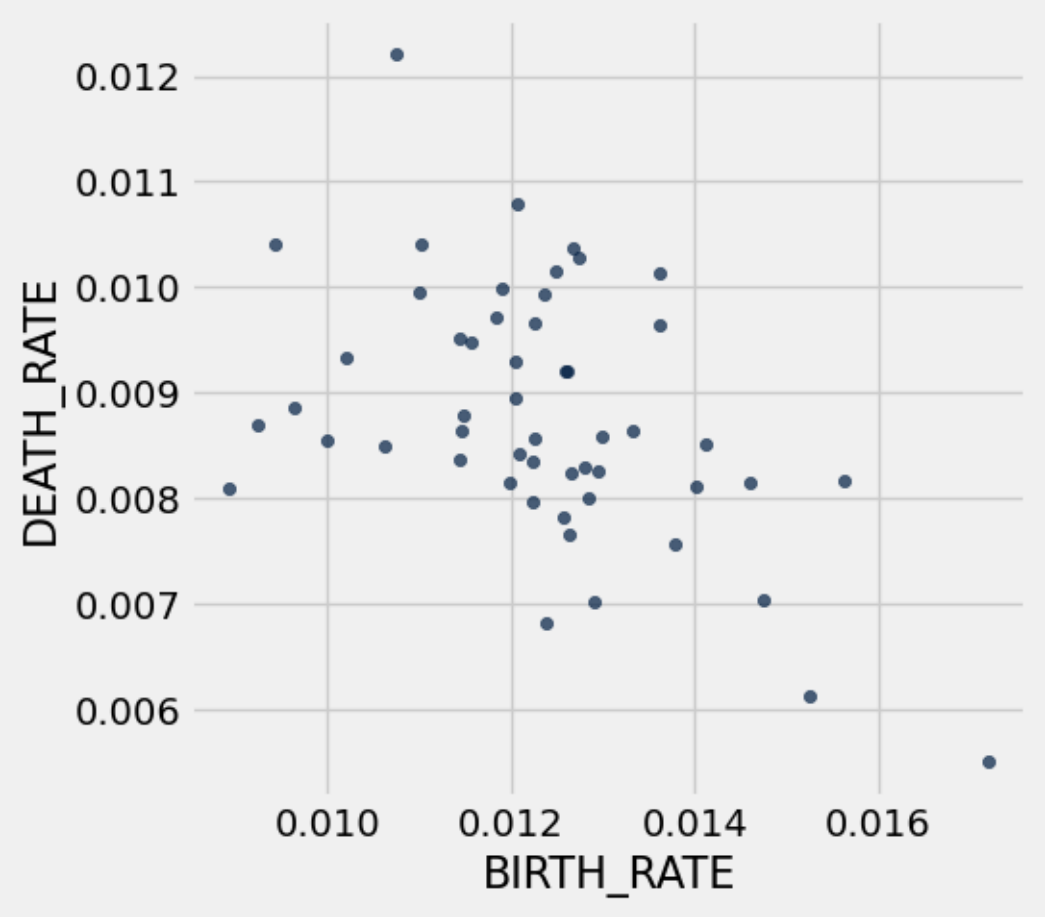

Question 5. In the code cell below, create a visualization that will help us determine if there is an association between birth rate and death rate during this time interval. It may be helpful to create an intermediate table containing the birth and death rates for each state. **(4 Points)**

Things to consider:

- What type of chart will help us illustrate an association between 2 variables?
- How can you manipulate a certain table to help generate your chart?
- Check out the Recommended Reading for this homework!

