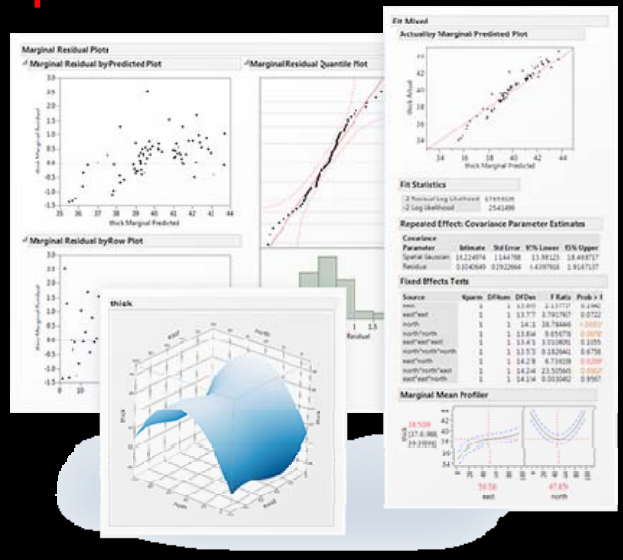
Data Exploration

What is JMP?

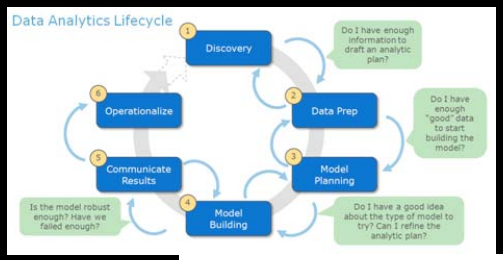
- Statistical Discovery Software from SAS
- Developed in 1989
- Comprehensive
  - Basic statistics and graphical summaries
  - Advanced tools and modeling techniques
- Extendible
  - Easy data import
  - Excel, R, Matlab, and SAS
- Tool for teaching, learning and doing
- *Visual, dynamic and interactive*

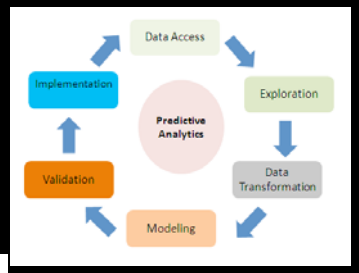


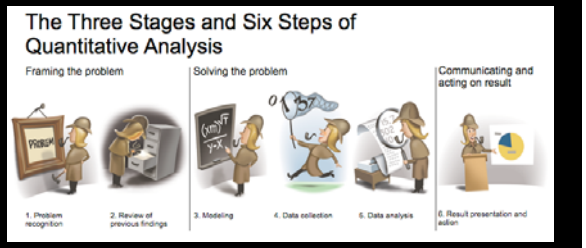
Courtesy of Mia Stephens: DSI 2014 Presentation

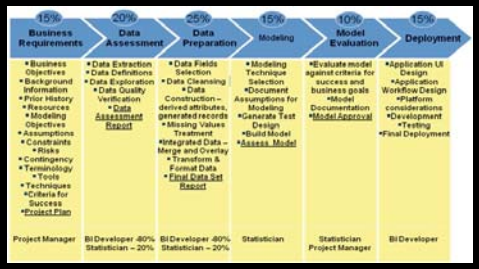
“the” analytics process

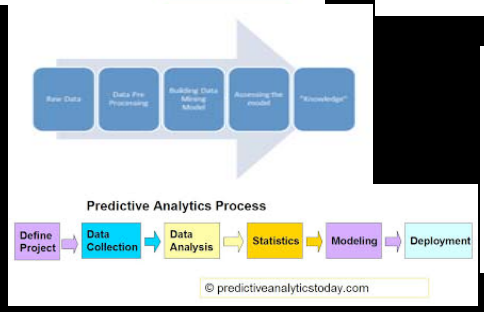
WILL THE REAL PROCESS PLEASE STAND UP

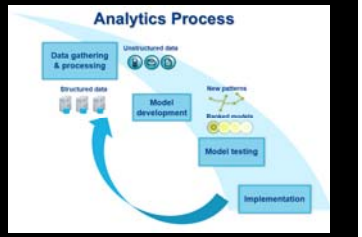


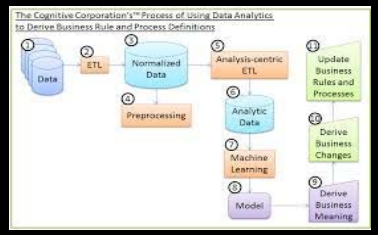


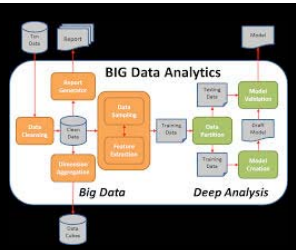


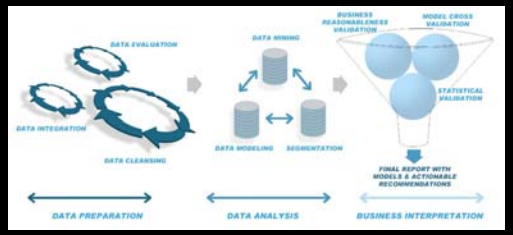


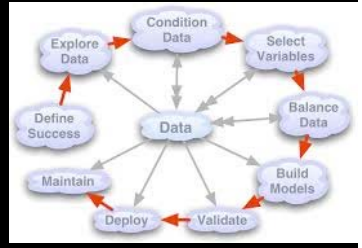












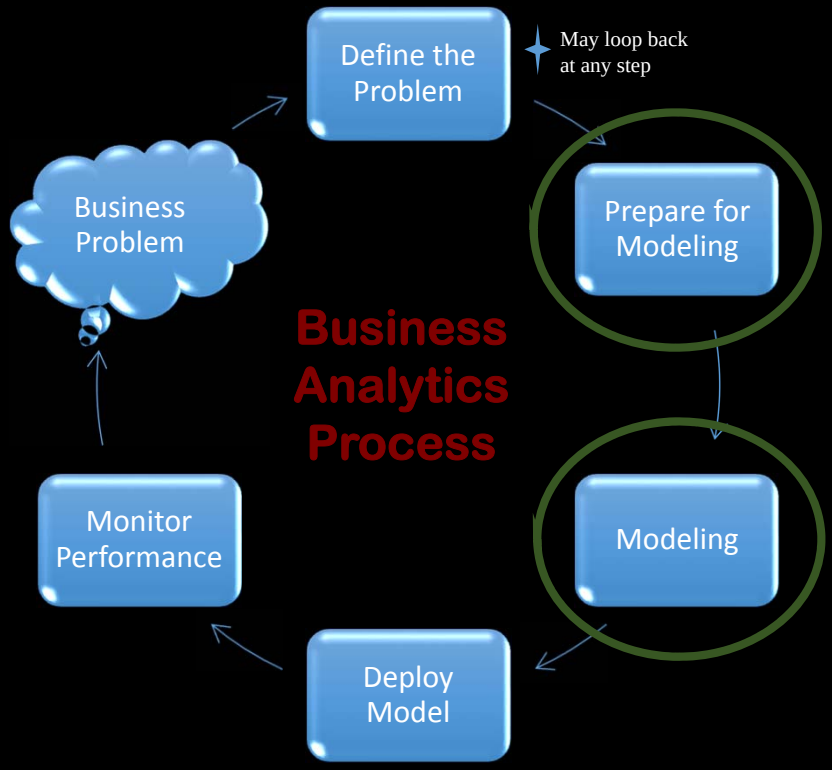
Courtesy of Mia Stephens: DSI 2014 Presentation

analytics

THE "GOOD" STUFF

What gets the most attention in the classroom?

Which topics do we spend the most time on?

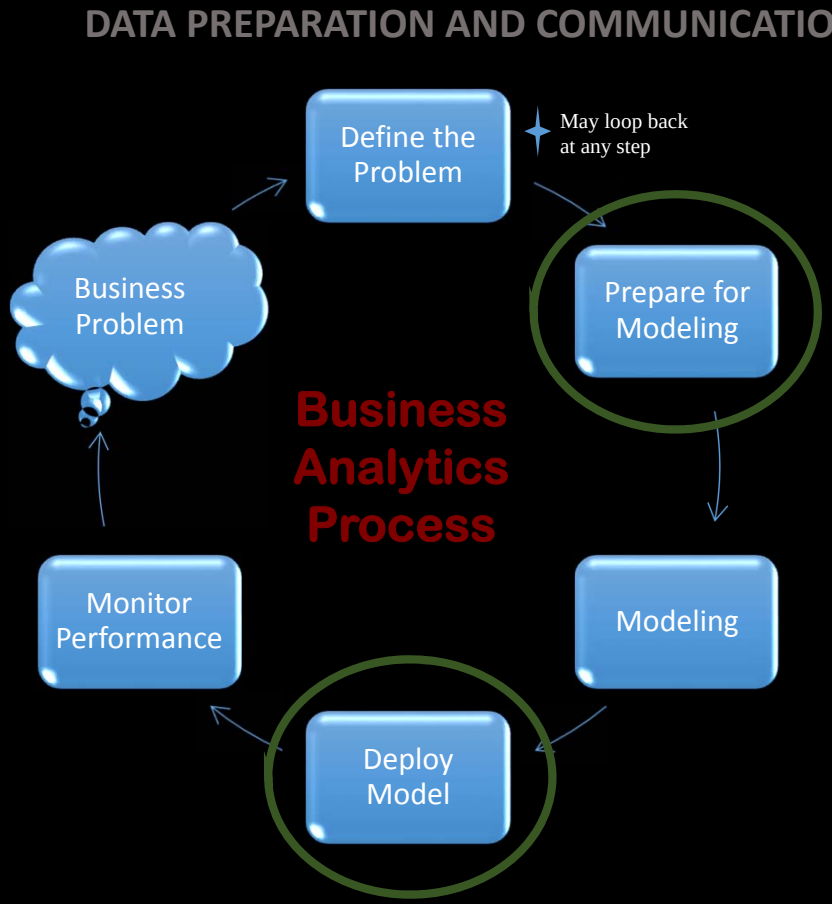


What is the most time consuming in practice?

From *Building Better Models with JMP Pro*, Grayson, Gardner and Stephens, 2015.

Courtesy of Mia Stephens: DSI 2014 Presentation

The “good” stuff



Data Preparation:

- Assess data quality
- Clean and transform data
- Examine and understand data
- Define features
- Create training, validation and test sets

Communication:

- Visualize results
- Interact with models
- Explore “what if” scenarios

Courtesy of Mia Stephens: DSI 2014 Presentation

Features in JMP

FOR DATA PREP AND COMMUNICATING RESULTS

- Columns Viewer
- Distribution
- Graph builder
- Data filter
- Missing Data Pattern and Tree Map
- Fit Y by X
- Recode
- Interactive Binning Add-In
- Dynamic Transformations
- Bubble plots
- Profiler
- Applications – one click
- HTML5

Courtesy of Mia Stephens: DSI 2014 Presentation

Exploratory Analysis

Working with Data

a. JMP Basics
b. Examining and Understanding Your Data
   - One column
   - Two columns
   - Many columns (Graph Builder)
c. Common Problems
   - Missing data
   - Messy data
   - Data in wrong form (from excel, data types, need to derive or transform)
   - Working with dates
   - Joining tables? Getting your data together for modeling!
d. Dimension reduction methods
   - Binning Data
   - Clustering Data
   - Principle Components Analysis and Variable Clustering
d. Exercises and Further Reading

From DRAFT *Building Better Models with JMP Pro*, Grayson, Gardner and Stephens, 2015.

Courtesy of Mia Stephens: DSI 2014 Presentation

Introduction

In this section we introduce some tools for working with and exploring data:

- Distribution – always a good starting point!
- Tabulate – for summarizing data
- Graph Builder – drag and drop, GUI graphing interface
- Mapping – create geographic maps with Graph Builder
- Data Filter – dynamically stratify data
- Tree Map – an alternative to Mosaic Plots
- Partition – use for exploring potentially important variables. Nice when there are many variables, nominal variables, missing values, and messy data.

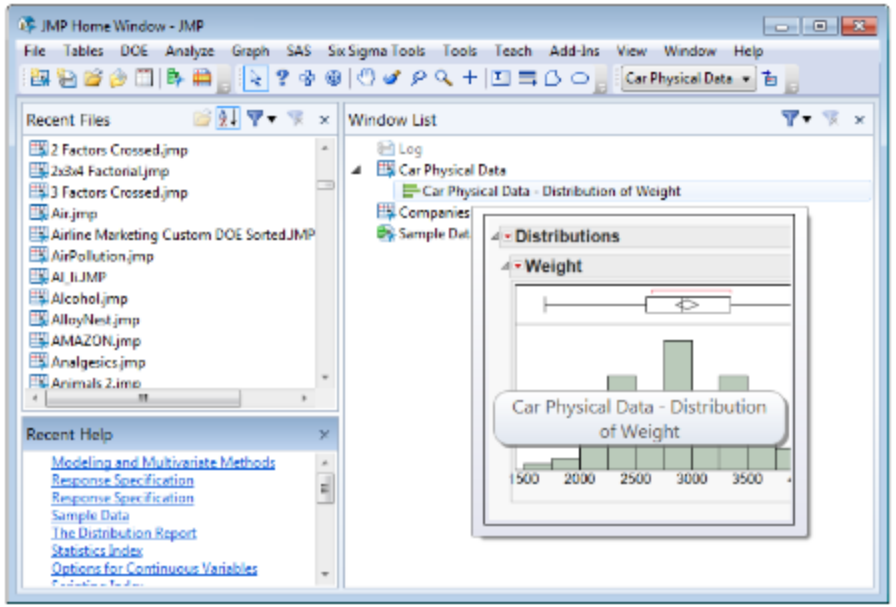
Courtesy of Mia Stephens: DSI 2014 Presentation

Navigating JMP® in Windows

This page gives information on the Windows JMP interface. For information on creating a new data table, opening data tables, and finding help within JMP see the page **Opening JMP and Getting Started**.

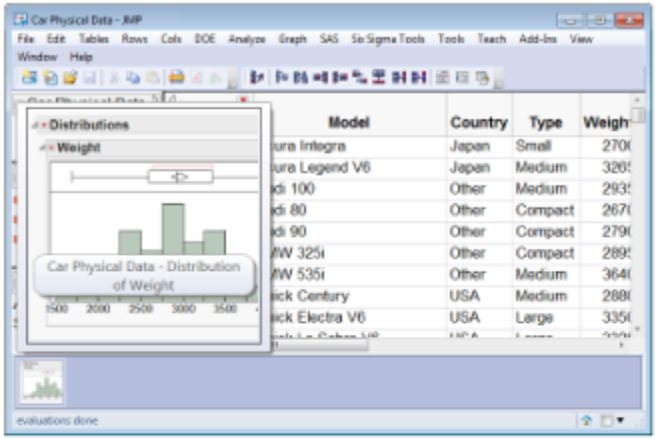
The Home Window

- When you first open JMP, you'll see the **Tip of the Day** window.
- You'll also see the new **JMP Home Window** (to the right), which provides access to:
  - Menus and toolbars (top).
  - Recently used files and help (on the left).
  - All open data tables and windows (on the right). Hover over an item in the list for a preview.
- The JMP Starter, Home Window and other features can be accessed via the **View** menu.



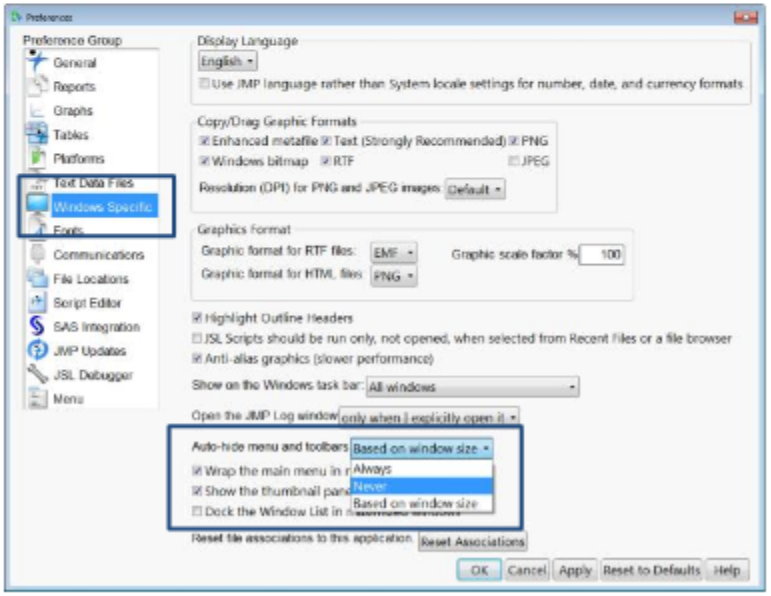
Navigation

- Data tables display thumbnails of open reports (bottom left). Hover over a thumbnail for a preview.
- Each data table and report provides icons (bottom, right corner) to facilitate navigation between windows:
  - Return to the Home Window (or, click CNTL – 1).
  - Go to the data table.
  - Select to arrange with other windows.

