



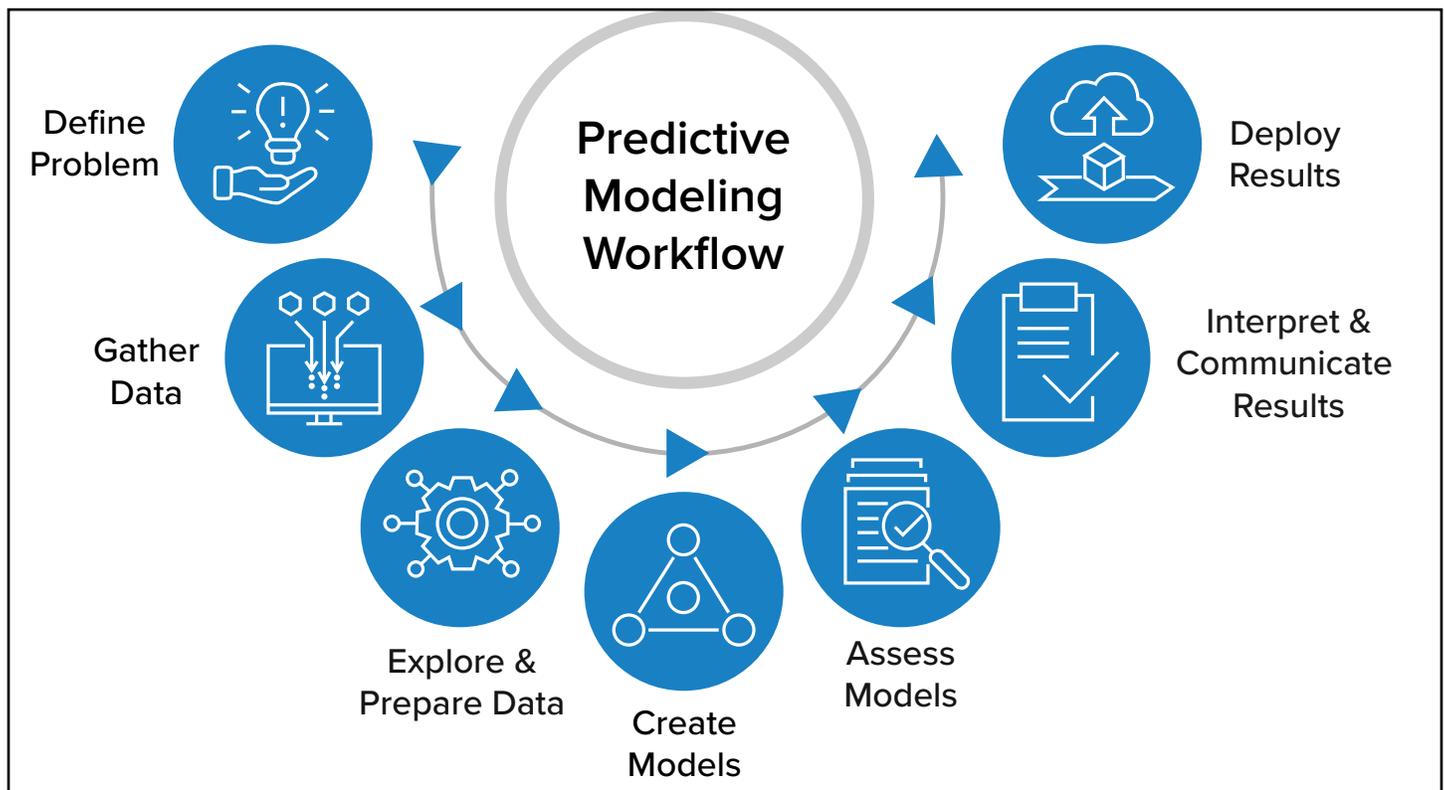
## Feature Engineering: 7 Techniques for Better Predictive Analytics & Machine Learning Models

# Feature Engineering: 7 Techniques for Better Predictive Analytics & Machine Learning Models

For years, feature engineering has been a popular concept in the world of data analytics due to its critical role in supporting successful machine learning and predictive analytics.

Feature engineering is the task of using knowledge about a process and its resulting data to extract properties, or features, that make predictive models work. Essentially, predictive models require input data to make predictions about the output of a process. The features or inputs you select to build a predictive model will have a major impact on the accuracy of the model’s predictions. To best apply feature engineering, knowledge of your process as well as your data are key to selecting features that will help you build accurate predictive models. The features you select should represent the underlying problem that you are trying to solve, making expertise in the process itself critical. At Minitab, we help practitioners like process experts, data scientists and business analysts leverage process knowledge to find data-driven solutions to solve their toughest business challenges.

In this whitepaper, we highlight seven feature engineering techniques and how you can successfully apply them using **Minitab Statistical Software**.



## Seven Feature Engineering Techniques

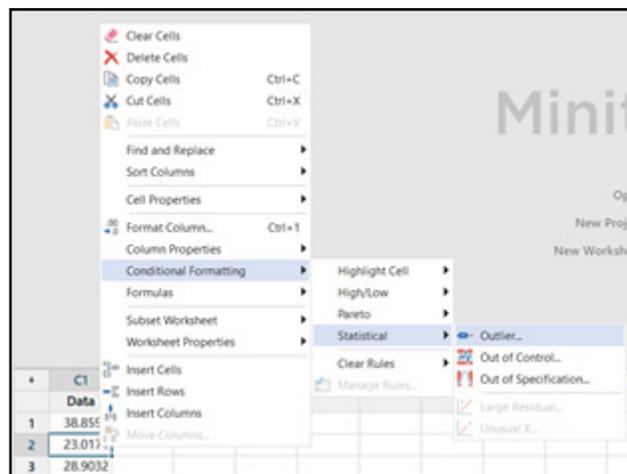
### 1. Handling Outliers

The presence of extreme outliers can have a significant impact on a model's performance. Outliers are generally defined as unusually high or low values, relative to the entire dataset. There are various approaches to identifying outliers and two common options are using the Interquartile Range, or IQR on a boxplot, or multiples of a simple standard deviation that is estimated from the data.

