Simulation Transcript
Understand and clean the data

1. You've uploaded the data in your CSV file to your project, and this is a preview. You can start to explore your data set here in preview or you can start using the data refinery tool. Let's begin. Select **Prepare data** to begin the data refinement.
2. It can take several minutes for the data refinery tool to read and refine the data. Before the data set was read-only and now it is editable so you can work with it. Notice there are three tabs at the top: **Data**, **Profile**, and **Visualizations**. Select the **Next arrow** to continue.

   You're ready to understand and clean the data

   Remember, this data is from auto insurance claims that the company approved. The goal of your project is to predict fraudulent claims. You must therefore examine the **predictability** of the data in **each column**. Will the data in the column help you predict fraud? Some columns will and some columns won't. It's up to you to decide. Let's see.

   Select **X** to close this window and continue.

3. The first four columns are identification numbers, such as the driver's ID and the policy ID. This data does not help predict who might commit fraud. They are simply identifiers. Let's be sure. Select the **Profile** tab.
4. The **Profile** tab in the data refinery tool provides descriptive statistics that the tool calculated from the data. It helps you analyze and decide what to keep and what to remove. Select the **Next arrow** to continue.
5. In looking at the statistics for the first four columns, there are 975 unique occurrences. If you see statistics that merely count the rows in a CSV file, then you need to remove that column. There is nothing interesting about the data and it's not predictable. Select the **Data** tab.
6. Let's delete them. Select the **three dots** next to the **HOUSEHOLD_ID** column. This gives you several options of what you can do to the column. Select **Remove column**.
7. Do the same for the other identifier data that won't help your project to predict fraudulent behavior. Select the **three dots** next to the **DRIVER_ID** column. Select **Remove column**.
8. Select the **three dots** next to the **POLICY_ID** column. Select **Remove column**.
9. Select the **three dots** next to the **CLAIM_ID** column. Select **Remove column**.
10. You're understanding and cleaning the data now! You've gone from 38 columns to 34 columns. Let's continue to understand and clean. Select the **Next arrow** to continue.
11. The **INCIDENT_CAUSE** column is interesting. It's about the cause of the auto accident in which 1 is for driver error, 2 is for natural causes, 3 is for other driver error, 4 is for a crime, and 5 is for other causes. It might help you predict fraud. Select the **Profile** tab to investigate further.
12. The data refinery tool is showing interesting statistics for the **INCIDENT_CAUSE** field. Notice there is a value for **Standard Deviation**. What does that mean? Select the **Next arrow** to continue.

    What is standard deviation?

Standard deviation is a statistical calculation that tells you how dispersed the data is in relation to the mean.

- The mean is the average of a data set.
- A low standard deviation means that data is clustered around the mean.
- A high standard deviation indicates data are more spread out.

For you, any column on the **Profile** tab that has a value in the **Standard deviation** field could contain interesting data for predictions on your project. Keep this data.

Select **X** to close this window and continue.

13. Let's continue. Select the **Data** tab to investigate more columns of data.
14. Notice the **DESCRIPTION** column is blank. You may come across missing data when working on a project. It's not helpful. Select the **three dots** and select **Remove column**.
15. The next column for **CLAIM_STATUS** is interesting. It has data about the status of the claim in which 1 is open, 2 is approved, 3 is paid, and 4 is flagged for fraud. This column could definitely help you predict fraud, so keep it. Select the **Next arrow** to continue.
16. You can keep the **ODOMETER_AT_LOSS** column because it could be interesting to consider the car's mileage at the time of the accident. Select the **Next arrow** to continue.
17. Next, review the **LOSS_EVENT_TIME** and **CLAIM_INIT_TIME** columns. Select the **Next arrow** to continue.

    Remember to be curious!

    For this data science project, it's a good practice to keep columns that have date and time information. You never know what it could reveal in terms of predictions. For instance, if someone waited a long time to file a claim instead of filing it the day of the auto accident, what could that mean?

    Notice how you are being curious. You are like a detective following clues! Data scientists need to think this way.

    You can keep the **LOSS_EVENT_TIME** and **CLAIM_INIT_TIME** columns.

    Select **X** to close this window and continue.

