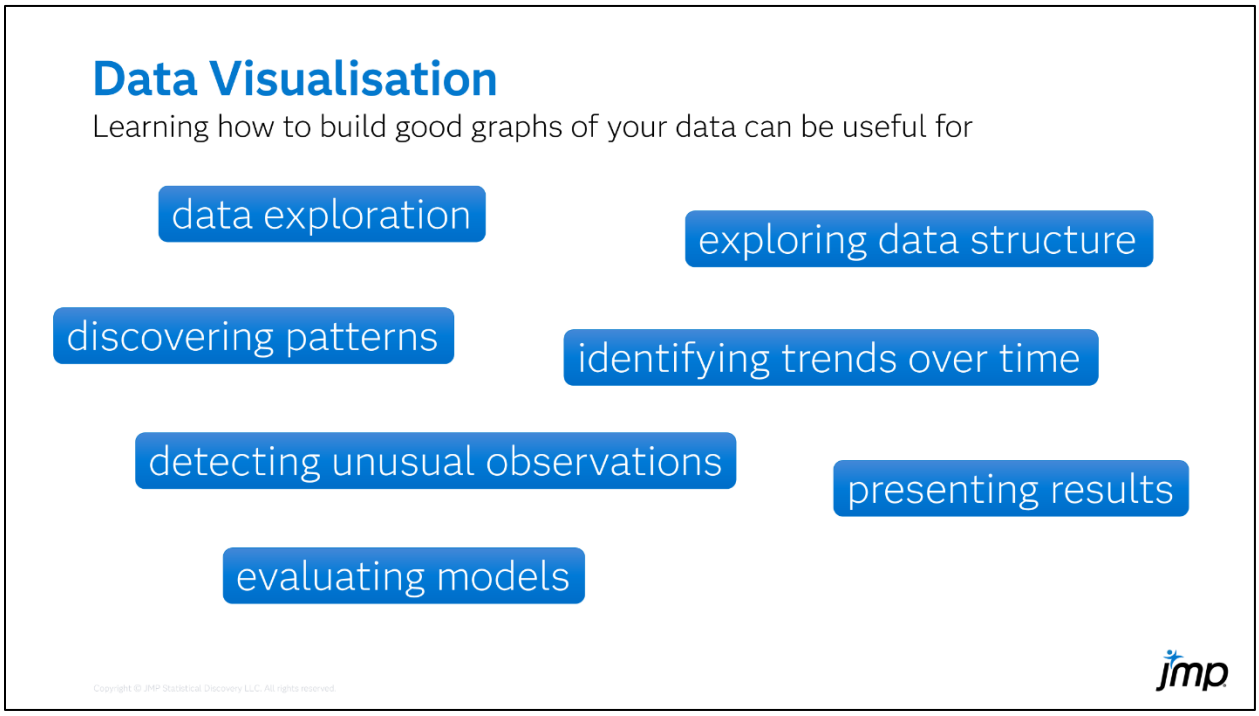
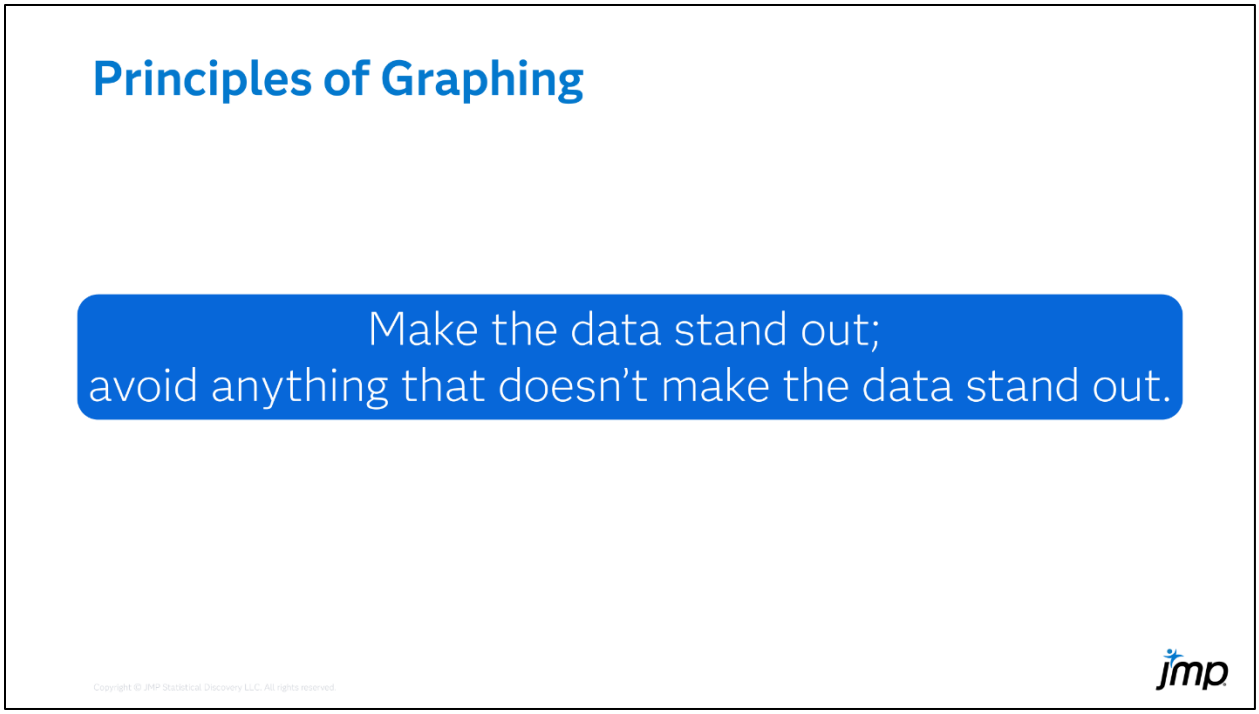
Visualising Data for Scientific Discovery: Foundations of Visual Analytics



There are many benefits to learning the skills needed to make good visualisations of your data, including:

- exploring your data
- discovering unexpected patterns
- exploring the hidden structure in your data
- detecting unusual observations or interesting data points
- finding time trends
- evaluating the fit of your statistical models
- presenting your results clearly and concisely: a picture is worth a thousand words!

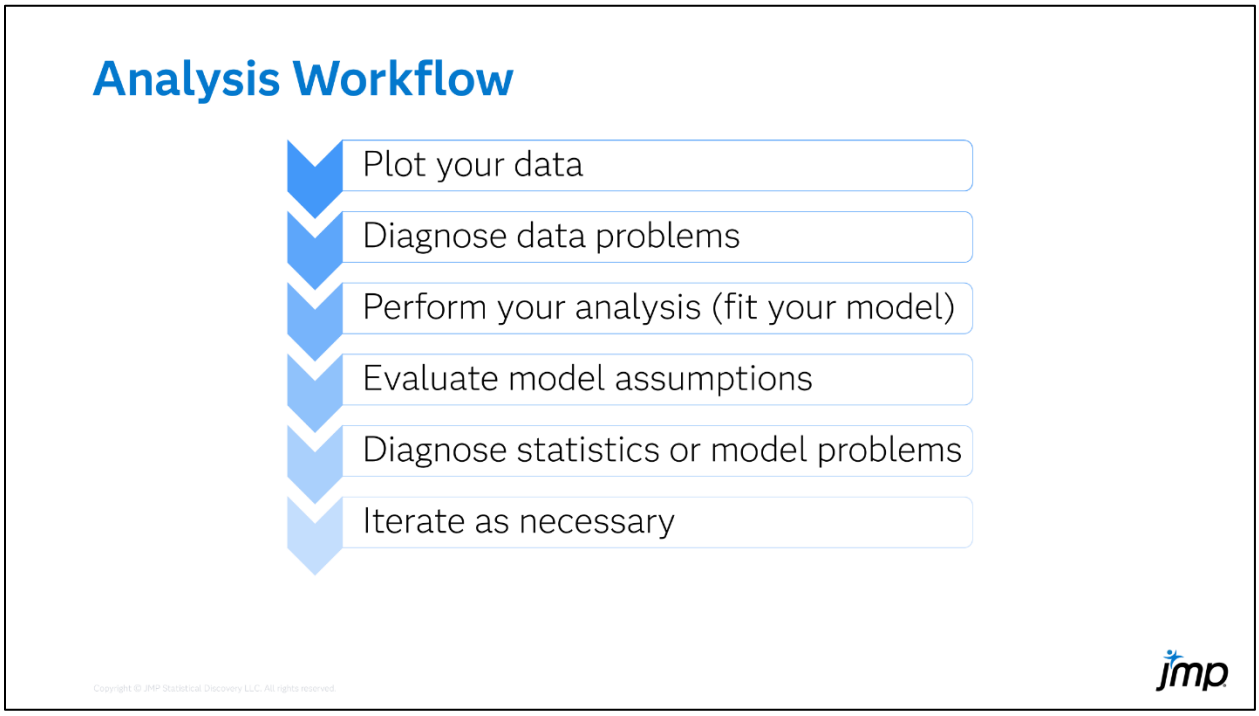


There are a number of principles for making good graphs that you can use to make good graphs. The overarching principle is to make the data stand out and avoid anything superfluous to making the data stand out. Other graphing principles follow.

- Use visually prominent graphical elements to show the data.
- Overlapping plotting symbols must be readable.
- Do not clutter the interior of the graph frame.
- Deemphasize grid lines.
- Visual clarity must be preserved under reduction and reproduction.
- Proofread graphs.
- Draw the data to scale.
- Do not show changes in one dimension by area or volume.
- Use a common baseline wherever possible.
- Graphing data should be an iterative, experimental process.
- Don't require the reader to make calculations.
- If showing improvement, plot improvement rather than before and after.

- A large amount of quantitative information can be packed into a small region.
- Use a log scale to understand percent changes or multiplicative factors.
- Showing data on a log scale can cure skewness.
- Avoid deceptive double-y axes.
- Choose an aspect ratio that shows variation in the data.
- All axes require scales.
- The horizontal axis should increase from left to right and the vertical axis from bottom to top.

In general, graphs are useful when you are analyzing data and when you need to present results. You might make many different, complicated graphs when you are exploring data, fitting models, and interpreting results. You might then summarize your findings with graphs to show others. While presenting graphs to others, follow the principles to make sure you communicate clearly and concisely.

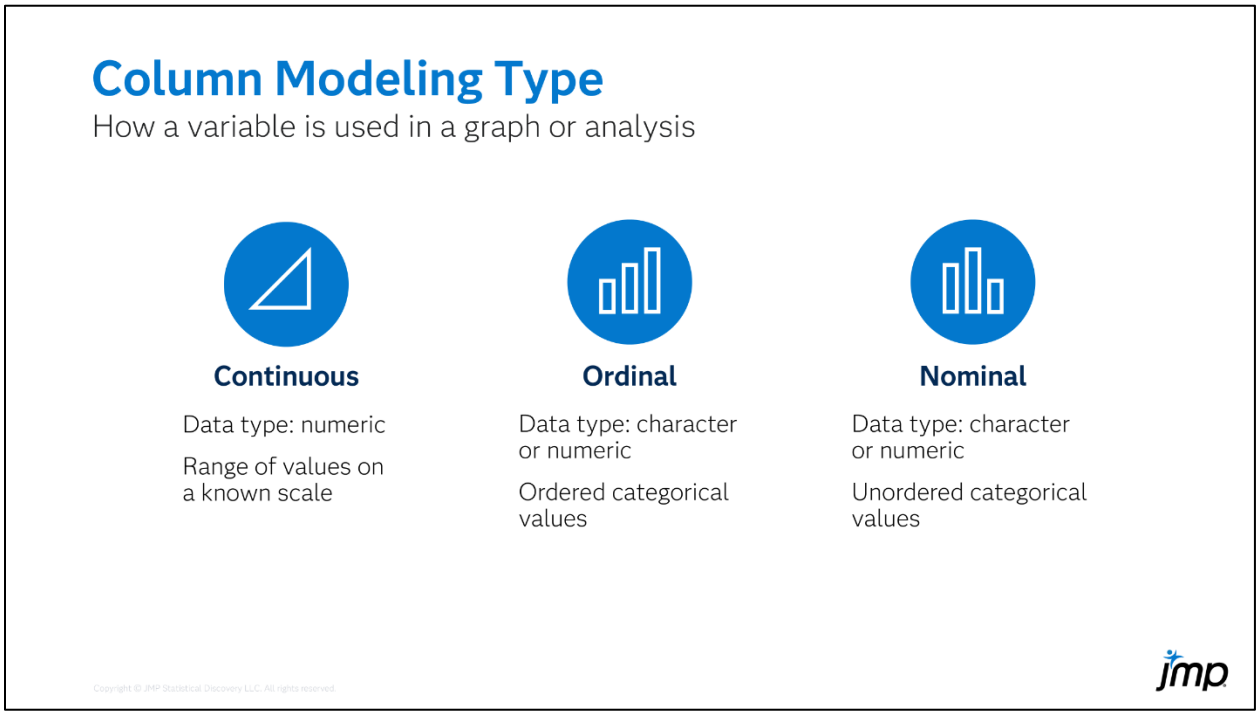


It is useful to follow a workflow in the process of statistical modeling of a dataset.

1. Plot your data to discover patterns and understand the nature of the data. Many problems are easily found this way, as well as meaningful and insightful information.

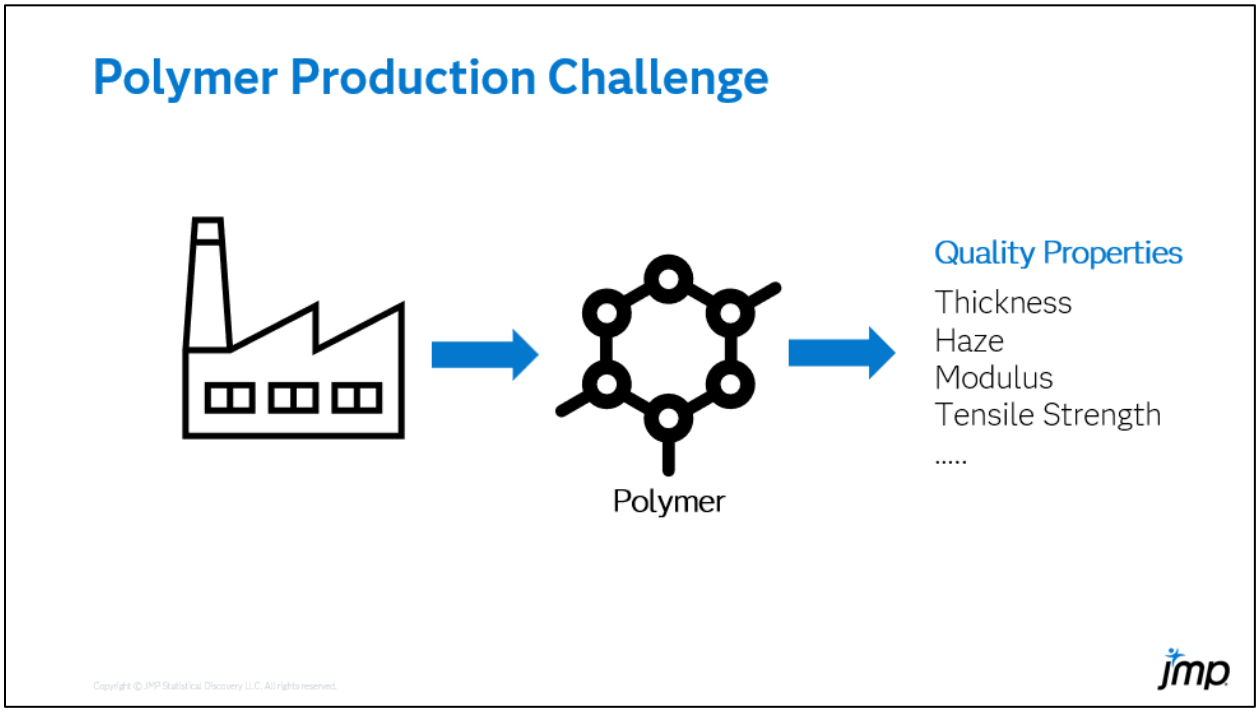
2. Diagnose any data problems before you proceed with your analysis. Correct any errors and consider excluding data that might be contaminated and invalid for your purpose.
3. Perform your analysis, both the planned analysis and any unplanned analysis that might be suggested by discoveries in the data plots.
4. Evaluate the assumptions of the model.
5. Diagnose statistics or model problems that might confuse or compromise the meaning, interpretation, or decisions based on the analysis.
6. Repeat as necessary if you encounter problems at any stage of your workflow until you are satisfied that you have a valid answer.

Graphing data is especially useful in steps 1, 2, and 4.

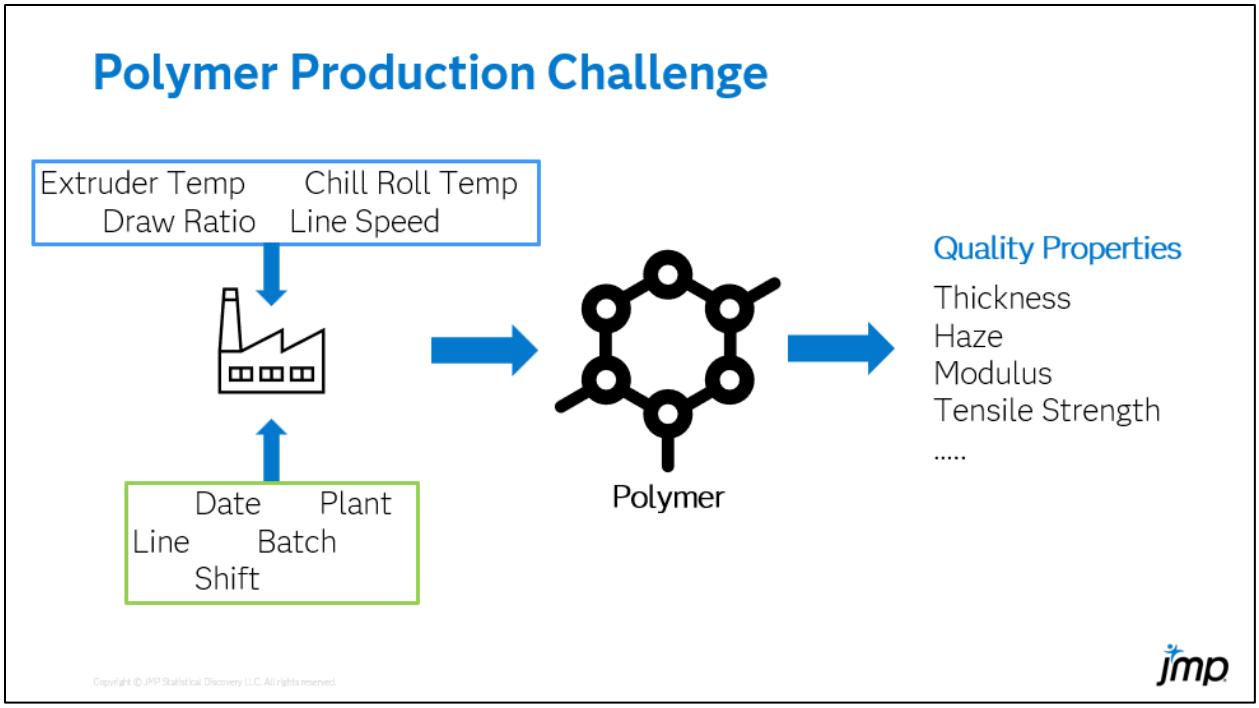


Different types of data lend themselves to different types of graphs. There are three types of data used for graphing. *Continuous* data are numeric data measured over a range of values on a known scale. Examples of continuous variables include concentration, temperature, absorbance, and reaction time. *Ordinal* data are either character or numeric data with a finite number of levels or categories. These categories contain an inherent ordering. Examples of ordinal variables are the size of shirt you are wearing (Small, Medium, Large) purity categories, yield categories, or demographic variables like batch number in order of processing. *Nominal* data are also either character or numeric data with a finite number of

levels, but these categories have no order. Examples include the color of shirt you are wearing (white, blue, black), the type of catalyst, film type, shift (during the workday), or which plant produced the chemical. The type of data you have determines the type of graph you can create.



A company faces quality issues with one of their core polymer film production processes. Due to low yield and occasional occurrence of defects they face friction with customers and have to justify the lost revenue of scrapped batches to their stakeholders. A team of engineers were tasked to investigate how to fix the issues with low yield and track down the occurrence of defects in production.



To illustrate the concepts of graphing data, we will use simulated data collected from a polymer film process. This rich dataset contains information on each film produced, including:

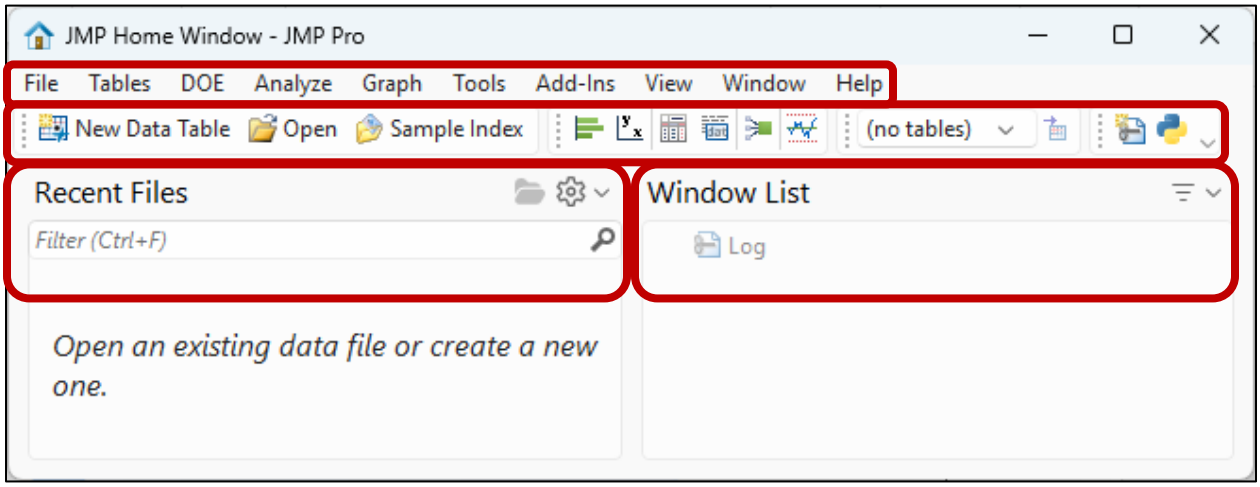
- Production data such as date, which plant (two plants), which line in the plant (two lines in each plant), which batch of consumables, which shift produced the film, and which film type was produced (PET, PP, PE).
- Process data such as extruder temperature, chill roll temperature, draw ratio, and line speed.
- Properties of the film, such as thickness, haze, tensile modulus, opacity, coefficient of friction, water evaporation transmission rate, oxygen transmission rate, yield, and defects per 1000 m of film.

