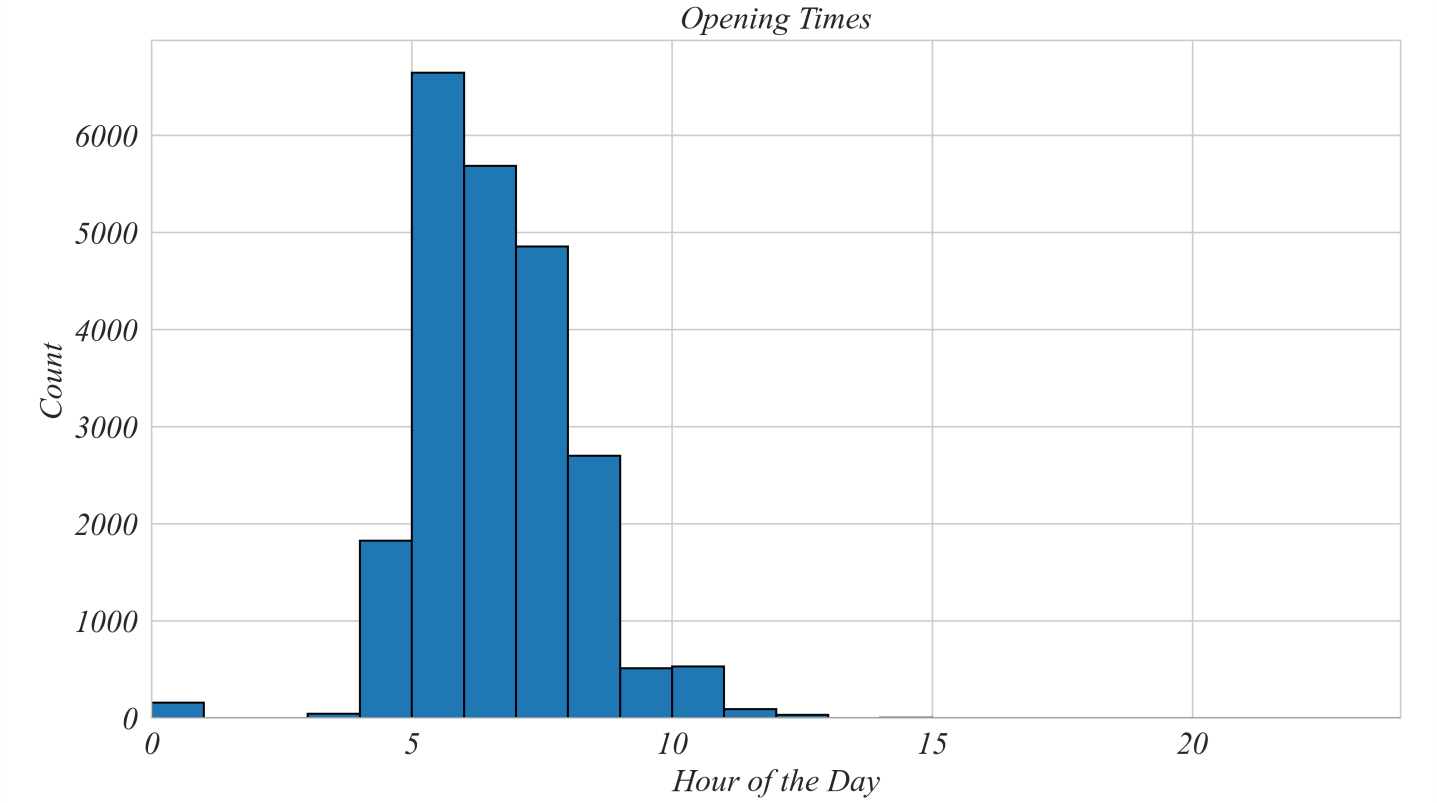
Part 1.4 | Filtering Data

Logical Filters

When opening a new cafe, one of the many decisions to make is which hours it should operate.

