Restructuring of DoE Data for Pigment Stability Optimisation

John Steele

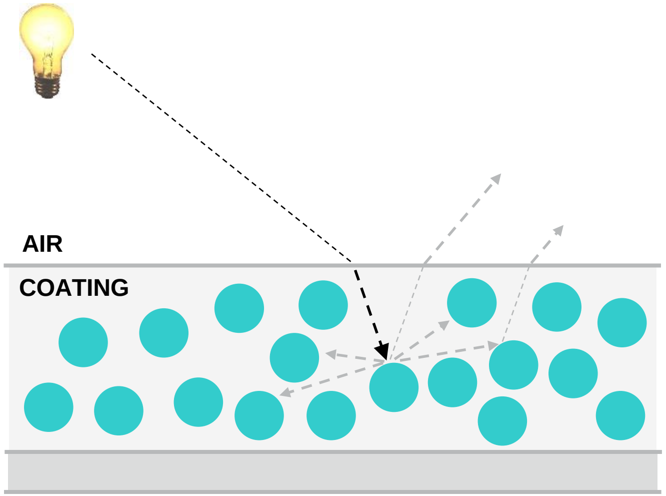
JMP Discovery Summit

March 2024

Technical Background

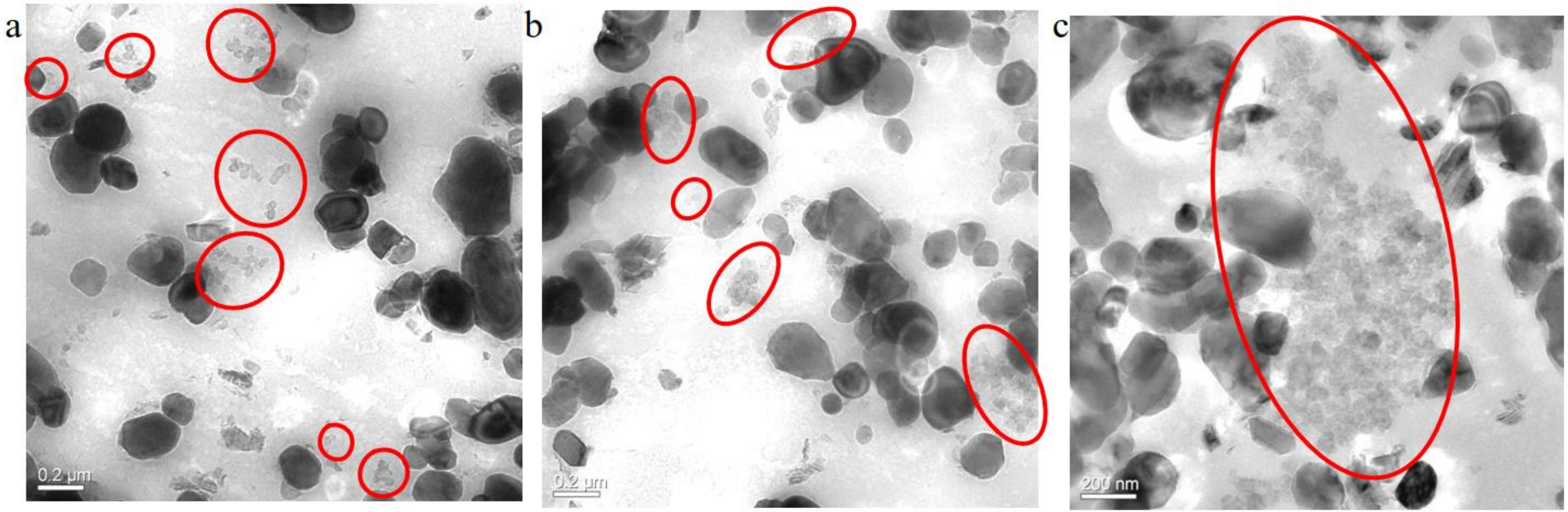
Light Scattering

AkzoNobel



- The opacity and colour of a paint film is a result of how it scatters and absorbs light
- The refraction of light is based on a pigment's refractive index
- *But* the overall amount of scattering that occurs is based on the size and number of pigment particles present in a film
- If particles aren't stable, the pigment will "flocculate"
  - Thus, there will be a change in the scattering behaviour
  - This impacts the desired opacity and colour of a paint system

Particle Flocculation



- Particle stabilisation is a complex subject involving a variety of electrostatic and steric interactions

- When we develop a new paint formulation, we need to ensure that pigments are stable (especially to outside forces) so that we can consistently deliver the target colour and opacity performance

Why is this Important?

- Mixing
  - Changes in colour on mixing and stirring could lead to the paint colour not being what the customer paid for!

- Shear
  - Different methods of application apply different levels of shear force when painting
  - This could result in the paint being a different colour depending on whether you use a roller or a brush!

- Mixing and stirring of paint could result in unwanted color changes due to particle destabilization and flocculation

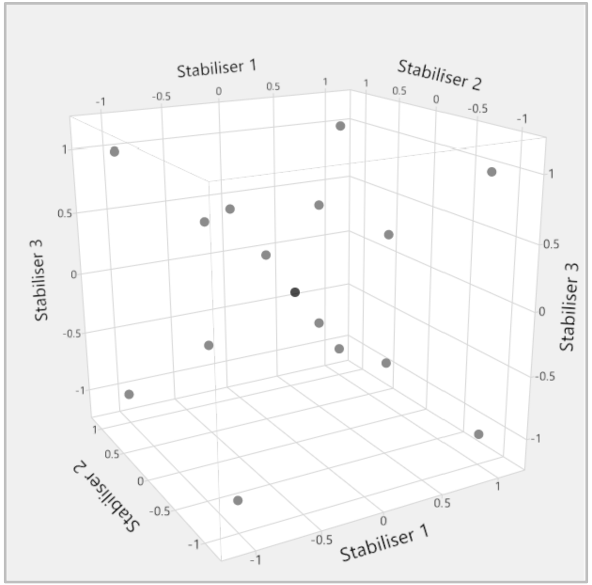



The Problem

Initial DoE

- A DoE was performed to investigate the impact of 3 different stabilisers in a fixed paint system
  - 17 runs examining interactions between the stabilisers
- Pigment stability was tested for 5 different materials added to these runs
  - Tested for shear and mixing stability
  - Tested at 3 different time points
- Design and testing plan all followed a sensible structure
- However…