Demonstration

In this demonstration, you will import the polymer film data from Excel into JMP using the JMP Excel Import Wizard. Next, you'll explore the data one variable at a time, learning about the context of the data from the production variables and the shape of the data from the other variables. Any data problems can be easily corrected using JMP's data cleaning functions. Finally, you'll look at relationships between two or more variables using graphs.

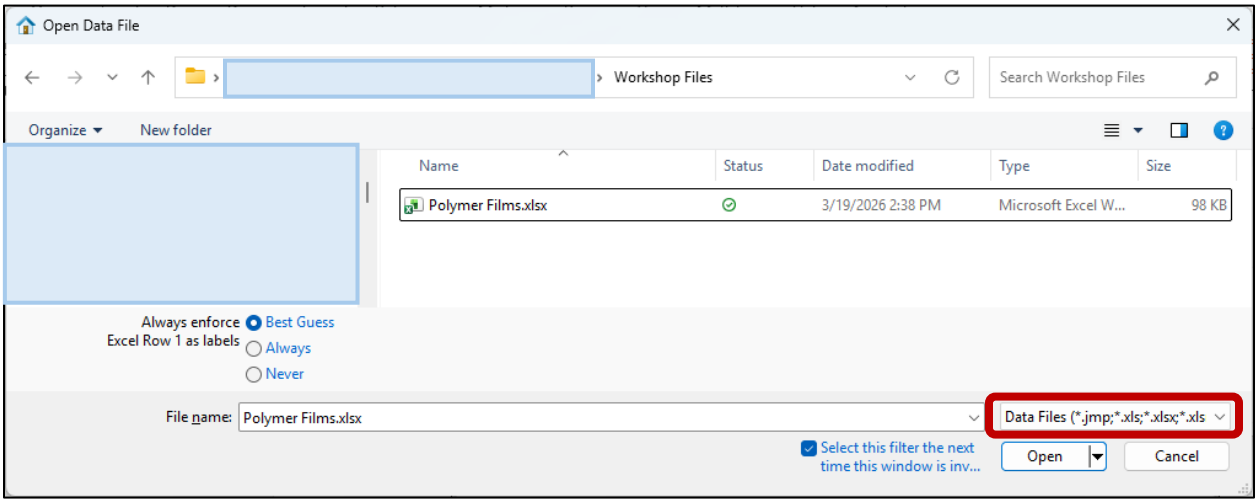
Data Import

1. Open JMP.



The JMP Home Window consists of a menu bar, toolbars with buttons for quick access to common functions, a region containing recent files that can be used to quickly access data tables and other JMP files, and a region containing any open JMP windows.

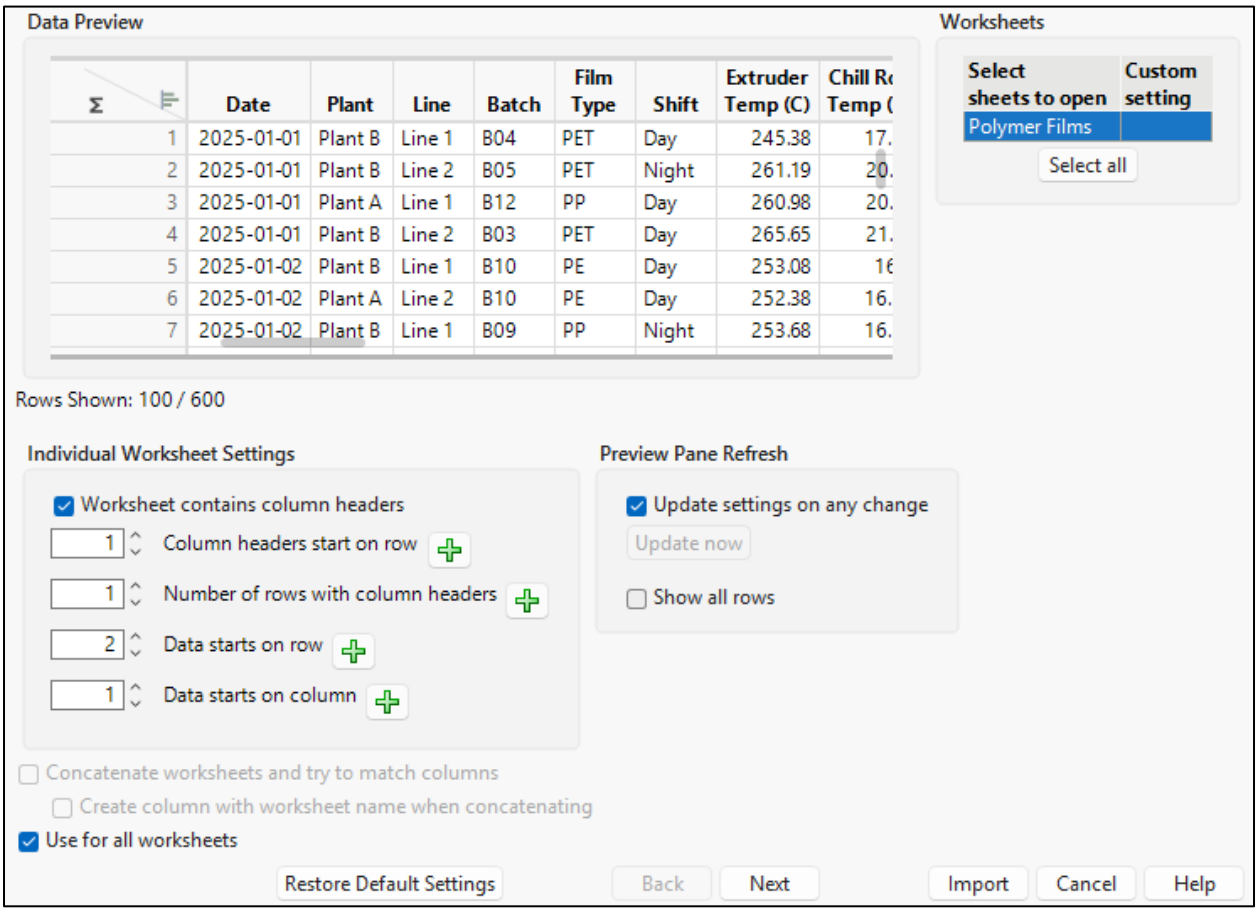
2. Select **File** > **Open**.
3. Navigate to the folder with the course data.



If you do not see the Excel file in the list, click the drop-down menu in the bottom right and select **Data Files**.

4. Select **Polymer Films.xlsx**.
5. Click **Open**.

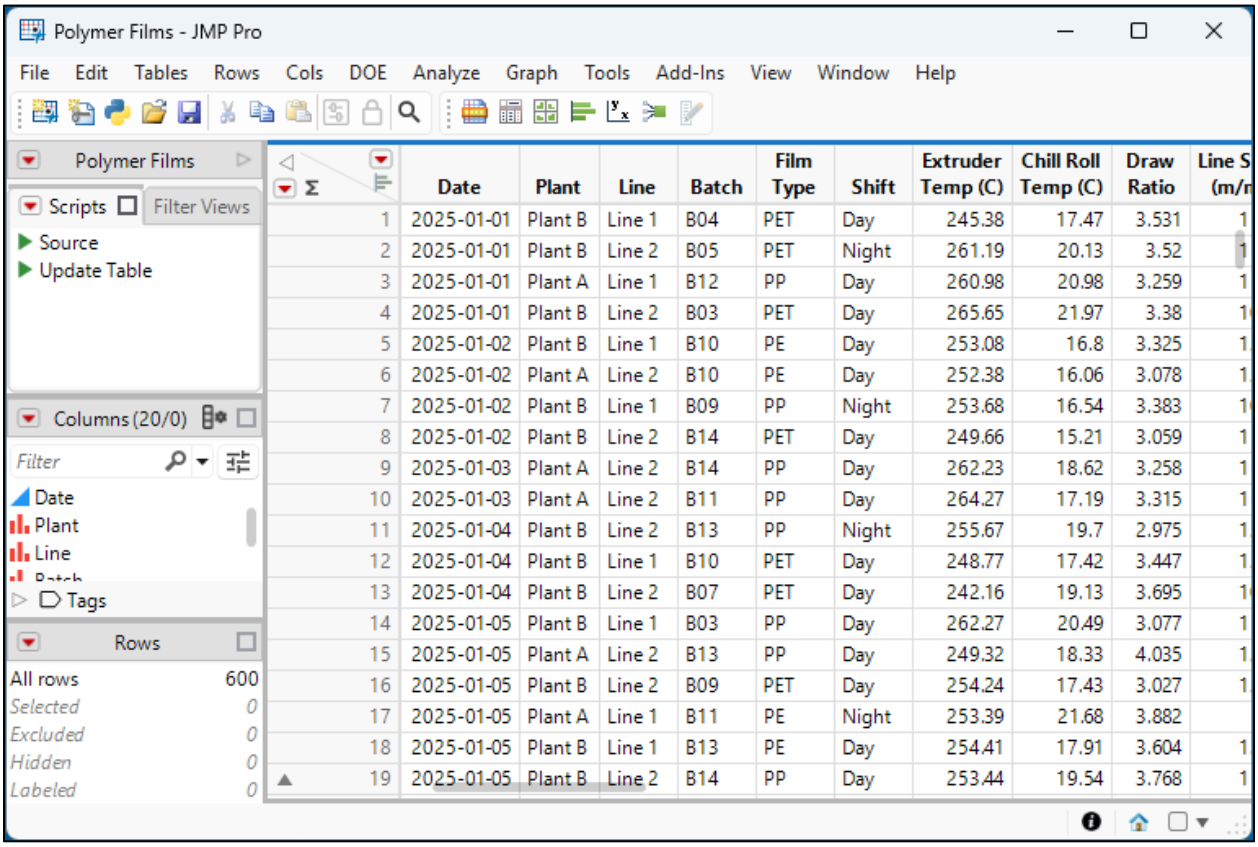
6. Click **Restore Default Settings**.



The Excel Import Wizard allows you to easily specify the format of your data in Excel so that it is imported correctly into JMP. You can select worksheets in the Excel workbook, import column headers as variable names, import any number of rows as column header information, import only certain columns and rows, as well as combine worksheets into one data table if needed.

It is important to note that while Excel is cell-based, that is, data of any type can live in a column in Excel, JMP is column-based, that is, all data in the same column needs to have the same type of data. JMP expects columns to represent variables and rows to represent observations of those variables, so data should be in tabular format. Because Excel doesn't require data in a tabular format, the Excel Import Wizard is useful in importing only the necessary data into JMP with minimal reformatting needed.

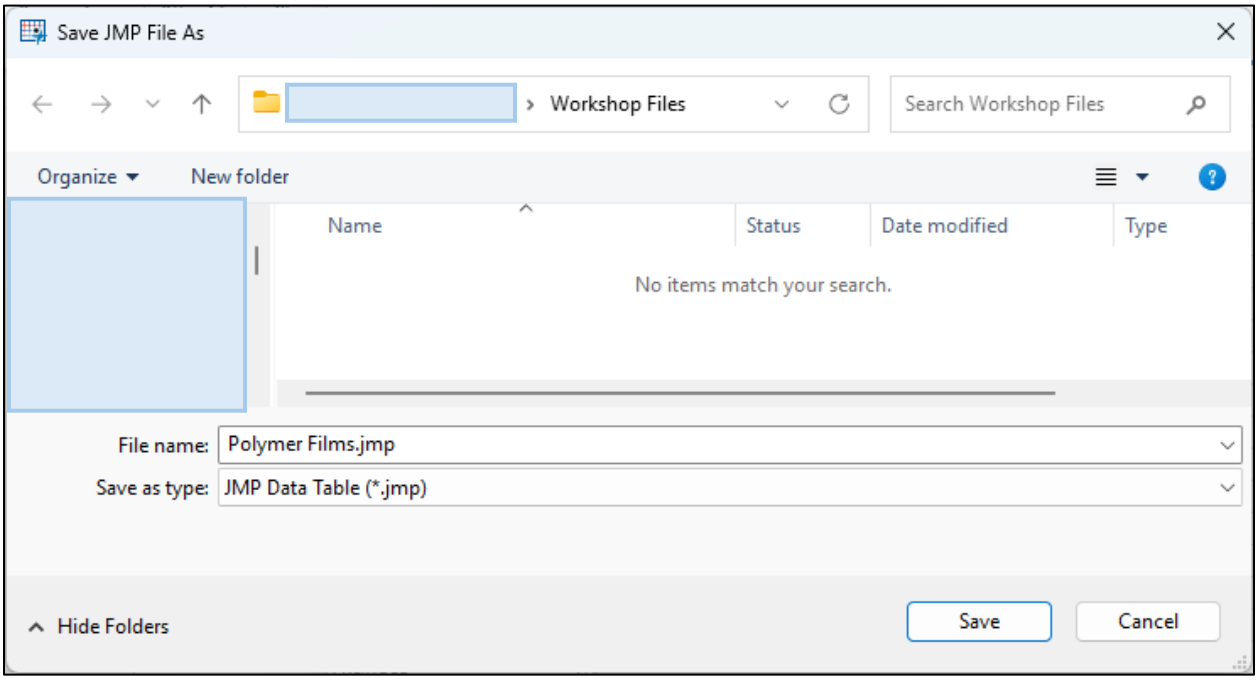
7. Click **Import**.



	Date	Plant	Line	Batch	Film Type	Shift	Extruder Temp (C)	Chill Roll Temp (C)	Draw Ratio	Line S (m/r
1	2025-01-01	Plant B	Line 1	B04	PET	Day	245.38	17.47	3.531	1
2	2025-01-01	Plant B	Line 2	B05	PET	Night	261.19	20.13	3.52	1
3	2025-01-01	Plant A	Line 1	B12	PP	Day	260.98	20.98	3.259	1
4	2025-01-01	Plant B	Line 2	B03	PET	Day	265.65	21.97	3.38	1
5	2025-01-02	Plant B	Line 1	B10	PE	Day	253.08	16.8	3.325	1
6	2025-01-02	Plant A	Line 2	B10	PE	Day	252.38	16.06	3.078	1
7	2025-01-02	Plant B	Line 1	B09	PP	Night	253.68	16.54	3.383	1
8	2025-01-02	Plant B	Line 2	B14	PET	Day	249.66	15.21	3.059	1
9	2025-01-03	Plant A	Line 2	B14	PP	Day	262.23	18.62	3.258	1
10	2025-01-03	Plant A	Line 2	B11	PP	Day	264.27	17.19	3.315	1
11	2025-01-04	Plant B	Line 2	B13	PP	Night	255.67	19.7	2.975	1
12	2025-01-04	Plant B	Line 1	B10	PET	Day	248.77	17.42	3.447	1
13	2025-01-04	Plant B	Line 2	B07	PET	Day	242.16	19.13	3.695	1
14	2025-01-05	Plant B	Line 1	B03	PP	Day	262.27	20.49	3.077	1
15	2025-01-05	Plant A	Line 2	B13	PP	Day	249.32	18.33	4.035	1
16	2025-01-05	Plant B	Line 2	B09	PET	Day	254.24	17.43	3.027	1
17	2025-01-05	Plant A	Line 1	B11	PE	Night	253.39	21.68	3.882	
18	2025-01-05	Plant B	Line 1	B13	PE	Day	254.41	17.91	3.604	1
19	2025-01-05	Plant B	Line 2	B14	PP	Day	253.44	19.54	3.768	1

The data have been imported as a JMP table. There are 20 columns (variables) and 600 rows (observations). Save this table.

8. Select **File** > **Save As**.



9. Click **Save**.

Graphs for One Variable

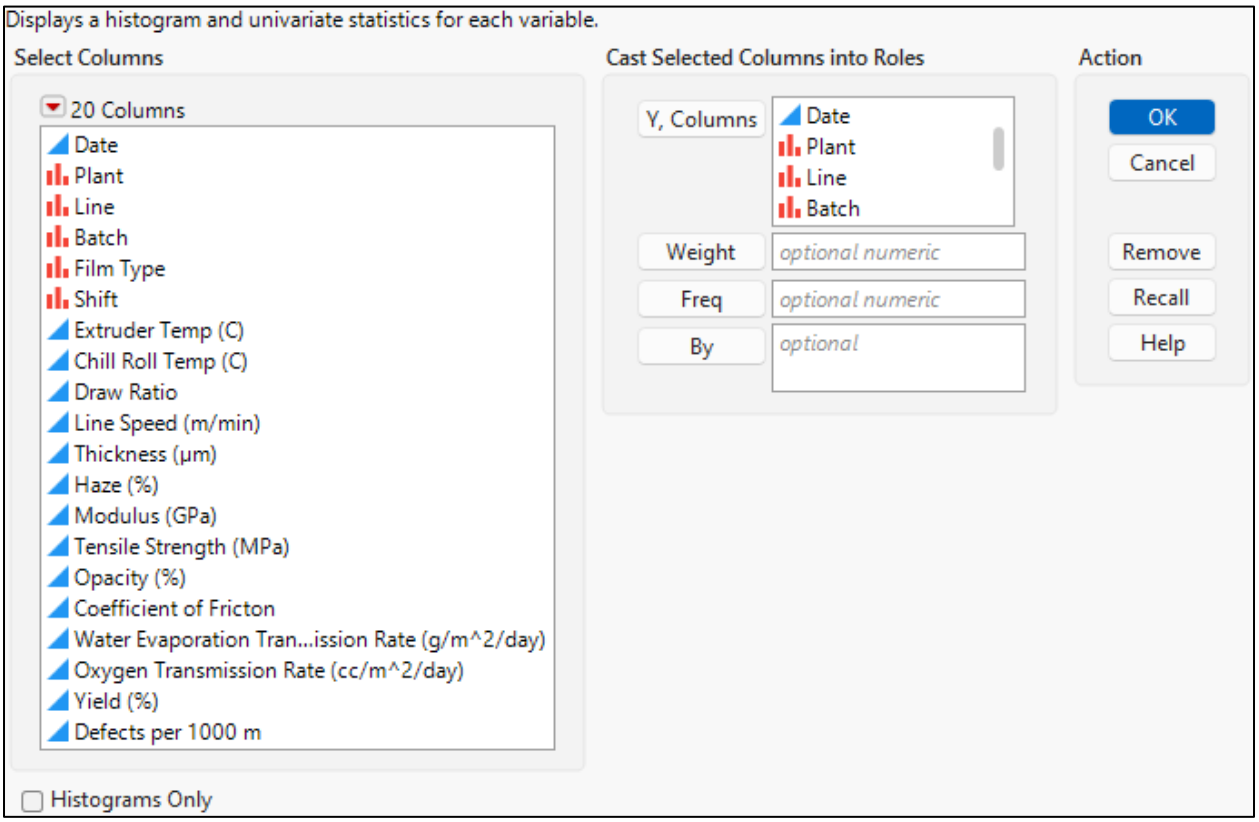
In this section, you will learn about dot plots, bar charts, histograms and boxplots, used to display data from one variable. In JMP, you can use the Distribution platform to examine multiple variables at the same time, or Graph Builder to control the graphs and layout.

Distribution

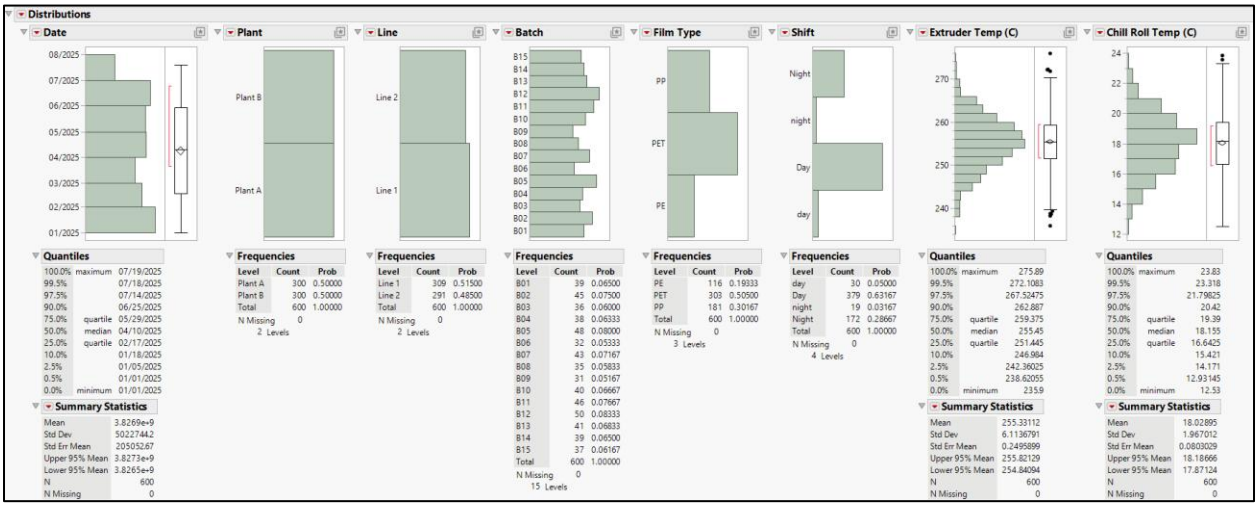
How many levels of each categorical variable exist? What proportion of each film type was manufactured? What do the distributions of data look like for the continuous variables? Are there any missing values for any variable? Are there any potential outlier observations? Do the distributions of some variables depend on other variables?

