ECON 0150 | Economic Data Analysis

The economist's data analysis pipeline.

Part 1.5 | Filtering Data

A New US Coffee Shop

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

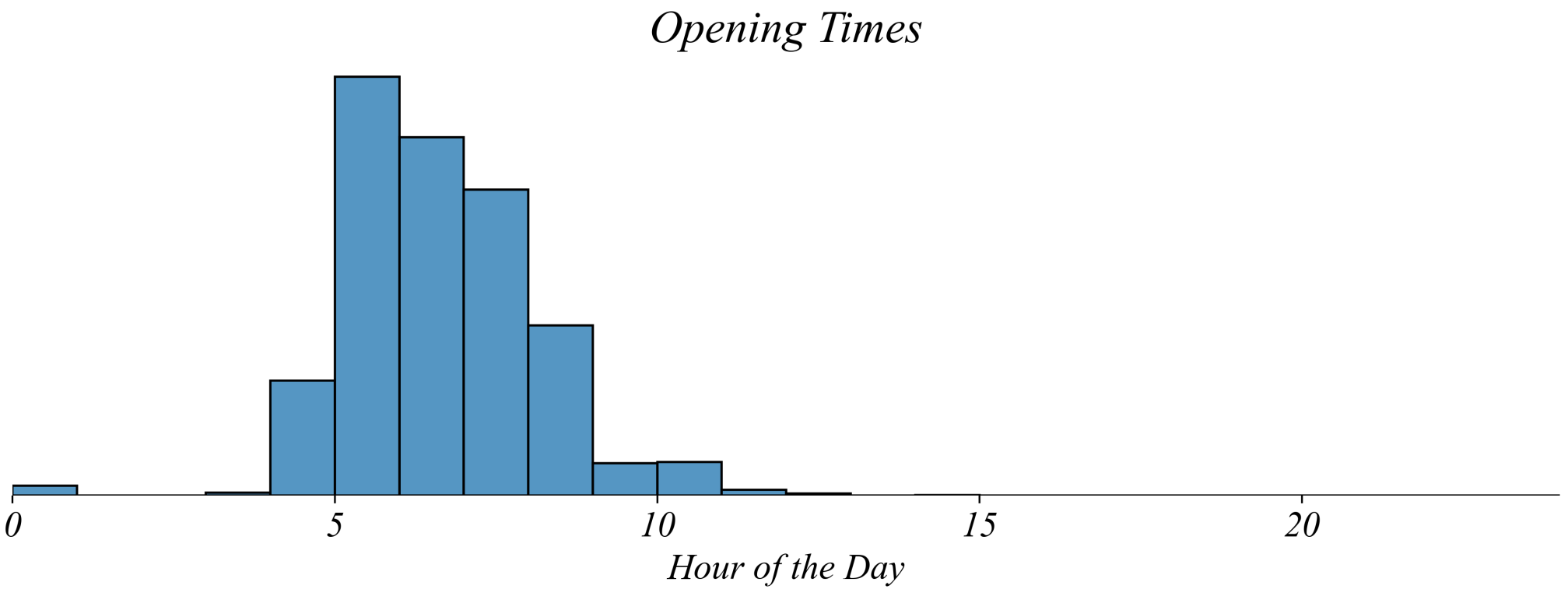
- *The coffee shop is opening in update New York, near the border to Canada.*
- *You're asked to help make some decisions about how to run the shop when it opens.*
- *The dataset* `Starbucks_Location_Hours.csv` *contains information about Starbucks coffee shops globally.*

A New Coffee Shop

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

A New Coffee Shop

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

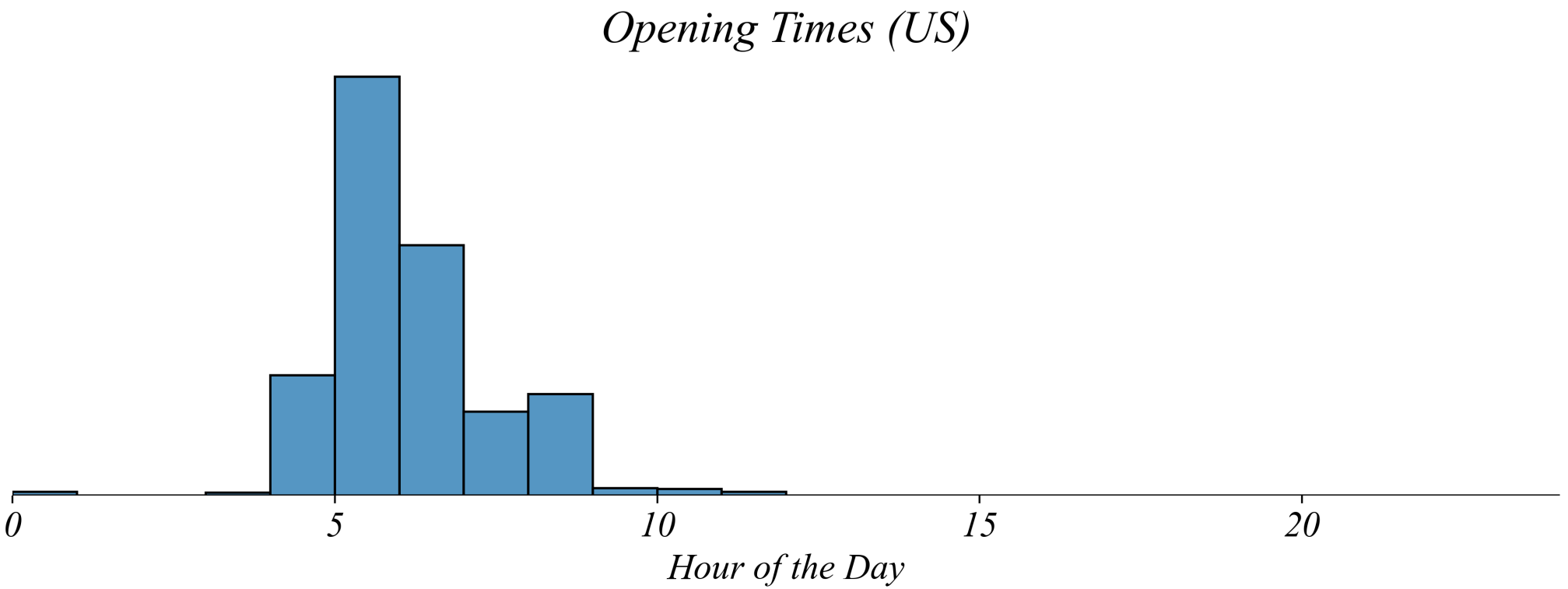


> so it seems best to open sometime in the morning… makes sense

> 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?

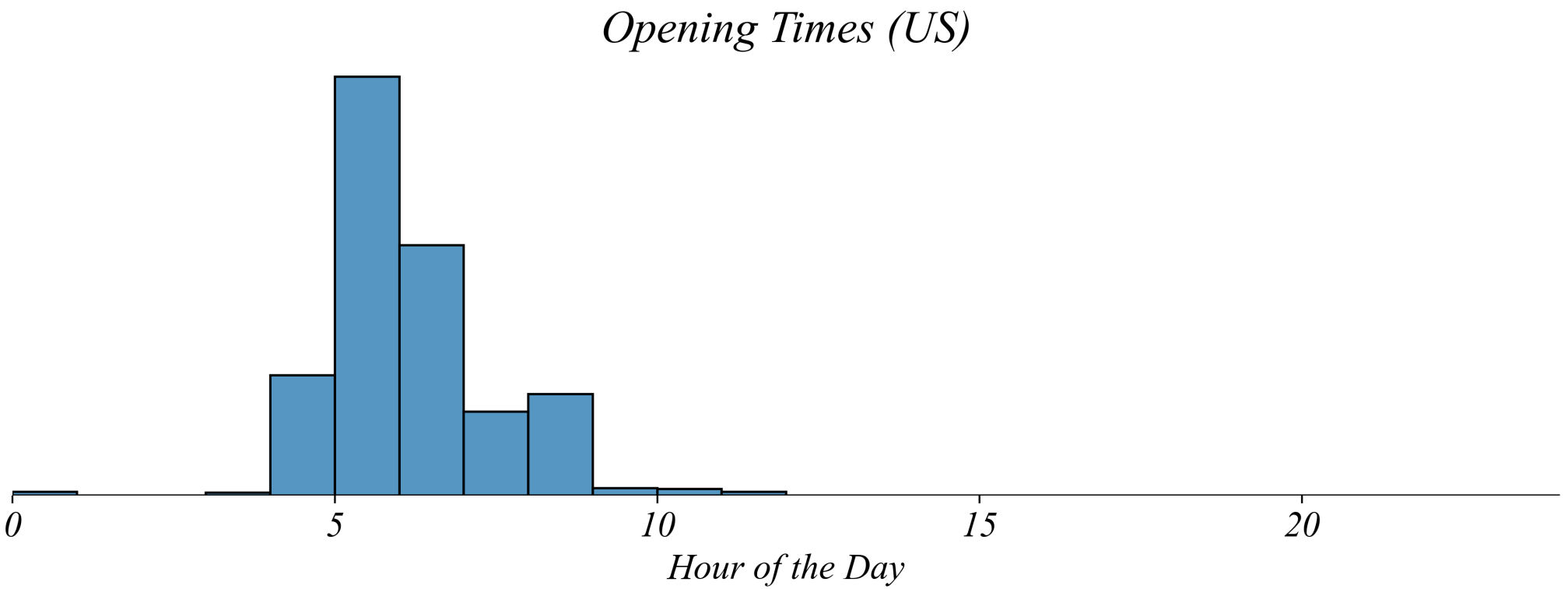


> *here we've filtered for US locations*

> *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?



