

JMP Discovery Summit October 2020

Automating the Data Curation Workflow

Mia Stephens

JMP Principal Product Manager

Jordan Hiller

JMP Senior Systems Engineer



Abstract

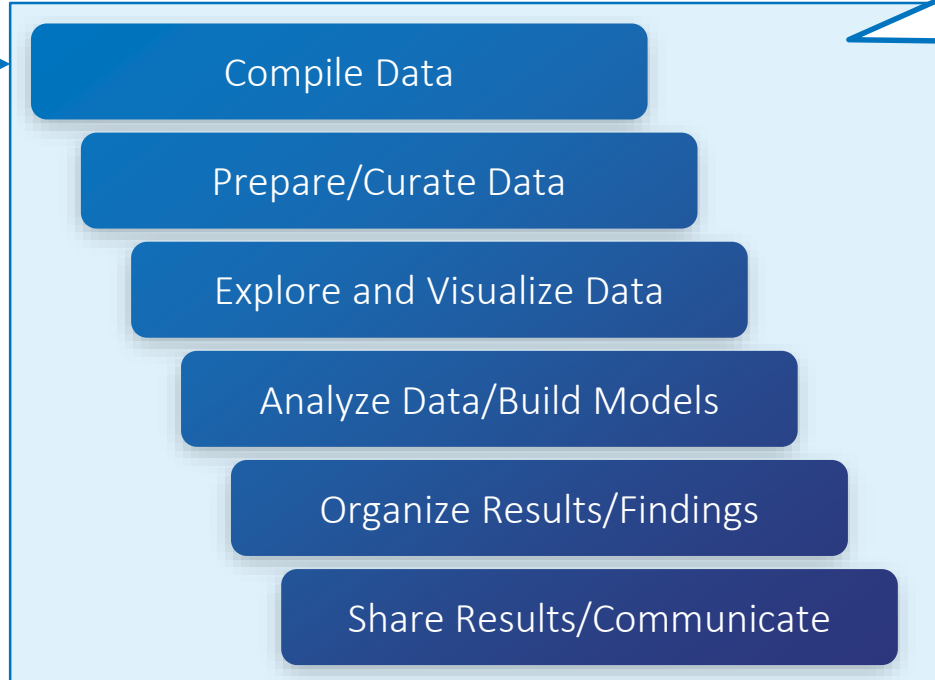
For most data analysis tasks, a lot of time is spent up front – importing data and preparing it for analysis. Because we often work with datasets that are regularly updated, automating our work using scripted repeatable workflows can be a real time saver. There are three general sections in an automation script: data import, data curation, and analysis/reporting. While the tasks in the first and third sections are relatively straightforward -- point-and click to achieve the desired result, and capture the resulting script -- data curation can be more challenging for those just starting out with scripting. In this talk we review common data preparation activities, discuss the jsl code necessary to automate the process, and provide advice for generating jsl code for data curation via point-and-click.

Outline

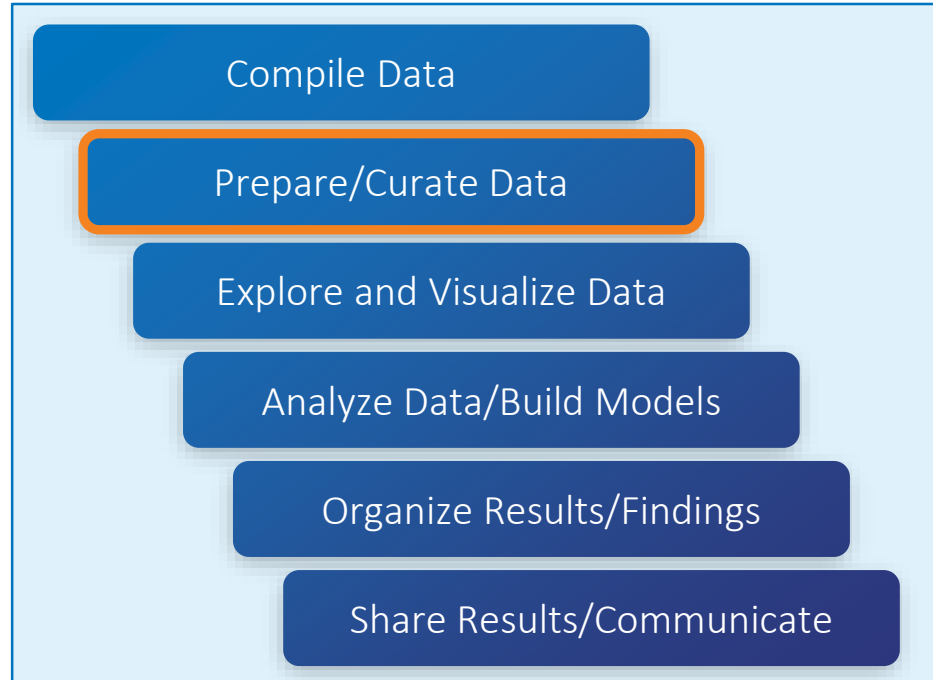
- Analytic Workflow
- What is Data Curation?
- The Need for Reproducibility
- How to in JMP 15
- Sneak Peak at JMP 16 Action Recording and the Enhanced Log mode

