ECON 0150 | Economic Data Analysis

The economist's data analysis pipeline.

Part 1.5 | Filtering Data

Example 1.5 | Starbucks Hours

Use Starbucks_Location_Hours.csv to inform a new shop's hours.

```
# Load the data
data = pd.read_csv(file_path + file_name)
```

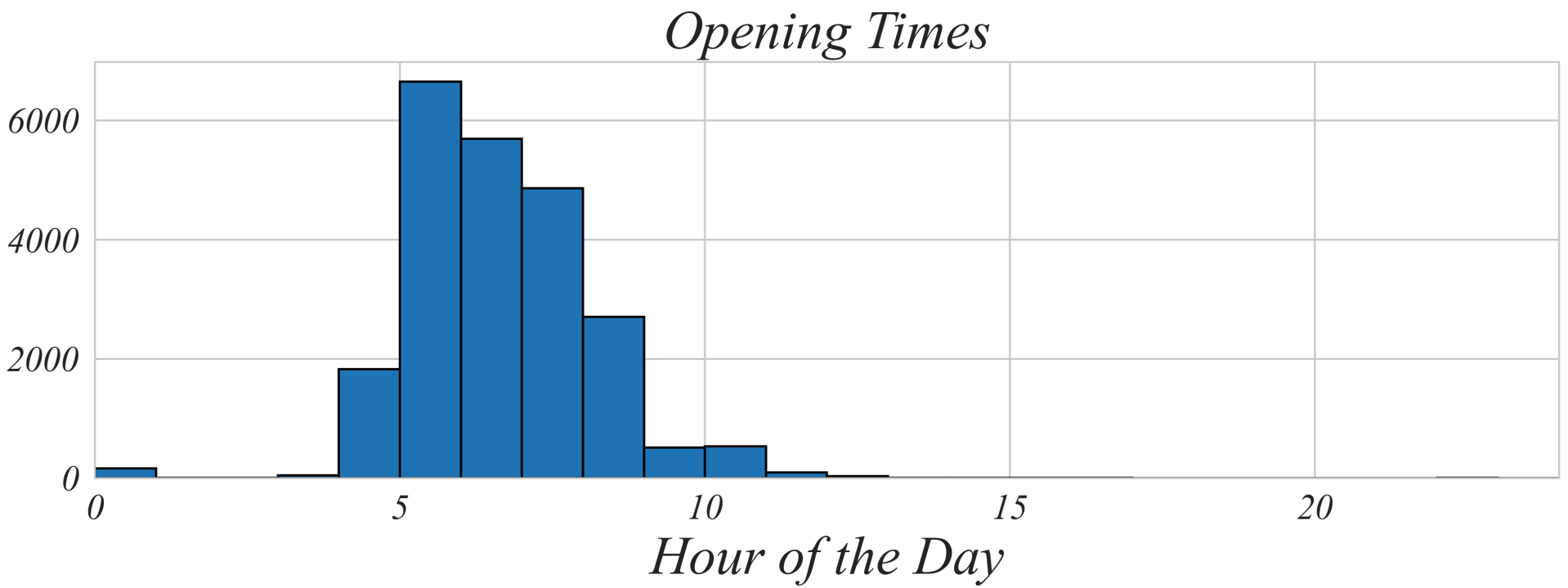
Example 1.5 | Starbucks Hours

Use Starbucks_Location_Hours.csv to inform a new shop's hours.

```
# Load the data
data = pd.read_csv(file_path + file_name)

# Create histogram
plt.hist(data['open'], bins=20)
```