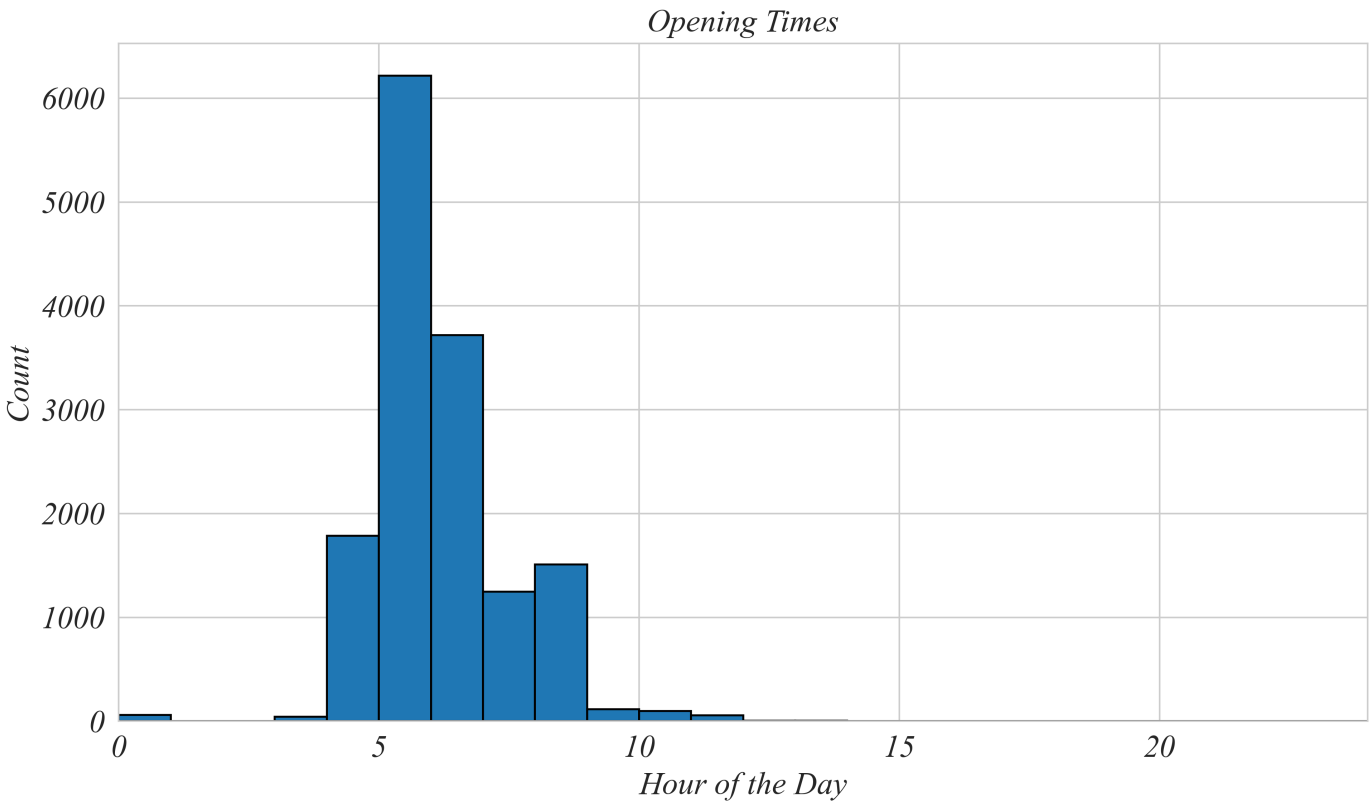
The operating hours of existing cafes can help inform this decision. This table shows the operating hours of Starbucks in the US as of January 2020.

Just looking at the table, what opperating hours are the most common? Well it's not easy to see. Instead, lets set up a histogram. Plotting opening time, it's easier to see that the most common opening time is 5AM.



This doesn't mean that opening at 5 AM is best. Some locations may operate in areas that might not need the same hours of opperation. We can select subsets of the locations by **filtering** our dataset, using **logical expressions** — statements that are either true or false. Only rows which satisfy the condition are kept in the dataset.