1. Select **Analyze** > **Distribution**.

2. Select all the columns in the **Select Columns** list on the left, then click **Y, Columns**.



3. Click **OK**.



For each variable in the dataset, JMP provides a graph and summary statistics.

If the variable has a nominal or ordinal modeling type (for example, **Plant** or **Film Type**), JMP provides a *bar chart*. A bar chart displays the number of observations for each categorical level of the variable. You can see that the number of observations for each plant is about the same, while there are more PET films than PP or PE films in the dataset.

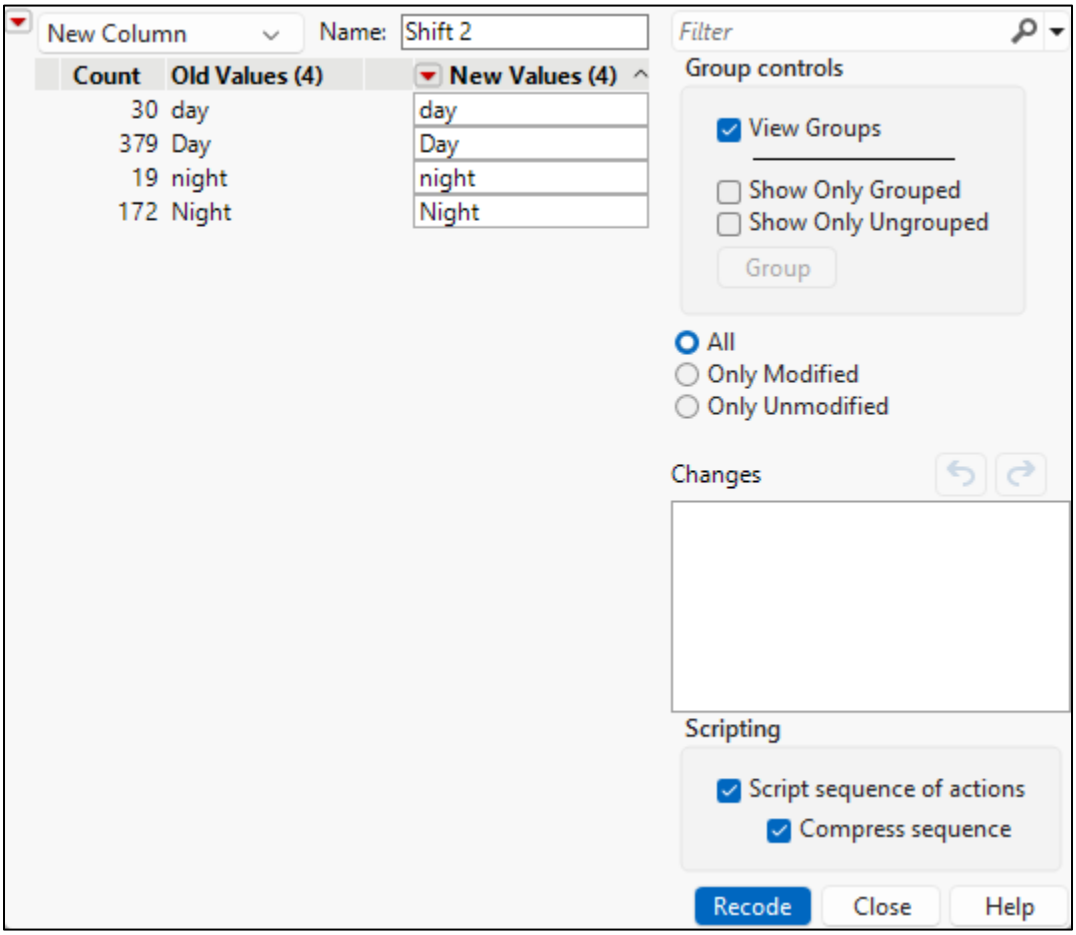
How many levels of each categorical variable are there?

Plant: 2, **Line**: 2, **Batch**: 15, **Film Type**: 3, **Shift**: 4.

The values of the **Shift** column don't look right. There are four levels, but some entries are capitalized and some are not. They should all have the same case. You can easily change all the erroneous data. First, set up the report to automatically recalculate when the data have changed.

4. Click the red triangle next to **Distributions** and select **Redo** > **Automatic Recalc**.
5. Return to the data table and select the **Shift** column.

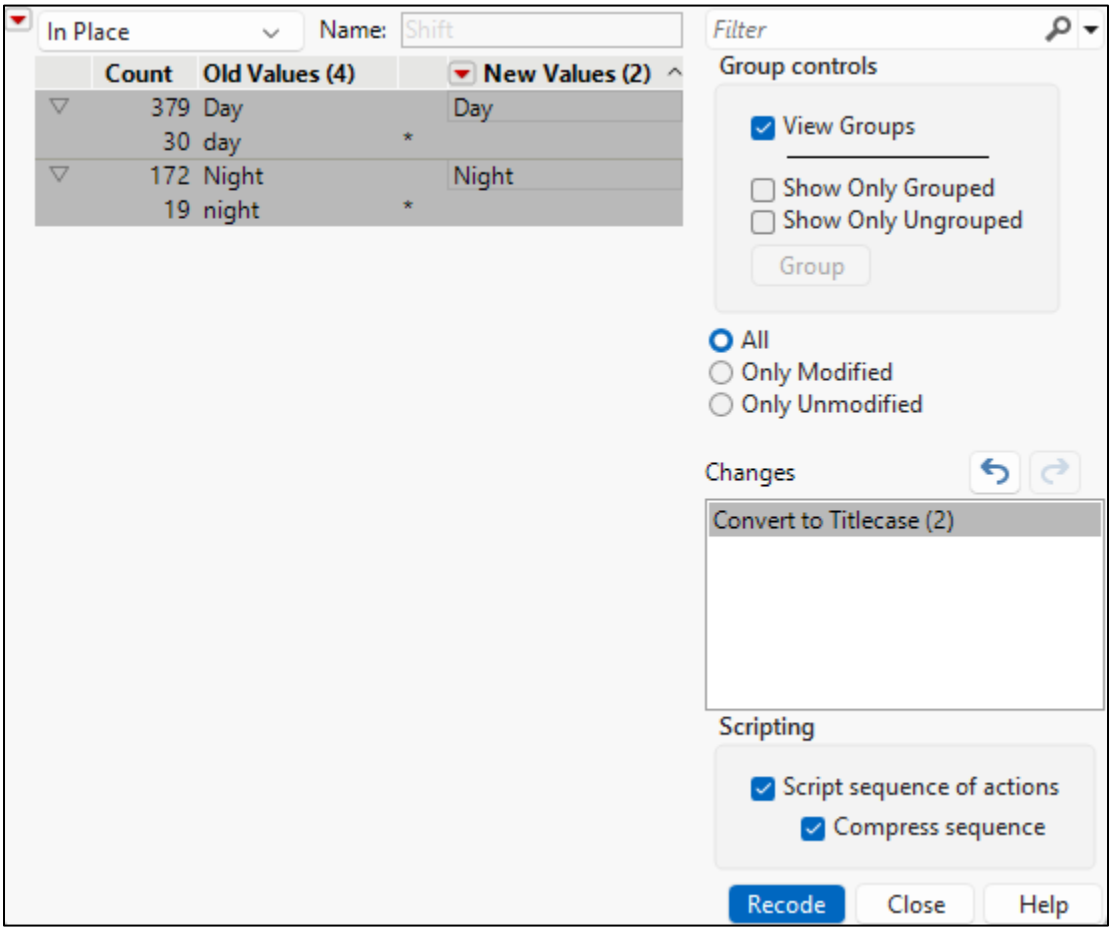
6. Select **Cols** > **Recode**.



You can click in the **New Values** boxes and type new values or use the helper options in the red triangle menu.

7. Click the red triangle in the upper left corner of the window and select **Convert to Titlecase**.

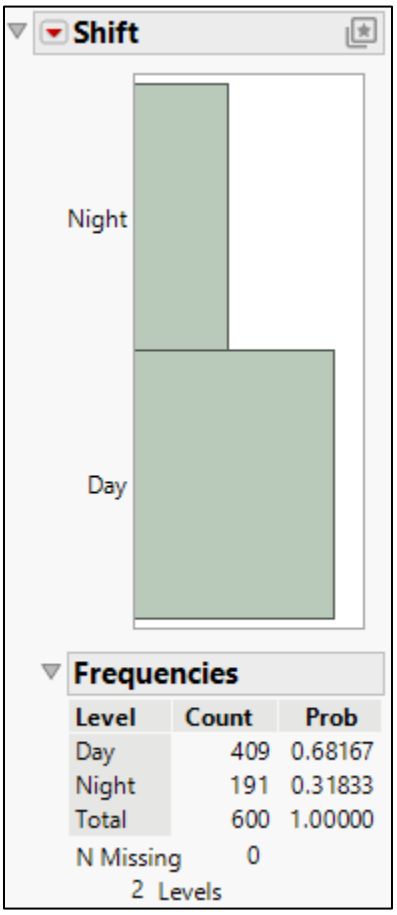
8. Click the **New Column** drop-down and select **In Place**.



9. Click **Recode**.

10. Deselect the **Shift** column.

11. Return to the **Distribution** report window.



Shift now has the correct two levels.

What proportion of each film type was manufactured?

The table below the **Film Type** graph has the proportions: PE: about 20%, PET: about 50%, PP: about 30%.

If the variable has a continuous modeling type (for example, **Date** or **Extruder Temp**), JMP provides a *histogram*. A histogram is a bar chart, where the length of the bar represents the number of observations in an interval. For example, each bar in the **Date** histogram represents one month; each bar in the **Extruder Temp** histogram represents 2 degrees C. From a histogram, you can see the shape of a distribution.

For the continuous variables, JMP provides both a histogram and a *boxplot*. A boxplot is a simpler graph for one variable. It shows the five number summary of the data: minimum, 25th percentile (first quartile), median (50th percentile, second quartile), 75th percentile (third quartile), and maximum. In the Distribution platform,

JMP displays an *outlier boxplot*, which highlights observations outside of fences based on the median and distance between the third and first quartiles.

What do the distributions of data look like for the continuous variables?

The shape of **Date** is fairly uniformly distributed; the shapes of **Extruder Temp** through **Line Speed** are fairly bell-shaped. Other variables, like **Thickness** or **Haze**, have more than one mode (bump). Multiple modes often indicate there is more than one source of variation in the data.

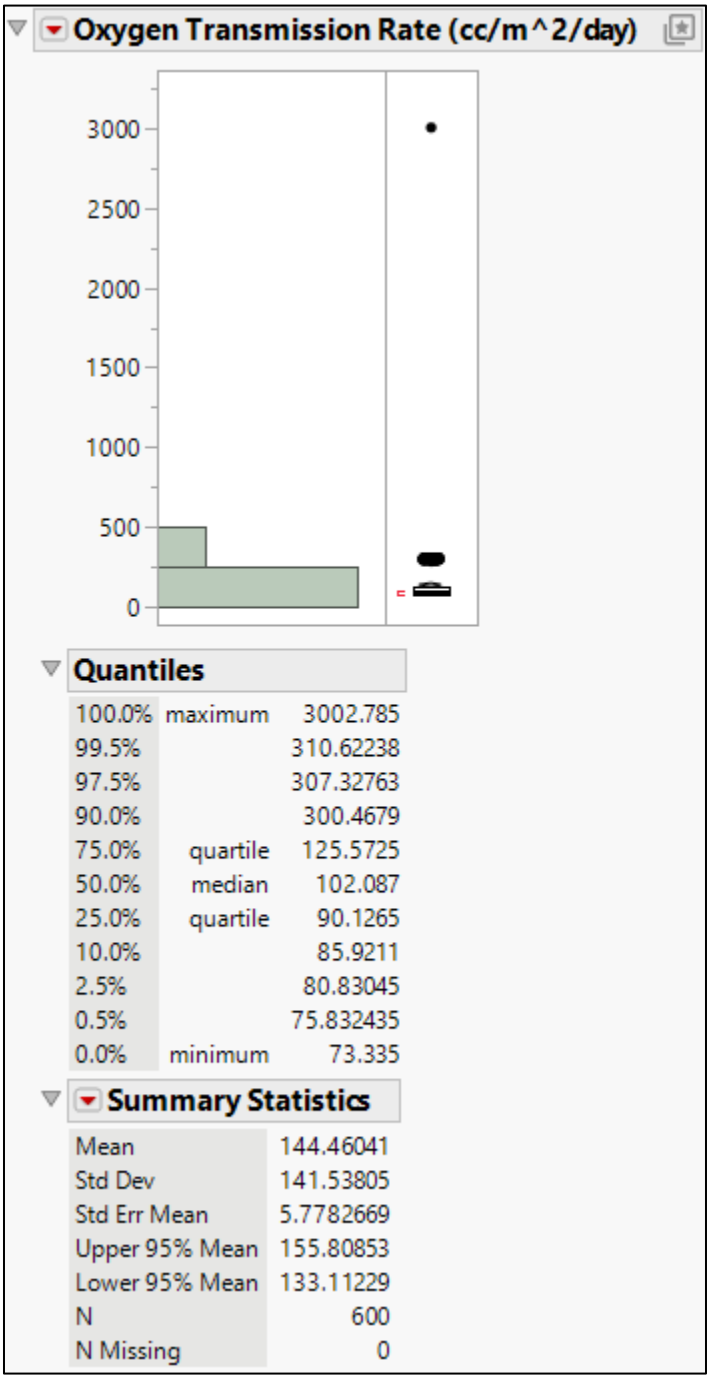
Are there any missing values for any variable?

The reports below the histograms indicate how many missing data exist (**N Missing**). Scroll across to see that no variables have missing values.

Are there any potential outlier observations?

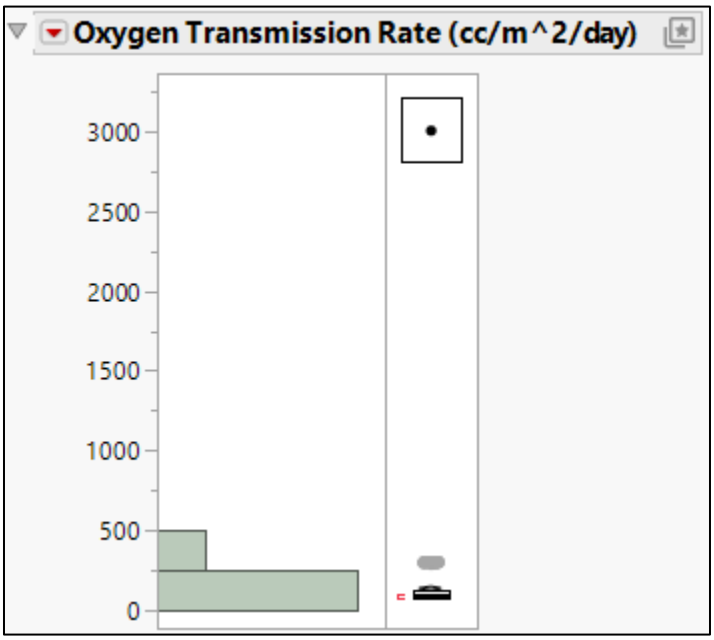
Outliers are observations that do not belong to the same distribution as other data. They were generated by another process. You can use histograms and outlier boxplots to detect potential outliers. Next, you can use your expertise or that of the person who collected the data to determine the reason for any anomalous data values.

12. Scroll through the continuous variables, looking for strange behavior.



There is one observation for **Oxygen Transmission Rate** which is an order of magnitude higher than other observations.

13. Select the observation in the graph.



The row has been selected in the data table.

14. Return to the data table and select **Rows** > **Next Selected**.

Water Evaporation Transmission ...	Oxygen Transmission Rate (cc/m^2/day)	Yield (%)	Defects per 1000 m
5.5533	3002.785	80.227	1.757

After discussion with the engineer, it was determined that the data point was incorrectly transcribed.

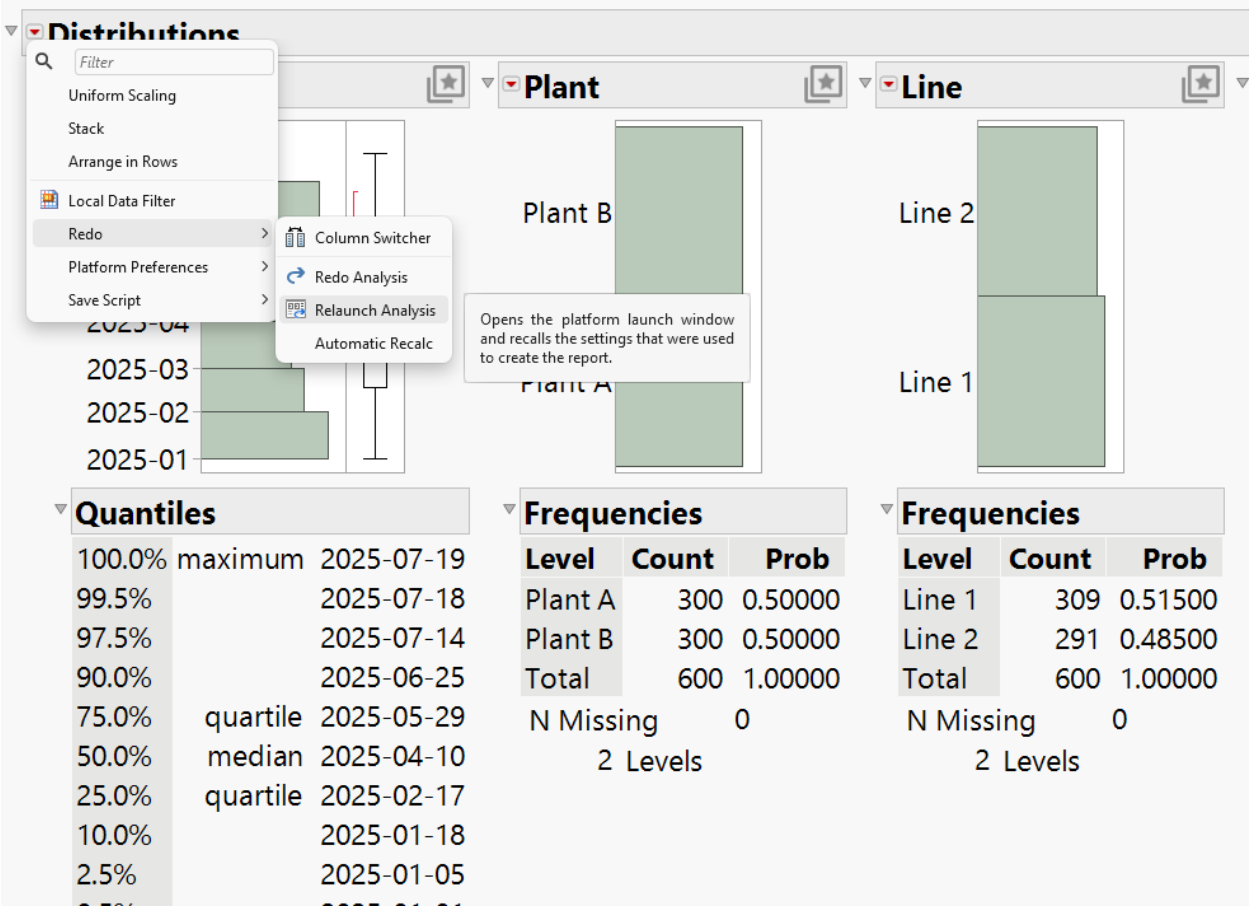
15. Double-click the cell for **Oxygen Transmission Rate** in row 99 and change 3002.785 to 302.785.

