Rotman

Master of
Management
Analytics

INTRO TO JMP – PART 3

Bootcamp (https://tdmdal.github.io/mma-jmp/)

September 17, 2020 Prepared by Jay / TDMDAL

Rotman School of Management
UNIVERSITY OF TORONTO

Plan

- Session 1
  - Workflow overview
  - Basic data manipulation

- Session 2
  - Join data tables
  - JMP graphing

- Session 3
  - **Modelling**
  - JMP Journal
  - JMP Scripting Language

Modeling in JMP

- Linear regression (done)
    - predict a continuous variable

- Logistic regression
    - predict categorical variable (i.e. a classification problem)
        - binomial logistic regression: the categorical variable has binary outcomes (e.g., 0, 1)

- K-mean clustering
    - a method to partition observations (into clusters)

- Outliers, missing values, and patterns

Binomial Logistic Regression

- let Y be the binary outcome variable
  - e.g. $\{0, 1\} = \{fail,\ success\}$

- Let $p = prob(Y = 1)$; $\frac{p}{1-p}$ is then the odds of being 1 (or success)

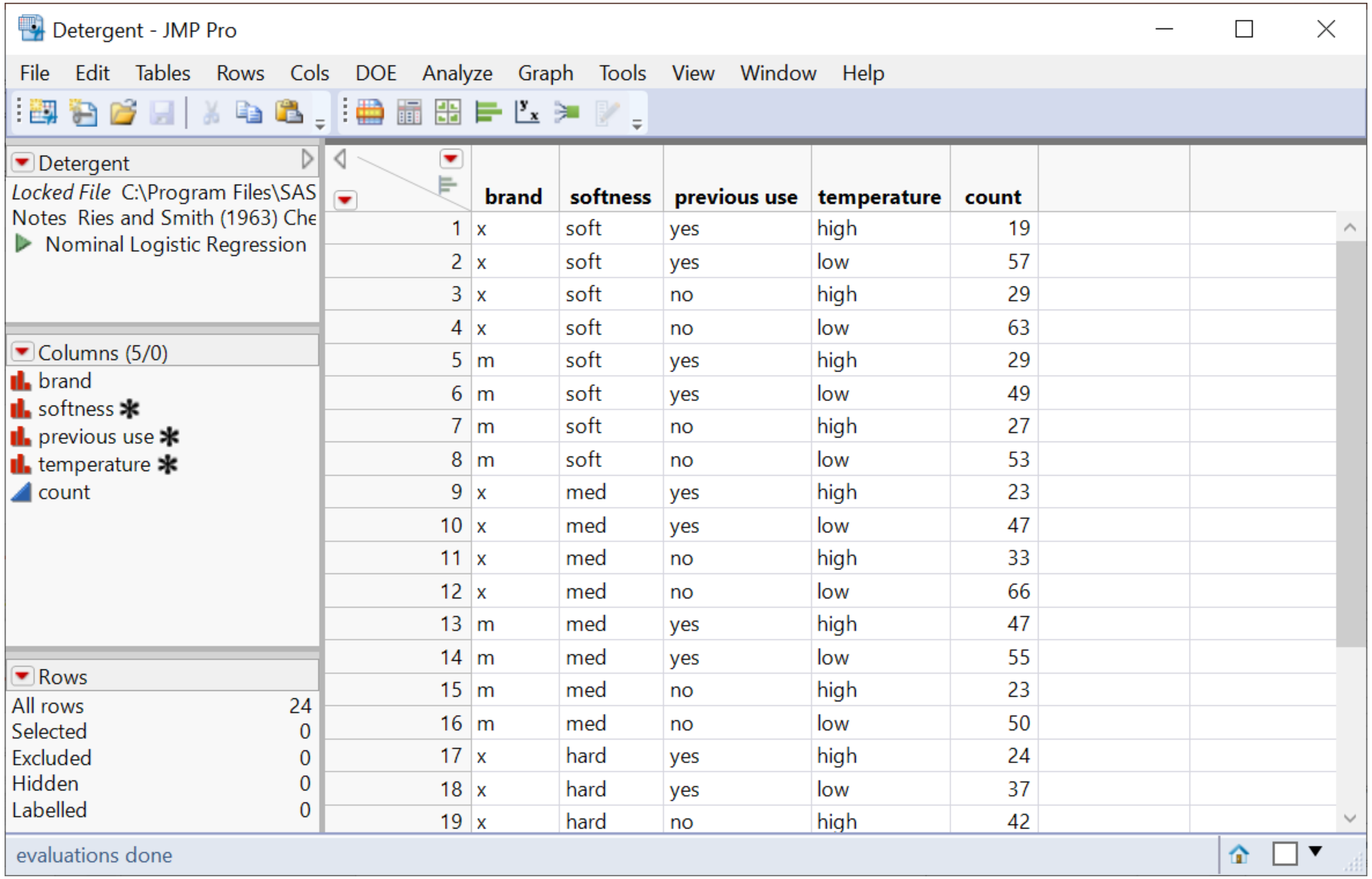
- Binary logistic regression models the logit-transformed probability as a linear relationship with the predictor variables
  - maximum likelihood estimation

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/

Binomial Logistic Regression (Demo): Data

- Preference for a brand of detergent (Ries and Smith, 1963)
  - Detergent.jmp
- Survey Questions
  - which brand do you prefer, x or m
  - water softness
  - previous user of m
  - water temperature


Detergent - JMP Pro
File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help
evaluations done