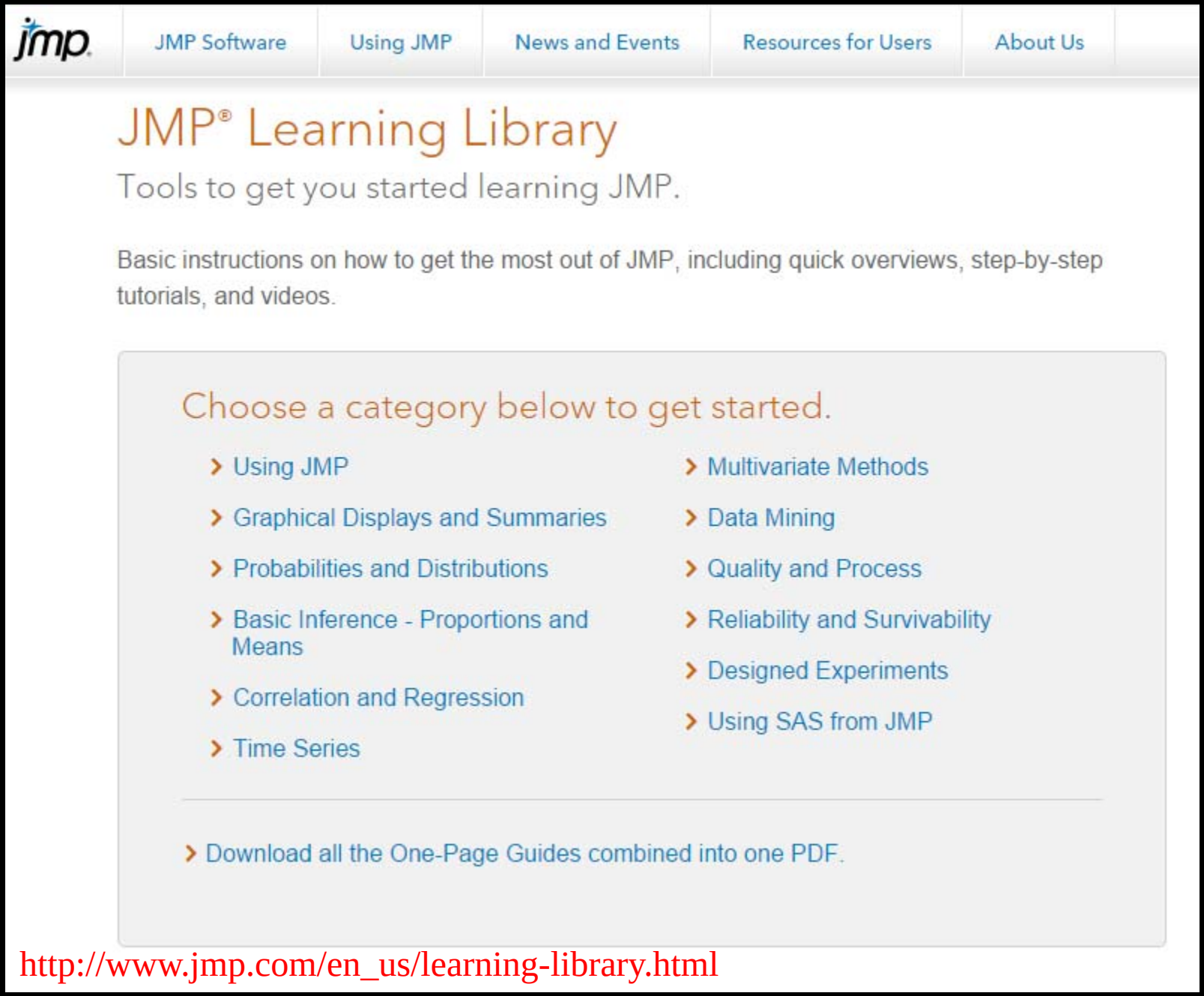


Tips:

- Each window has a JMP menu and toolbar.
- To view a hidden menu click **Alt**, or hover on the report window where the menu would normally appear.
- To have menus and toolbars always display, go to **File > Preferences > Windows Specific** and change **Auto-hide menus and toolbars** from **Based on window size** to **Never**.
- The JMP Home Window and Tip of the Day appear by default when JMP is opened. To change the default windows, use **File > Preferences > General**.

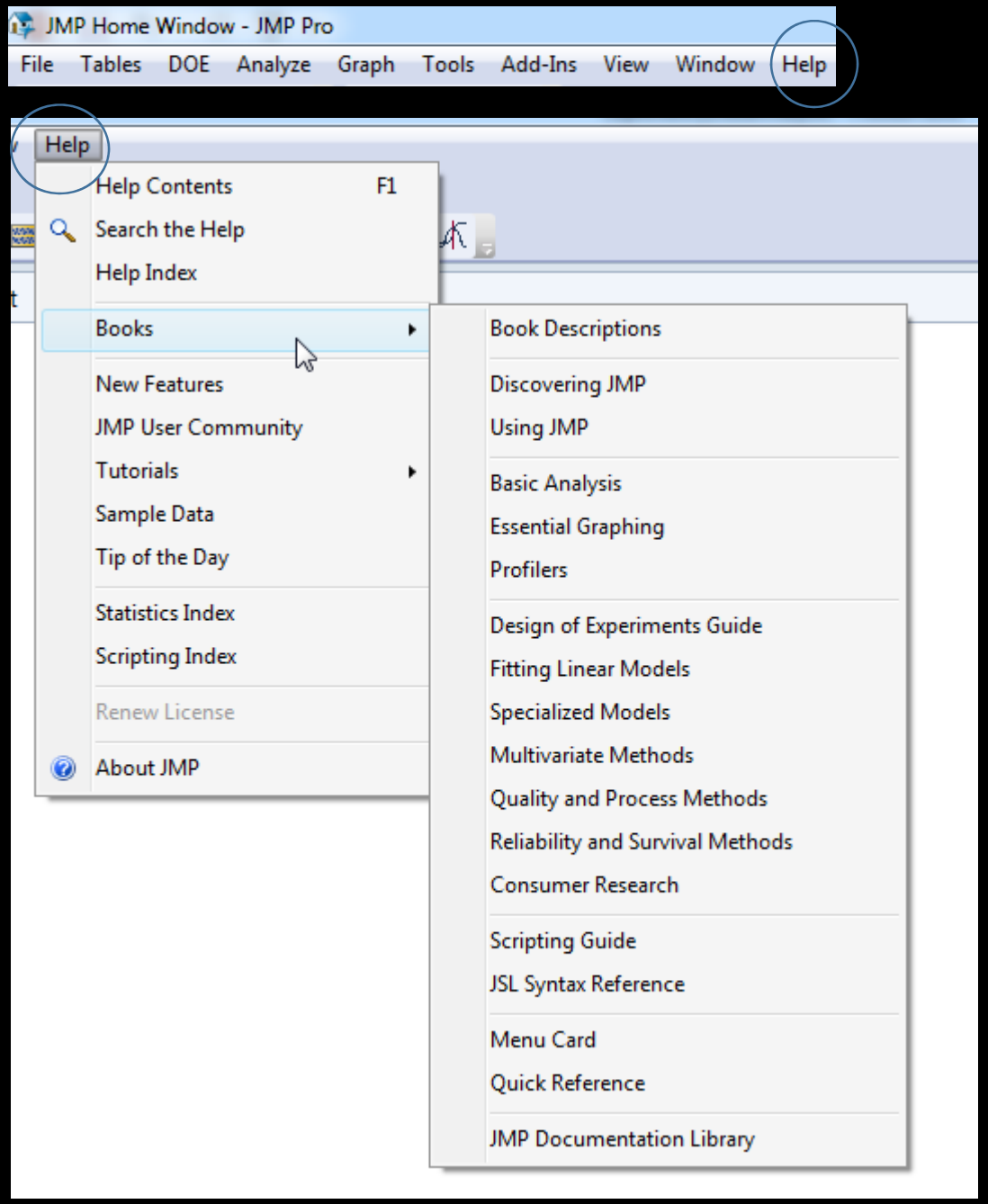


Note: For more details, see the book ***Using JMP*** (under **Help > Books**).



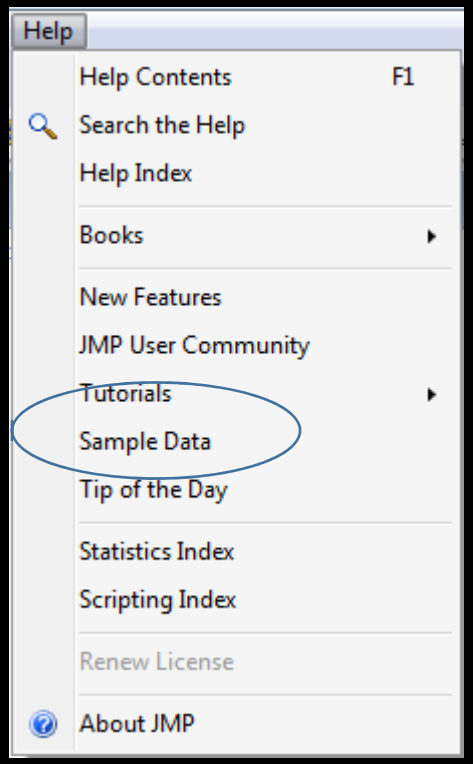
http://www.jmp.com/en_us/learning-library.html

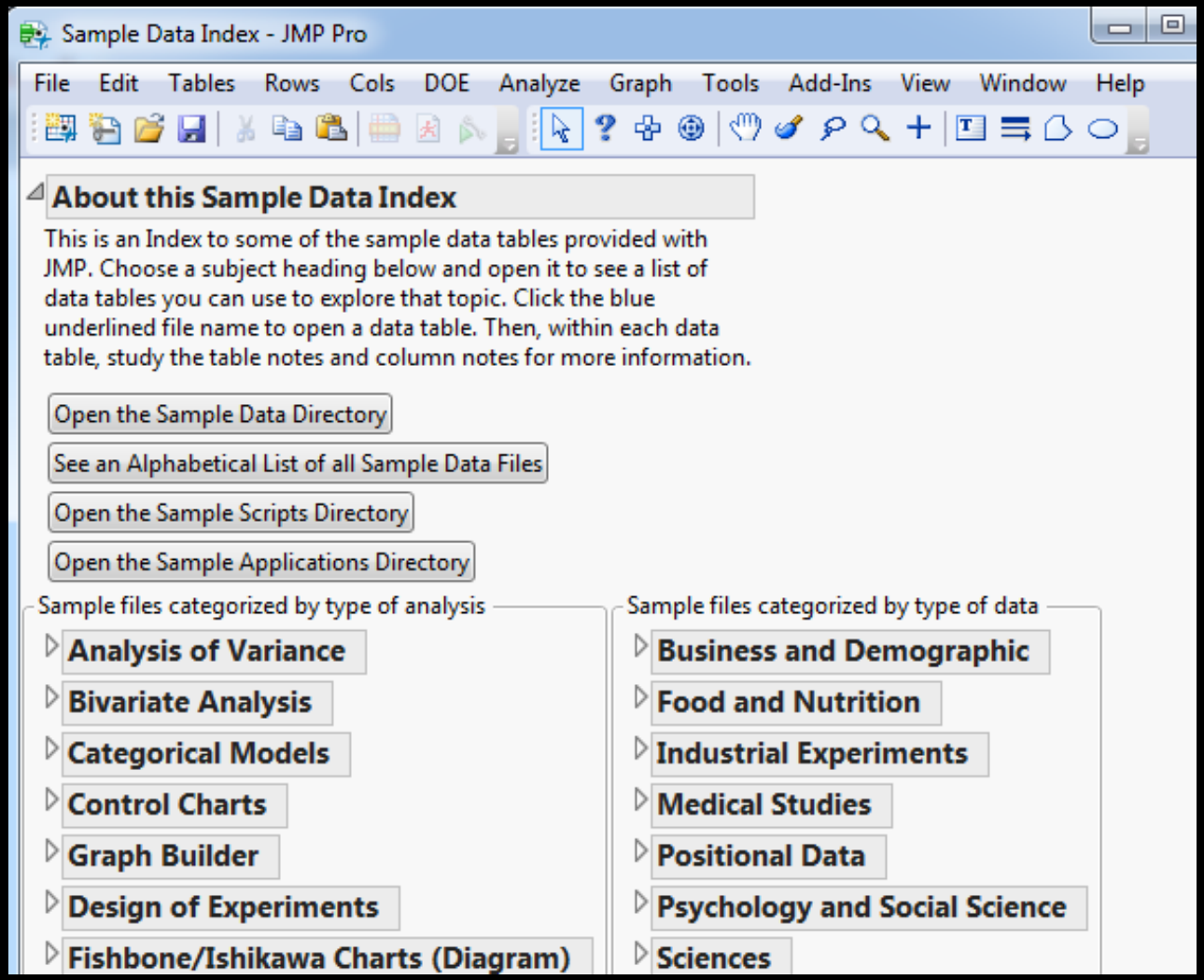
JMP Help



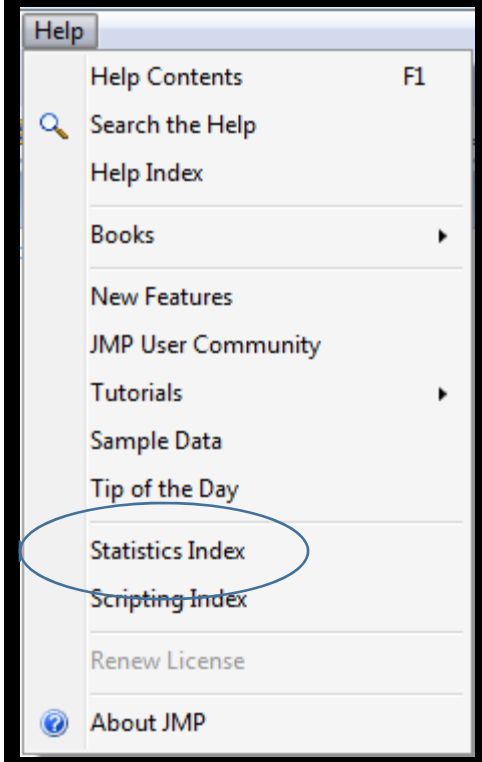
Courtesy of Mia Stephens: DSI 2014 Presentation

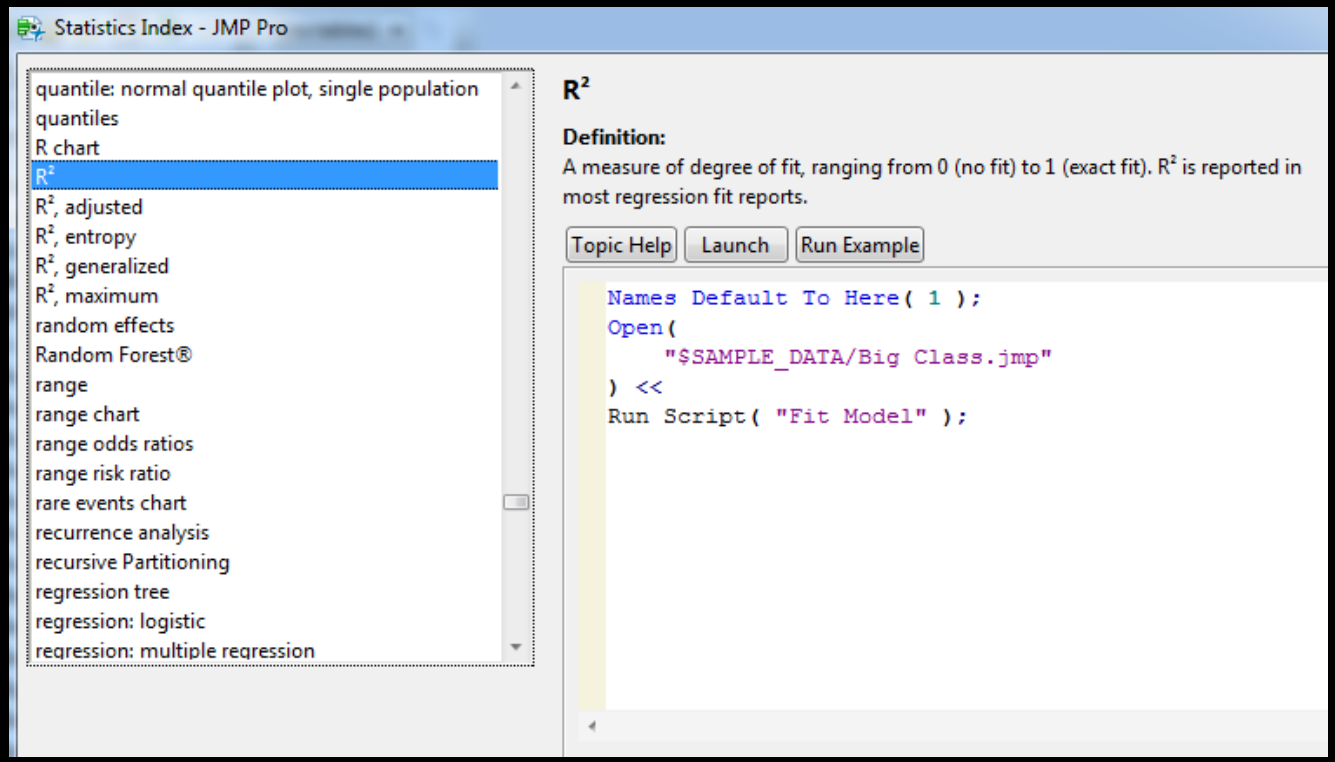
Sample Data


Help
Help Contents F1
Search the Help
Help Index
Books
New Features
JMP User Community
Tutorials
Sample Data
Tip of the Day
Statistics Index
Scripting Index
Renew License
About JMP