A New Coffee Shop: Filter by Category

Q. What hours should a US coffee shop operate?



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

> filter only the US locations

A New Coffee Shop: Filter by Category

Q. What hours should a US coffee shop operate?

> filter only the US locations

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

A New Coffee Shop: Filter by Category

Q. What hours should a US coffee shop operate?

> filter only the US locations

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

A New Coffee Shop: Filter by Category

Q. What hours should a US coffee shop operate?

> filter only the US locations

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

# Select the rows with True
us_data = data[data['COUNTRY_CODE'] == 'US']
```

A New Coffee Shop: Filter by Category

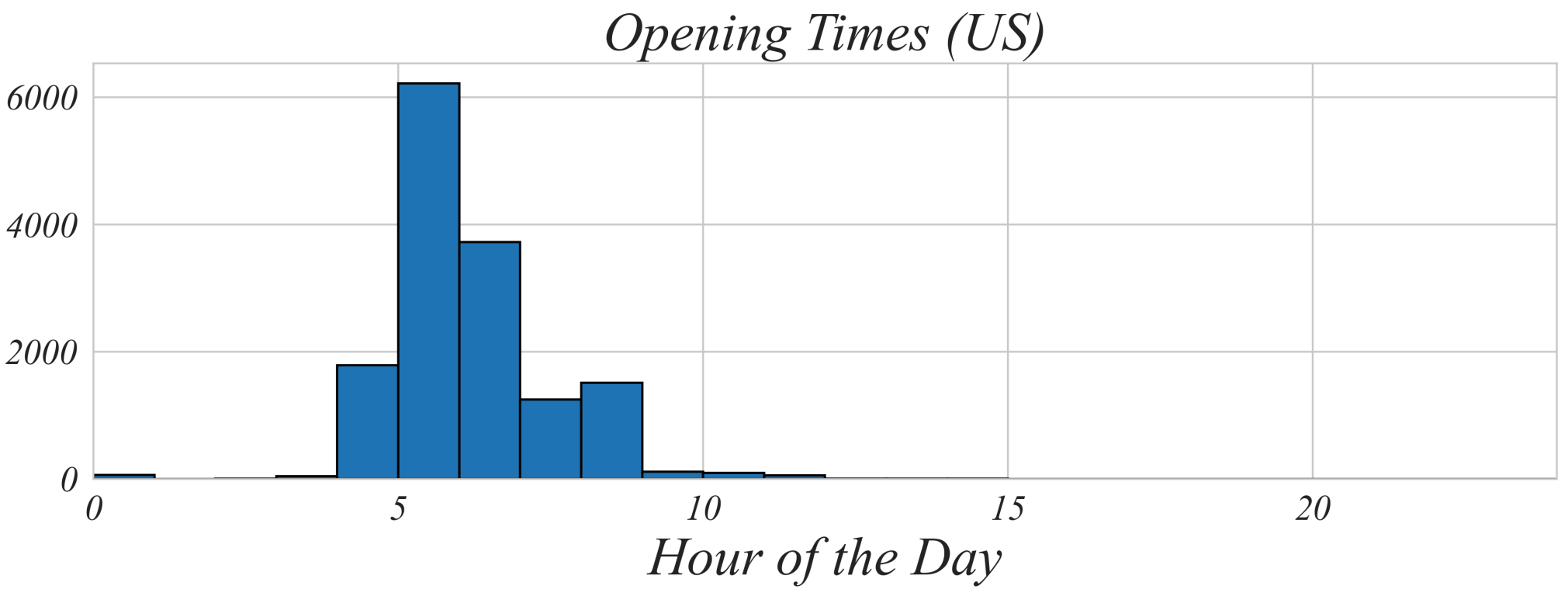
Q. What hours should a US coffee shop operate?

> filter only the US locations

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

# Select the rows with True
us_data = data[data['COUNTRY_CODE'] == 'US']

# Create histogram
plt.hist(us_data['open'], bins=20)
```

A New Coffee Shop: Filter by Category

Q. What hours should a US coffee shop operate?



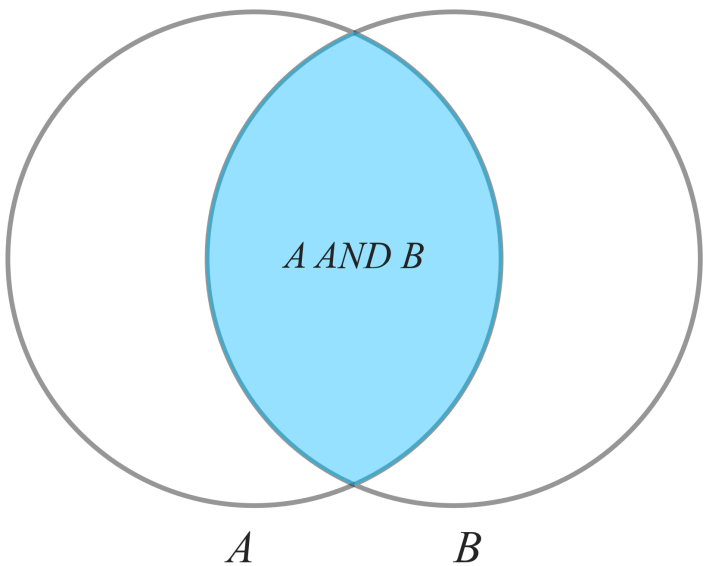
> is there something different between the US and Canada?

*> filter for **BOTH** countries*

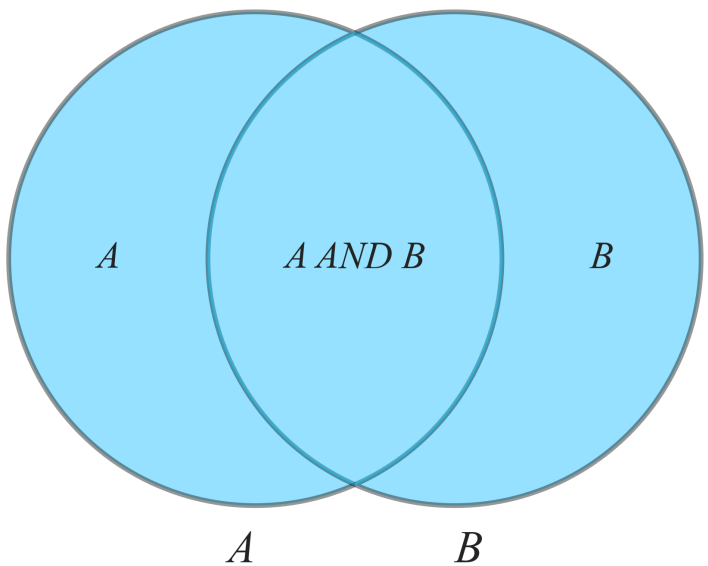
A New Coffee Shop: Filter by Category

Lets us some Boolian logic :)

AND
(python: &)
Both terms



OR
(python: |)
Either term



> is there something different between the US and Canada?