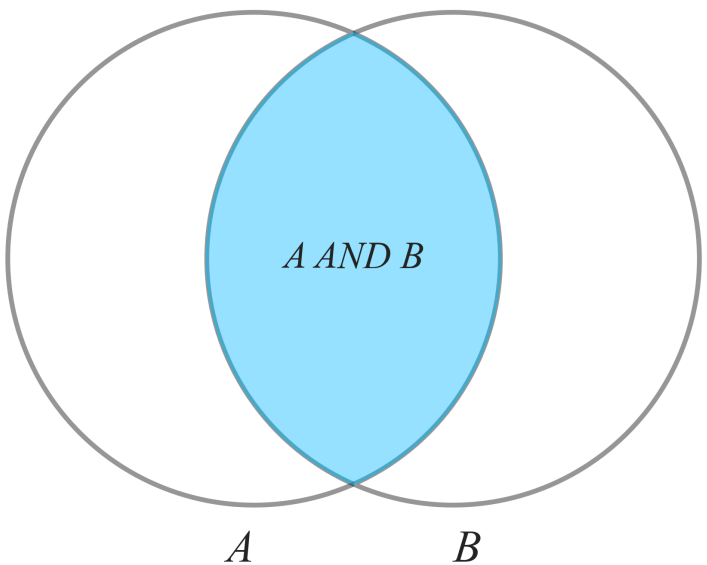
> *but maybe we should look at Canadian shops too…*

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

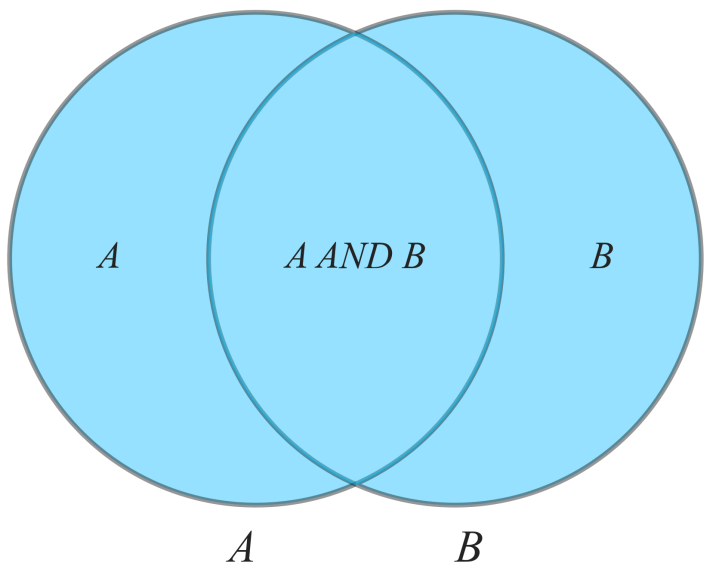
A New Coffee Shop: Filter by Category

Lets us some Boolian logic :)

AND
(python: &)
Both terms



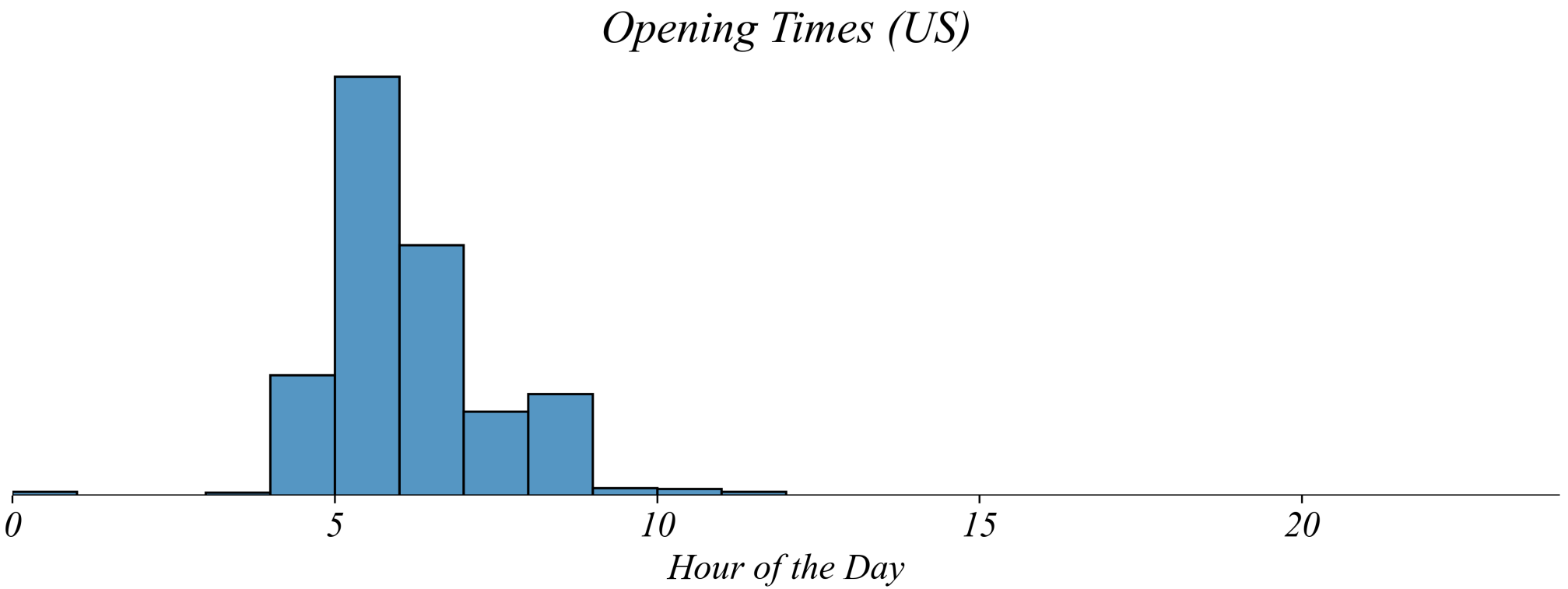
OR
(python: |)
Either term



> is there something different between the US and Canada?

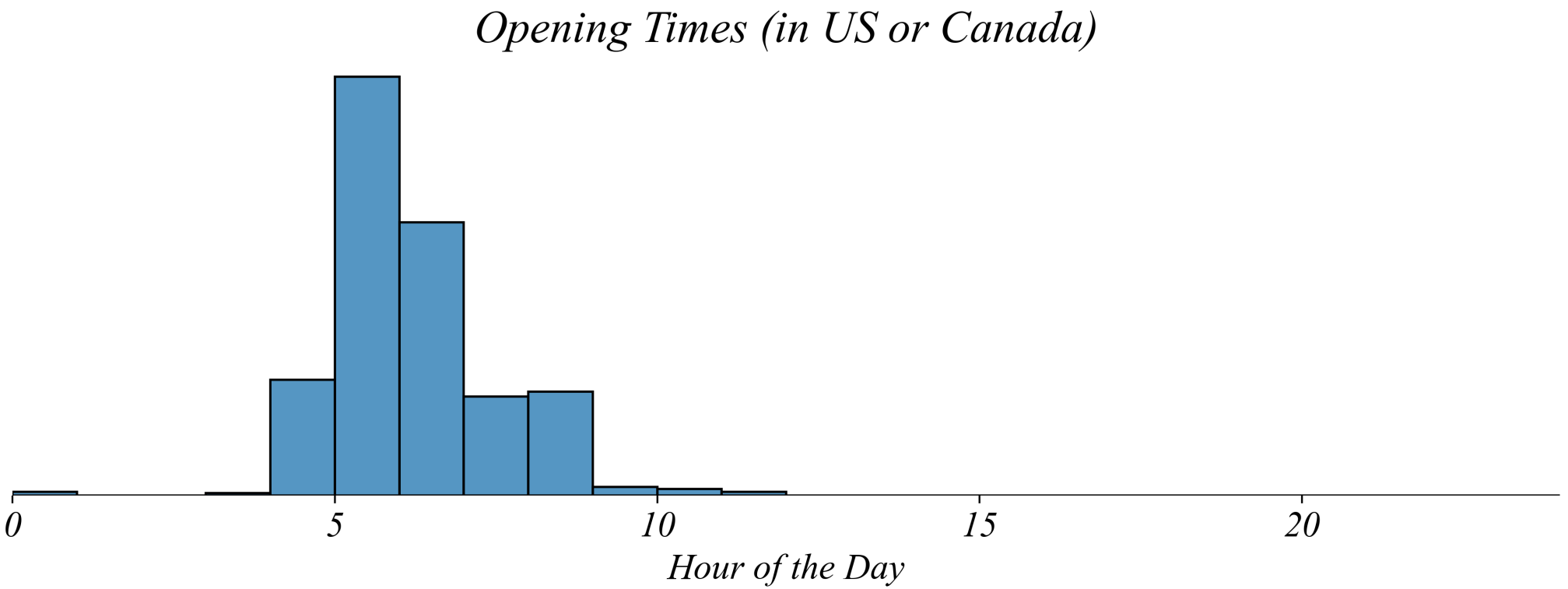
A New Coffee Shop: Filter by Category

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



A New Coffee Shop: Filter by Category

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



> 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?