Sample Data Index - JMP Pro
File Edit Tables Rows Cols DOE Analyze Graph Tools Add-Ins View Window Help
About this Sample Data Index
This is an Index to some of the sample data tables provided with JMP. Choose a subject heading below and open it to see a list of data tables you can use to explore that topic. Click the blue underlined file name to open a data table. Then, within each data table, study the table notes and column notes for more information.
Open the Sample Data Directory
See an Alphabetical List of all Sample Data Files
Open the Sample Scripts Directory
Open the Sample Applications Directory
Sample files categorized by type of analysis
Analysis of Variance
Bivariate Analysis
Categorical Models
Control Charts
Graph Builder
Design of Experiments
Fishbone/Ishikawa Charts (Diagram)
Sample files categorized by type of data
Business and Demographic
Food and Nutrition
Industrial Experiments
Medical Studies
Positional Data
Psychology and Social Science
Sciences

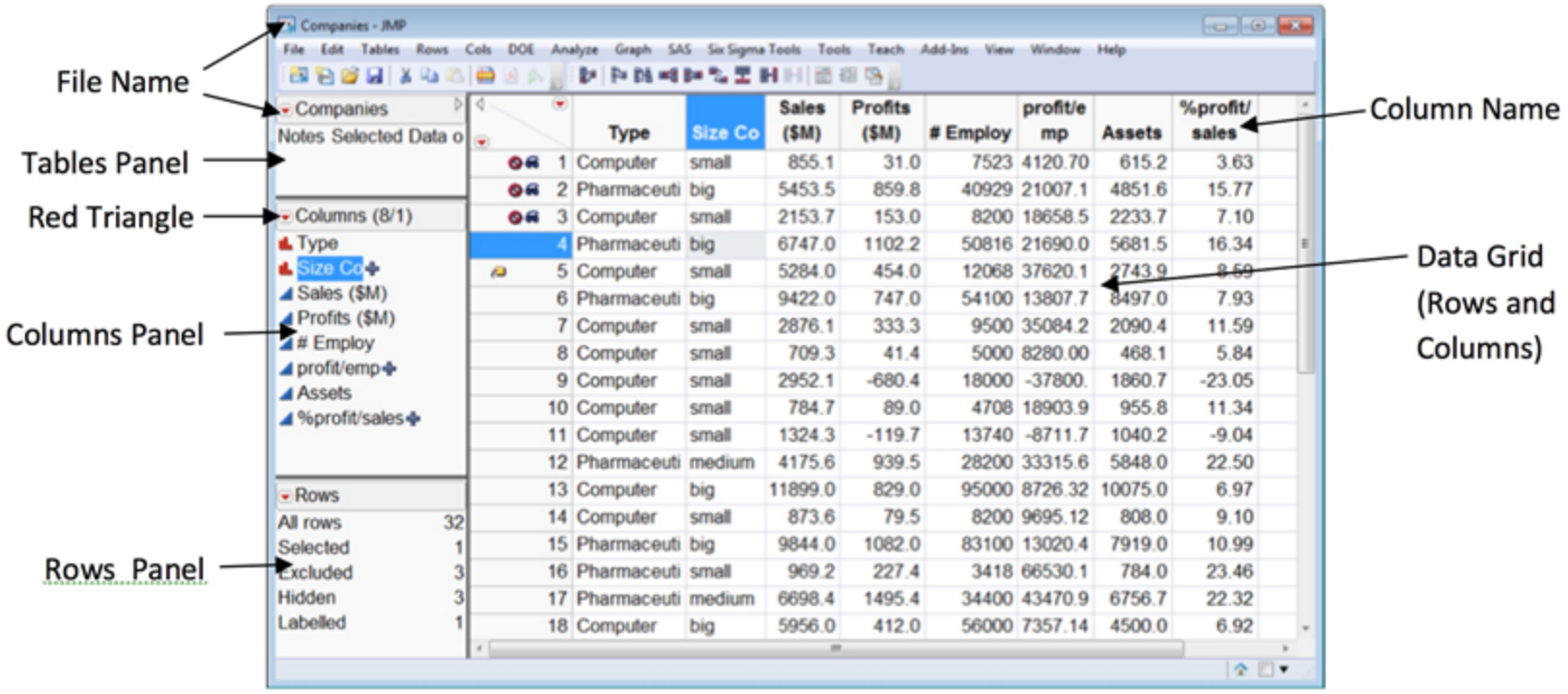
Statistics Index





Courtesy of Mia Stephens: DSI 2014 Presentation

Figure 3.1: Example Data Table in JMP



	Type	Size Co	Sales ($M)	Profits ($M)	# Employ	profit/emp	Assets	%profit/sales
1	Computer	small	855.1	31.0	7523	4120.70	615.2	3.63
2	Pharmaceuti	big	5453.5	859.8	40929	21007.1	4851.6	15.77
3	Computer	small	2153.7	153.0	8200	18658.5	2233.7	7.10
4	Pharmaceuti	big	6747.0	1102.2	50816	21690.0	5681.5	16.34
5	Computer	small	5284.0	454.0	12068	37620.1	2743.9	8.59
6	Pharmaceuti	big	9422.0	747.0	54100	13807.7	8497.0	7.93
7	Computer	small	2876.1	333.3	9500	35084.2	2090.4	11.59
8	Computer	small	709.3	41.4	5000	8280.00	468.1	5.84
9	Computer	small	2952.1	-680.4	18000	-37800.	1860.7	-23.05
10	Computer	small	784.7	89.0	4708	18903.9	955.8	11.34
11	Computer	small	1324.3	-119.7	13740	-8711.7	1040.2	-9.04
12	Pharmaceuti	medium	4175.6	939.5	28200	33315.6	5848.0	22.50
13	Computer	big	11899.0	829.0	95000	8726.32	10075.0	6.97
14	Computer	small	873.6	79.5	8200	9695.12	808.0	9.10
15	Pharmaceuti	big	9844.0	1082.0	83100	13020.4	7919.0	10.99
16	Pharmaceuti	small	969.2	227.4	3418	66530.1	784.0	23.46
17	Pharmaceuti	medium	6698.4	1495.4	34400	43470.9	6756.7	22.32
18	Computer	big	5956.0	412.0	56000	7357.14	4500.0	6.92

Data sets in JMP, called *data tables*, consist of a data grid with data (on the right) and panels (on the left) that define the data and the variables.

An example data table, **Companies.jmp**, is shown in Figure 3.1 (figure borrowed from the one page guide, JMP Data Tables). This data table, and others we'll use in this book, are available in the sample data directory (under **Help > Sample Data**).

Historical data were gathered on 5960 bank customers to determine whether a customer is a good or bad credit risk for a home equity loan. The data are stored in the file **Equity.jmp** in the sample data directory (under **Help > Sample Data**).

The response (or Y) variable is BAD, which is coded as 0 (good risk) or 1 (bad risk). The other variables are:

LOAN: The amount of the loan

MORTDUE: How much they need to pay on their mortgage

VALUE: Assessed valuation

REASON: Debt consolidation or home improvement (DebtCon or HomeImp)

JOB: Broad job category

YOJ: Years on the job

DEROG: Number of derogatory reports

DELINQ: The number of delinquent trade lines

CLAGE: Age of oldest trade line

NINQ: Number of recent credit inquiries

CLNO: Number of trade lines

DEBTINC: Debt to income ratio

Variables are coded in JMP with a Continuous, Ordinal or Nominal modeling type. This coding helps to make sure that JMP performs the correct analysis and produces appropriate graphs.

- Continuous variables, like **LOAN** and **MORTDUE**, have numeric values (e.g.; 2, 5, 3.35, 159.667,…). These variables must be in decimal format, an can have no special symbols or text.
- Nominal variables, like **BAD**, **REASON** and **JOB**, can have either numeric or character values, and represent unordered categories or labels (e.g.; the names of states, colors of M&Ms, machine numbers,…).
- Ordinal variables, which can also have either numeric or character values, represent ordered categories (e.g.; small, medium and large; 1-9 severity rating scales,…). This data set has no ordinal variables.

A first step in any analysis is to ensure that your variables have the correct modeling type. To change the modeling type, click on the icon next to the variable in the data table or in any analysis. JMP also allows for a number of data types in addition to modeling types. These can be changed using the **Column Info** window (right-click on a column name and select Column Info). Note that there are a number of column properties that can be specified from this window, including notes and formulas.

Exploring Data One Variable at a Time

We start by looking at the variables one at a time. We use summary statistics and graphical summaries to get familiar with our data and at the same time identify any potential issues.

For continuous variables, we're interested in summary statistics, such as the mean (the average), the standard deviation (the spread), minimum and maximum values, and the number of missing observations. We're also interested in the shape of the distribution. Is the distribution more or less symmetric? Is it skewed? Are there clusters of data or severe outliers?

For categorical data (nominal or ordinal modeling types) we're interested in the number of categories (or levels), the number of observations in each category, and the number of missing observations.

Two key tools in JMP for exploring variables one at a time are **Columns Viewer** and **Distribution**.

From DRAFT *Building Better Models with JMP Pro*, Grayson, Gardner and Stephens, 2015.

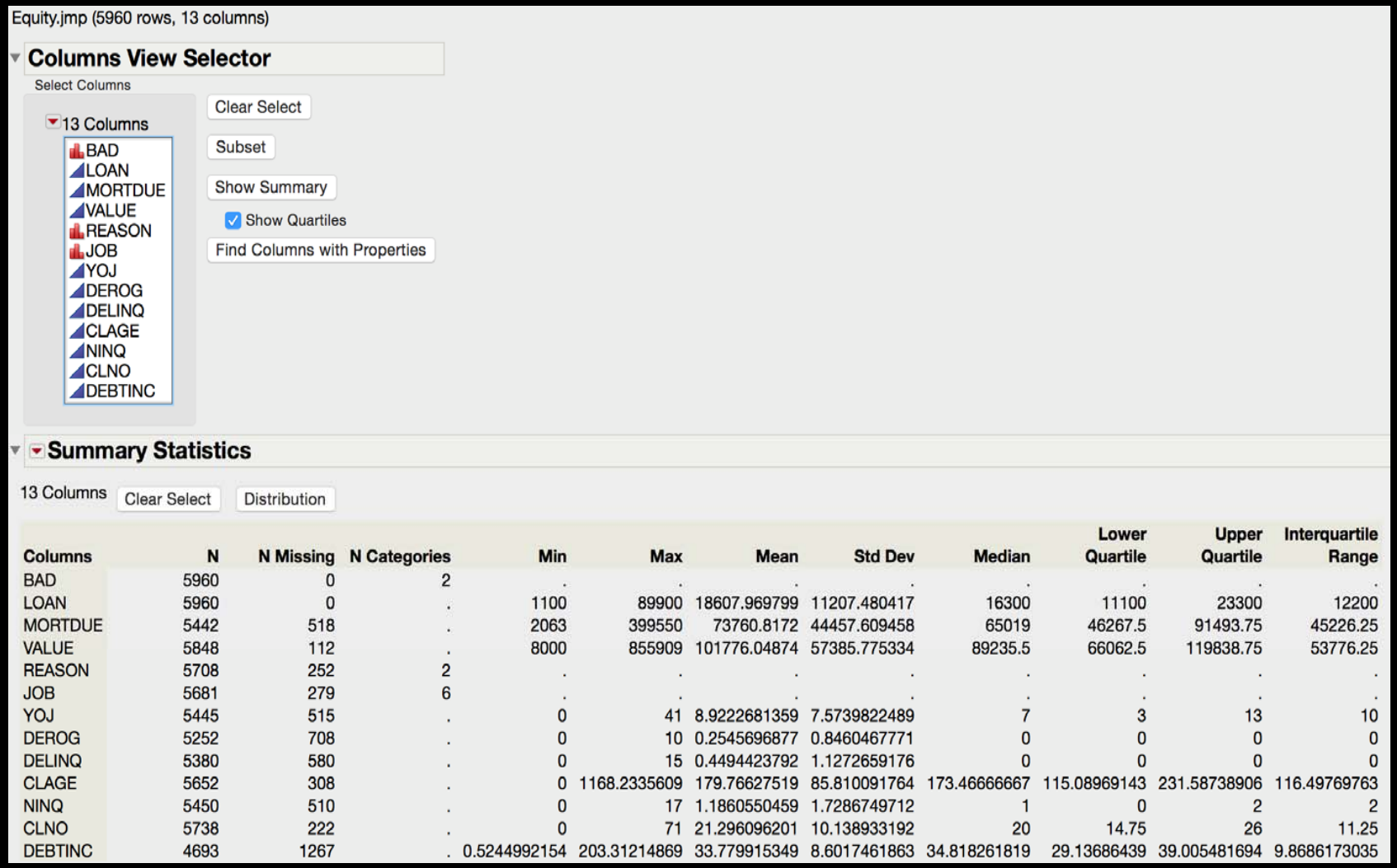
The Columns Viewer provides numeric summaries of our data, as shown in Figure 3.2, and is particularly useful with large data sets (with many variables). (Go to **Cols** > **Columns** Viewer. Select all of the variables from the **Select Columns** list, click **Show Quartiles**, then click **Show Summary**.)

For the three nominal variables, **BAD**, **REASON** and **JOB** we see the number of observations (N), the number of missing values (N **Missing**), and the number of categories (N **Categories**).

For the continuous variables we also see N and N Missing, plus a number of summary statistics. Of potential concern is the number of missing records, particularly for DEROG and DEBTINC.

The Min and Max are the range of values for the variables. Many of the continous variable have a minimum value of 0, and some of these have medians and quartiles that are also 0. This may or may not be an issue, but it should be investigated. The mean is the average value, while the median is the middle value (the 50th percentile). Big differences between these two values would be an indication of potential skewness.

Figure 3.2: Columns Viewer, Equity Data



Columns	N	N Missing	N Categories	Min	Max	Mean	Std Dev	Median	Lower Quartile	Upper Quartile	Interquartile Range
BAD	5960	0	2	.	.	.	.	.	.	.	.
LOAN	5960	0	.	1100	89900	18607.969799	11207.480417	16300	11100	23300	12200
MORTDUE	5442	518	.	2063	399550	73760.8172	44457.609458	65019	46267.5	91493.75	45226.25
VALUE	5848	112	.	8000	855909	101776.04874	57385.775334	89235.5	66062.5	119838.75	53776.25
REASON	5708	252	2	.	.	.	.	.	.	.	.
JOB	5681	279	6	.	.	.	.	.	.	.	.
YOJ	5445	515	.	0	41	8.9222681359	7.5739822489	7	3	13	10
DEROG	5252	708	.	0	10	0.2545696877	0.8460467771	0	0	0	0
DELINQ	5380	580	.	0	15	0.4494423792	1.1272659176	0	0	0	0
CLAGE	5652	308	.	0	1168.2335609	179.76627519	85.810091764	173.46666667	115.08969143	231.58738906	116.49769763
NINQ	5450	510	.	0	17	1.1860550459	1.7286749712	1	0	2	2
CLNO	5738	222	.	0	71	21.296096201	10.138933192	20	14.75	26	11.25
DEBTINC	4693	1267	.	0.5244992154	203.31214869	33.779915349	8.6017461863	34.818261819	29.13686439	39.005481694	9.8686173035

From DRAFT *Building Better Models with JMP Pro*, Grayson, Gardner and Stephens, 2015.

The **Distribution** button (top, in Figure 3.2) launches the **Distribution** platform for selected variables. This platform is also available from the **Analyze** menu.

