ECON 0150 | Economic Data Analysis

How economists analyze data.

Part 1.5 | Filtering Data

Example 1.5 | Starbucks Hours

Examine opening times in Starbucks_Location_Hours.csv

```python
1  # Load the data
2  data = pd.read_csv(file_path + file_name)
```
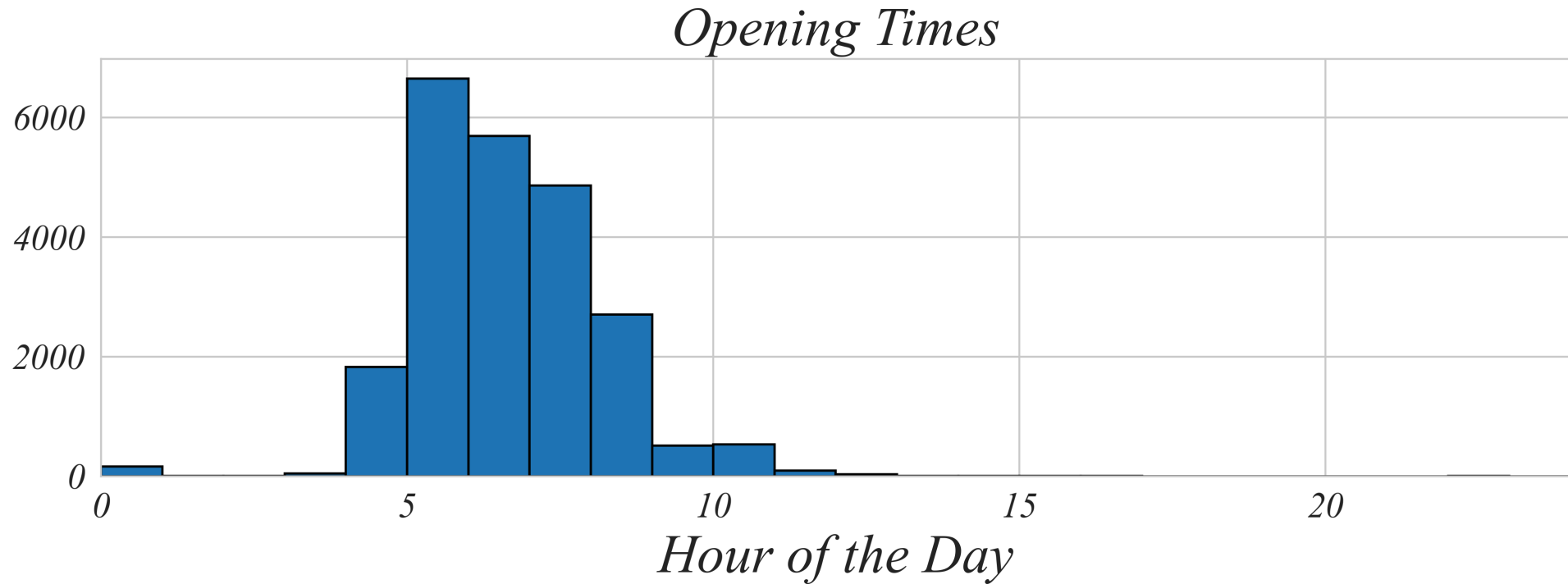
Example 1.5 | Starbucks Hours

Examine opening times in Starbucks_Location_Hours.csv

```
1  # Load the data
2  data = pd.read_csv(file_path + file_name)
3
4  # Create histogram
5  plt.hist(data['open'], bins=20)
```

Part 1.5 | Filtering Data by Category

Q. What hours should a US coffee shop operate?

Opening Times

> maybe there's something specific about the US though?

> filter only the US locations

Part 1.5 | Filtering Data by Category
Q. What hours should a US coffee shop operate?

> *filter only the US locations*

```
1  # Decide whether each row's country code is 'US'
2  data['COUNTRY_CODE'] == 'US'
```

Part 1.5 | Filtering Data by Category
Q. What hours should a US coffee shop operate?

> filter only the US locations

```
1 # Decide whether each row's country code is 'US'
2 # data['COUNTRY_CODE'] == 'US'
```