A New Coffee Shop: Filter by Category

Q. What hours should a US coffee shop operate?

> *filter for **BOTH** countries*

```
# 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')]
```

A New Coffee Shop: Filter by Category

Q. What hours should a US coffee shop operate?

> *filter for **BOTH** countries*

```
# 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()
data[data['COUNTRY_CODE'].isin(['US', 'CA'])]
```

A New Coffee Shop: Filter by Category

Q. What hours should a US coffee shop operate?

> *filter for **BOTH** countries*

```
# 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()
data[data['COUNTRY_CODE'].isin(['US', 'CA'])]
```

A New Coffee Shop: Filter by Category

Q. What hours should a US coffee shop operate?

> *filter for **BOTH** countries*

```
# 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'])]
```

A New Coffee Shop: Filter by Category

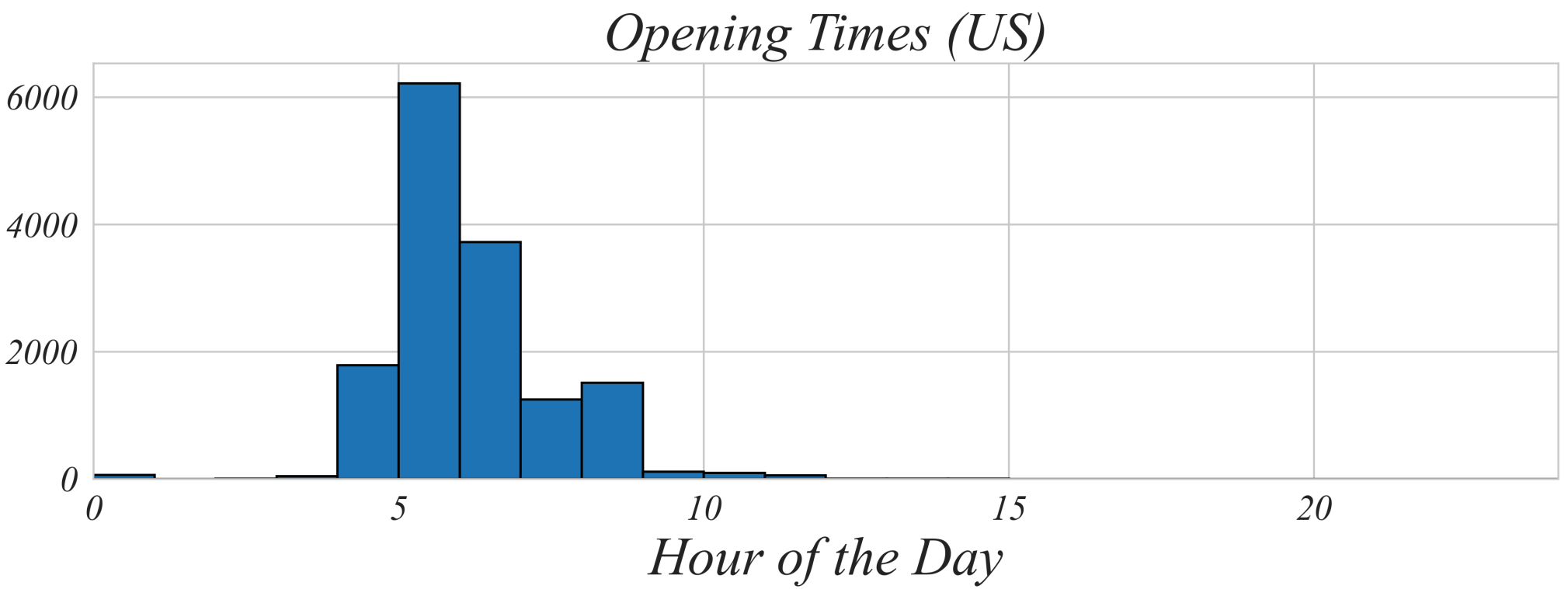
Q. What hours should a US coffee shop operate?

> *filter for **BOTH** countries*

```
# 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'])]

# Create histogram
plt.hist(us_ca_data['open'], bins=20)
```