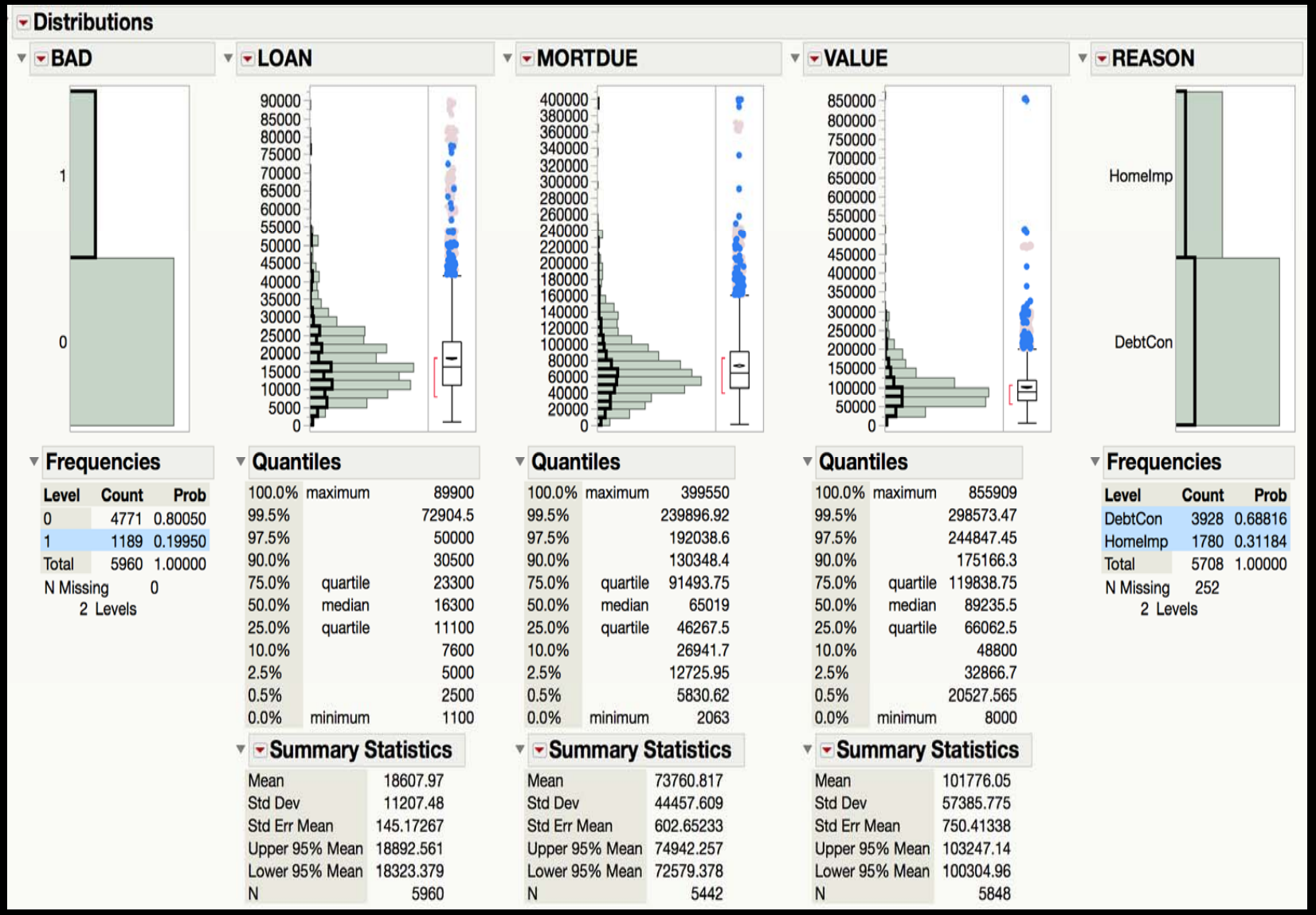
In Figure 3.3 we see distributions for the first five variables. (In **Analyze > Distribution**, select the variables and click **Y, Columns**, then click **OK**.) Categorical variables JMP produces bar charts and frequency distributions. For continuous variables JMP produces histograms, box plots, and summary statistics.

By default, the view is vertical. This view allows us to see distributions of more variables at one time. However, you can easily convert this to a horizontal view (click on the top red triangle and select **Stack**). To produce horizontal histograms from Distribution in the future, you can set a preference.

Back to the output in figure 3.3. Of the 5960 customers, nearly 20% were a bad credit risk. The mean loan amount is $18607, but we see some extreme amounts. In fact, we also see some extreme values for MORTDUE and VALUE – both of these variables appear right-skewed.

All of the graphs in JMP are dynamically linked to the data table and to every other graph. This allows us to explore potential relationships between variables. In Figure 3.3 we have clicked on the bar for bad risk customers (1). This highlights the bar, and we can see how the bad risk customers are distributed across the other variables. For example, we can see that more of the bad risk customers were consolidating debt than taking out a loan for home improvement.

Figure 3.3: Distribution Output, first 5 Variables



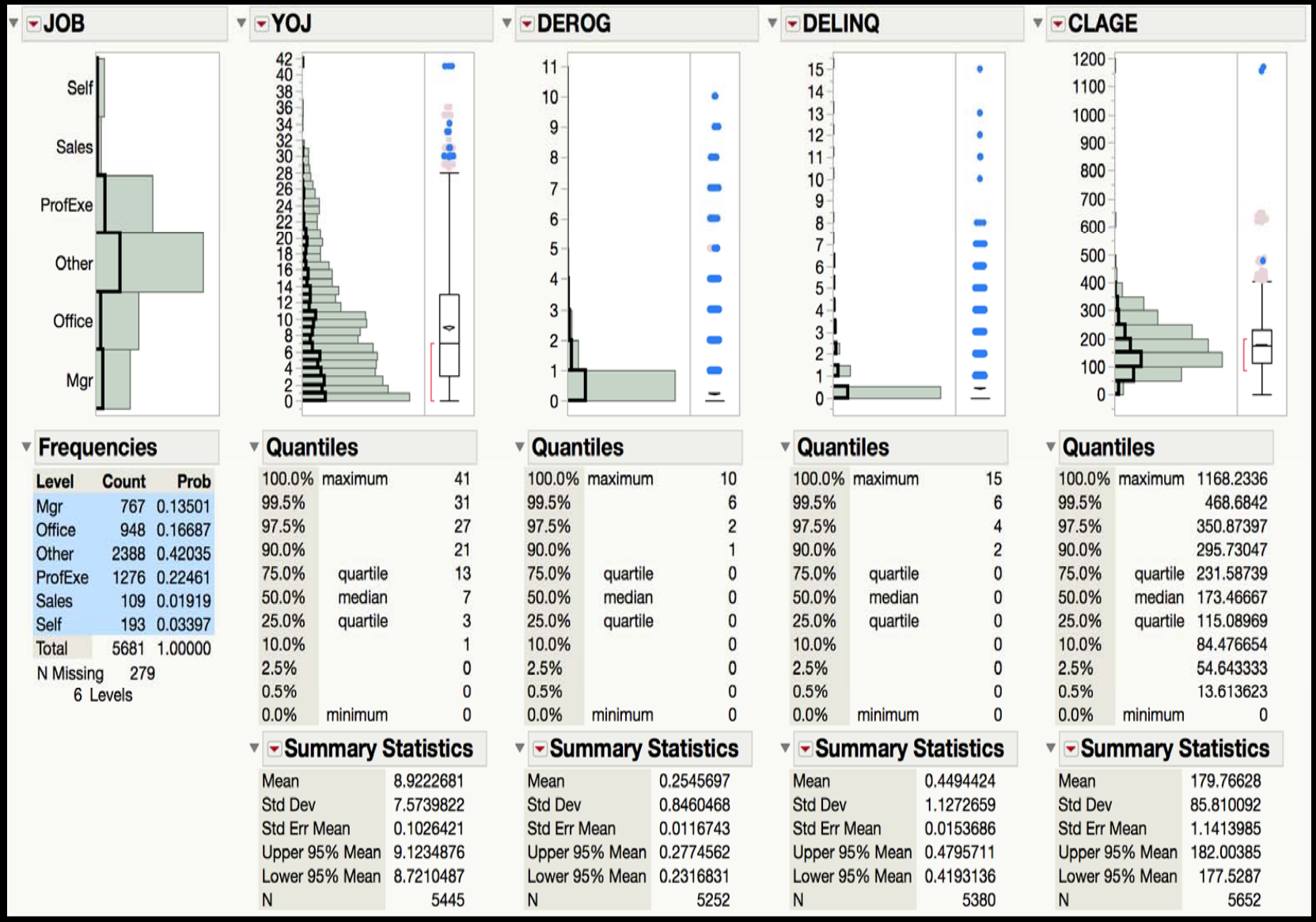
From DRAFT *Building Better Models with JMP Pro*, Grayson, Gardner and Stephens, 2015.

In Figure 3.4, we see the next 5 variables in the data set. Recall from Figure 3.2 that some of the continuous variables, like DEROG and DELINQ, had minimum values of zero. For these variables, we can clearly see the distributions using the histograms and quantiles. Most of the values are, in fact, zero.
In CLAGE we see another potential problem – 2 customers with loan ages near 1200 days (these can be seen in the box plot).

So far, we've identified four potential data quality issues:

- Missing values from many variables
- Skewed distributions, and a long right-tail in the distributions of some variables
- Messy data – Continuous variables with many zeros
- Outliers in CL Age

Figure 3.4: Distribution Output, Next 5 Variables



From DRAFT *Building Better Models with JMP Pro*, Grayson, Gardner and Stephens, 2015.

Exploring Data Two Variables at a Time

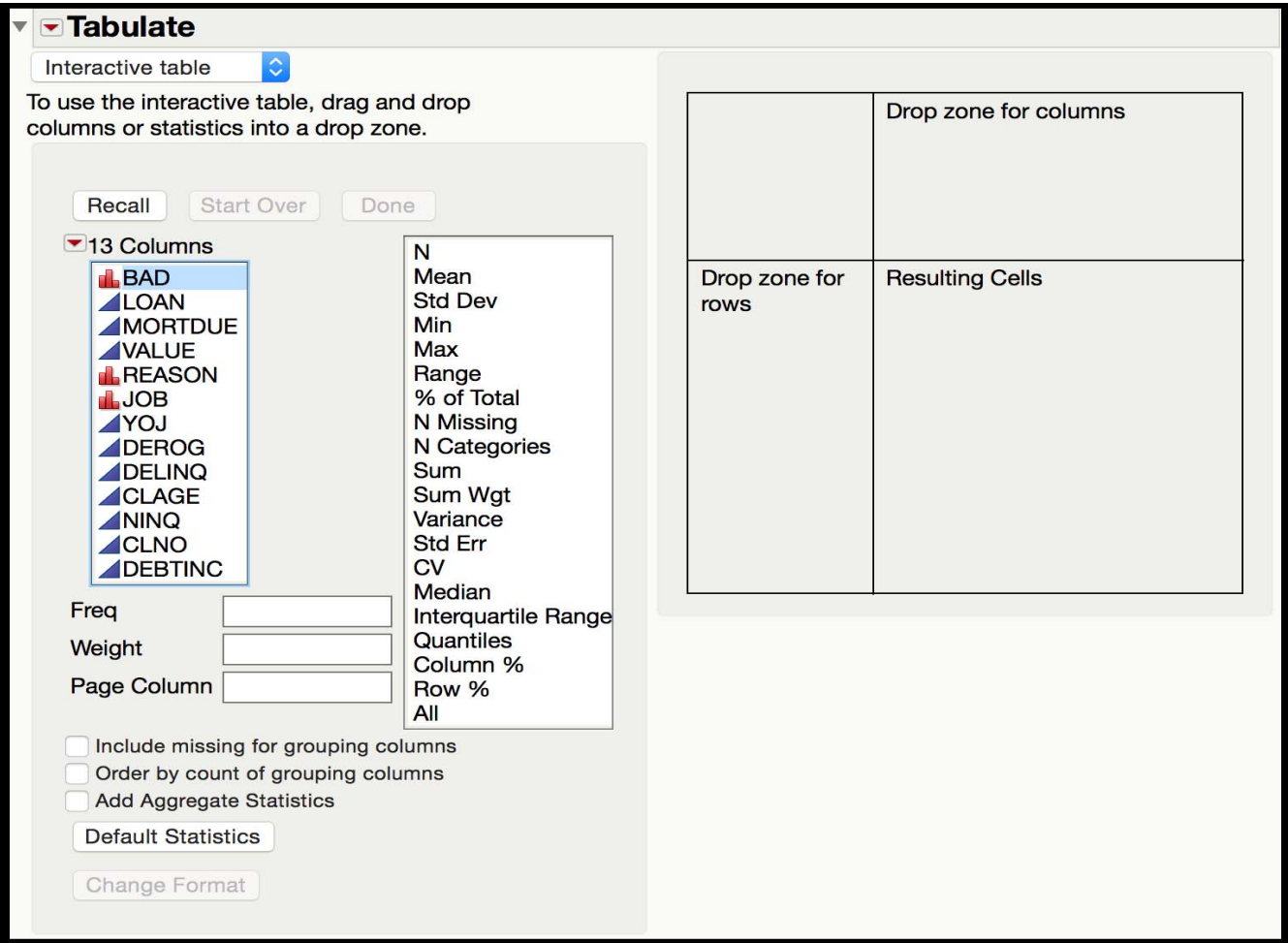
The dynamic linking allows us to start to get a feel for potential bivariate relationships between our variables. Three additional platforms for exploring variables two at a time are **Tabulate**, **Graph Builder**, and **Fit Y by X**.

From DRAFT *Building Better Models with JMP Pro*, Grayson, Gardner and Stephens, 2015.

Tabulate is a platform for dynamically summarizing data an constructing tables of descriptive statistics. The initial Tabulate window is shown in Figure 3.5.

Figure 3.5: Initial Tabulate Window

Drag variables from the columns list to the drop zone for columns and rows, then drag and drop statistic of interest into the results zone to add new statistics. Click the **Done** button to close the control panel and produce the final table.



In Figure 3.6 we see percentages for BAD for the two reasons and six job categories (drag **BAD** to **Drop Zone for Columns**, **REASON** to **Drop Zone for Rows**, drag **JOB** just below **REASON**, and then drag **Row %** to the results area). There are more loans for customers consolidating debt than home improvement, and two job categories, **Sales** and **Self**, appear to have higher risk of bad loans than the other categories. However, there are far fewer customers in these two job categories.

Figure 3.6: Tabular summary of BAD versus REASON and JOB

Tabulate

	BAD			
	0		1	
REASON	**N**	**Row %**	**N**	**Row %**
DebtCon	3183	81.03%	745	18.97%
HomeImp	1384	77.75%	396	22.25%
JOB				
Mgr	588	76.66%	179	23.34%
Office	823	86.81%	125	13.19%
Other	1834	76.80%	554	23.20%
ProfExe	1064	83.39%	212	16.61%
Sales	71	65.14%	38	34.86%
Self	135	69.95%	58	30.05%

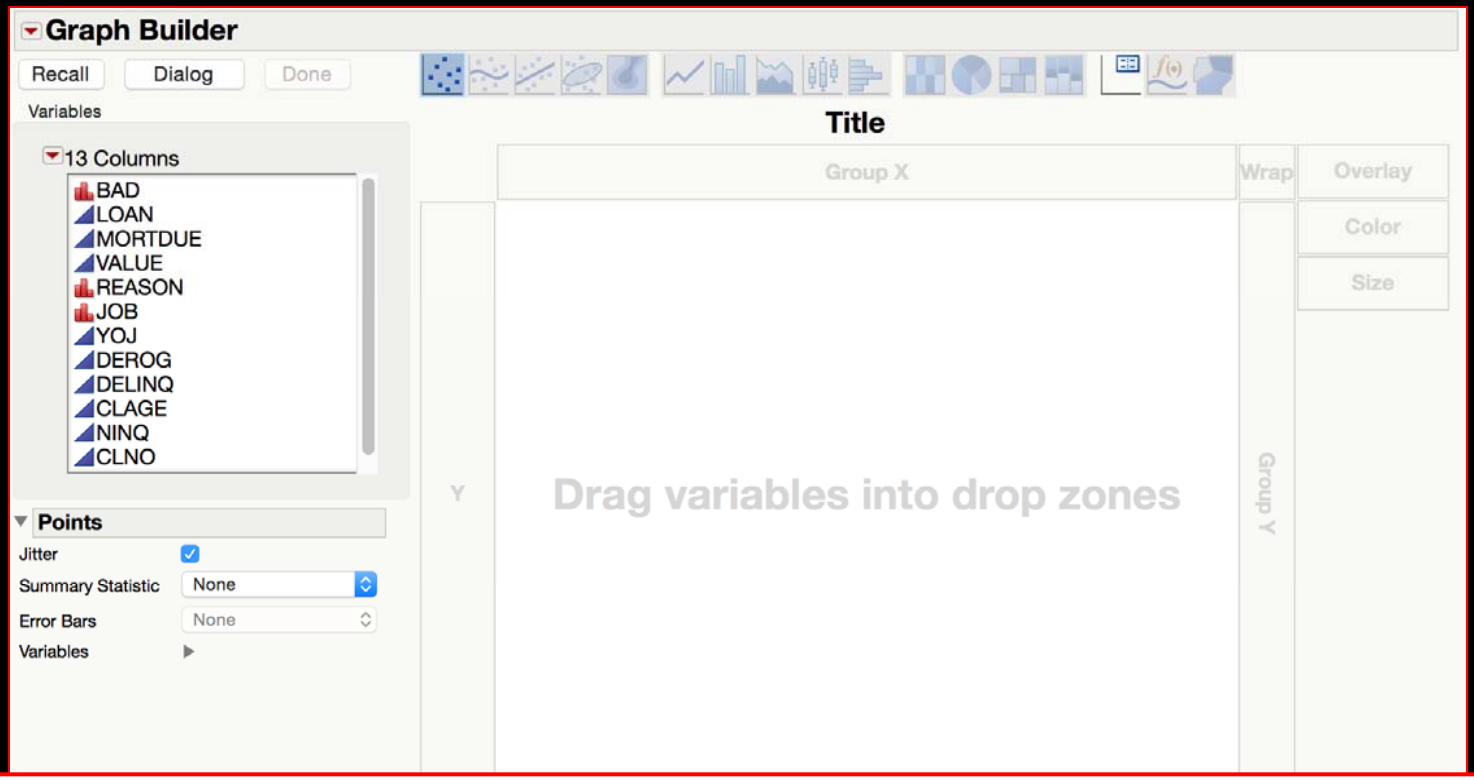
Figure 3.7: Means of Continuous Variables for Levels of BAD