Analytic workflow

What is the business problem?



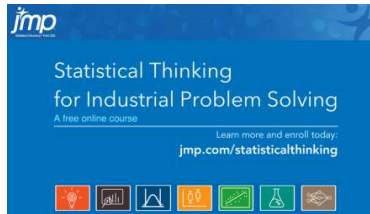
Analytic workflow




What is Data Curation?

- Ensuring that data are useful in driving analytic discoveries.
- Largely about data organization and cleanup.
- Addresses these common issues:
 - incorrect formatting
 - incomplete data
 - missing data
 - dirty or messy data

Borrowed from STIPS






Statistical Thinking for Industrial Problem Solving

A free online course

Learn more and enroll today:
jmp.com/statisticalthinking



Module 2: Exploratory Data Analysis

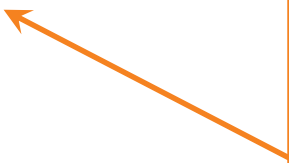


 Data Preparation for Analysis

What is Data Curation?

Common Issues

- Ensuring that data are useful in driving analytic discoveries.
- Largely about data organization and cleanup.
- Addresses these common issues:
 - incorrect formatting
 - incomplete data
 - missing data
 - dirty or messy data



Data are in the wrong form or format for analysis:

- Data table as a whole
- Individual variables
- Cosmetic

What is Data Curation?

Common Issues

- Ensuring that data are useful in driving analytic discoveries.
- Largely about data organization and cleanup.
- Addresses these common issues:
 - incorrect formatting
 - incomplete data
 - missing data
 - dirty or messy data

Lack of data

- On important variables
- On combinations of variables
- Not enough data (observations)

What is Data Curation?

Common Issues

- Ensuring that data are useful in driving analytic discoveries.
- Largely about data organization and cleanup.
- Addresses these common issues:
 - incorrect formatting
 - incomplete data
 - missing data
 - dirty or messy data

Values for variables not available

- Missing completely at random
- Missing at random
- Missing not at random

What is Data Curation?