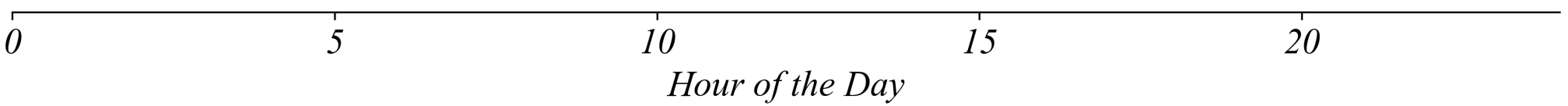
Opening Times (shops in US AND Canada)

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?

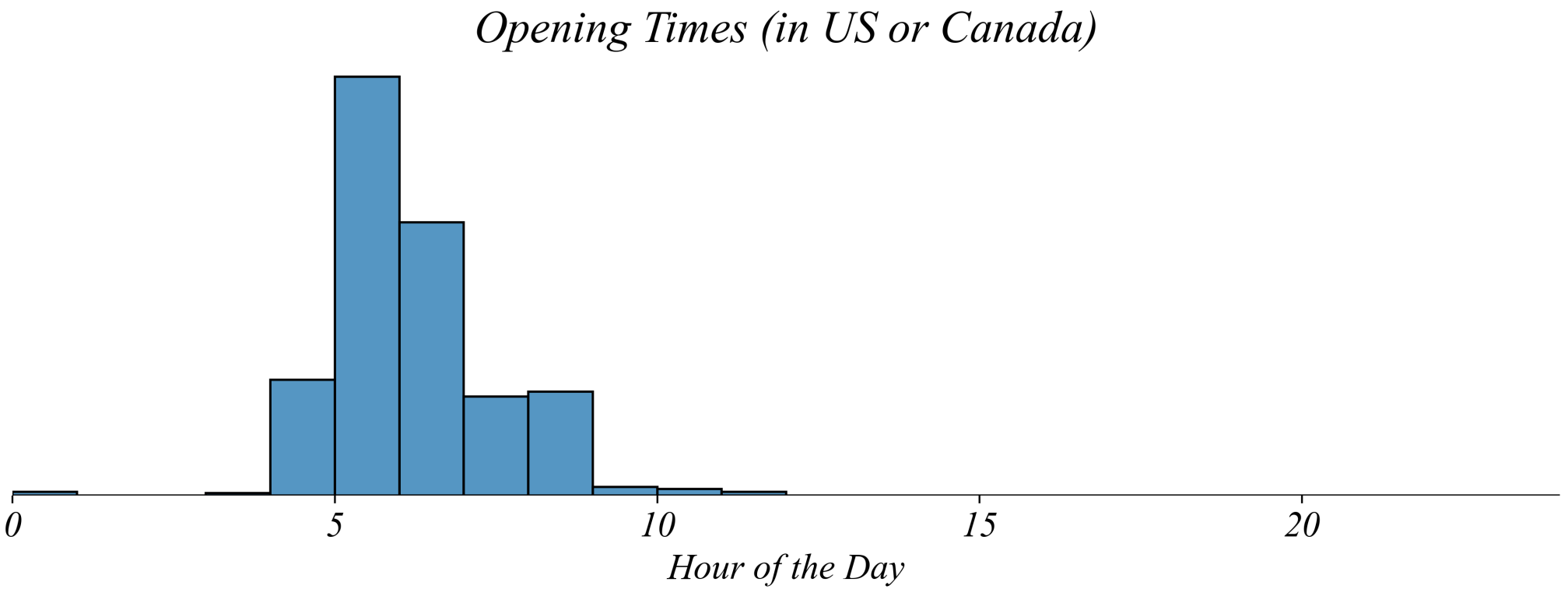
Opening Times (shops in US AND Canada)



