Summarizing Numerical Variables

Summarizing Numerical Variables

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

Numerical Variables: Histograms

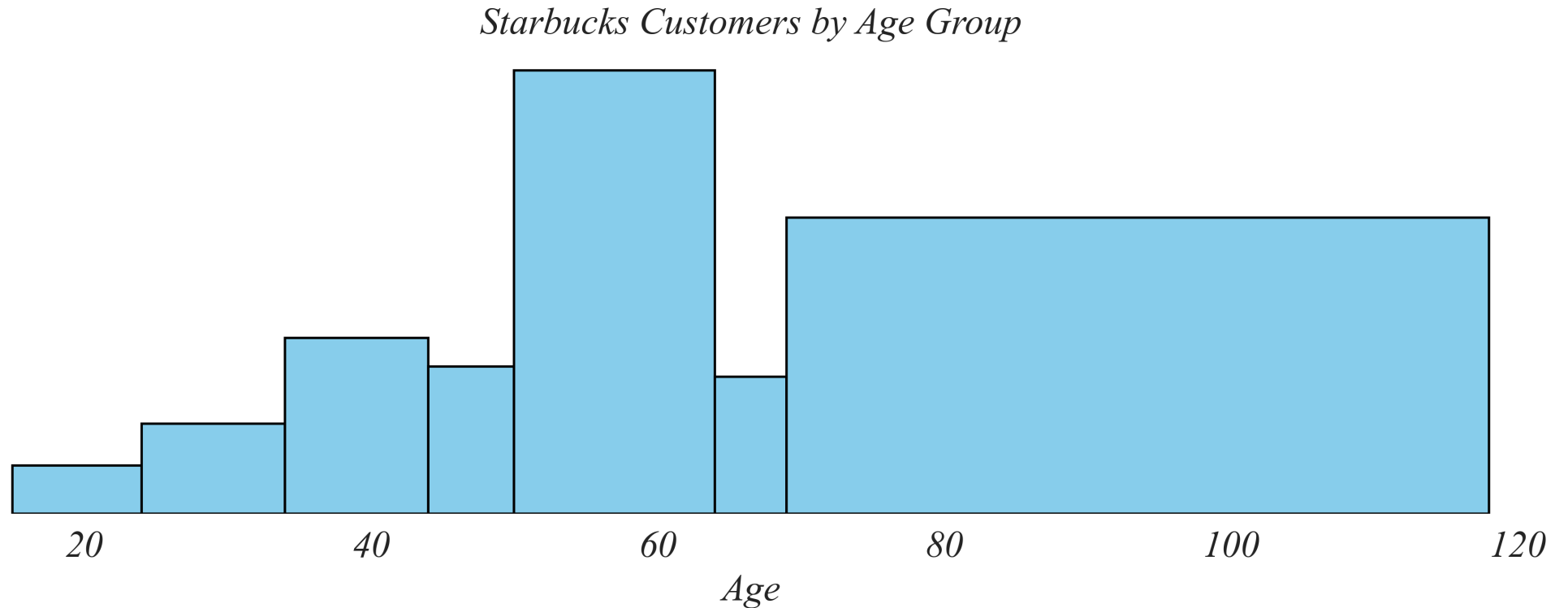
Q. Which age group has the most Starbucks customers?



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

Numerical Variables: Histograms

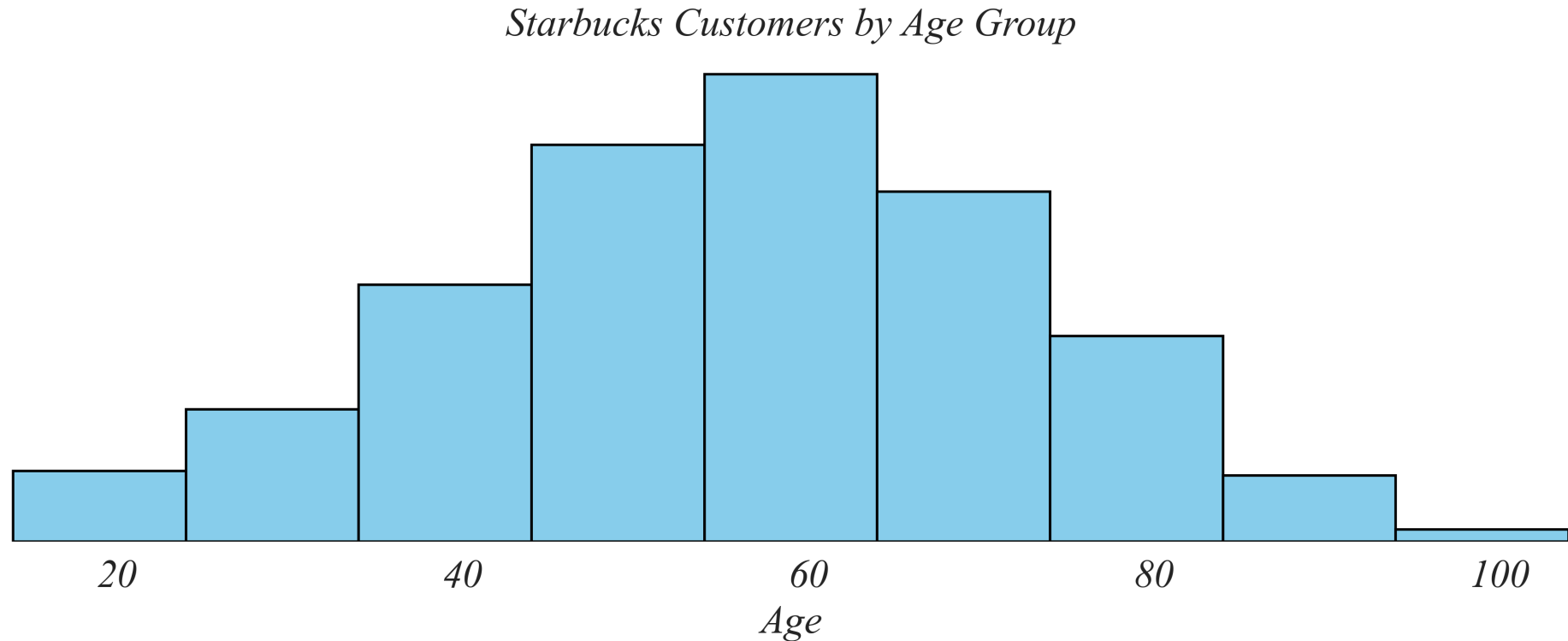
Q. Which age group has the most Starbucks customers?



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

Histograms: Use equal sized bins

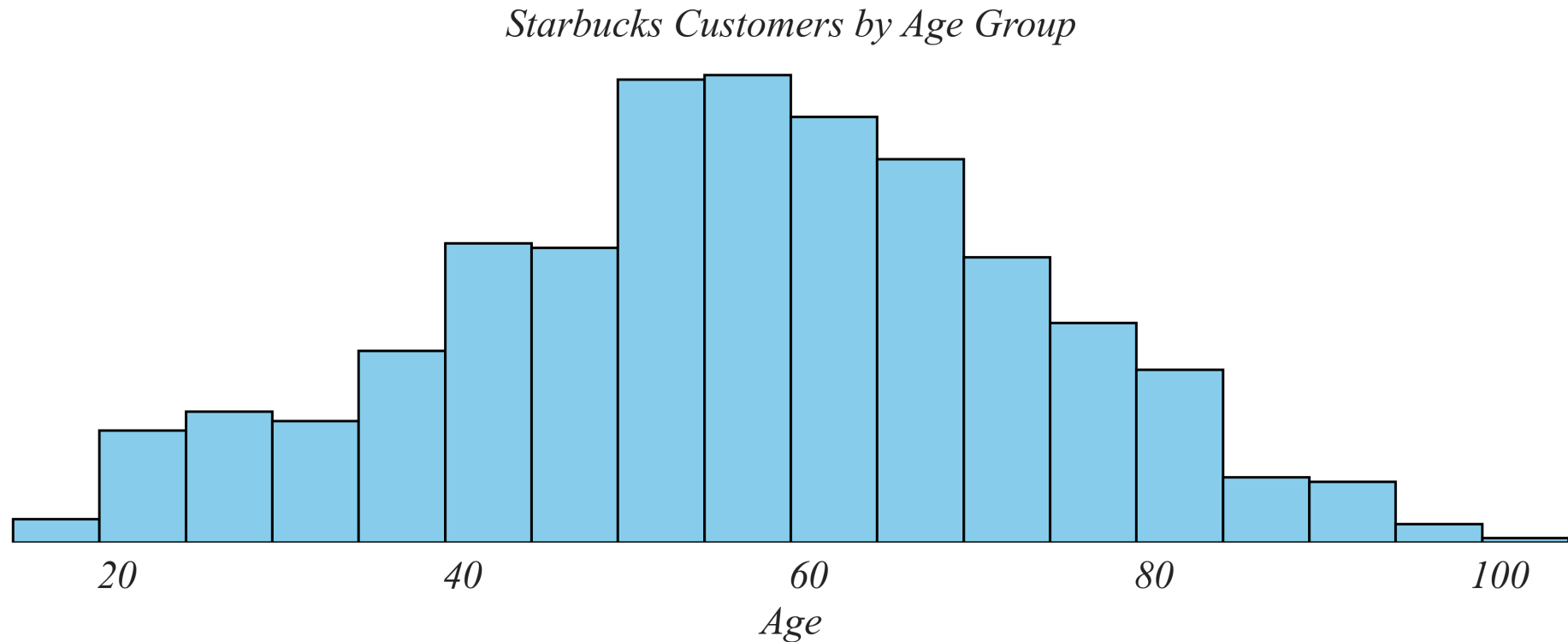
Q. Which age group has the most Starbucks customers?