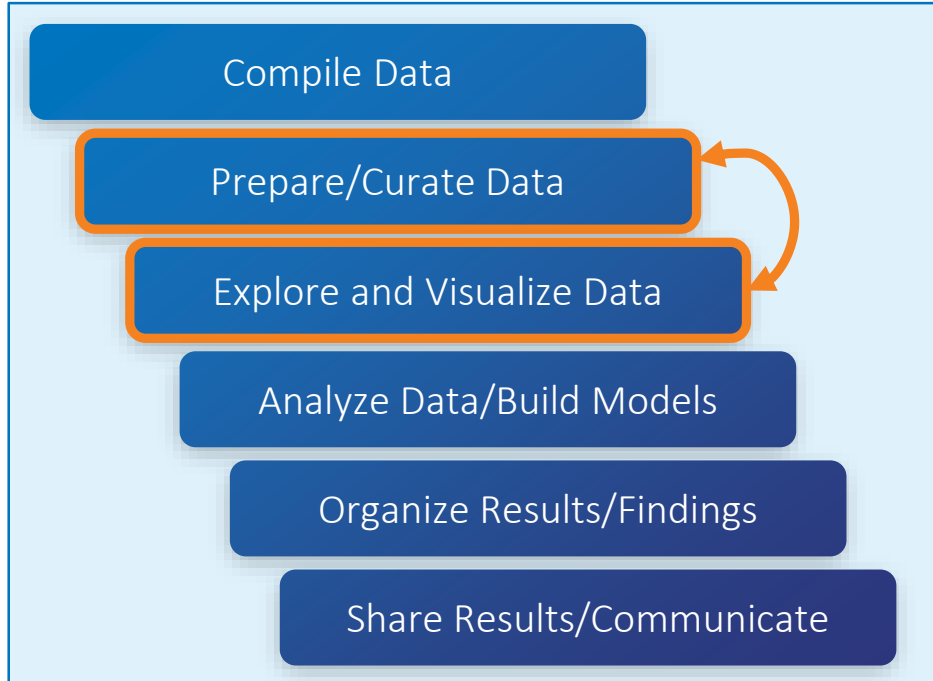
Common Issues

- Ensuring that data are useful in driving analytic discoveries.
- Largely about data organization and cleanup.
- Addresses these common issues:
 - incorrect formatting
 - incomplete data
 - missing data
 - dirty or messy data

Issues with observations or variables

- Incorrect
- Inconsistencies
- Inaccurate
- Errors, typos
- Obsolete
- Outdated
- Censored
- Truncated
- Redundant
- Duplicated

How Do You Identify Potential Issues?