Detergent

Locked File C:\Program Files\SAS

Notes Ries and Smith (1963) Che

Nominal Logistic Regression

Columns (5/0)

brand

softness ✱

previous use ✱

temperature ✱

count

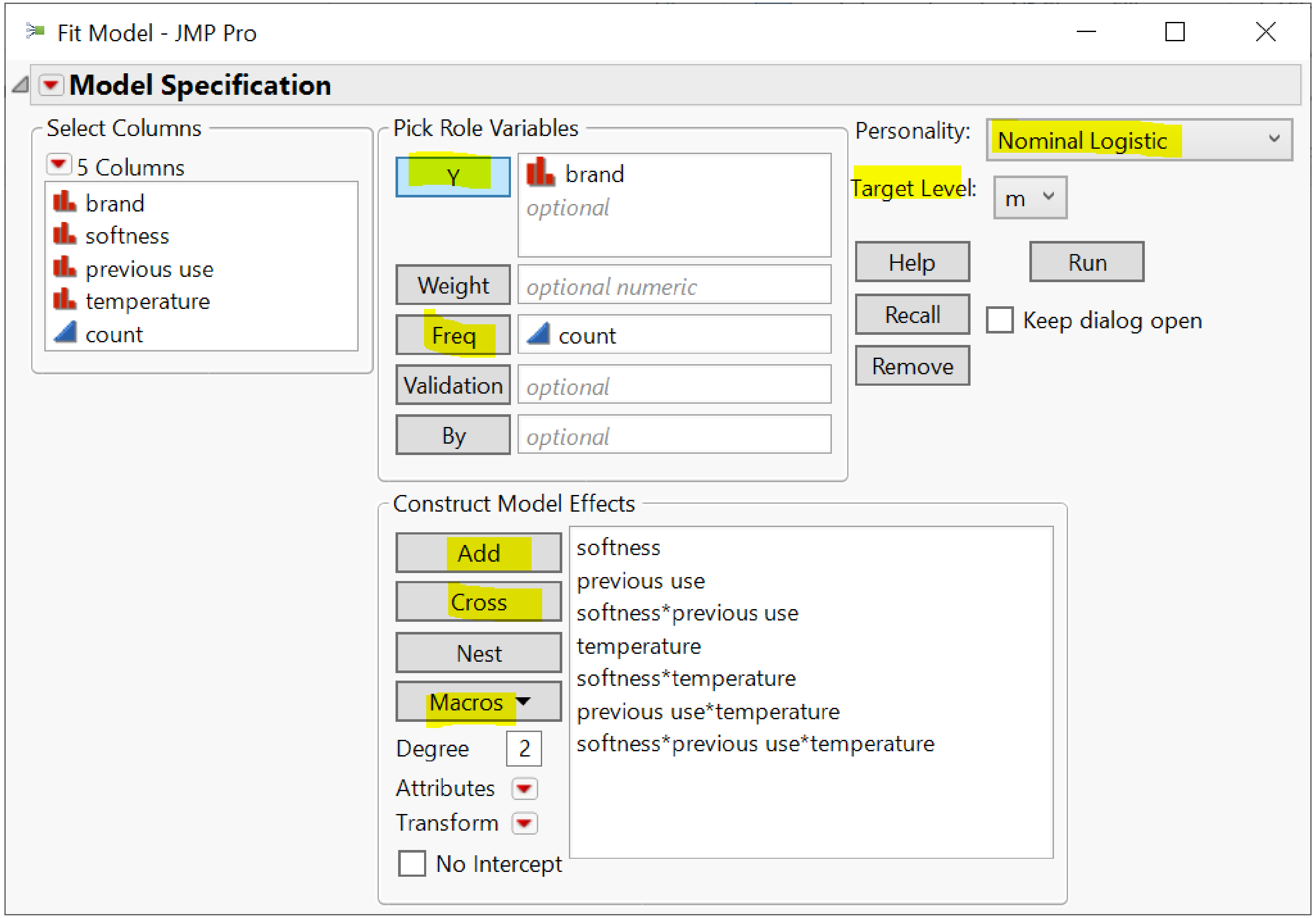
Rows

All rows	24
Selected	0
Excluded	0
Hidden	0
Labelled	0

	brand	softness	previous use	temperature	count
1	x	soft	yes	high	19
2	x	soft	yes	low	57
3	x	soft	no	high	29
4	x	soft	no	low	63
5	m	soft	yes	high	29
6	m	soft	yes	low	49
7	m	soft	no	high	27
8	m	soft	no	low	53
9	x	med	yes	high	23
10	x	med	yes	low	47
11	x	med	no	high	33
12	x	med	no	low	66
13	m	med	yes	high	47
14	m	med	yes	low	55
15	m	med	no	high	23
16	m	med	no	low	50
17	x	hard	yes	high	24
18	x	hard	yes	low	37
19	x	hard	no	high	42