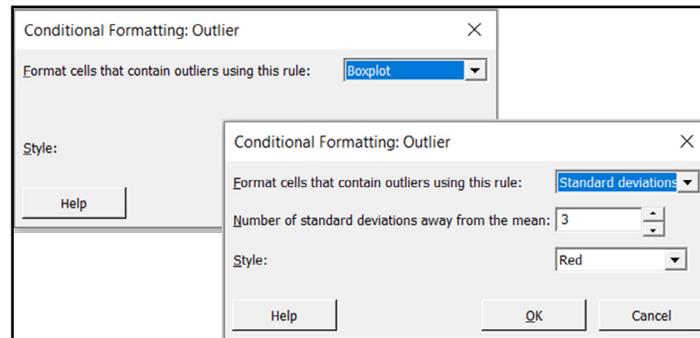
The boxplot method is an appealing option due to its simplicity and the fact that outliers can easily be visualized on a boxplot graph. Identifying outliers using multiples of the standard deviation is based on the concept that under a normal distribution, approximately 95% of the data falls within two standard deviations of the mean, and more than 99% of the data will fall within three standard deviations of the mean.

Minitab Statistical Software makes it easy to identify outliers using boxplot rules or multiples of the standard deviation.

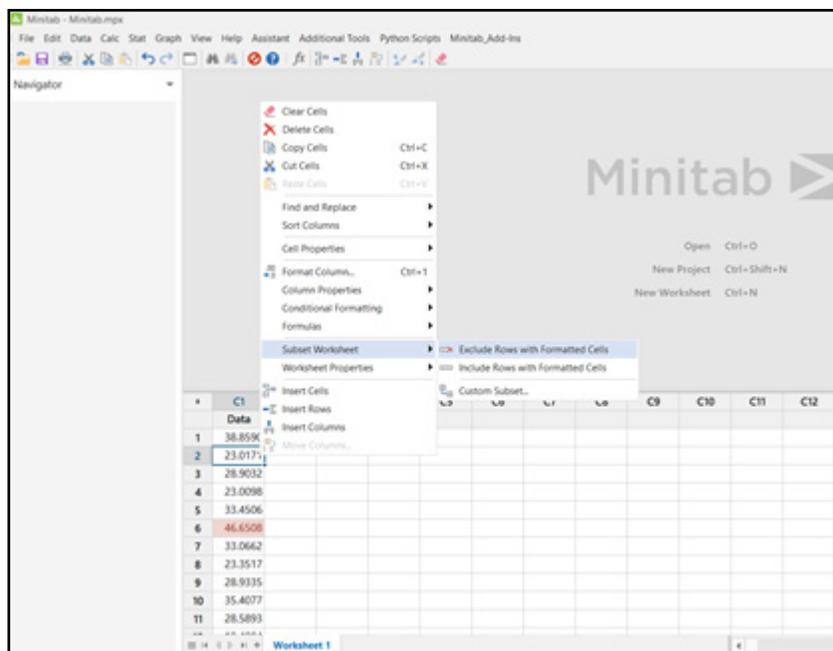
Simply right click on a data column in the worksheet and select **Conditional Formatting > Statistical > Outlier**.



In the new window, use the drop-down list to select a method of identifying outliers. If using the standard deviations option, the number of standard deviations specified in this window will be used to identify potential outliers by highlighting them in the worksheet.



Once the outliers are highlighted based on the rules you specified, right-click and choose **Subset Worksheet > Exclude Rows with Formatted Cells** to create a new worksheet that excludes the outliers.



To clarify, the recommendation is not to discard potential outliers. Outliers should be investigated. Try to identify the cause of any outliers and correct any data entry or measurement errors. Consider removing data values that are associated with abnormal, one-time events, otherwise known as special causes.

## 2. Binning or Bucketing

Binning or bucketing is the process of grouping individual data values into a smaller number of bins or groups. This technique is commonly applied to categorical data when some of the values occur very infrequently relative to other values in the dataset. The main purpose of binning is to reduce overfitting. An overfit model is overly adjusted to fit the peculiarities in the sample of data that is used to build the model, but overfit models do not generalize well to new data, making their predictions on future data less accurate. Depending on the type of model, binning may be required for an algorithm to converge.