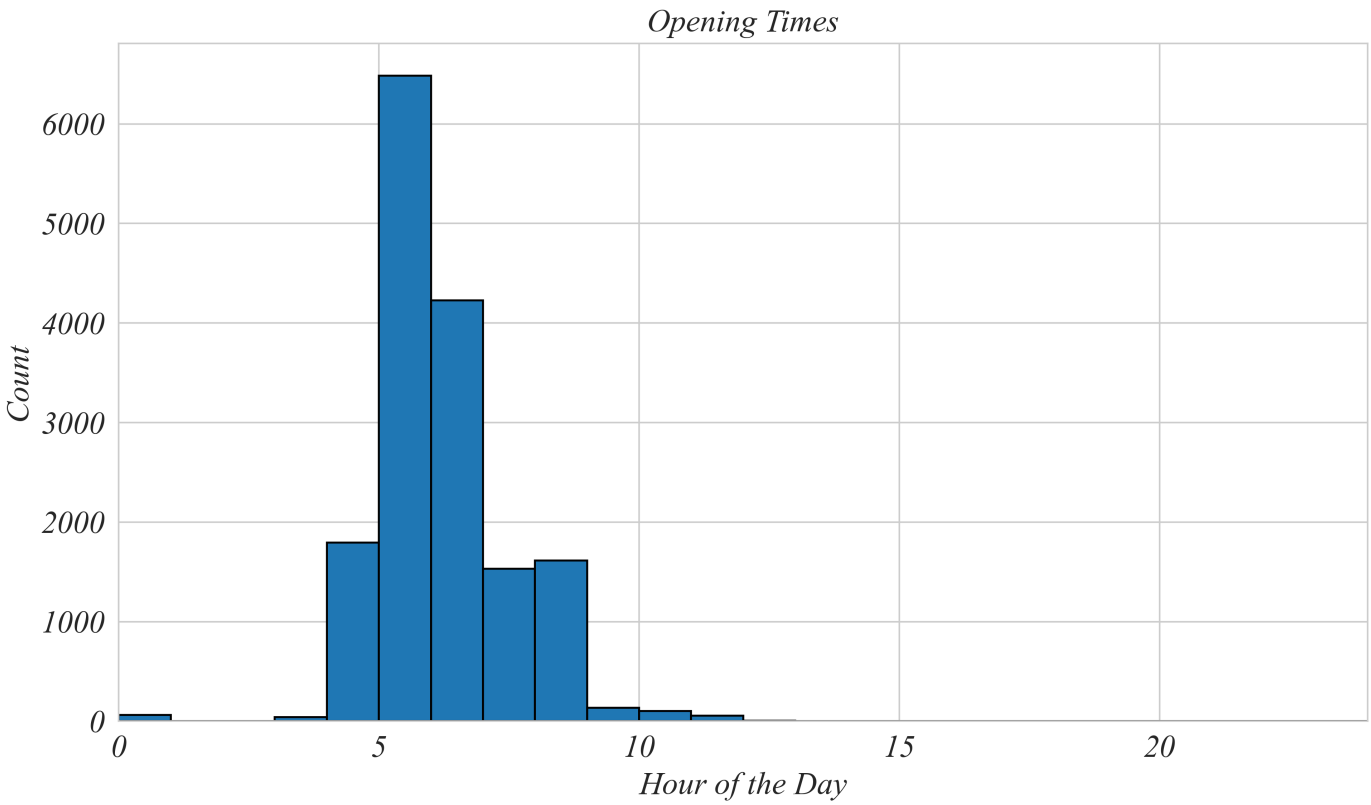
How would you write a logical expression that keeps only the locations that are inside the US? We would use COUNTRY_CODE = US.



The US seems to have later opening times than the global average. This could be for a number of reasons, one of which is some error with the way times are written in the data. But it could be that other countries have different waking patterns that the US, maybe due to the way time zones are layed out.