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

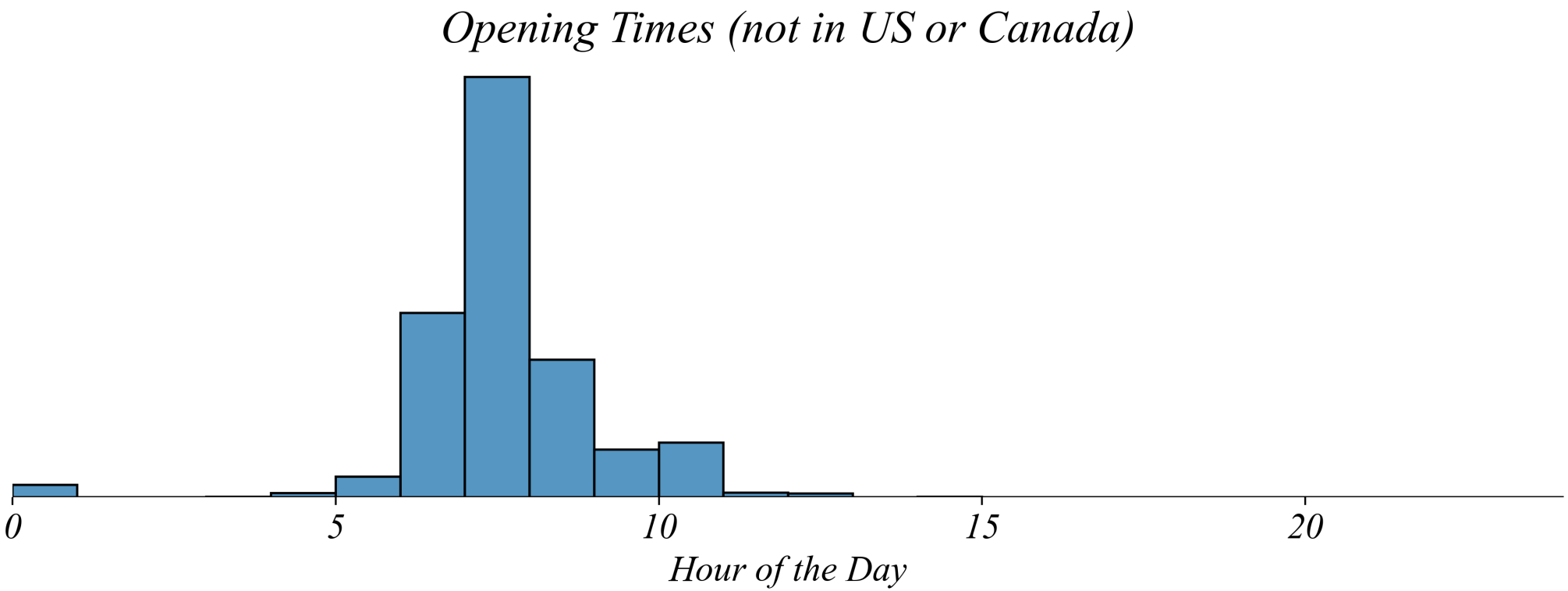
A New Coffee Shop: Filter by Category

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



A New Coffee Shop: Filter by Category

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



> 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.*

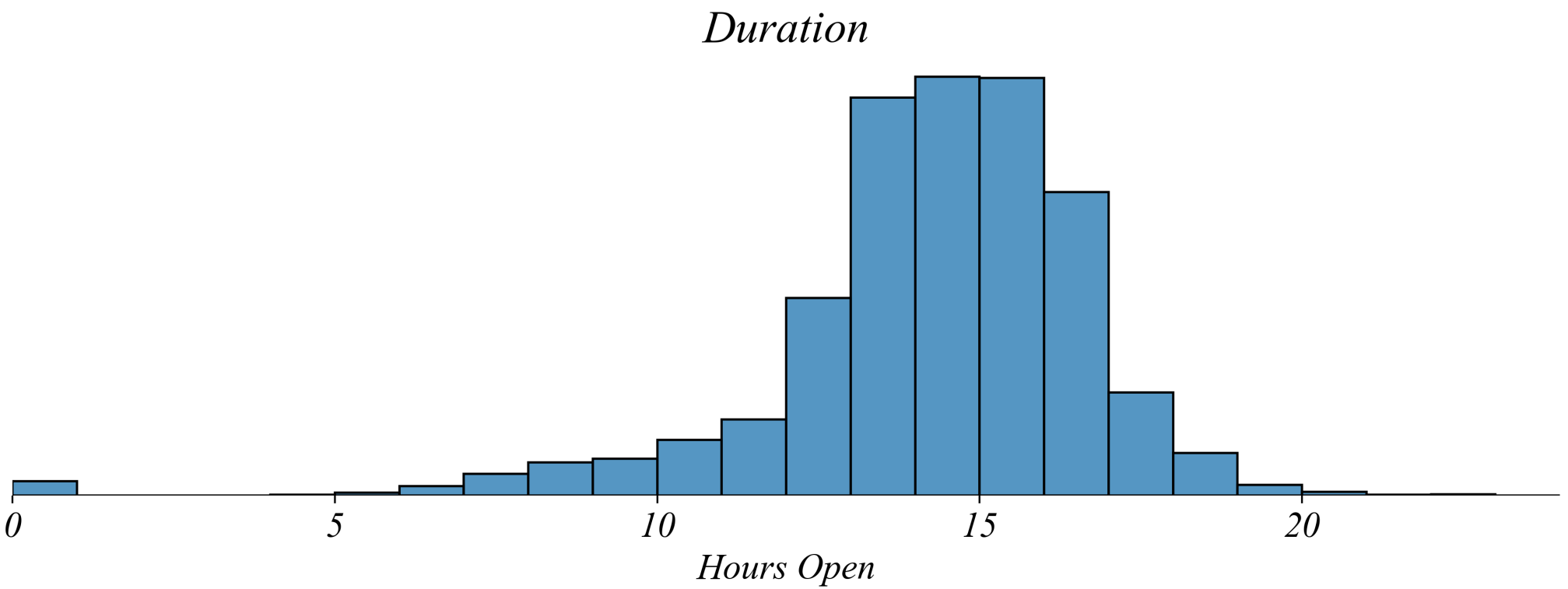
A New Coffee Shop

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

- *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?

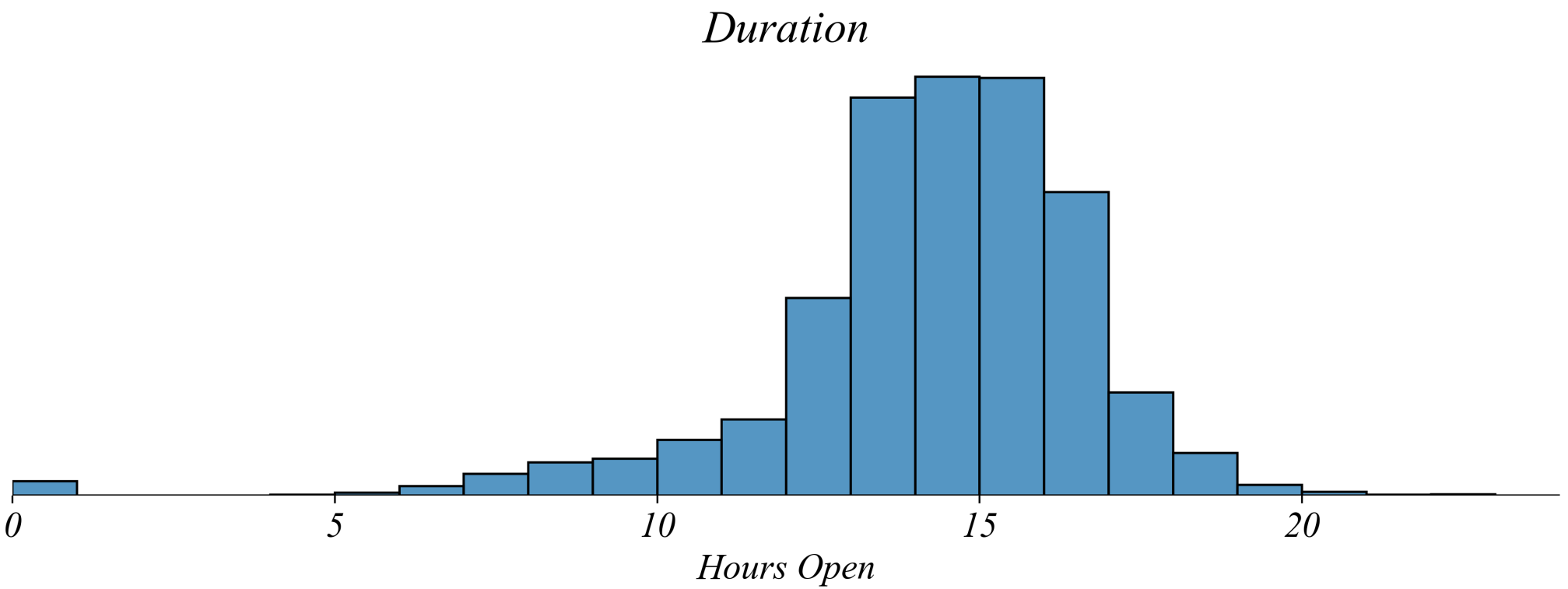


> so most shops stay open for around 15 hours

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

A New Coffee Shop

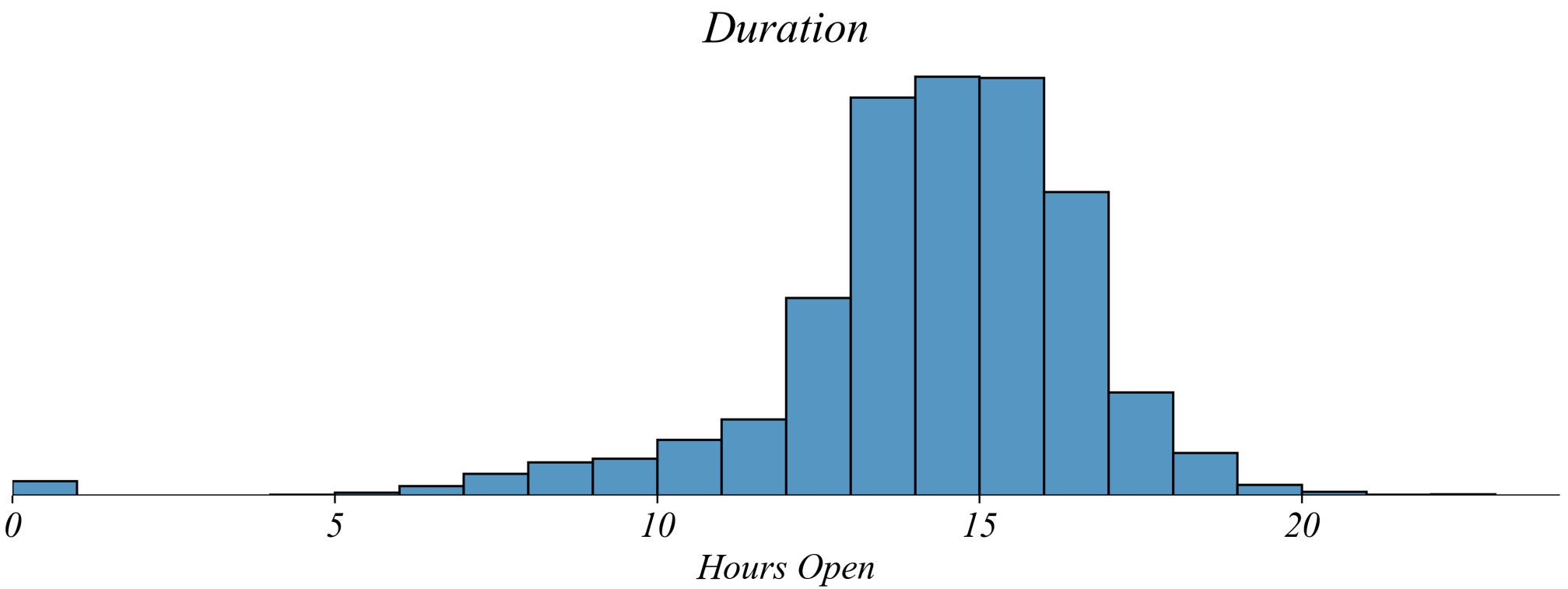
Q. How long might be good for the coffee shop to stay open?