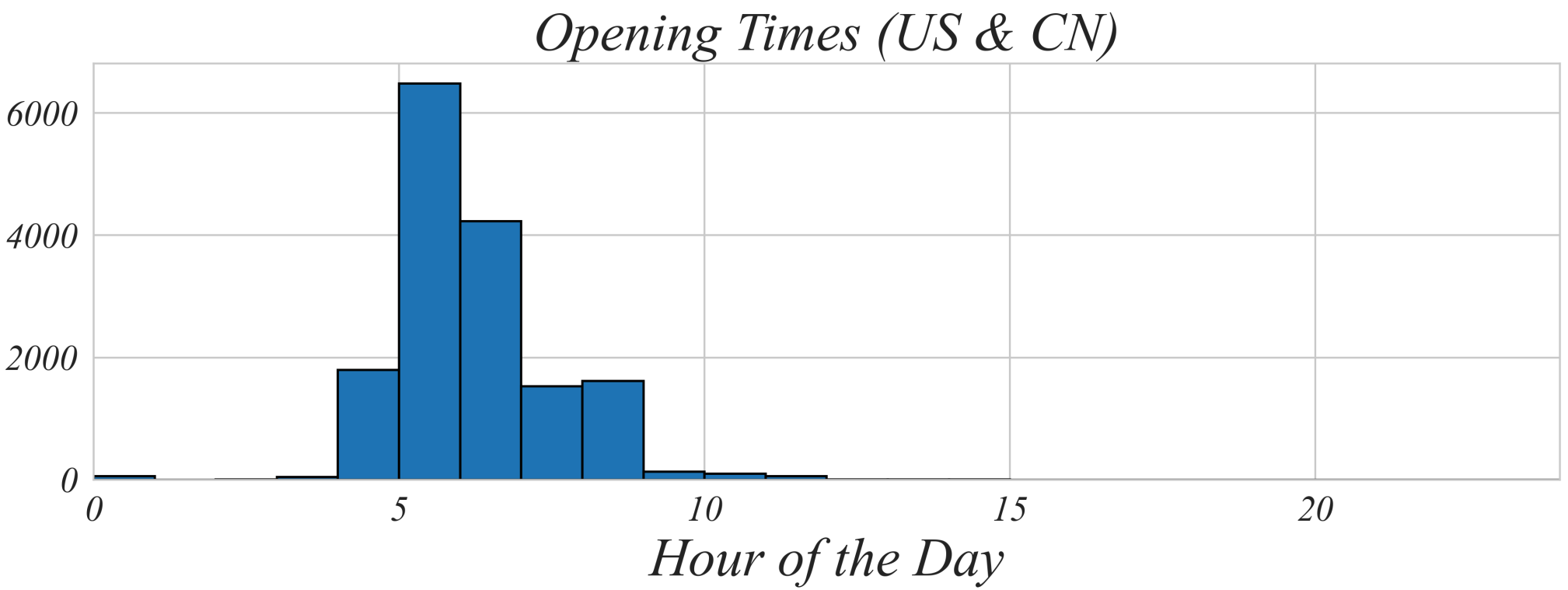
A New Coffee Shop: Filter by Category

Q. What hours should a US coffee shop operate?



A New Coffee Shop: Filter by Category

Q. What hours should a US coffee shop operate?



A New Coffee Shop: Filter 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.

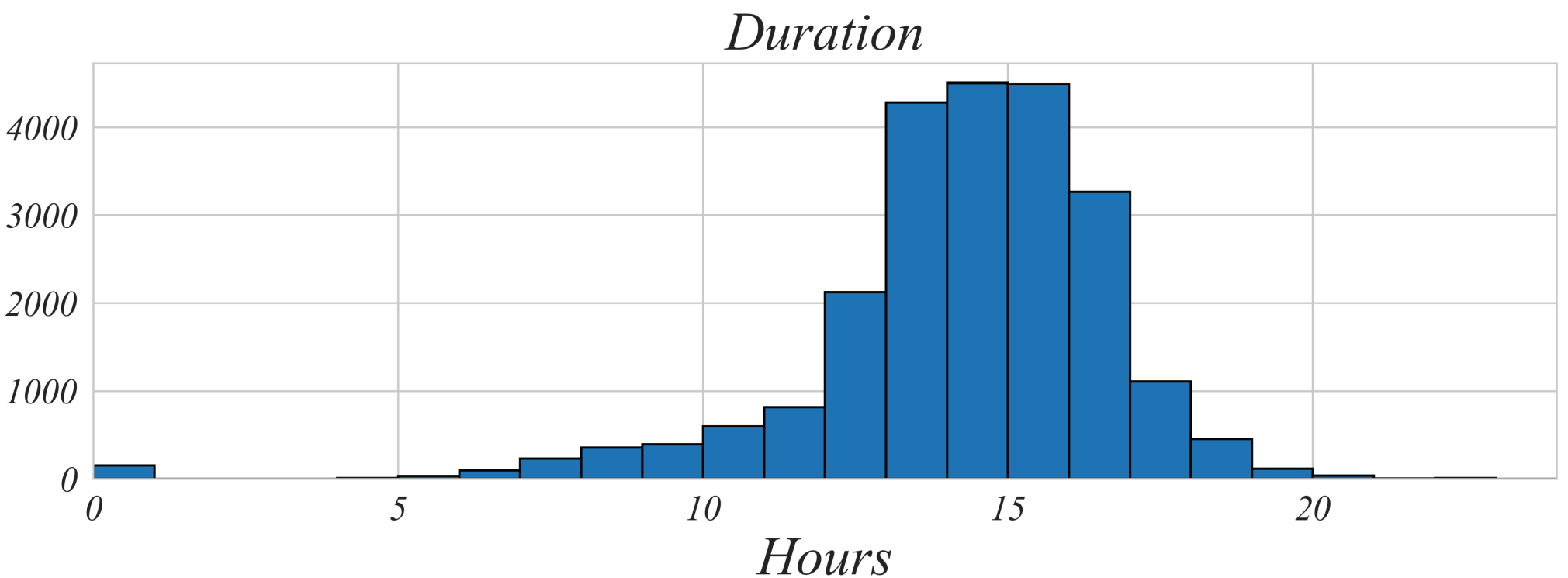
Symbol	Python	Example
=	==	`df[df['age'] == 25]`
≠	!=	`df[df['age'] != 25]`
<	<	`df[df['age'] < 25]`
>	>	`df[df['age'] > 25]`
≤	<=	`df[df['age'] <= 25]`
≥	>=	`df[df['age'] >= 25]`