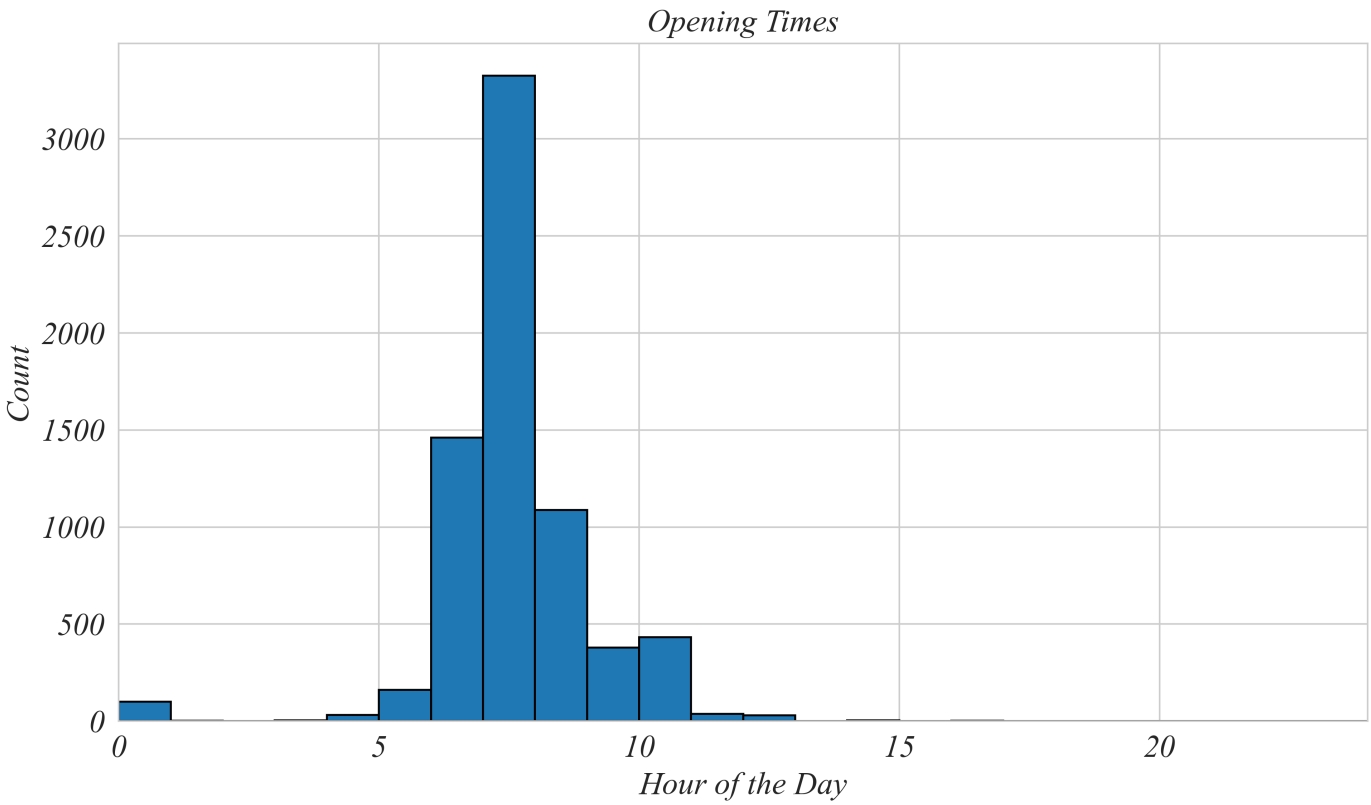
Combining Filters

Filters can be used on their own or can be combined to look at subsets of subsets of the data. Logical expressions are combined using logical operators like AND and OR. Let's use a logical operator to see cafes from both the US and Canada in one histogram.



COUNTRY_CODE==US OR COUNTRY_CODE==CA returns records that meet either one of those conditions, which is what we want. If a cafe is in Canada but not the US, we would say that it's in the US or Canada. If we wanted to return records that meet *both* conditions at the same time, we would use COUNTRY_CODE==US AND COUNTRY_-CODE==CA. AND tells us all records that meet both conditions. Since cafes can't be in both countries at the same time, no rows are returned.

We can also see the opening times in countries not in CA or the US.



Excel Exercise

The first thing to do is to turn this data into a table. Select the data by hitting Command-D on Mac or maybe Control-D on Windows. This should select all the data. Then hit Control-T to create a table. Since you've already selected the data, it's already selected in the table window, so just hit create.

The first thing to do is translate the time into an hour. Tables are super nice. We just need to right click to create a new column. Excel sometimes has control issues, so you'll need to create the column to the left of the time columns for some reason. When working with Excel tables, it automatically applies your work to all cells in the column. Apply to one cell using the following:

```
=HOUR(cell)
```

Then rename the column "OPENING HOUR". Then do the same thing for Closing Hour. We can create duration by taking the difference between the columns.

To create a histogram of opening times, all we need to do is select the column, then insert our histogram. We'll need to change the bins to make it look nice. Give it a nice title. Then double click on the table and go to "Add Chart Element", add a Primary Horizontal Axis Title, maybe called Time of Day.

We can save this figure by right clicking on the upper corners of the figure and selecting "Save as Picture". I would urge you to create a folder titled "In Class Figures" in your Data folder. Then maybe call this figure "Starbucks_Opening_Times_All.png".

Then we want to filter data. Excel tables make this nice. All we have to do is click on the column we wish to filter and choose filter options.

To select only the locations in the US, choose "countryCode", filer as "Equals", then enter US. We automatically see the table update. This takes away the rows for locations that aren't in the US, which also hides our table. To fix this we'll create a new Sheet, call it Figures, and copy and paste our figure there. We have to unfilter our data to do this. But once we're done we can move it over, then refilter our data. Save this figure as maybe "Starbucks_Opening_-Times_US.png"

We can also use multiple filters. Lets filter on the US and Canada. We go in, filter on the US, and hit enter. This filters the data and brings up another option for filtering. If we choose equals and CA and hit enter, we get no entries.

This is because the automatic filtering option is AND, which means we're selecting all the locations that are in both the US and Canada, which is zero. So we have to switch from AND to OR.

Going back to the figure, we can retitle it and save it as "Starbucks_Opening_Times_US_CA.png".

Now we want to create a figure showing opening times for all countries not in the US or Canada.

Lets go into the countryCode column and clear filters to start over. We can choose to filter on "Does not equal" and enter US. This is all the locations not in the US. What might we do next?

