ECON 0150 | Economic Data Analysis

The economist's data analysis pipeline.

Part 2.3 | Filtering Data

A New US Coffee Shop

Lets use Starbucks_Location_Hours.csv to inform a new shop's hours.

- *The coffee shop is opening in update New York, near the border to Canada.*

- *You're asked to help make some decisions about how to run the shop when it opens.*

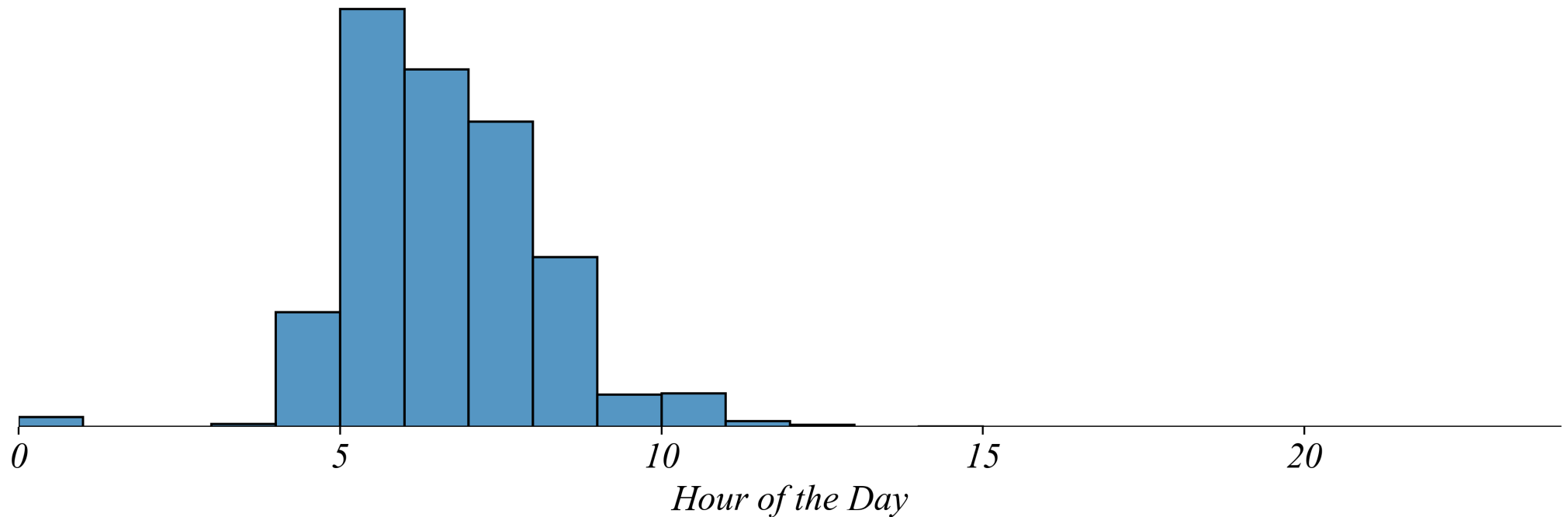- *The dataset* `Starbucks_Location_Hours.csv` *contains information about Starbucks coffee shops globally.*

A New Coffee Shop

Q. When might be a good time for the coffee shop to open?
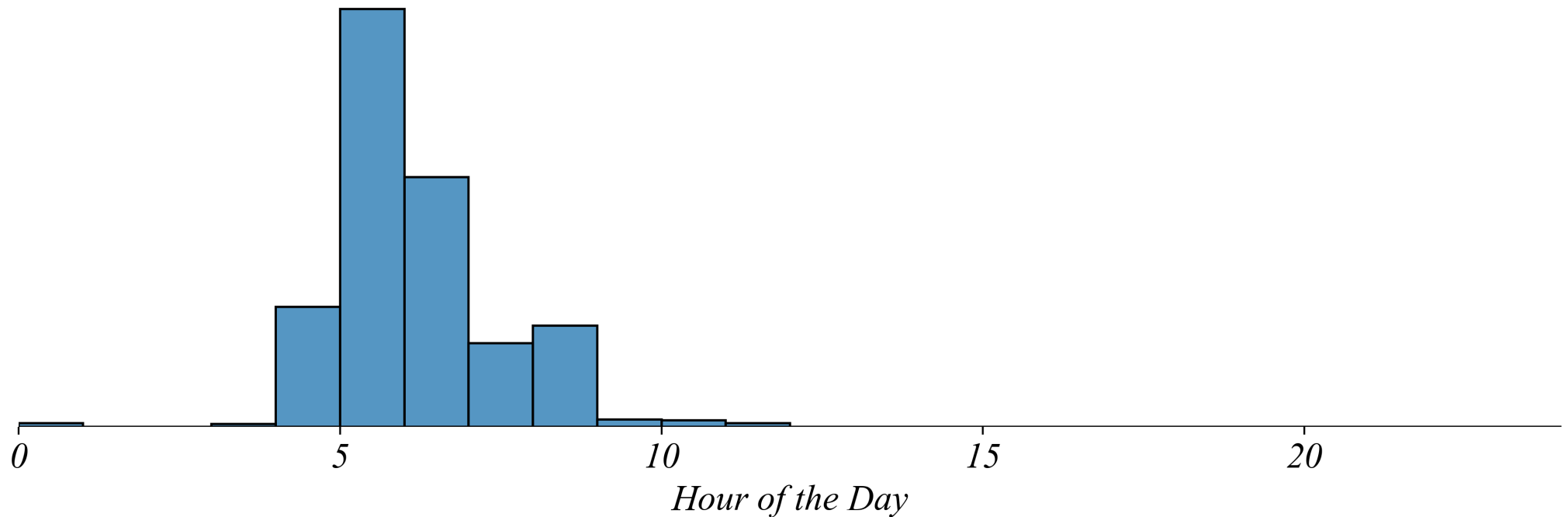
A New Coffee Shop

Opening Times

Hour of the Day

> *so it seems best to open sometime in the morning… makes sense*

> *but what if there's something specific about US coffee drinkers though?*

A New Coffee Shop: Filter by Category

Q. When might be a good time for the coffee shop to open?

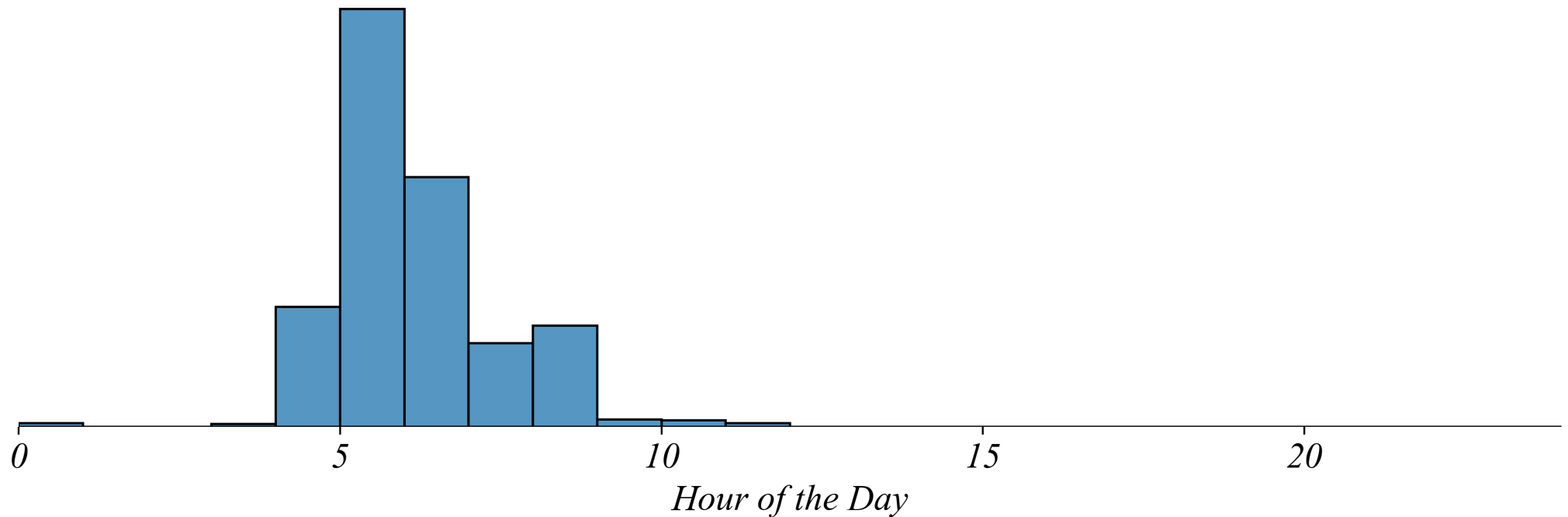Opening Times (US)