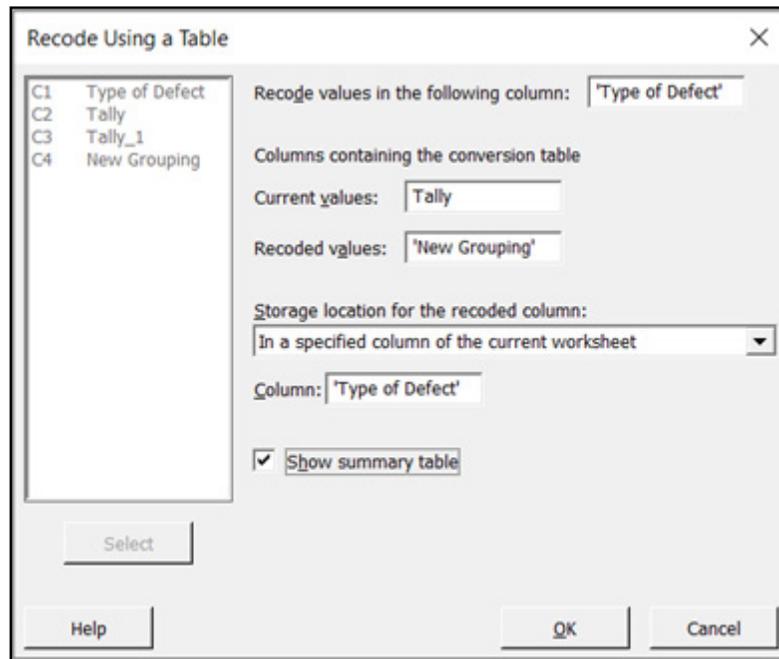
In Minitab, a conversion table offers a convenient way to recode categories. First, we can use **Stat > Tables > Tally Individual Variables** and choose the option to store results in a worksheet. This option will store a list of each unique value in the column vector you provide, along with the frequency counts for each value, in the active worksheet.

In the example below, Tally shows each unique value, and Tally\_1 shows the counts.

	C1-T	C2-T	C3	C4-T	C5	C6
	Type of Defect	Tally	Tally_1	New Grouping		
1	Smear	Crack	2	Other		
2	Discoloration	Dent	3	Other		
3	Discoloration	Mislabeled	3	Other		
4	Scratch	Deformity	4	Other		
5	Discoloration	Torn	4	Other		
6	Smear	Smear	111	Smear		
7	Smear	Discoloration	126	Discoloration		
8	Discoloration	Scratch	132	Scratch		
9	Scratch					

The values: *crack*, *dent*, *mislabeled*, *deformity* and *torn* appear only a few times in this dataset, relative to the other types of defects. To combine the less frequent values into a single *other* category, the new groupings can be defined in the worksheet as shown in C4-T above. The information in the worksheet can now be used as a conversion table to recode the infrequent values in the original column.

In Minitab, that option is available by choosing **Data > Recode > Use Conversion Table**:



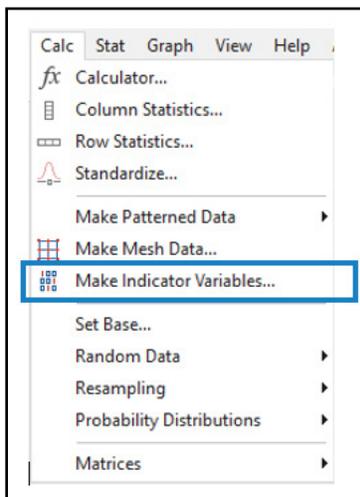
The approach described above could also be used to reduce the number of levels in a categorical grouping variable. For example, if the data contains granular information about the US states and a regional indicator would provide more meaningful information, a conversion table can be used to redefine individual states into north, south, east and west designations instead.

Minitab also offers an easy way to bin numeric data in this menu: **Data > Recode > To Numeric**, then use the drop-down list to select **Recode ranges of values**. This option allows you to define the end points of each new bin.

### 3. Encoding Categorical Predictors

Encoding is a technique applied to categorical predictors to create numeric representations of each unique text value in a column. Converting text-based predictors to numbers allows algorithms to understand and process the information to build predictive models. Even though Minitab's supervised learning methods such as regression and CART can handle categorical predictors automatically, it may still be desirable to encode categorical predictors to use in other analysis.

In Minitab, the feature that easily converts categorical predictors into numeric representations is called **Make Indicator Variables** and the option is available in the Calc menu.



This option can be used to create columns of 0s and 1s from a categorical variable. For each value in a categorical grouping, an indicator variable has a value of 0 if the value is not represented in that row or observation, and 1 if the category is represented in that row.

↓	C1-T	C2	C3	C4
	Category	Category_Maybe	Category_No	Category_Yes
1	Yes	0	0	1
2	No	0	1	0
3	Maybe	1	0	0
4	Maybe	1	0	0
5	No	0	1	0
6	Yes	0	0	1
7				

In the above illustration, the original categorical variable is in C1, which has three unique values: *yes*, *no* and *maybe*, and C2, C3 and C4 are the indicators generated using Minitab's point-and-click interface.

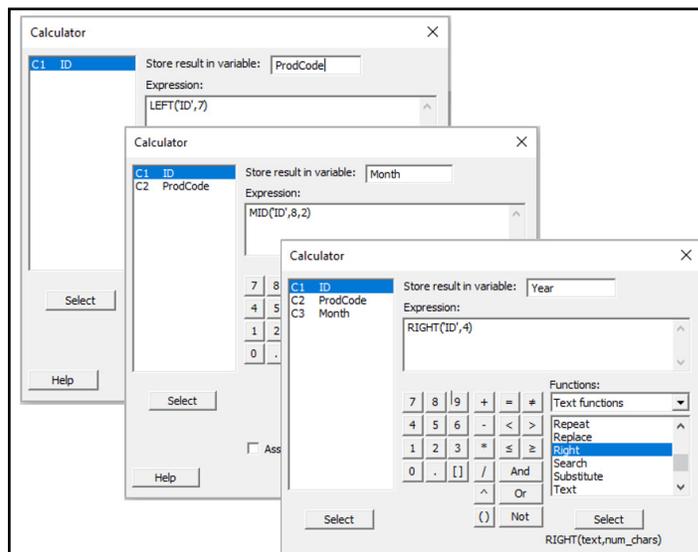
## 4. Feature Splitting

Splitting a feature or a text string can be a useful way to extract relevant information to use in model building. Oftentimes when data is imported from a database, several pieces of information may be merged into a single text string within a column. As an example, consider the ID column below, where the first seven characters represent a product code, and the last eight characters represent the day, month and year.