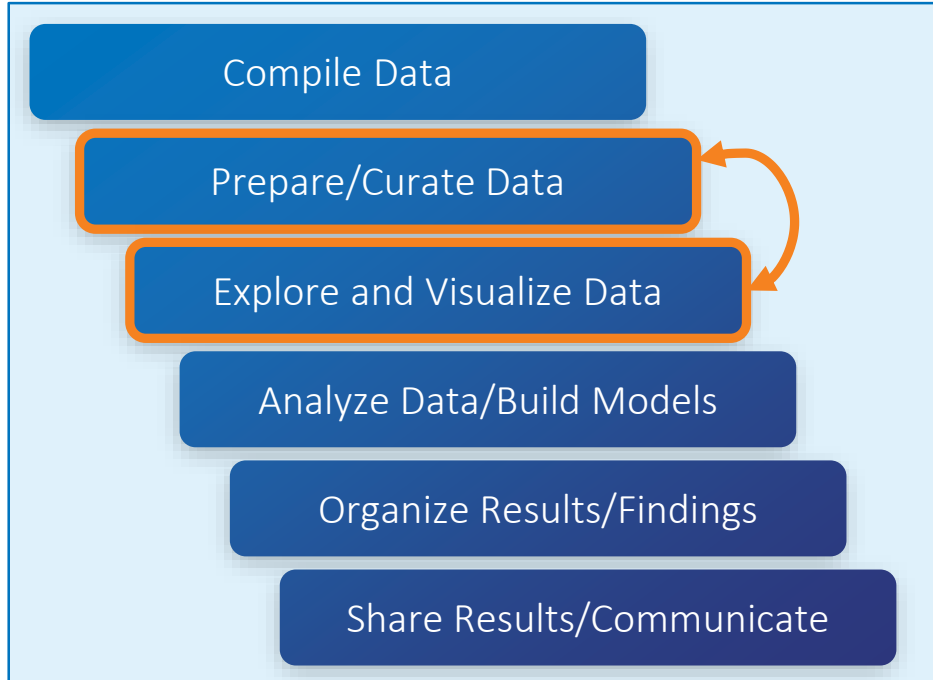


1. Scan the data table for obvious issues

Example: Components

facility	batch number	part number	customer number	batch size	number scrapped	pressure	humidity	dwell	temp	speed	process	vacuum	supplier	scrap rate
FabTech	1044	18769	37386-M2	500	815	16	54	105	107	130	2 on	1 off	Anderson Cox Worley Hersh	0.18
FabTech	10036	18769	35752-C2	500	33	15	44	75	100	80	1 off	1 off	Hersh	0.024
2 FabTech	10039	16935	25513-C1	1000	61	15	46	80	-	95	1 off	Cox	0.061	
3 FabTech	10040	16935	47210-X2	200	23	16	40	80	106	93	1 off		0.115	
4 FabTech	10041	18769	37938-M	200	29	16	40	75	106	73	2 on	Hersh	0.145	
5 FabTech	10042	18769	37386-M2	1000	59	15	46	80	-	105	1 off	Hersh	0.059	
6 FabTech	10043	16935	37938-M	200	28	15	N/A	76	107	73	2 on	Cox Inc.	0.14	
7 FabTech	10044	18769	35752-C2	5000	70	15	46	75	-	130	1 off	Cox	0.014	
8 FabTech	10045	18769	35752-C2	200	0	15	46	75	-	130	1 off	Cox	0	
9 FabTech	10046	16935	37938-M	1000	48	15	45	70	-	83	1 off	Hersh	0.048	
10 FabTech	10047	18769	35816-M5	500	65	15	43	75	106	88	2 on		0.13	
11 FabTech	10048	2283	37386-M2	5000	630	16	43	66	106	88	2 on	Hersh	0.126	
12 FabTech	10049	16935	35752-C2	500	58	15	50	80	104	85	1 off	Hersh	0.116	
13 FabTech	10050	2283	35752-C2	5000	520	15	50	70	105	80	1 off	Hersh	0.104	
14 FabTech	10051	2283	25513-C1	5000	485	15	45	70	105	75	1 off	Cox	0.097	
15 FabTech	10052	2283	35752-C2	500	39	15	N/A	75	-	80	1 off	Hersh	0.078	
16 FabTech	10053	16935	37938-M	200	-4	15	45	70	-	120	1 off	Worley	-0.02	
17 FabTech	10054	2283	37938-M	1000	19	15	45	76	-	120	1 off	Hersh	0.019	
18 FabTech	10055	18769	37938-M	500	10	15	43	70	-	110	1 off	Hersh	0.02	
19 FabTech	10056	16935	35752-C2	500	16	15	41	72	-	100	1 off	Cox	0.032	
20 FabTech	10057	16935	35752-C2	1000	24	15	42	74	-	105	1 off	hersh	0.024	
21 FabTech	10058	-	37386-M2	1000	29	15	42	70	-	93	1 off	Hersh	0.029	
22 FabTech	10059	16935	35752-C2	200	6	15	45	76	-	108	1 off	Hersh	0.03	
23 FabTech	10060	16935	37386-M2	1000	23	15	45	72	-	108	1 off	Hersh	0.023	
24 FabTech	10061	18769	35816-M5	200	6	15	42	75	-	105	1 off	Hersh	0.03	
25 FabTech	10062	16935	37386-M2	200	-6	15	45	70	-	110	1 off	Hersh	-0.03	
26 FabTech	10063	16935	37386-M2	500	29	15	44	90	-	103	1 off	Hersh	0.058	
27 FabTech	10064	18769	37386-M2	5000	305	15	44	91	-	103	1 off	Cox	0.061	
28 FabTech	10065	2283	37386-M2	500	46	15	44	78	104	76	1 off	Hersh	0.092	
29 FabTech	10066	2283	37386-M2	200	17	15	44	69	-	76	2 on	Hersh	0.085	
30 FabTech	10067	16935	35816-M5	500	12	15	47	78	-	120	1 off	Cox	0.024	

How Do You Identify Potential Issues?