> 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?

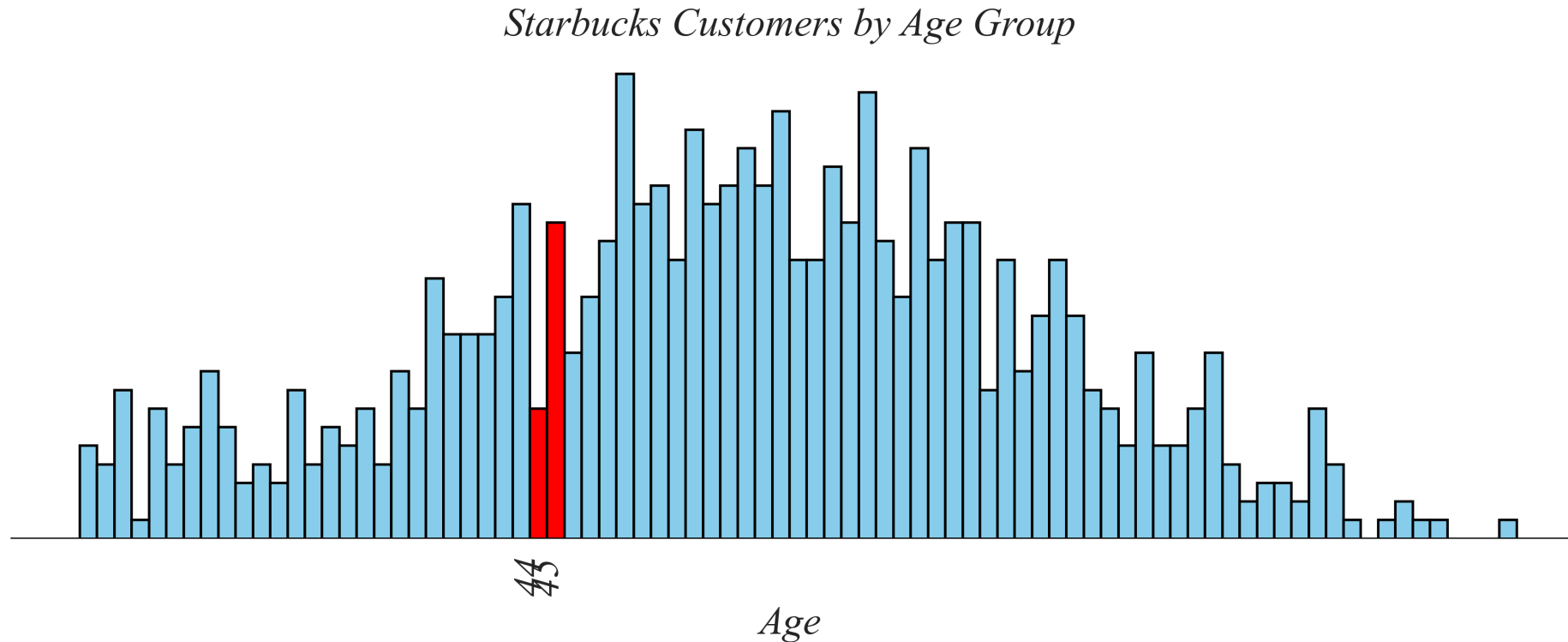


> *what if we take this even further?*

> *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?

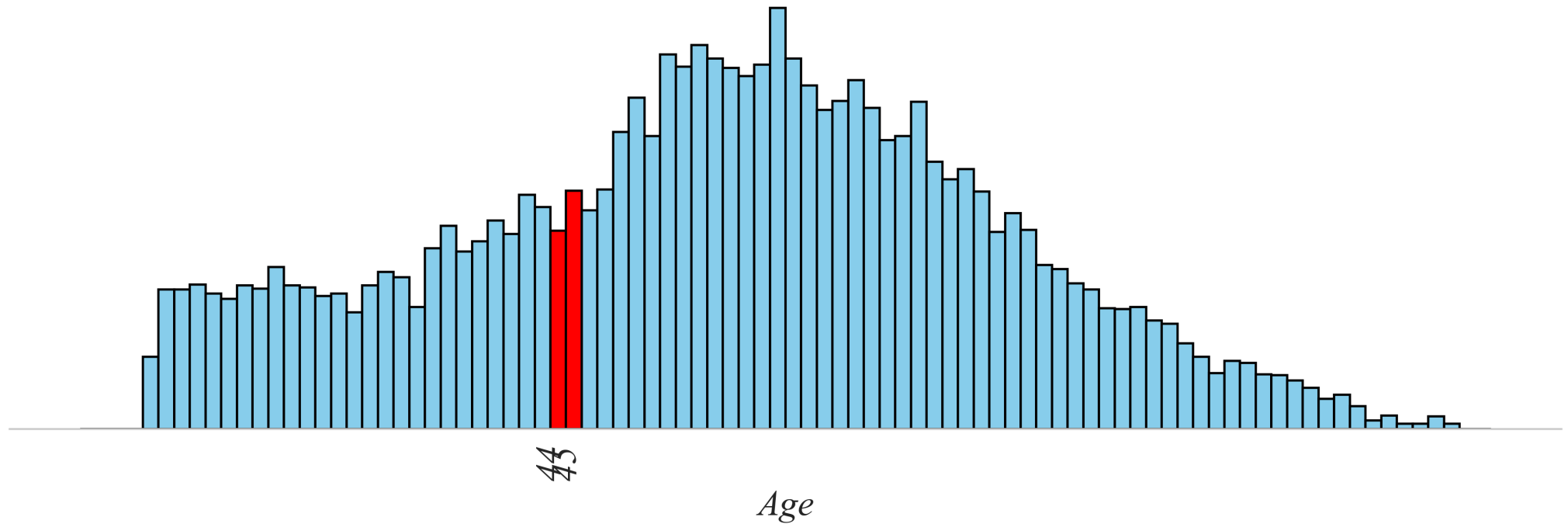


- > *we can go too far, introducing statistical noise. how do we fix the problem?*
- > *increase the sample size or the bin width!*

Histograms: Balance resolution vs noise

Q. Which age group has the most Starbucks customers?

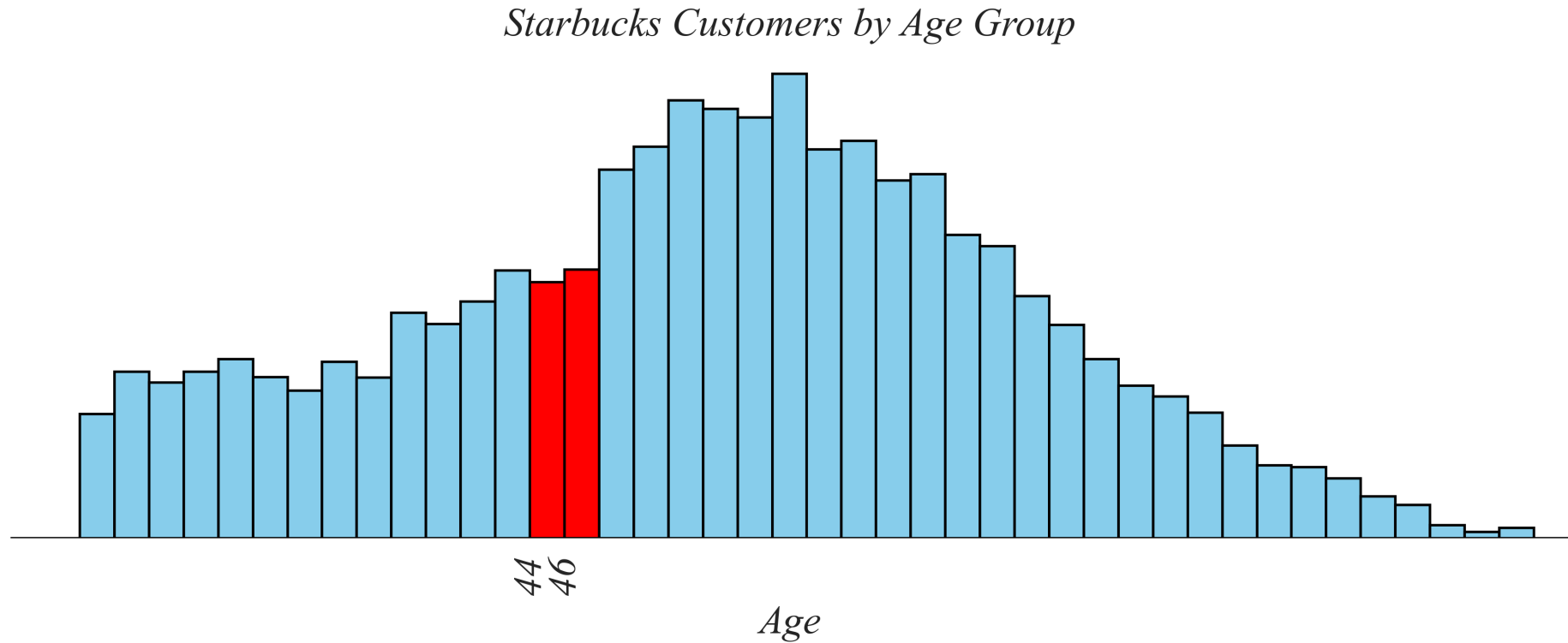
Starbucks Customers by Age Group



> 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!