Binomial Logistic Regression (Demo): Fit

- Analyze > Fit Model



Fit Model - JMP Pro

Model Specification

Select Columns

5 Columns

- brand
- softness
- previous use
- temperature
- count

Pick Role Variables

Y: brand

Weight: *optional numeric*

Freq: count

Validation: *optional*

By: *optional*

Personality: Nominal Logistic

Target Level: m

Construct Model Effects

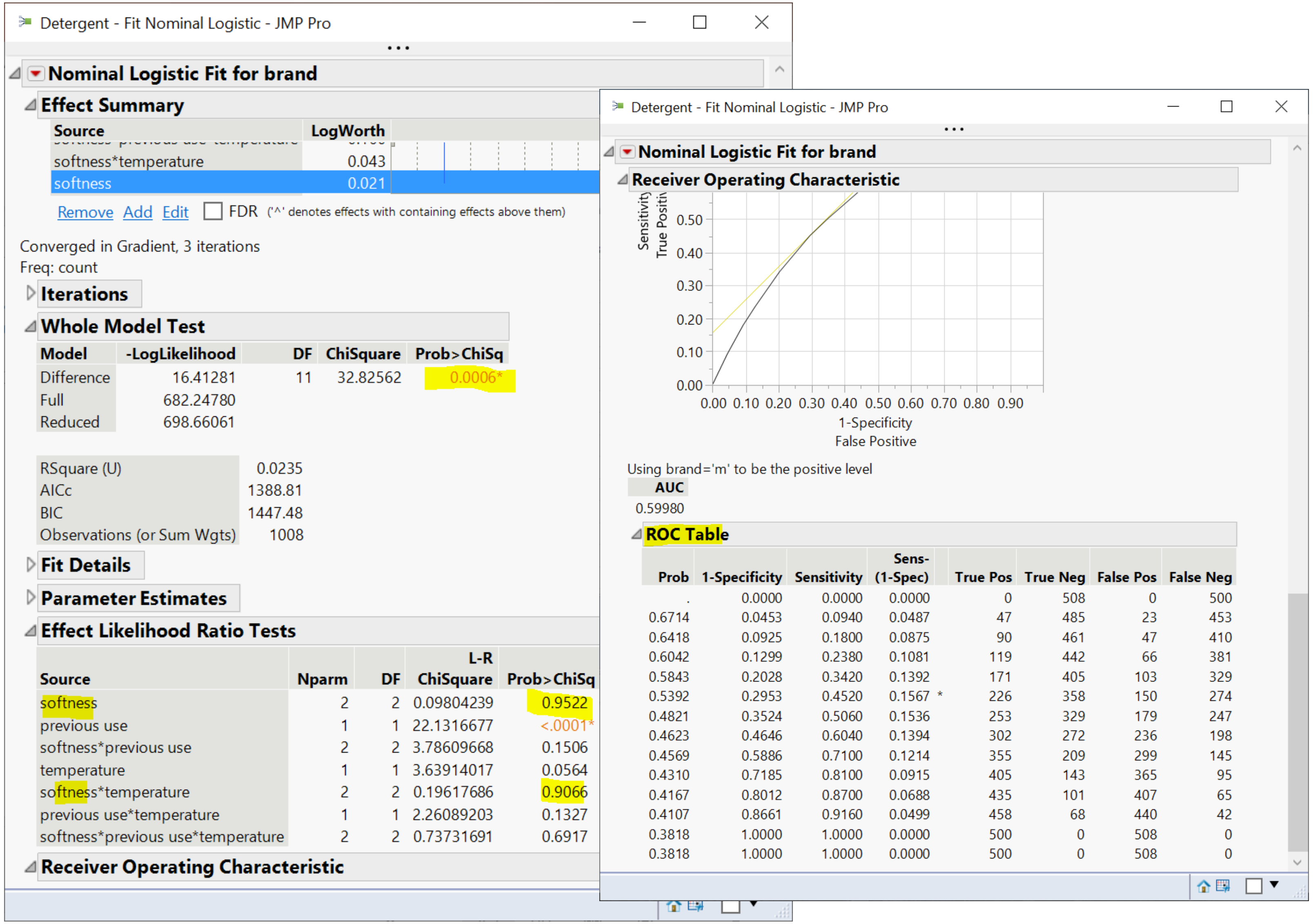
Degree 2

□ No Intercept

softness
previous use
softness*previous use
temperature
softness*temperature
previous use*temperature
softness*previous use*temperature

Binomial Logistic Regression (Demo): Report

- Overall model fit is significant
- Softness doesn't seem to contribute too much
- ROC Table
  - [sensitivity, specificity, etc.]



Nominal Logistic Fit for brand

Effect Summary

Source	LogWorth
softness*temperature	0.043
softness	0.021

('^' denotes effects with containing effects above them)

Converged in Gradient, 3 iterations

Freq: count

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	16.41281	11	32.82562	0.0006*
Full	682.24780			
Reduced	698.66061			

RSquare (U)	0.0235
AICc	1388.81
BIC	1447.48
Observations (or Sum Wgts)	1008

Fit Details

Parameter Estimates

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
softness	2	2	0.09804239	0.9522
previous use	1	1	22.1316677	<.0001*
softness*previous use	2	2	3.78609668	0.1506
temperature	1	1	3.63914017	0.0564
softness*temperature	2	2	0.19617686	0.9066
previous use*temperature	1	1	2.26089203	0.1327
softness*previous use*temperature	2	2	0.73731691	0.6917

Receiver Operating Characteristic

Nominal Logistic Fit for brand

Receiver Operating Characteristic

Using brand='m' to be the positive level

AUC
0.59980

ROC Table

Prob	1-Specificity	Sensitivity	Sens-(1-Spec)		True Pos	True Neg	False Pos	False Neg
.	0.0000	0.0000	0.0000		0	508	0	500
0.6714	0.0453	0.0940	0.0487		47	485	23	453
0.6418	0.0925	0.1800	0.0875		90	461	47	410
0.6042	0.1299	0.2380	0.1081		119	442	66	381
0.5843	0.2028	0.3420	0.1392		171	405	103	329
0.5392	0.2953	0.4520	0.1567	*	226	358	150	274
0.4821	0.3524	0.5060	0.1536		253	329	179	247
0.4623	0.4646	0.6040	0.1394		302	272	236	198
0.4569	0.5886	0.7100	0.1214		355	209	299	145
0.4310	0.7185	0.8100	0.0915		405	143	365	95
0.4167	0.8012	0.8700	0.0688		435	101	407	65
0.4107	0.8661	0.9160	0.0499		458	68	440	42
0.3818	1.0000	1.0000	0.0000		500	0	508	0
0.3818	1.0000	1.0000	0.0000		500	0	508	0