1. Scan the data table for obvious issues
2. Explore data one variable at time

Columns Viewer

Summary Statistics

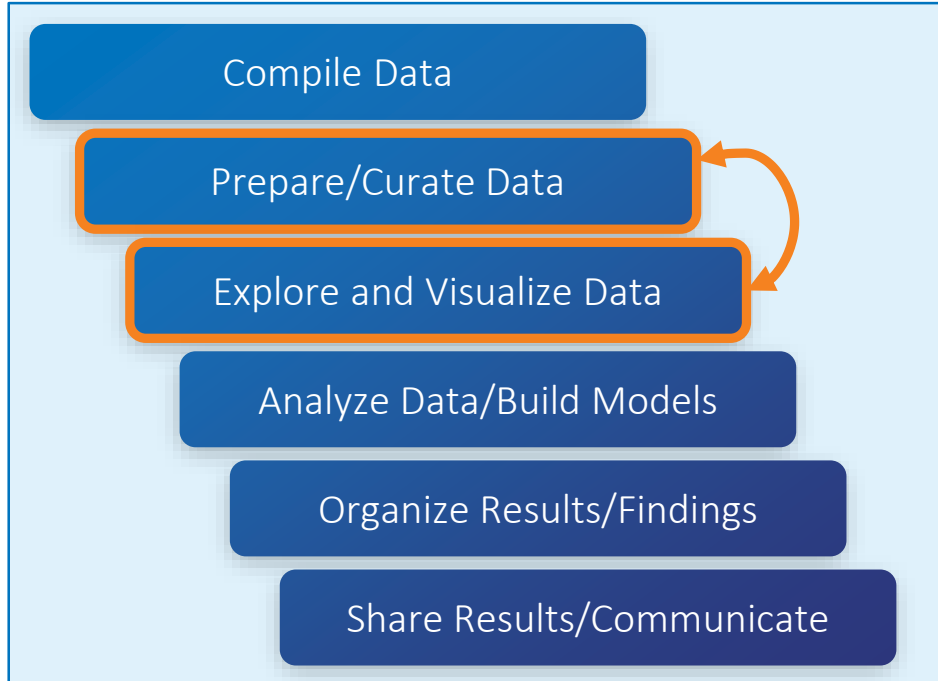
15 Columns

Columns	N	N Missing	N Categories	Min	Max	Mean	Std Dev	Median	Lower Quartile	Upper Quartile	Interquartile Range
facility	369	0	1
batch number	369	0	.	10038	10430	10231.165312	114.26265171	10228	10132	10329.5	197.5
part number	365	4	.	2283	18769	14848.852055	6447.4827416	18769	16935	18769	1834
customer number	369	0	20
batch size	369	0	4
number scrapped	369	0	.	-6	815	100.22764228	169.0031264	33	17	84	67
pressure	367	2	.	14.9	16.4	15.50	0.3211012828	15.46	15.26	15.70	0.44
humidity	368	1	37
dwell	366	3	.	57	105	78.844262295	7.802374586	78	74	82.25	8.25
temp	104	265	.	103	107	105.2	1.0072176978	105.0	104.5	106.0	1.5
speed	368	1	.	4	130	94.1	16.124410679	92.5	83.0	105.0	22.0
process	368	1	.	1	2	1.1494565217	0.35702331	1	1	1	0
vacuum	368	1	2
supplier	359	10	10
scrap rate	369	0	.	-0.03	0.176	0.0669756098	0.0400774651	0.057	0.036	0.09	0.054