16. Deselect all rows and columns.

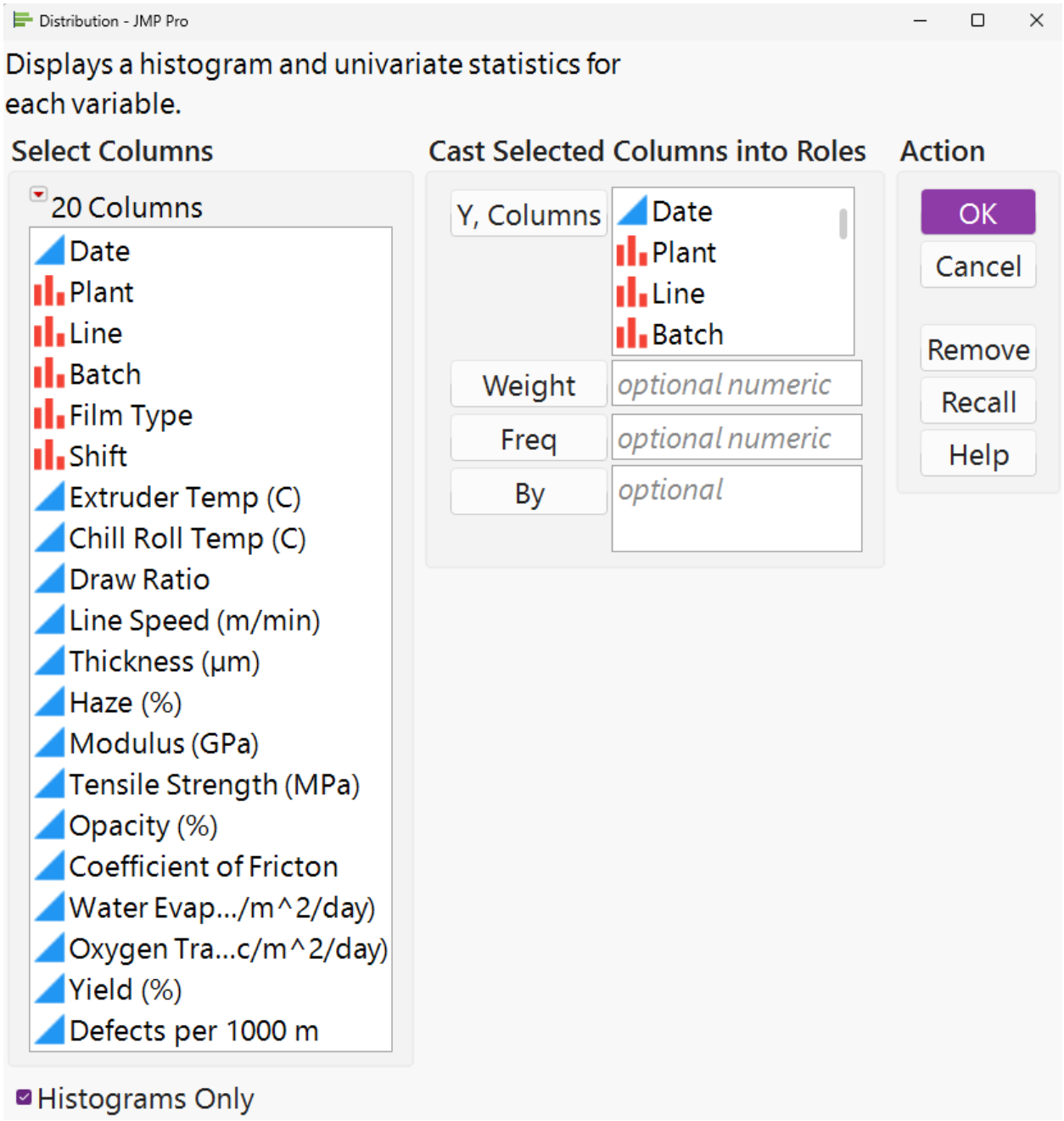
Do the distributions of some variables depend on other variables?

Some of the histograms for continuous variables have multiple modes. You can use JMP's interactivity to see if the modes are due to one or more of the other variables.

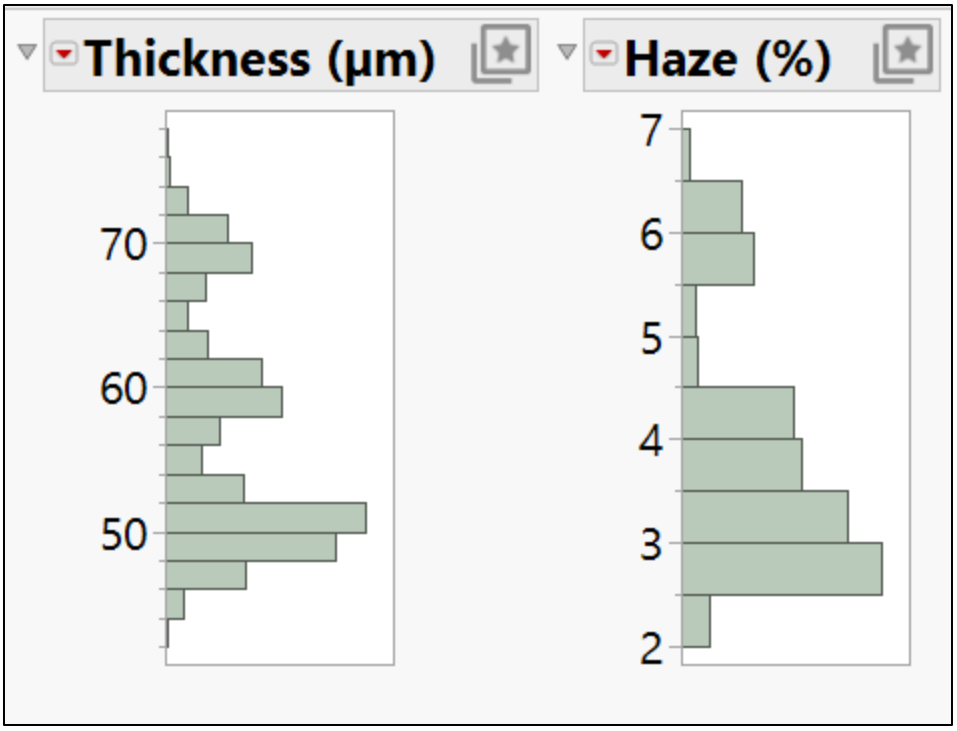
17. To simplify the visualizations, the platform is going to relaunched with a simplified view. Click the red button and go to Redo > Relaunch Analysis.



18. On the Dialog window, go to the bottom left and select the 'Histograms Only' option.

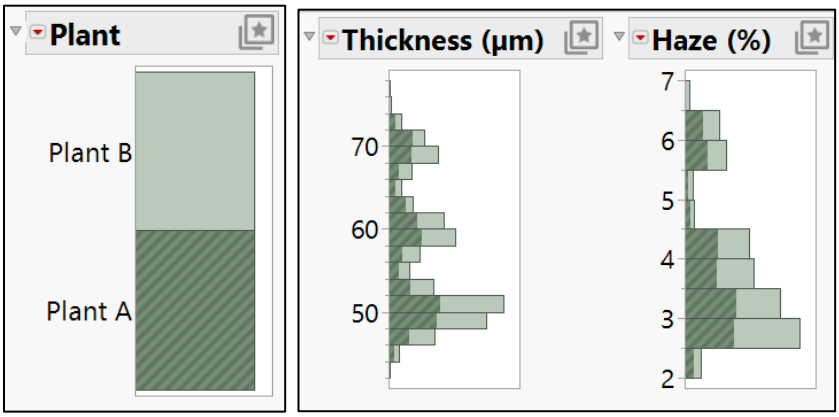


19. Select 'OK'.
20. Scroll to the histograms for **Thickness** and **Haze**.



The variables have multiple modes.

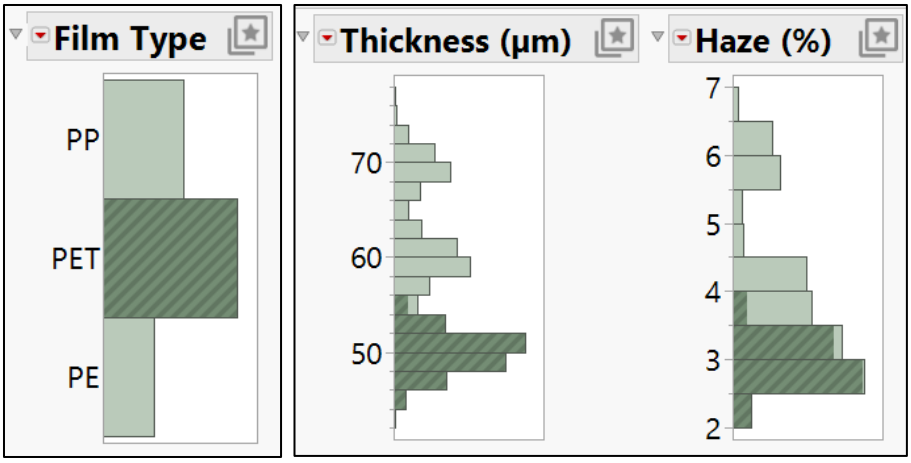
21. Click the bar for **Plant A** in the **Plant** histogram.



When you click a bar in a bar chart or histogram, all the rows for that level of the variable are selected in the data table. Those rows are therefore highlighted in all other open graphs.

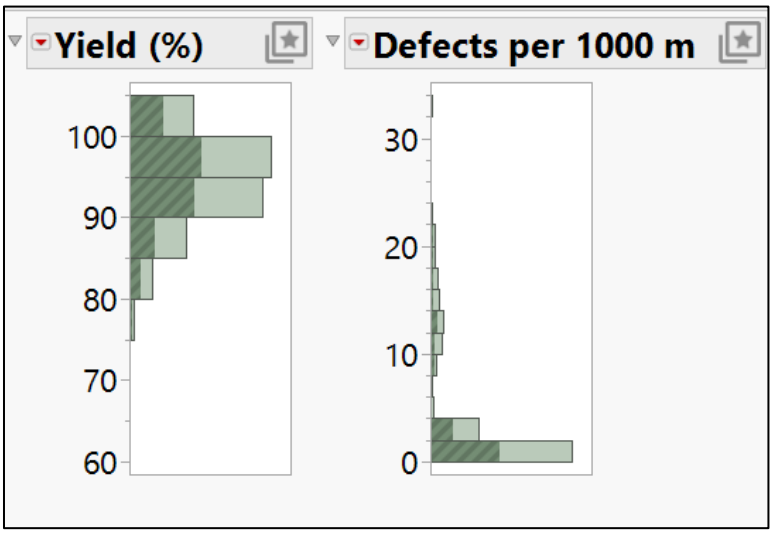
In this case, the patterns in the histograms for **Thickness** and other variables do not appear to be caused by different **Plants**.

22. Select the **PET** bar in the **Film Type** report.



The strong signals in these variables (and many others) seem to be coming from different film types.

Examine the **Yield** and **Defects** histograms.



Some of the poor yielding and defect-ridden lots are using the PET film type, but many are not. We must look elsewhere for a cause.

Next, use the Graph Builder platform to examine the data.

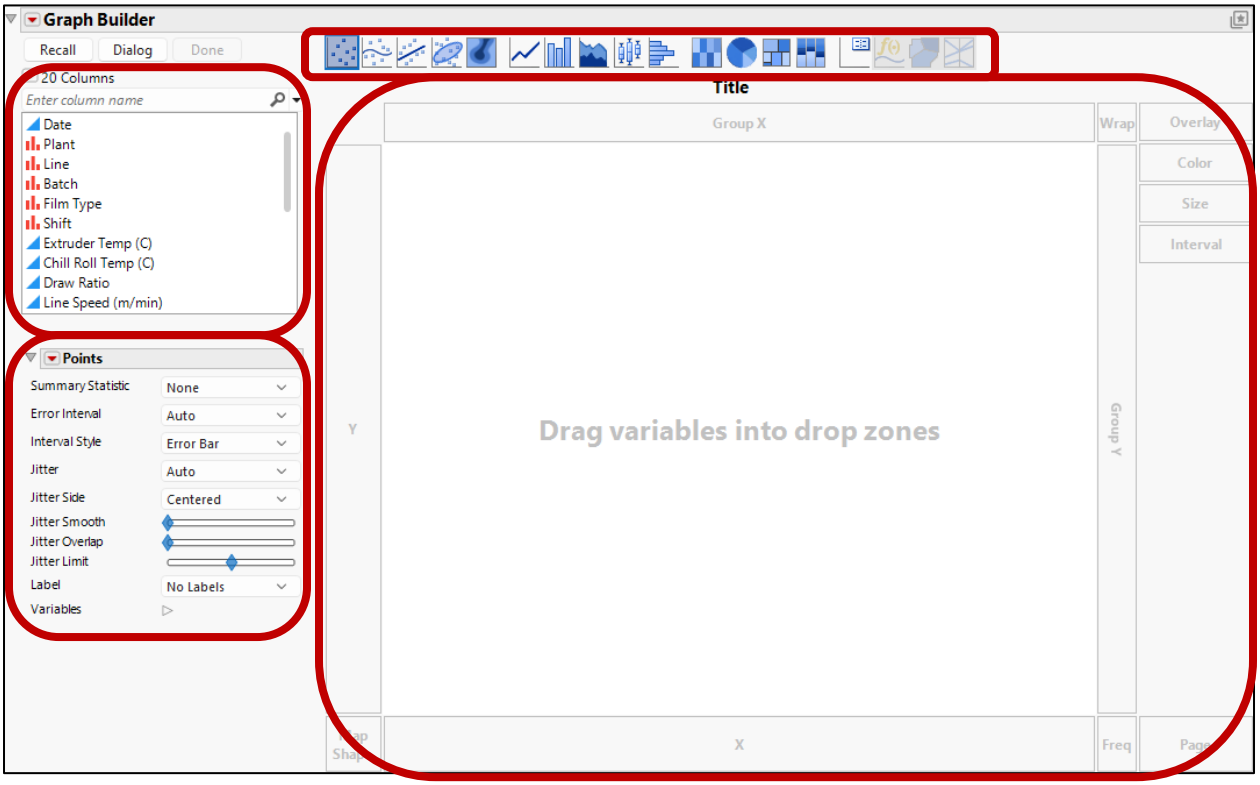
23. Press and hold the Ctrl key (Windows) or Command key (Mac), then click the **PET** bar in the **Film Type** report to deselect the rows.
24. Return to the data table and select **File** > **Save**.

Graph Builder

In the past, defects were tied to poor maintenance practices of the processing equipment. We can use Graph Builder to examine the relationship between yield and defects and other variables.

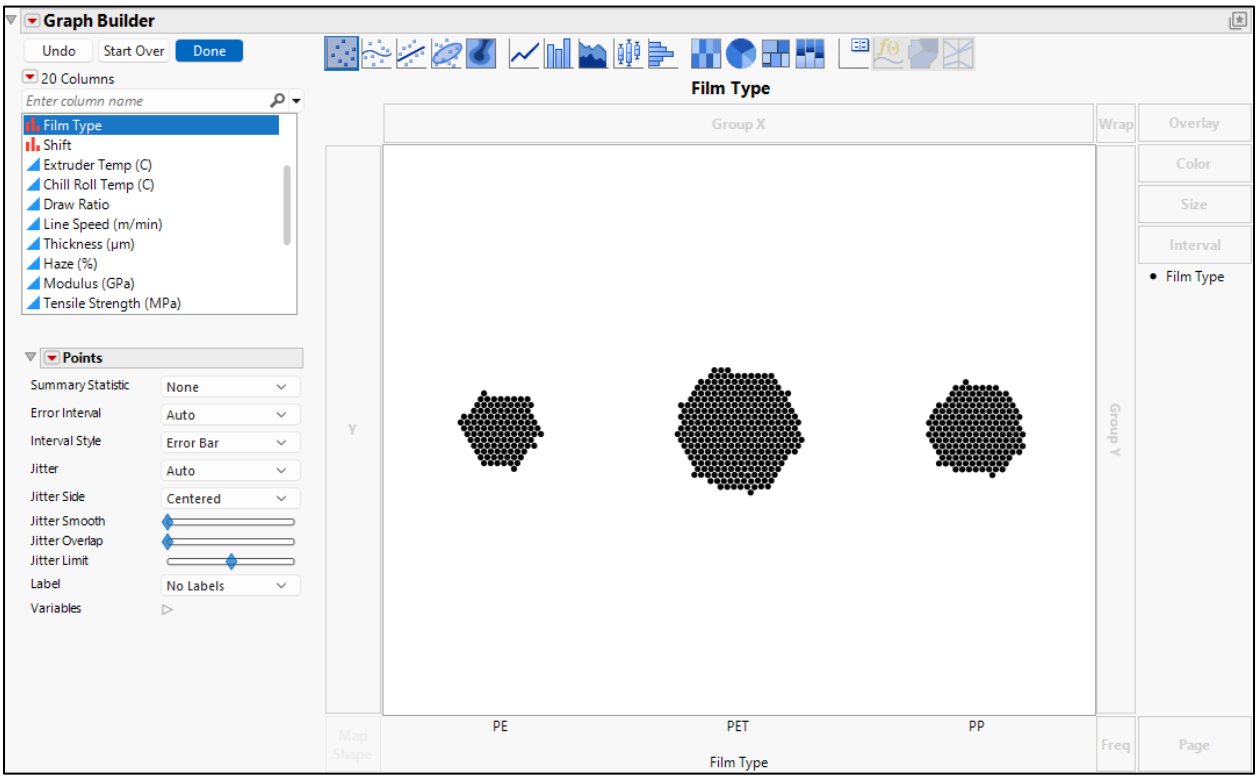
Graph Builder is a powerful, interactive platform for data visualization. You do not need to know the type of graph you want to create before beginning in Graph Builder. Let's look at some simple graphs before using the platform to answer our research question.

1. Select **Graph** > **Graph Builder**.



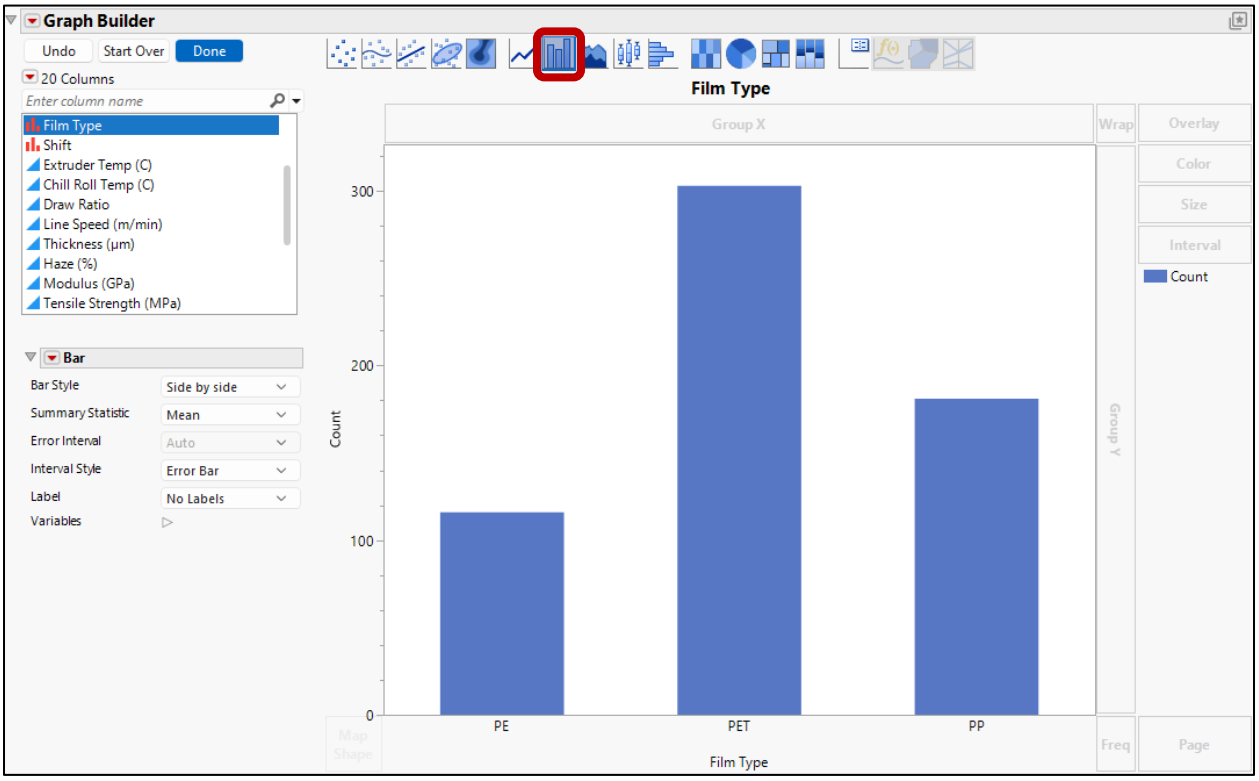
The report window contains regions for selecting variables (top left), modifying graphing elements (bottom left, will change as variables are added to roles), selecting graphing elements (top of graphing area), and displaying the graph (middle of window).

2. Drag **Film Type** to the X drop zone.



A dot plot is shown. There is a dot for each row in the data table, so the size of the dot mass represents the number of observations in each level of **Film Type**.

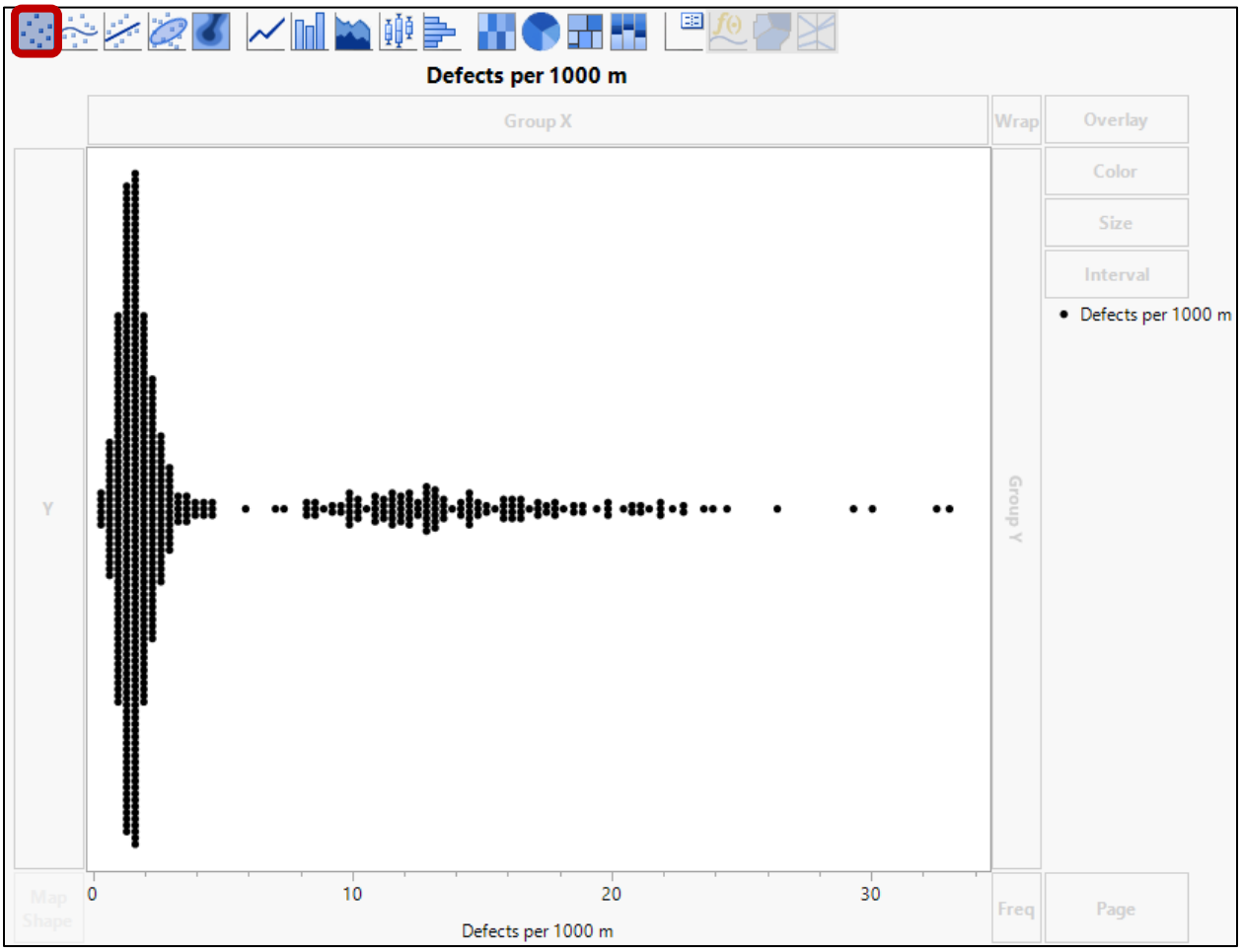
3. Click the Bar element icon in the element toolbar above the graph.