Your Turn (Hands-on)

- Do the same analysis without the softness variable
- Save the analysis script in the data table
- Challenge: How to construct a table of correct classification rate at each probability cutoff

$$correct\ classification\ rate = \frac{true\ positive + true\ negative}{total\ \#\ of\ predictions}$$

K-Means Cluster Analysis

- A method to partition n observations into k clusters
  - such that total within-cluster sum of squares (between observations to cluster centroid) is minimized

- A **cluster** refers to a collection of data points aggregated together because of certain similarities

- Need to set k
  - There are methods to help you decide the value of k

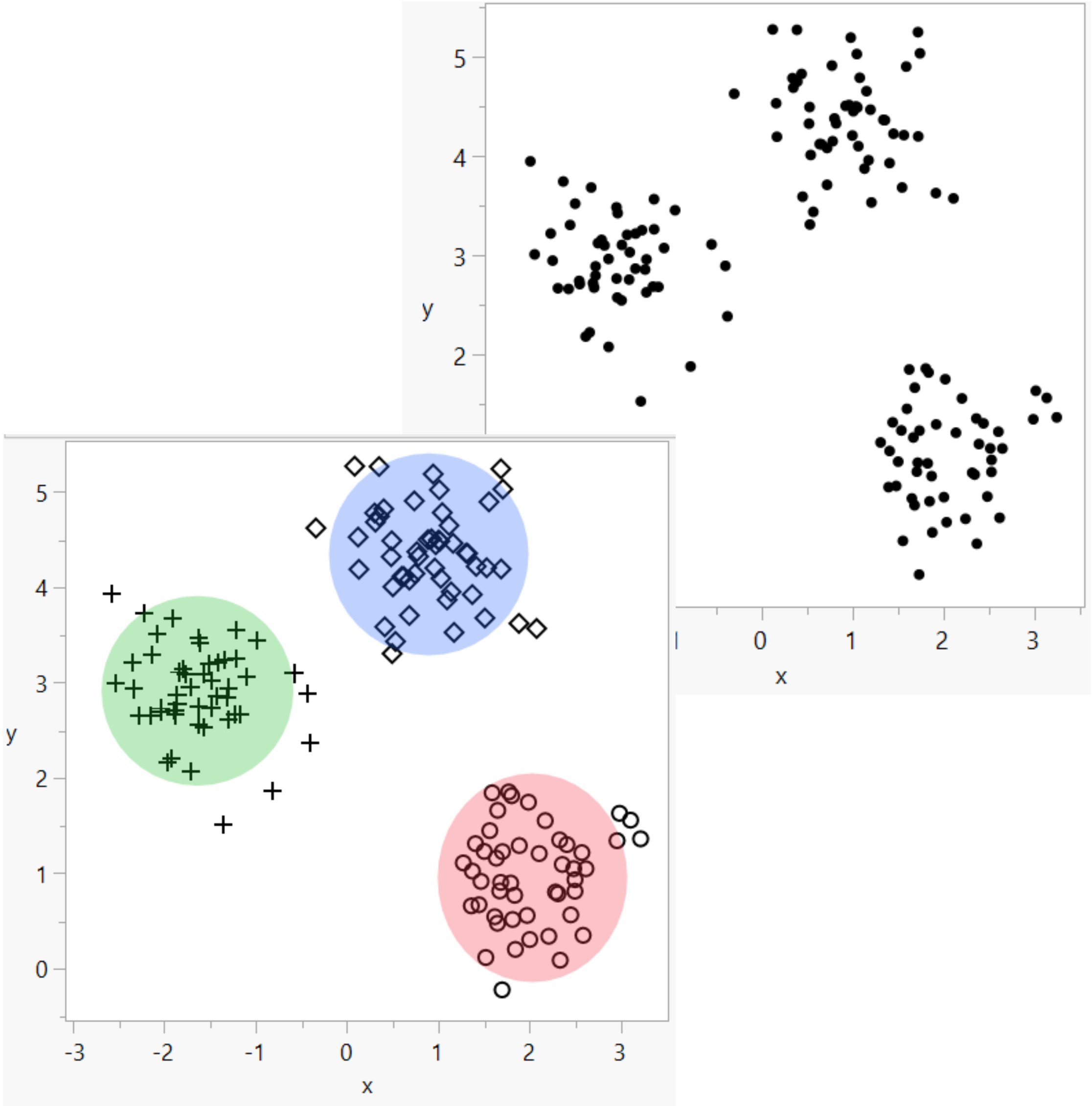
K-Means Cluster Analysis: An Example