Distribution

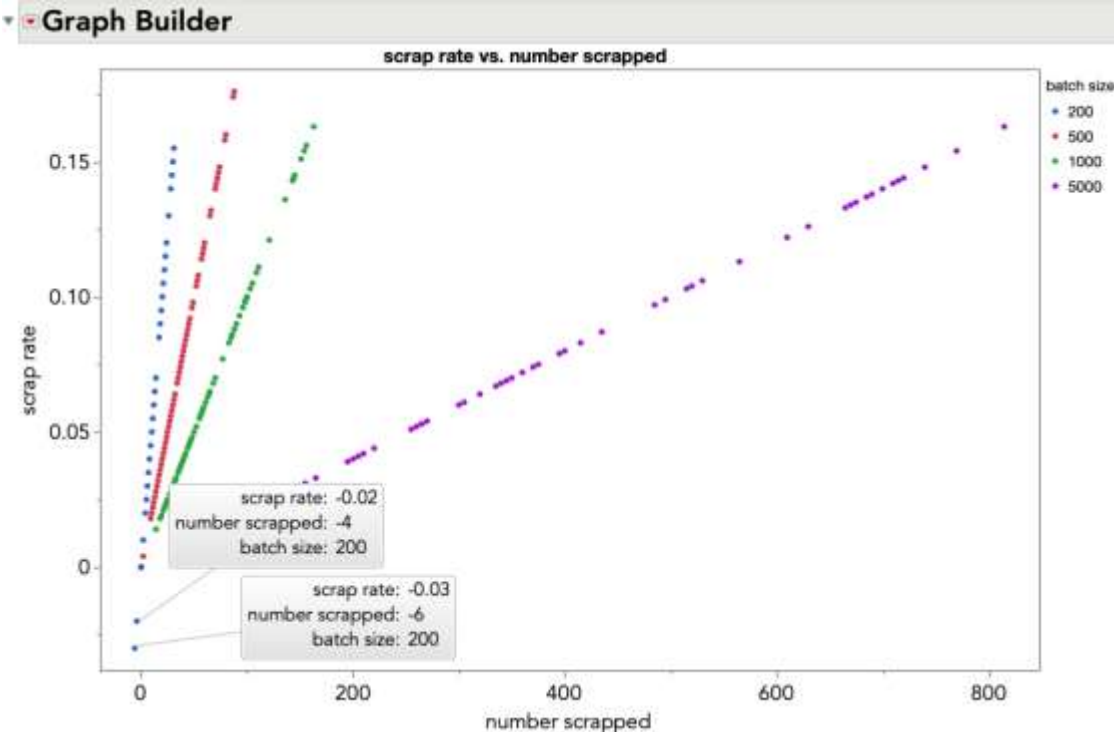


How Do You Identify Potential Issues?

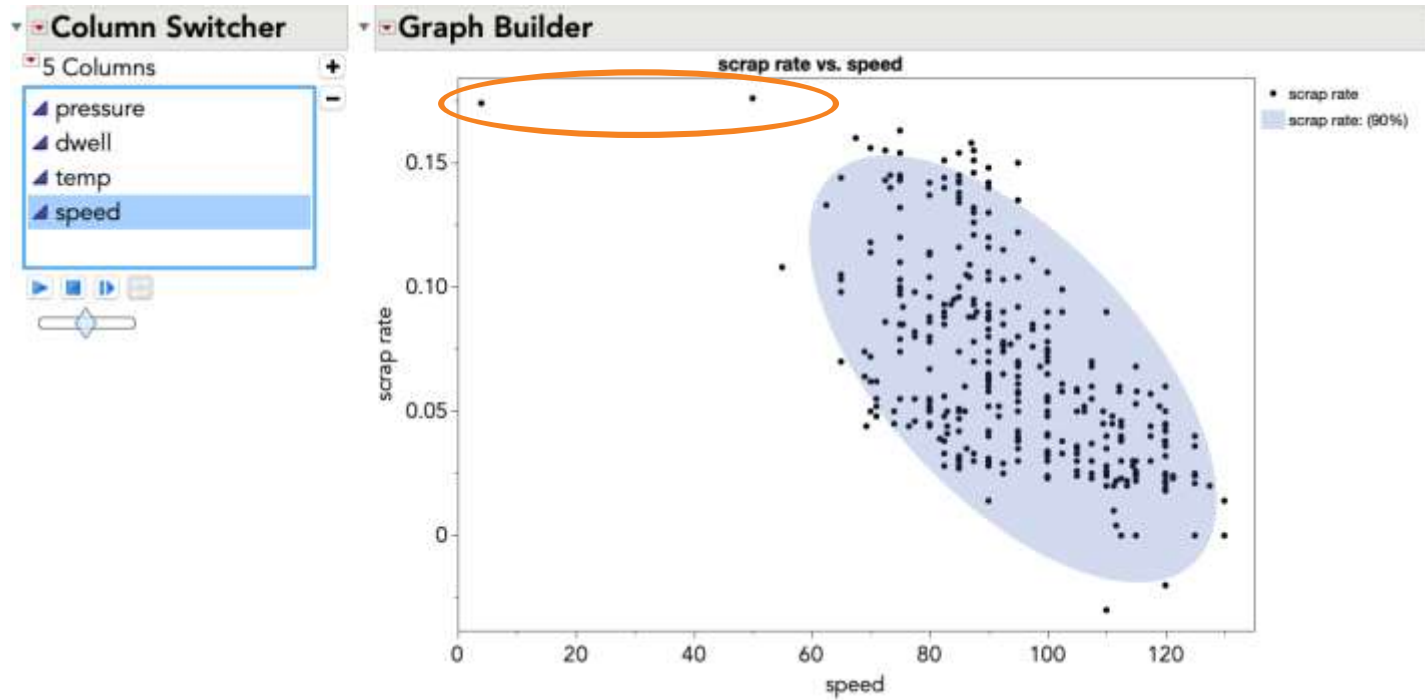


1. Scan the data table for obvious issues
2. Explore data one variable at a time
3. Explore data two or more variables at a time