The bar chart shows the number of observations within each category. It is easier for human brains to compare length than area, so a bar chart is usually preferable to the dot plot or a pie chart.

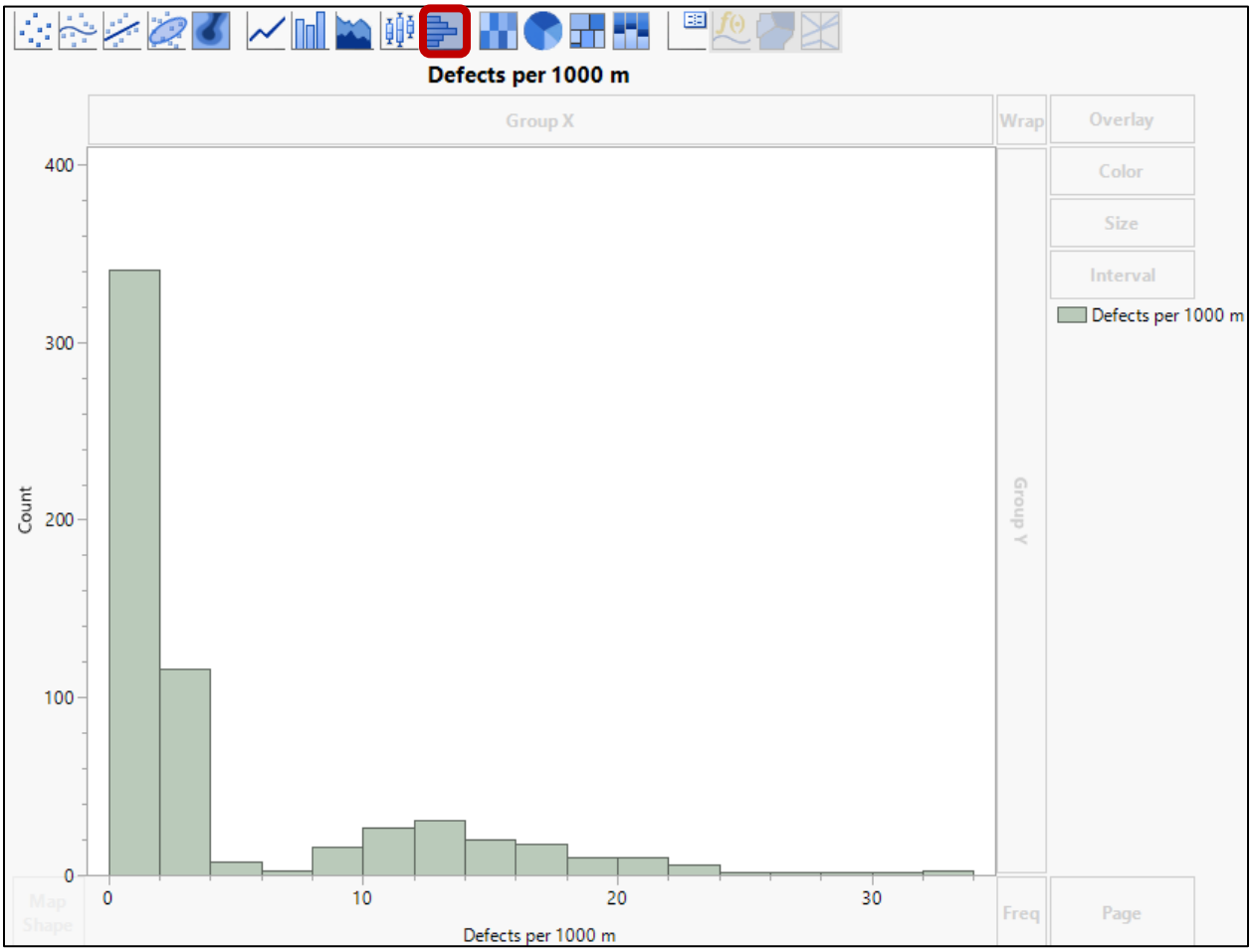
4. Drag **Defects per 1000 m** to the X drop zone, replacing **Film Type**.

5. Click the Points element icon.



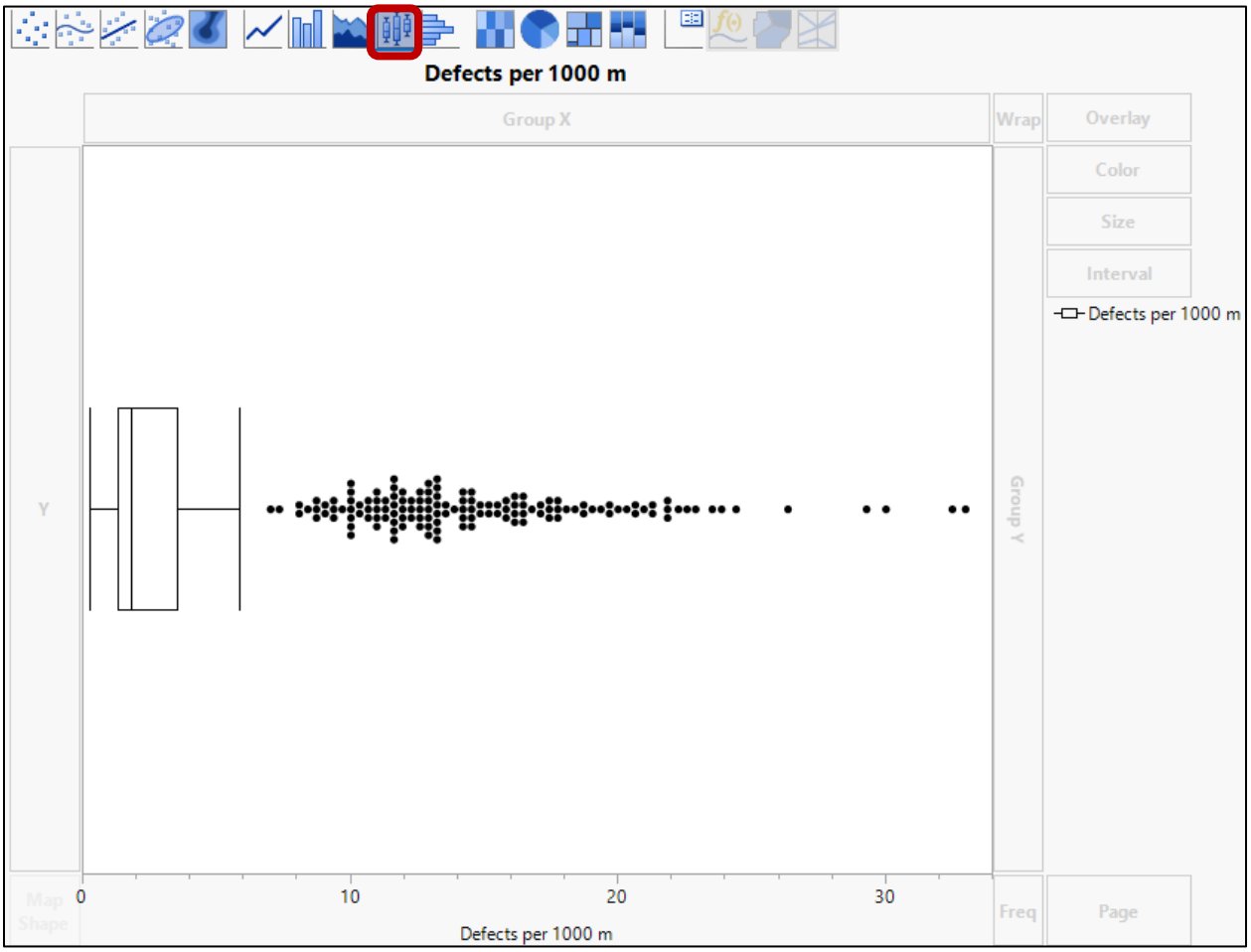
A dot plot is shown. There is one dot for each row in the data table. The dots are *jittered* so you can visualize how many observations there are for each value of **Defects per 1000 m**.

6. Click the Histogram element icon.



The histogram bins the data. You can control the bin width using the Grabber tool. (Select **Tools** > **Grabber**, then click and drag up or down on the bars.)

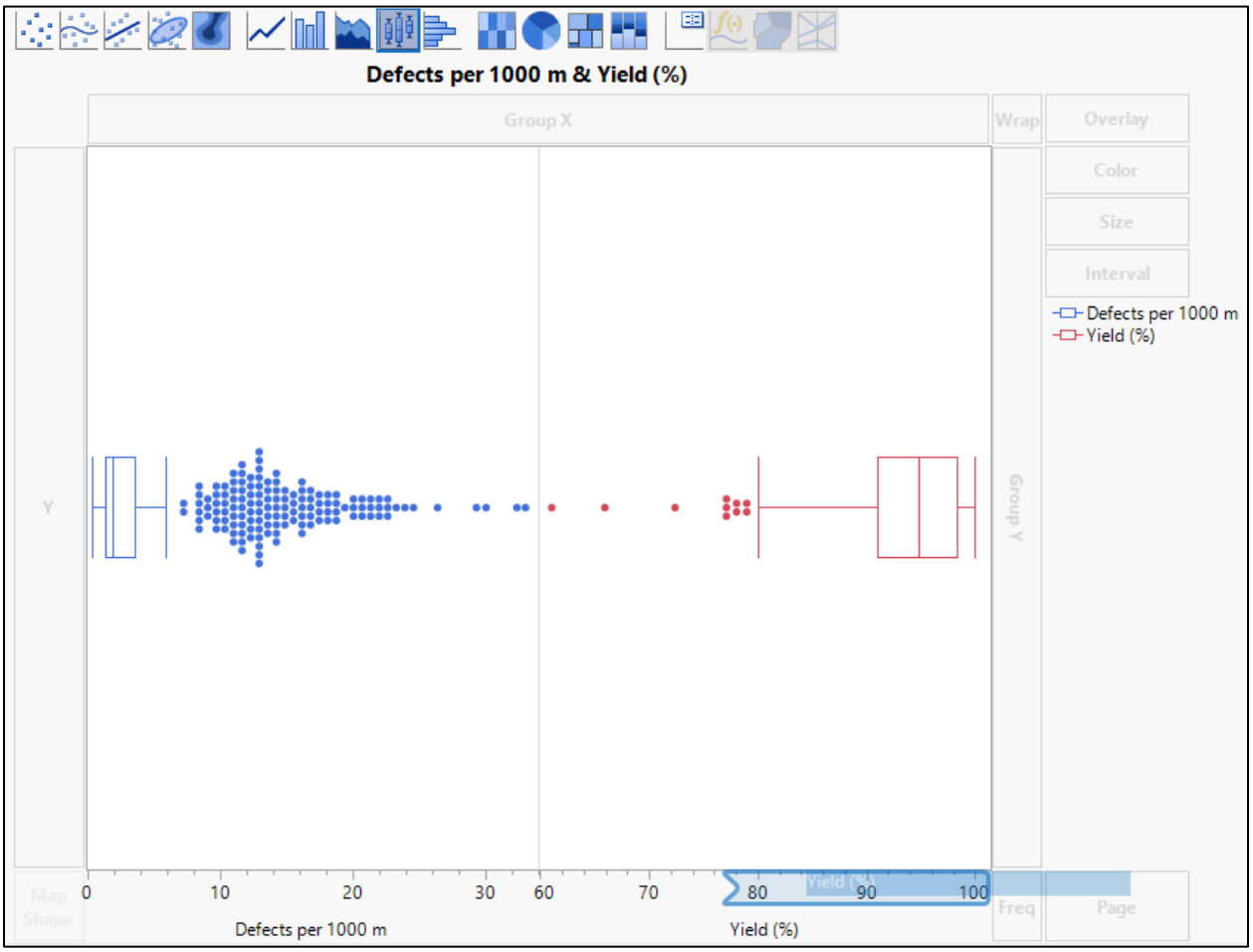
7. Click the boxplot element icon.



This outlier boxplot shows the five number summary of the data (min, 25th percentile, median (50th percentile), 75th percentile, and max) along with rules for detecting potential outliers.

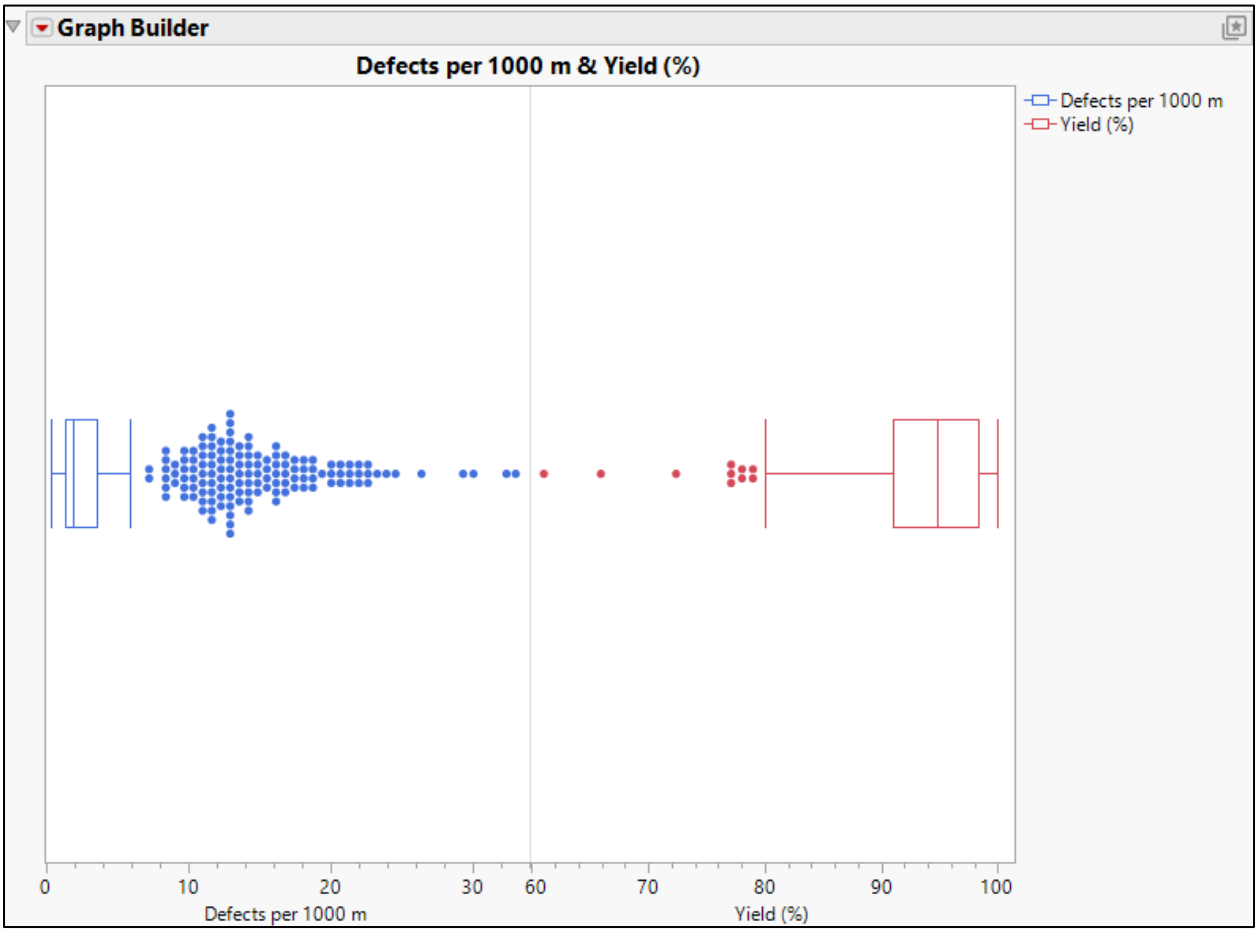
Next, examine the distribution of **Defects per 1000 m** and **Yield** at the same time.

8. Drag **Yield** to the right side of the X drop zone, so both variables are plotted.



9. Release the mouse button.

10. Click **Done**.



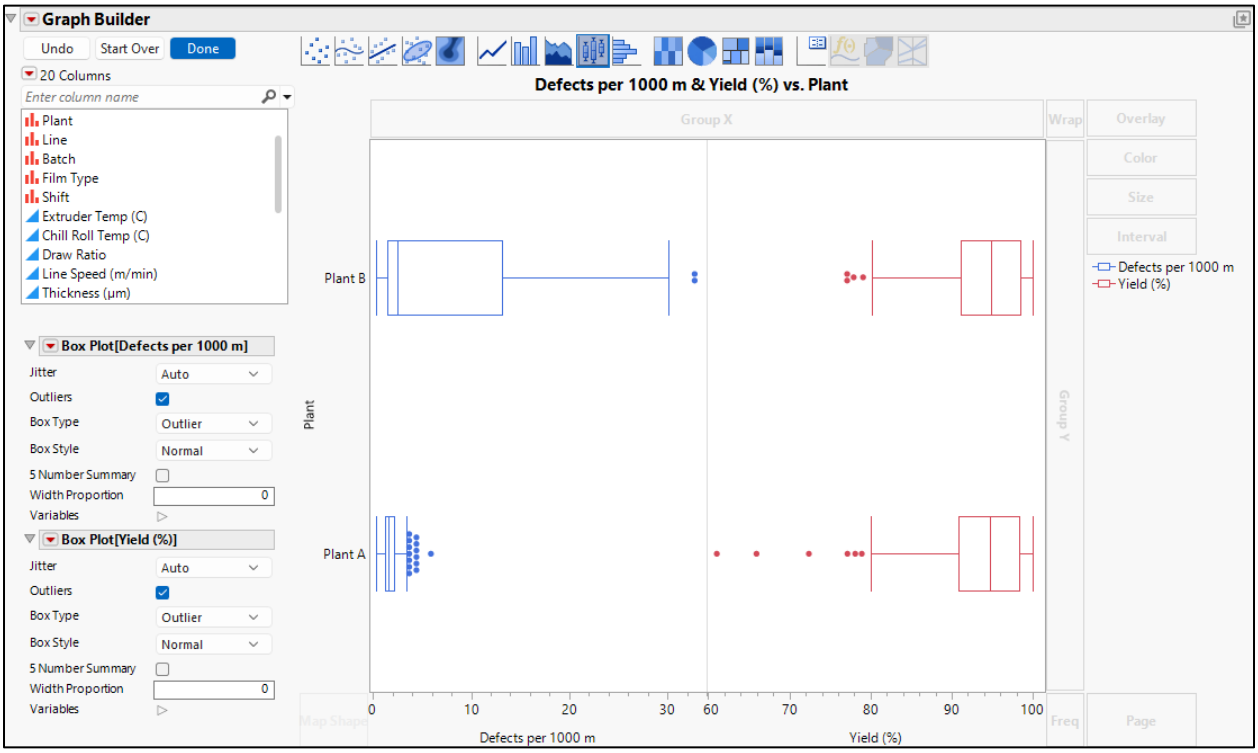
Now that you know how to make simple graphs for one variable, let's try to answer the research question.

Graphs for Two and More Variables

You can use Graph Builder to explore graphs that show the relationships between two variables. Again, the appropriate graph type depends on the modeling type of the variables. Let's look at a boxplot (continuous variable) for each level of a grouping variable (nominal or ordinal variable).