Hour of the Day

> *here we've filtered for US locations*

> *so it seems US Starbucks open earlier*

A New Coffee Shop: Filter by Category

Q. When might be a good time for the coffee shop to open?
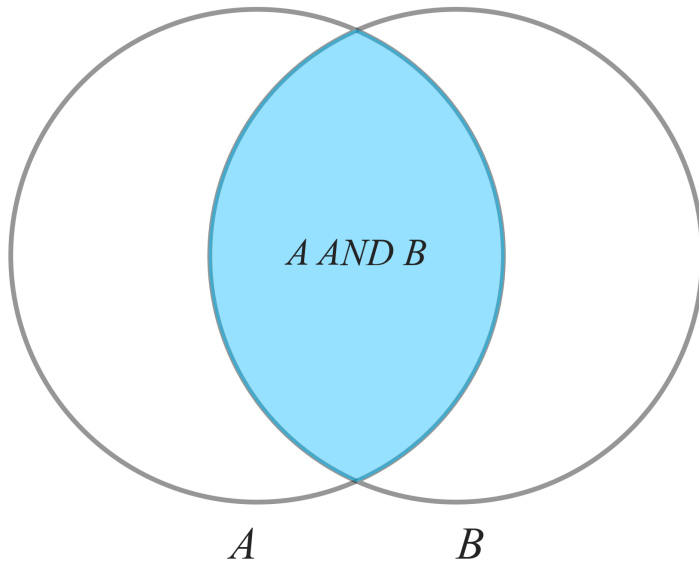
Opening Times (US)

Hour of the Day

> but maybe we should look at Canadian shops too…

*> let filter for **BOTH** countries*

A New Coffee Shop: Filter by Category

Lets us some Boolian logic :)

AND
(python: &)
Both terms

A AND B

A   *B*

OR
(python: |)
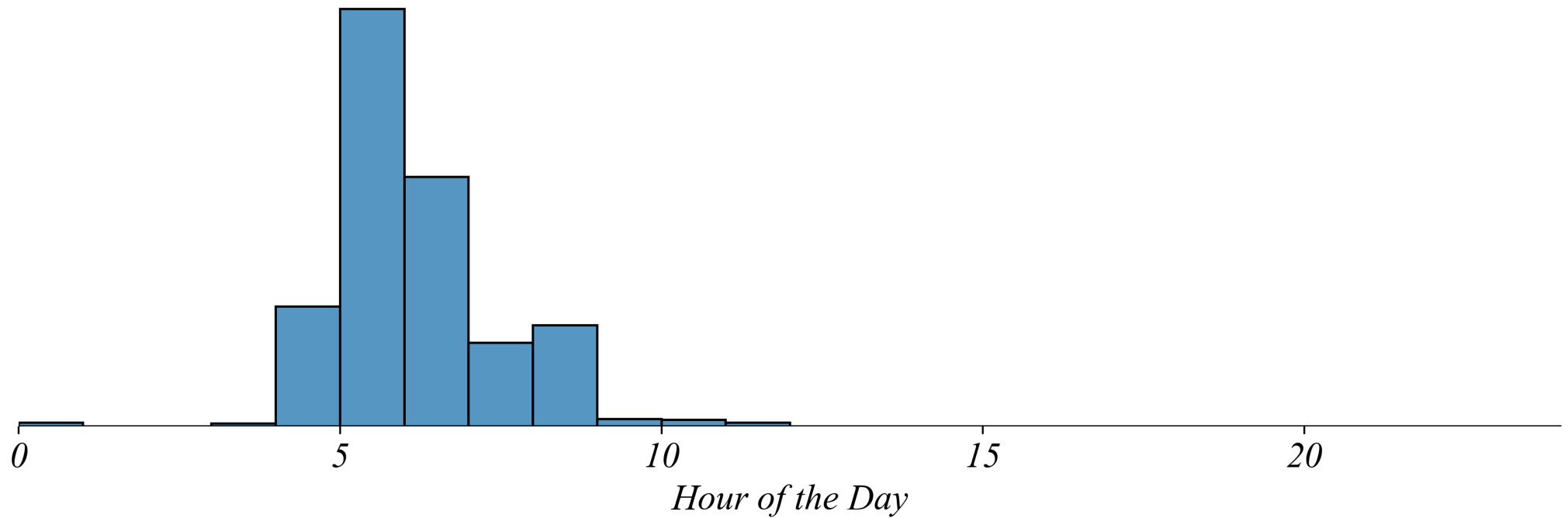Either term

A   *A AND B*   *B*

A   *B*
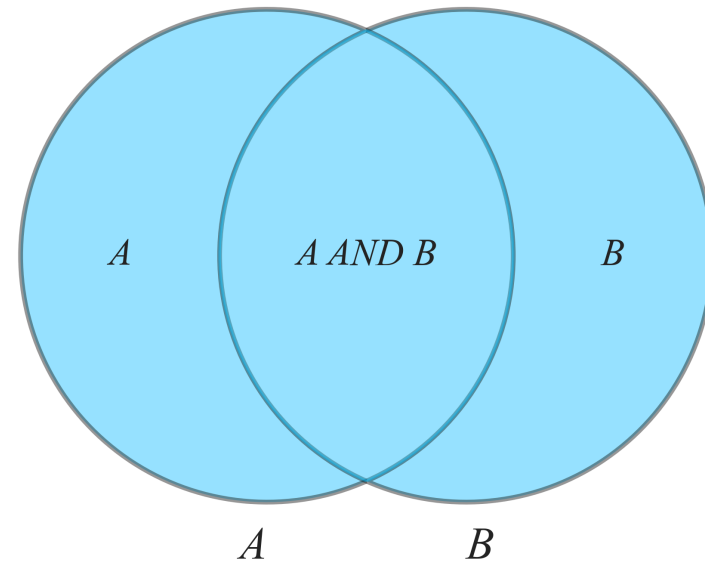
> *is there something different between the US and Canada?*

A New Coffee Shop: Filter by Category

Q. When might be a good time for the coffee shop to open?

Opening Times (US)

Hour of the Day

A New Coffee Shop: Filter by Category

Q. When might be a good time for the coffee shop to open?

Opening Times (in US or Canada)

Hour of the Day

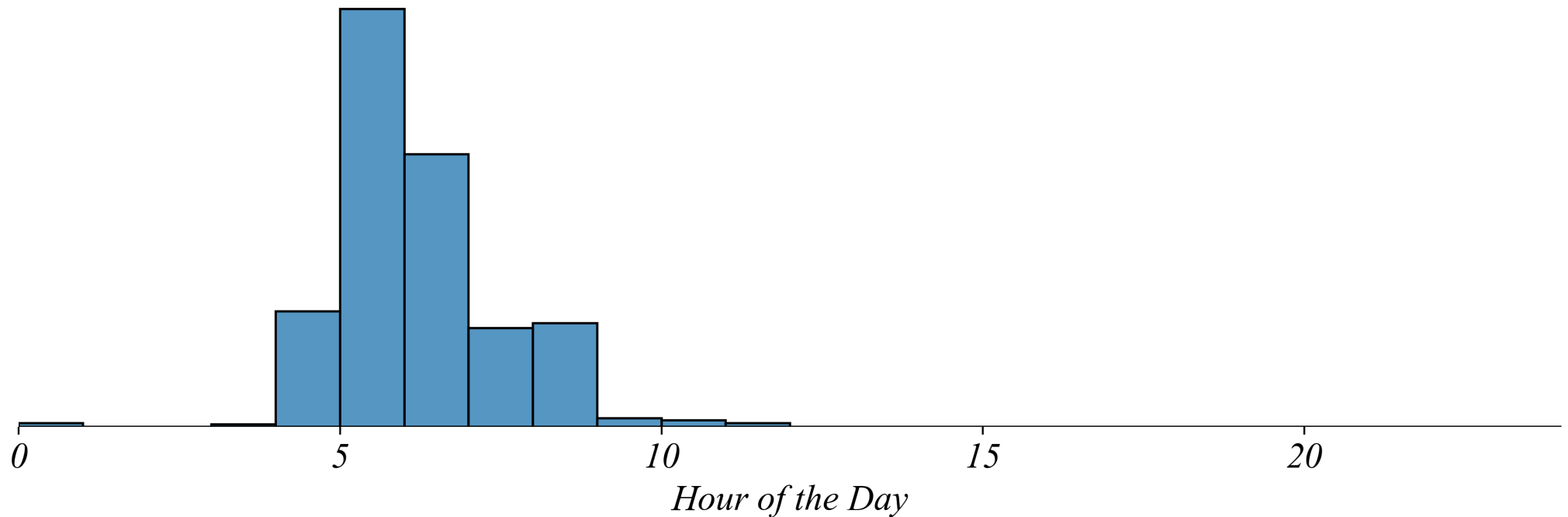*> so not much difference between when shops in the US and Canada open*

A New Coffee Shop: Filter by Category

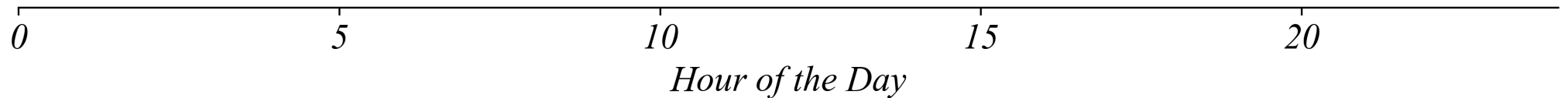*Q. When might be a good time for the coffee shop to open?*

Opening Times (shops in US AND Canada)

What would this histogram look like?