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.*

Exercise 1.2: Histograms

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

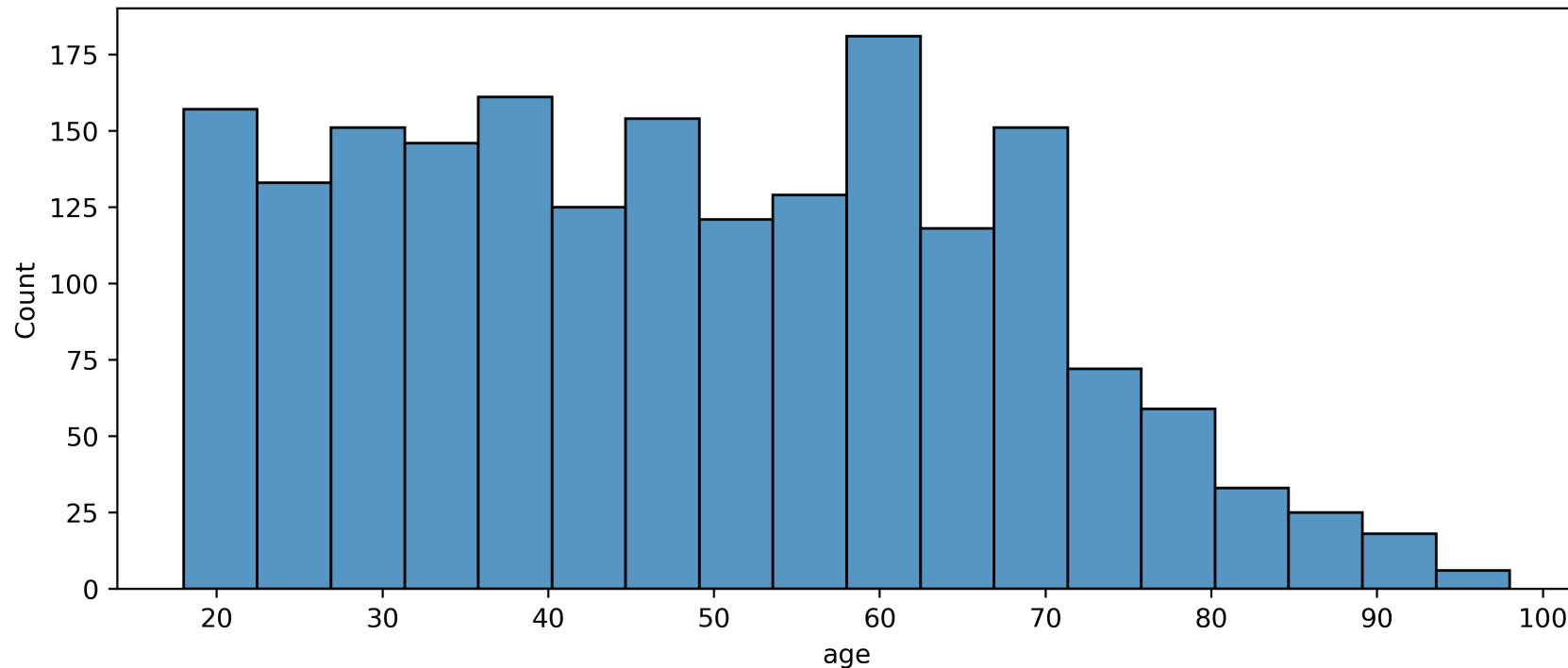
Lets use the data to examine whether customers between 45 - 55 years old spend the most among customers making less than \$40k.

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?

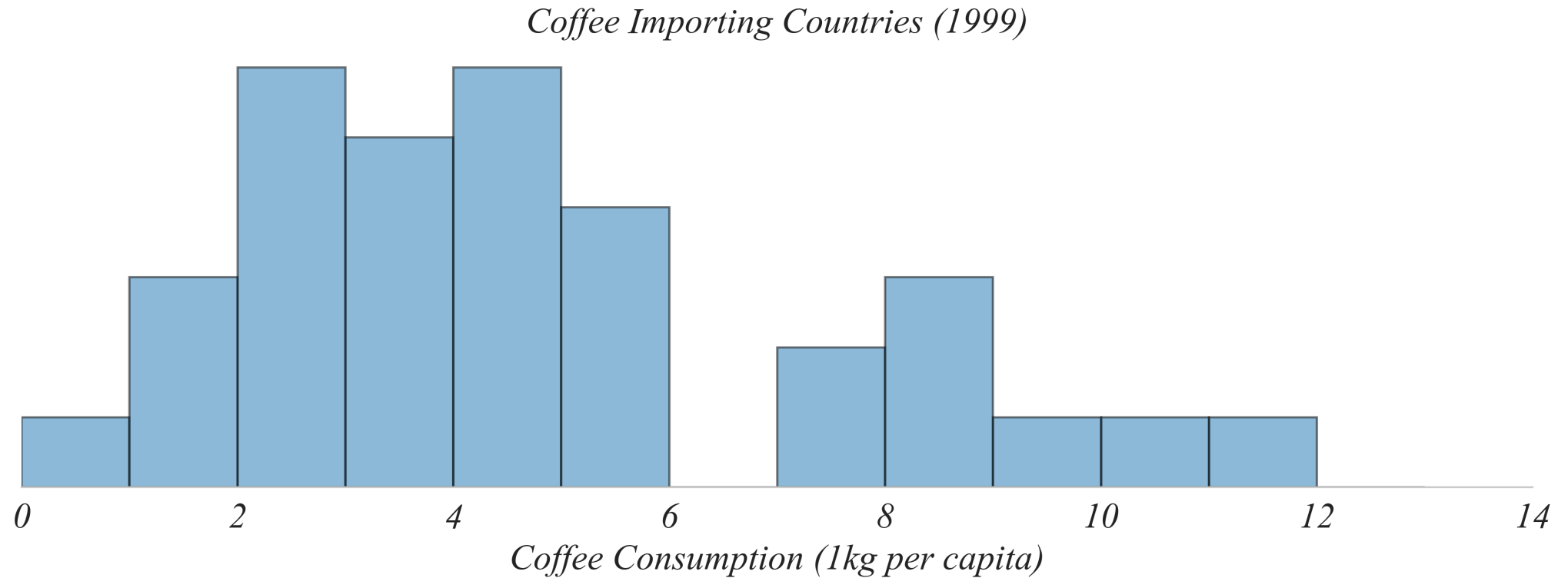
```
1 # Histogram with 5 year bins
2 sns.histplot(customers, x='age', bins=range(20,100,5))
```



```
1 # Save Figure
2 plt.savefig('exercise_1_2_1.png')
```