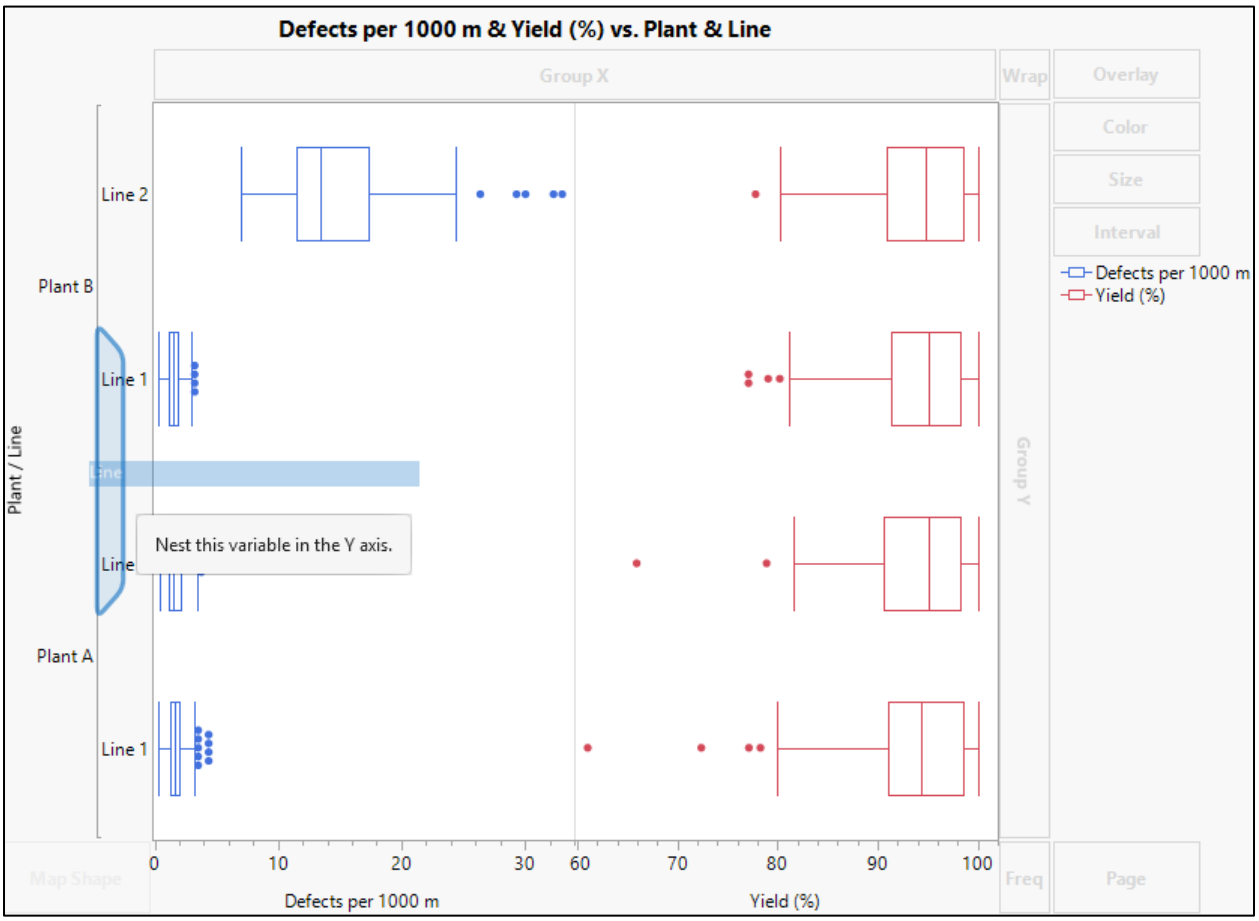
Graph Builder

1. Click the red triangle next to **Graph Builder** and select **Control Panel**.
2. Drag **Plant** to the Y drop zone.



The two plants have different distributions of **Defects per 1000 m**. The three batches with low yield are all in Plant A, otherwise the distributions of **Yield** are similar across the plants. What about the lines in each plant?

3. Drag **Line** to the Y drop zone just inside the Y axis.



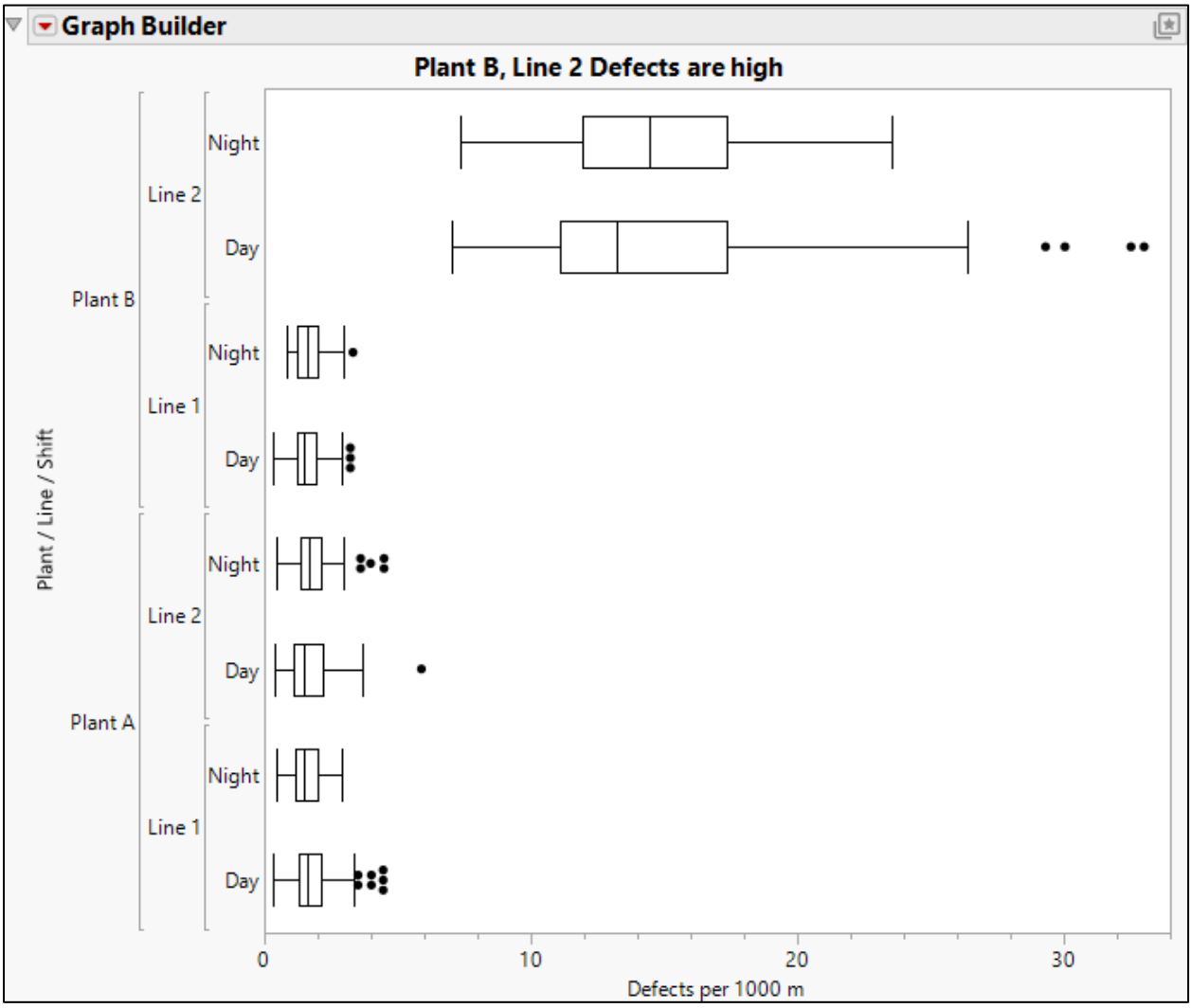
Line has been nested within **Plant**. There seems to be a defect problem for Plant B, Line 2. The distribution of yield seems to be the same. Remove **Yield** to focus on **Defects per 1000 m**, then add **Shift** to the nested structure in the Y drop zone.

4. Drag **Yield** away from the X drop zone to a neutral area.
5. Drag **Shift** to the Y drop zone just inside the Y axis.
6. Click **Done**.

   Remove the legend as the information it provides is already in the graph title and axis labels. Add an informative graph title.

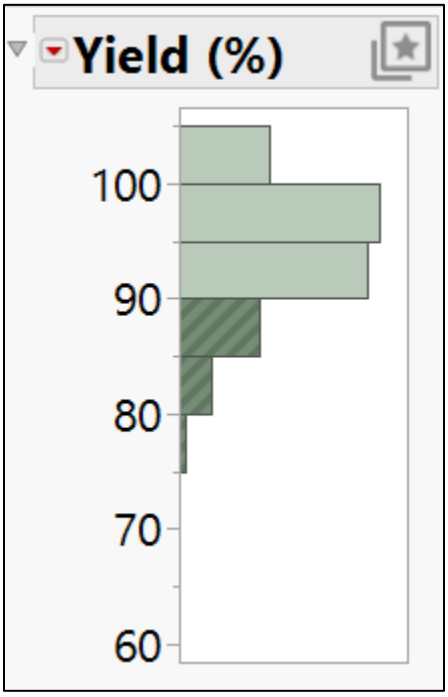
7. Click the red triangle next to **Graph Builder** and select **Show** > **Legend**.

8. Click the graph title and enter Plant B, Line 2 Defects are high.

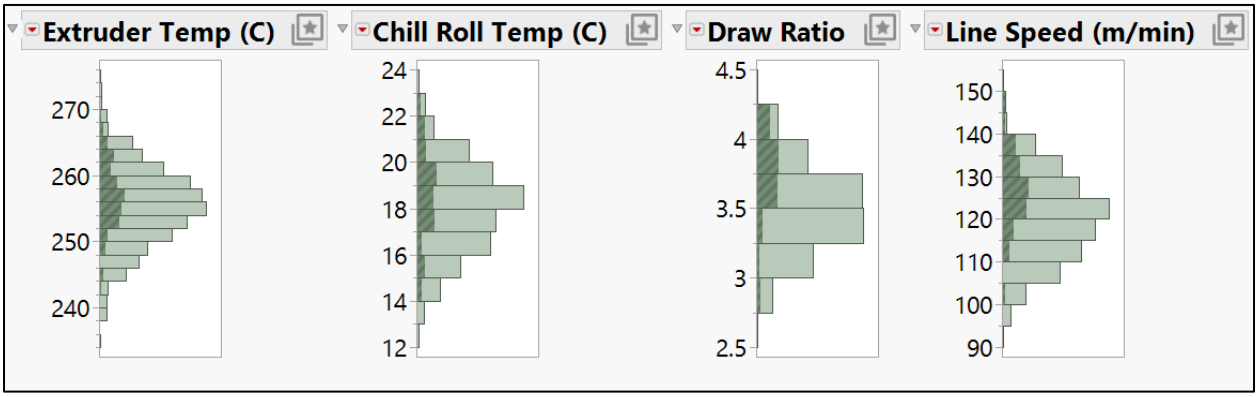


Next, investigate low-yielding batches.

9. Return to the **Distribution** report and select the bars for low yielding batches. Here we have selected those with yield below 90%.



10. Examine the other histograms to look for possible relationships.



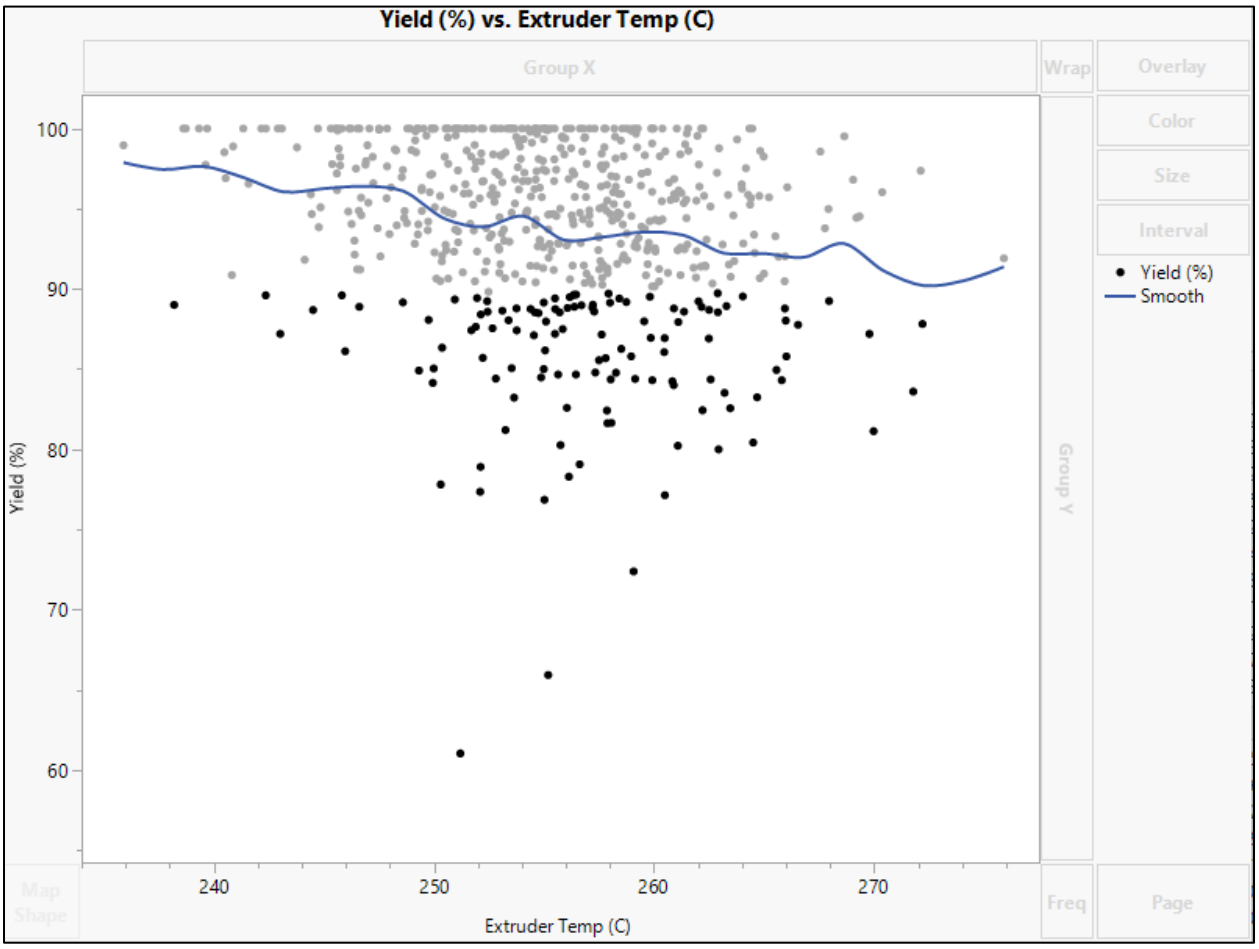
Perhaps there is a relationship between **Yield** and **Draw Ratio** and **Line Speed.** Examine with Graph Builder.

11. Select **Graph** > **Graph Builder**.

12. Drag **Yield** to the Y drop zone.

13. Drag **Extruder Temp** to the X drop zone.

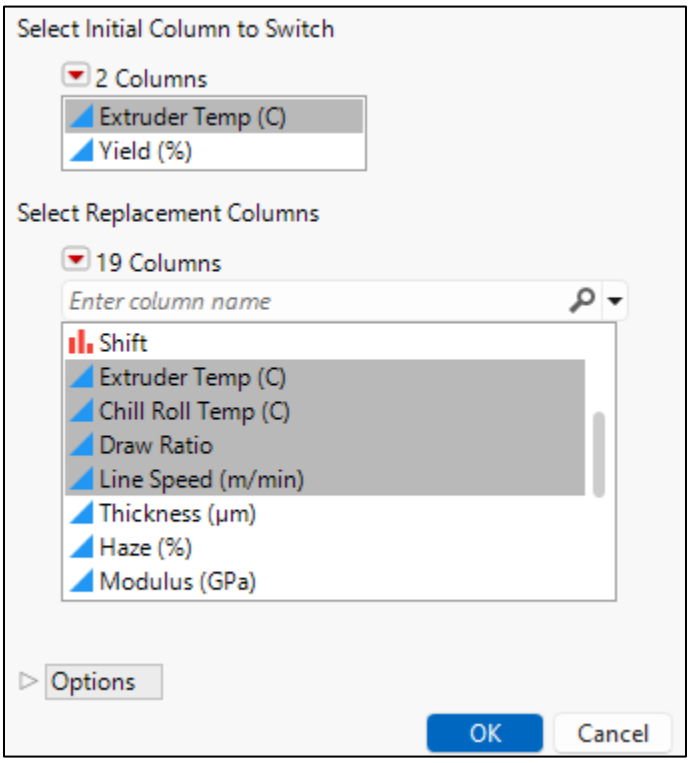
14. Click **Done**.



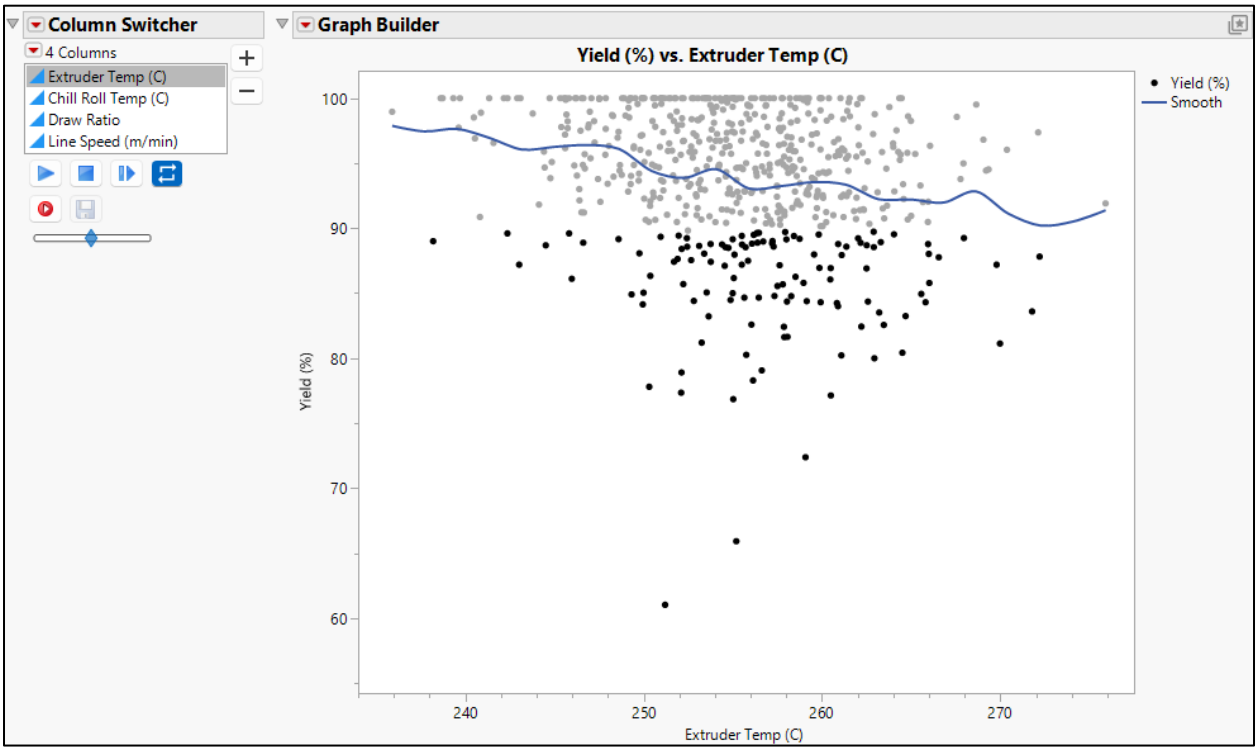
There is not a strong relationship.

15. Click the red triangle next to **Graph Builder** and select **Redo** > **Column Switcher**.
16. In the **Select Initial Column to Switch** box, ensure **Extruder Temp** is selected.