Part 1.5 | Filtering Data by Category

Q. What hours should a US coffee shop operate?

> *filter only the US locations*

```
1  # Decide whether each row's country code is 'US'
2  # data['COUNTRY_CODE'] == 'US'
3
4  # Select the rows with True
5  us_data = data[data['COUNTRY_CODE'] == 'US']
```
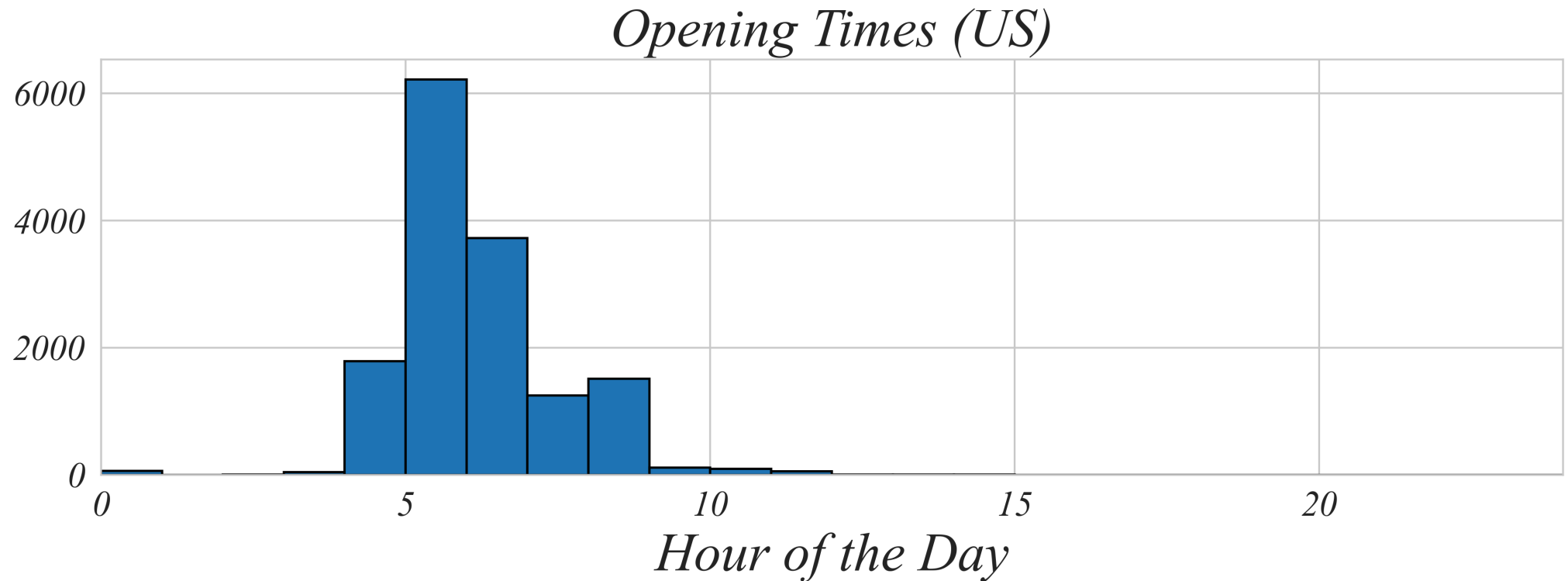
Part 1.5 | Filtering Data by Category
Q. What hours should a US coffee shop operate?

> *filter only the US locations*

```python
1  # Decide whether each row's country code is 'US'
2  # data['COUNTRY_CODE'] == 'US'
3
4  # Select the rows with True
5  us_data = data[data['COUNTRY_CODE'] == 'US']
6
7  # Create histogram
8  plt.hist(us_data['open'], bins=20)
```

Part 1.5 | Filtering Data by Category
Q. What hours should a US coffee shop operate?
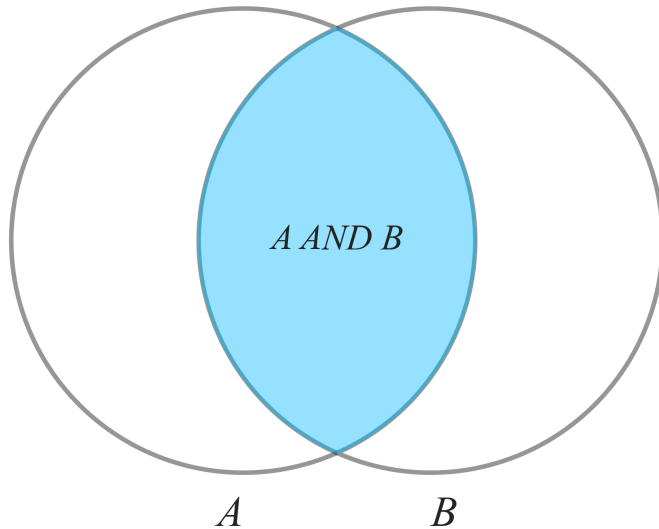
Opening Times (US)

> *is there something different between the US and Canada?*
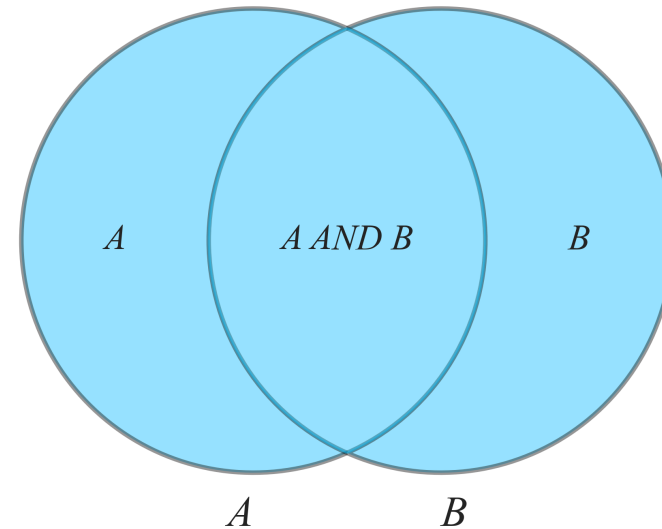
> *filter for **BOTH** countries*

Part 1.5 | Filtering Data by Category
Lets us some Boolian logic :)

AND
(python: &)
Both terms

OR
(python: |)
Either term

A AND B

A

A AND B

B

A   B

A   B

> is there something different between the US and Canada?

Part 1.5 | Filtering Data by Category
Q. What hours should a US coffee shop operate?

> *filter for **BOTH** countries*

```python
1  # Find the data in either the US or in Canada (CA)
2
3  # Method 1: Using OR operator
4  data[(data['COUNTRY_CODE'] == 'US') | (data['COUNTRY_CODE'] == 'CA')]
```

Part 1.5 | Filtering Data by Category
Q. What hours should a US coffee shop operate?

> *filter for **BOTH** countries*

```
1  # Find the data in either the US or in Canada (CA)
2
3  # Method 1: Using OR operator
4  data[(data['COUNTRY_CODE'] == 'US') | (data['COUNTRY_CODE'] == 'CA')]
5
6  # Method 2: Using isin()
7  data[data['COUNTRY_CODE'].isin(['US', 'CA'])]
```

Part 1.5 | Filtering Data by Category
Q. What hours should a US coffee shop operate?