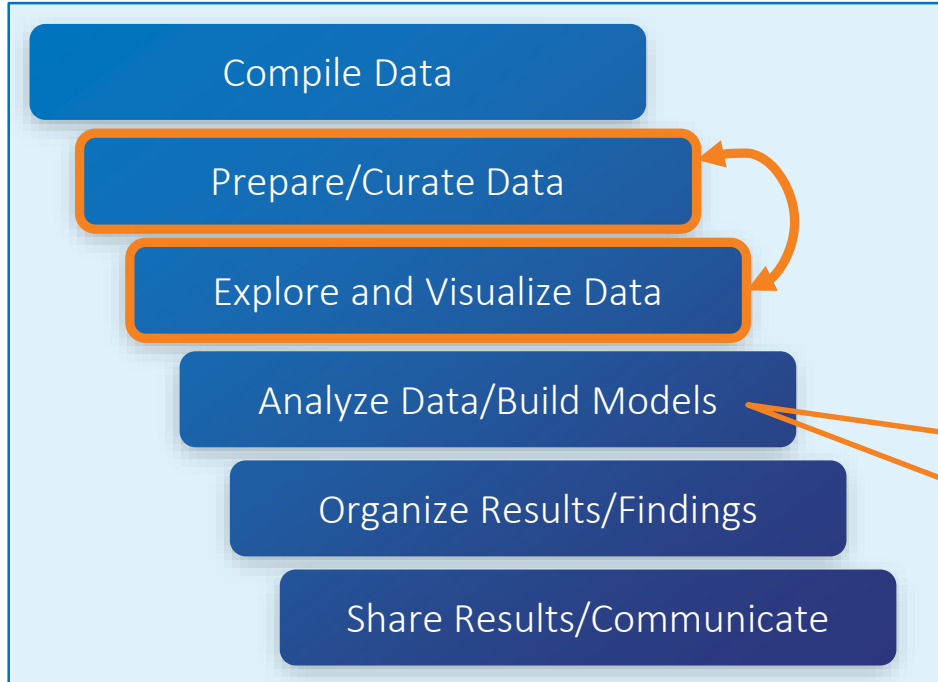
Graph Builder



Graph Builder with Column Switcher



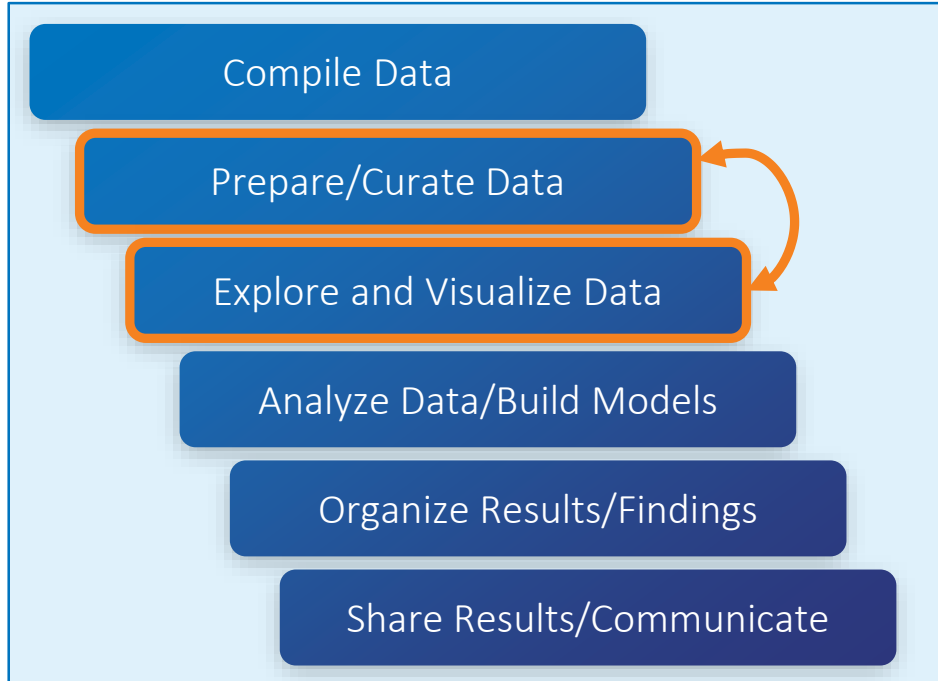
How Do You Identify Potential Issues?