Numerical Variables: Histograms

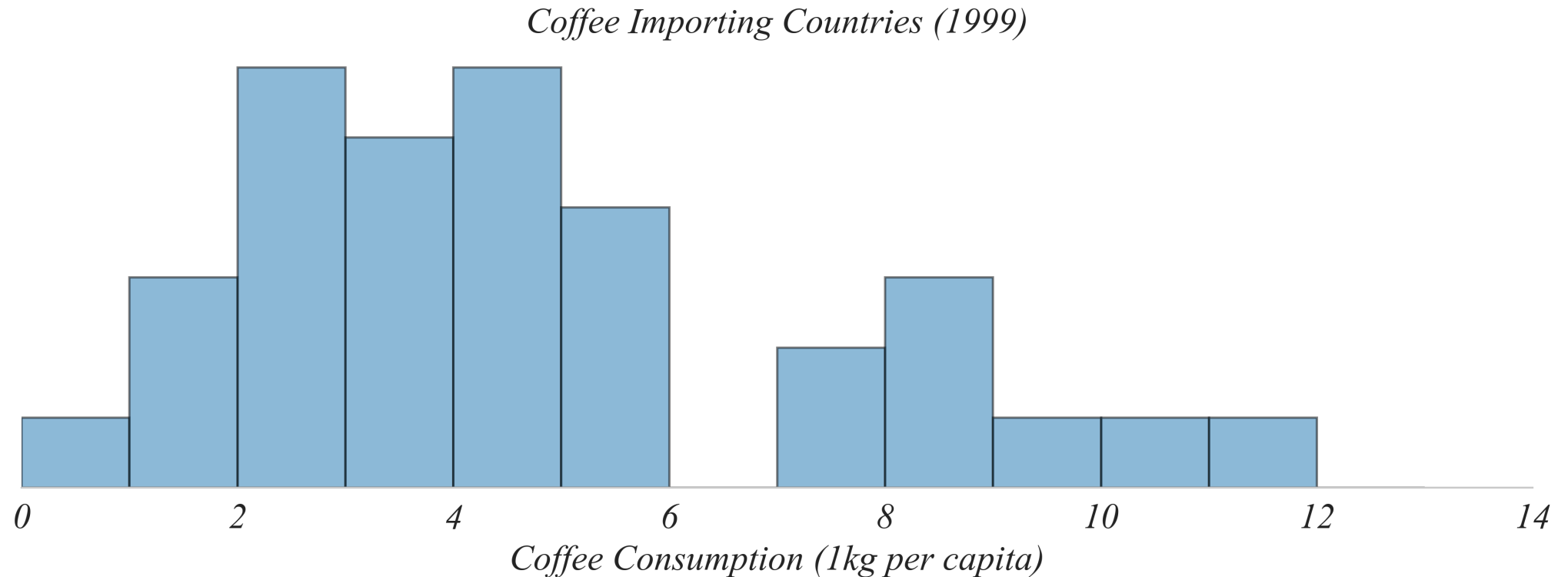
Q. Which countries drank an average amount of coffee?



> *histogram bins make it impossible to see the exact values*

Numerical Variables: Histograms

Q. Which countries drank the most coffee in 1999?

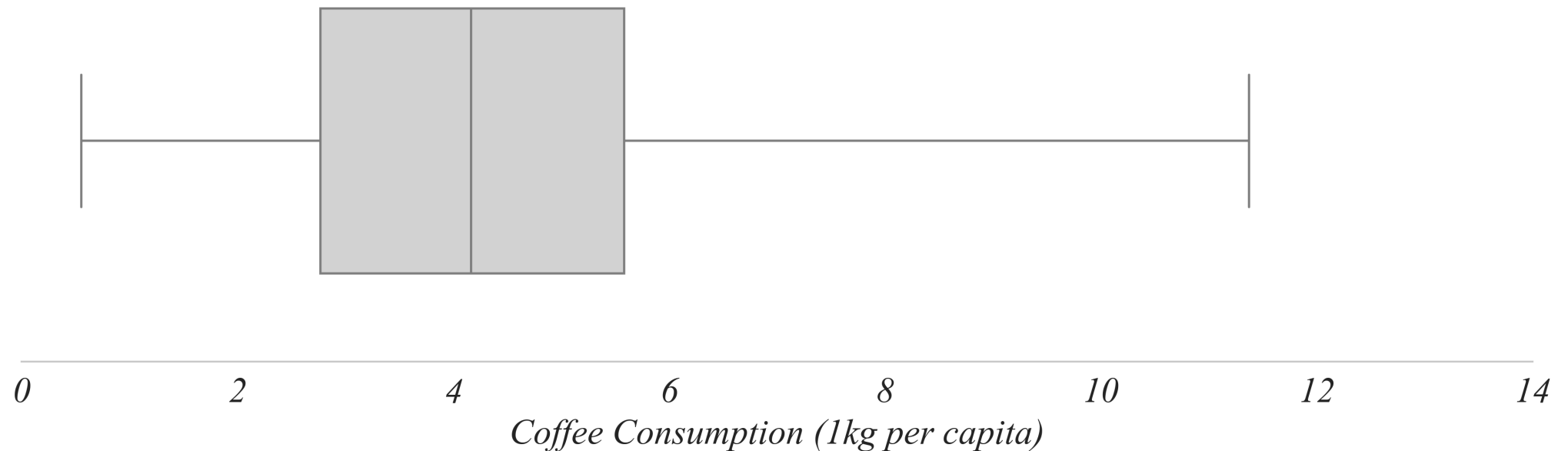


> *again, histograms make it difficult to see statistical measures*

Numerical Variables: Boxplots

Q. Which countries drank the most coffee in 1999?

Coffee Importing Countries (1999)

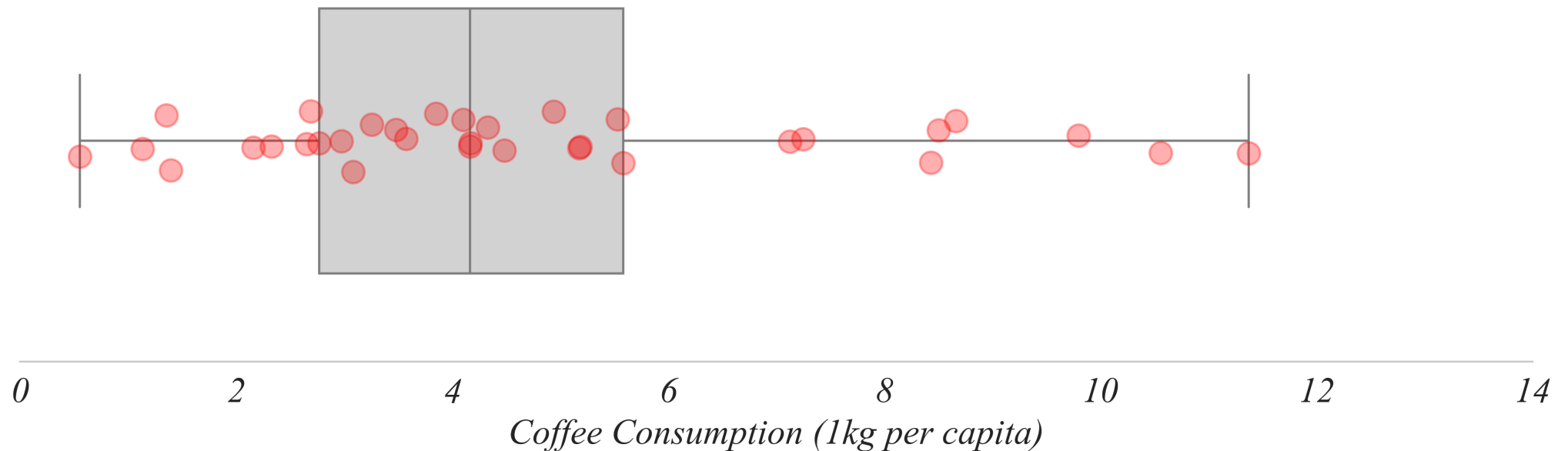


- > *as we'll see, boxplots can tell us about quartiles*
- > *but boxplots are still pretty unclear for our question*

Boxplots + Stripplots

Q. Which countries drank the most coffee in 1999?

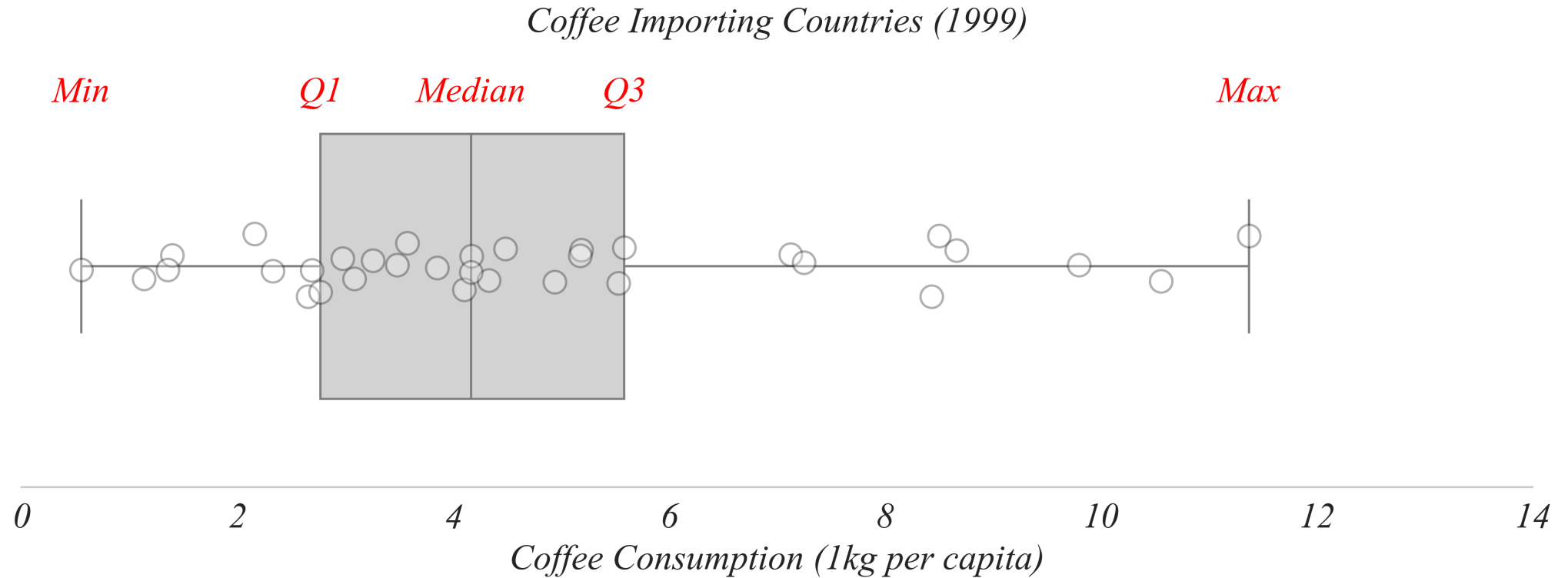
Coffee Importing Countries (1999)



- > *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?