> lets filter for coffee shops that open before 7 AM

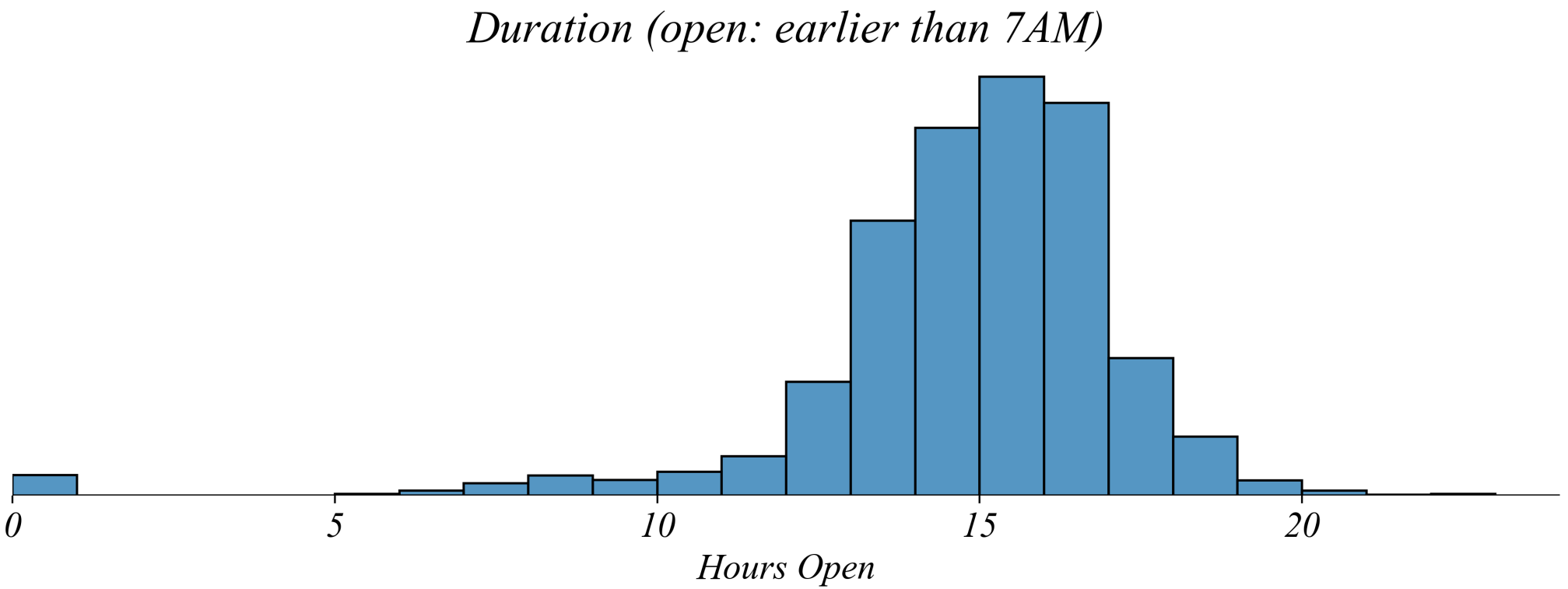
A New Coffee Shop: Filter by Inequality

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



A New Coffee Shop: Filter by Inequality

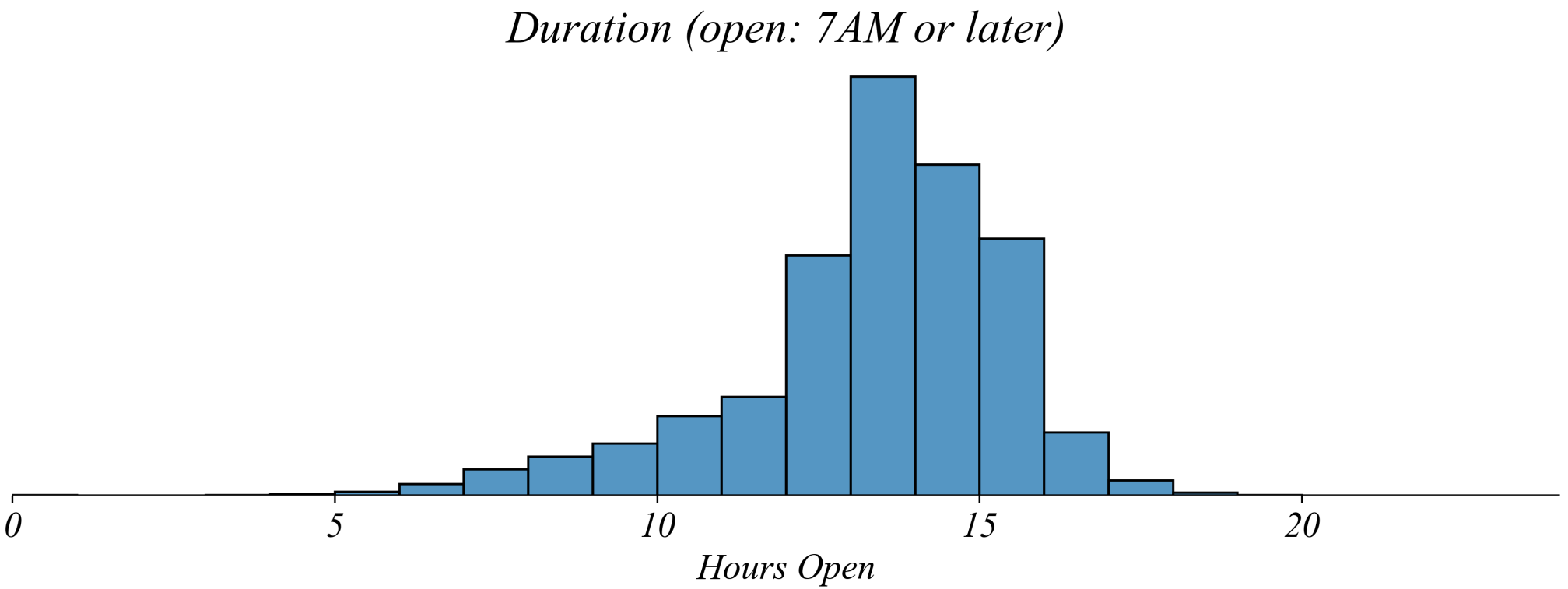
Q. How long might be good for the coffee shop to stay open?



> 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?

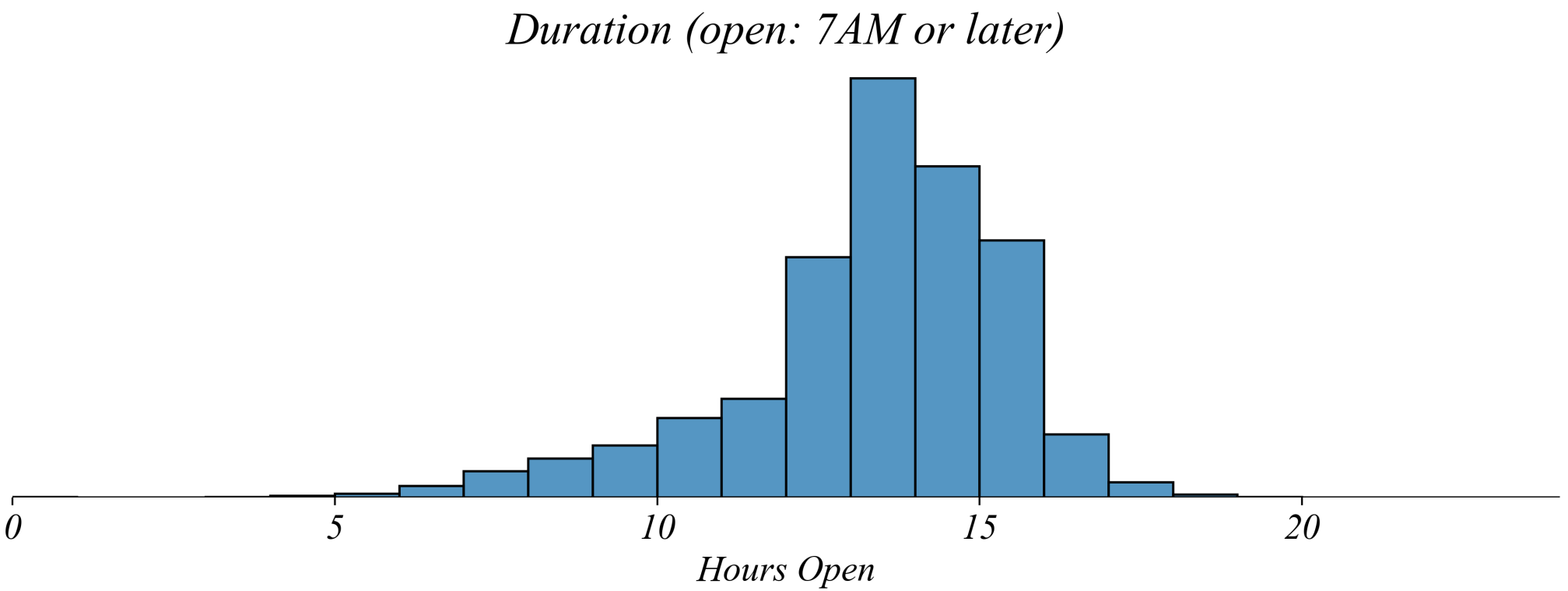


> but here we're looking at all shops globally!