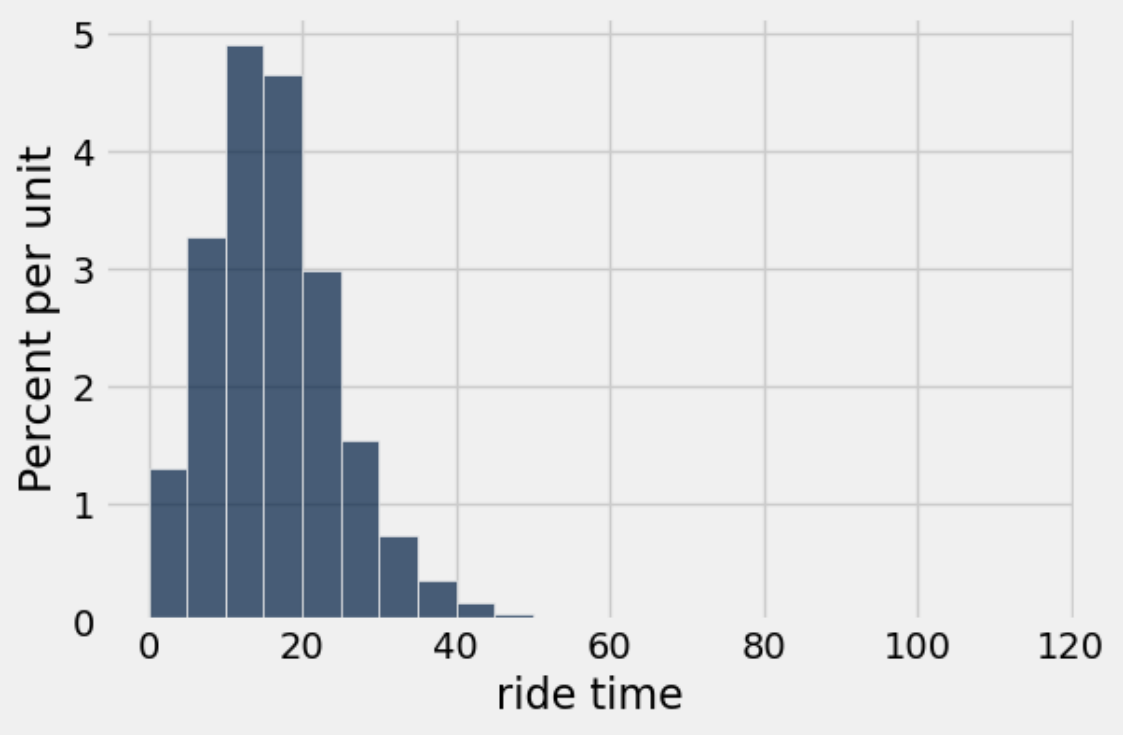
In [49]:
```
# In this cell, use birth_rates and death_rates to generate your visualization
birth_rates_2015 = pop.column('BIRTHS') / pop.column('2015')
death_rates_2015 = pop.column('DEATHS') / pop.column('2015')
Table().with_columns("BIRTH_RATE", birth_rates_2015, "DEATH_RATE", death_rates_2015).scatter("
```



Question 1. Produce a histogram that visualizes the distributions of all ride times in Boston using the given bins in `equal_bins`. **(4 Points)**

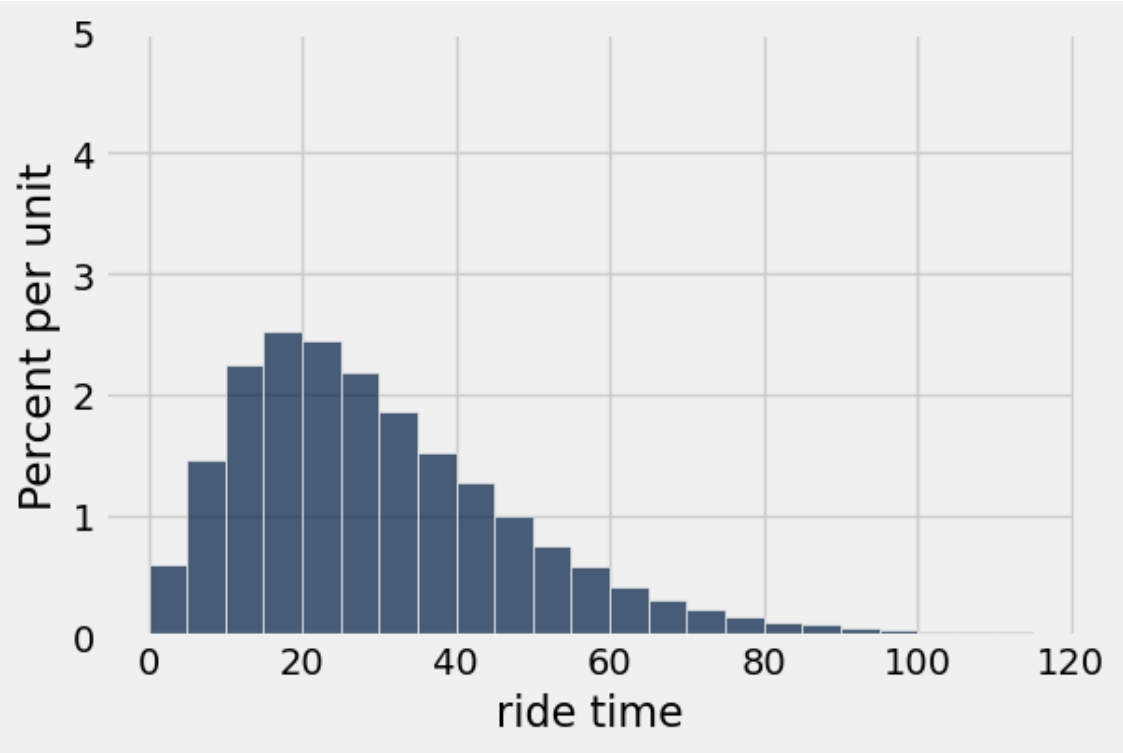
Hint: See Chapter 7.2 if you're stuck on how to specify bins.

```
In [79]: equal_bins = np.arange(0, 120, 5)
         boston.hist("ride time", bins=equal_bins)
```