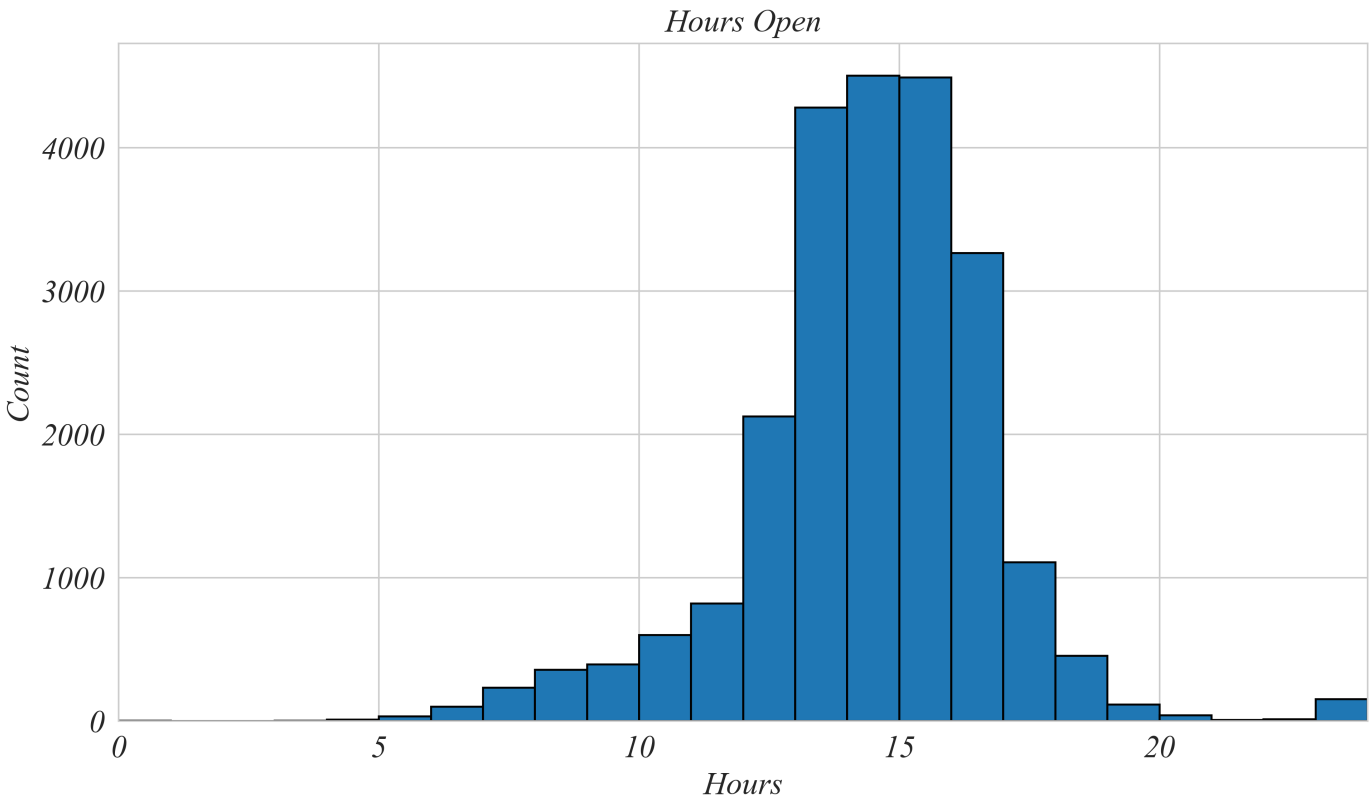
We should also choose "Does not equal" and enter CA. But do we use AND or OR?

If we choose OR like last time, it will give us all the locations that are not in the US OR not in Canada. So if a location is in Canada it will definitely not be in the US, so will be included.

But if we use AND, then it's saying that it must not be in the US and also not in Canada. This is all the locations outside both the US and Canada, which is what we're looking for.

Inequalities

In addition to opening time, cafes also need to decide how long to stay open every day. The dataset includes how long each store is open in addition to when it opens. What do you think the median operating duration is?



Although it's a little difficult to see it the median in this type of figure, the median duration — the middle value — is 14 hours a day.

Another way we can filter the data is using inequalities, another type of data filter, allowing us to see if such long operating hours are typical of all cafes, regardless of when they open.

Filters can use logical expressions based on inequalities like < (less than) or ≥ (greater than or equal to). Each of these relationships functions similar to the English language definition.

Symbol	Meaning
=	Equal to
≠	Not equal to
>	Greater than
≥	Greater than or equal to
<	Less than
≤	Less than or equal to