Stabiliser 1	Stabiliser 2	Stabiliser 3
-1	-1	-1
-1	-1	1
-1	1	1
1	-1	-1
1	-1	1
1	1	-1
-1	1	-1
0	0	0
-0.5	-0.5	0.5
-0.5	0.5	0.5
0.5	0.5	0.5
0.5	-0.5	-0.5
0.5	-0.5	0.5
0.5	0.5	-0.5
-0.5	0.5	-0.5
1	1	1
0	0	0

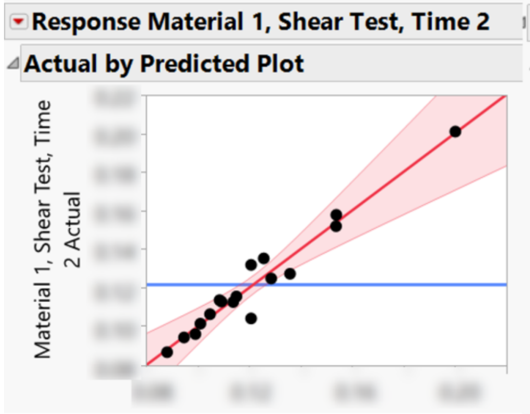


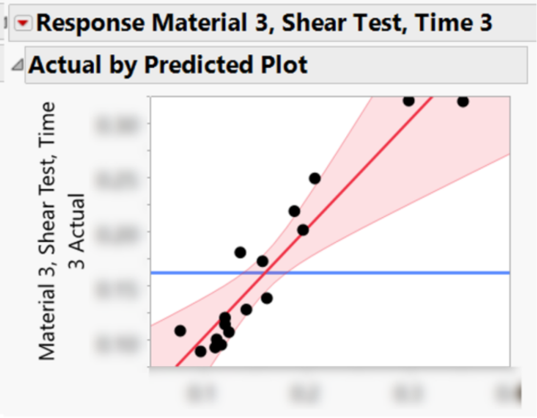
Initial Analysis

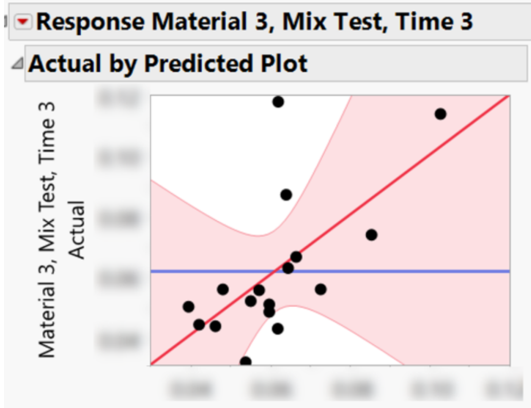
- Each response was analysed separately for each combination of:
  - Test
  - Material
  - Time

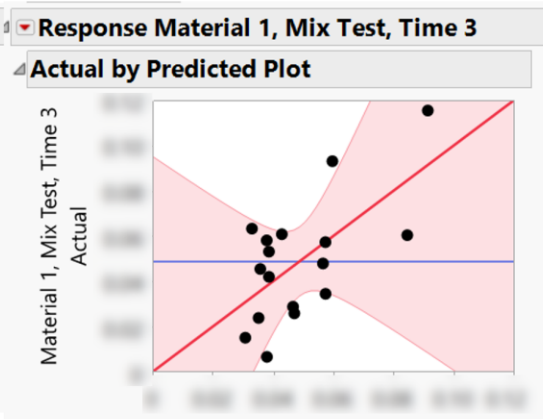
- A total of 25 responses
  - Some of these modelled quite well
  - Others did not…
  - No real pattern to which of these categories a response would fall

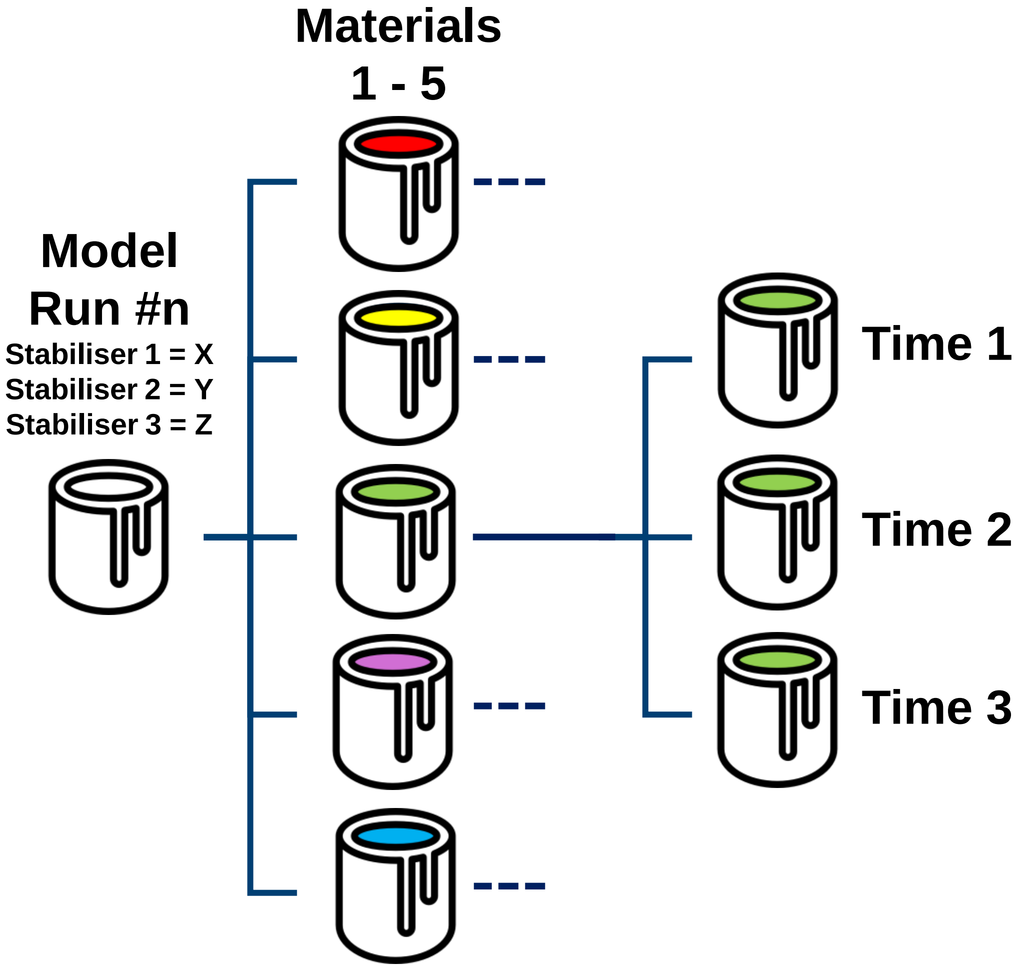
- Some responses had a very small range of values, while others had a very large range











AkzoNobel

- These combinations of time and material are repeated for both the shear test and the mix test
- Both the material and the time are actually ***factors***
- Including these ***"hidden factors"*** within the model would dramatically increase our data range
- But how do we go about doing this?