A New Coffee Shop: Filter by Category

Q. When might be a good time for the coffee shop to open?

Opening Times (shops in US AND Canada)

```
0          5          10         15         20
```
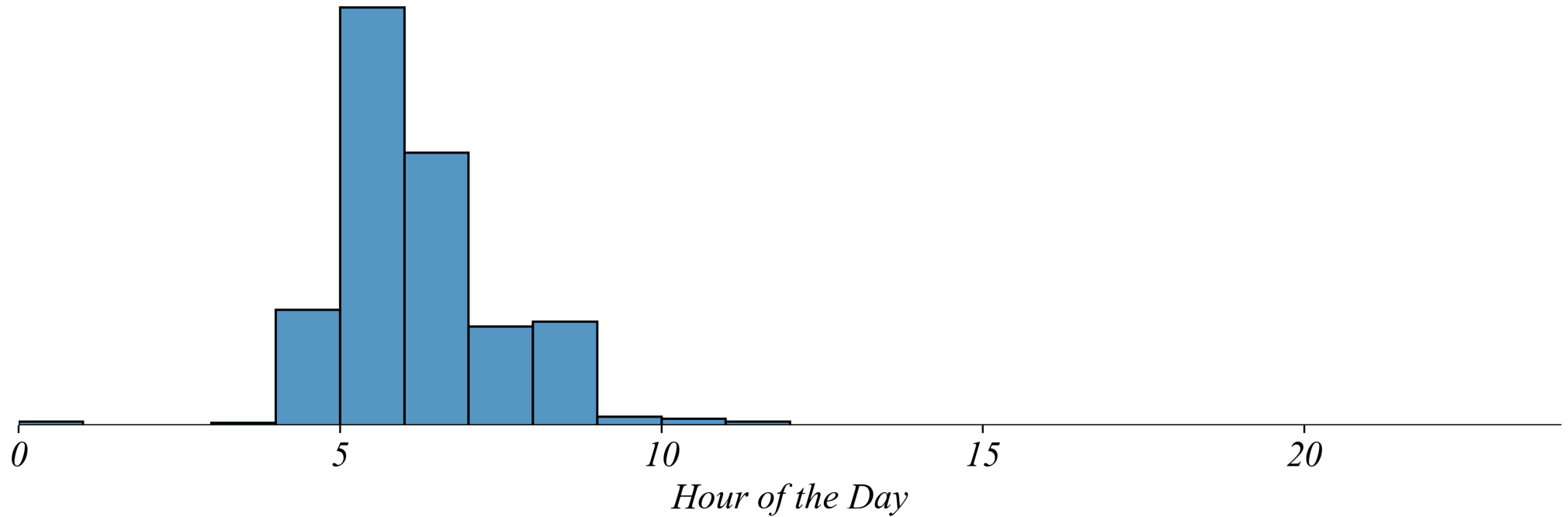
Hour of the Day

> *no data! no coffee shop can be in the US AND Canada!*

A New Coffee Shop: Filter by Category

Q. When might be a good time for the coffee shop to open?

Opening Times (in US or Canada)

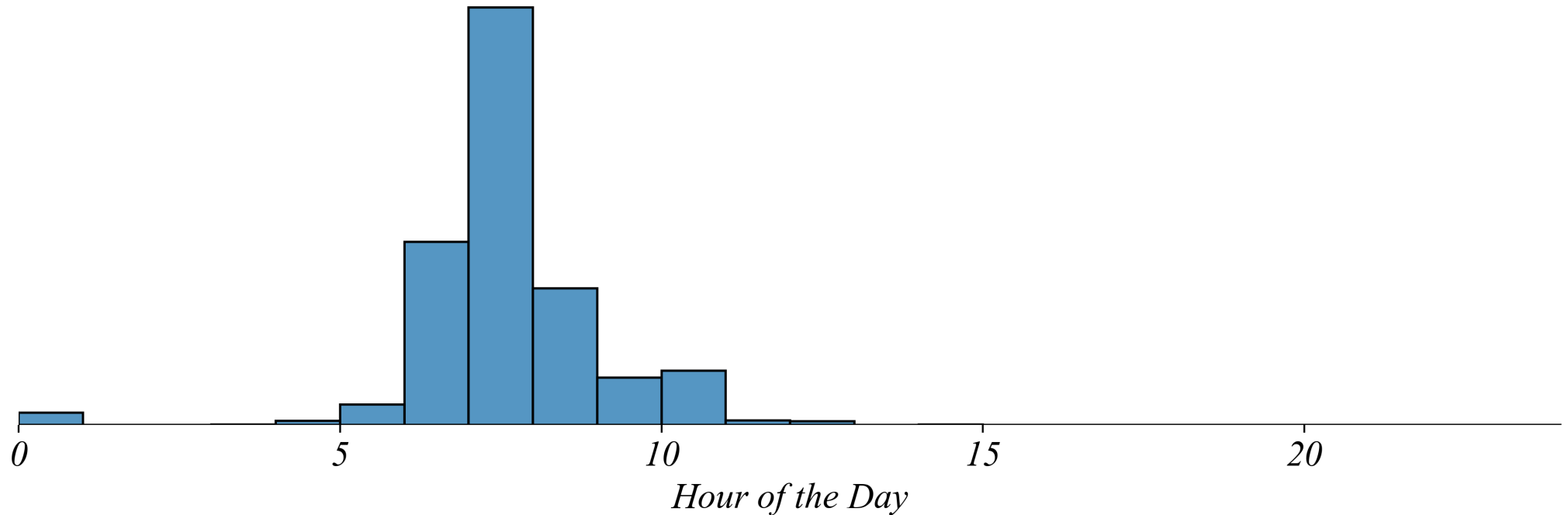Hour of the Day

A New Coffee Shop: Filter by Category

Q. When might be a good time for the coffee shop to open?

Opening Times (not in US or Canada)

Hour of the Day

> *so coffee shops in US and Canada open much earlier than the rest of the world*

A New Coffee Shop

Q. When might be a good time for the coffee shop to open?

- *So lets open at 5 AM.*

A New Coffee Shop

- *So lets open earlier than 7 AM.*
- *How long should we stay open?*

A New Coffee Shop

Q. How long might be good for the coffee shop to stay open?

Duration

Hours Open
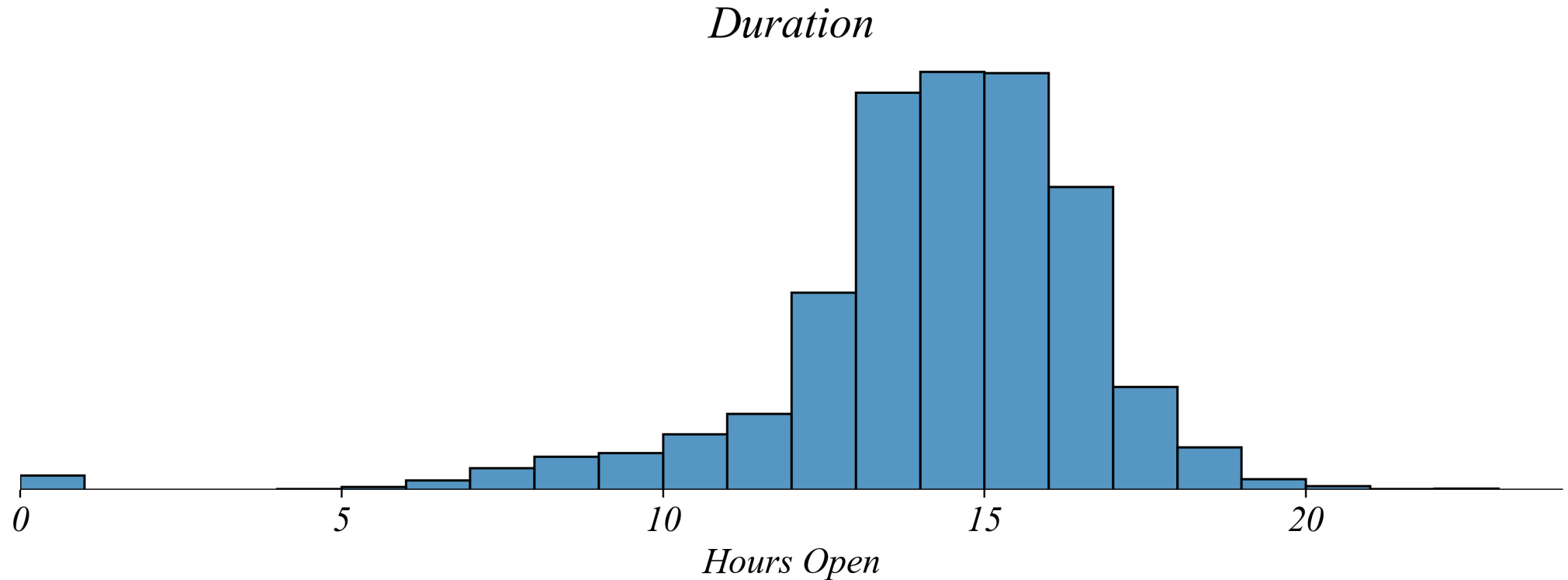
> *so most shops stay open for around 15 hours*