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

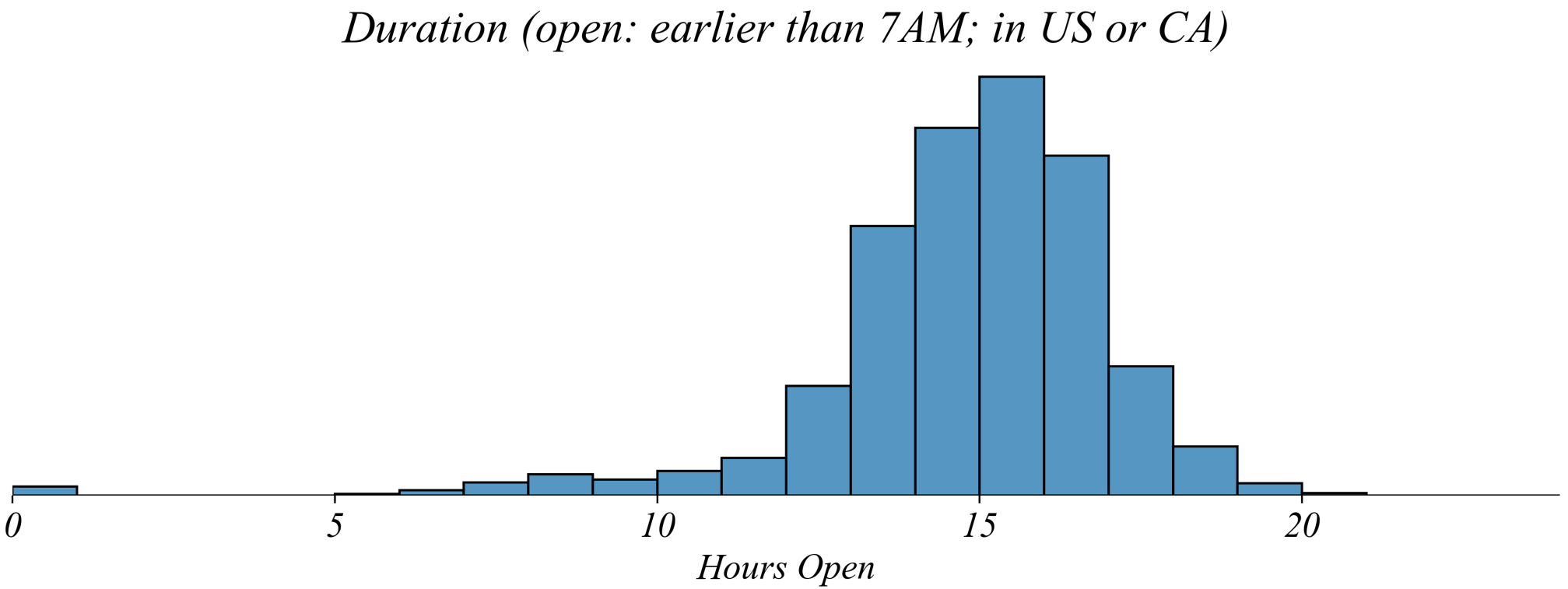
A New Coffee Shop: Filter by Inequality

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



A New Coffee Shop: Filter by Inequality

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

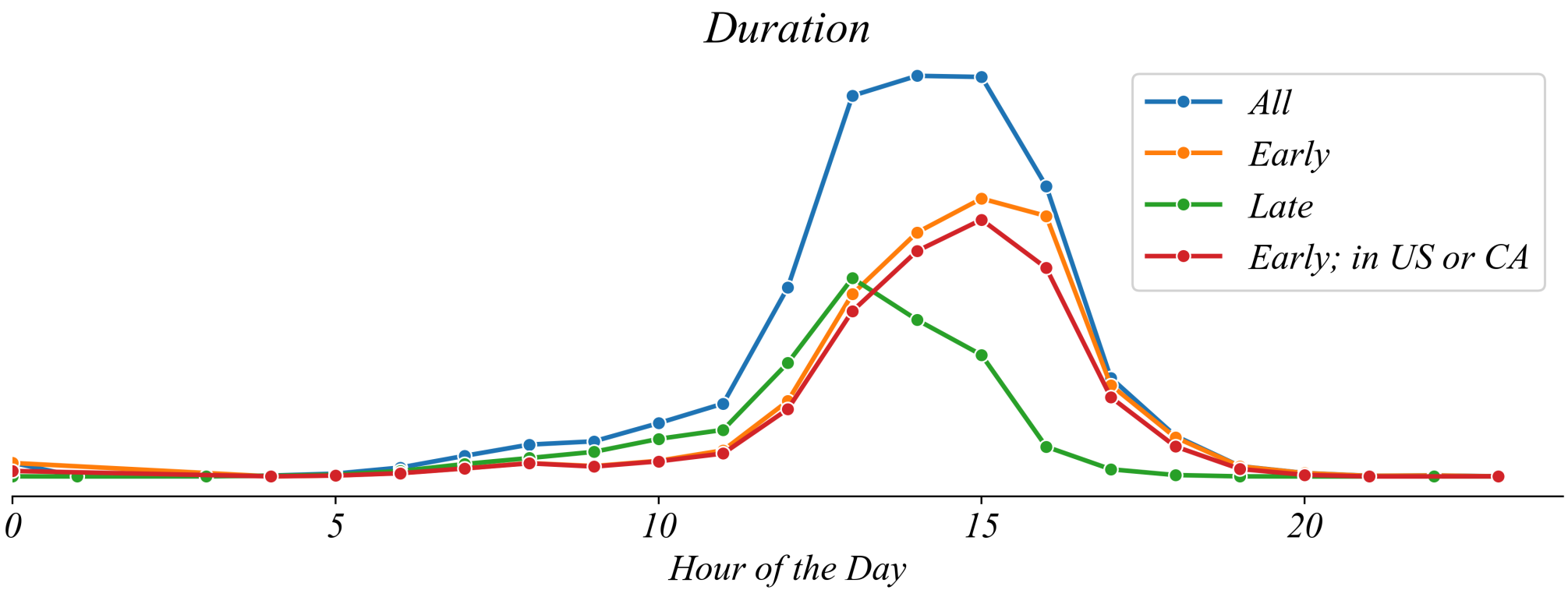


> *shops that open early will stay open longer in the US or Canada*

> *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?



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.

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

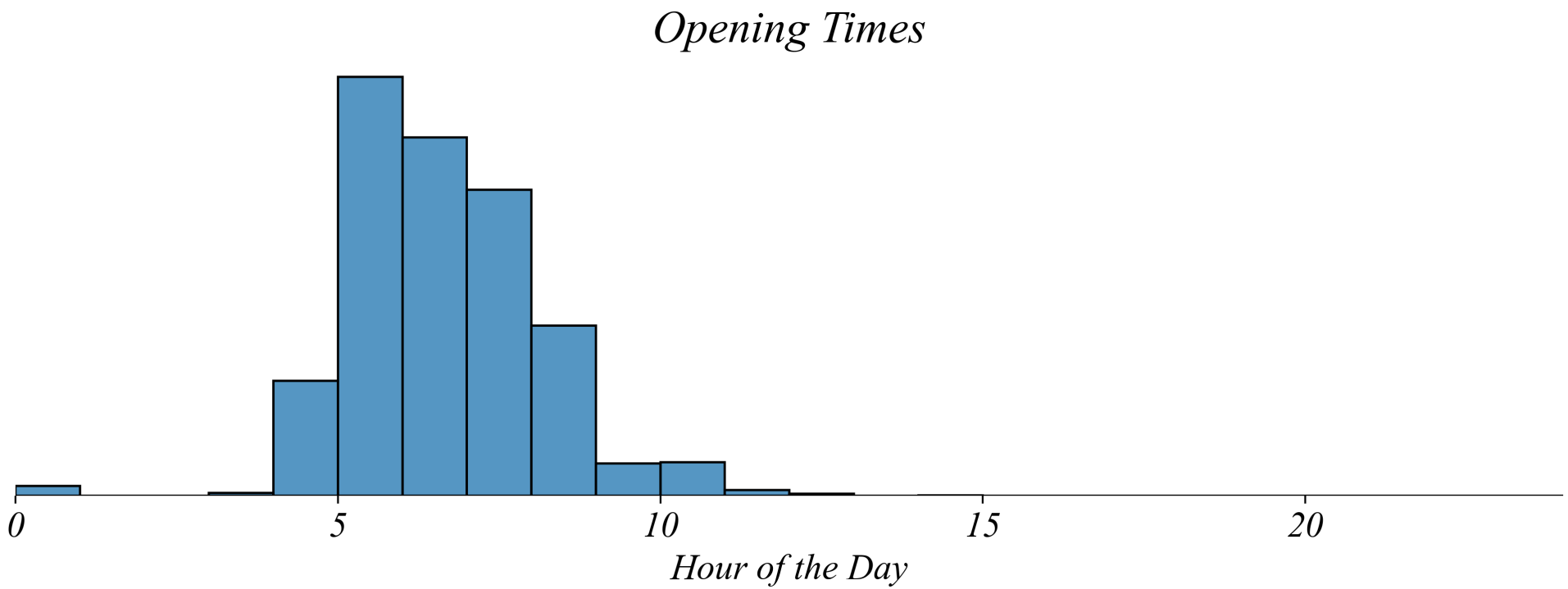
Coffee Shop Hours

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



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?