> *filter for **BOTH** countries*

```
1  # Find the data in either the US or in Canada (CA)
2
3  # Method 1: Using OR operator
4  # data[(data['COUNTRY_CODE'] == 'US') | (data['COUNTRY_CODE'] == 'CA')]
5
6  # Method 2: Using isin()
7  data[data['COUNTRY_CODE'].isin(['US', 'CA'])]
```

Part 1.5 | Filtering Data by Category
Q. What hours should a US coffee shop operate?

> *filter for **BOTH** countries*

```python
# Find the data in either the US or in Canada (CA)

# Method 1: Using OR operator
# data[(data['COUNTRY_CODE'] == 'US') | (data['COUNTRY_CODE'] == 'CA')]

# Method 2: Using isin() and define a new dataset
us_ca_data = data[data['COUNTRY_CODE'].isin(['US', 'CA'])]
```

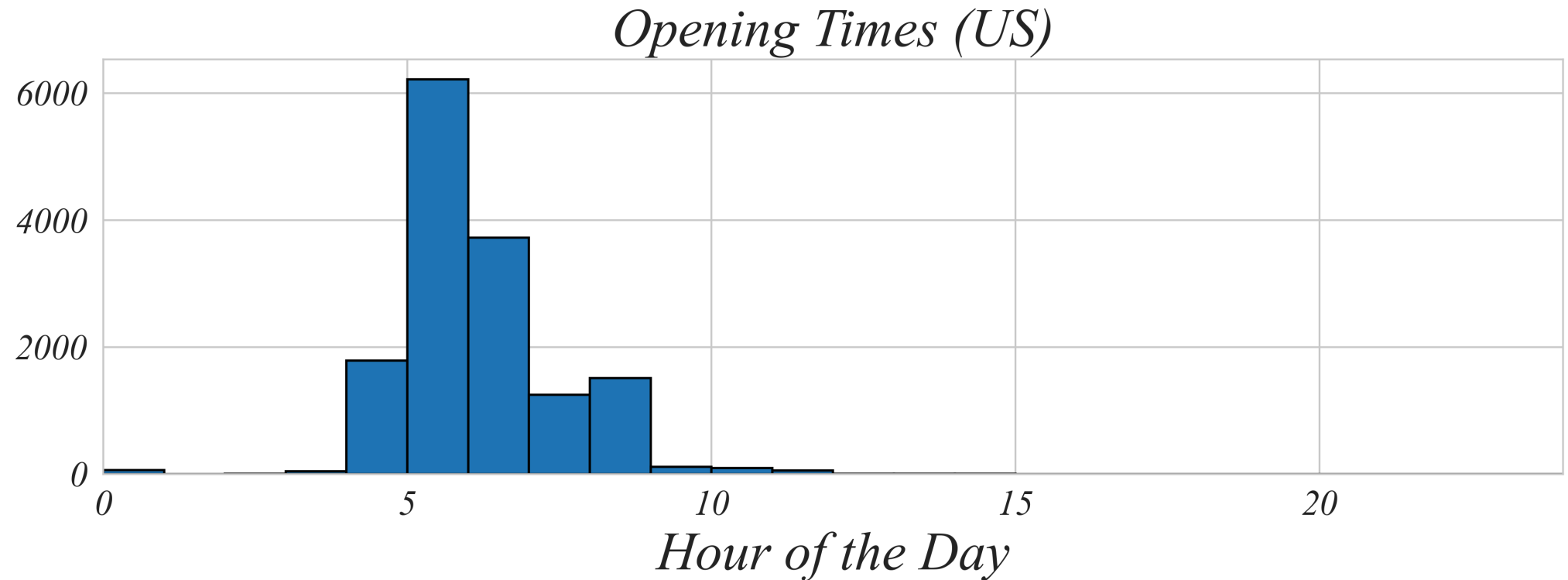# Part 1.5 | Filtering Data by Category

Q. What hours should a US coffee shop operate?

> *filter for **BOTH** countries*

```python
1   # Find the data in either the US or in Canada (CA)
2
3   # Method 1: Using OR operator
4   # data[(data['COUNTRY_CODE'] == 'US') | (data['COUNTRY_CODE'] == 'CA')]
5
6   # Method 2: Using isin() and define a new dataset
7   us_ca_data = data[data['COUNTRY_CODE'].isin(['US', 'CA'])]
8
9   # Create histogram
10  plt.hist(us_ca_data['open'], bins=20)
```
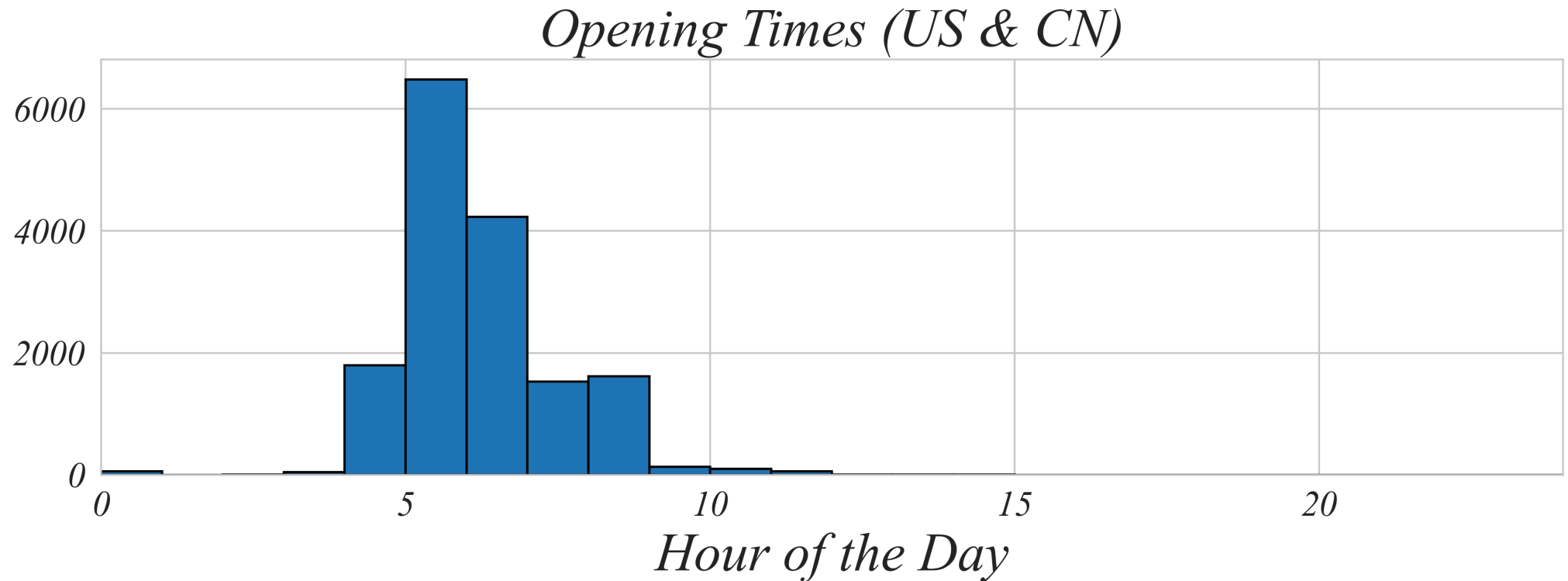
Part 1.5 | Filtering Data by Category