↓	C1-T
	ID
1	38BEE2712292020
2	38BED2712302020
3	38BEG2712302020
4	38BET2712312020
5	39BEE2812312020

Introducing the above ID column as a categorical predictor in a model would result in each unique value in the column being treated as a separate group, which would not be helpful. A more logical approach would be to separate the product code, the month, the year and introduce each variable separately as a predictor to assess their effects on the process output.

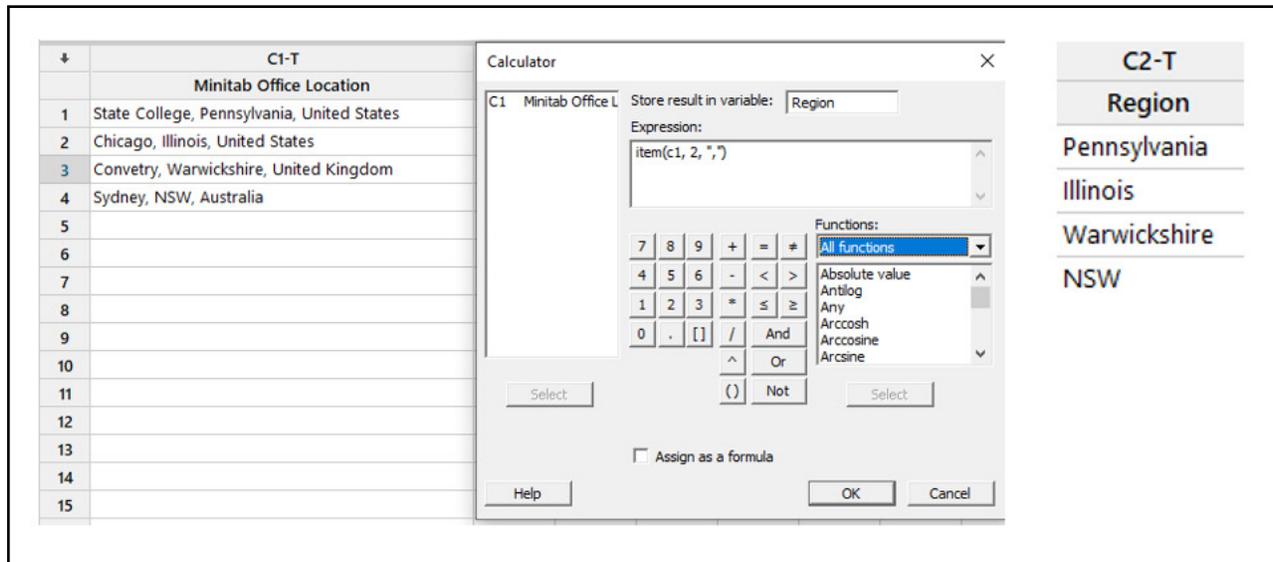
Minitab’s calculator, which is available by selecting **Calc > Calculator**, includes functions to extract values from text strings. Specifically, the [LEFT](#), [RIGHT](#), and [MID](#) functions are useful in specifying the characters to extract. In the example below, these functions were used to extract the product code, which is based on the first seven characters from the left, the month, represented as two digits in the middle of the text string starting with the 8th character, and the year, which is represented by the last four digits.



The Minitab calculator functions list in the above window provides an example of the syntax for applying each function. A more detailed explanation of each function can be found [here](#). For example, to extract the four rightmost characters, the RIGHT function is used. Within the parenthesis, we add the column with the original text string, a comma and the number of characters to extract beginning from the right side. Minitab stores the extracted values in a new column.

↓	C1-T	C2-T	C3-T	C4-T
	ID	ProdCode	Month	Year
1	38BEE2712292020	38BEE27	12	2020
2	38BED2712302020	38BED27	12	2020
3	38BEG2712302020	38BEG27	12	2020
4	38BET2712312020	38BET27	12	2020
5	39BEE2812312020	39BEE28	12	2020

To extract values from a text string based on their position between the delimiters that you specify, the [ITEM](#) and [WORD](#) functions in the calculator can be extremely useful. In the example below, the calculator expression extracts the region, which is in the second position between comma delimiters.