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

Coffee Shop Hours: filter for US locations

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

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

Coffee Shop Hours: filter for US locations

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

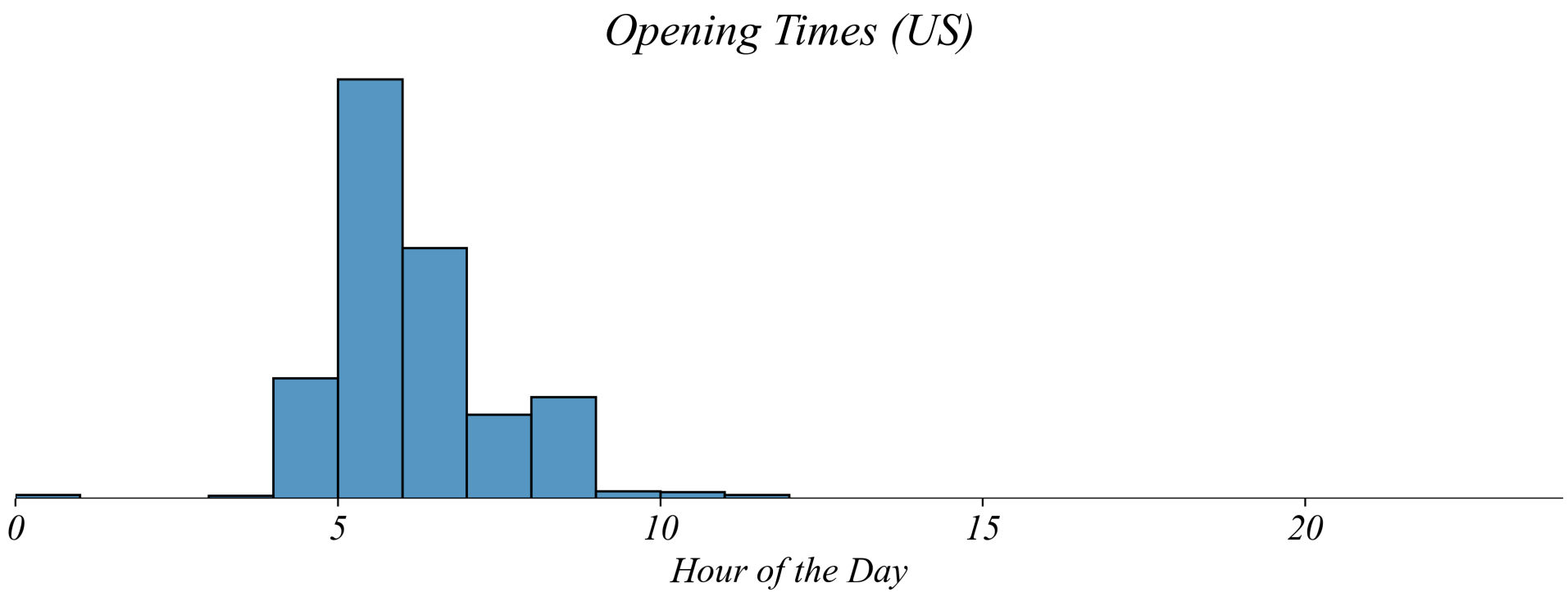
```
# Decide whether each row's country code is 'US'
# data['COUNTRY_CODE'] == 'US'

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

Coffee Shop Hours: filter for US locations

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

```
1 # Histogram of US locations
2 plt.hist(us_data['open'], bins=20)
```



Coffee Shop Hours: shops in either US or CA

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

```
# 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')]
```

Coffee Shop Hours: shops in either US or CA

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

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

Coffee Shop Hours: shops in either US or CA

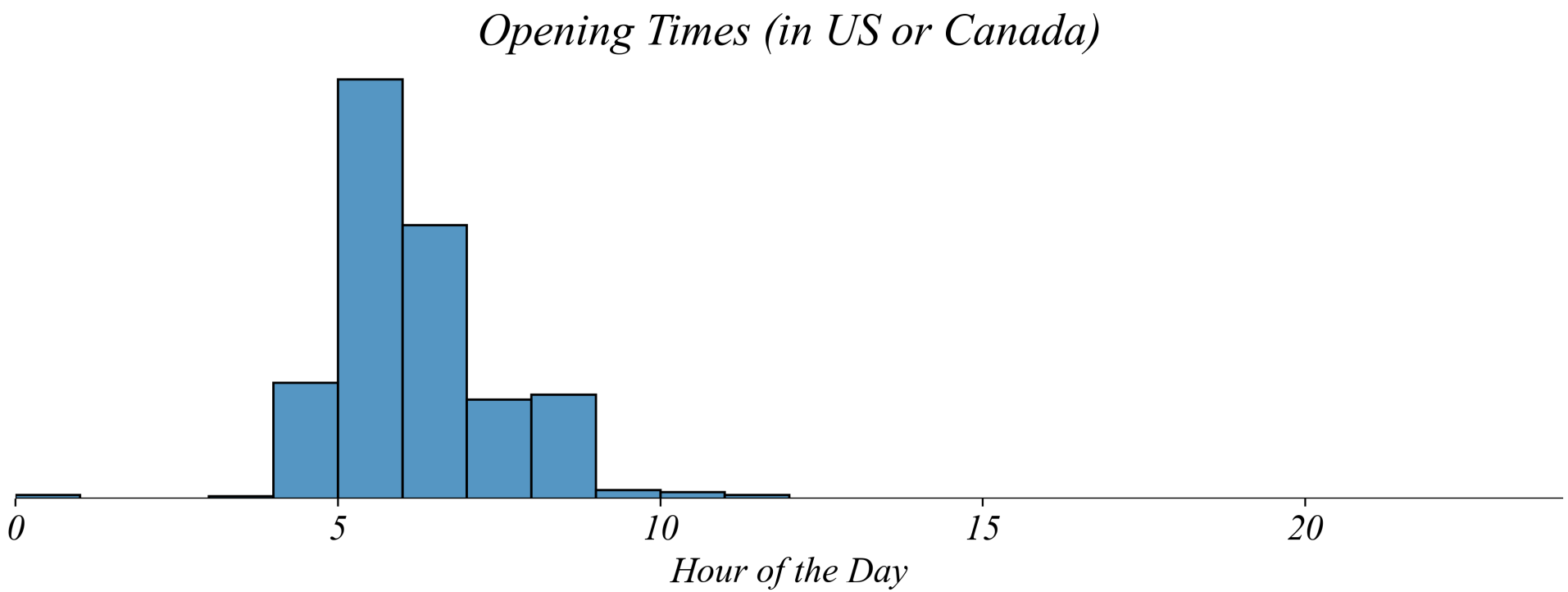
Q. When might be a good time for a US coffee shop to open?