Q. What hours should a US coffee shop operate?

Opening Times (US)

Hour of the Day

Part 1.5 | Filtering Data by Category

Q. What hours should a US coffee shop operate?

Opening Times (US & CN)

Hour of the Day

Part 1.5 | Filtering Data by Category

Q. What would the histogram for the following filtered data look like?

```
1  data[(data['COUNTRY_CODE'] == 'US') & (data['COUNTRY_CODE'] == 'CN')]
```

> *it would contain no data!*

Part 1.5 | Filtering Data by Inequality

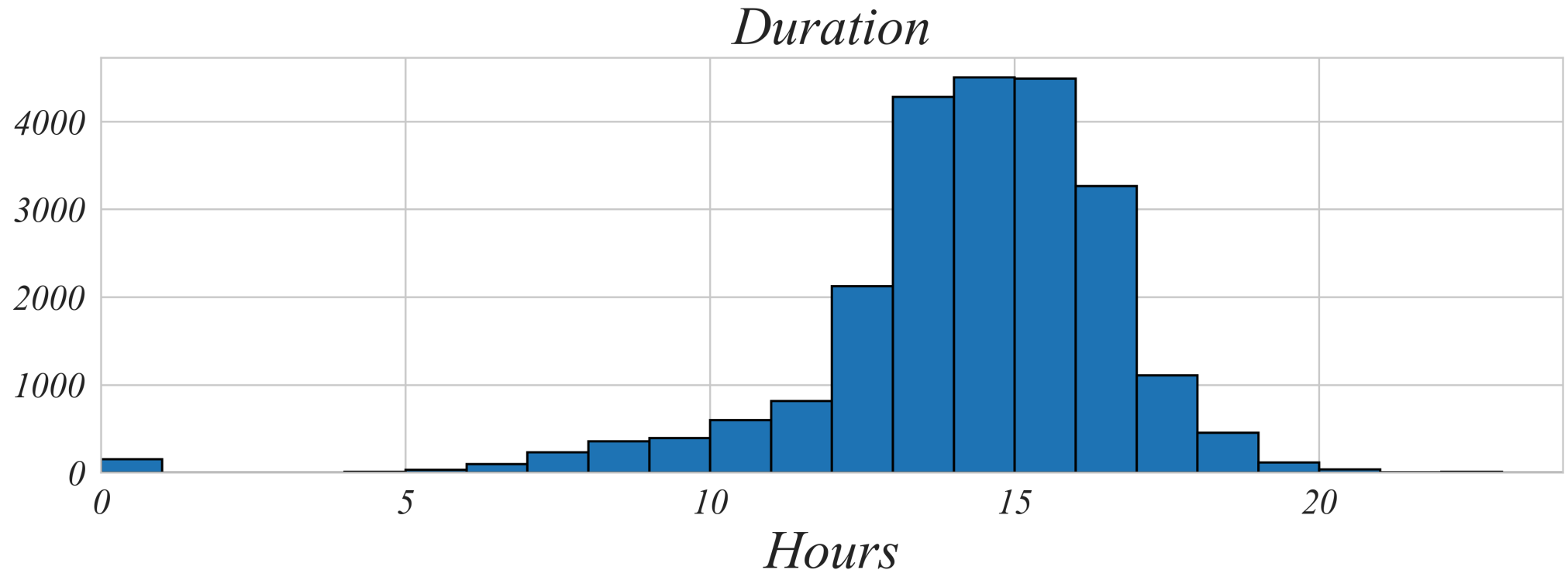*Filtering continuous data requires using inequalities.*

Symbol	Python	Example
$=$	==	df[df['age'] == 25]
$\neq$	!=	df[df['age'] != 25]
$<$	<	df[df['age'] < 25]
$>$	>	df[df['age'] > 25]
$\leq$	<=	df[df['age'] <= 25]
$\geq$	>=	df[df['age'] >= 25]