Question 2. Now, produce a histogram that visualizes the distribution of all ride times in Manila using the given bins. **(4 Points)**

```
In [81]: equal_bins = np.arange(0, 120, 5)
         manila.hist("ride time", bins=equal_bins)
         # Don't delete the following line!
         plt.ylim(0, 0.05);
```



Question 6. Identify one difference between the histograms, in terms of the statistical properties. > *Hint*: Without performing any calculations, can you comment on the average or skew of each histogram? **(4 Points)**

For the boston histogram, the distribution of data indicates a fairly symmetric pattern, with a marginal right-ended tail and a peak in bin 10-15. Thus, the average of the boston dataset ride time would be around the 15-20 bin. Based on the manila histogram, the percent per unit (ride time in this case) sharply increases in the 0 to 20 bins, and peaks in between the 15-20 min bin. From that point, in bins 20-100, the rate of change slowly decreases, which skews the data more right, as more data points are spread out on the right of the supposed median, forming a right-ended tail and therefore increasing the average. Thus, I would predict the average to be in between bins 30-40.

Question 7. Why is your solution in Question 6 the case? Based on one of the following two readings, why are the distributions for Boston and Manila different? **(4 Points)**

- Boston reading
- Manila reading

Hint: Try thinking about external factors of the two cities that may be causing the difference! The readings provide some potential factors – try to connect them to the ride time data.

Based on the Boston article/guide, the climate data highlighted the unstable and fairly precipitous nature of Boston's weather, with some months averaging over 3 inches of rain throughout the year. The guide also indicated the below-zero temperatures which were typical of winter in Boston. Based on these observations, some potential factors that might influence the data from the histogram are the fact that more people are probably using cars to get to places within the city, since more road-friendly options such as walking, biking, and public transit may be either cumbersome or delayed by bad weather. The influx of cars on the road may also increase ride times due to traffic or simply bad weather. Manila, on the other hand, has a rooted issue in poor infrastructure and frequent traffic. The article above highlights 3 main reasons for this phenomena, which are poorly enforced traffic lanes, lights, and laws, constant delays in public transit, and poor U-turn choke points. The combination of these factors may explain the outlier data points in ride times, which are unpredictable due to poor traffic at popular hours of the day. For example, during rush hour, which occurs in the morning and evening, more people need to commute to and from work or school, thus causing a high spike in ride times, which are amplified with delayed public transit and poor road infrastructure.

Question 2. State at least one reason why you chose the histogram from Question 1. **Make sure to clearly indicate which histogram you selected** (ex: "I chose histogram A because …"). **(5 Points)**

I chose histogram C because the original scatter plot indicated a large concentration of data points between -1 and 0, with a gradually-decreasing tail on the right. Thus, histogram C makes sense since it indicates a peak in between -1 and 0, and a slow decline/skew right, which corresponds to the tail. The peak on histogram C also explains the more broad distribution of each corresponding y values for each x, indicating a taller range of data points for each x below 0, thus causing the peak in between -1 and 0.

Question 4. State at least one reason why you chose the histogram from Question 3. **Make sure to clearly indicate which histogram you selected** (ex: "I chose histogram A because …"). **(5 Points)**

I chose histogram B because if we look at the scatter plot and analyze the y-axes (as the independent variable), then there is a gap in between the data points, which is reflected in histogram B, which has a gap at 0. Also, since the data is condensed between -1.5 to -0.5 and 0.5 to 1.5 respectively, the correct histogram should reflect the concentration in between those values, which histogram B does. The other histograms indicate values outside that range, which is incorrect.