Tabulate

		BAD	
		0	1
LOAN	Mean	19028.11	16922.12
MORTDUE	Mean	74829.25	69460.45
YOJ	Mean	9.15	8.03
DEROG	Mean	0.13	0.71
DELINQ	Mean	0.25	1.23
CLAGE	Mean	187.00	150.19
NINQ	Mean	1.03	1.78
CLNO	Mean	21.32	21.21
DEBTINC	Mean	33.25	39.39

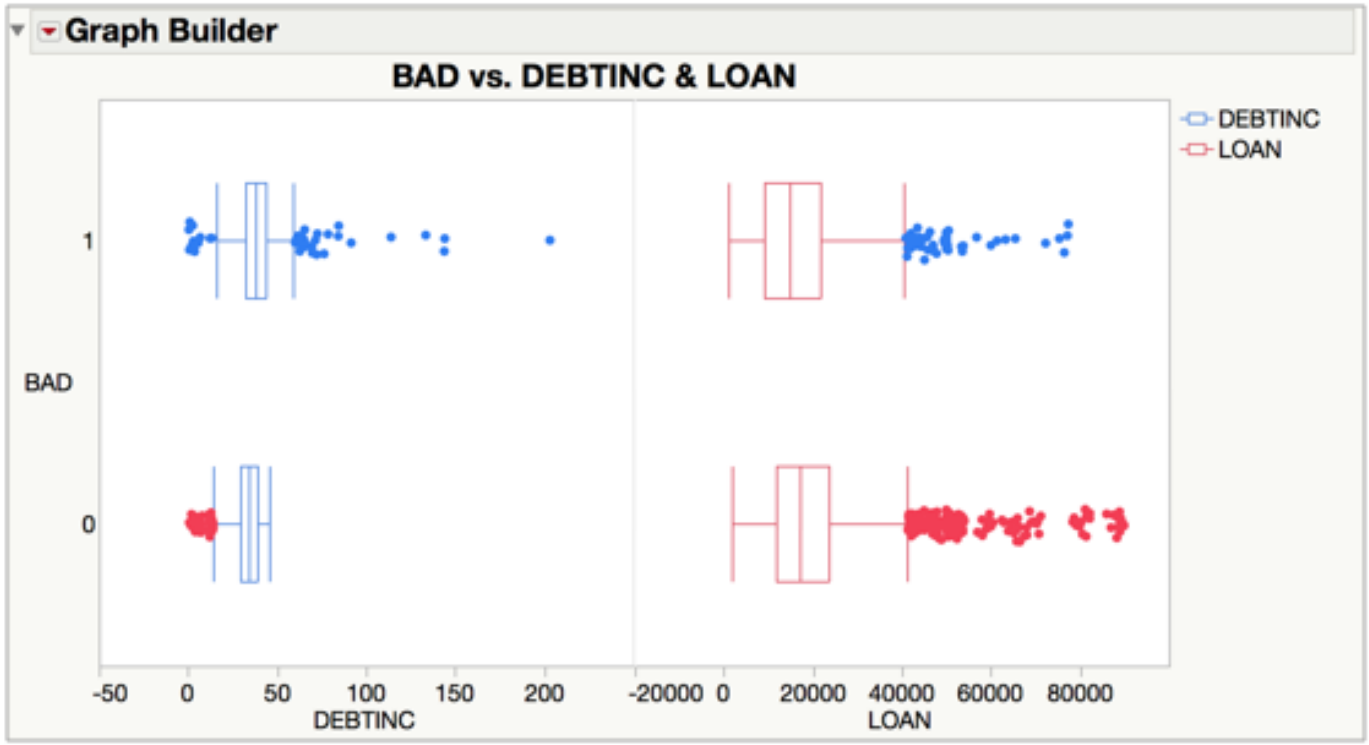
The means of the continuous variables for the two categories of **BAD** are shown in Figure 3.7. There appear to be some potential differences. For example, the loan amounts are higher on average for low risk loans, while the debt to income ratio is higher for high risk loans. These bivariate summaries highlight some potentially important variables.

Figure 3.8: Initial Graph Builder Window



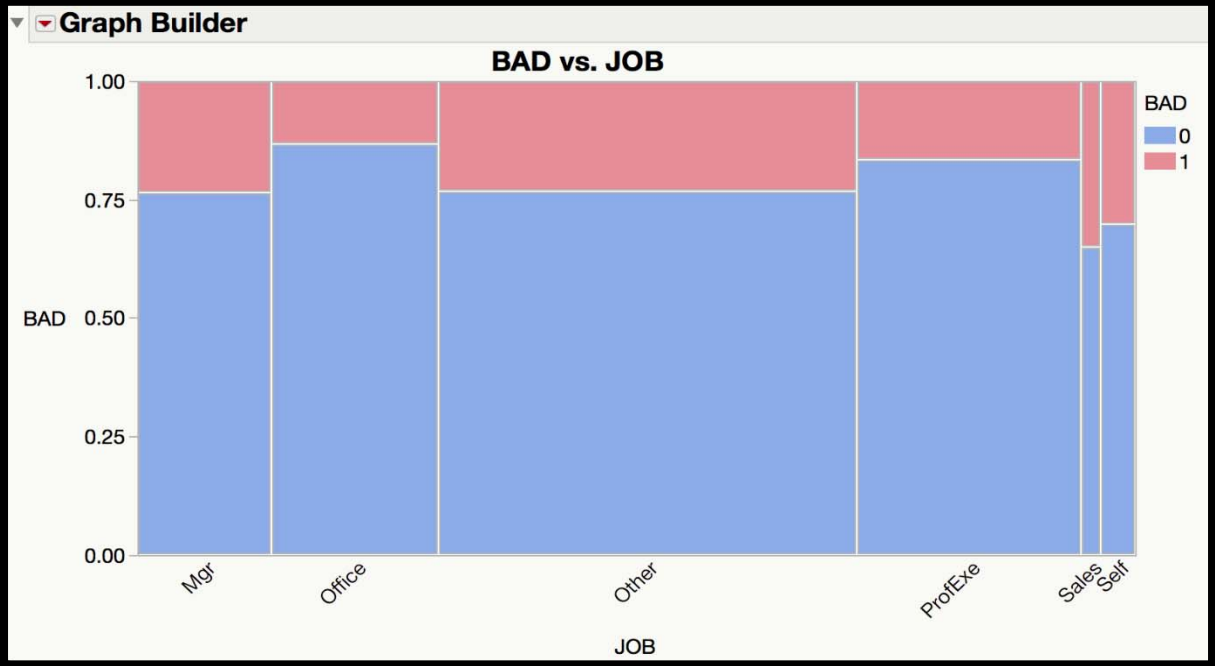
A graphical version of **Tabulate** is the **Graph Builder**. This is the first option under the **Graph** menu, and is a flexible graphing and exploratory platform. The initial window is shown in Figure 3.8. Like Tabulate, Graph Builder uses drop zones for specifying the variables to display. The primary zones are Y, and X, but there are other zones along the top and right that allow you to add additional information to the graph. The icons across the top can be used to change the graph type displayed, and additional options for the selected graph element are provided on the bottom left, below the variable list.

Figure 3.9: Box Plots for DEBTINC and LOAN versus BAD



In Figure 3.9 we see box plots for DEBTINC and LOAN (drag **BAD** to **Y**, **DEBTINC** to **X**, and **LOAN** after **DEBTINC** on the **X** zone, then click **Done**). We can see that more of the high risk customers had a high debt to income ratio, and more of the high loan amounts were for low risk customers.

Figure 3.10: Mosaic Plot of JOB versus BAD



When the X variable is categorical, different graph types are available. In Figure 3.10 we see a plot of BAD versus JOB (drag **BAD** to **Y**, **JOB** to **X**, select the **Mosaic** graph element from the icons at the top, then click **Done**).

This plot, called a mosaic plot, is an efficient way to graphically explore the relationship between two categorical variables. The width of the bars across the bottom shows the frequency of the job categories relative to one another. The bars are broken down, in this case, by **BAD** = 0 (bottom) and **BAD** = 1 (top). From this, we can that there are very few customers in sales or self-employed, the undefined **Other** category has the most customers, and the lowest risk customers are office workers and professionals or executives. (Hint: double-click on the x-axis to change the axis settings – here, we've changed the **Label Orientation** to **Angled** to display all of the labels).

Figure 3.11: Fit Y by X Dialog Window

To formally analyze the relationship between two variables we use the **Fit Y by X** platform from the **Analyze** menu. Like most platforms in JMP, **Fit Y by X** is contextual, producing analyses based on the modeling types of the variables selected.

The Fit Y by X dialog window is shown in Figure 3.11. The icons on the bottom left indicate what type of analysis JMP will provide based on the modeling type of the response and the factor. In this example, we have selected a nominal response (BAD), so the two available analyses are logistic regression and contingency. We have selected one continuous factor and one nominal factor. JMP will produce both logistic regression (for the continuous factor) and contingency (for the nominal factor).

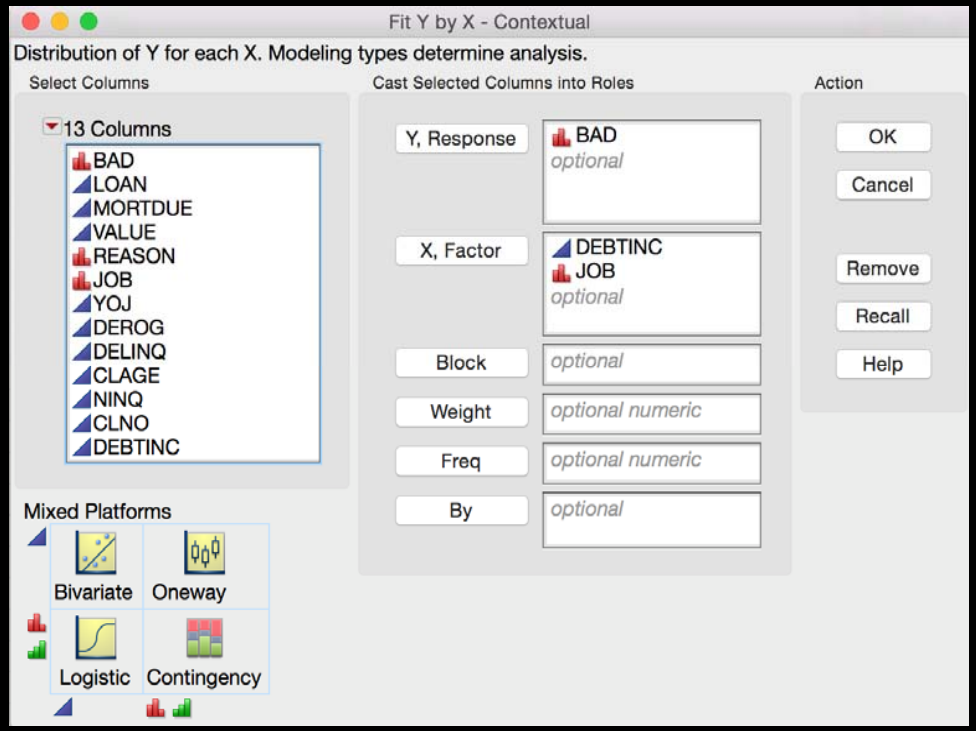


Figure 3.12: Fit Y by X, BAD versus LOAN and JOB

In Figure 3.12 we see partial results for both analyses. The mosaic plot is the same plot we created using the **Graph Builder**, but summary information and statistical output are also provided, along with many options under the red triangle. The logistic plot shows what happens to the probability that a customer will be low risk (**BAD** = 0) as the loan amount increases. Since the slope of the curve is positive, the probability increases (meaning that the probability that the customer will be a bad risk (**BAD** = 1) decreases. We'll provide background technical details on logistic regression in Chapter 5.

If our response were continuous, our two analysis options from **Fit Y by X** are bivariate. Bivariate provides (correlation and regression) and oneway (two-sample t-Test and ANOVA).

For an example, we use the **Boston Housing.jmp** data from the **Sample Data** directory. The response variable of interest, **mvalue**, is the median home value (in $1000) for homes in the Boston area in the 1970's. Two potential predictors are the number of rooms (**rooms**) and a dummy variable indicating whether the town tracks the Charles River (**chas**).

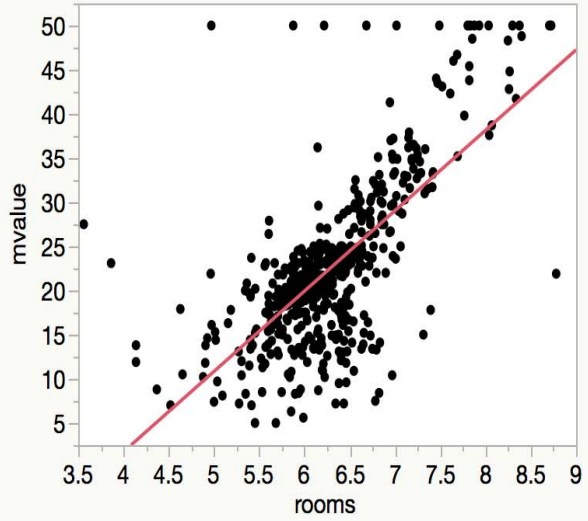
On the left in Figure 3.13 we see partial results from the bivariate analysis, where **mvalue** is the response and **rooms** is the factor. Here, we've fit a regression line (select **Fit Line** from the red triangle). On average, as the number of rooms increases the median value of homes also increases. But, this plot provides additional information. For example, we see a number of towns with median home values of 50 ($50,000). It is likely that this was an upper cutoff value of some sort.

On the right in Figure 3.13 we see partial results from the oneway analysis with **mvalue** as the response and **chas** as the factor. We have selected **Quantiles** and **t Test** from the top red triangle (and have selected **Points Jittered** from the **red triangle > Display Options**). If the factor selected has more than two levels, then ANOVA would be available. The p-value for the t-Test indicates that there is a significant difference in median home value for **chas** = 0 versus **chas** = 1 (the p-value for the two-tailed test is reported under Prob > |t|). This graph also provides additional information. For example, the number of points and the width of the x-axis labels indicates that there are for more towns that do not track the Charles River.

Figure 3.13: Fit Y by X Output for the Boston Housing Data

Fit Group

Bivariate Fit of mvalue By rooms



Linear Fit

Linear Fit

mvalue = -34.67062 + 9.102109*rooms

Summary of Fit

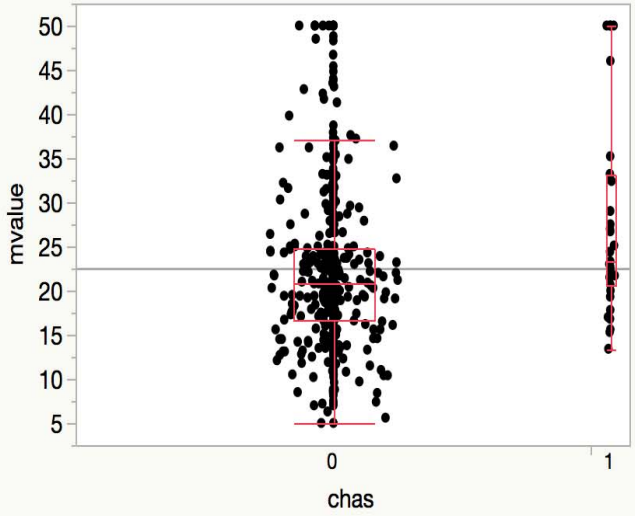
Lack Of Fit

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-34.67062	2.649803	-13.08	<.0001*
rooms	9.102109	0.419027	21.72	<.0001*

Oneway Analysis of mvalue By chas



Quantiles

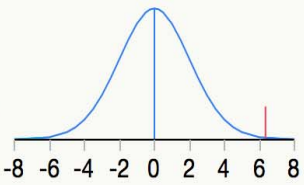
Level	Minimum	10%	25%	Median	75%	90%	Maximum
0	5	12.52	16.6	20.9	24.8	33.4	50
1	13.4	16.32	20.7	23.3	33.2	50	50

t Test

1-0

Assuming unequal variances

Difference	6.3462	t Ratio	3.113291
Std Err Dif	2.0384	DF	36.87641
Upper CL Dif	10.4768	Prob > |t|	0.0036*
Lower CL Dif	2.2155	Prob > t	0.0018*
Confidence	0.95	Prob < t	0.9982



From DRAFT *Building Better Models with JMP Pro*, Grayson, Gardner and Stephens, 2015.