Part 1.5 | Filtering Data by Inequality

Q. How long do locations stay open?

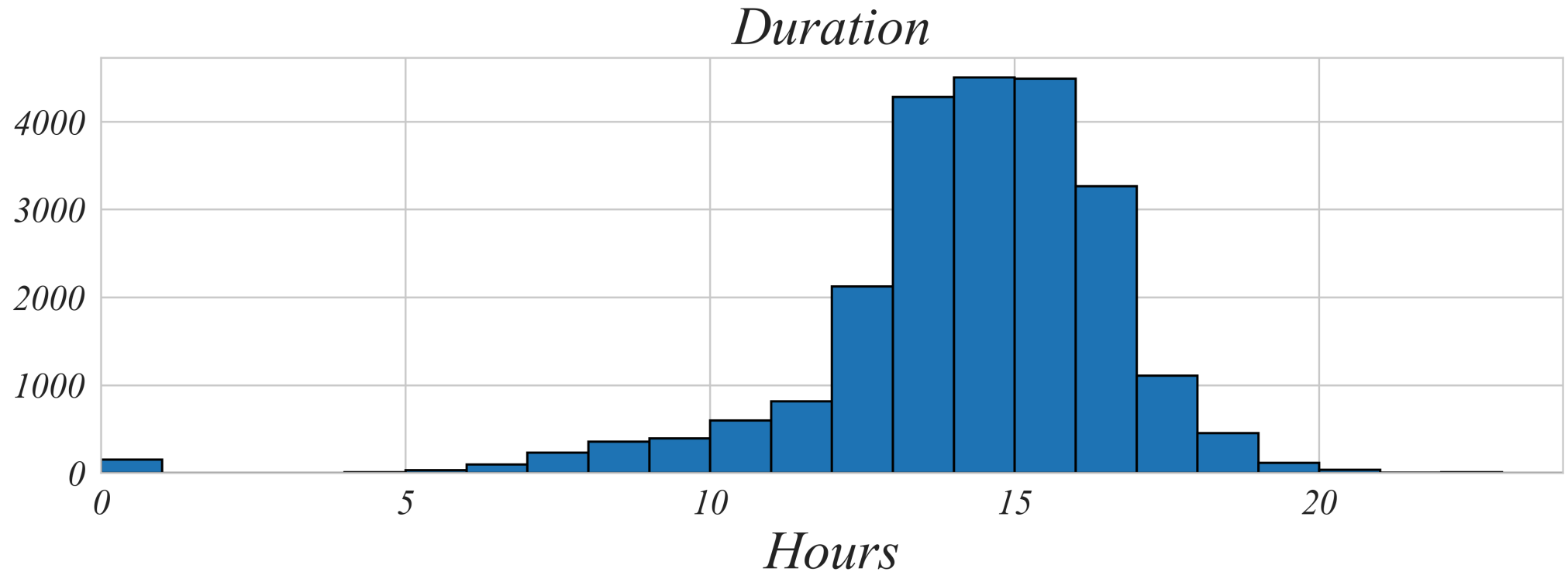# Part 1.5 | Filtering Data by Inequality