17. In the **Select Replacement Columns**, select **Extruder Temp** through **Line Speed**.

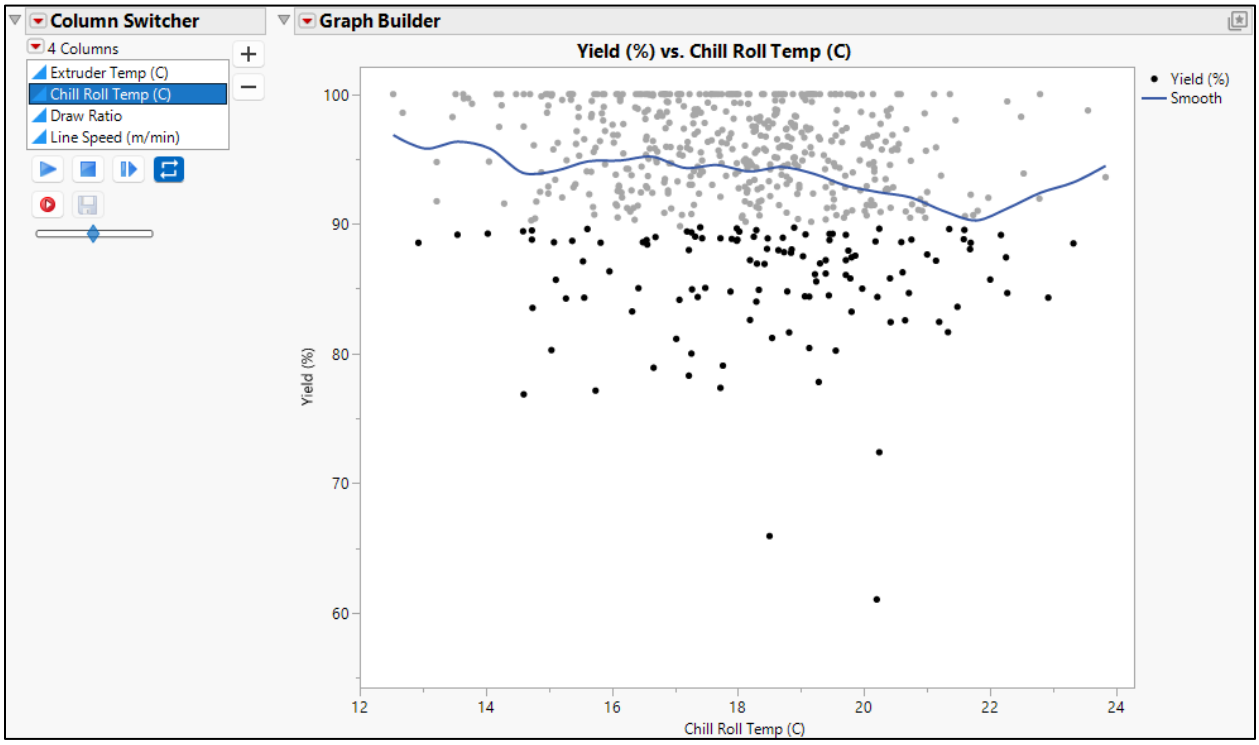


18. Click **OK**.

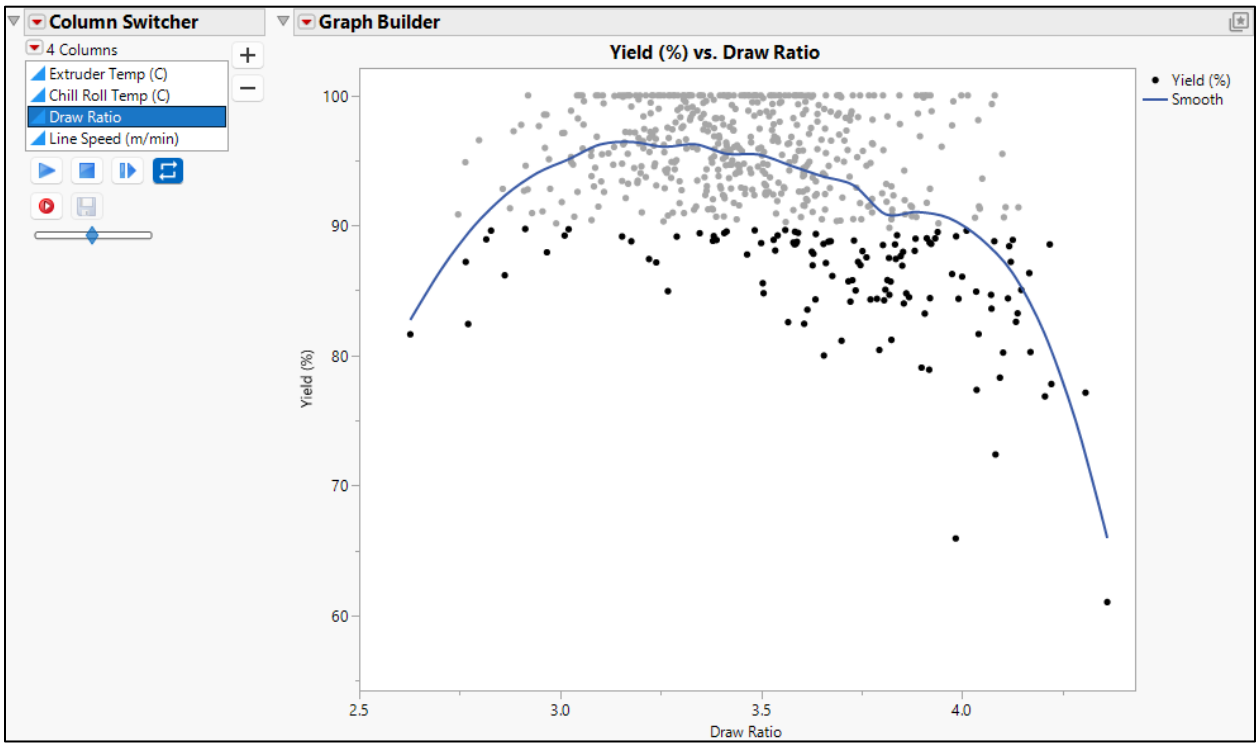


You can use the Column Switcher to quickly examine the other relationships.

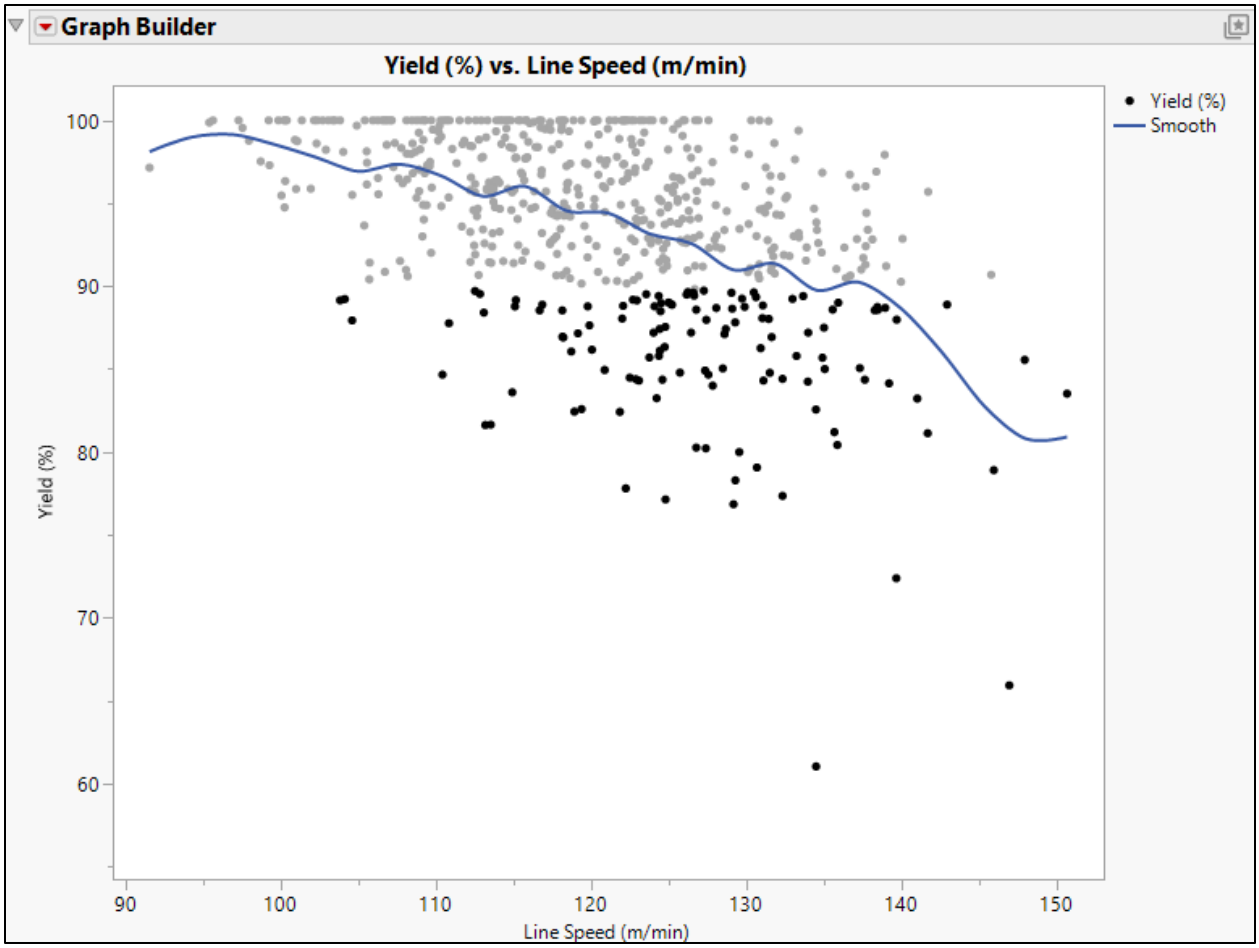
19. Click **Chill Roll Temp**, then **Draw Ratio**, then **Line Speed**.



There is not a strong relationship.

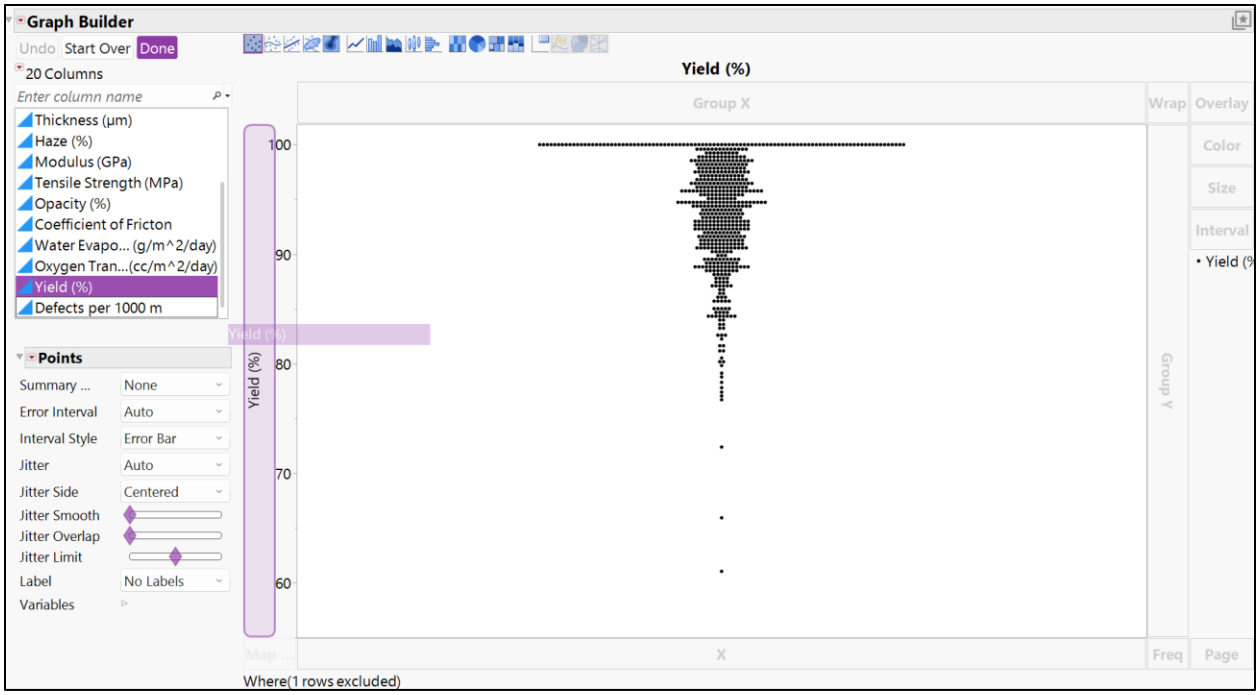


It looks like low or high draw ratio is related to low yield.

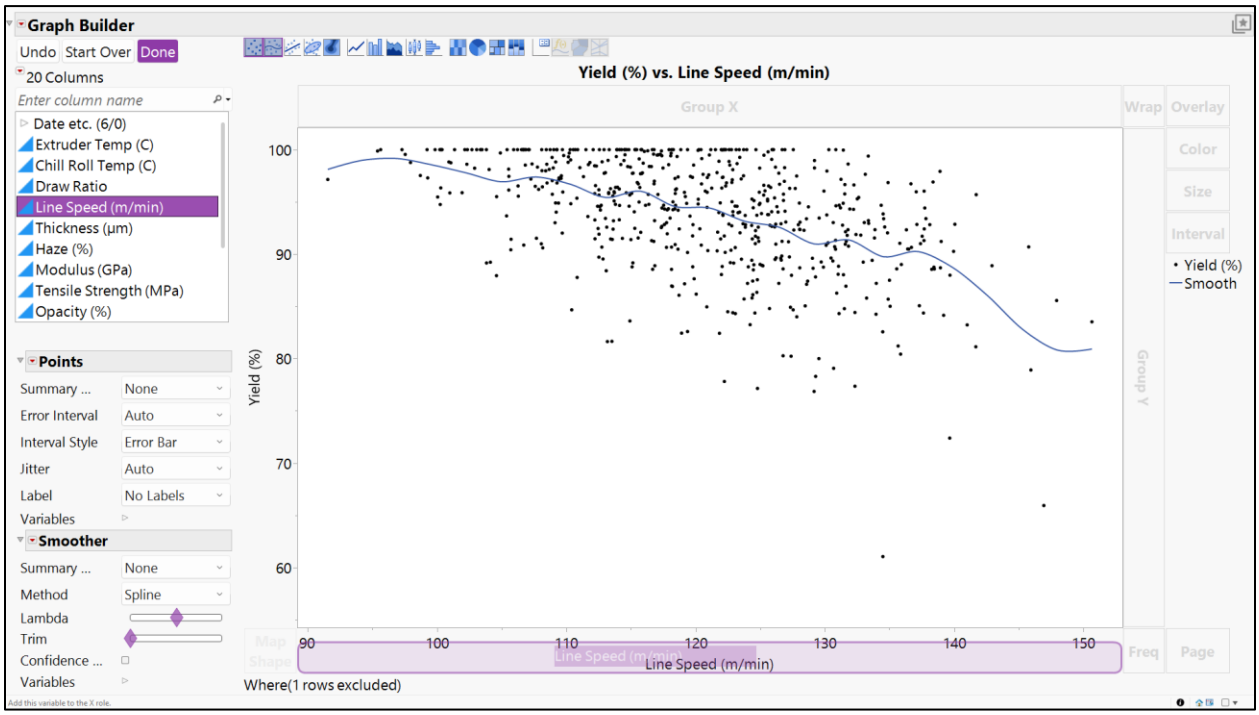


It seems as if high line speed is related to lower yields. Let's use the data interactivity to explore if there is a relationship between the line speed and the draw rate.

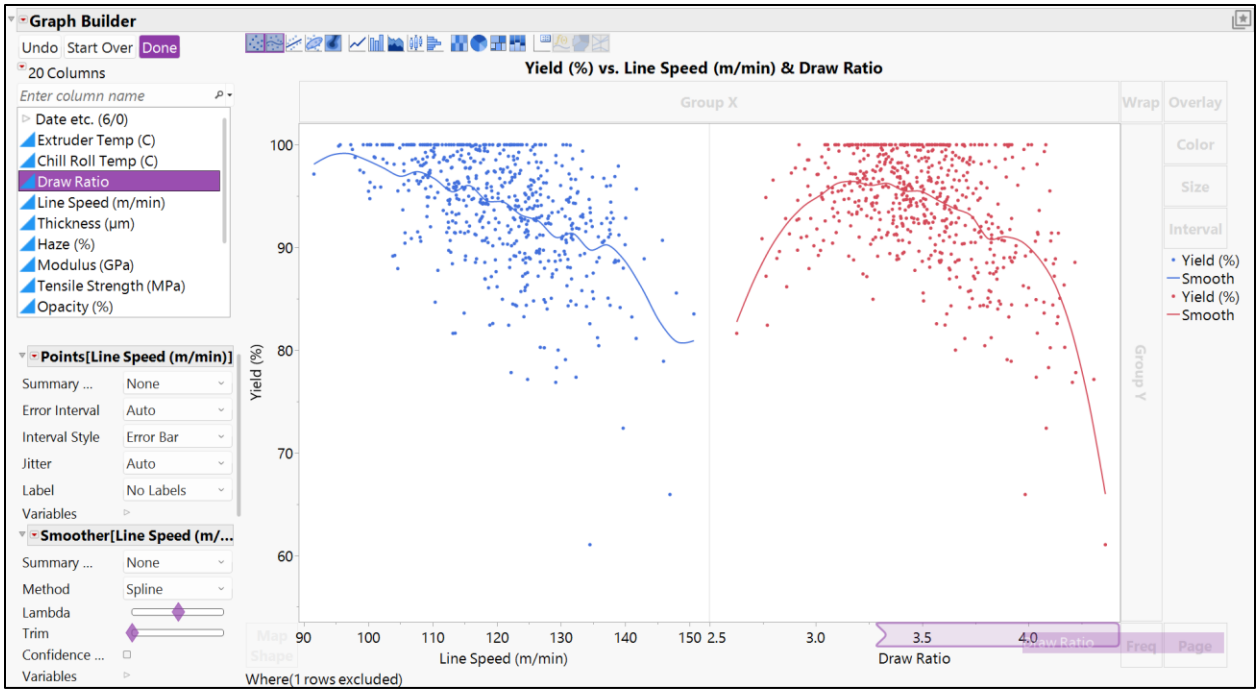
20. Press **Start Over** on the Graph Builder window.
21. Drag **Yield** to the Y axis.



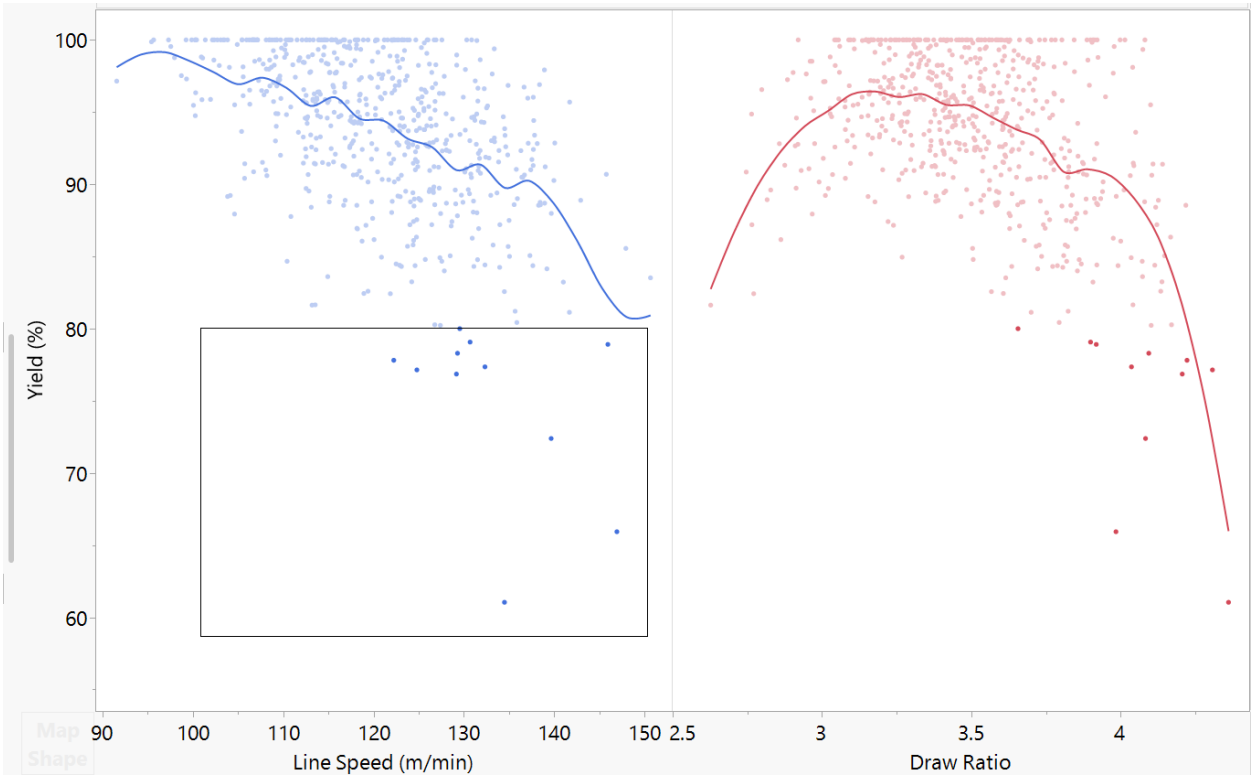
22. Drag **Line Speed** to the X axis.



23. Drag **Draw Rate** to the right side of the X drop zone, so both variables are plotted.



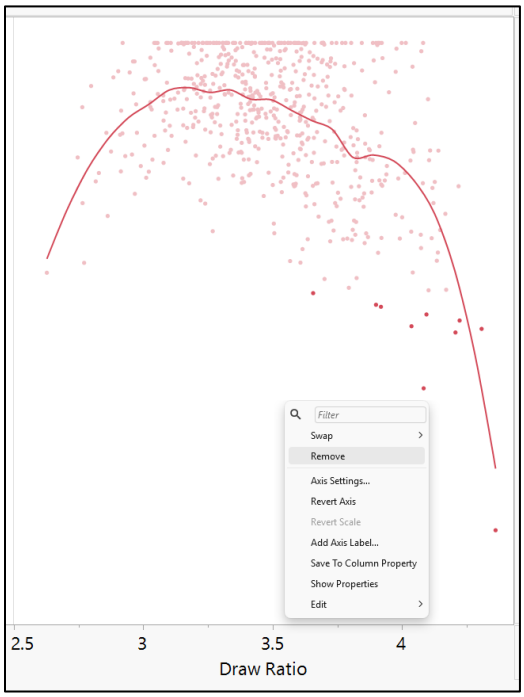
24. Drag and select the points on the Line Speed graph where Yield is less than 80.