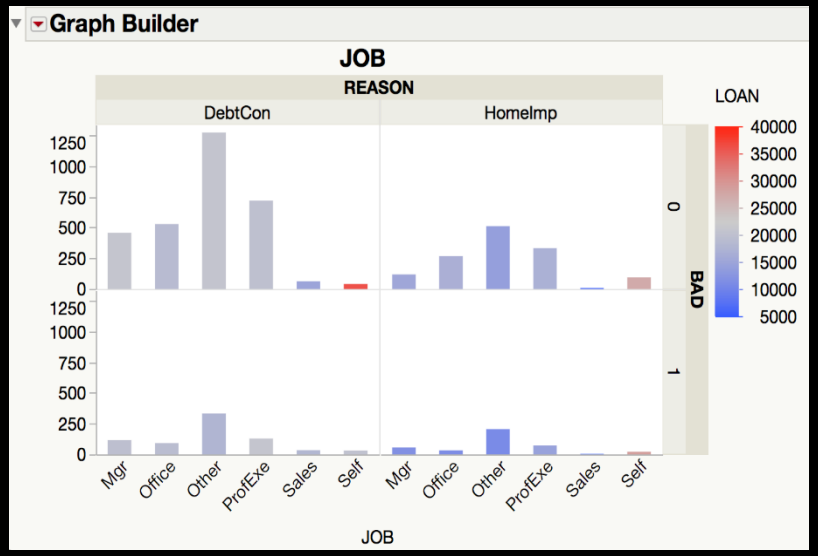
Exploring Data Three or More Variables at a Time

Tabulate and Graph Builder, introduced in the last section, can also be used to explore more than two variables at a time. For example, in Figure 3.14 (left) we see a tabular summary of BAD by REASON and JOB within REASON, along with average loan amounts. On the right in Figure 3.14 we have a graphical view of the same four variables, where the bar heights represent the number of customers and the bar colors represent the average loan amount.

Figure 3.14: Tabulate and Graph Builder with Multiple Variables

Tabulate

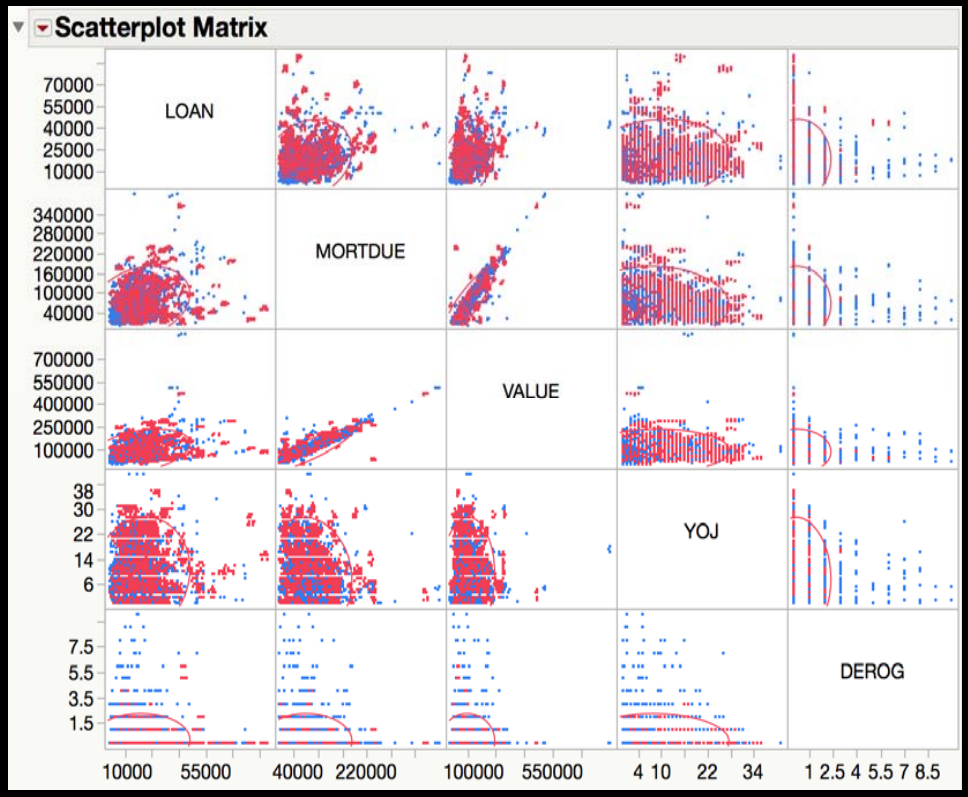
		BAD		
		0	1	LOAN
REASON	JOB	Row %	Row %	Mean
DebtCon	Mgr	0.80	0.20	20799.30
	Office	0.85	0.15	19056.29
	Other	0.79	0.21	20266.08
	ProfExe	0.85	0.15	20151.83
	Sales	0.65	0.35	16102.06
	Self	0.56	0.44	29389.04
HomeImp	Mgr	0.68	0.32	14194.83
	Office	0.89	0.11	16313.95
	Other	0.71	0.29	13175.42
	ProfExe	0.82	0.18	16594.07
	Sales	0.67	0.33	5308.33
	Self	0.82	0.18	27649.57



From DRAFT *Building Better Models with JMP Pro*, Grayson, Gardner and Stephens, 2015.

A popular view of potential relationships between continuous variables is a scatterplot matrix (from **Analyze > Multivariate Methods > Multivariate**). A scatterplot matrix for the first five continuous variables is shown in Figure 3.15. The points are colored by BAD = 0 (red) and BAD = 1 (blue). Each box graphically shows the correlation, or the linear association, between two variables. The first row contains scatterplots between LOAN and each of the other variables. A density ellipse is drawn in each scatterplot to provide an overall view of the relationship between each pair of variables.

Figure 3.15: Equity Scatterplot Matrix

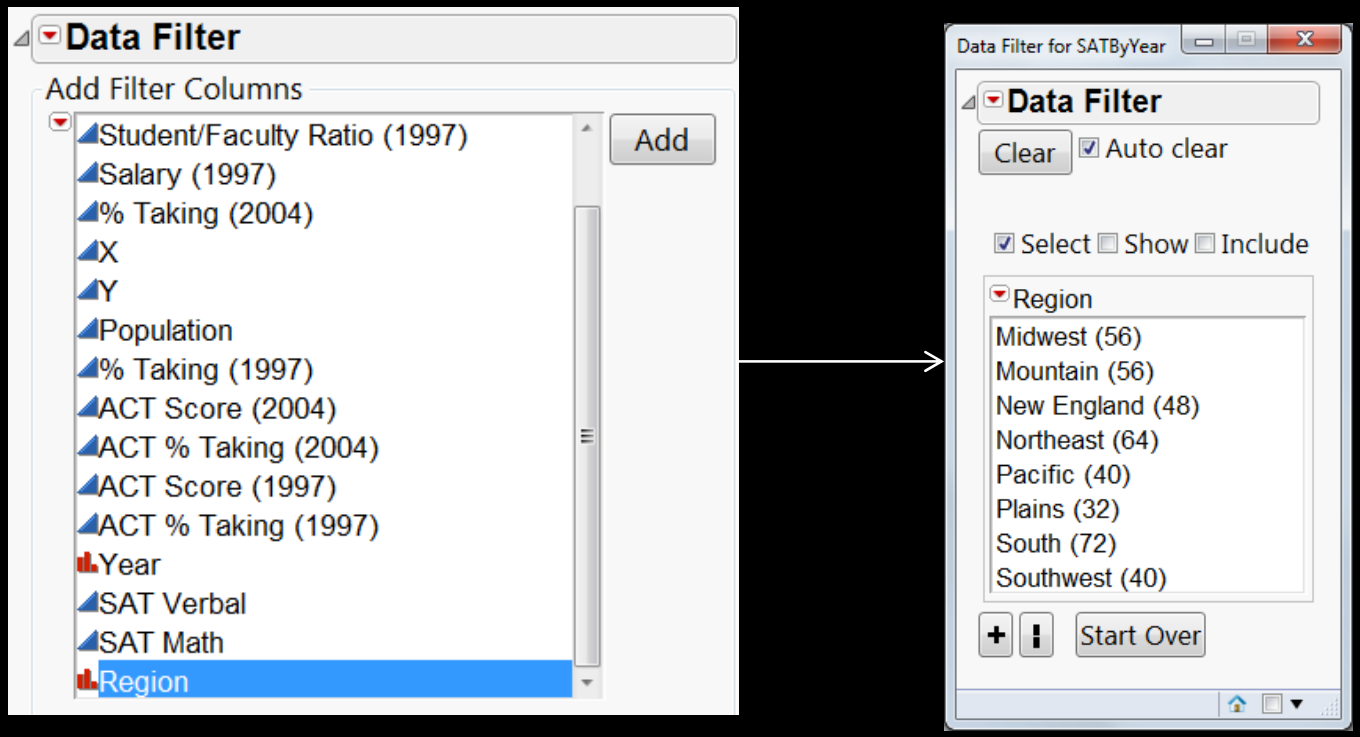


In this plot, we see that there is a strong relationship between MORTDUE and VALUE, and weak relationships between the other variables. We also see the extreme values for MORTDUE and VALUE across the other variables.

Other Useful Features

Data Filter

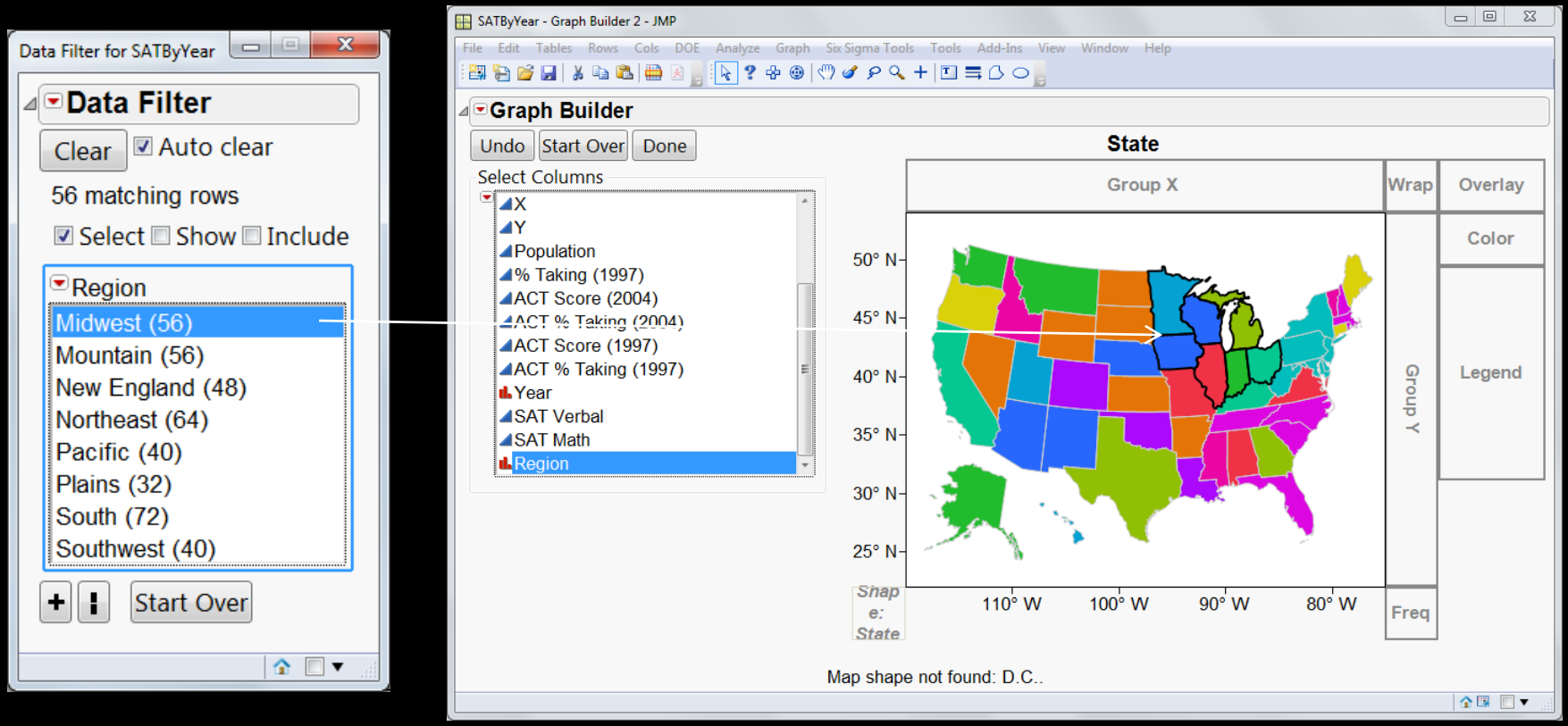
- The data filter (**Rows > Data Filter**) allows you to dynamically stratify data by values of one or more filtering variables.
- From an open data table select the data filter. Then, select a variable of interest, and click Add.



Courtesy of Mia Stephens: DSI 2014 Presentation

Data Filter

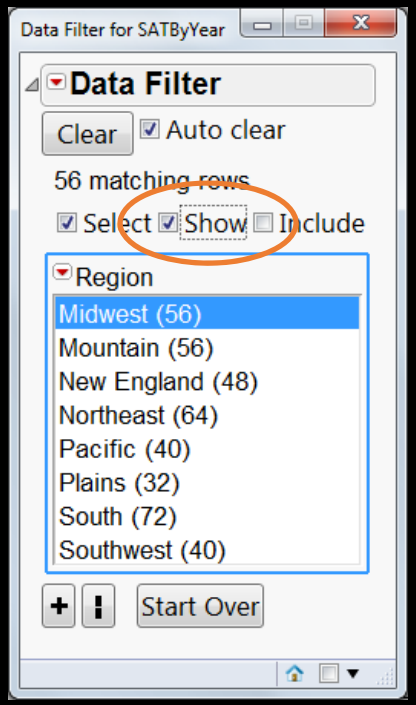
- Select a value of the filtering variable (or a range of values if the variable is continuous). The corresponding values will be selected in the data table and every open graph.

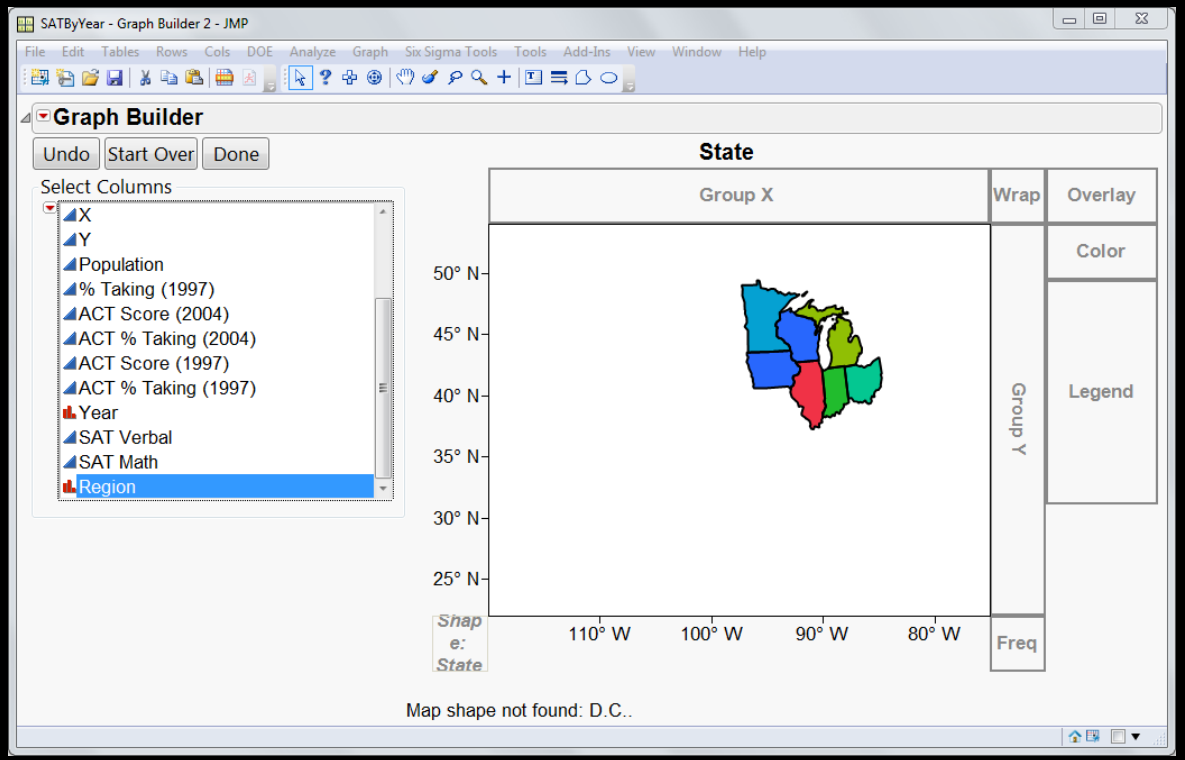


Courtesy of Mia Stephens: DSI 2014 Presentation

Data Filter

- Click **Show** in the Data Filter. All other observations will be hidden in the data table.
- In Graph Builder, only the selected value(s) of the filter variable(s) will display.