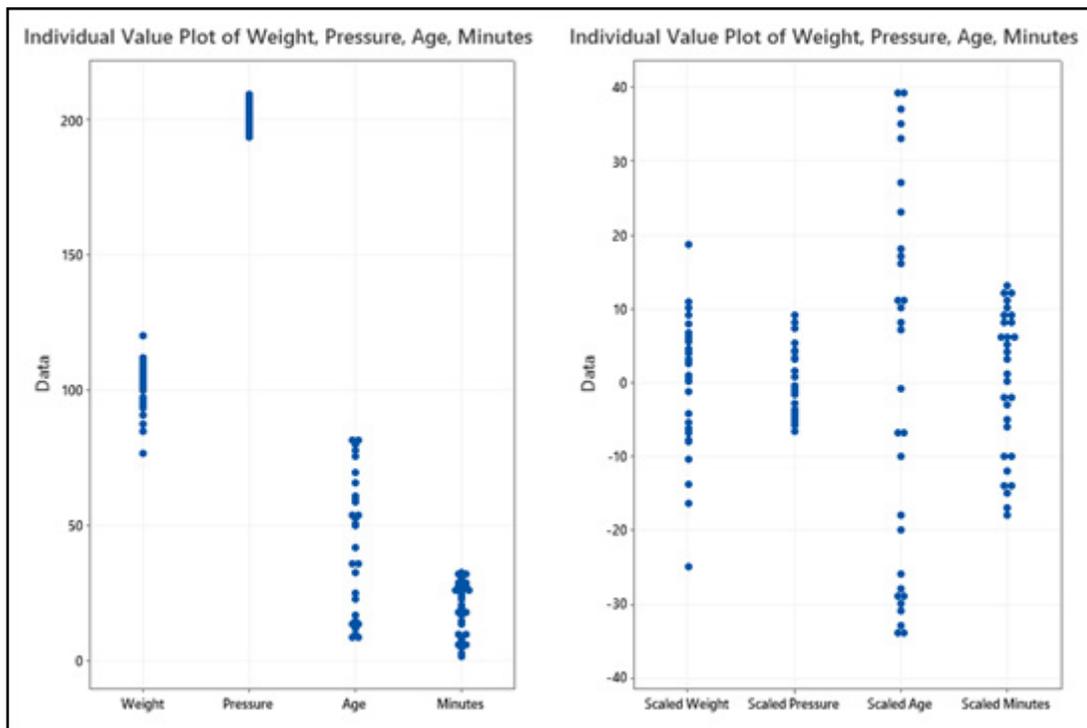


The screenshot shows a Minitab workspace with three columns: C1-T (Minitab Office Location), C2-T (Region), and C3-T (Region). The C1-T column contains text strings like "State College, Pennsylvania, United States". The C2-T column contains the extracted region names: "Pennsylvania", "Illinois", "Warwickshire", and "NSW". A calculator window is open, showing the expression `item(c1, 2, ',')` and the result stored in the variable "Region". The calculator window also shows a list of functions and a numeric keypad.

## 5. Scaling

Scaling or standardizing predictors refers to changing the original data values so that all predictors are on the same scale. Most of the time, inputs to a process are on different scales. For example, weight, age, pressure and time in minutes. Scaling can be helpful in making comparisons, ensuring one predictor doesn't appear more important than others merely because of the magnitude of its scale and scaling can help address [multicollinearity](#).

As an example, a comparison of the four variables on the left side below, which are on different scales, is not very informative in terms of visualizing differences in variation. A comparison of the scaled variables on the right side makes visualizing differences in variation much easier.



While Minitab's linear and logistic regression models include options to [standardize continuous predictors](#) programmatically as part of the analysis, there are also built-in features to control the scaling directly on the data in the worksheet.

In Minitab, this function is called standardization and is available by choosing **Calc > Standardize**. This menu provides several common scaling options like subtracting the mean or subtracting the mean and dividing by the standard deviation.

The screenshot shows the Minitab Standardize dialog box. The 'Input column(s):' field contains 'Weight Pressure Age Minutes'. The 'Store results in:' field contains 'c6 c7 c8 c9'. The 'Subtract mean' option is selected. The 'First:' field is set to 0.0 and the 'Second:' field is set to 1.0. The 'Make range from start to end' option is also selected, with 'Start:' set to -1.0 and 'End:' set to 1.0. Below the dialog box is a table showing the original and scaled data for four rows.

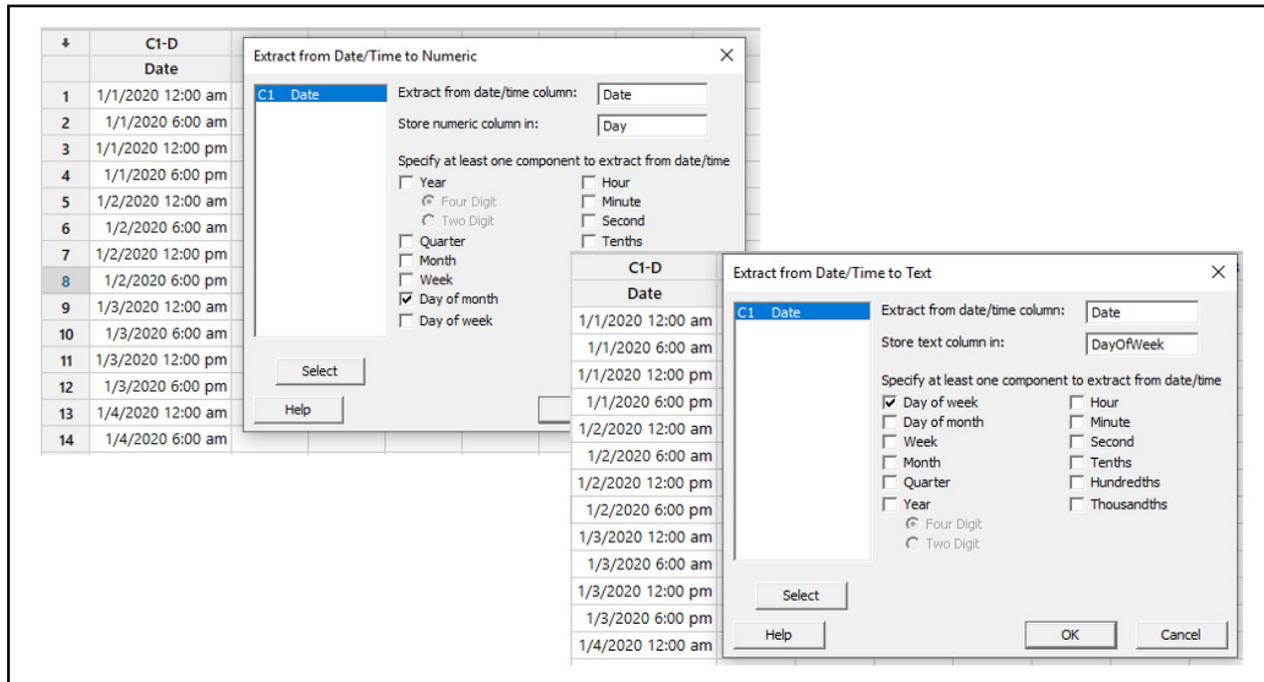
#	C1	C2	C3	C4	C5	C6	C7	C8	C9
	Weight	Pressure	Age	Minutes		Scaled Weight	Scaled Pressure	Scaled Age	Scaled Minutes
1	95.413	199.839	53	3		-6.0238	-0.31153	10.2	-16.8333
2	106.096	209.349	54	14		4.6585	9.19806	11.2	-5.8333
3	108.119	197.524	76	5		6.6817	-2.62648	33.2	-14.8333
4	112.376	194.772	33	8		10.9384	-5.37907	-9.8	-11.8333