> *does that mean we should stay open for 15 hours?*

A New Coffee Shop

Q. How long might be good for the coffee shop to stay open?

Duration

Hours Open

> *lets filter for coffee shops that open before 7 AM*

A New Coffee Shop: Filter by Inequality

Q. How long might be good for the coffee shop to stay open?
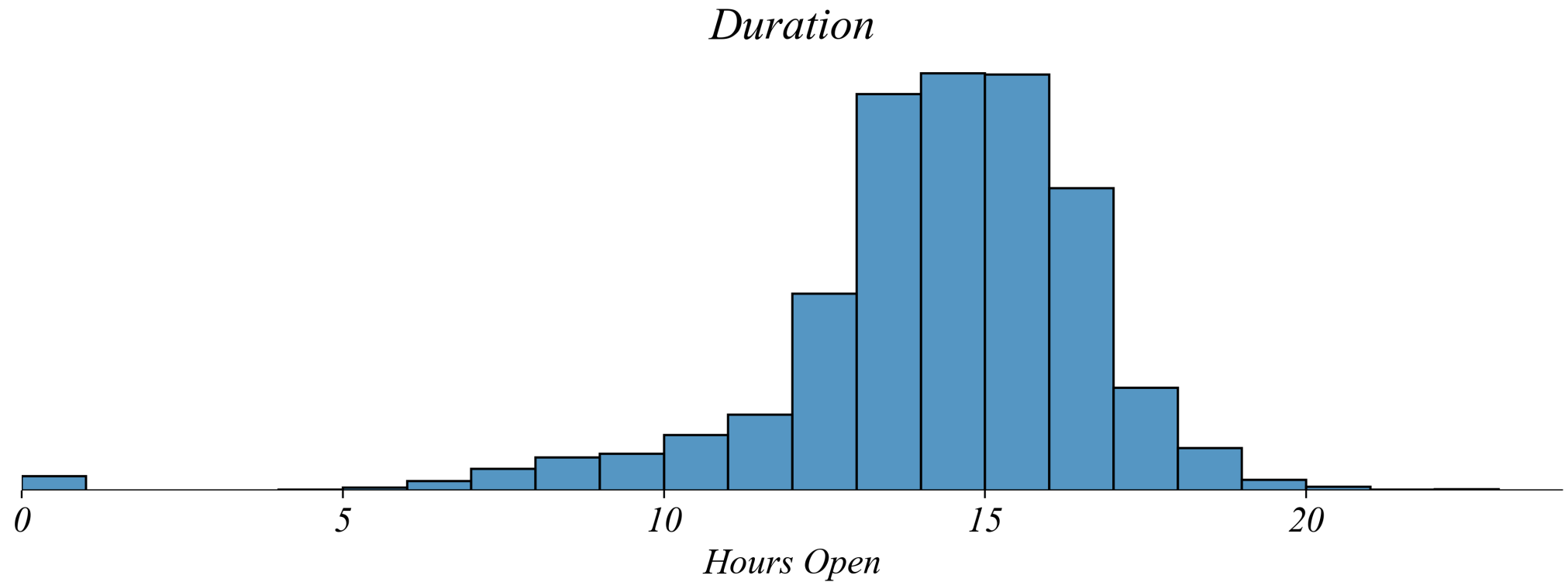
Duration

Hours Open

A New Coffee Shop: Filter by Inequality

Q. How long might be good for the coffee shop to stay open?

Duration (open: earlier than 7AM)

Hours Open
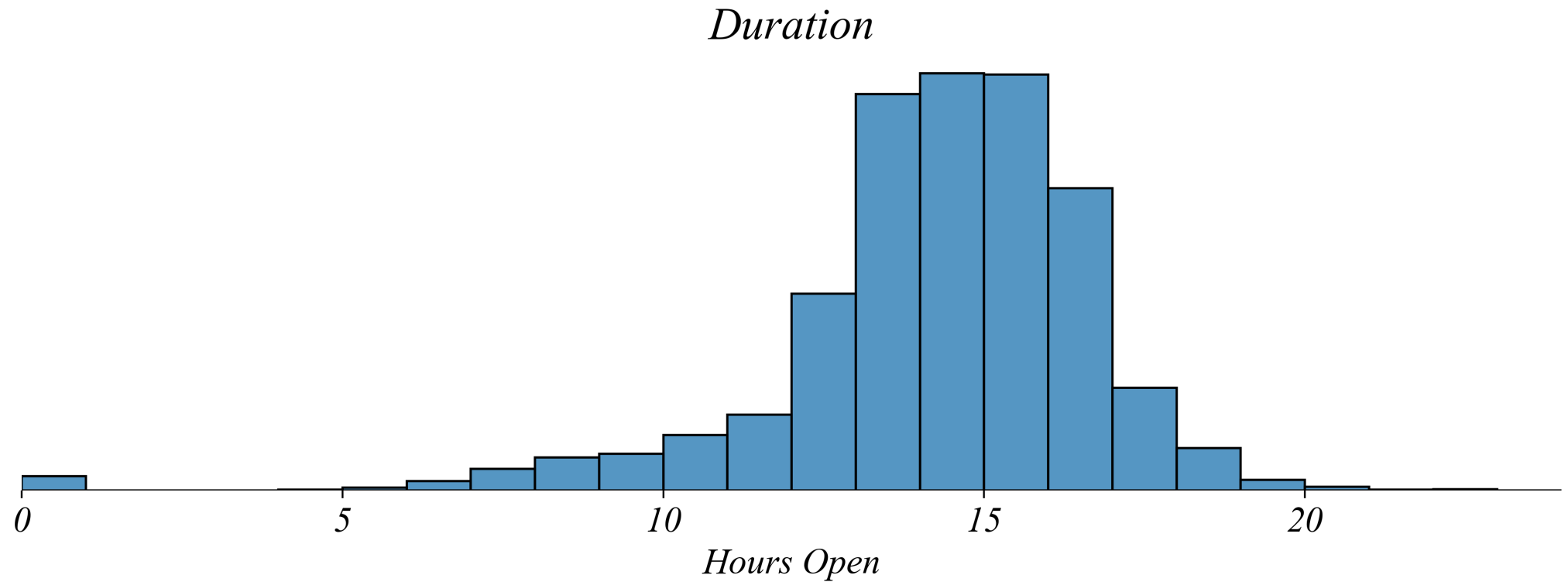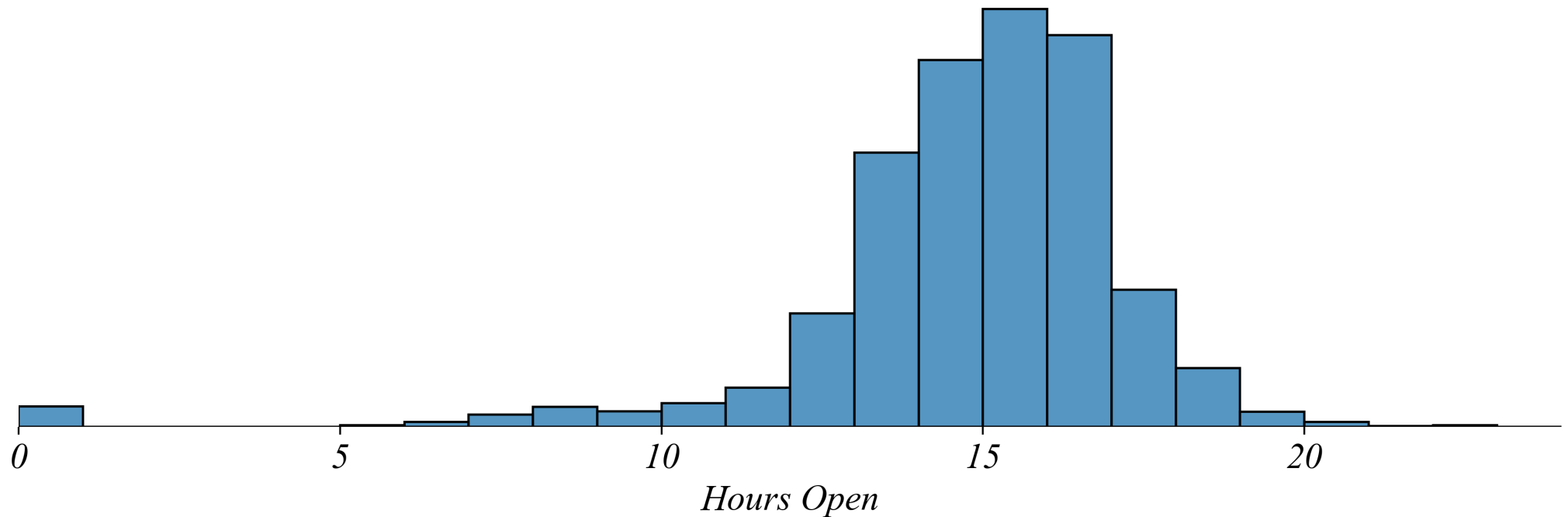
> *so shops that open early stay open longer*