Q. How long do locations stay open?

Duration

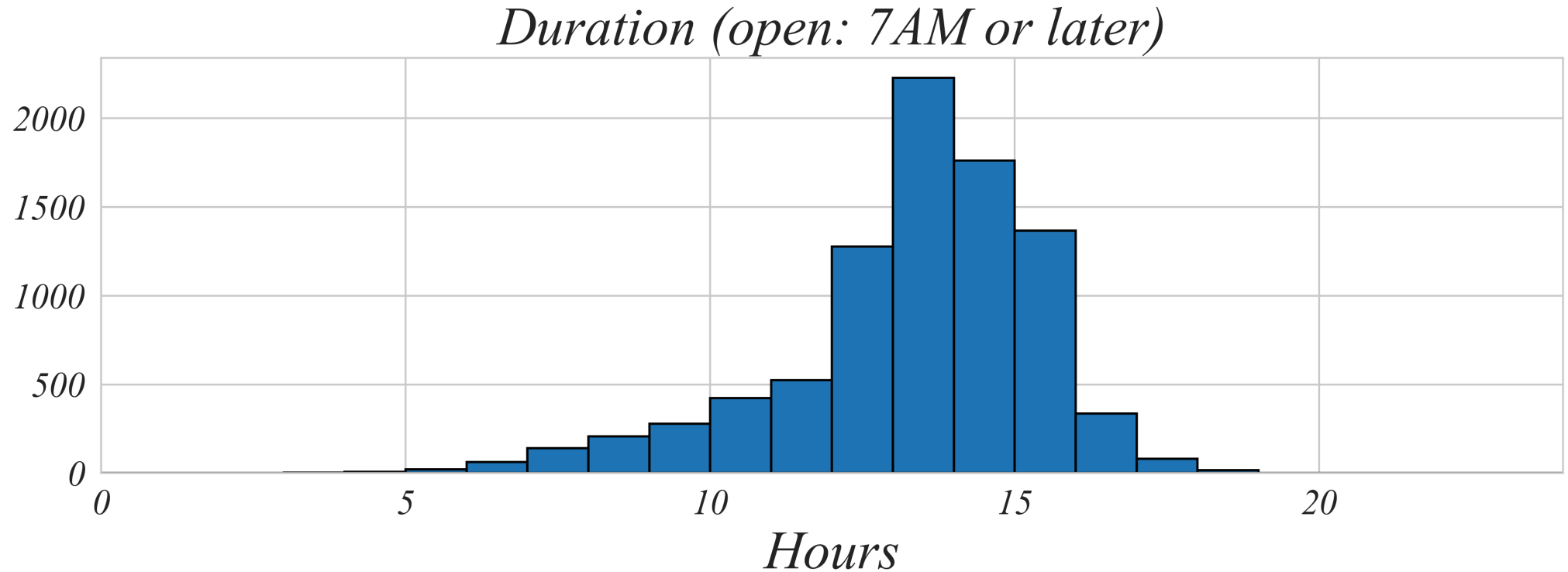> *but is it different by opening time?*

Part 1.5 | Filtering Data by Inequality

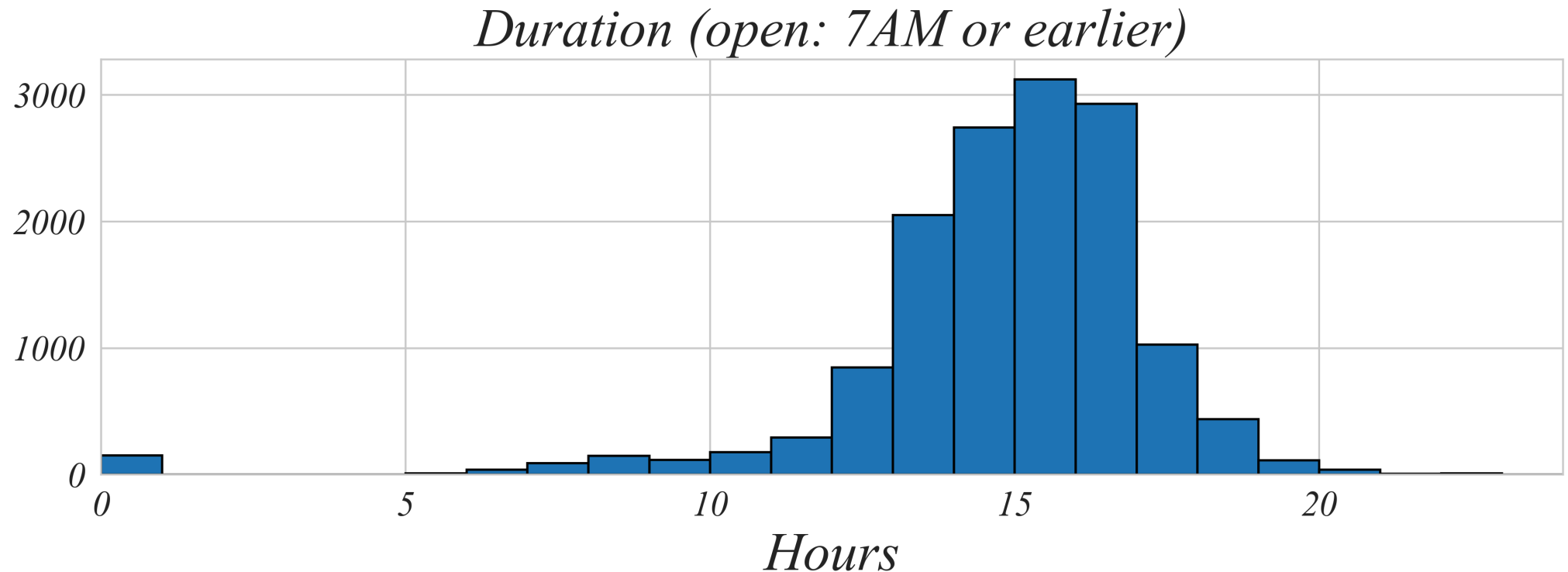*Q. Do locations that open later in the morning stay open fewer hours?*

Duration

Part 1.5 | Filtering Data by Inequality

Q. Do locations that open later in the morning stay open fewer hours?

Duration (open: 7AM or later)

Part 1.5 | Filtering Data by Inequality

Q. Do locations that open later in the morning stay open fewer hours?

Duration (open: 7AM or earlier)

> *we can also compare to locations that open early*

Part 1.5 | Filtering Data by Inequality

Q. Do locations that open later in the morning stay open fewer hours?

Duration (open: 7AM or earlier)
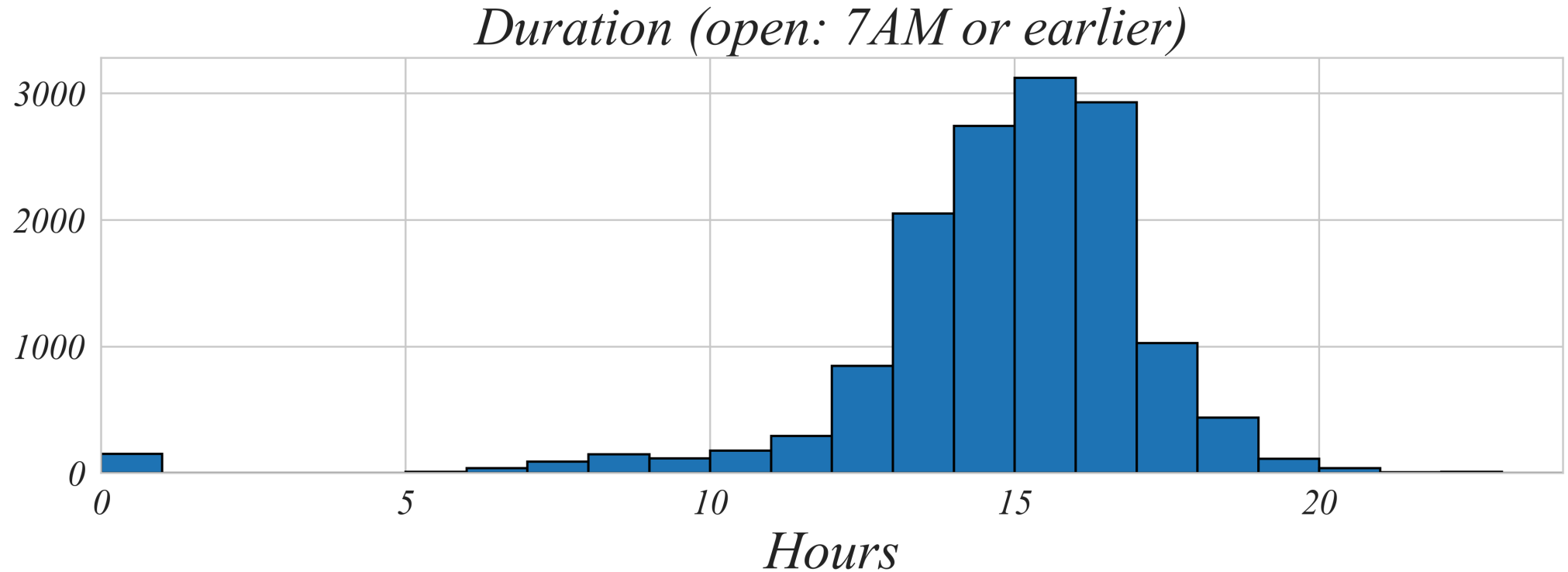
> *but maybe this differs country?*