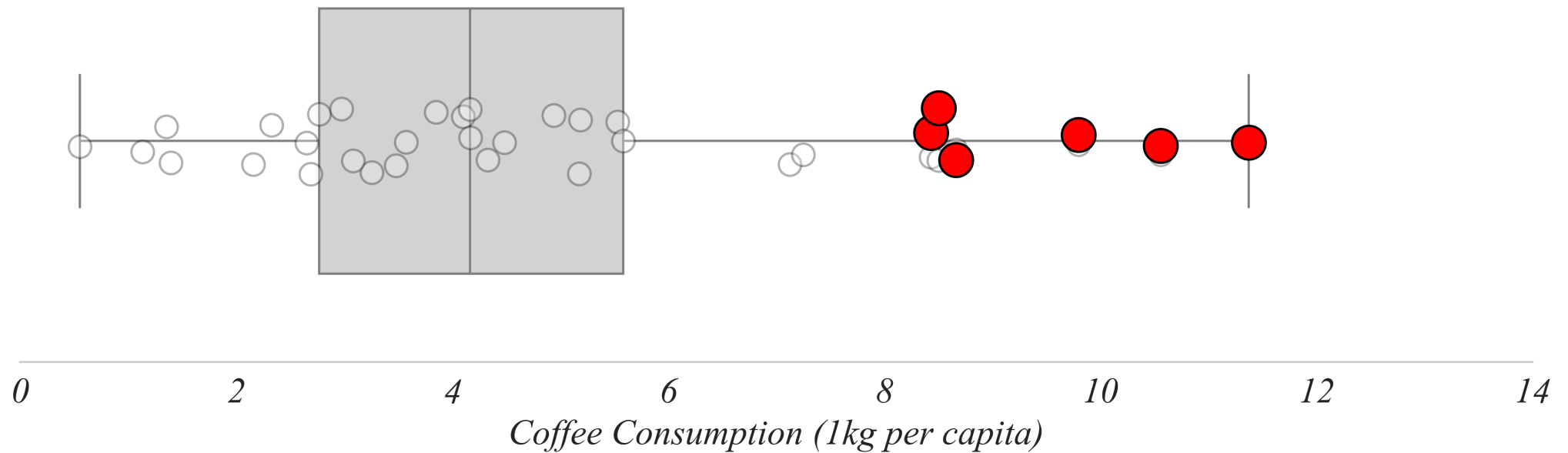
Coffee Importing Countries (1999)



Boxplots + Stripplots

Which countries consumed more than 8 kg per capita?

Coffee Importing Countries (1999)

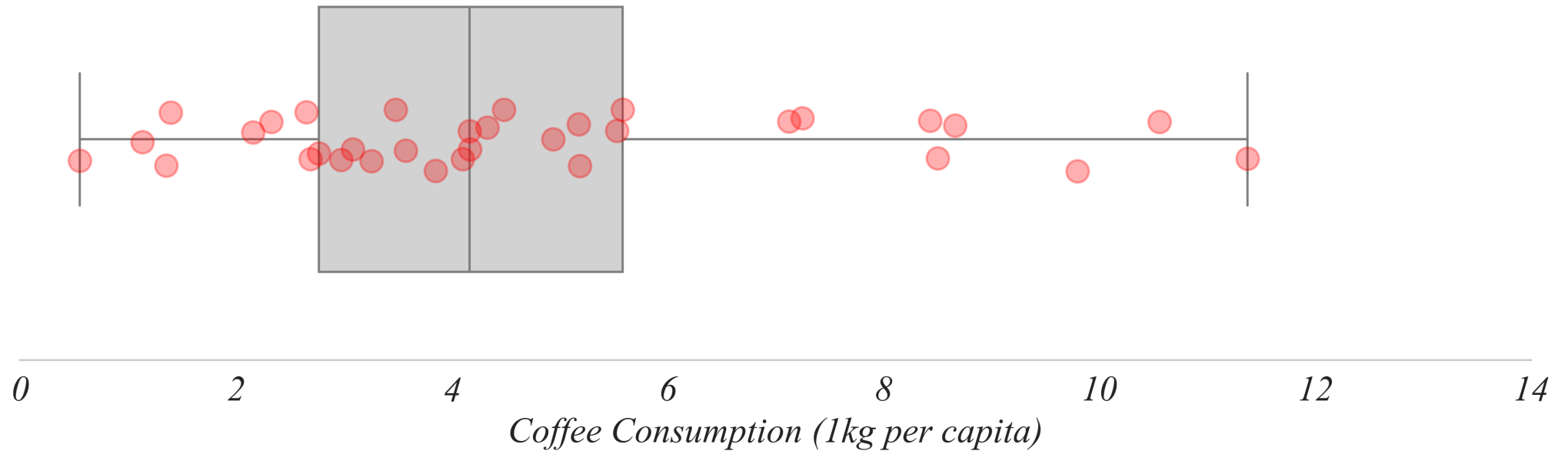


> *we can highlight the relevant subsets of the data*

Boxplots + Stripplots

Which country consumed the most coffee per capita?

Coffee Importing Countries (1999)

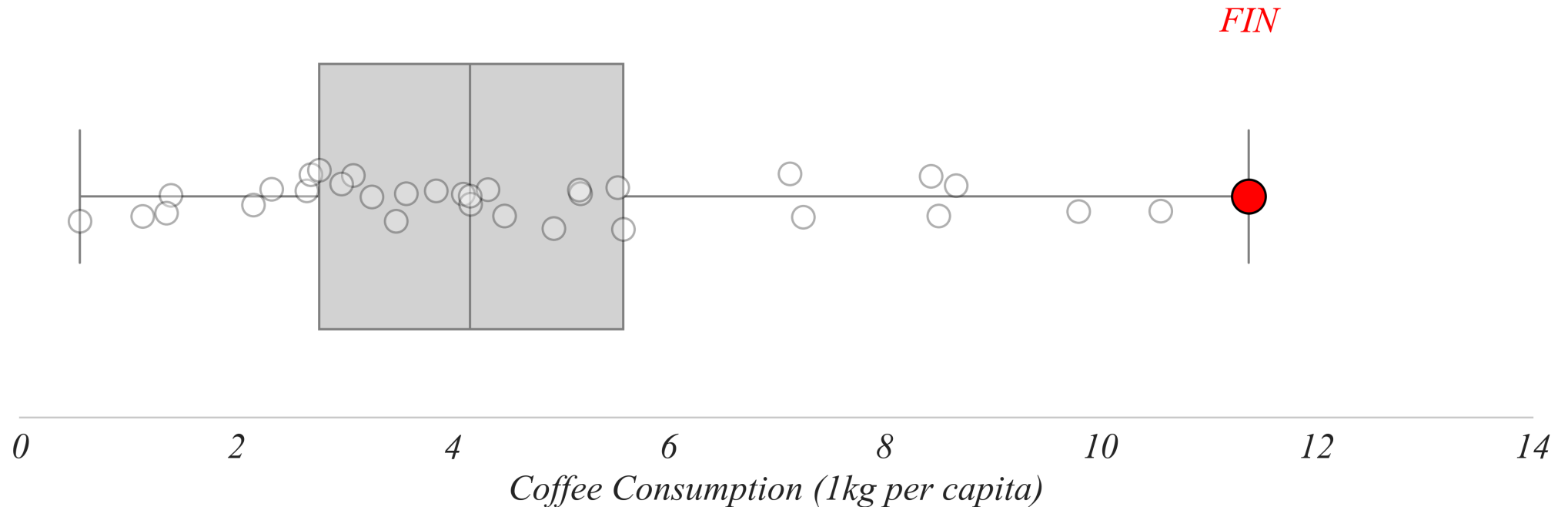


> we can find the exact values according to quartiles

Boxplots + Stripplots

Which country consumed the most coffee per capita?