- Observations: 150 2-d points

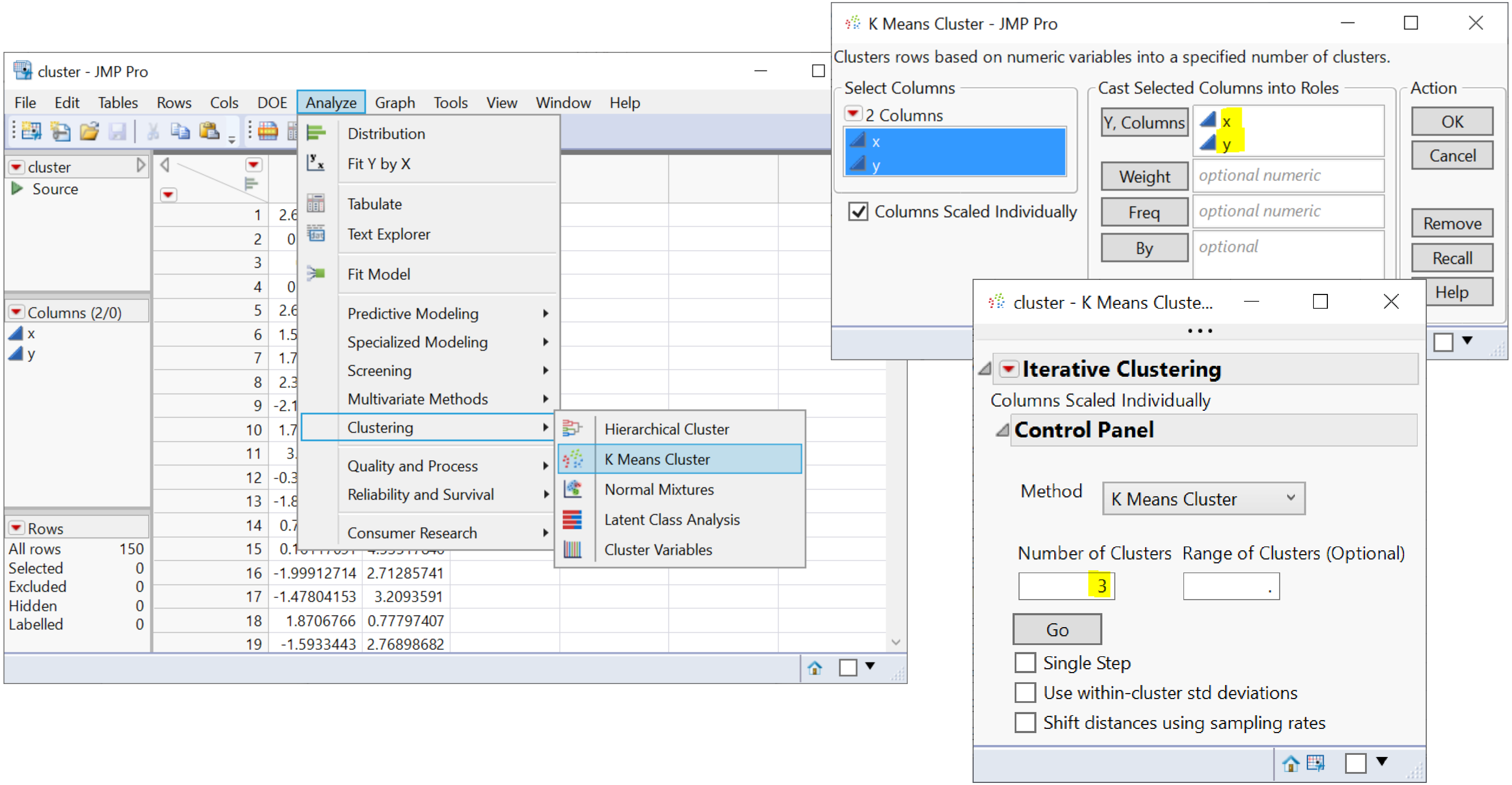
- Set $k = 3$
  - partition each observation to one of the 3 clusters $S = \{S_1, S_2, S_3\}$

- K-means clustering algorithm finds 3 clusters such that

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{3} \underbrace{\sum_{x \in S_i} \|x - \mu_i\|^2}_{\text{Within-cluster sum of squares}}$$



K-mean Cluster Platform (Demo)



	x	y
16	-1.99912714	2.71285741
17	-1.47804153	3.2093591
18	1.8706766	0.77797407
19	-1.5933443	2.76898682

Clusters rows based on numeric variables into a specified number of clusters.

Y, Columns: x, y

☑ Columns Scaled Individually

Iterative Clustering

Columns Scaled Individually

Control Panel

Method: K Means Cluster

Number of Clusters: 3

Range of Clusters (Optional): .