AkzoNobel

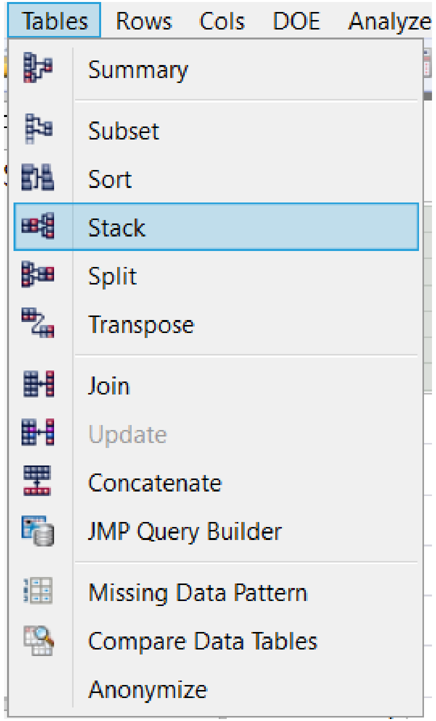
Restructuring the Data

Restructuring
Step 1

- First the data needs to be "stacked"

- This takes the 25 separate columns and converts the data into only 2 columns
  - One with the column header label
  - One with the actual data

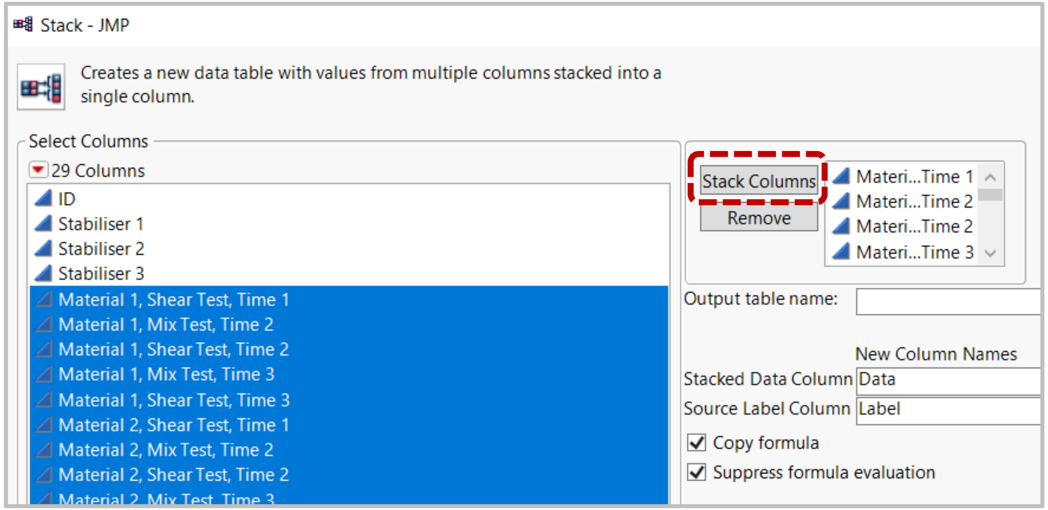
- JMP has a variety of tools for the restructuring of data under the "Tables" menu
  - I personally find Stack to be the most useful.

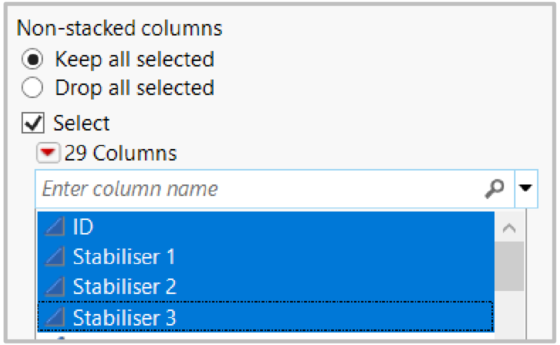


Restructuring
Step 1

- We select the data columns we want to stack and add them to the stack columns list

- For the "non-stacked columns" we want to select only the existing factors, and the run IDs
  - This prevents the unnecessary duplication of data





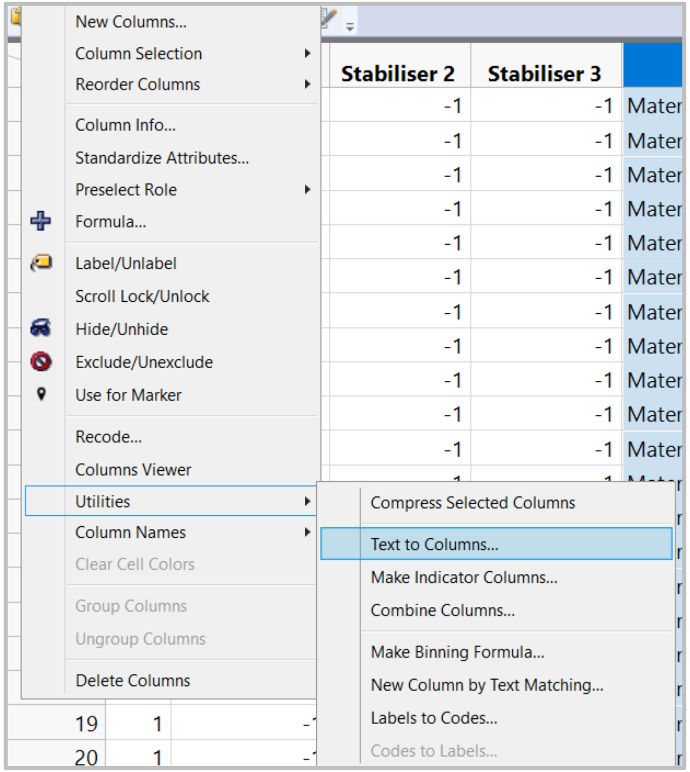
Restructuring
Step 1 - Output

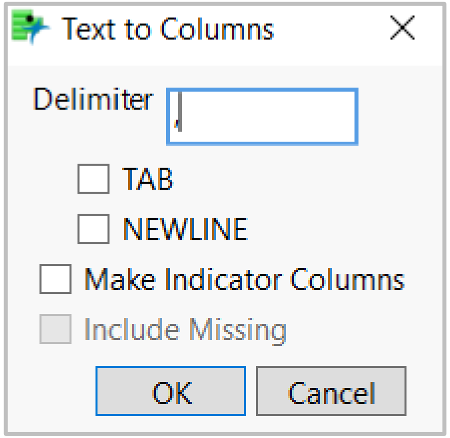
	ID	Stabiliser 1	Stabiliser 2	Stabiliser 3	Label	Result
1	1	-1	-1	-1	Material 1, Shear Test, Time 1	[illegible]
2	1	-1	-1	-1	Material 1, Mix Test, Time 2	[illegible]
3	1	-1	-1	-1	Material 1, Shear Test, Time 2	[illegible]
4	1	-1	-1	-1	Material 1, Mix Test, Time 3	[illegible]
5	1	-1	-1	-1	Material 1, Shear Test, Time 3	[illegible]