1. Scan the data table for obvious issues
2. Explore data one variable at a time
3. Explore data two or more variables at a time

- More advanced tools (Explore Outliers/Missing)
- You will identify more issues here.

How Do You Identify Potential Issues?



1. Scan the data table for obvious issues
 2. Explore data one variable at a time
 3. Explore data two or more variables at a time
- Make notes of issues
 - Reshape/clean data as you go along
 - ***Capture curation steps for reproducibility***

Reproducible Data Curation

The Need for Reproducibility

- Many data sets get updated regularly with new data
- The same data curation steps need to be performed after each update
- Rather than point-and-click for each update, it is more efficient to have a script that performs all the curation steps
- Curation scripts also document the process steps
- Goal: Show how to generate a reproducible data curation script by point-and-click

Reproducible Data Curation

How do we obtain a script to address the issues we've identified, so we can reproduce as needed?

- For some curation tasks, JSL code can be generated directly by point-and-click in JMP 15:
 - Recode
 - New Formula Column
 - Stack, Split, Join, Concatenate, Update

Reproducible Data Curation

How do we obtain a script to address the issues we've identified, so we can reproduce as needed?

- Other tasks do not generate their own JSL code in JMP 15.
- The [Data Cleaning Script Assistant](#) Add-in helps with:
 - Changing Data Types, Modeling Types, and Column Formats
 - Rename, Reorder, and Delete Columns
 - Setting Column Properties, such as spec limits and value labels

Reproducible Data Curation

How do we obtain a script to address the issues we've identified, so we can reproduce as needed?

- New in JMP 16:
 - Action recording and enhanced log mode
 - The JSL code for most data curation activities will be generated and logged automatically

JMP 15 Cheat Sheet

Task	Point and Click Tool	Getting the script
Fix Character Data Values	Recode	Red Triangle
Reshape Data Tables	Stack, Split, Sort, Join, Concatenate, Update	Source Script
Subset Rows	Row Selection > Select Where ; Subset	Red Triangle ; Script Assistant > Subset Selected Rows
Transformations and other Derived Variables	Formula Editor or Create New Formula Column	Right Click > Copy Column
Reorder Columns	Data Table	Script Assistant > Victor
Delete Columns	Data Table	Script Assistant > Victor
Rename Variables	Column Info Window	Script Assistant > Victor
Change Data Type	Column Info Window	Script Assistant > Victor
Change Modeling Type	Column Info Window	Script Assistant > Victor
Change Display Format	Column Info Window	Script Assistant > Format Copier
Set Column Properties (Value Ordering et al.)	Column Info Window	Script Assistant > Property Copier

*Not an exhaustive list

Demo



JMP 15 Cheat Sheet

Task	Point and Click Tool	Getting the script
Fix Character Data Values	Recode	Red Triangle
Reshape Data Tables	Stack, Split, Sort, Join, Concatenate, Update	Source Script
Subset Rows	Row Selection > Select Where ; Subset	Red Triangle ; Script Assistant > Subset Selected Rows
Transformations and other Derived Variables	Formula Editor or Create New Formula Column	Right Click > Copy Column
Reorder Columns	Data Table	Script Assistant > Victor
Delete Columns	Data Table	Script Assistant > Victor
Rename Variables	Column Info Window	Script Assistant > Victor
Change Data Type	Column Info Window	Script Assistant > Victor
Change Modeling Type	Column Info Window	Script Assistant > Victor
Change Display Format	Column Info Window	Script Assistant > Format Copier
Set Column Properties (Value Ordering et al.)	Column Info Window	Script Assistant > Property Copier

*Not an exhaustive list

Summary

- Data curation begins with an exploratory and iterative approach to identifying problems
- Automation of the data curation workflow saves time and enables reproducibility
- For automation, the sequence of data operations is translated from point-and-click to a JSL script
- In JMP 15, use both built-in tools and the [Data Cleaning Script Assistant Add-in](#) to obtain the JSL script
- In JMP 16, curation scripts are captured automatically in the Enhanced Log

Thank you