A New Coffee Shop: Filter by Inequality

Q. How long might be good for the coffee shop to stay open?

Duration (open: 7AM or later)

Hours Open

> but here we're looking at all shops globally!

> our shop is opening in the US near Canada, so lets filter by country too

A New Coffee Shop: Filter by Inequality

Q. How long might be good for the coffee shop to stay open?

Duration (open: 7AM or later)

Hours Open

A New Coffee Shop: Filter by Inequality

Q. How long might be good for the coffee shop to stay open?
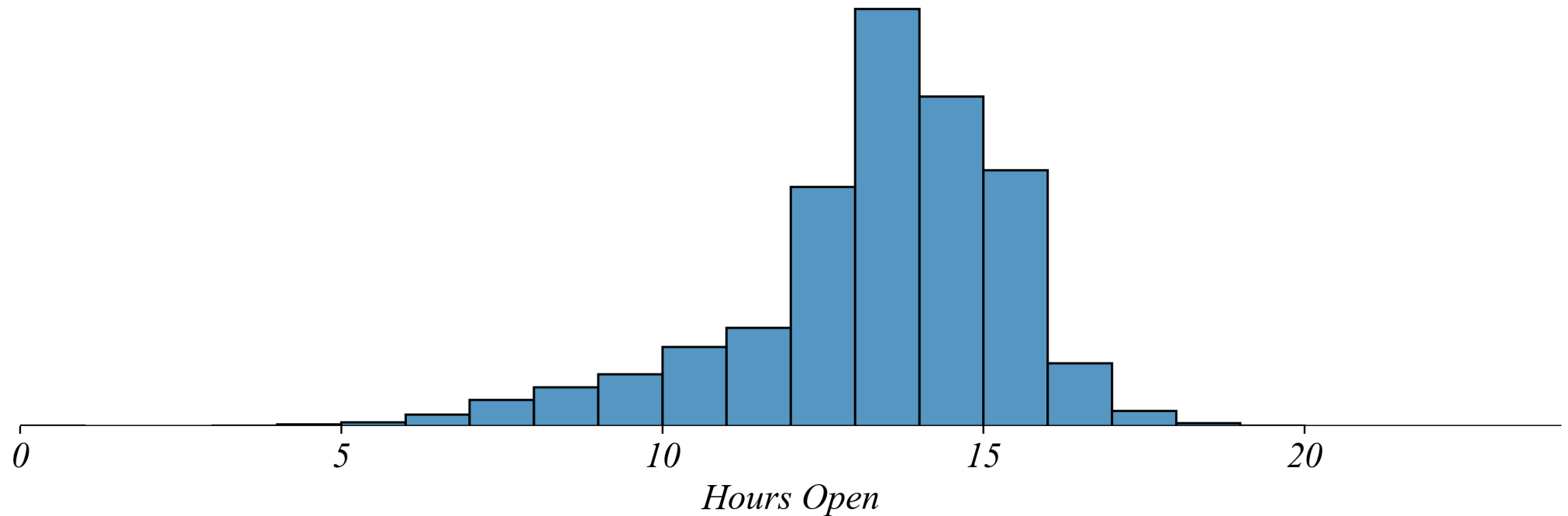
Duration (open: earlier than 7AM; in US or CA)
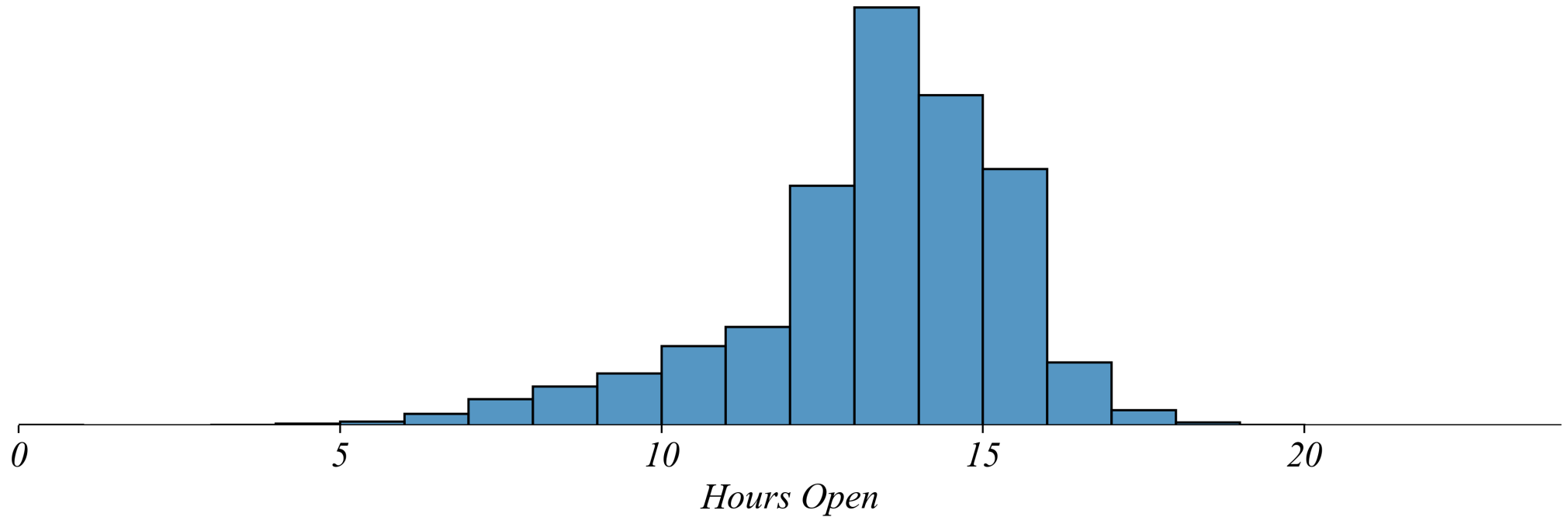

Hours Open
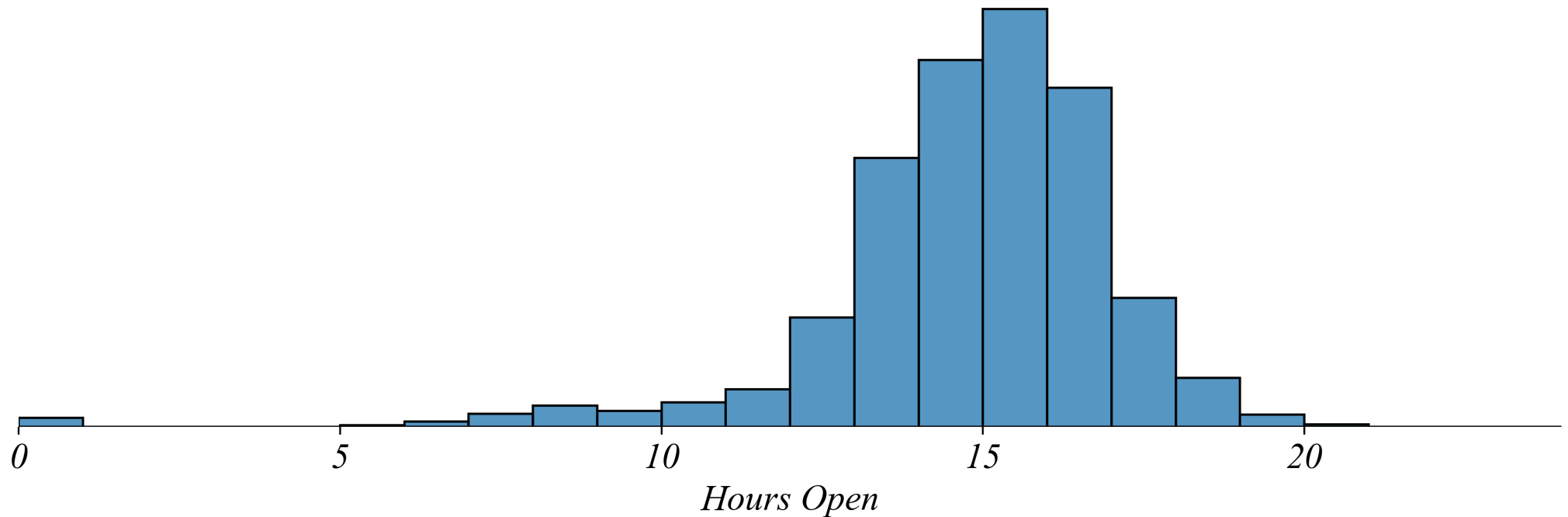
> *shops that open early will stay open longer in the US or Canada*