□ Single Step

□ Use within-cluster std deviations

□ Shift distances using sampling rates

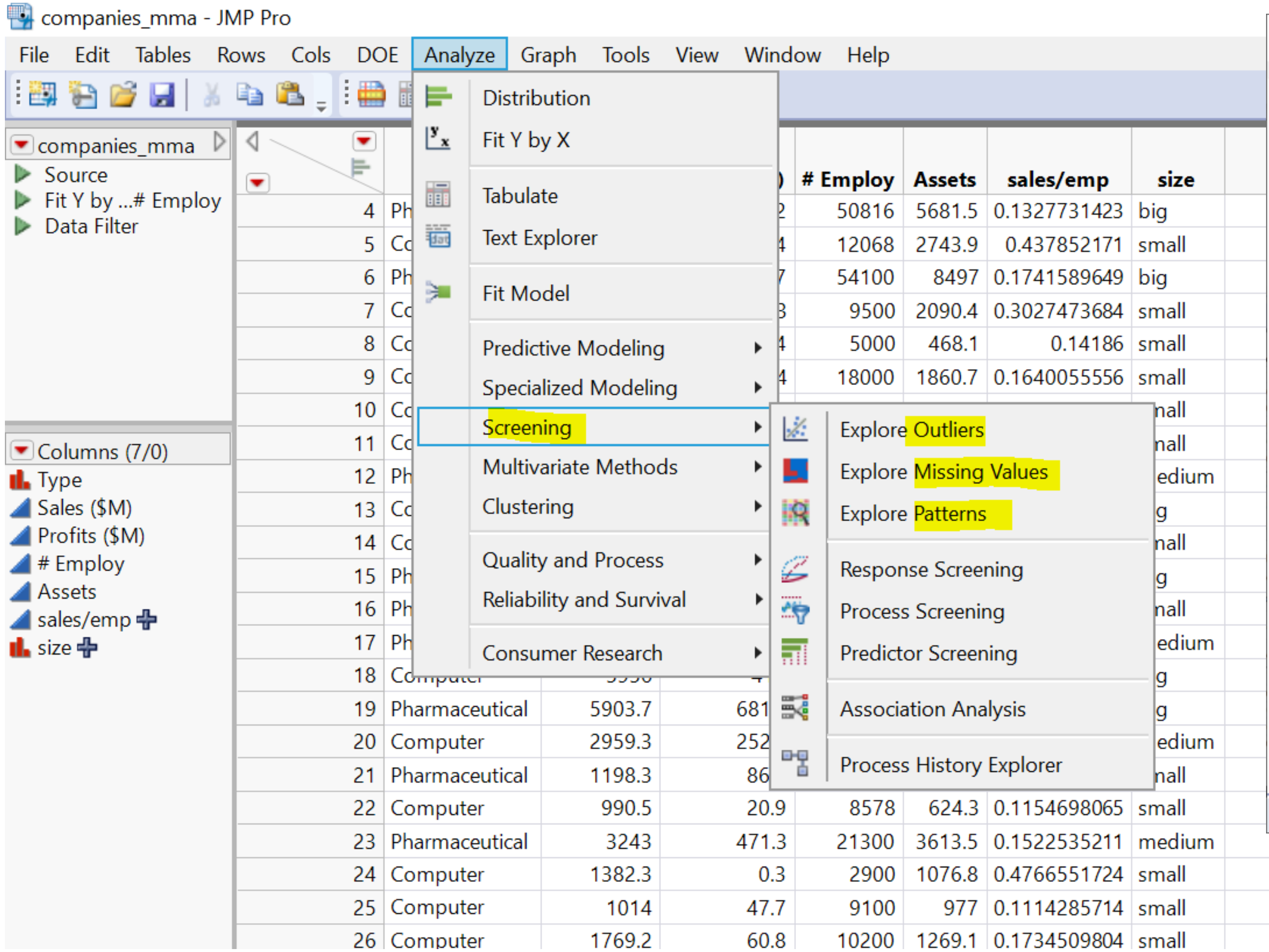
Your Turn (Hands-on)

- Import the country_risk.xlsx data (data/basics/country_risk.xlsx)
  - note that it’s an Excel file and column header starts at row 2

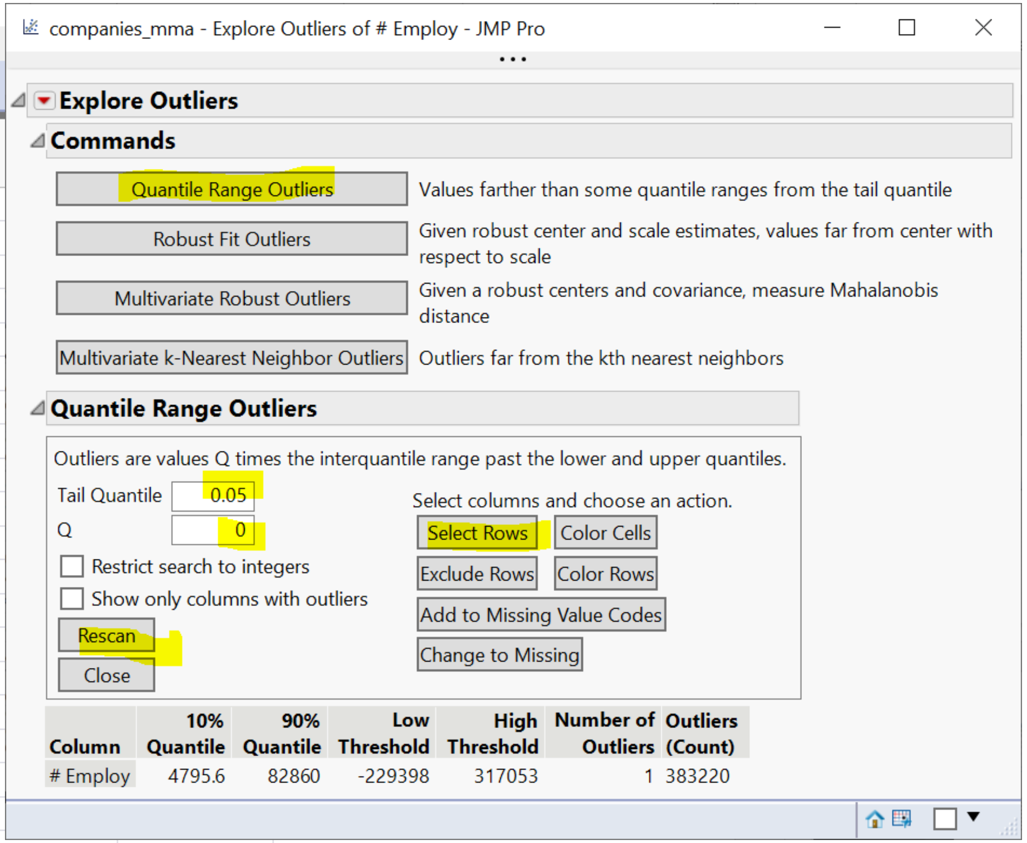
- Perform a pair-wise correlation analysis across the following 5 variables
  - Corruption, Peace, Legal, GDP Growth, Population
  - Note that Corruption and Legal variables are highly correlated
  - hint: use the Multivariate platform
    - Menu: ***Analyze -> Multivariate Methods - > Multivariate***

- Perform a K-means cluster analysis
  - As a start, use Peace, Legal and GDP Growth as factors; and set k=3
  - Produce a scatterplot matrix
  - Can you label each cluster (high-risk, medium-risk, etc.)?