A New Coffee Shop: Filter by Inequality

Q. How long do locations stay open?

A New Coffee Shop: Filter by Inequality

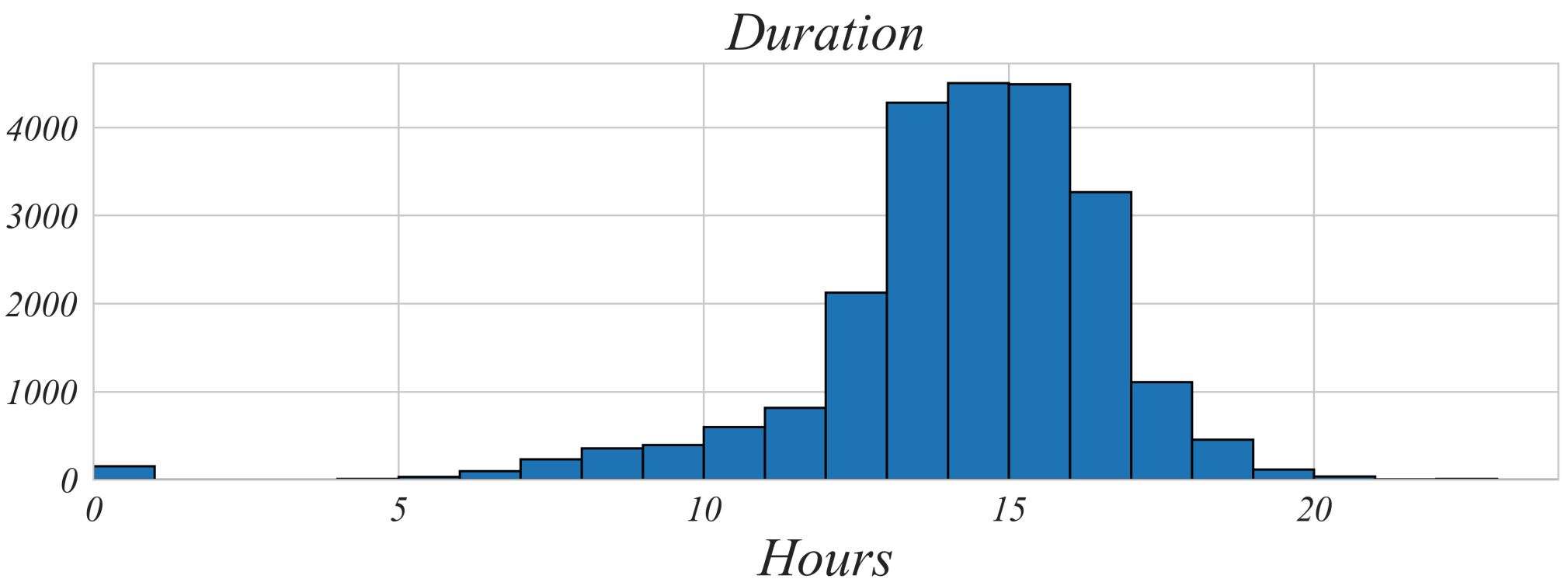
Q. How long do locations stay open?



> but is it different by opening time?