Not only can multiple variables be scaled at the same time, but Minitab also offers the flexibility to specify the values used for scaling (Subtract first value, then divide by the second) or to set the entire range for each variable to the end points specified in the above window (Make range from start to end).

## 6. Extracting Date

While date and time data may contain important information for predicting and controlling a target or response variable, in its raw form, date/time data may not be useful in predictive modeling. Depending on formatting, date/time data may be treated as numeric (many applications, including [Minitab](#), store date/time data as fractions of days) or each unique value may be treated as a grouping variable, neither of which is desirable or informative. Minitab includes a menu that was exclusively designed to simplify the extraction of date/time data into a usable format.

**Data > Date Time > Extract from Date/Time to Numeric** or **Extract from Date/Time to Text** include several built-in functions to extract only the feature information that is relevant for modeling a process. The decision on what information to extract is highly dependent on the process itself, therefore process knowledge is indispensable. For example, the hour of the day could be a useful feature in predicting an airline's flight delays. As another example, extracting the day of the week could be helpful in determining whether certain days of the week are associated with a higher defect rate in a process.



Values that are extracted into a numeric format can be introduced as continuous predictors in a model, whereas values that are extracted to a text format are useful as categorical predictors. Minitab can also extract elapsed time between events and offers many other date/time manipulation features in the calculator. For a more detailed guide to date/time manipulation in Minitab, see our related [Tips & Tricks for Date/Time Data](#) post.

## 7. Missing Value Imputation

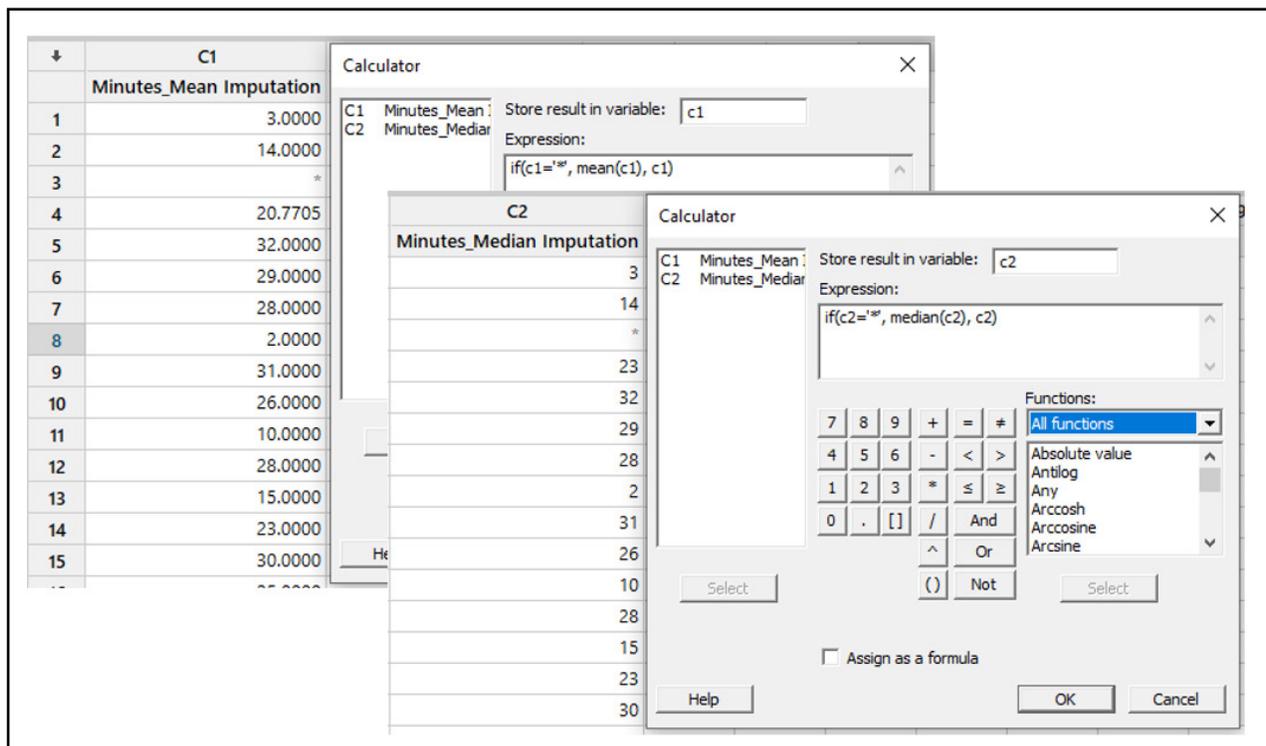
Certain algorithms, such as linear or logistic regression require complete records in each row to process the information and build a model. When values are missing, one possible approach is to discard any records with missing values, like Minitab Statistical Software does automatically. When the amount of data is limited, discarding partial information could affect the model’s ability to detect truly significant effects. Another option is to calculate the missing observations based on other values in each feature or column, which is simple to do using Minitab’s calculator.