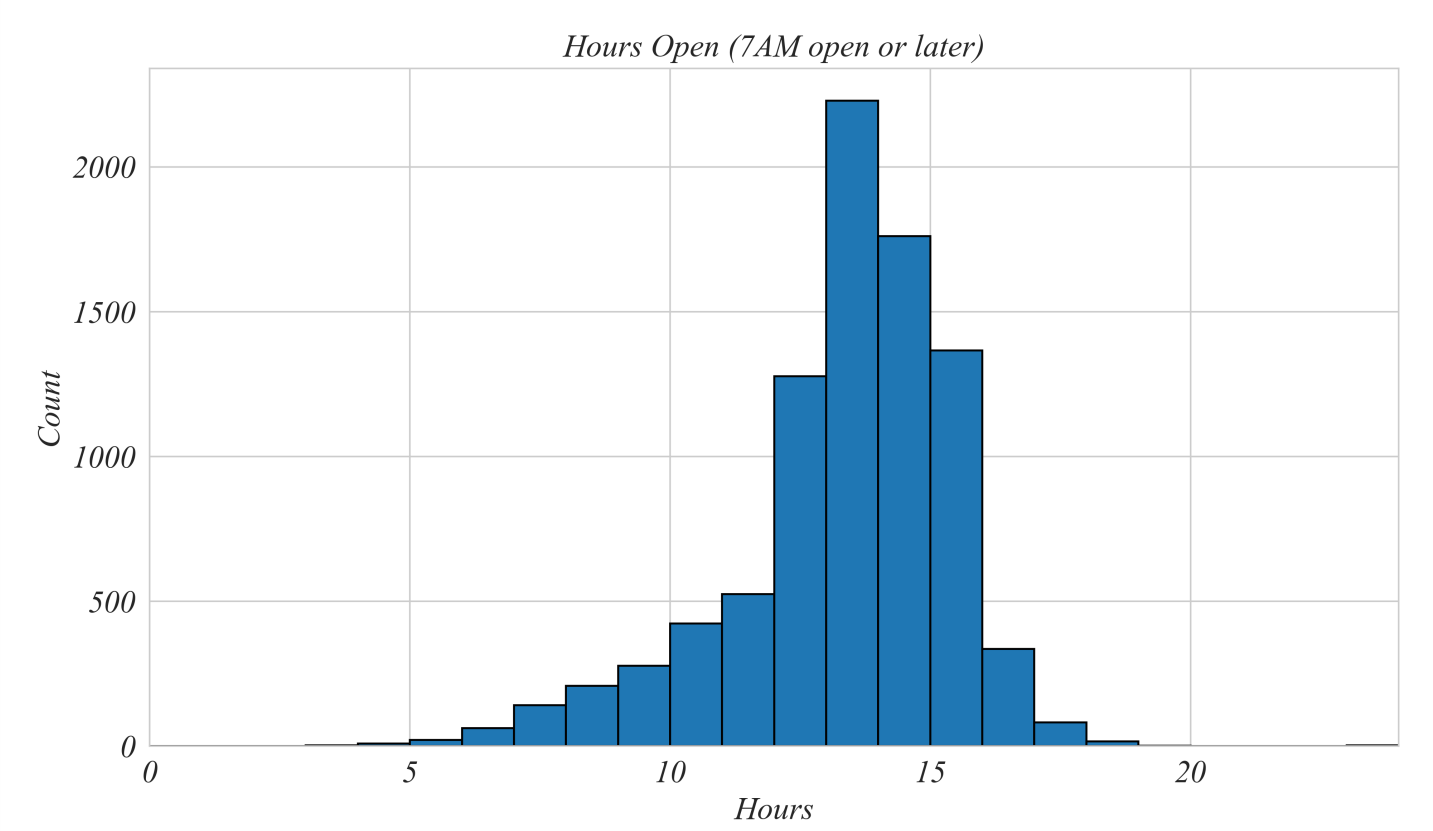
Suppose we want all of the cafes that open at 7 AM or later. How might we construct a logical fiter to select these cafes?