Coffee Importing Countries (1999)

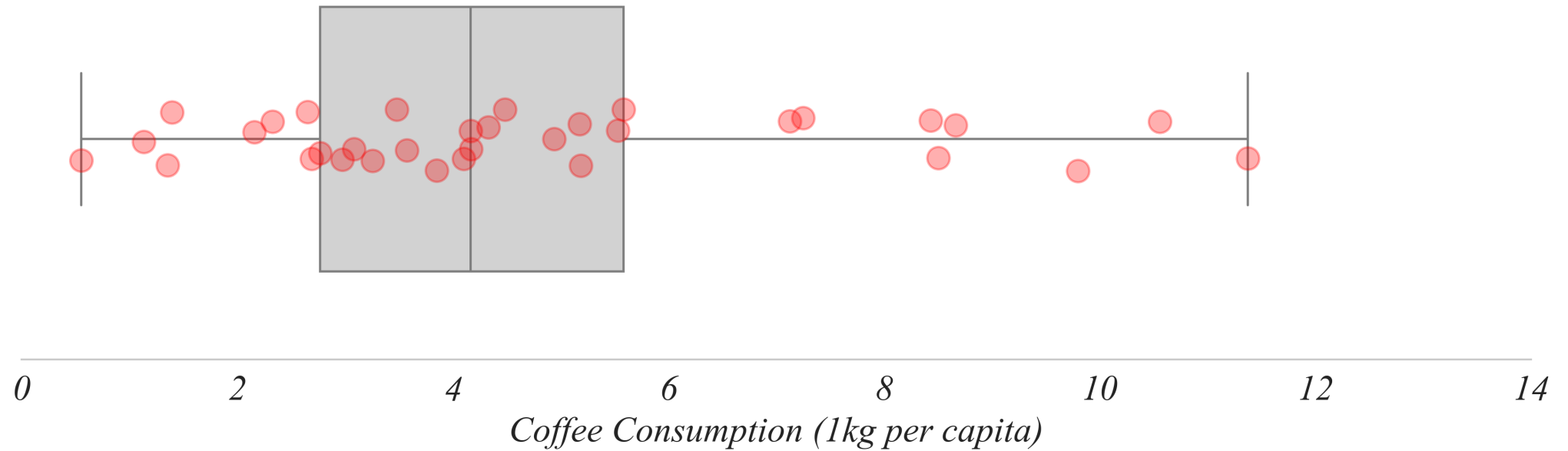


- > *we can find the exact values according to quartiles*
- > *Finland consumed the most coffee per capita in 1999*

Boxplots + Stripplots

Which country consumed the least coffee per capita?

Coffee Importing Countries (1999)

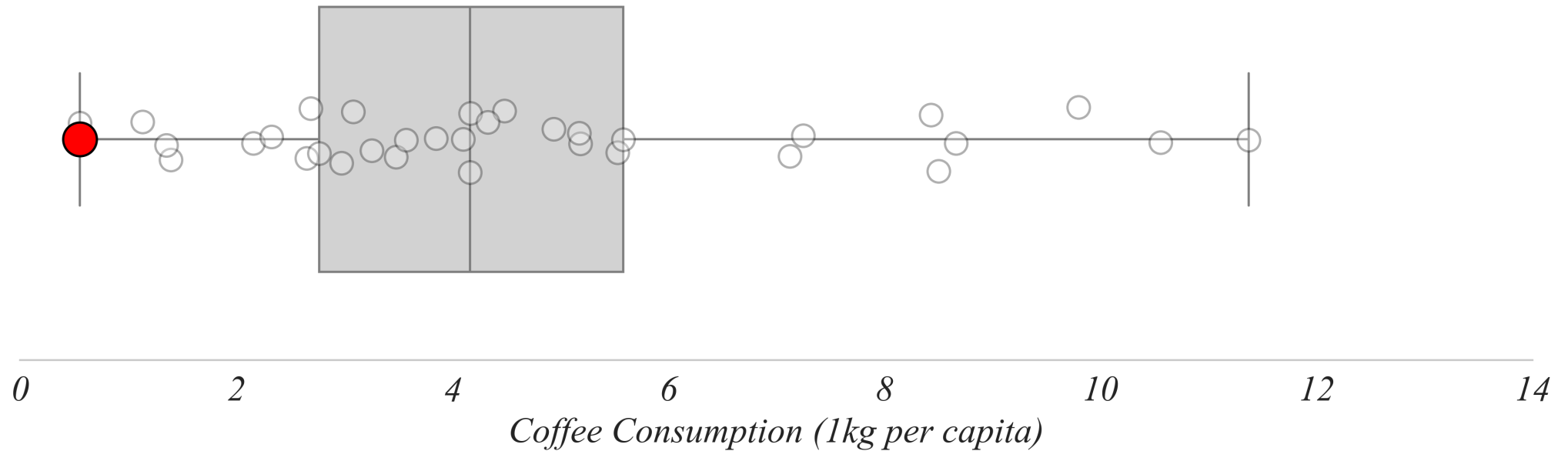


Boxplots + Stripplots

Which country consumed the least coffee per capita?

Coffee Importing Countries (1999)

RUS

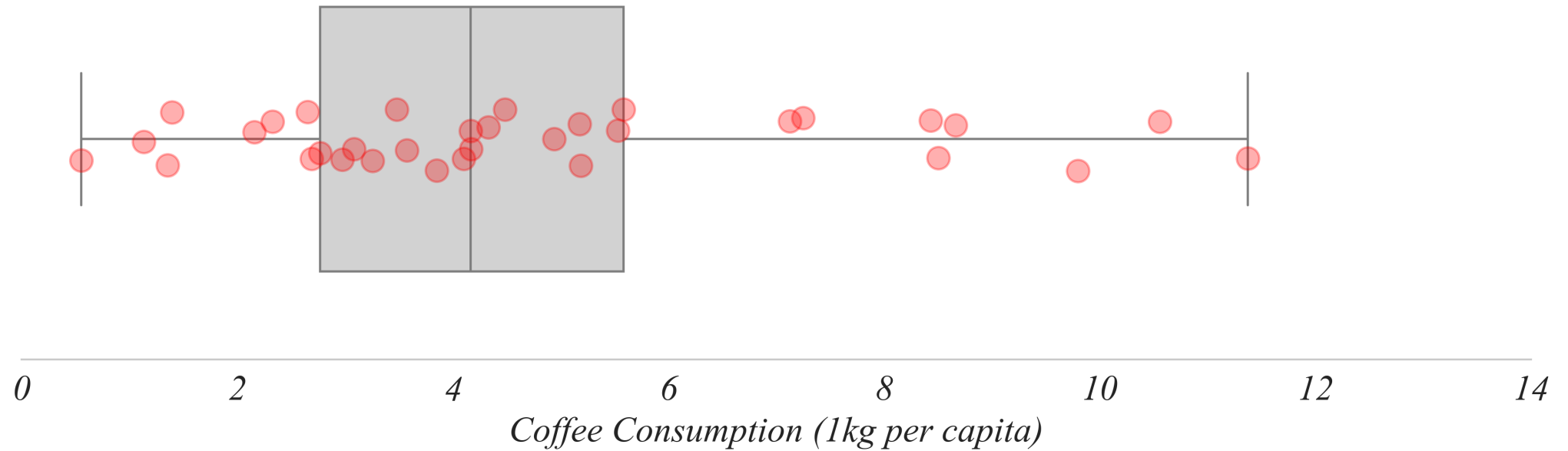


> Russia consumed the least coffee per capita in 1999

Boxplots + Stripplots

How about the median?

Coffee Importing Countries (1999)

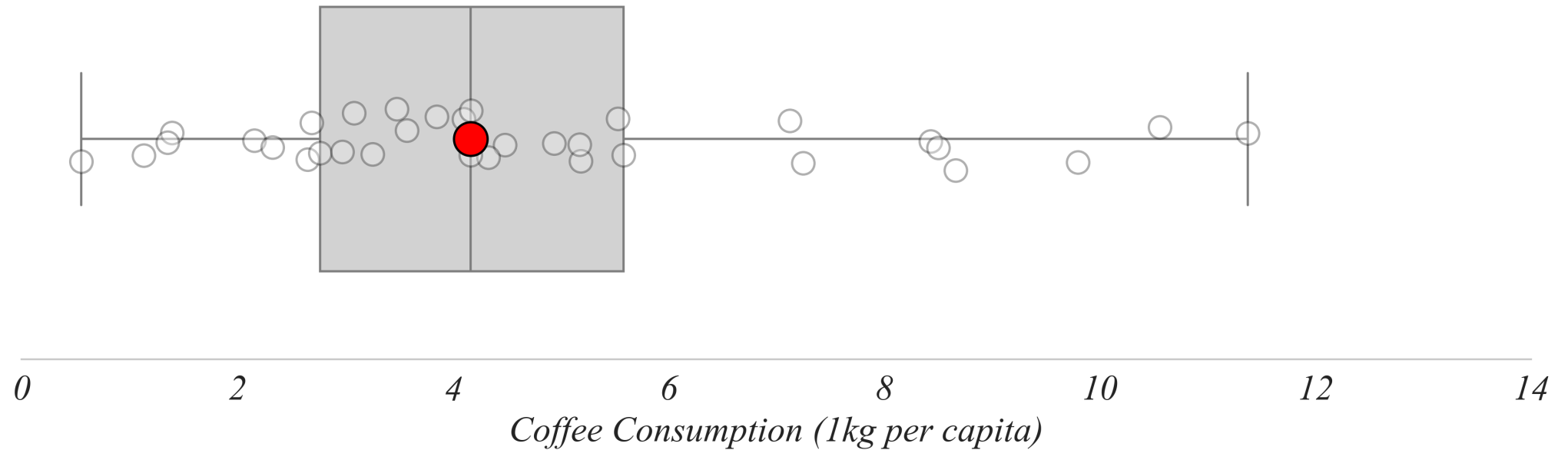


Boxplots + Stripplots

How about the median?