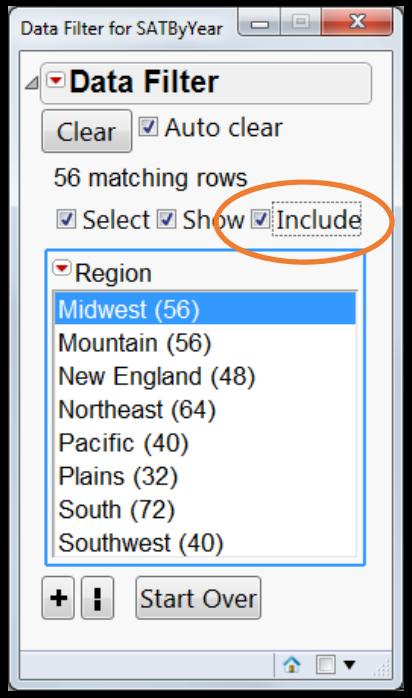


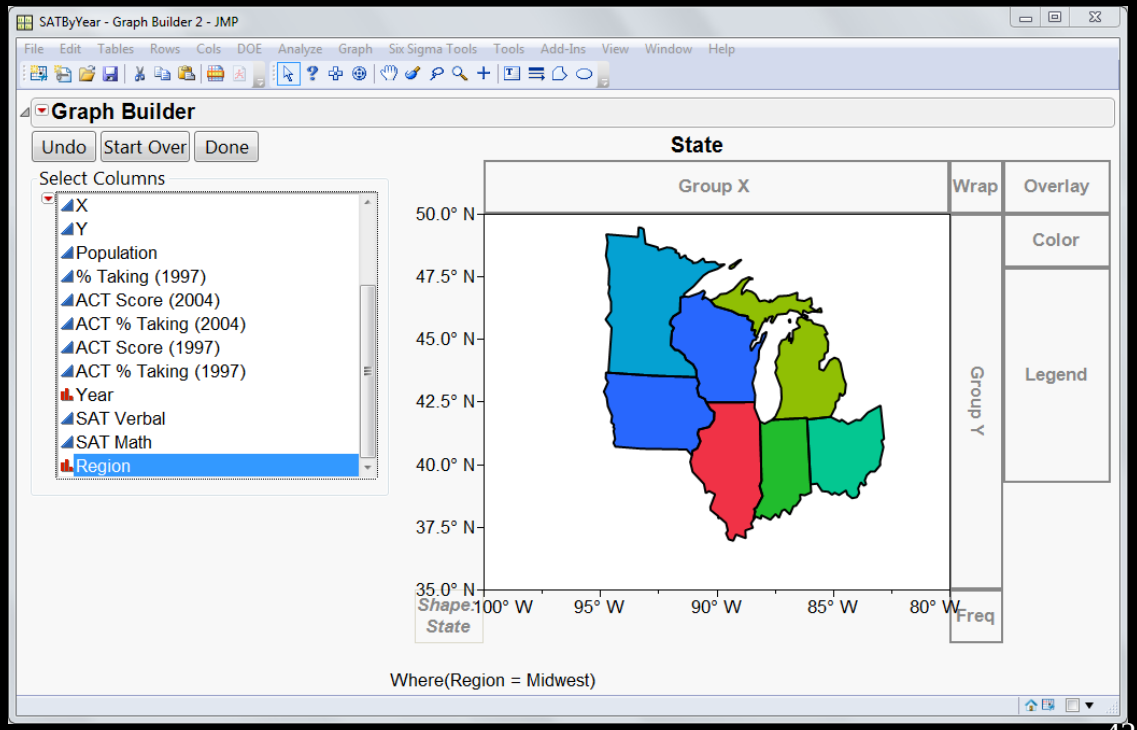


Courtesy of Mia Stephens: DSI 2014 Presentation

Data Filter

- Click **Exclude** in the Data Filter. All other observations will be excluded in the data table.
- Graph Builder will automatically update – only selected values will be included in the graph and calculations.

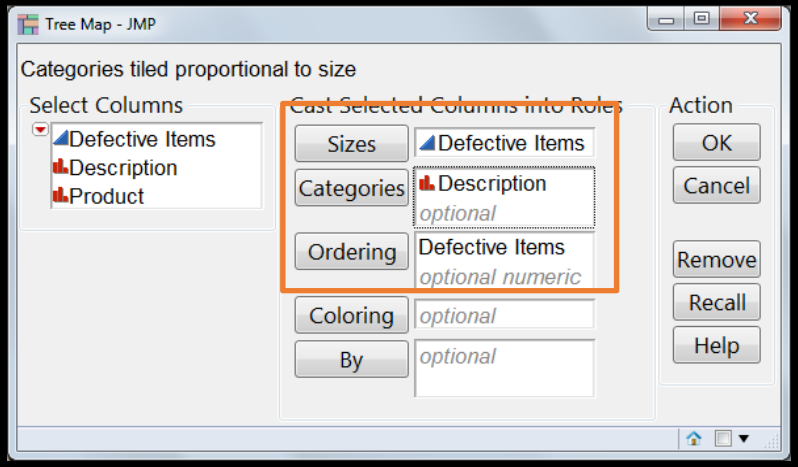


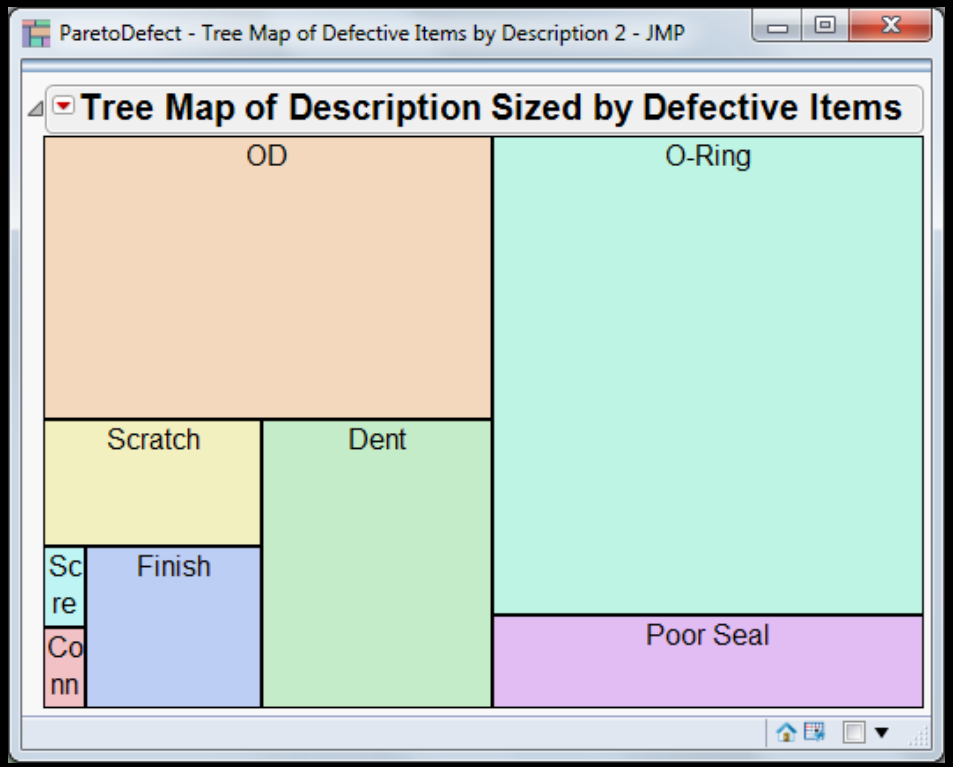


Courtesy of Mia Stephens: DSI 2014 Presentation

Tree Maps

- Tree maps (**Graph > Tree Map**) are an alternative to Mosaic Plots for displaying nominal data.
- Select Categories, a Size variable, and an Ordering variable (Example: **ParetoDefect.jmp**).

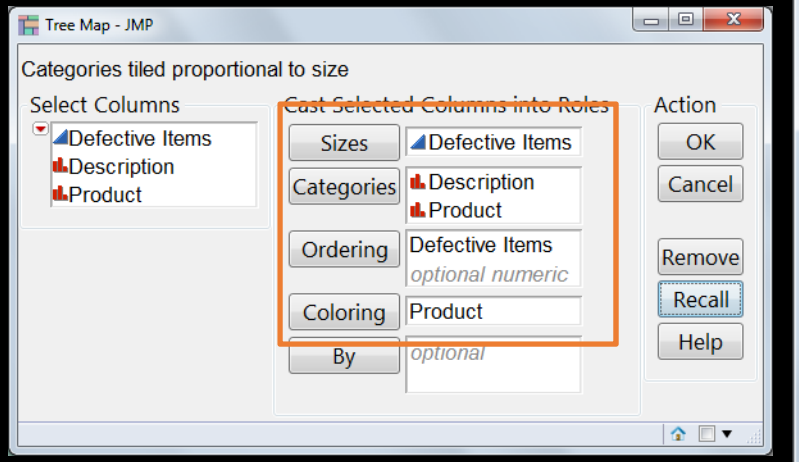


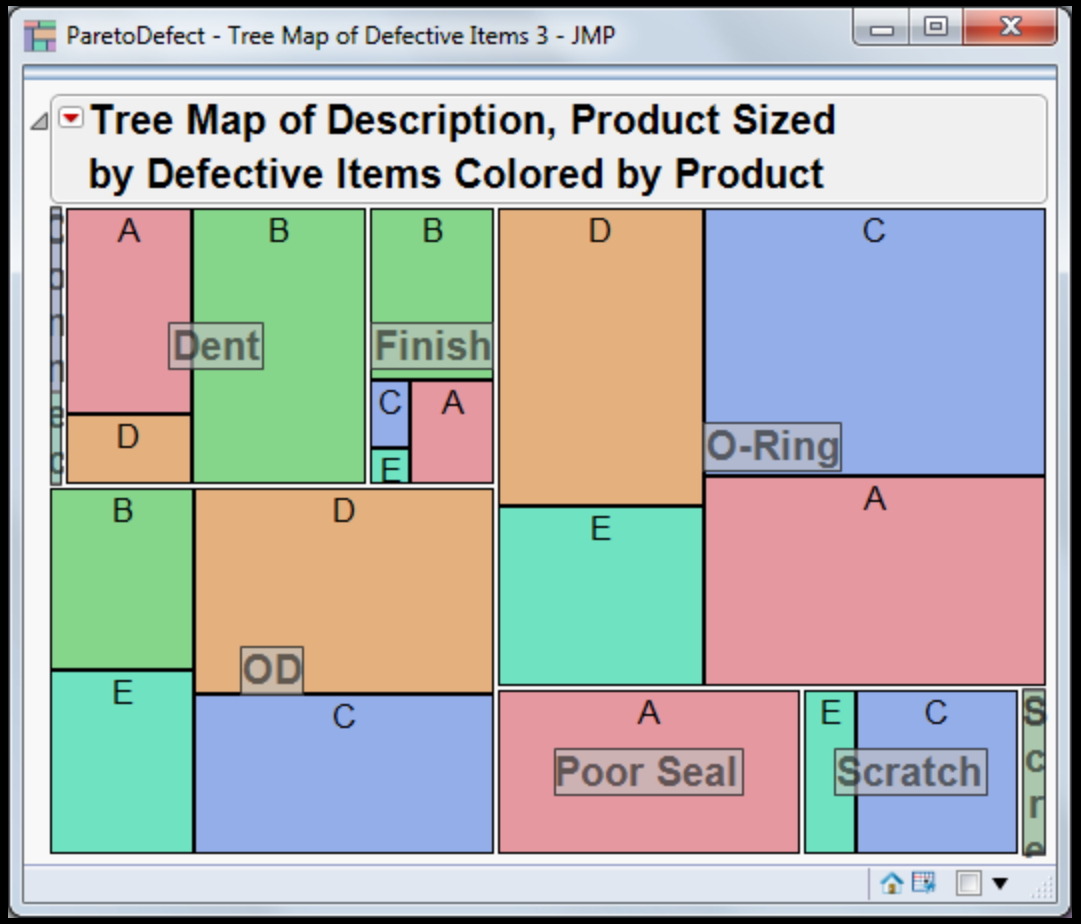


Courtesy of Mia Stephens: DSI 2014 Presentation

Tree Maps

- Add a second Category variable, and add a Coloring variable:

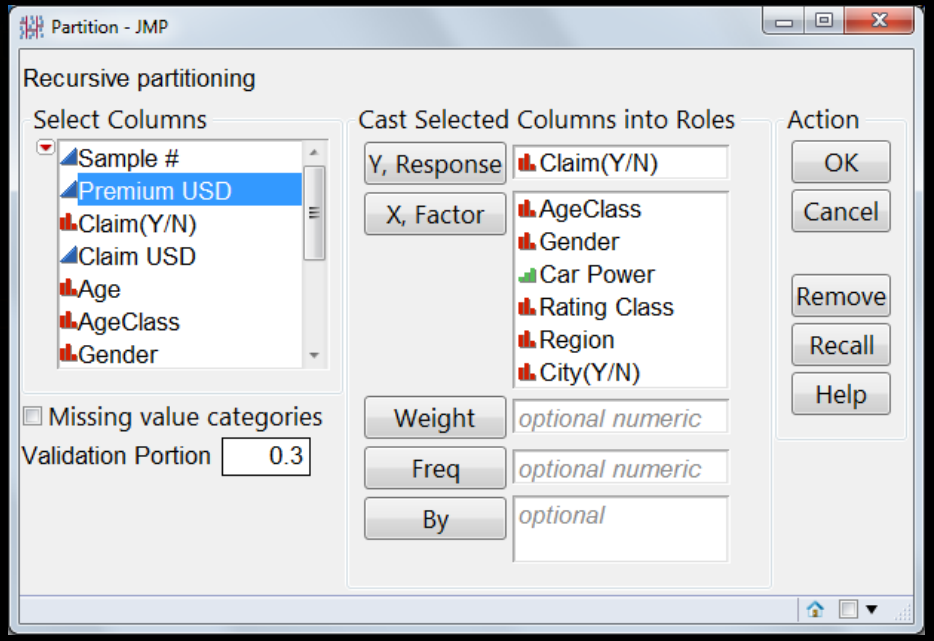




Courtesy of Mia Stephens: DSI 2014 Presentation

Introduction to Partition

- Partition (**Analyze > Modeling > Partition**) is a data mining tool, used for data exploration and for building predictive models.
- We will introduce the tool here, and will revisit the topics again later.
- Example: **Auto Raw Data.jmp** in the Sample Data Directory.

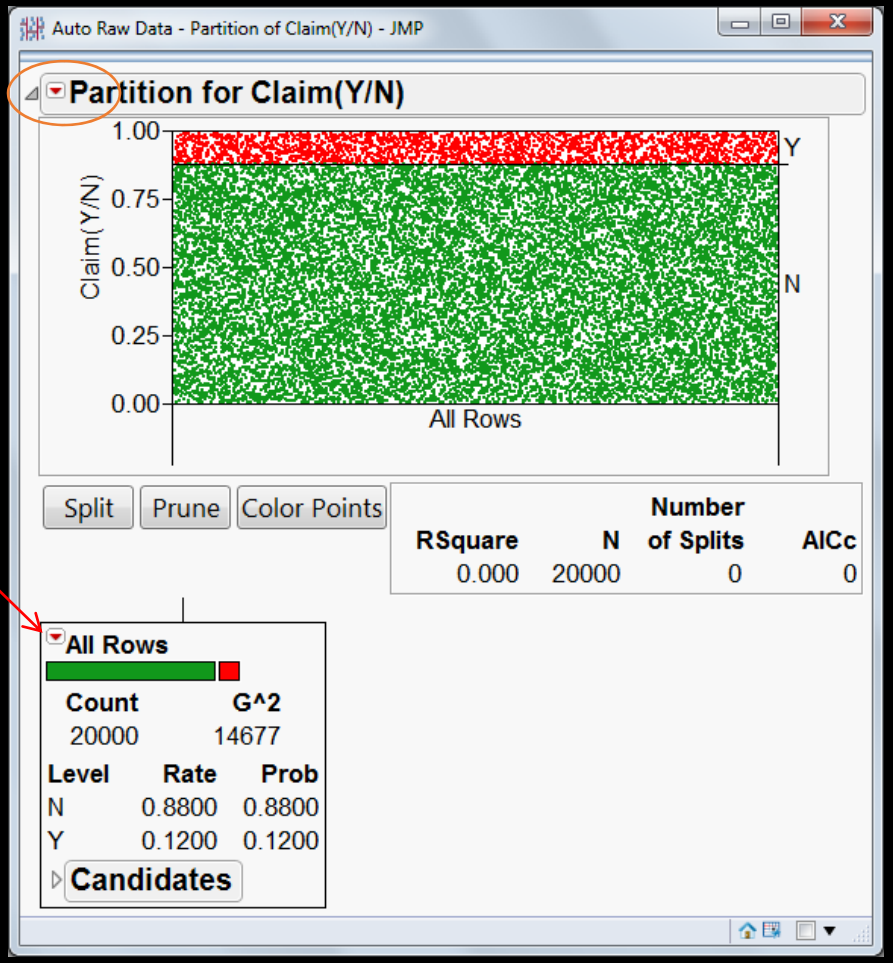


Courtesy of Mia Stephens: DSI 2014 Presentation

Introduction to Partition

- JMP displays a graph with a line drawn at the overall response rate.