- This converts the individual column headers for the “hidden factors” and test combinations into a set of string data
- These now appear alongside all the original, initial factors
- However, we still need to split these string into something that we can use as separate sets of factor data

Restructuring
Step 2

- We can split these strings using the "Text to Columns" function
  - Very similar to the function of the same name in Excel
- This can be found under the "Utilities" section of the "Columns" menu
- If we specify our delimited as a comma, it will split our string data into its separate components





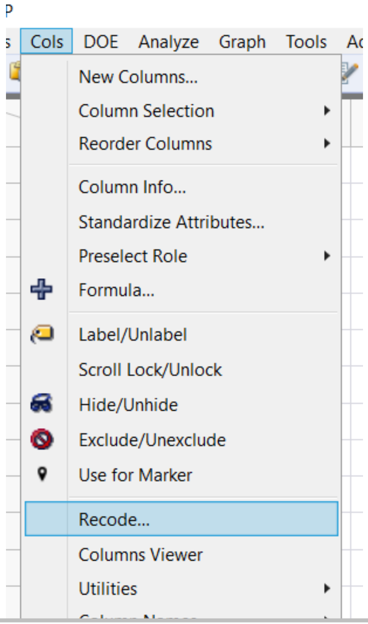
Restructuring
Step 2 – Output

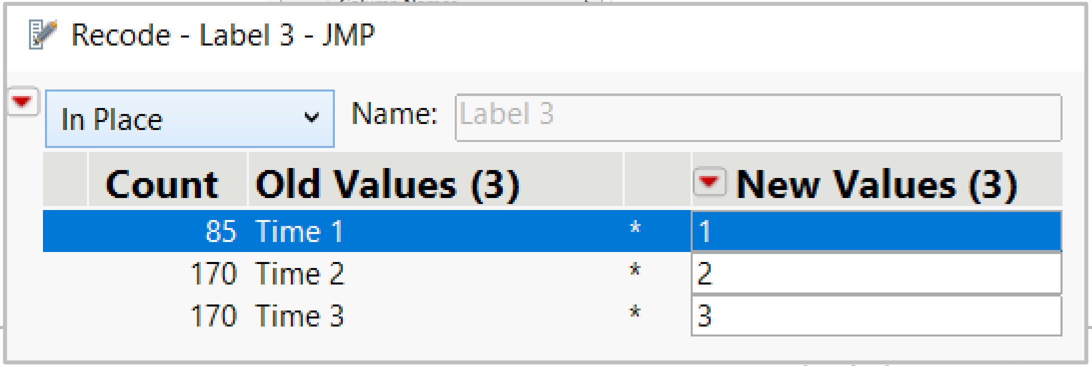
Label	Label 1	Label 2	Label 3	Result
Material 1, Shear Test, Time 1	Material 1	Shear Test	Time 1	[illegible]
Material 1, Mix Test, Time 2	Material 1	Mix Test	Time 2	[illegible]
Material 1, Shear Test, Time 2	Material 1	Shear Test	Time 2	[illegible]
Material 1, Mix Test, Time 3	Material 1	Mix Test	Time 3	[illegible]
Material 1, Shear Test, Time 3	Material 1	Shear Test	Time 3	[illegible]

- The original data remains, but additional columns have been added to contain the separated factor data
- We can now reformat this and tidy it up ready for use

Restructuring
Step 3

- Since the time factor is actually a numeric value, we need to change it from this string format
- We now have 425 data rows, so we don’t want to do this manually
- The recode tool in the columns menu is a quick and efficient way to do this





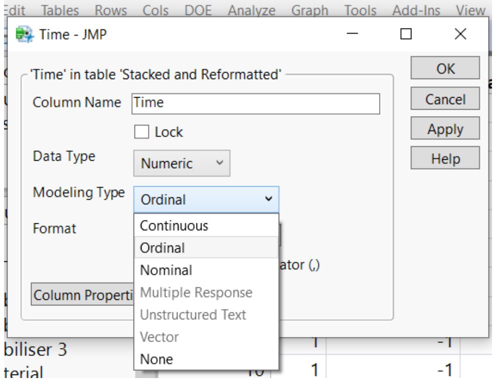
Restructuring
Step 3

AkzoNobel

- The recoded values will still be entered as “character” values and will need changing to numeric values via the column info menu

- This set of tests only have 3 different time-points so we can potentially consider changing the type to *ordinal* numeric for the purpose of analysis so that its options in the analysis profiler are discrete categoric factors rather than a continuous numeric range

- The end result is a table with 5 factors, 1 column defining the test type, and 1 column defining the result