Coffee Importing Countries (1999)

USA

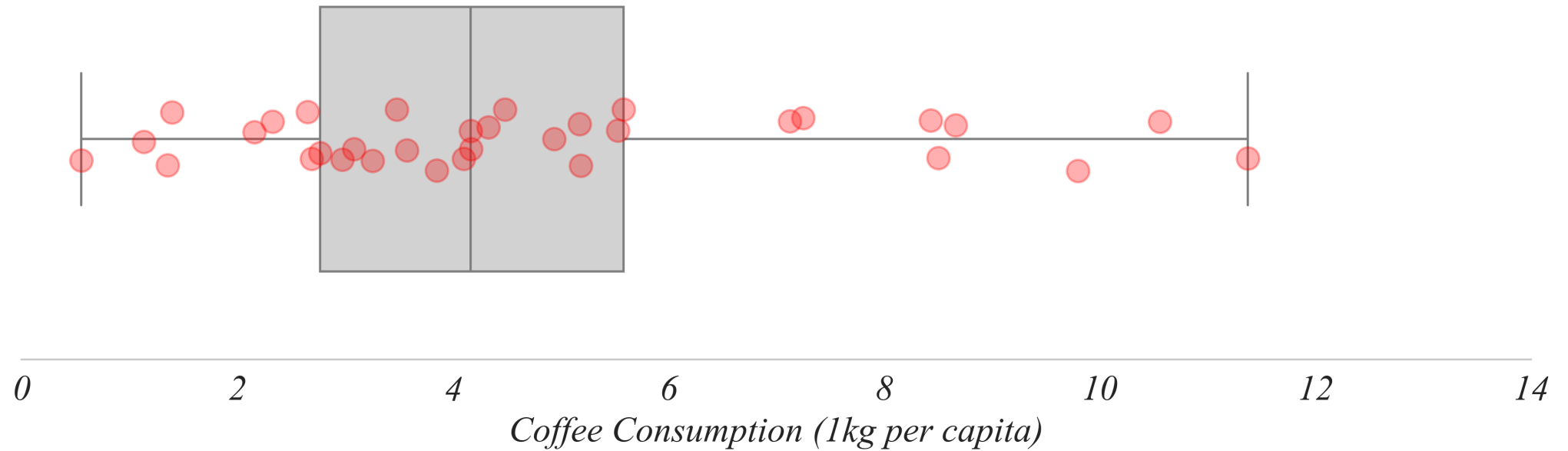


> the US!

Boxplots + Stripplots

Which country consumes more than exactly 25% of countries?

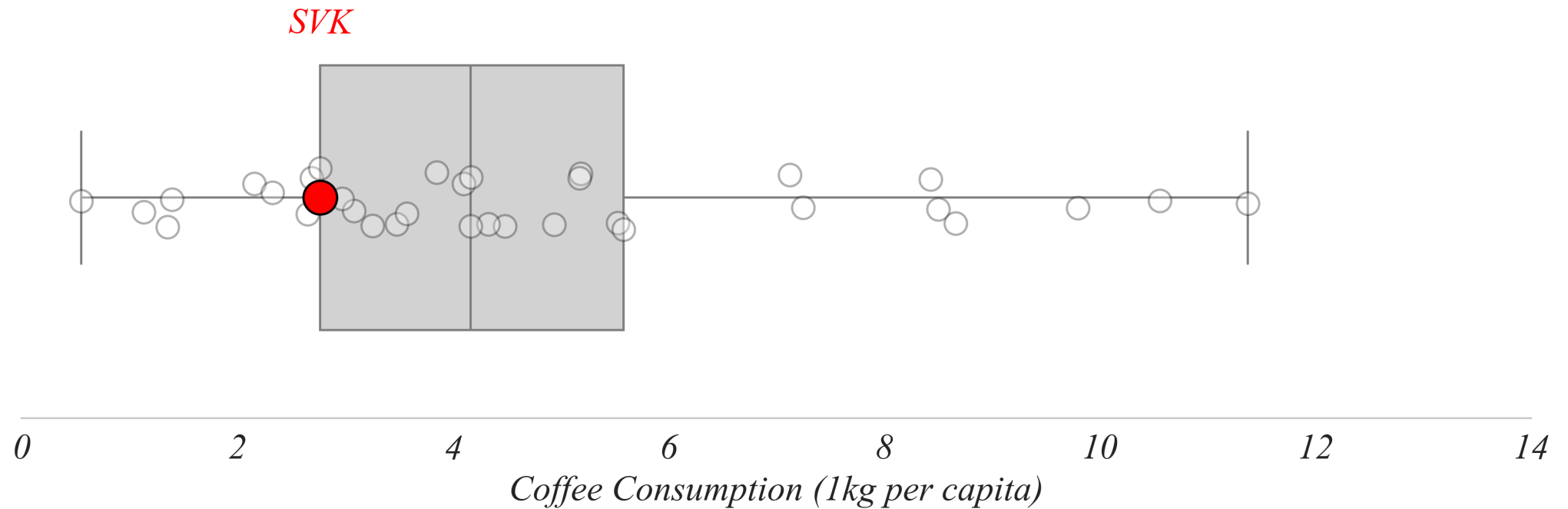
Coffee Importing Countries (1999)



Boxplots + Stripplots

Which country consumes more than exactly 25% of countries?

Coffee Importing Countries (1999)

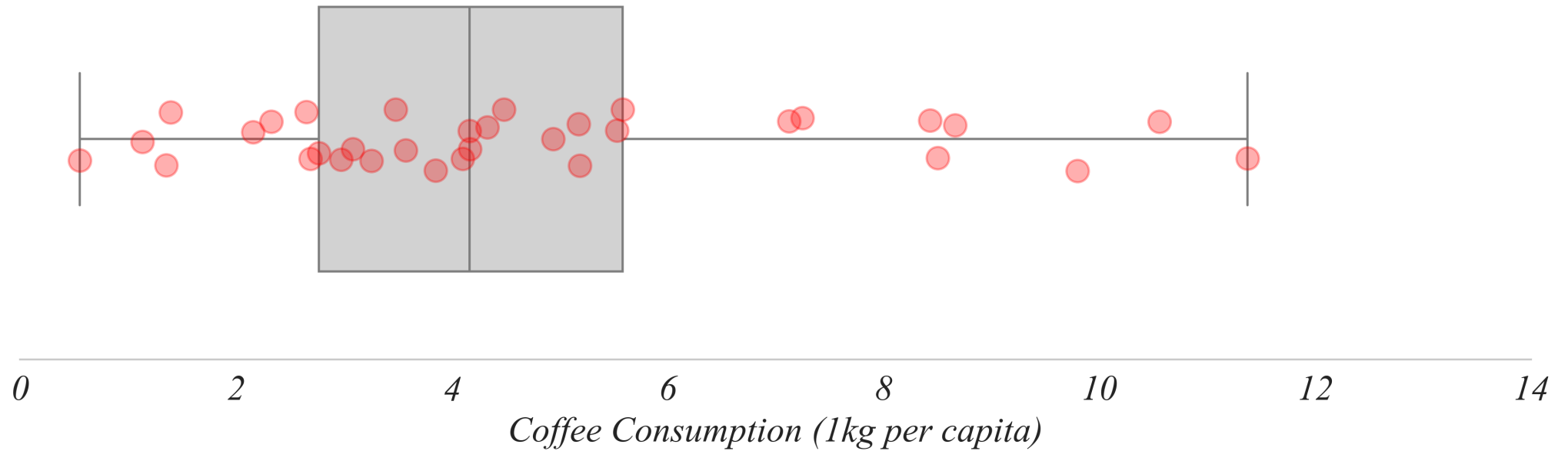


> *Slovakia!*

Boxplots + Stripplots

Which country consumes more than exactly 75% of countries?