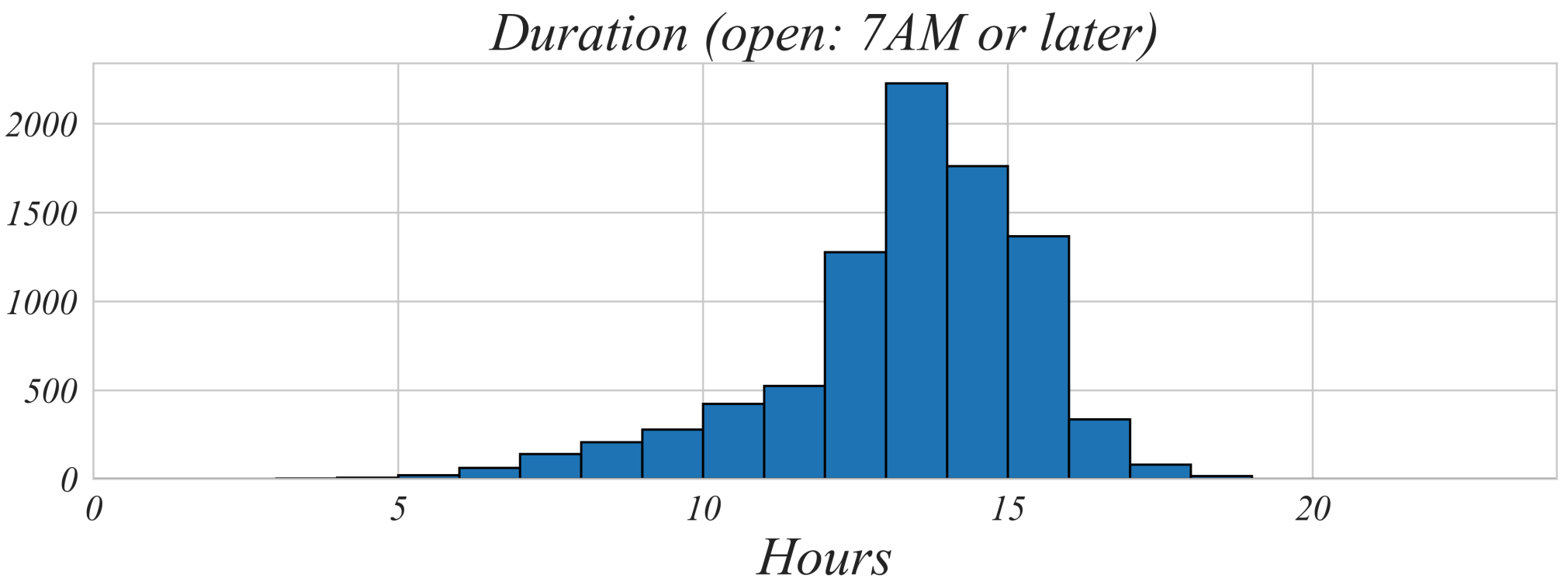
A New Coffee Shop: Filter by Inequality

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



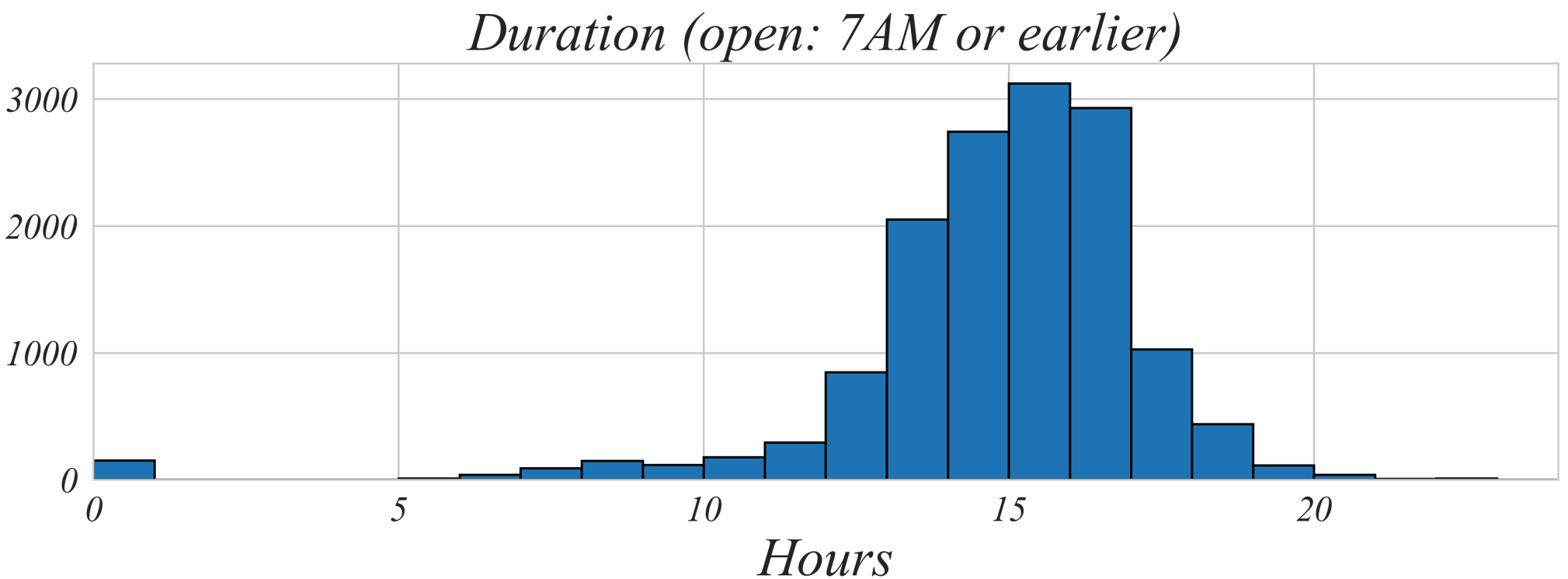
A New Coffee Shop: Filter by Inequality

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



A New Coffee Shop: Filter by Inequality