For numeric imputation, the missing values could easily be replaced with the column mean or median using a logical if/then statement in Minitab’s calculator. The syntax is easy to break down:

Logical statement: IF(this thing is true, do this, otherwise do this)

Calculate the mean or median: mean(column) or median(column)

In Minitab’s calculator, we can combine the logical and mean or median expressions into a single statement as shown below. If the value in a given column is missing (denoted by \* for numeric data), use the mean or median value to replace the missing value, otherwise if the value is not missing, leave things as they are:

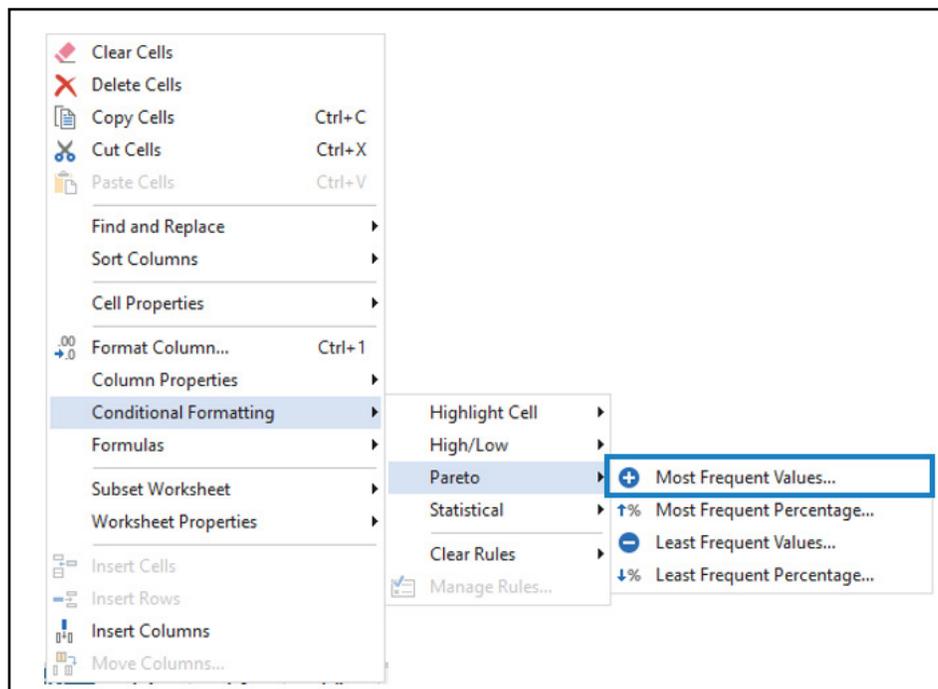


	C1	C2
	<b>Minutes_Mean Imputation</b>	<b>Minutes_Median Imputation</b>
1	3.0000	
2	14.0000	
3	*	
4	20.7705	
5	32.0000	
6	29.0000	3
7	28.0000	14
8	2.0000	*
9	31.0000	23
10	26.0000	32
11	10.0000	29
12	28.0000	28
13	15.0000	2
14	23.0000	31
15	30.0000	26
...	...	10
		28
		15
		23
		30

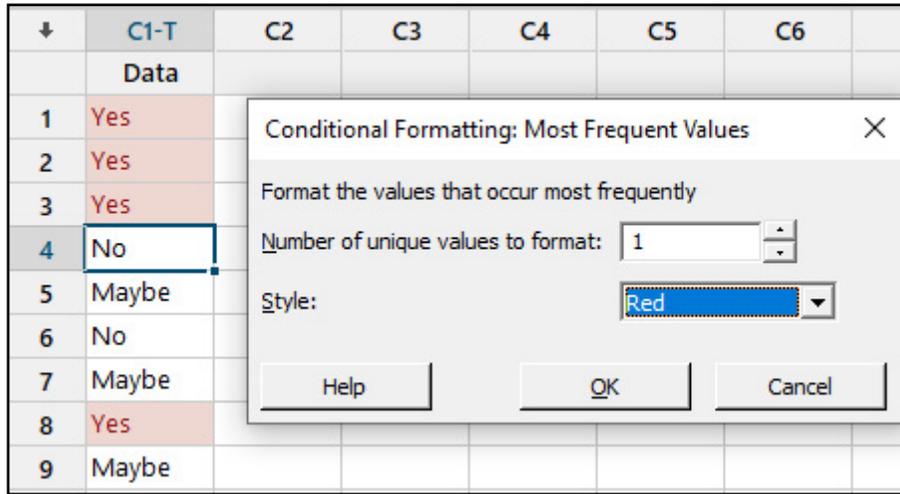
Using either the mean or median, all missing values in the column are instantly replaced:

