

Calculating Thresholds Using IQR for Outlier Detection

Question and Screenshot:

The screenshot shows a coding exercise titled "Removing outliers" from a course on "Exploratory Data Analysis in Python". The exercise is worth 25 XP and is at step 3 of 4. The instructions are:

- Find the 75th and 25th percentiles, saving as `price_seventy_fifth` and `price_twenty_fifth` respectively.
- Calculate the IQR, storing it as `prices_iqr`.
- Calculate the upper and lower outlier thresholds.
- Remove the outliers from `planes`.

The code editor shows the following code:

```
1 # Find the 75th and 25th percentiles
2 price_seventy_fifth = planes["Price"].quantile(0.75)
3 price_twenty_fifth = planes["Price"].quantile(0.25)
4
5 # Calculate iqr
6 prices_iqr = price_seventy_fifth - price_twenty_fifth
7
8 # Calculate the thresholds
9 upper = ____ + (____ * ____ )
10 lower = ____ - (____ * ____ )
```

The IPython Shell output shows:

```
<script.py> output:
Interquartile Range (IQR) for Price: 7096.25
```

The input field for the next step is empty.

Question Explanation:

The task involves calculating the upper and lower thresholds for identifying outliers in the flight prices dataset. The thresholds are computed using the interquartile range (IQR) and percentiles.

Code Solution:

```
# Find the 75th and 25th percentiles
price_seventy_fifth = planes["Price"].quantile(0.75)
price_twenty_fifth = planes["Price"].quantile(0.25)
```

```
# Calculate iqr
```

```
prices_iqr = price_seventy_fifth - price_twenty_fifth
```

```
# Calculate the thresholds
```

```
upper = price_seventy_fifth + (1.5 * prices_iqr)
```

```
lower = price_twenty_fifth - (1.5 * prices_iqr)
```

Solution Explanation:

1. The `quantile()` function calculates the 75th and 25th percentiles of the 'Price' column.
2. The IQR is the difference between the 75th and 25th percentiles, representing the spread of the middle 50% of the data.
3. The upper threshold is computed as the 75th percentile plus 1.5 times the IQR, while the lower threshold is the 25th percentile minus 1.5 times the IQR.
4. These thresholds are used to identify outliers in the dataset.