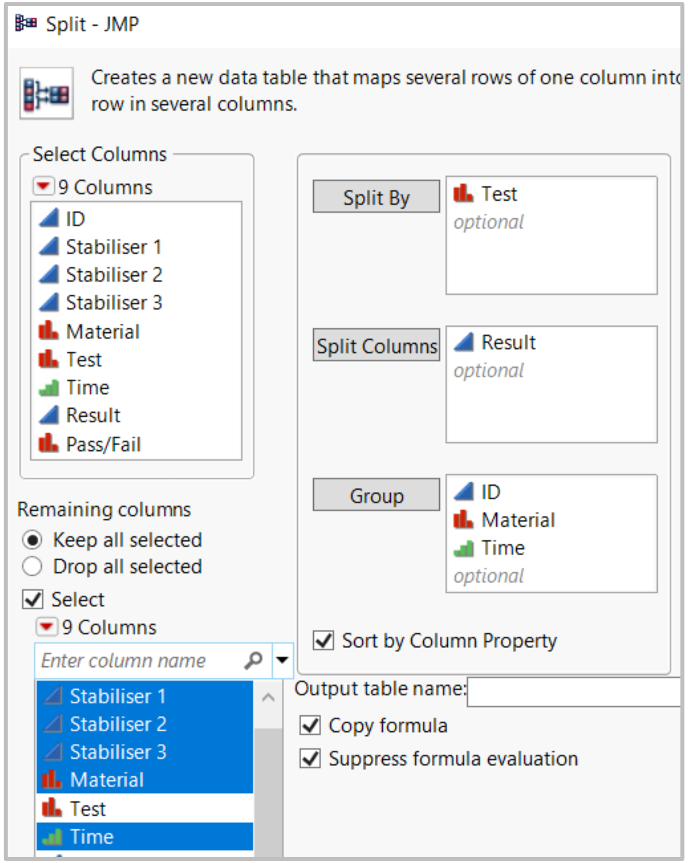


Material	Test	Time	Result
Material 1	Shear Test	1	[illegible]
Material 1	Mix Test	2	[illegible]
Material 1	Shear Test	2	[illegible]
Material 1	Mix Test	3	[illegible]

Restructuring
Step 4 (Optional)

- Potentially we can use the split function in the tables menu to reformat the data so that we have a separate, labelled column for each different result

- This isn't required (but can be useful from an interpretability perspective) as when we analyse the data we can use the fit model's "by" option to separate our data based on the individual test type

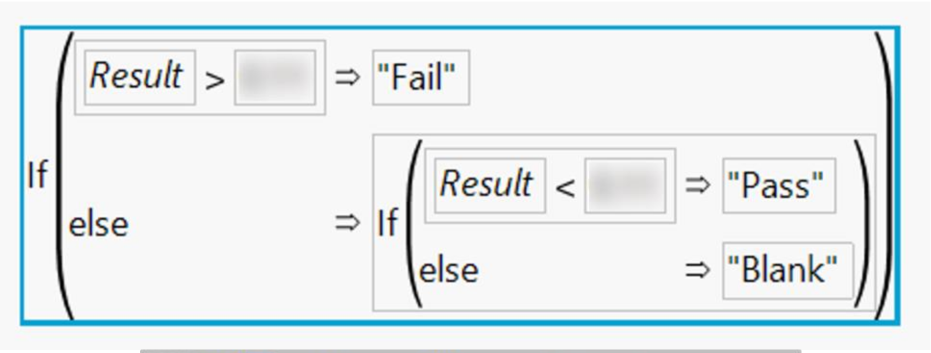


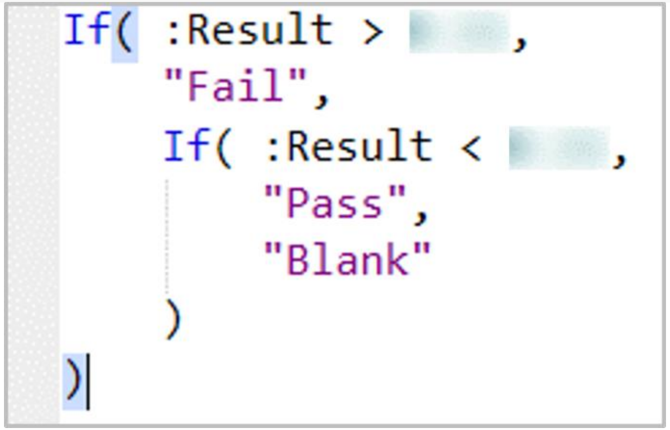
Restructuring
Step 4 (Optional) – Output

ID	Material	Time	Stabiliser 1	Stabiliser 2	Stabiliser 3	Mix Test	Shear Test
1	Material 1	1	-1	-1	-1	[illegible]	[illegible]
1	Material 1	2	-1	-1	-1	[illegible]	[illegible]
1	Material 1	3	-1	-1	-1	[illegible]	[illegible]
1	Material 2	1	-1	-1	-1	[illegible]	[illegible]
1	Material 2	2	-1	-1	-1	[illegible]	[illegible]
1	Material 2	3	-1	-1	-1	[illegible]	[illegible]

Adding Pass/Fail Conditions

- For these tests, the specific result is usually less important than whether it gives a pass or a fail
- We can use a formula column to translate the numerical results into categories based on the pass/fail thresholds
- Can be built using the formula tool, or coded manually
- Can also be set up using “make binning formula” under the columns’ utilities menu

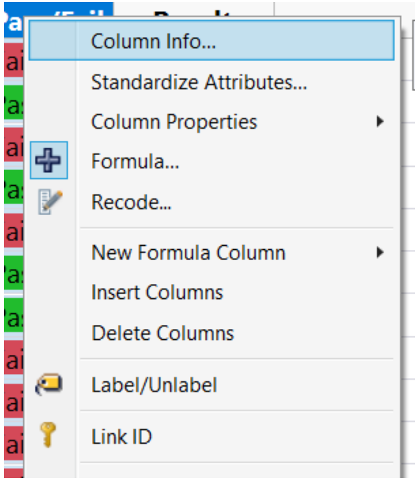


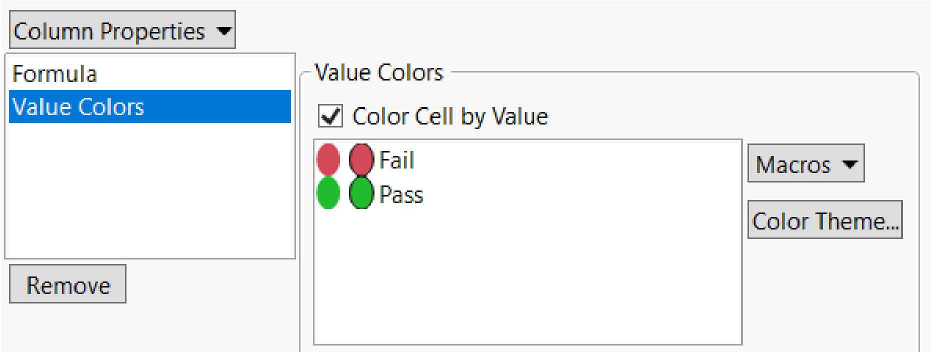


Adding Pass/Fail Conditions

- Using the column info options it is also possible to colour the cells based on their contents

- Select “value colours” from the column properties menu
  - Assign colours
  - Make sure “colour by cell value” is selected