Part 1.5 | Filtering Data by Inequality

Q. Does the hours open at early open locations differ by country?

Duration (open: 7AM or earlier; US & CA)
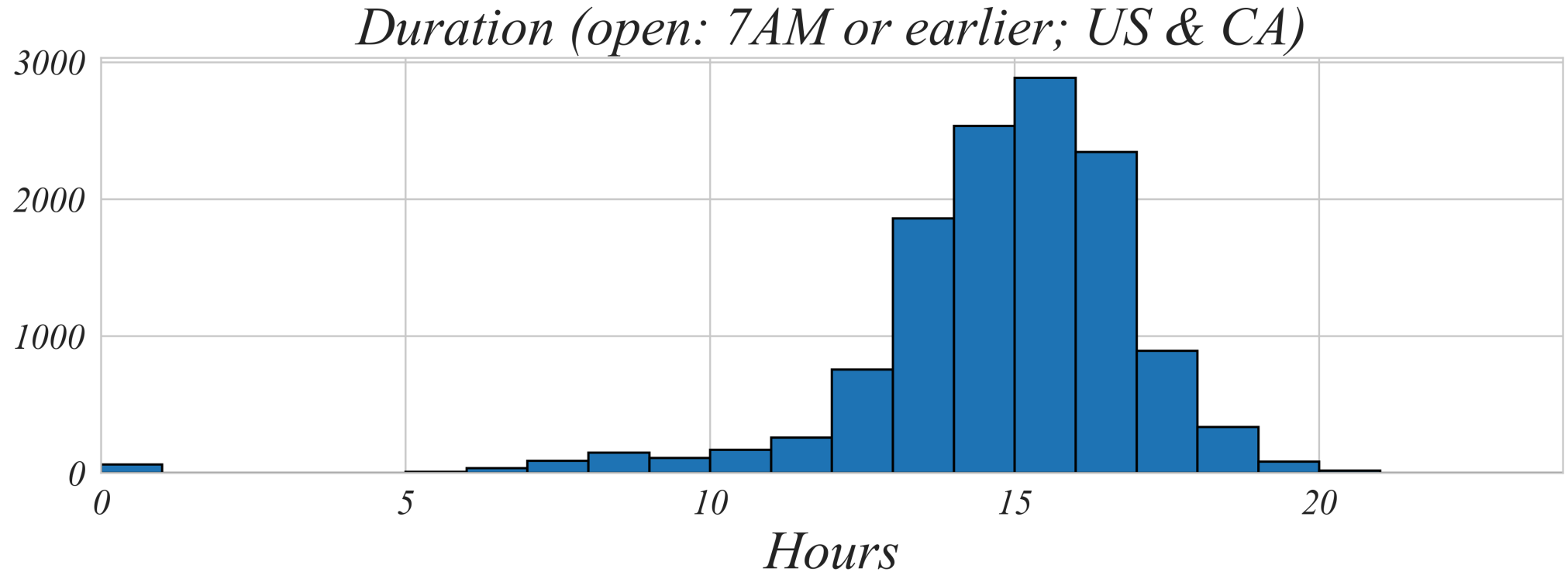
> *maybe there's a more systematic way of showing these differences*
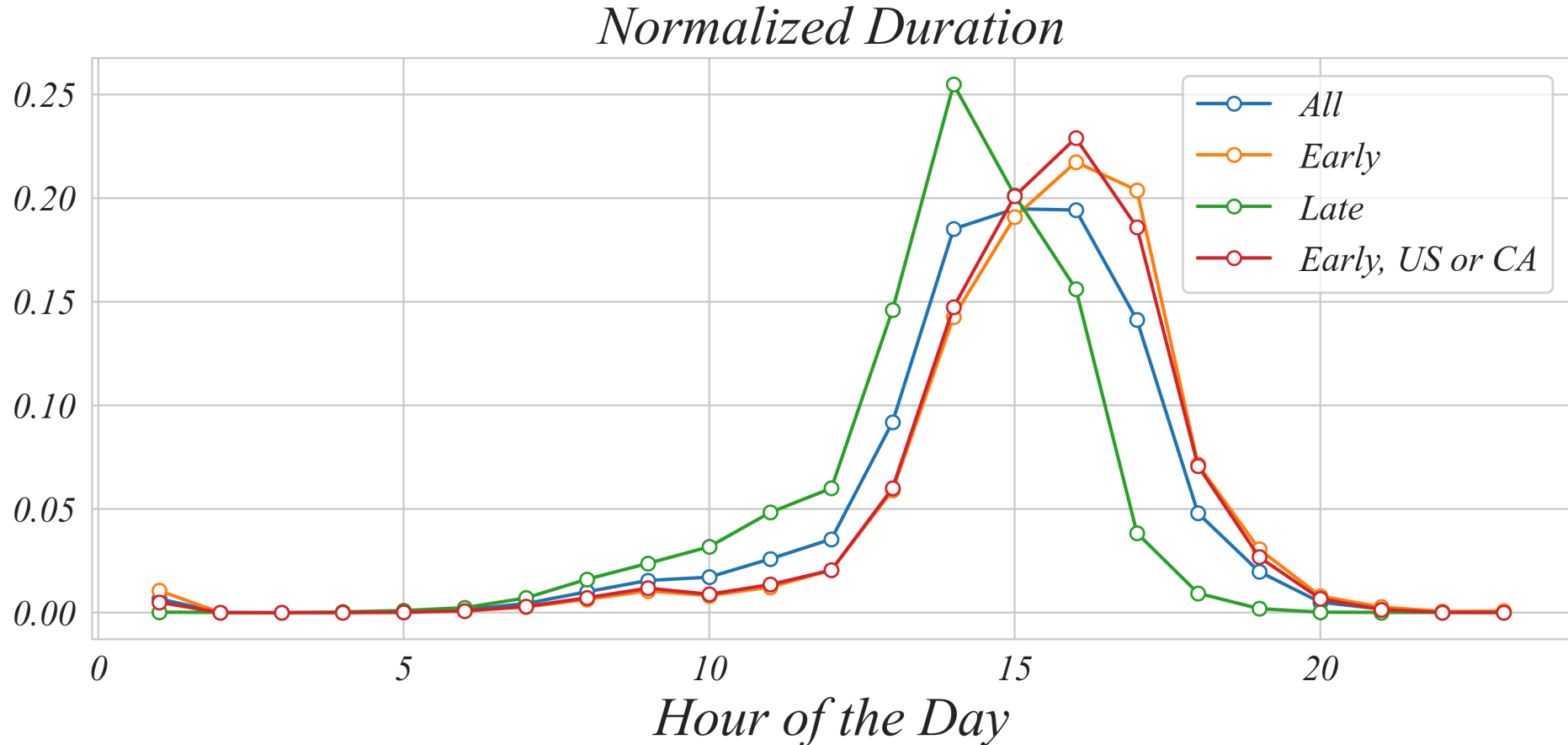
Part 1.5 | Filtering Data by Inequality