C1	C2
Minutes_Mean Imputation	Minutes_Median Imputation
3.0000	3
14.0000	14
20.7852	23
20.7705	23
32.0000	32
29.0000	29

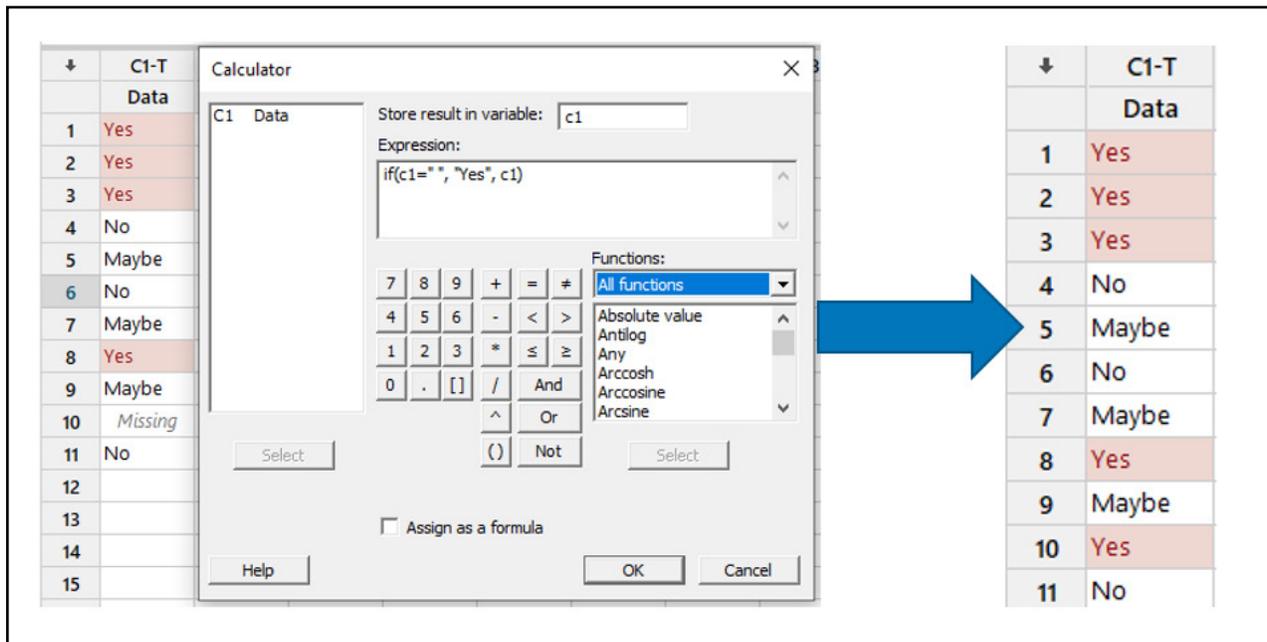
For imputation of text features, missing values could be recoded to a different category. Using a conversion table as previously discussed in the Binning or Bucketing section in this whitepaper. As another option, it may be desirable to recode missing text values to reflect the most frequently occurring value in the column. While there are several ways to identify the most frequently occurring text value, Minitab's [conditional formatting](#) options provide a shortcut. Simply right click on the text column and choose **Conditional Formatting > Pareto > Most Frequent Values**.



In the Most Frequent Values window, type in the number of values to format. In this case, we are interested only in the one most common value, so we use 1 and the corresponding values are highlighted in the worksheet.



To replace missing values with the most frequently occurring value of “Yes”, the calculator’s IF logical statement can be used once again. Notice that a space between double quotes is the syntax used to define a missing value in the calculator.



## Get Started Now

Now that you're familiar with these feature engineering techniques, you're ready to start applying them to your predictive analytics and machine learning projects. And remember: the key to building accurate predictive models is applying knowledge of your processes and your data in your feature engineering technique selection.

## About Minitab

Minitab has spent the past 50 years helping companies and organizations drive cost containment, enhance product and service quality and boost customer satisfaction with our comprehensive, easy-to-use, best-in-class suite of data analysis and process improvement tools. Thousands of businesses and institutions worldwide work with Minitab because of our unique and integrated approach of solutions analytics, empowering you to make better decisions that drive business excellence.

Interested in accessing additional guidance with your predictive analytics projects? Contact us to get help from our expert statisticians, consultants and world-class technical support today.



Talk to Minitab

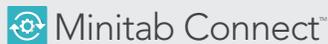
[Download Your Free Trial of Minitab Now](#)

Data Analysis



Powerful statistical software everyone can use

Data Transformation



Data access, automation, and governance for comprehensive insights

Predictive Modeling



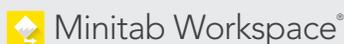
Machine learning and predictive analytics software

Online Stat Training



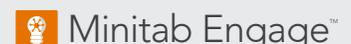
Master statistics and Minitab anywhere with online training

Visual Business Tools



Visual tools to ensure process and product excellence

Project Ideation & Execution



Start, execute, measure and report on improvement projects in real time

Minitab® and the Minitab® logo are registered trademarks of Minitab, LLC in the United States and other countries. See [minitab.com/legal/trademarks](https://minitab.com/legal/trademarks) for more information.

©2021, Minitab, LLC. All Rights Reserved.