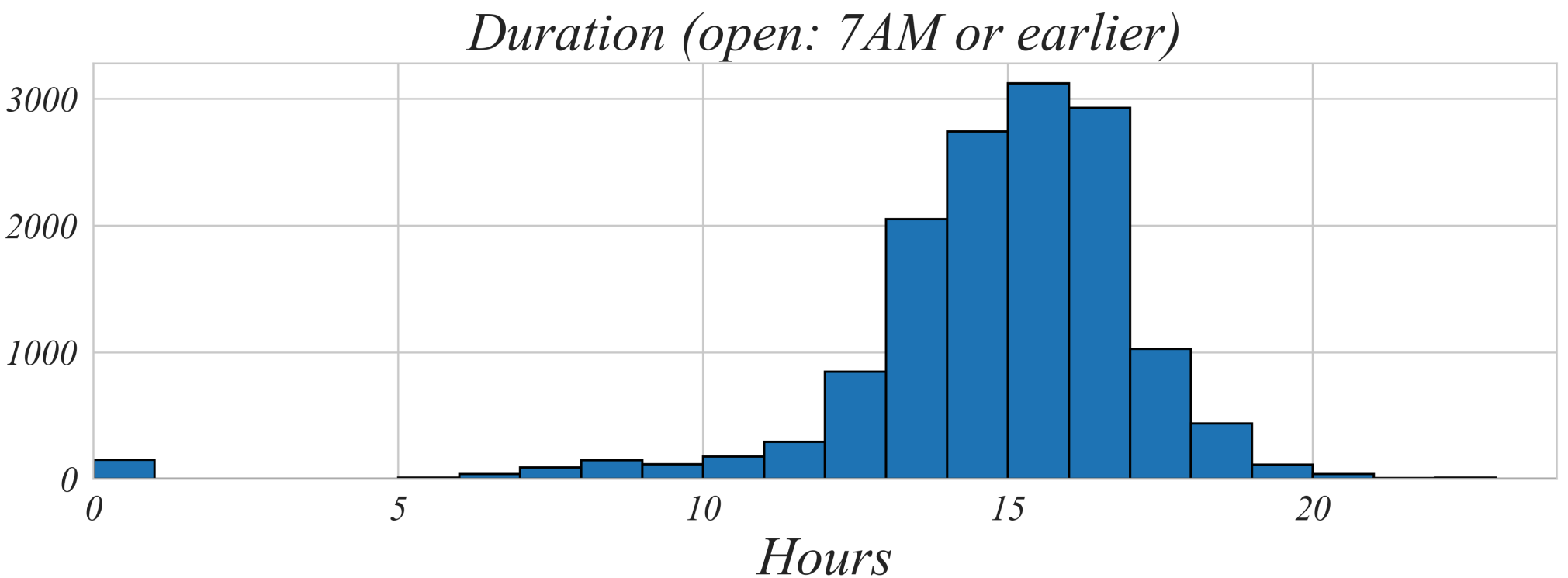
Q. Do locations that open later in the morning stay open fewer hours?



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

A New Coffee Shop: Filter by Inequality

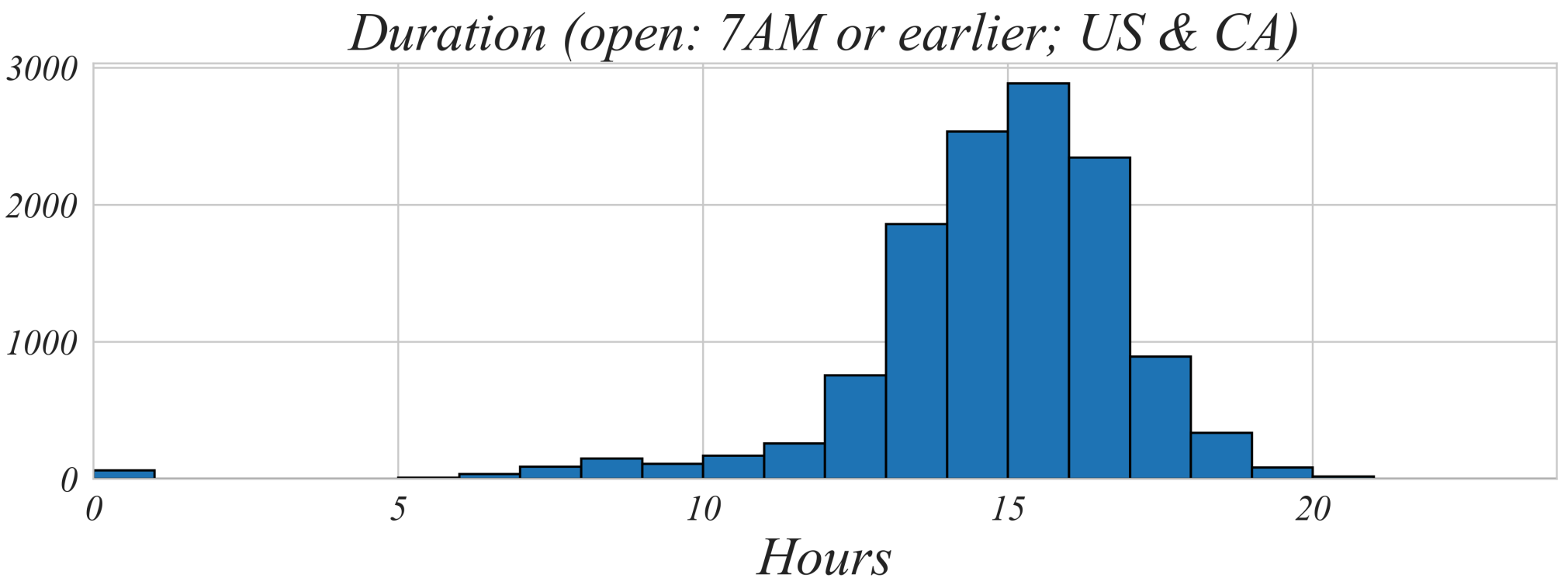
Q. Do locations that open later in the morning stay open fewer hours?