Adding Pass/Fail Conditions
End Result

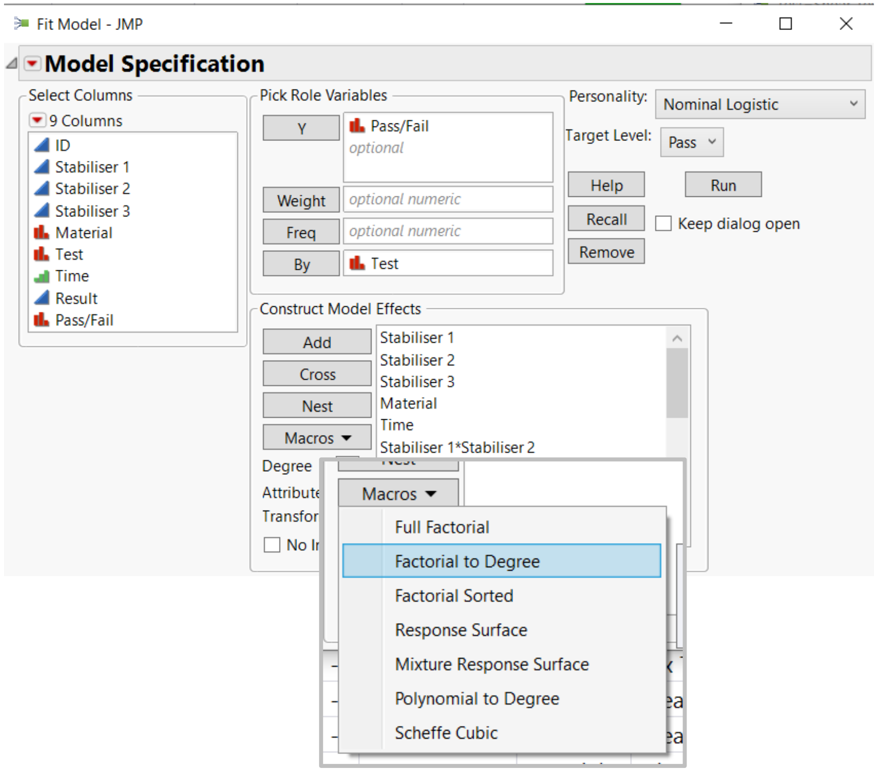
ID	Stabiliser 1	Stabiliser 2	Stabiliser 3	Material	Test	Time	Pass/Fail
1	-1	-1	-1	Material 1	Shear Test	1	Fail
1	-1	-1	-1	Material 1	Mix Test	2	Pass
1	-1	-1	-1	Material 1	Shear Test	2	Fail
1	-1	-1	-1	Material 1	Mix Test	3	Pass
1	-1	-1	-1	Material 1	Shear Test	3	Fail

Modelling the Data

Logistic Regression Modelling
Inputs

AkzoNobel

- Logistic regression is a type of categorisation model
  - Excellent for our Pass/Fail data
  - Model type automatically assigned by JMP when categoric data is added as a response (Y)
- Use "By" to split the data into two separate models based on the test label
- Factor interactions can be quickly added using the "Factorial to Degree" option under Macros
  - Uses the degree specified in the Degree box
  - For this model I used degree = 3 to give information on possible three factor interactions



Logistic Regression Modelling
Outputs – Model Quality

- Data shown here is for the “mix test” data

- Logistic regression models have a “confusion matrix” output
  - Shows how well the model classifies the categories
    - Similar to standard predicted vs. actual plots
  - For this model 2 rows are predicted as passes, but are actually failures
    - More useful that R^2 values for this type of model
    - 2 mis-categorisations out of 170 data points is ~1.2%

Confusion Matrix

Training

Actual	Predicted Count	
Pass/Fail	Pass	Fail
Pass	142	0
Fail	2	26

Actual	Predicted Rate	
Pass/Fail	Pass	Fail
Pass	1.000	0.000
Fail	0.071	0.929

RSquare (U)	0.9434
AICc	190.968
BIC	311.616
Observations (or Sum Wgts)	170

Logistic Regression Modelling

Outputs – Effect Summary

- One of the main questions from the team performing this work was “what are the main drivers and impacts on our performance?”

- The model Effect Summary lists the factors and interactions that are having the biggest effect on the result
  - Which factors have significant interactions with which other factors?
  - Which factors and interactions are unimportant?