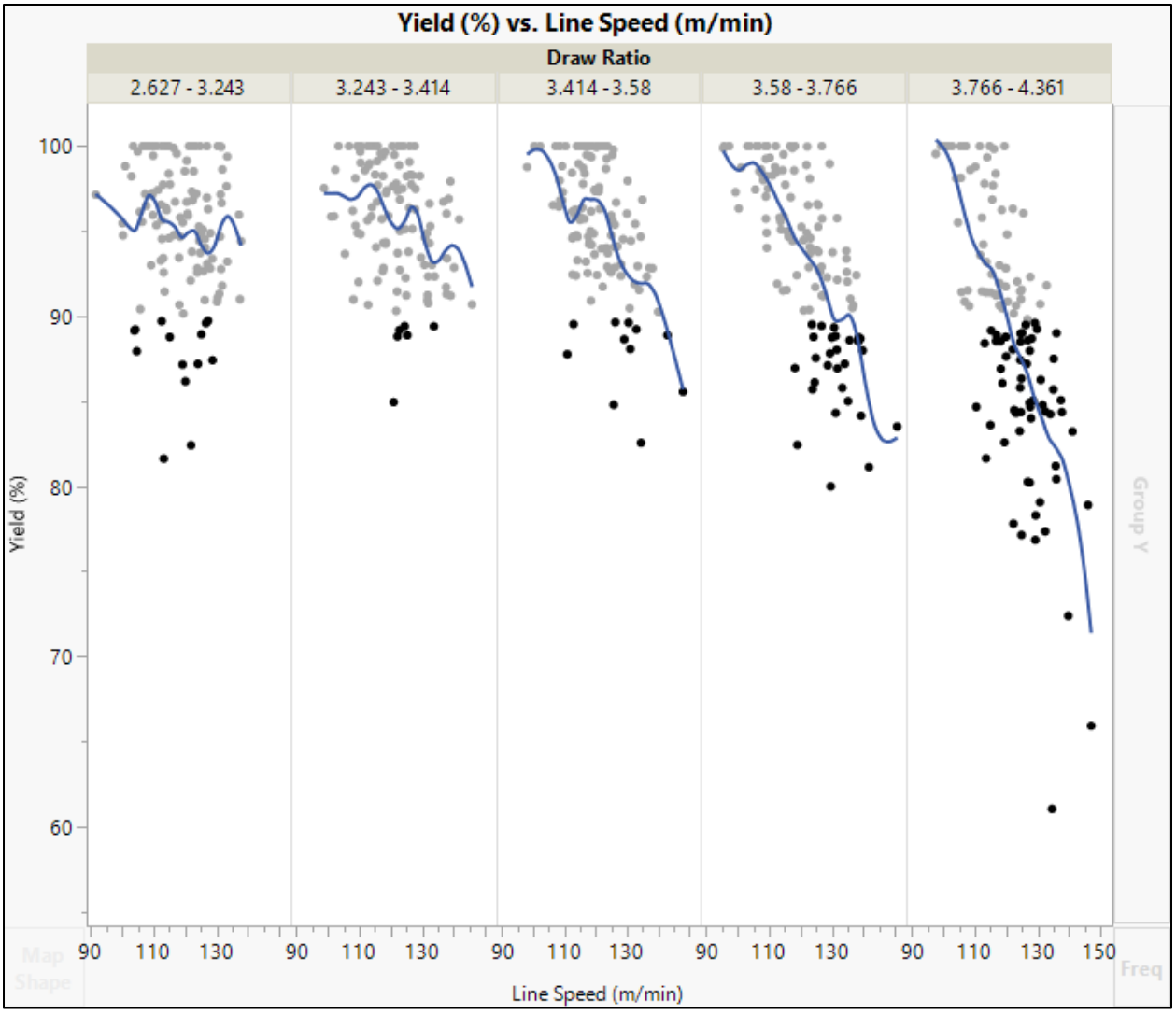
The points selected in the Line Speed are all high settings (over 120 m/min), causing yields below 80%. Looking at the Draw Ratio, the same points that use a Line Speed of

over 120 m/min also use a high draw ratio over ~3.5. This indicates that there might be an interaction between Line Speed and Draw Ratio leading to a decline in Yield.

25. Right Click the X axis of the Draw Ratio and select ‘Remove’.



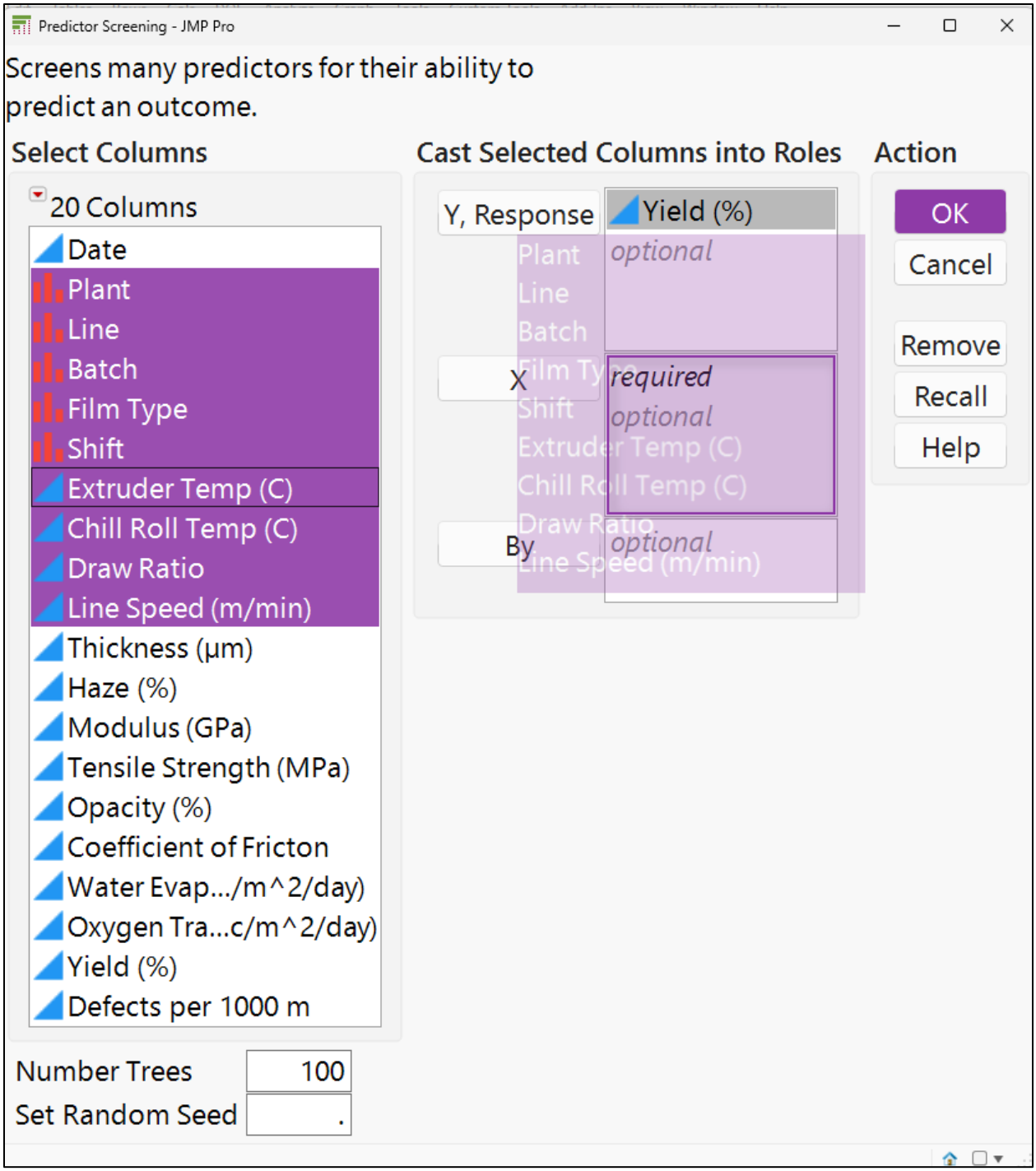
26. Drag **Draw Ratio** to the Group X drop zone.



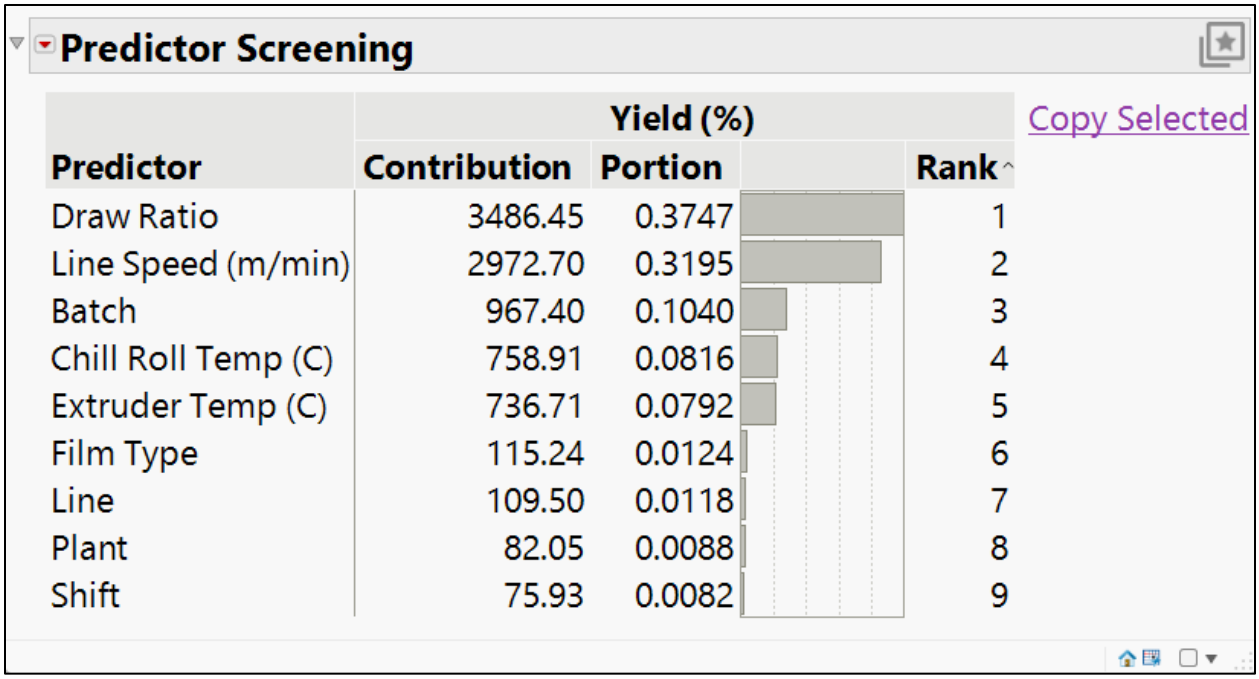
This graph shows the interaction effect of **Line Speed** and **Draw Ratio**. The effect of **Line Speed** changes with the value of **Draw Ratio**.

To quickly identify variables/predictors in a process, the Predictor Screening platform can quickly rank important variables.

27. Select **Analyze > Screening > Predictor Screening**



28. Drag **Yield** into **Y, Response**
29. Drag from **Plant to Line Speed** into **X**
30. Press **OK.**



Predictor Screening

Copy Selected

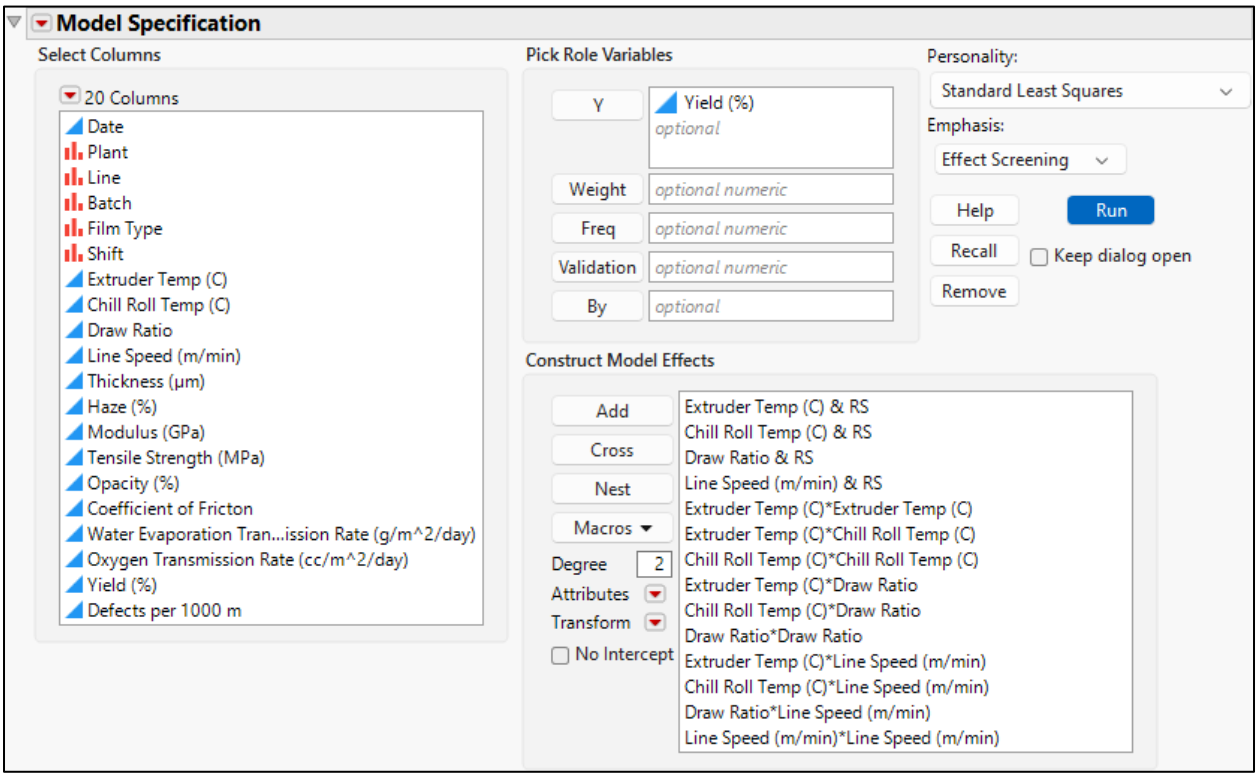
Predictor	Yield (%) Contribution	Portion	Rank
Draw Ratio	3486.45	0.3747	1
Line Speed (m/min)	2972.70	0.3195	2
Batch	967.40	0.1040	3
Chill Roll Temp (C)	758.91	0.0816	4
Extruder Temp (C)	736.71	0.0792	5
Film Type	115.24	0.0124	6
Line	109.50	0.0118	7
Plant	82.05	0.0088	8
Shift	75.93	0.0082	9

Looking at the ranking of the effect of the predictors on the Yield, Draw Ratio and Line Speed account for a big amount of the change in Yield, with a portion of 0.375 and 0.319 respectively.

JMP has a wide array of analytical capabilities as well as strong graphing capabilities. Let's fit a linear model to the yield as a function of the four continuous process variables. JMP's Prediction Profiler is a powerful interactive graph to visualize a statistical model.

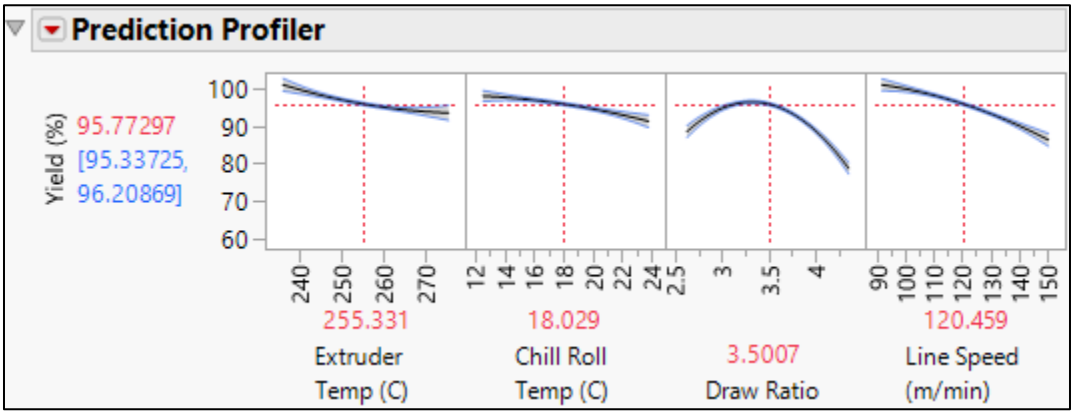
31. Select **Analyze** > **Fit Model**.
32. Select **Yield**, then click **Y, Response**.

33. Select **Extruder Temp** through **Line Speed**, then click **Macros** > **Response Surface**.



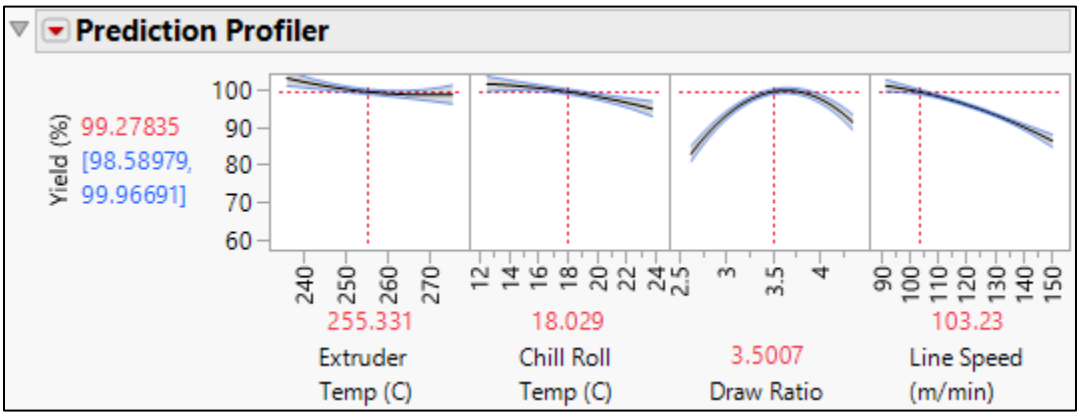
34. Click **Run**.

35. Scroll to the bottom of the report to view the **Prediction Profiler**.



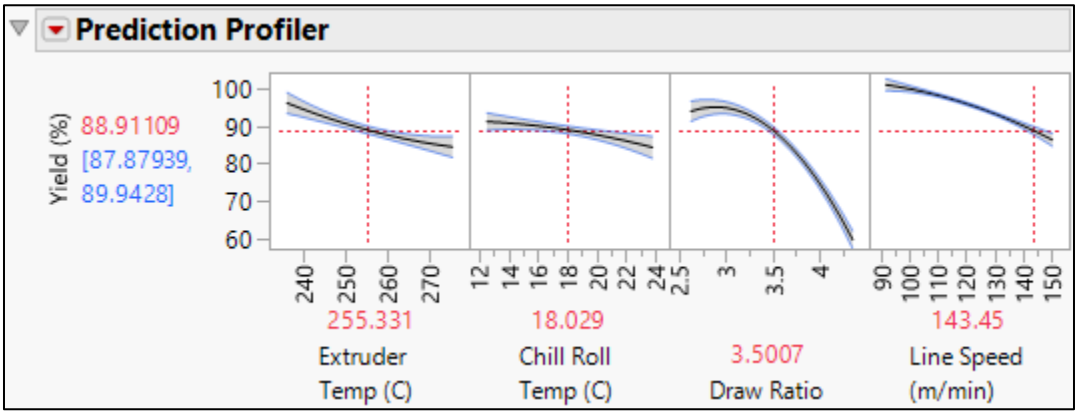
The profiler shows cross sections of the fitted response surface. There is one row for each response variable, here just one. There is a panel for each factor in the model. The horizontal red dashed line is the predicted value of the response at the current settings of the factors, shown by the vertical dashed lines. You can drag the vertical dashed lines to see how the predictions change with values of the factors. You can also view the interaction between **Draw Ratio** and **Line Speed**.

36. Drag the current value of **Line Speed** to a lower level.



At a low line speed, higher draw ratio is better (higher yield).

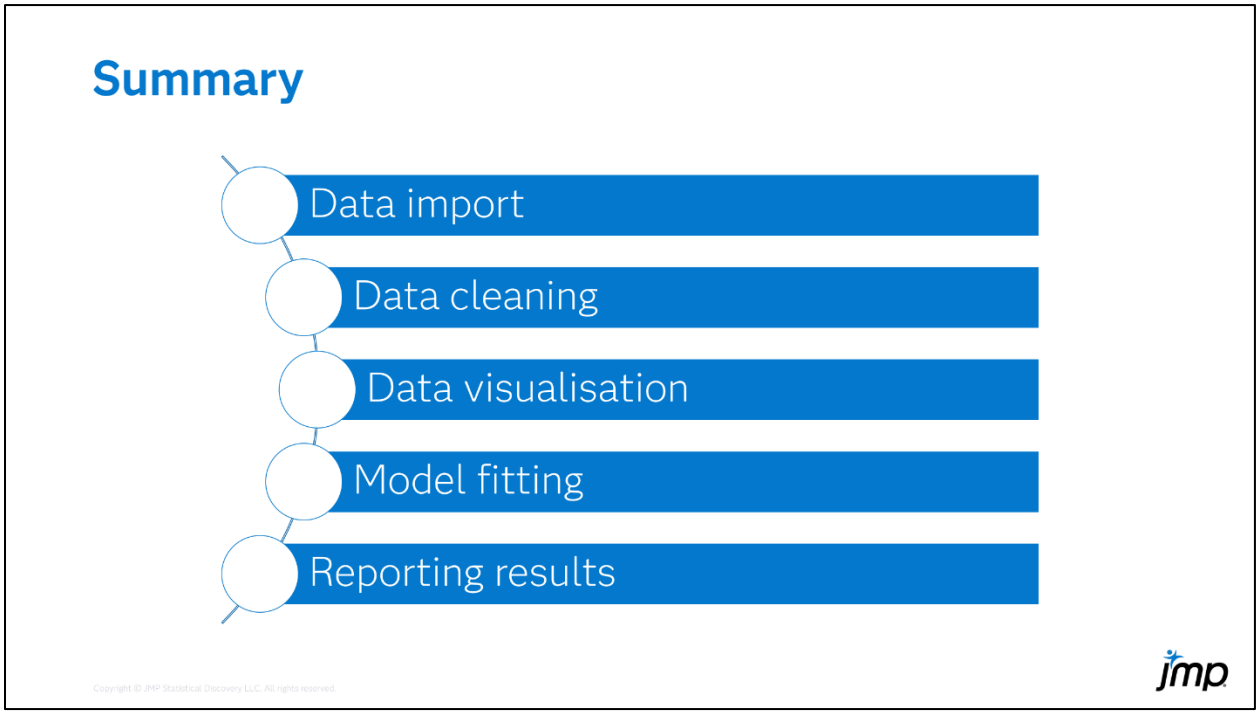
37. Drag the current value of **Line Speed** to a higher level.



At a high line speed, lower draw ratio is better.

Note that there are some combinations of the factors that result in a predicted yield above 100%. This is a result of the model assumption of normality of the residual values. There are statistical models that do not use a normal distribution, but they are beyond the scope of this workshop.

Summary



Now that you know the basic principles of graphing data, you can use JMP to import data from Excel and other data sources, clean the data, and visualize the information you have gathered. JMP also offers powerful platforms for statistical modeling and sharing results with both JMP users and those who don't have a JMP license.