SHOW DATASET

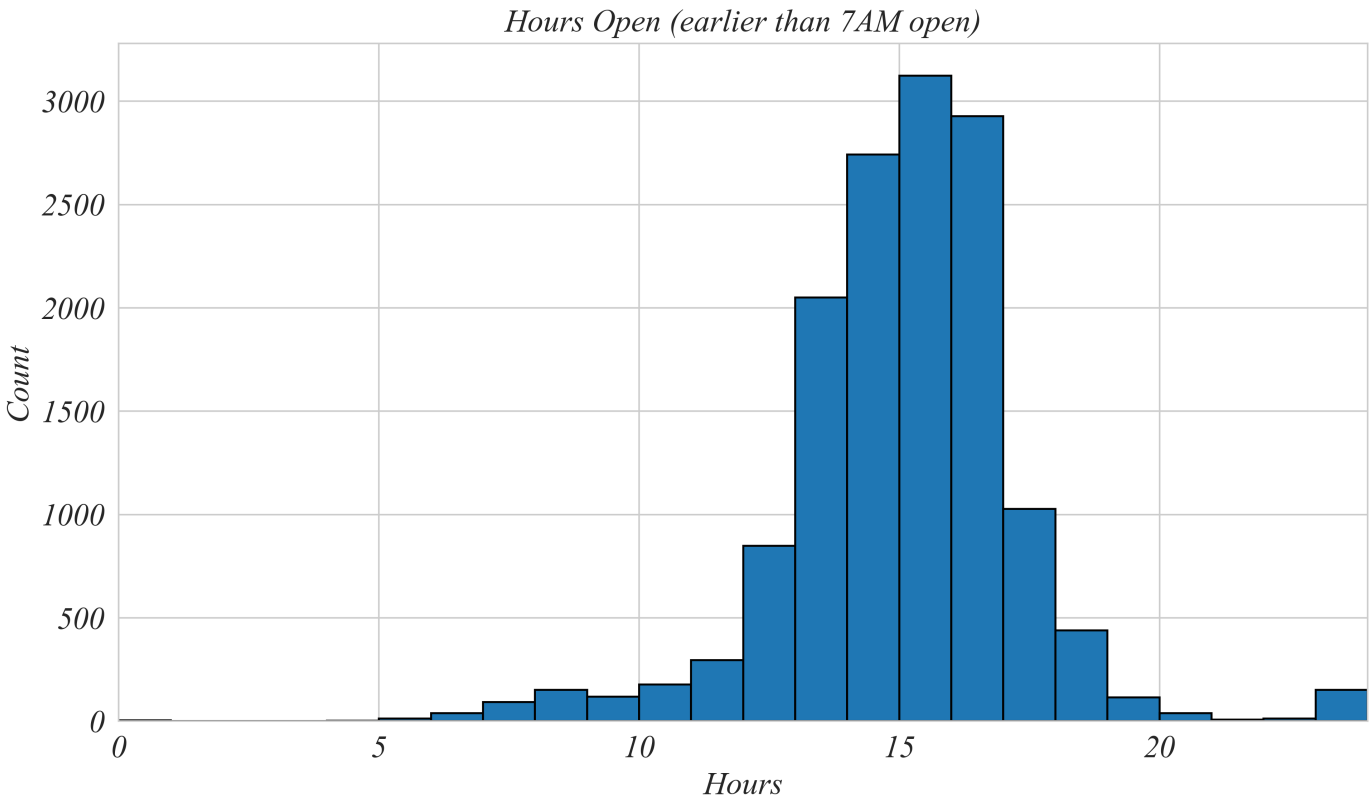
This would keep cafes that open later than 7 AM but would filter out the cafes opening right at 7 AM.

In this dataset, times are stored as the hour in the day — so 6 AM is stored as 6 and 3:30 PM is stored as 15.5. Stores that open at 7 AM or later correspond to the logical expression OPEN ≥ 7.

Is there a significant difference in the operating hours of stores that open at 7 AM or later vs. stores that open earlier?