18. It's important to save as you work. And save often! In the upper right, select the **Save** icon.
19. Next, the **POLICE_REPORT** column is important and could be useful. Notice that the data type is called Integer. This means it's either a 0 (meaning no) or a 1 (meaning yes). Either a police report was filed or it wasn't. You can keep this column. Select the **Next arrow** to continue.
20. Let's move ahead to the **CLAIM_AMOUNT** column. How much did it cost to fix the car after the accident? This is very interesting for your project. Your company pays these claims. Keep this column. Select the **Next arrow** to continue.
21. Next, review the **FLAG_FOR_FRAUD_INV** column. Select the **Next arrow** to continue.

    Claims have been flagged for fraud

The **FLAG_FOR_FRAUD_INV** column is unique. The data comes from the team of fraud investigators who reviewed and concluded whether the person's claim on that row was not fraud (0) or fraud (1). It was entered later, at the end of the year. These are known answers for you. This data is a training set for the tool to build a supervised training model. IBM Watson Studio will look at this column, randomly select data, and build patterns for you.

You can keep the **FLAG_FOR_FRAUD_INV** column. Also, remember it for later when you visualize the data.

Select **X** to close this window and continue.

22. Next, the **PRIMARY_DRIVER_ID** column has identifier data. Remember that ID-type data is not going to be predictable for this project. Select the **three dots** and select **Remove column**.
23. It's important to save as you work. And save often! In the upper right, select the **Save** icon.
24. Let's move ahead to the **FIRST_NAME** and **LAST_NAME** columns. Could this data help you predict fraud? Select the **Next arrow** to continue.

Be careful with personal information (PI)

Some columns in your data set represent personal information (PI) about the fictitious person who made the auto claim that was paid by your company.

Personal information is any information that relates to an identifiable, living individual. When collected, personal information can lead to identifying a particular person. For example, first and last names, phone numbers, email addresses, license plate numbers, and birth dates are all personal information. You must be aware of such data on projects and whether it's permissible to use.

For this data science project, if you think about it, personal information is not predictable for fraud. Someone's name or email address will not help you predict fraud. We'll remove this data for you.

Select **X** to close this window and continue.

25. So far, you've removed columns with data that's missing or not predictable. There's another step to clean the data. You must examine the **data type** of each column to be sure it's correct. The data refinery tool has already done some corrections. Select the **Next arrow** to continue.
26. Notice that the data type for **ODOMETER_AT_LOSS** is correctly showing as Decimal. Select the **Next arrow** to continue.
27. But, the data type for **LOSS_EVENT_TIME** is String. Strings are a sequence of letters, digits, punctuation, and so on such as a mailing address. This column needs to be correct to be a Date so the tool can make calculations. Select the **three dots**, select **Convert column type**.
28. The left window allows you to customize the data type for your column. From the **Type** drop-down menu, select **Date**.
29. Under the **Select the order...** list, you'll see that **mdy** is now selected. This means the data will display as month (m), day (d), and year (y). Select **Apply**.

30. Notice your update saved and the **LOSS_EVENT_TIME** column now correctly shows the data type as Date. The rest of the data types are going to be updated for you, so you can move ahead. Select the **Next arrow** to continue.
31. Now you're ready to save your work and create a refinery job in the tool. This creates a separate CSV file that has your changes. On the top tool bar, to the right of where you save, select **Save and create a job**.
32. To create a job for your project, start by typing "Predicting fraud in auto insurance" in the **Name** field and then press **Enter**.
33. You can leave the **Description** field blank. Select **Next**.
34. In the **Configure** section, you can accept the defaults. Select **Next**.
35. In the **Schedule** section, you can keep the schedule off. Note that if you were working long-term on this project, you may want to set the job to run at midnight if it takes a long time so it's ready when you start work. Select **Next**.
36. In the **Notify** section, you can keep notifications off. Select **Next**.
37. Notice your original CSV file that you uploaded is under **Source**. Under **Target** is your new CSV file that the data refinery tool is creating for you. It has "_shaped" in the file name. This is the file to access going forward for your project. Select **Create and run**.
38. Notice the notification! The job was successfully created. Select the **Next arrow** to continue.

You successfully cleaned the data set by removing missing columns and columns with data that's not predictable. You also corrected the data type showing for a column for proper calculations, and created the job in the tool.