Q. Does the hours open at early open locations differ by country?

Duration
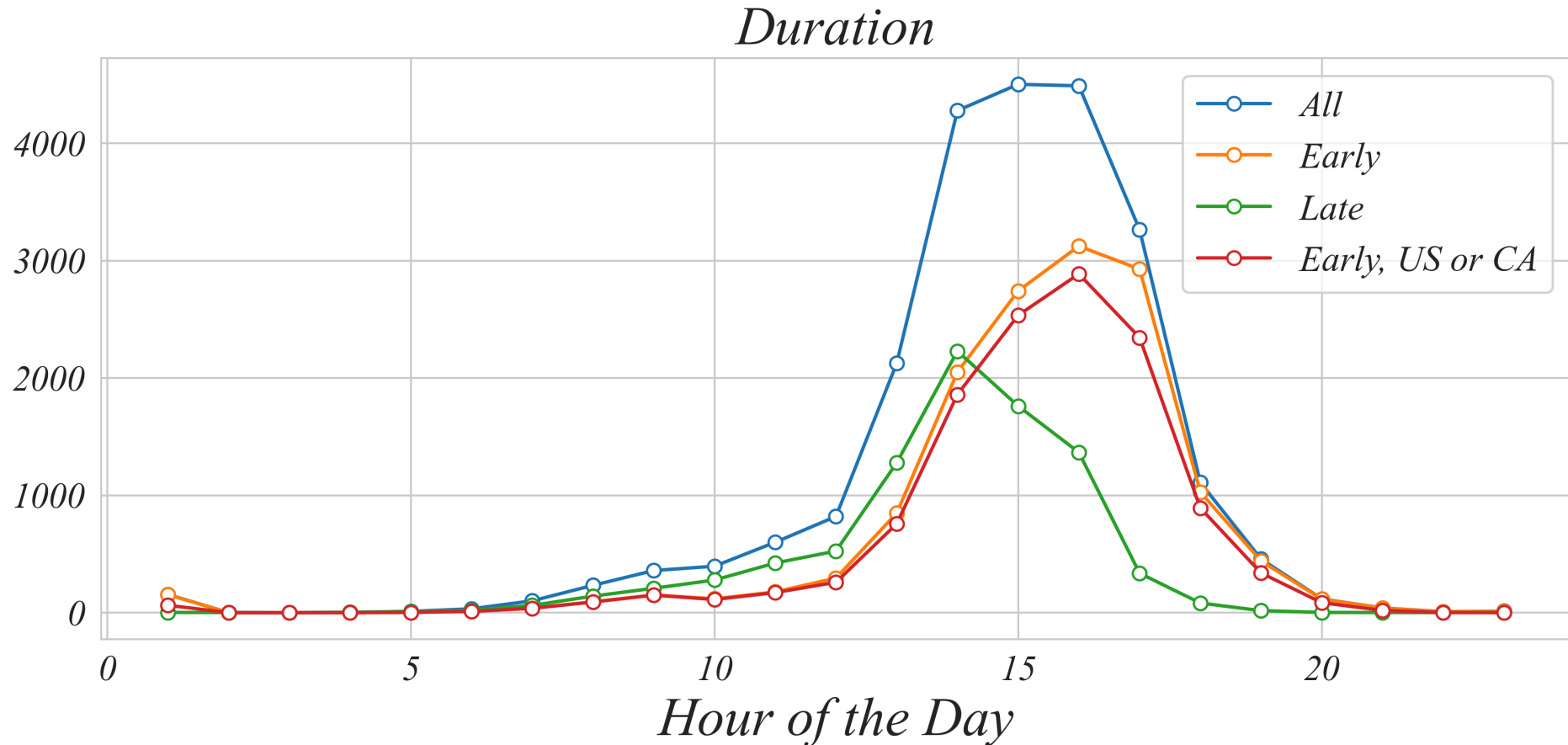
> *normalizing gives us a better picture of the distribution*

Part 1.5 | Filtering Data by Inequality

Q. Does the hours open at early open locations differ by country?

Normalized Duration

> *normalizing the distributions allows us to compare filter*

Taylor | ECON 0150 | Week 3 | Class B