Using a Linear Model with Non-Linear Data

Oftentimes, real world data is not linear. However, in certain situations, we can still use linear models on this non-linear data. We have two primary ways we can do this. First, we can restrict the domain we are looking at. Second, we can remove outliers that prevent our data from being linear. Let's start by looking at how we restrict the domain. To do this, first, look for a section of data on the scatterplot with a linear association. Next, check that less than half of the data is being excluded. Lastly, identify and state the domain that has a linear association using the largest and smallest x-values of the linear data. If we want to apply a linear model by removing outliers, first, look for part of the association that could be linear. Then make sure only one or two points are ignored as possible outliers. Finally, identify the outlier points that will be ignored by their coordinate pairs.

After you have done one of these two processes, you will want to create a linear equation to represent that dataset. To do this, first choose two points on the line and calculate the slope using the slope formula. The guide points should not be data points. For this step, it's helpful to choose round numbers such as integers or easy decimals. Next, find the y-intercept by either extending the line of best fit to the y-axis and observing the intercept or solving for the y-intercept by substituting the slope and a guide point into slope-intercept form. Finally, write the final equation in slope-intercept form.

So when do we restrict the domain versus ignoring outliers? Let's first look at restricting the domain. Say we have this dataset here. It's mostly linear, but it starts to trend down at the end. If we restrict our domain to negative 4 to positive 3, that allows us to ignore these points here. Now we have a linear dataset to work with, and we can fit a linear equation to it, like this.

What about removing outliers. Well this data set is mostly linear, but a couple datapoints actually make it more U-shaped. If we remove just those 2 points, we end up with a dataset that looks like this, which again, we can fit to a linear model.

Alright, now that we've seen each of those situations, let's head over to the whiteboard to look at a few more examples.

This first question reads, "Identify the domain on the following scatterplot that can be analyzed using a linear model." Well, looking at the data as it exists right now, it's kind of U-shaped, but if we only look at the data between about x equals 3 and x equals 8, these data are linear. So if we analyze on the domain 3, 8, we have a linear data set we can analyze. Let's look at another one.

This question reads, "Identify the outliers that could be ignored on the following scatterplot to allow analysis with a linear model." Again, right now this data set is not really linear, but if we ignore this data point and this data point, the remaining data actually are linear. Alright, let's look at one more example.

This one reads, "Cara is going on a road trip with her friends. She wants to make sure she saves as much money on gas as she can, so she wants to know how many miles per gallon she is getting when she drives her car at 55 mph. Her data is shown in the table. Create a scatter plot to visualize the data and determine if the association is linear or nonlinear. Determine how to create a linear model for the data by stating a domain or identifying outliers to ignore." So let's begin by plotting this data set on this scatterplot, starting with our first point at a speed of 20 miles per hour and a gas usage of 10 miles per gallon, so that's going to be at 20, 10, right here. The next data point is at 30, 19, which is here. The next

point is at 40, 20, which is here. The next one is at 40, 23, here. Next is at 50, 22, which is right here. The next point is at 50, 25, which is right here. The next point is at 60, 26, right here. And the last point is at 75, 15, which is right here. So right now we see that the data has kind of an arc shape, but if we drop this outlier and this outlier, the data that remain are pretty much linear, like that. So even though the whole data set isn't completely linear, since Cara is trying to analyze the point at 55 miles per hour, that's inside the linear portion of our data set.