Simulation Transcript
Visualize the data

1. It's time to see what story the data is telling. Select the **Visualizations** tab.
2. The tool offers many types of charts. Select the **double arrow** icon.
3. Look at all the options for chart types! Keep in mind that a visualization will not start off as perfect. You must experiment to try different charts. For your project, a good, initial visualization to explore data is a scatter plot chart. Select **Scatter plot**.
4. Select the columns of data you want to visualize. Consider your x-axis. It's your independent variable. You want this to be the new column you added because you want to see if claims over $10K are classified as 0 for no fraud or 1 for fraud. Under **X-axis**, select **EXCESSIVE_CLAIM_AMOUNT**.
5. Consider your y-axis. It's your dependent variable. For your project, it needs to be a dollar amount so you can see how many claims over $10K fall under 0 for no fraud or 1 for fraud. Under **Y-axis**, select **CLAIM_AMOUNT**.
6. Notice the initial visualization! You have a distribution of dots in your scatter plot chart. The dots above **0** represent all claims under $10K. The dots above **1** represent all claims over $10K. This is interesting so far. Select the **Next arrow** to continue.
7. But, you must dive deeper. Remember fraud investigators concluded which claims were actual fraud and added a column to the data set. These are the "answers" you need. Let's add this data in color-coding. Under **Color map**, select **FLAG_FOR_FRAUD_INV**.
8. This reveals something interesting in the scatter plot! The blue dots are individual claims that investigators concluded to *not be fraud*. The pink dots were concluded by investigators to *be fraud*. Select the **Next arrow** to continue.
9. This is an interactive visualization in IBM Watson Studio. Let's explore! Hover over different areas of the chart to view some data points. Select the **Next arrow** to continue.
10. Select the **blue area** on the y-axis just above 0.
11. First, this means there are some claims *less than* $10K that *are not* fraud. This claim for $1,720.00 is blue. Blue means it was deemed to not be fraud by the investigators. Now select the **pink area** on the y-axis just above there.
12. Second, this means there are some claims *less than* $10K that *are* fraud. This claim for $2,730.00 is pink. Pink means it was deemed to be fraud by the investigators. Interesting! Now select the **pink area** above the 1 just above 40,000.000.
13. Third, you can see there are a lot of claims *greater than* $10K that *are fraud* because pink means it was deemed to be fraud by the investigators. This means the business sponsor's hypothesis that claims over $10K could indicate fraud is correct. Select the **Next arrow** to continue.
14. But, look! There are also claims *greater than* $10K that are *not fraud* because blue means it was deemed to not be fraud by the investigators. So, not all claims over $10K are fraudulent. Interesting! Select the **Next arrow** to continue.
15. This is an informative visualization. You can use it to conclude these insights and share it with the business sponsor, so select **Save Visualization to project** in the upper right to save it for future viewing.

You successfully created a scatter plot visualization in the data refinery tool to display data so that you can use it to conclude insights.