Outliers, Missing Values, and Patterns



				# Employ	Assets	sales/emp	size
4	Ph		2	50816	5681.5	0.1327731423	big
5	Co		4	12068	2743.9	0.437852171	small
6	Ph		7	54100	8497	0.1741589649	big
7	Co		3	9500	2090.4	0.3027473684	small
8	Co		4	5000	468.1	0.14186	small
9	Co		4	18000	1860.7	0.1640055556	small
10	Co						nall
11	Co						nall
12	Ph						edium
13	Co						g
14	Co						nall
15	Ph						g
16	Ph						nall
17	Ph						edium
18	Computer		4				g
19	Pharmaceutical	5903.7	681				g
20	Computer	2959.3	252				edium
21	Pharmaceutical	1198.3	86				nall
22	Computer	990.5	20.9	8578	624.3	0.1154698065	small
23	Pharmaceutical	3243	471.3	21300	3613.5	0.1522535211	medium
24	Computer	1382.3	0.3	2900	1076.8	0.4766551724	small
25	Computer	1014	47.7	9100	977	0.1114285714	small
26	Computer	1769.2	60.8	10200	1269.1	0.1734509804	small



Explore Outliers

Commands

Quantile Range Outliers — Values farther than some quantile ranges from the tail quantile

Robust Fit Outliers — Given robust center and scale estimates, values far from center with respect to scale

Multivariate Robust Outliers — Given a robust centers and covariance, measure Mahalanobis distance

Multivariate k-Nearest Neighbor Outliers — Outliers far from the kth nearest neighbors

Quantile Range Outliers

Outliers are values Q times the interquantile range past the lower and upper quantiles.

Tail Quantile 0.05

Q 0

□ Restrict search to integers

□ Show only columns with outliers

Select columns and choose an action.

Column	10% Quantile	90% Quantile	Low Threshold	High Threshold	Number of Outliers	Outliers (Count)
# Employ	4795.6	82860	-229398	317053	1	383220

Note: I should perhaps put this slide somewhere between data manipulation and modelling

Plan

- Session 1
  - Workflow overview
  - Basic data manipulation

- Session 2
  - Join data tables
  - JMP graphing

- Session 3
  - Modelling
  - **JMP Journal**
  - **JMP Scripting Language**

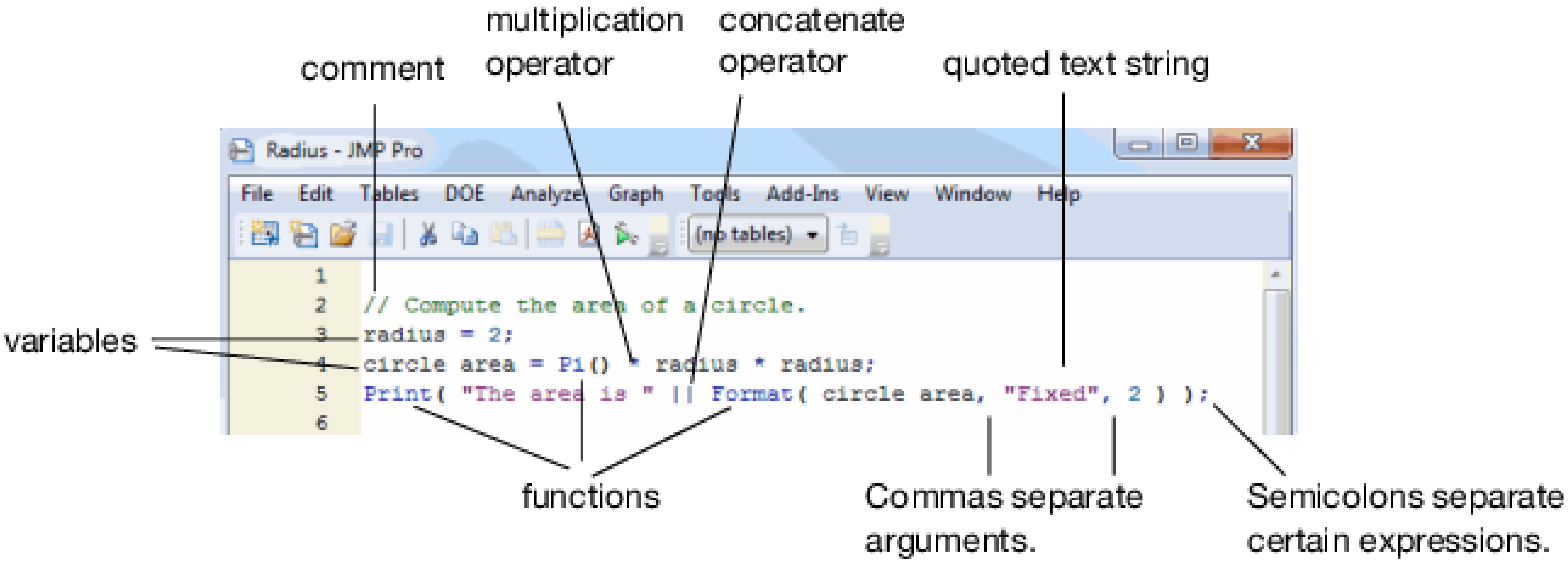
JMP Journal – Communicate Your Results

- Create a JMP journal when you want to present your results

- A JMP journal combine two kind of presentations
  - Static: embed output of JMP (graphs and reports), fixed at a moment in time
  - Dynamic: built from outlines containing text and buttons (links) that organize data tables and reports

- Getting-started resources
  - Dmitry’s video about JMP Journal on Quercus (6 mins)
  - Creating, Using and Sharing JMP Journals (45 mins)

JMP Script Language (JSL)



```
// Compute the area of a circle.
radius = 2;
circle area = Pi() * radius * radius;
Print( "The area is " || Format( circle area, "Fixed", 2 ) );
```

comment

multiplication operator

concatenate operator

quoted text string

variables

functions

Commas separate arguments.