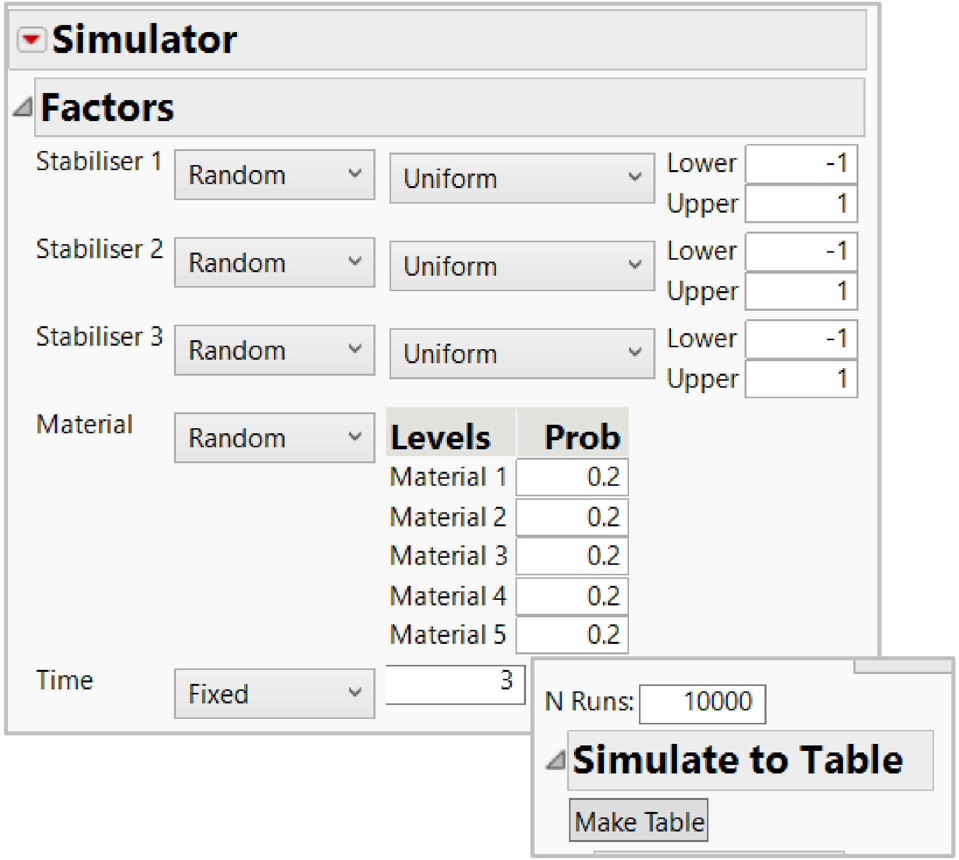
Effect Summary

Source	LogWorth	PValue	
Stabiliser 1*Stabiliser 2*Time	153.654	0.00000	
Stabiliser 2*Material	93.157	0.00000	
Stabiliser 1*Stabiliser 3*Material	75.264	0.00000	
Stabiliser 2*Stabiliser 3*Material	46.105	0.00000	
Stabiliser 3*Time	15.814	0.00000	
Stabiliser 1*Material	12.414	0.00000	^
Stabiliser 3	10.119	0.00000	^
Stabiliser 1*Stabiliser 2	4.800	0.00002	^
Material	3.784	0.00016	^
Stabiliser 2*Stabiliser 3	2.627	0.00236	^
Stabiliser 1*Time	1.853	0.01401	^
Stabiliser 2*Stabiliser 3*Time	0.955	0.11083	
Material*Time	0.535	0.29183	
Stabiliser 3*Material*Time	0.167	0.68136	
Stabiliser 1*Stabiliser 3	0.122	0.75448	^
Stabiliser 1	0.030	0.93285	^
Stabiliser 1*Material*Time	0.022	0.95042	
Stabiliser 2	0.002	0.99495	^
Stabiliser 1*Stabiliser 3*Time	0.001	0.99819	
Stabiliser 2*Time	0.000	0.99999	^
Stabiliser 2*Material*Time	.	.	
Stabiliser 1*Stabiliser 2*Material	.	.	
Stabiliser 1*Stabiliser 2*Stabiliser 3	.	0.00000	
Stabiliser 3*Material	.	.	
Time	.	0.00000	

Logistic Regression Modelling

Outputs – Data Simulation

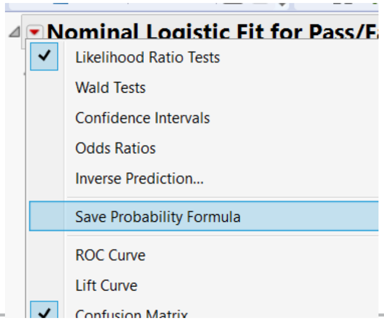
- For logistic regression, the contour profiler gives us options for simulating large amounts of data based on our model
- Gives expected results based on the model
- Can restrict the range of factors to be included



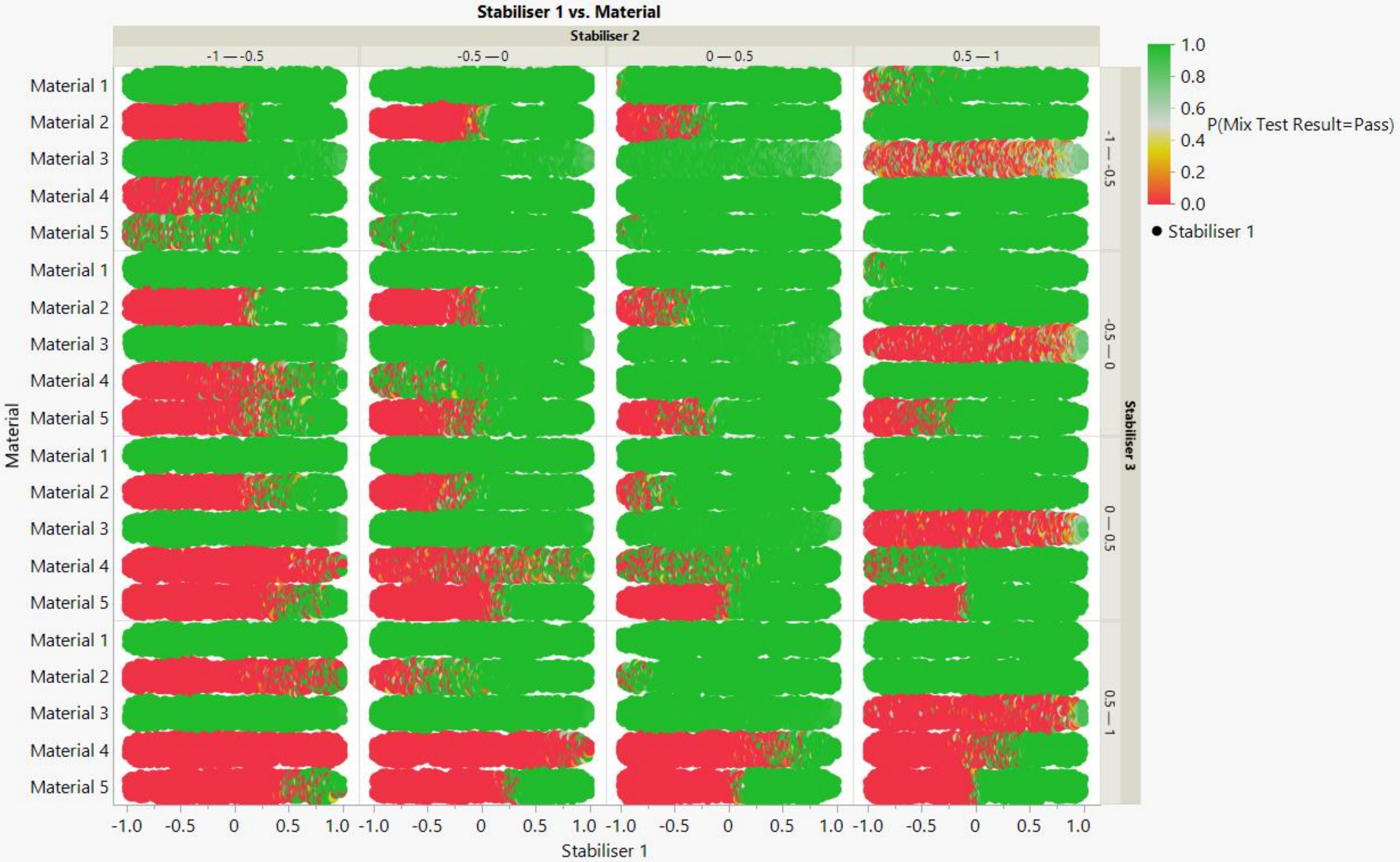
Logistic Regression Modelling
Outputs – Data Simulation

