Simulation Transcript
Refine the data

1. Here is your "_shaped" CSV file that the data refinery tool created. Notice you're working with 19 columns of data now after your cleaning efforts. It's time to refine it based on your business sponsor's hypothesis that claims over $10K may be fraudulent. Select the **Next arrow** to continue.
2. You are creating a new column that depicts which values in the column **CLAIM_AMOUNT** are greater than $10K. Select the **three dots** in the **CLAIM_AMOUNT** column and select **New step...**
3. A **CLEANSE** menu displays for the **CLAIM_AMOUNT** column. Select **Calculate**.
4. You want to view data from the auto claims that were paid by your company that are greater than $10K. Select the drop-down below **Operator**.
5. Now, select **Is greater than**.
6. Type "10000" in the **Value** field. Then, check the box to **Create a new column for results**.
7. It's time to select a name for your new column. Type "EXCESSIVE_CLAIM_AMOUNT" in the **New column name** field and press **Enter**.
8. You want to view this new column as the right-most in the data set, so you can keep the default under **New column position**. Select **Apply**.
9. Your new column called **EXCESSIVE_CLAIM_AMOUNT** is now added at the end of the data set! Select the **Next arrow** to continue.
10. Let's examine it. Notice the values are either **true** or **false**. This is the Boolean data type. It will be easier to view this data if you change it from Boolean to Integer. Select the **three dots** next to the **EXCESSIVE_CLAIM_AMOUNT** column.
11. Select **Convert column type...**
12. This menu gives you options for converting the data type of your column. From the **Type** drop-down menu, select **Integer**.
13. Keep everything as it is here and select **Apply**.
14. Now the data in your last column shows as **0** or **1**. **0** means the claim on that row is not over $10K. **1** means the claim on that row is over $10K. This will be very helpful data to determine if the hypothesis that claims over $10K may be fraudulent. Select the **Next arrow** to continue.

You successfully added a new column to refine the data set so you can investigate the hypothesis that claims over $10K may help predict fraud.