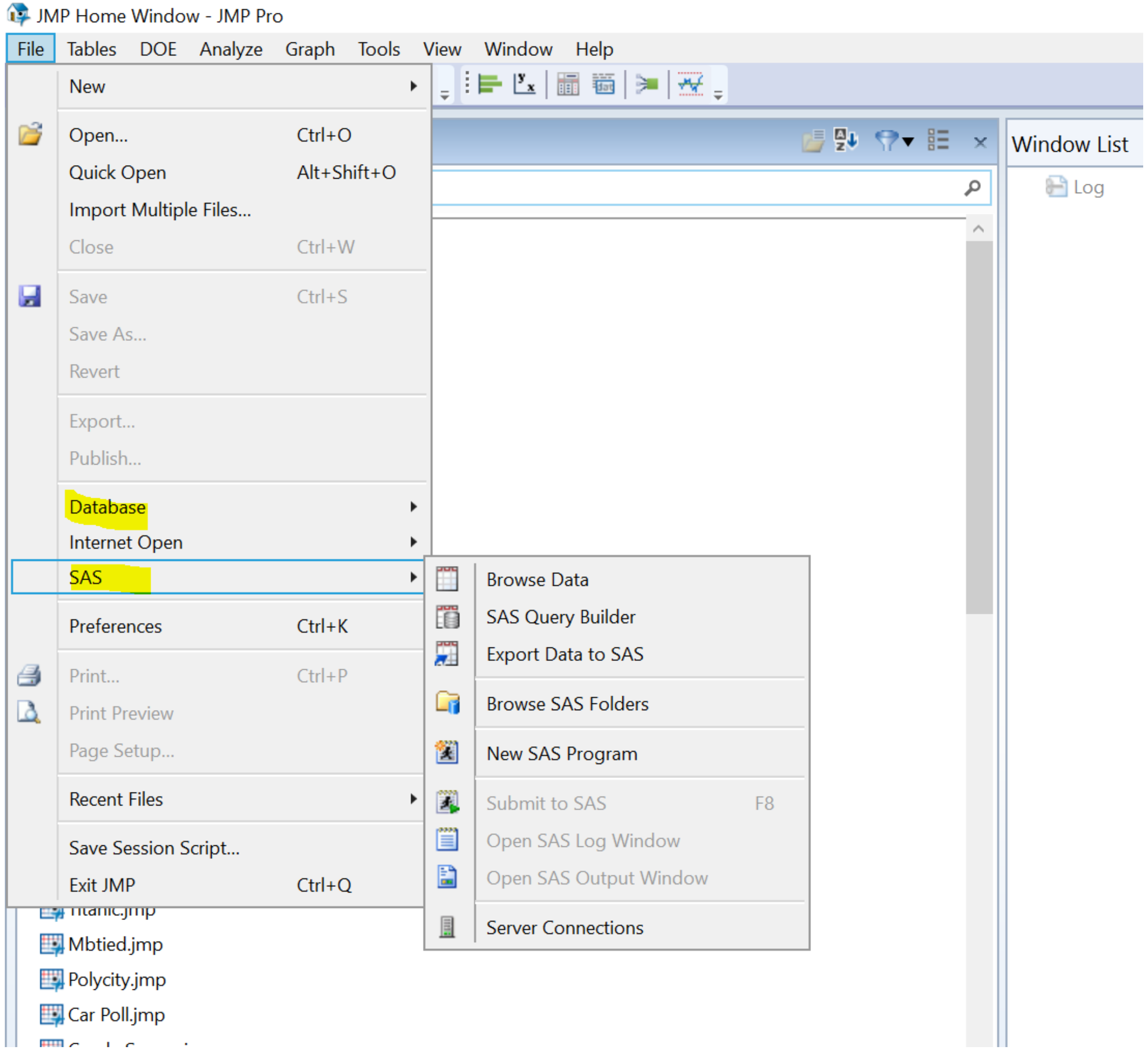
Semicolons separate certain expressions.

https://www.jmp.com/support/help/en/15.2/#page/jmp/jsl-building-blocks.shtml#

Connect to SAS and Database Systems

- SAS
  - Learning resource: Using SAS from JMP

- Database


JMP Home Window - JMP Pro
File
Tables
DOE
Analyze
Graph
Tools
View
Window
Help
New
Open... Ctrl+O
Quick Open Alt+Shift+O
Import Multiple Files...
Close Ctrl+W
Save Ctrl+S
Save As...
Revert
Export...
Publish...
Database
Internet Open
SAS
Preferences Ctrl+K
Print... Ctrl+P
Print Preview
Page Setup...
Recent Files
Save Session Script...
Exit JMP Ctrl+Q
Browse Data
SAS Query Builder
Export Data to SAS
Browse SAS Folders
New SAS Program
Submit to SAS F8
Open SAS Log Window
Open SAS Output Window
Server Connections
Window List
Log
Titanic.jmp
Mbtied.jmp
Polycity.jmp
Car Poll.jmp