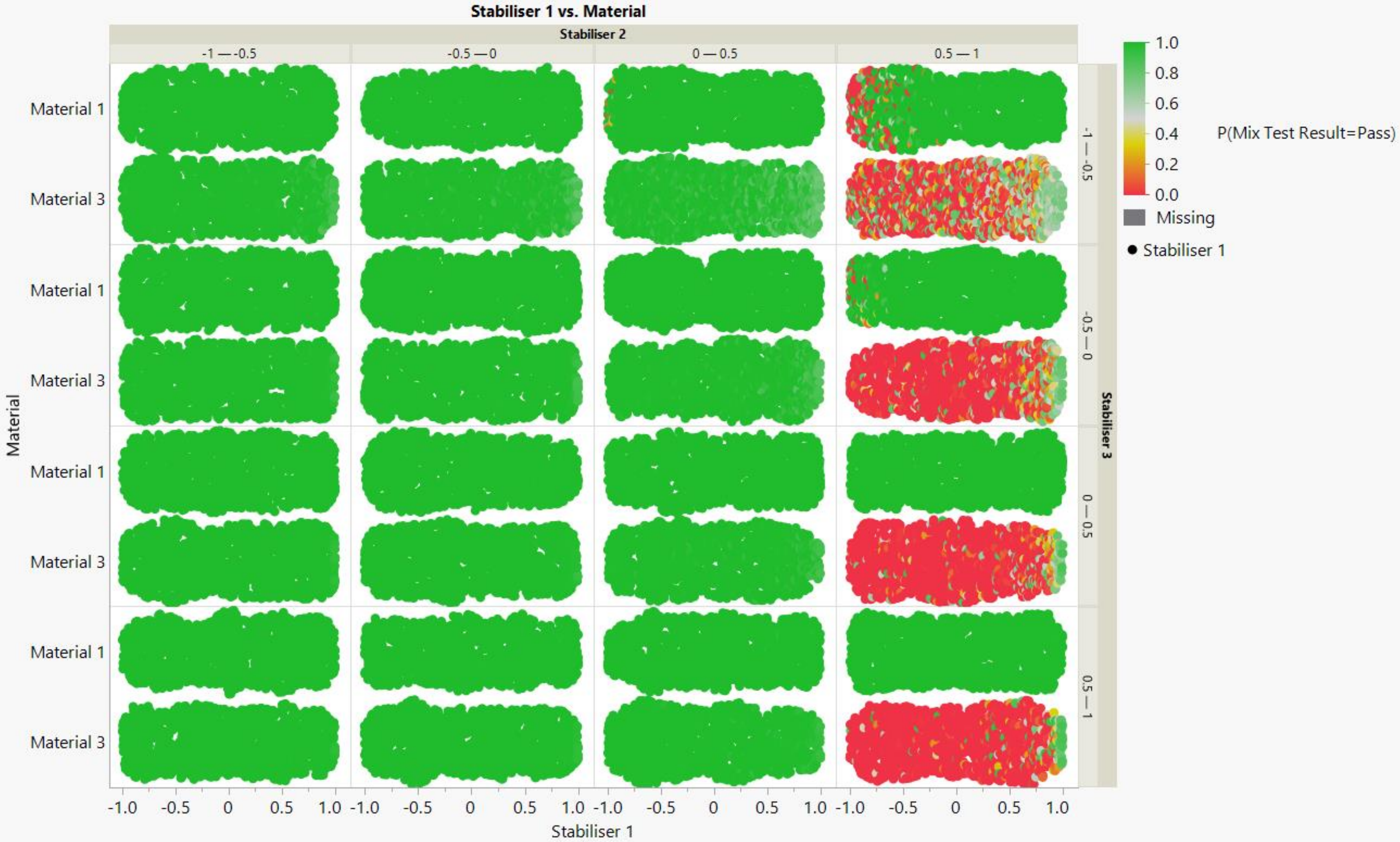
Stabiliser 1	Stabiliser 2	Stabiliser 3	Material	Time	P(Mix Test Result=Pass)	P(Mix Test Result=Fail)
0.9540191549	0.7611923888	-0.047480704	Material 3	3	0.6181572412	0.3818427588
-0.957937326	0.5926631363	-0.590259448	Material 3	3	0.0061678739	0.9938321261
0.4842737811	0.3429494957	-0.599460405	Material 1	3	1	3.690998e-65

- Rather than giving *just* a pass/fail result, the data simulation gives the *probability* that a simulated set of factors will fall into a given category
  - We can also get this information for our original data table by selecting “save probability formula” via the red triangle
- With this data we can visualise where certain factor combinations lead to failures




Stabiliser 1 vs. Material
Stabiliser 2
-1 — -0.5
-0.5 — 0
0 — 0.5
0.5 — 1
Stabiliser 3
Stabiliser 1
1.0
0.5
0
-0.5
-1.0
Material 1
Material 2
Material 3
Material 4
Material 5
Material
P(Mix Test Result=Pass)
1.0
0.8
0.6
0.4
0.2
0.0
Stabiliser 1


Stabiliser 1 vs. Material
Stabiliser 2
-1 — -0.5
-0.5 — 0
0 — 0.5
0.5 — 1
Material
Material 1
Material 2
Material 3
Material 4
Material 5
Stabiliser 3
-1 — -0.5
-0.5 — 0
0 — 0.5
0.5 — 1
Stabiliser 1
-1.0
-0.5
0
0.5
1.0
P(Mix Test Result=Pass)
1.0
0.8
0.6
0.4
0.2
0.0
Stabiliser 1



Stabiliser 1 vs. Material
Stabiliser 2
-1 — -0.5
-0.5 — 0
0 — 0.5
0.5 — 1
Material 1
Material 3
Material 1
Material 3
Material 1
Material 3
Material 1
Material 3
Material
-1 — -0.5
-0.5 — 0
0 — 0.5
0.5 — 1
Stabiliser 3
-1.0 -0.5 0 0.5 1.0 -1.0 -0.5 0 0.5 1.0 -1.0 -0.5 0 0.5 1.0 -1.0 -0.5 0 0.5 1.0
Stabiliser 1
1.0
0.8
0.6
0.4
0.2
0.0
P(Mix Test Result=Pass)
Missing
Stabiliser 1

Summary

- The results and outputs of your experimental design may contain hidden factors you didn't originally consider
- An initial, poor-quality analysis doesn't mean your data doesn't have value
- JMP's table tools offers speedy and efficient options to restructure your data
- JMP's column tools allow for further restructuring and adjustment of your data
- There are modelling and visualisation options beyond basic multiple linear regression

Questions?