> *this is hard to see: maybe there's a more systematic way of showing differences*

A New Coffee Shop: Filter by Inequality

Q. How long might be good for the coffee shop to stay open?

Duration

Legend:
- All
- Early
- Late
- Early; in US or CA

Hour of the Day
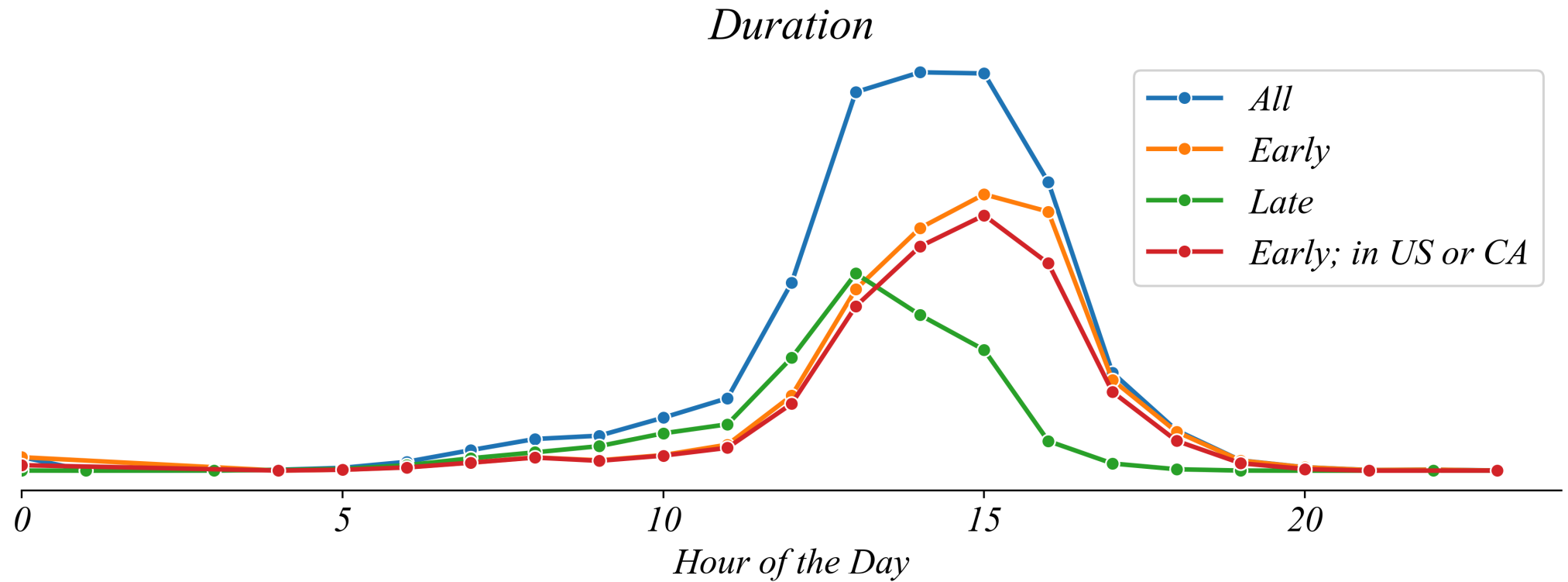
Exercise 1.5 | Coffee Shop Hours

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

- *The coffee shop is opening in update New York, near the border to Canada.*

- *You're asked to help make some decisions about how to run the shop when it opens.*

- *The dataset* `Starbucks_Location_Hours.csv` *contains information about Starbucks coffee shops globally.*

Coffee Shop Hours: load the data

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

```python
1  # Load the data
2  data = pd.read_csv(file_path + file_name)
```
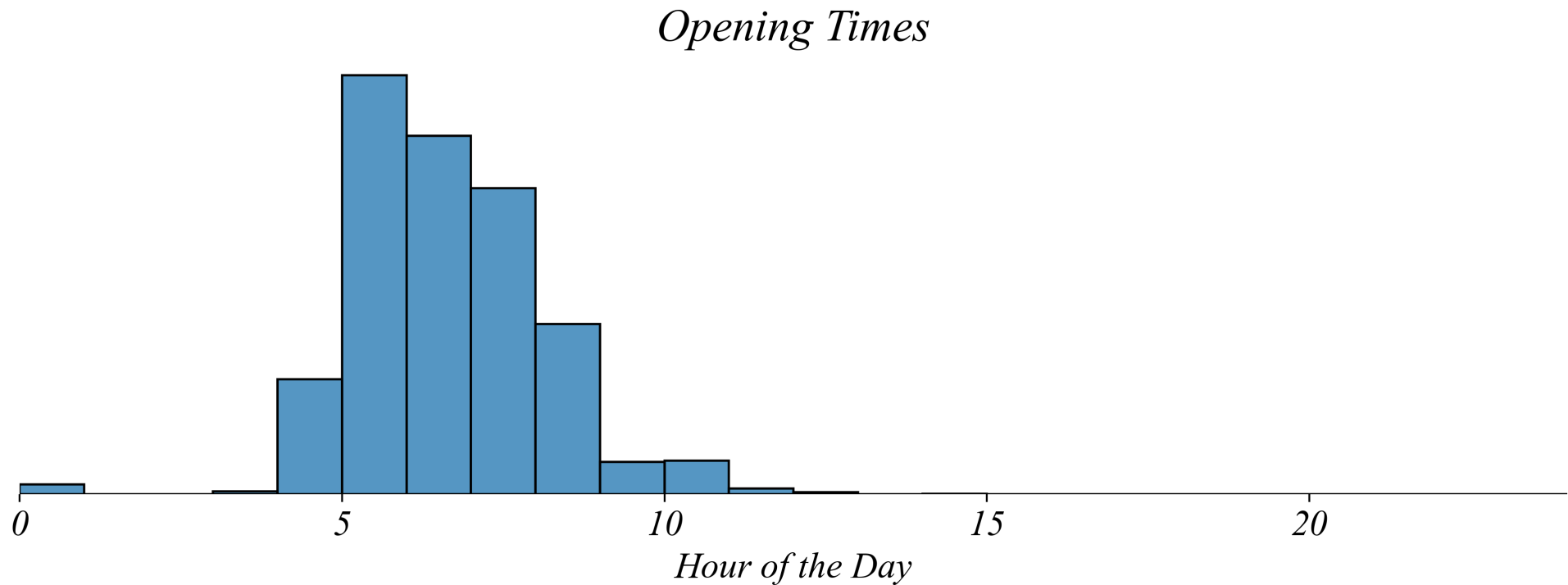
Coffee Shop Hours

Q. When might be a good time for the coffee shop to open?

Coffee Shop Hours: plot all opening times

Q. When might be a good time for the coffee shop to open?

```
1  # Histogram
2  sns.histplot(hours, x='open', bins=range(0,25,1))
```

Opening Times

Hour of the Day

Filtering Data by Category

Filtering categorical data requires logical operations.