Coffee Importing Countries (1999)

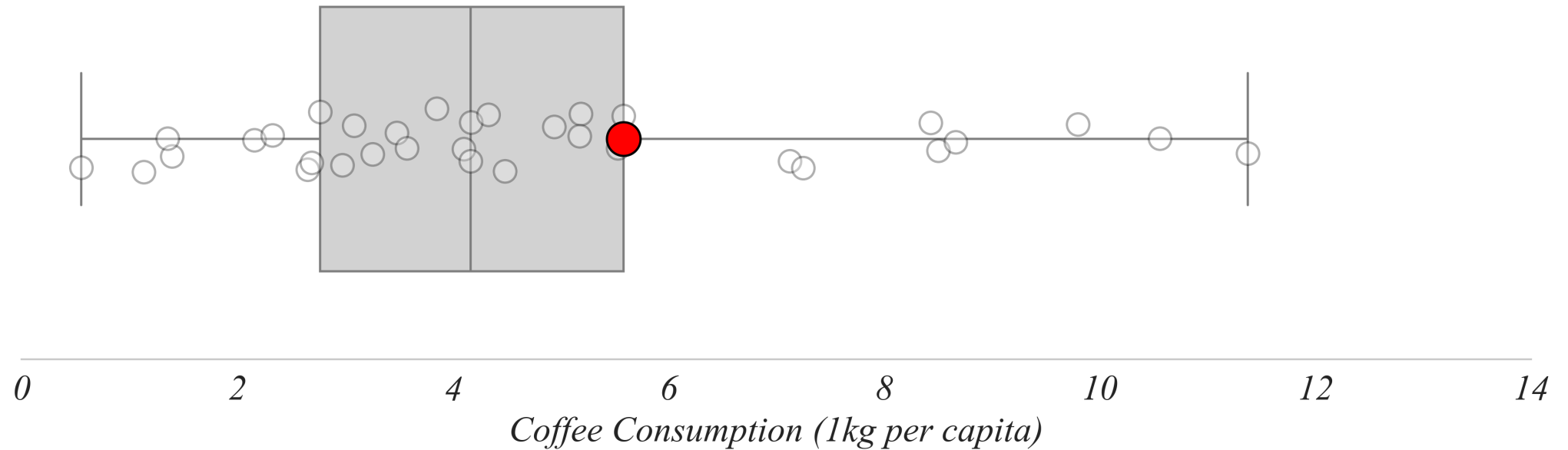


Boxplots + Stripplots

Which country consumes more than exactly 75% of countries?

Coffee Importing Countries (1999)

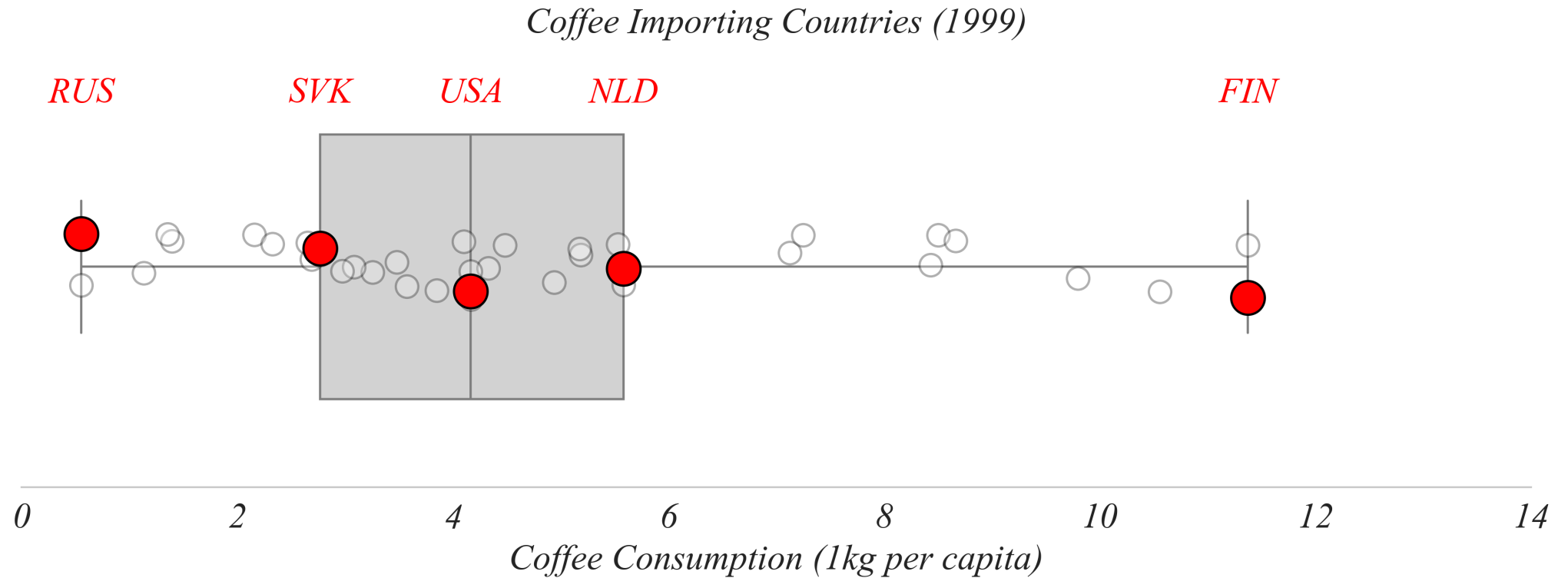
NLD



> 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.*

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.

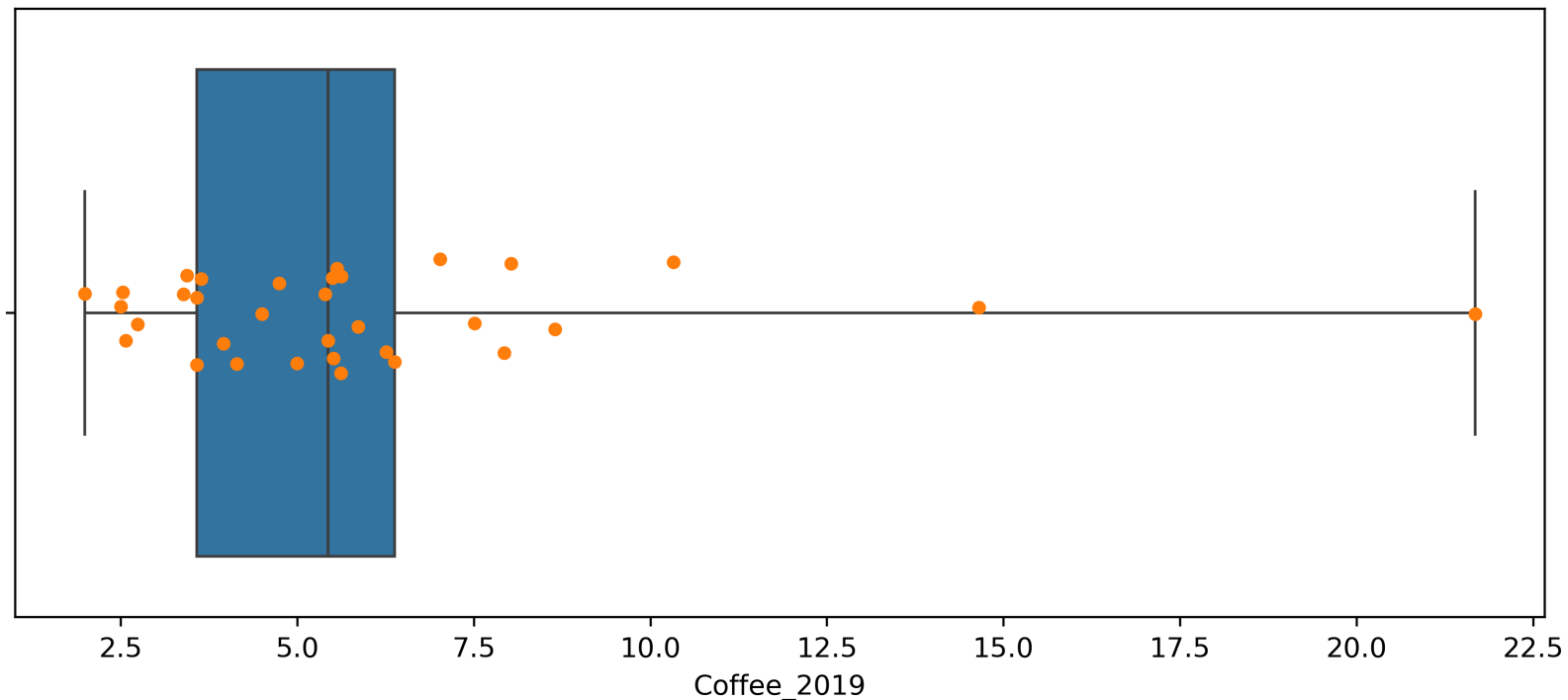
- ***Data:*** *Coffee_Per_Cap_2019.csv*

Exercise 1.2: Boxplots + Stripplots

Show the distribution of coffee consumption per capita in 2019.

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

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



```
1 # Save Figure
2 plt.savefig('exercise_1_2_2.png')
```