

Finding Outliers Using IQR - Step 3/4

Learn / Courses / Introduction to Statistics In Python

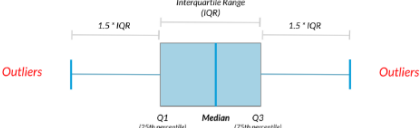
← Course Outline →

Daily XP 1006

Exercise

Finding outliers using IQR

Outliers can have big effects on statistics like mean, as well as statistics that rely on the mean, such as variance and standard deviation. Interquartile range, or IQR, is another way of measuring spread that's less influenced by outliers. IQR is also often used to find outliers. If a value is less than $Q1 - 1.5 \times IQR$ or greater than $Q3 + 1.5 \times IQR$, it's considered an outlier. In fact, this is how the lengths of the whiskers in a `matplotlib` box plot are calculated.



In this exercise, you'll calculate IQR and use it to find some outliers. `pandas` as `pd` and `numpy` as `np` are loaded and `food_consumption` is available.

Instructions 3/4 25 XP 3 4

- Calculate the lower and upper cutoffs for outliers of `emissions_by_country`, and store these as `lower` and `upper`.

Take Hint (-7 XP)

script.py

Light Mode

```
1 # Calculate total co2_emission per country: emissions_by_country
2 emissions_by_country = food_consumption.groupby('country')
  ['co2_emission'].sum()
3
4 # Compute the first and third quantiles and IQR of
  emissions_by_country
5 q1 = np.quantile(emissions_by_country, 0.25)
6 q3 = np.quantile(emissions_by_country, 0.75)
7 iqr = q3 - q1
8
9 # Calculate the lower and upper cutoffs for outliers
10 lower = ----
11 upper = ----
```

Run Code Submit Answer

IPython Shell

Slides

```
Angola      412.99
Argentina   2172.48
Armenia     1189.93
...
Uruguay     1634.91
Venezuela   1184.10
Vietnam     641.51
Zambia      225.30
Zimbabwe    350.33
Name: co2_emission, Length: 130, dtype: float64
First Quartile (Q1): 446.66
Third Quartile (Q3): 1111.1525000000001
Interquartile Range (IQR): 664.4925000000001

Countries with outlier CO2 emissions:
country
Argentina    2172.4
Name: co2_emission, dtype: float64

In [1]:
```

Figure 3: Screenshot showing the calculation of lower and upper cutoffs for outliers using IQR.

Question

Outliers can have big effects on statistical measures such as mean, variance, and standard deviation. Interquartile range (IQR) is another way of measuring spread that is less influenced by outliers. In this step, you'll calculate the lower and upper cutoffs for outliers of `emissions_by_country` and store these as `lower` and `upper`.

****Instructions:****

1. Calculate the lower cutoff for outliers using $Q1 - 1.5 * IQR$.
2. Calculate the upper cutoff for outliers using $Q3 + 1.5 * IQR$.
3. Print the lower and upper cutoffs for identifying outliers.

Corrected Code Solution

```
import pandas as pd
import numpy as np

# Calculate total CO2 emissions per country
emissions_by_country = food_consumption.groupby('country')
['co2_emission'].sum()

# Compute Q1 and Q3
q1 = np.quantile(emissions_by_country, 0.25)
q3 = np.quantile(emissions_by_country, 0.75)
iqr = q3 - q1

# Calculate lower and upper cutoffs for outliers
lower = q1 - 1.5 * iqr
upper = q3 + 1.5 * iqr

# Print the results
print(f"First Quartile (Q1): {q1}")
print(f"Third Quartile (Q3): {q3}")
print(f"Interquartile Range (IQR): {iqr}")
print(f"Lower cutoff for outliers: {lower}")
print(f"Upper cutoff for outliers: {upper}")
```

Answer Explanation

1. ****Lower and Upper Cutoffs:**** The cutoffs are calculated as:
 - Lower = $Q1 - 1.5 * IQR$.
 - Upper = $Q3 + 1.5 * IQR$.These thresholds help identify values that deviate significantly from the bulk of the data.
2. ****Results:**** The lower and upper cutoffs are printed, providing a numerical range for identifying outliers in the `emissions_by_country` dataset.

3. **Output Analysis:** Any country with total CO2 emissions below the lower cutoff or above the upper cutoff is considered an outlier. These cutoffs assist in isolating extreme values for further investigation.