Logic	Python	Example
Equals	==	`data[data['shop'] == 'A']`
Unequal	!=	`data[data['shop'] != 'A']`
NOT	~	`data[~(data['shop'] == 'A')]`
In list	.isin()	`data[data['shop'].isin(['A', 'B'])]`
AND	&	`(data['shop'] == 'A') & (data['open'] < 7)`
OR	|	`(data['shop'] == 'A') | (data['open'] < 7)`

Coffee Shop Hours: filter for US locations

Q. When might be a good time for the coffee shop to open?

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

Coffee Shop Hours: filter for US locations

Q. When might be a good time for the coffee shop to open?

```python
1  # Decide whether each row's country code is 'US'
2  us_codes = hours['COUNTRY_CODE'] == 'US'
3
4  # Select the rows with True
5  us_hours = hours[us_codes]
```
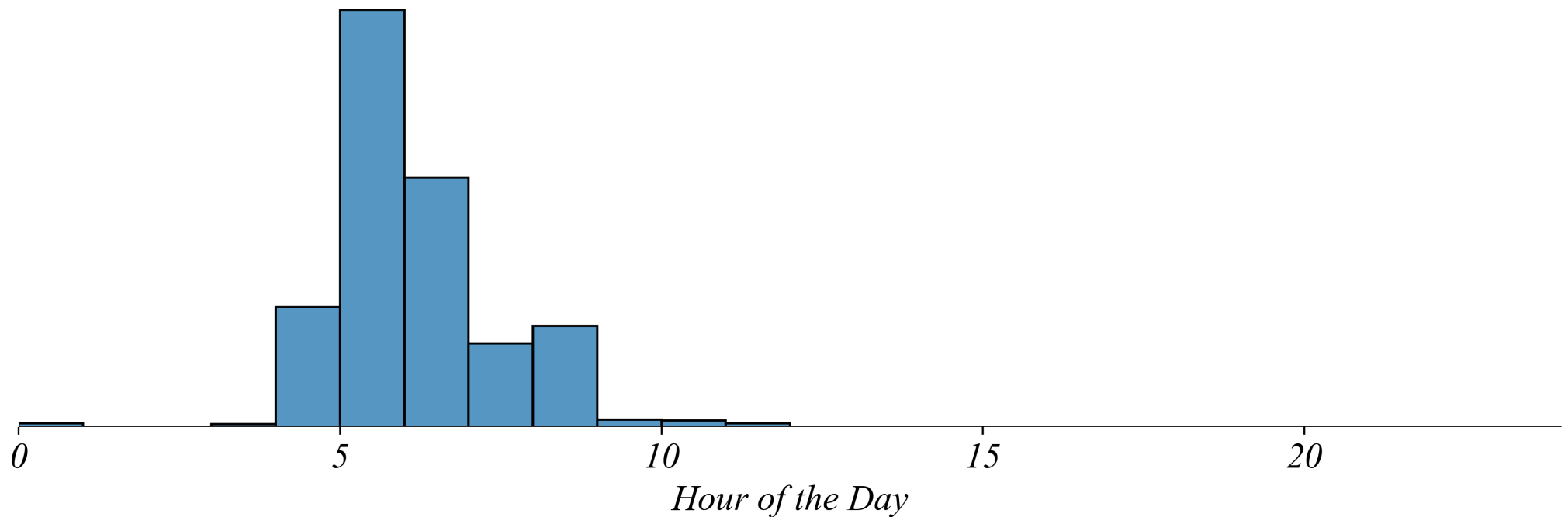
Coffee Shop Hours: filter for US locations

Q. When might be a good time for the coffee shop to open?

```python
1  # Histogram of US locations
2  sns.histplot(us_hours, x='open', bins=range(0,25,1))
```


Opening Times (US)

Coffee Shop Hours: shops in either US or CA

Q. When might be a good time for a US coffee shop to open?

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

Coffee Shop Hours: shops in either US or CA

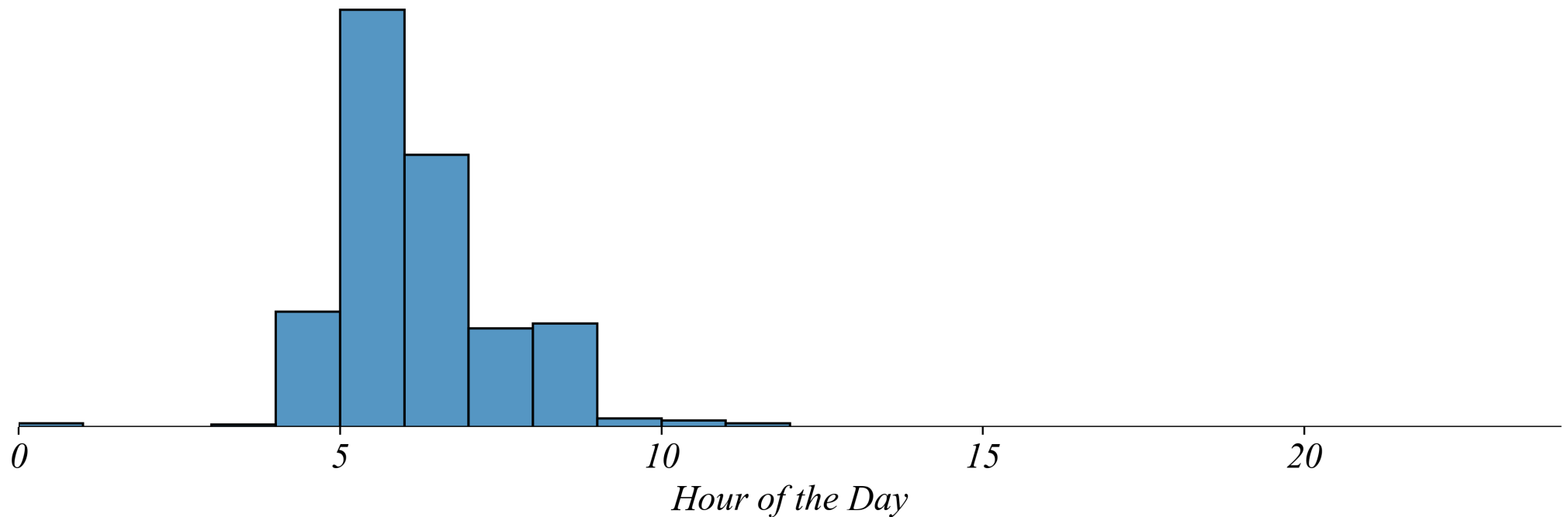*Q. When might be a good time for a US coffee shop to open?*

```
1  # Find the data in either the US or in Canada (CA)
2  # Method 2: Using isin()
3  us_ca = hours[hours['COUNTRY_CODE'].isin(['US', 'CA'])]
```

Coffee Shop Hours: shops in either US or CA

Q. When might be a good time for a US coffee shop to open?

```
1  # Create histogram
2  sns.histplot(us_ca, x='open', bins=range(0,25,1))
```


Opening Times (in US or Canada)

Hour of the Day

Coffee Shop Hours: shops in either US or CA

What would this dataset look like?

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

> *it would contain no data!*

Opening Times (shops in US AND Canada)

```
0                5               10               15               20
```

Hour of the Day
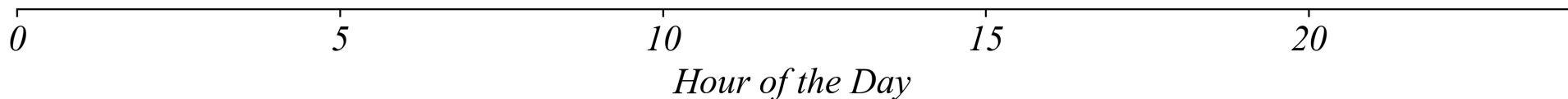
Filtering Data by Inequality

Filtering numerical data requires inequalities.

Symbol	Python	Example
=	==	`data[data['open'] == 7]`
≠	!=	`data[data['open'] != 7]`
<	<	`data[data['open'] < 7]`
>	>	`data[data['open'] > 7]`
≤	<=	`data[data['open'] <= 7]`
≥	>=	`data[data['open'] >= 7]`