```
# Find the data in either the US or in Canada (CA)
# Method 2: Using isin() and define a new dataset
us_ca_data = data[data['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 plt.hist(us_ca_data['open'], bins=20)
```



Coffee Shop Hours: shops in either US or CA

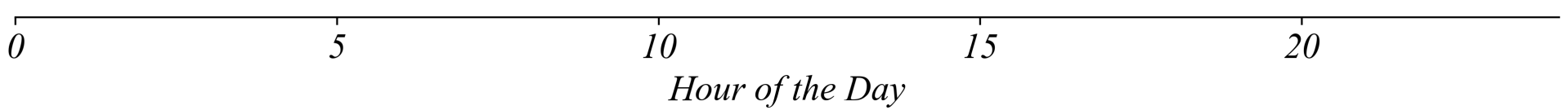
Q. When might be a good time for a US coffee shop to open?

What would this dataset look like?

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

> *it would contain no data!*

Opening Times (shops in US AND Canada)



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

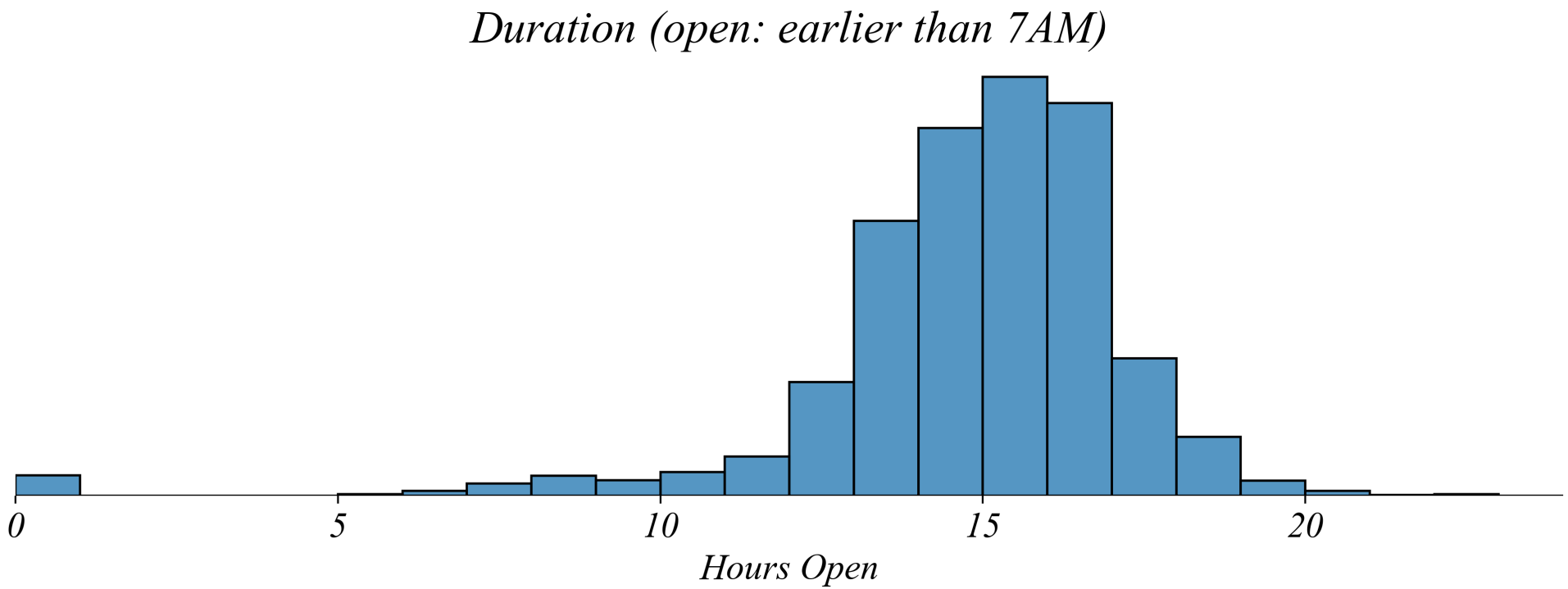
Q. How long might be good for the coffee shop to stay open?

```
# Filter for shops that open before 7 AM
early_data = 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?

```
# Create histogram of duration for early-opening shops
plt.hist(early_data['duration_hr'], bins=20)
```



Coffee Shop Hours: combine filters

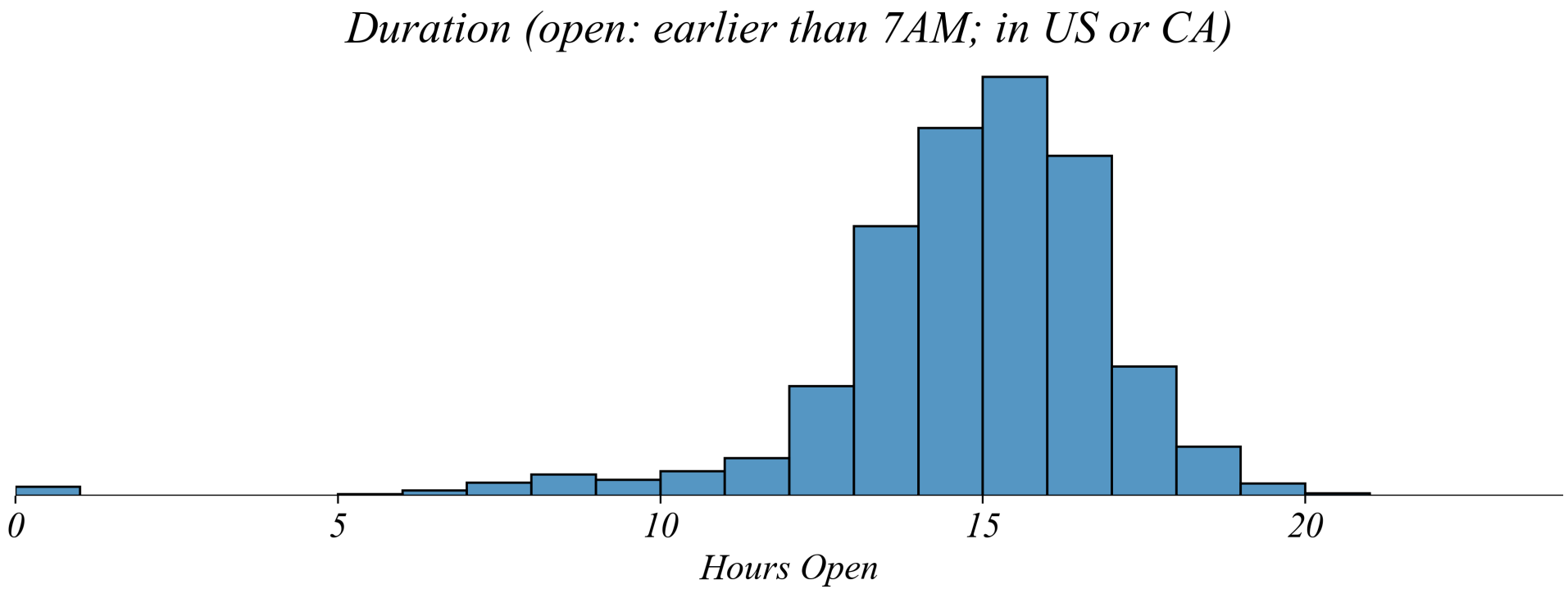
Q. How long might be good for the coffee shop to stay open?

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

Coffee Shop Hours: combine filters

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

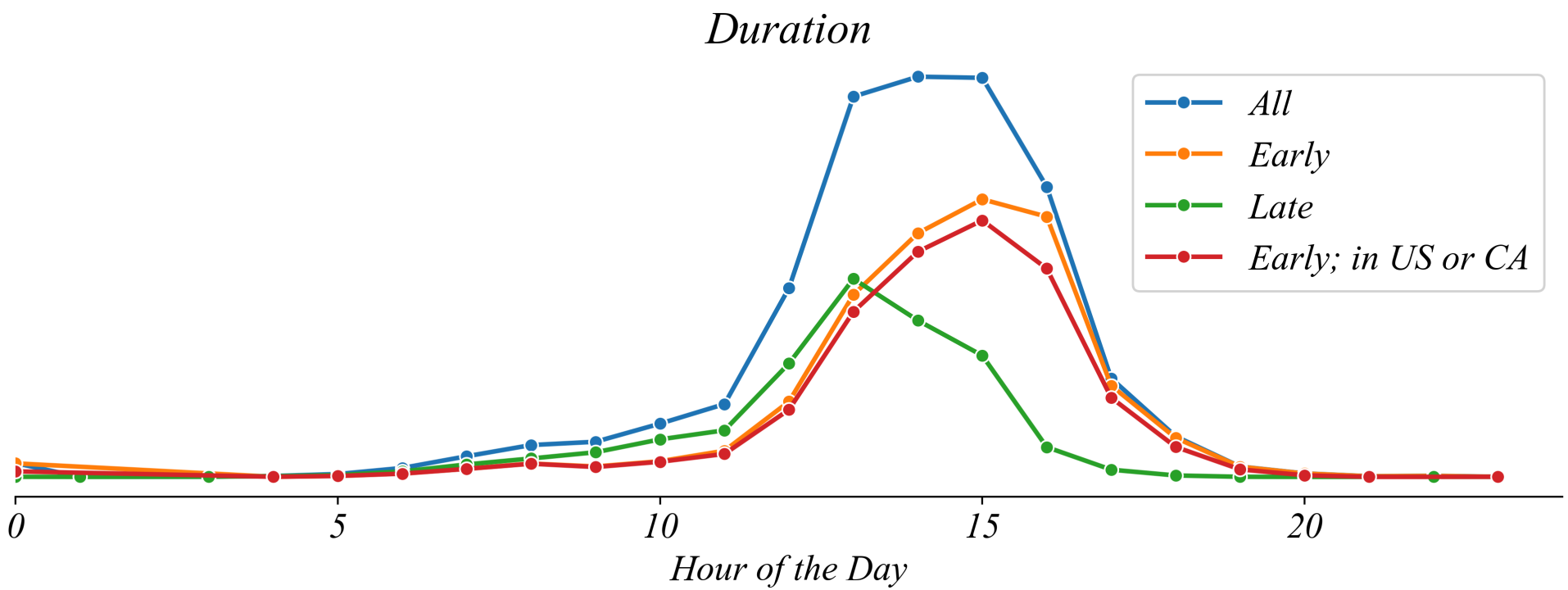
```
# Create histogram of duration for early-opening US/CA shops
plt.hist(early_us_ca_data['duration_hr'], bins=20)
```



Coffee Shop Hours: compare opening times

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

```
# Compare early vs all shops in US/CA
plt.hist(us_ca_data['duration_hr'], bins=20, alpha=0.5, label='All US/CA')
plt.hist(early_us_ca_data['duration_hr'], bins=20, alpha=0.7, label='Early US/CA')
plt.legend()
```



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