Click on the top red triangle and select **Display Options > Show Split Prob** to display the overall response rate.



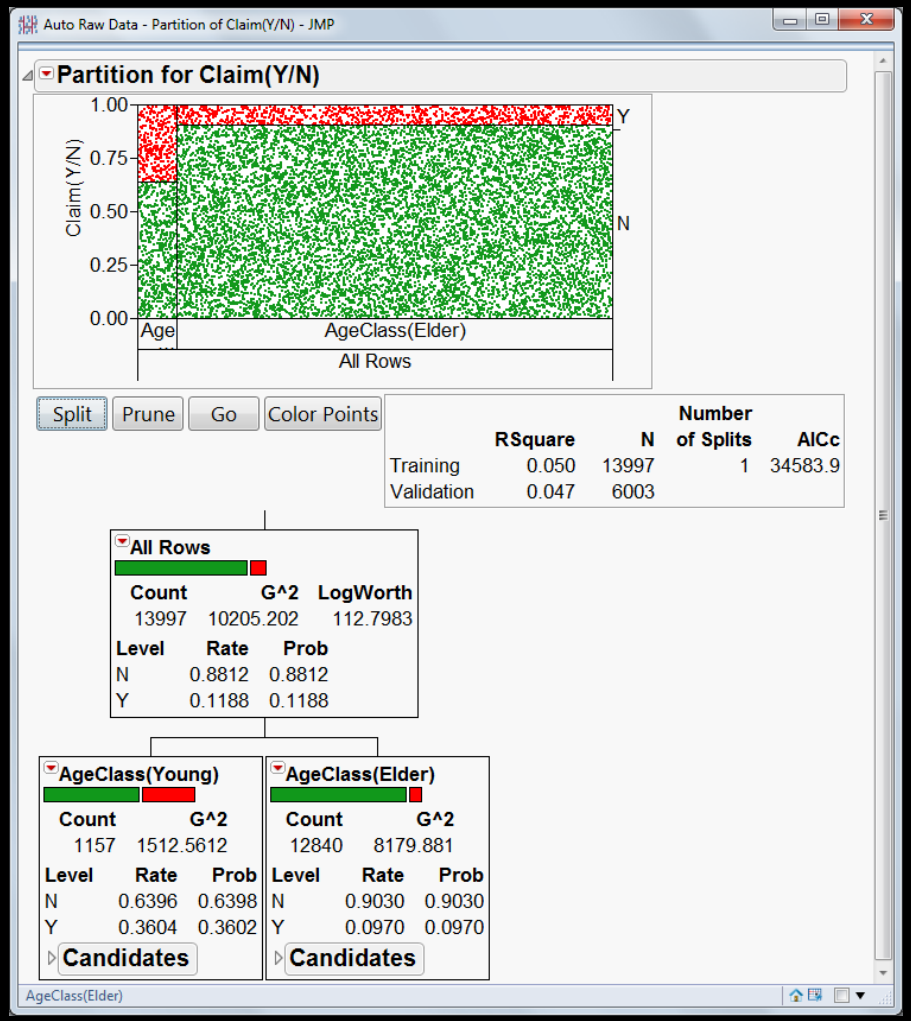
Courtesy of Mia Stephens: DSI 2014 Presentation

Introduction to Partition

- Click the **Split** button. The original observations will be split into two nodes, or leaves.

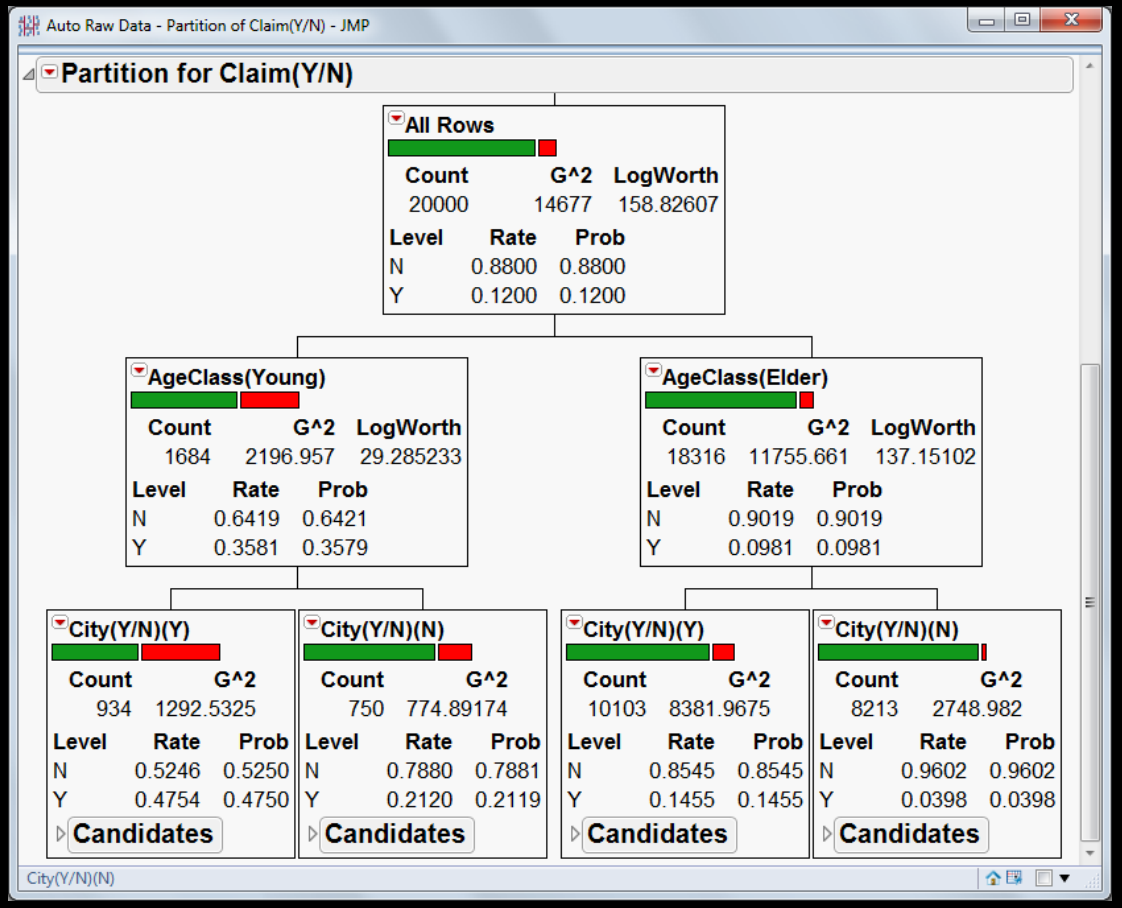
Interpretation:

- In the left leaf, corresponding to AgeClass = Young, the probability that there is a claim is 0.3604.
- In the right leaf, corresponding to AgeClass = Elder, the probability that there is a claim is 0.097.



Introduction to Partition

- Click the **Split** button again, and then a third time. The original observations will be split into two nodes, or leaves.



Saving Results

jmp.com/learn rev 07/2012

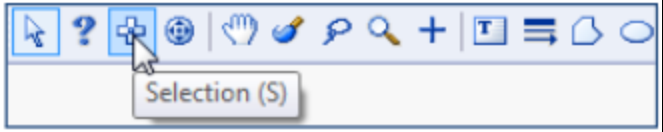
Saving JMP® Results

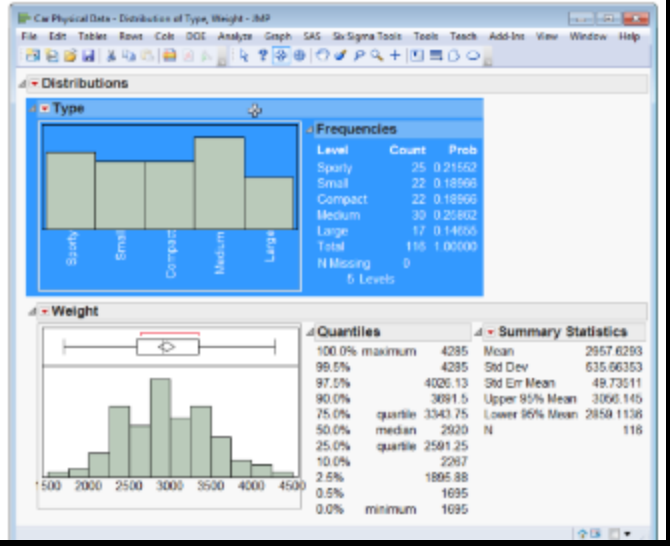
This page gives information on saving JMP output and results.

The Selection Tool – Copy and Paste to Another Program

1. From any JMP output window, click on the **selection tool** in the toolbar or use the keyboard shortcut (S).
2. Click on the content you'd like to copy – selected content is highlighted. Click near the edge of the report to select all content. To extend a selection, hold the shift key.
3. Click **Edit > Copy** (or **Control-C**).
4. Open the program where you'd like to paste the content, and select **Paste**. To paste as an object, select **Paste > Paste Special** and from the list select **Picture (Enhanced Metafile).**

   If using a Mac, select **Paste**, or use **Paste Special > PDF** for high-quality graphics.





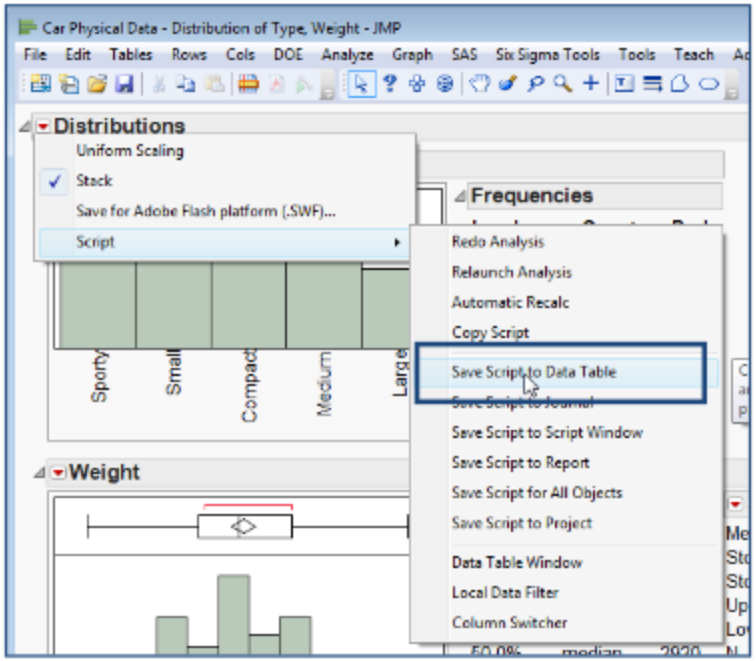
http://www.jmp.com/academic/learning_library.shtml

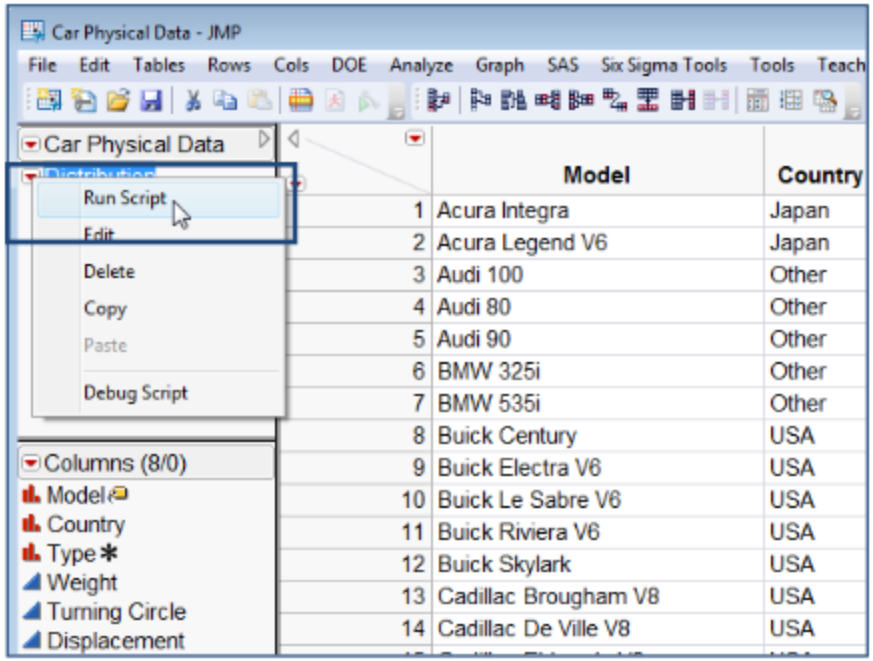
Courtesy of Mia Stephens: DSI 2014 Presentation

Save Your Work Using Scripts

You can save the steps taken to produce a report as a JSL script (JSL is "JMP Scripting Language"). This enables you to recreate the report at any time.

1. From any JMP output window, click the **top red triangle** and select **Script > Save Script to Data Table**.
2. The saved script will appear in the **tables panel** of the data table.
3. To run the script, click on the red triangle and select **Run Script**. To edit (change the name), select **Edit**.
4. Select **File > Save** to save the saved script and any other changes to the data table.





Notes: Select **Edit > Save Selection As** (or **File > Export** on the Mac) to save the selection in a variety of graphical formats, including JPG, EPS, SVG and GIF. To change default graphic formats, use **File > Preferences > Windows Specific** (or **JMP > Preferences > Mac OS Settings** on the Mac).

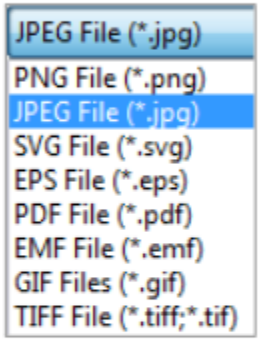
Courtesy of Mia Stephens: DSI 2014 Presentation

Publication-Quality Graphics – Windows

Saving Graphic Images

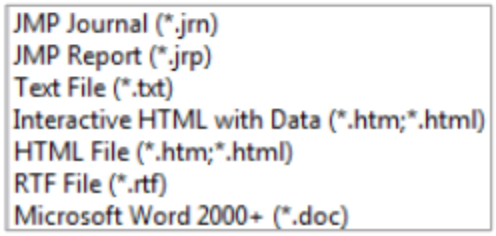
1. From any JMP output window, click on the **selection tool** in the toolbar or use the keyboard shortcut (S).
2. Click on the content you'd like to copy – selected content is highlighted. Click near the edge of the report to select all content. To extend a selection, hold the shift key.
3. Select **Edit > Save Selection As**.
4. Under **Save as Type** (at the bottom of the window), select the desired graphic format.

   For printed media (presentations, journals, etc.) a vector image, such as **EMF** (enhanced metafile) is recommended. This format preserves transparency, can be edited in image-editing software, and can be scaled or resized without losing clarity.



5. If a bitmap graphic format (PNG, JPEG, GIF or TIFF) is selected, change the **Image DPI Setting** to 300 for a high-resolution graphic.
6. Enter a file name, specify the folder to save the file to, and click **Save**.

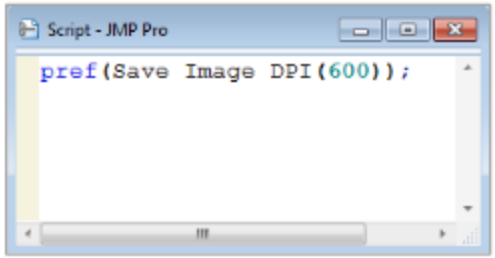
Note: To save all of the output in the active window, use **File > Save As**. In addition to the graphics formats, several report formats are available.



Courtesy of Mia Stephens: DSI 2014 Presentation

Saving Graphic Images

1. In JMP, go to **File > New > Script**.
2. Type the following code in this window: *pref(Save Image DPI(600));*
3. Right-click, and select **Run Script**.



This will change the default for all saved graphics to 600 (or whatever number you plug into the parentheses) until you change the default.

Tips:

- Edit the graph in JMP prior to saving in a graphics format. For example, change the size of the graph frame, change marker sizes and/or colors, and edit axes, legends and titles prior to saving.
- When copying and pasting into Microsoft Office products, use **Paste Special > Picture Enhanced Metafile** to paste output as vector images.
- The EPS vector format does not preserve graph transparency (EMF does).
- JMP report elements that use Open GL (surfaces, scatterplot 3-D, etc.) cannot be output as vector images.

Courtesy of Mia Stephens: DSI 2014 Presentation

Summary

In this section we introduced some tools for working with and exploring data:

- Tabulate – for summarizing data, a drag and drop pivot table
- Graph Builder – dynamically graph data, explore many variables at a time.
- Mapping – geographic maps from Graph Builder, also available from Bubble Plot
- Data Filter – dynamically stratify data
- Tree Map – an alternative to Mosaic Plots
- Partition – Use to explore potentially important variables when there are many variables, nominal variables, missing values, and messy data

Courtesy of Mia Stephens: DSI 2014 Presentation