Coffee Shop Hours: filter for early opening shops

Q. How long might be good for the coffee shop to stay open?
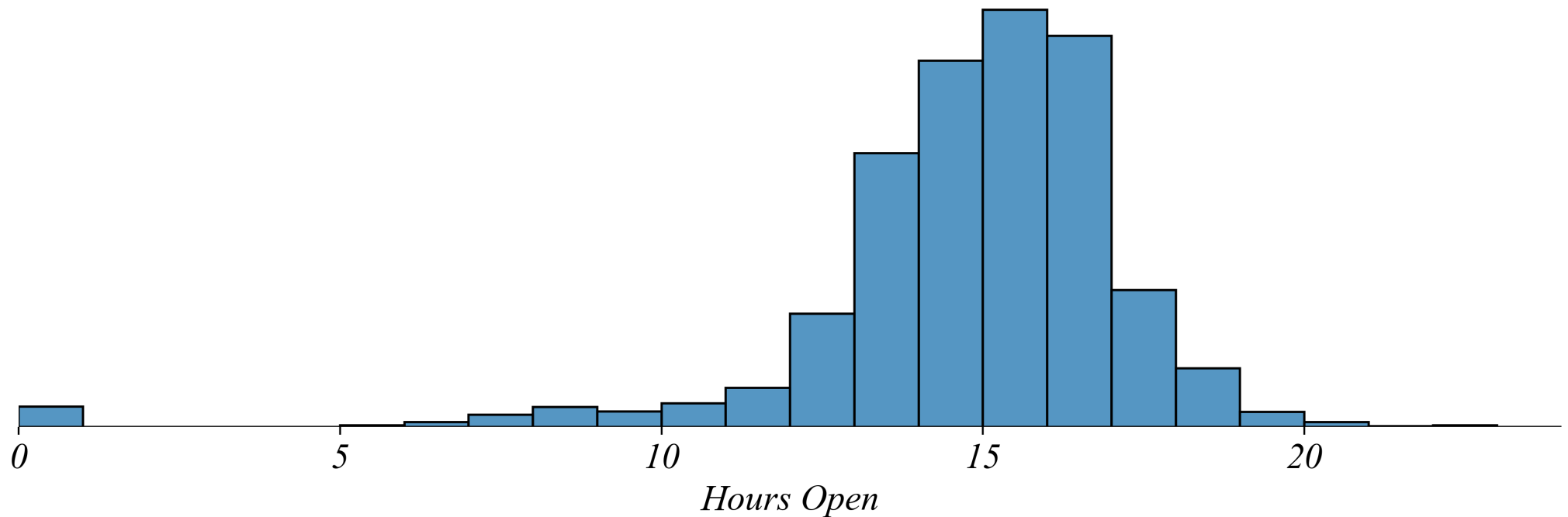
```python
1  # Filter for shops that open before 7 AM
2  early_hours = hours[hours['open'] < 7]
```

Coffee Shop Hours: filter for early opening shops

Q. How long might be good for the coffee shop to stay open?

```
1  # Create histogram of duration for early-opening shops
2  sns.histplot(early_hours, x='duration_hr', bins=range(0,25,1))
```

Duration (open: earlier than 7AM)

Hours Open

Coffee Shop Hours: combine filters

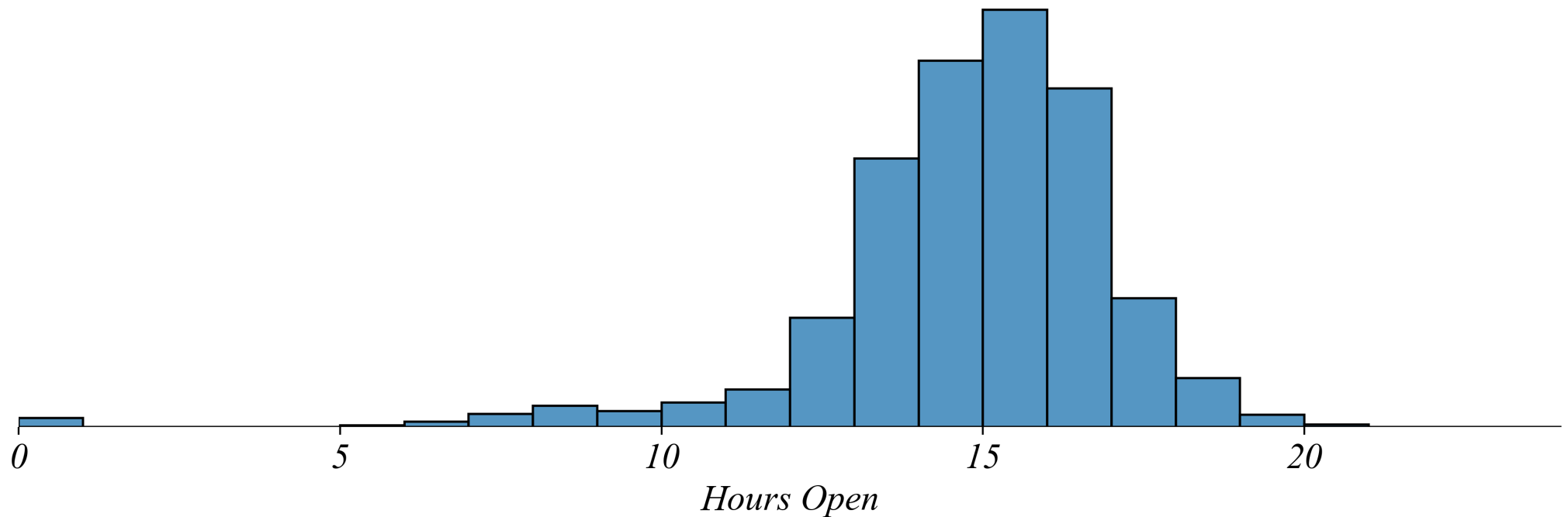*Q. How long might be good for the coffee shop to stay open?*

```python
# Filter for shops that open early AND are in US or Canada
early_us_ca = hours[(hours['open'] < 7) & (hours['COUNTRY_CODE'].isin(['US', 'CA']))]
```

Coffee Shop Hours: combine filters

Q. How long might be good for the coffee shop to stay open?

```
1  # Create histogram of duration for early-opening US/CA shops
2  sns.histplot(early_us_ca, x='duration_hr', bins=range(0,25,1))
```

Duration (open: earlier than 7AM; in US or CA)

Hours Open

Coffee Shop Hours: compare opening times

Q. How long might be good for the coffee shop to stay open?

```python
1  # Compare early vs all shops in US/CA
2  sns.histplot(us_ca, x='duration_hr', bins=range(0,25,1), label='All US/CA')
3  sns.histplot(early_us_ca, x='duration_hr', bins=range(0,25,1), label='Early US/CA')
4  plt.legend()
```


Duration

Legend:
- *All*
- *Early*
- *Late*
- *Early; in US or CA*

Hour of the Day
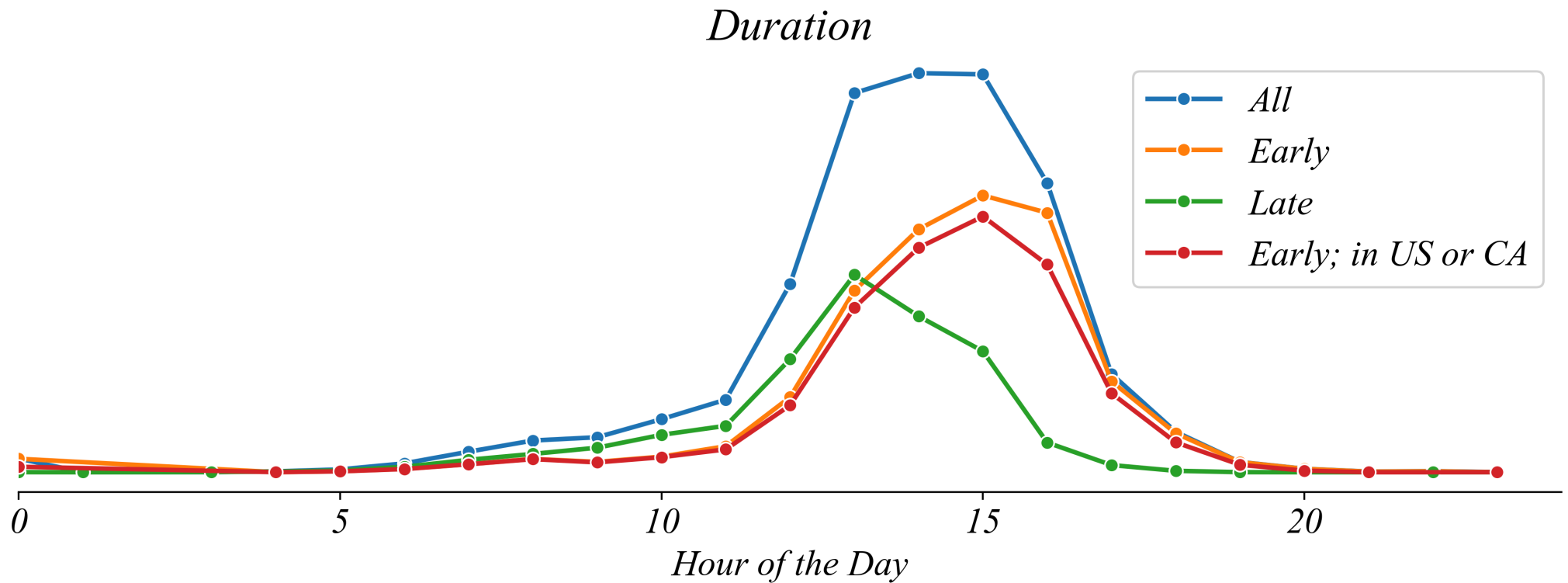
Coffee Shop Hours: recommendation

Q. How long might be good for the coffee shop to stay open?

- ***Opening time****: Before 7 AM (around 5-6 AM)*
- ***Duration****: About 16-17 hours (based on early-opening US/CA shops)*
- ***Closing time****: Around 9-11 PM*

> this matches what successful coffee shops do in similar markets