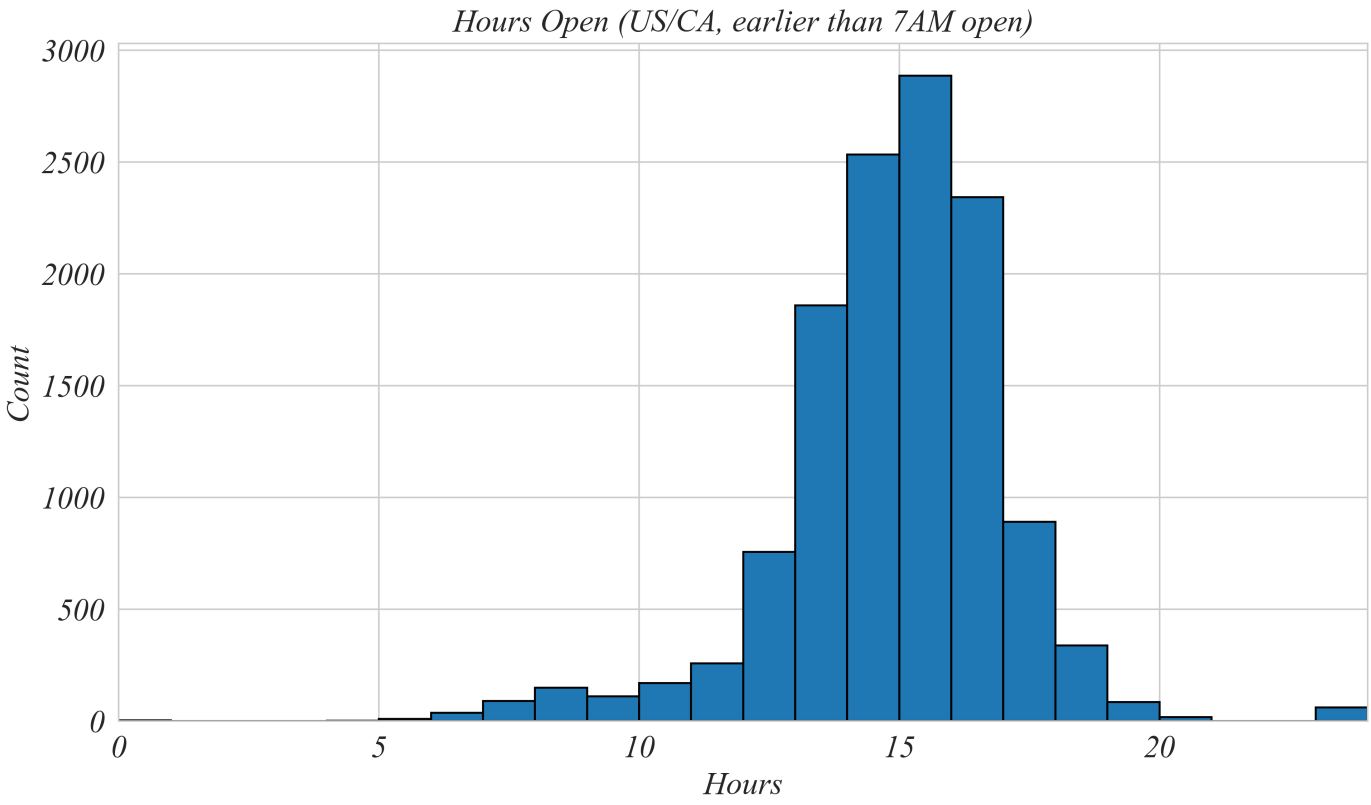


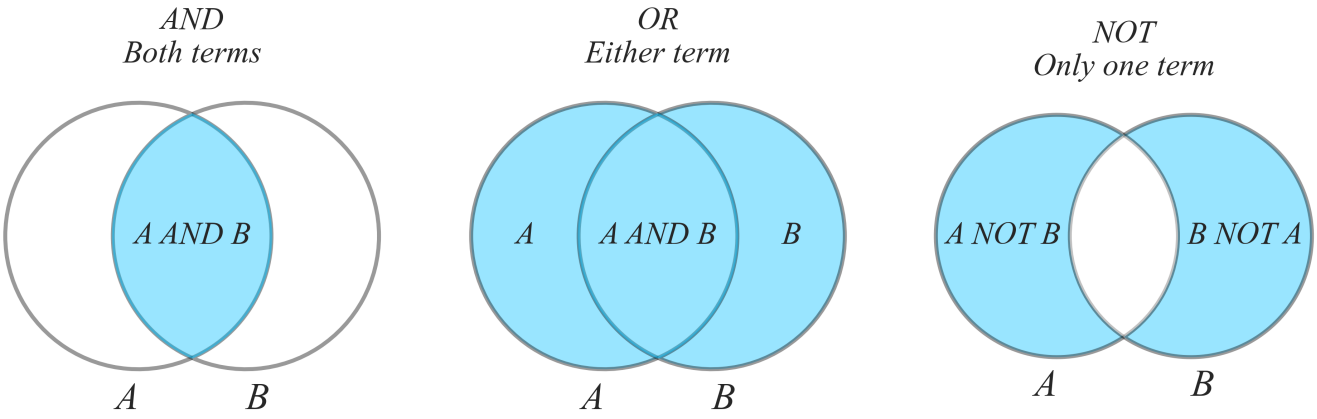
This filter shows us stores that open at 7 AM or later, which tend to stay open 13 hours a day — fewer hours than the early opening shops.



This filter shows us stores that open before 7 AM. They tend to stay open for around 15 hours a day. Stores that open later tend to have shorter hours than a typical Starbucks cafe.

We can also combine a filter on opening time with other filters to explore the operating hours of other particular segments. This figure shows the duration of shops opening before 7 AM in the US and Canada.


Hours Open (US/CA, earlier than 7AM open)
Count
3000
2500
2000
1500
1000
500
0
0
5
10
15
20
Hours


AND
Both terms
A AND B
A
B
OR
Either term
A
A AND B
B
A
B
NOT
Only one term
A NOT B
B NOT A
A
B

Excel Exercise

Create a histogram of Duration. You can see some negative values. What's going on here? If we scroll down to a negative value you'll see that if a location closes at midnight or later, their closing time gets recorded the next day. In these cases, we've actually counted backwards.

Lets fix this. We can filter duration on whether it's positive or not. We can see that if a location opens at 8 and closes at 0, it's been open for 16 hours. We can create a new column that adds 24 hours to the column if it's negative.

```
=IF(Duration<0, 24+Duration, Duration)
```

Then we can call this new column Duration Clean. We'll create a second histogram for this.

Then we can filter on whether the store opened before 7 or not.