> *but maybe this differs country?*

A New Coffee Shop: Filter by Inequality

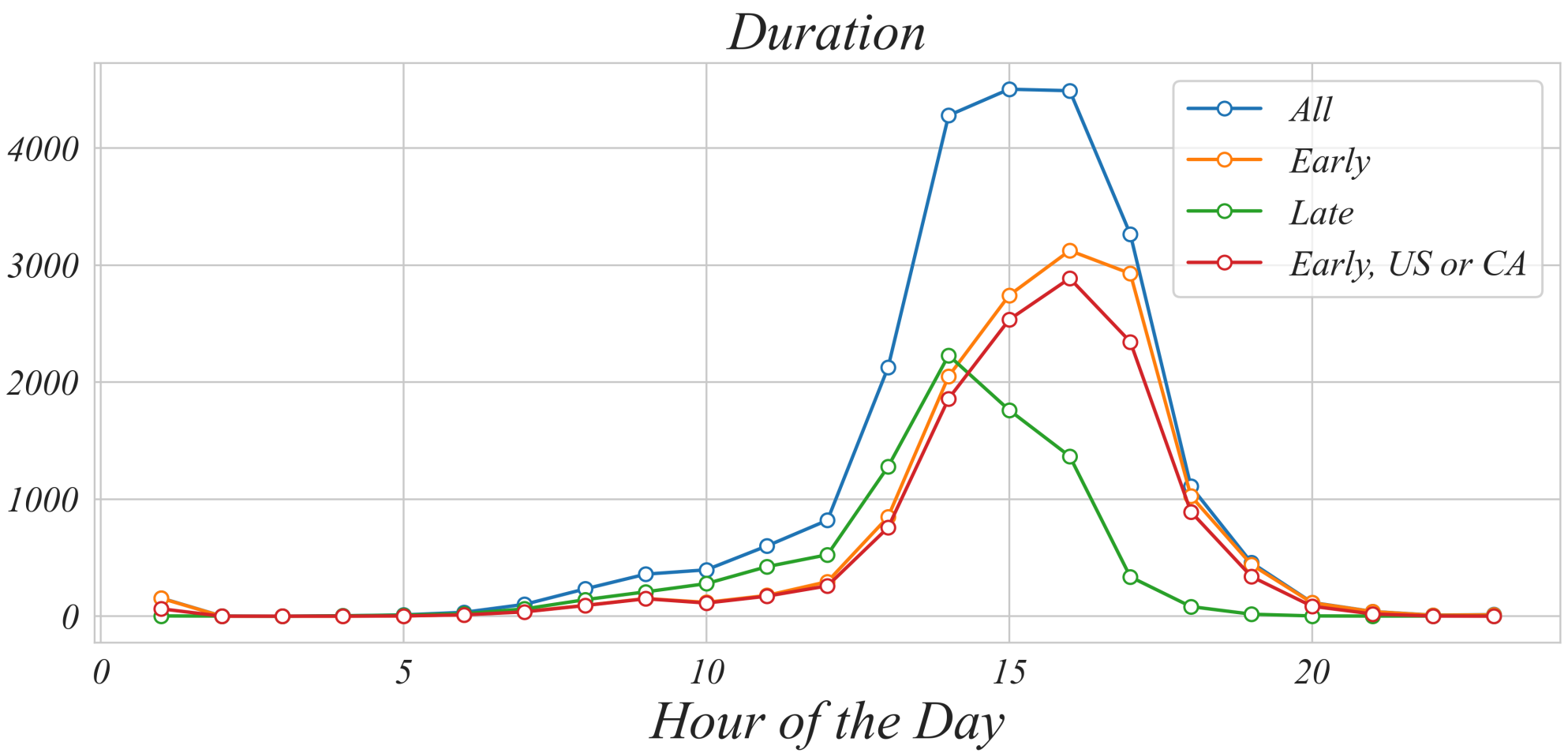
Q. Does the hours open at early open locations differ by country?



> *maybe there's a more systematic way of showing these differences*

A New Coffee Shop: Filter by Inequality

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



> but some filters have more shops, making it hard to compare

A New Coffee Shop: Filter by Inequality

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

> normalizing the distributions allows us to compare between filters